GWC 2004

Second International WordNet Conference, GWC 2004
Brno, Czech Republic, January 20–23, 2004
Proceedings (CD-ROM version)

P. Sojka, K. Pala, P. Smrž, Ch. Fellbaum, P. Vossen (Eds.)

GWC 2004

Second International WordNet Conference, GWC 2004
Brno, Czech Republic, January 20–23, 2004
Proceedings

Masaryk University, Brno

Volume Editors

Petr Sojka
Faculty of Informatics, Masaryk University, Brno
Department of Programming Systems and Communication
Botanická 68a, CZ-602 00 Brno, Czech Republic
Email: sojka@informatics.muni.cz

Karel Pala, Pavel Smrž
Faculty of Informatics, Masaryk University, Brno
Department of Information Technologies
Botanická 68a, CZ-602 00 Brno, Czech Republic
Email: {pala,smrz}@informatics.muni.cz

Christiane Fellbaum
Department of Psychology, Princeton University
Princeton, New Jersey 08544, U.S.A.
Email: fellbaum@princeton.edu

Piek Vossen
Irion Technologies BV
Bagijnhof 80, P.O. Box 2849, 2601 CV Delft, The Netherlands
Email: Vossen@irion.nl

CATALOGUING IN PUBLICATION NATIONAL LIBRARY OF CZECH REPUBLIC

GWC 2004 (Brno, Česko)
GWC 2004 : second international WordNet conference : Brno, Czech Republic, January 20-23, 2004 : proceedings / P. Sojka ... [et al.] (eds.). – 1st ed. – Brno : Masaryk University, 2003. – xi, 359 s.
ISBN 80-210-3302-9

81’322 * 004.652:81’374
computational linguistics
lexical databases
proceedings of conferences

004.4/.6 Computer programming, programs, data
81 - Linguistics
ISBN 80-210-3302-9

Printed in the Czech Republic

Typesetting: Camera-ready by Petr Sojka from source files provided by authors. Data conversion by Petr Sojka and Aleš Horák.

Preface

A couple of years ago, the idea of an international meeting of the Global WordNet Association seemed truly daring and a bit fantastic. When it happened, two of us recall sitting on the excursion bus in Mysore, still overwhelmed at the number and quality of the submissions, the high attendance in the wake of September 11, and the outstanding local hospitality. Out of the blue, a participant across the aisle casually offered to "host the next meeting." Our spontaneous reaction was giddy laughter and the response that we had not even started to dream of a repeat.

Here we are exactly two years later, in a very different part of the world, and it seems like the kind of family meeting that everyone knows comes around inevitably in well-defined intervals and that is accepted unquestioningly. We are delighted that work on wordnets is being carried out in more countries and in an ever increasing number of languages. We cherish the common goals of developing wordnets, carrying out research, and building applications in a spirit of sharing that makes this community so special in a highly competitive world.

The Program Committee had a difficult job to select 34 oral and 16 poster presentations. Many people worked hard to make this conference successful. Our deep gratitude goes to the members of the Program Committee, local organizers, helpers and sponsors.

November 2003 Christiane Fellbaum

Organization

GWC 2004 was organized by the Faculty of Informatics, Masaryk University, in co-operation with the Global WordNet Association. The conference webpage is located at `http://www.fi.muni.cz/gwc2004/`

Program Committee

Pushpak Bhattacharya (IIT Mumbai, India)
Orhan Bilgin (Sabanci University, Istanbul, Turkey)
Paul Buitelaar (DFKI Saarbruecken, Germany)
Dan Cristea (University of Iasi, Romania)
Bento Carlos Dias da Silva (UNESP, Soa Paolo, Brasil)
Dominique Dutoit (University of Caen and Memodata, France)
Aleš Horák (Masaryk University, Brno, Czech Republic)
Chu-Ren Huang (Academica Sinica, Taipei, Republic of China)
Adam Kilgarriff (University of Brighton, England)
Karin Kipper (University of Pennsylvania, USA)
Claudia Kunze (Tuebingen University, Germany)
Bernardo Magnini (IRST, Trento, Italy)
Palmira Marrafa (University of Lisbon, Portugal)
Simonetta Montemagni (ILC-CNR, Pisa, Italy)
Grace Ngai (Polytechnical University, Hong Kong)
Karel Pala (Masaryk University, Brno, Czech Republic)
German Rigau (Basque Country University, Spain)
Pavel Smrž (Masaryk University, Brno, Czech Republic)
Sofi a Stamou (University of Patras, Greece)
Felisa Verdejo (U.N.E.D. Madrid, Spain)
Michael Zock (LIMSI-CNRS, Orsay, France)

Organizing Committee

Aleš Horák, Dana Komárková *(Secretary)*, Karel Pala *(Chair)*, Martin Povolný, Anna Sinopalnikova, Pavel Smrž, Petr Sojka *(Proceedings)*

Panel named *Figurative Language in WordNets and other Lexical Resources, and their Applications* is organized by Antonietta Alonge and Birte Lönneker.

Supported by:

Global WordNet Association

Table of Contents

I Figurative Language in WordNets and other Lexical Resources, and their Applications (Panel)

Metaphors in the (Mental) Lexicon 3
Christiane Fellbaum (Princeton University, USA)

Clustering of Word Senses 4
Eneko Agirre (University of the Basque Country, Donostia, Spain)

Sense Proximity versus Sense Relations 5
Julio Gonzalo (Universidad Nacional de Educación a Distancia, Madrid, Spain)

Implications of an AI Metaphor Understanding Project 7
John Barnden (University of Birmingham, UK)

Building and Extending Knowledge Fragments 8
Wim Peters (University of Sheffield, UK)

The *Heart* of the Problem: How Shall We Represent Metaphors in Wordnets? 10
Antonietta Alonge (Università di Perugia, Italy), Birte Lönneker (University of Hamburg, Germany)

Why WordNet Should Not Include Figurative Language, and What Would Be Done Instead 11
Patrick Hanks (Berlin-Brandenburg Academy of Sciences, Germany)

II Papers

Approximating Hierarchy-Based Similarity for WordNet Nominal Synsets using Topic Signatures 15
Eneko Agirre (University of the Basque Country, Donostia, Spain), Enrique Alfonseca (Universidad Autonoma de Madrid, Spain), Oier Lopez de Lacalle (University of the Basque Country, Donostia, Spain)

The MEANING Multilingual Central Repository 23
J. Atserias, L. Villarejo (Universitat Politècnica de Catalunya, Barcelona, Catalonia), G. Rigau, E. Agirre (University of the Basque Country, Donostia, Spain), J. Carroll (University of Sussex, UK), B. Magnini (ITC-irst, Trento, Italy), P. Vossen (Irion Technologies B.V., Delft, The Netherlands)

Russian WordNet 31
Valentina Balkova (Russicon Company, Russia), Andrey Sukhonogov (Petersburg Transport University, Moscow, Russia), Sergey Yablonsky (Petersburg Transport University, Moscow and Russicon Company, Russia)

ArchiWordNet: Integrating WordNet with Domain-Specifi c Knowledge 39
Luisa Bentivogli (ITC-irst, Trento, Italy), Andrea Bocco (Politecnico di Torino, Italy), Emanuele Pianta (ITC-irst, Trento, Italy)

Extending WordNet with Syntagmatic Information 47
Luisa Bentivogli, Emanuele Pianta (ITC-irst, Trento, Italy)

Exploiting ItalWordNet Taxonomies in a Question Classifi cation Task 54
Francesca Bertagna (Istituto di Linguistica Computazionale, CNR, Pisa, Italy)

Morphosemantic Relations In and Across Wordnets 60
Orhan Bilgin, Özlem Çetinoğlu, Kemal Oflazer (Sabanci University, Istanbul, Turkey)

A Prototype English-Arabic Dictionary Based on WordNet 67
William J. Black, Sabri El-Kateb (UMIST, Manchester, UK)

Automatic Assignment of Domain Labels to WordNet 75
Mauro Castillo, Francis Real (Universitat Politècnica de Catalunya, Barcelona, Spain), German Rigau (University of the Basque Country, Donostia, Spain)

Creation of English and Hindi Verb Hierarchies and their Application to Hindi WordNet Building and English-Hindi MT 83
Debasri Chakrabarti, Pushpak Bhattacharyya (Indian Institute of Technology, Mumbai, India)

Procedures and Problems in Korean-Chinese-Japanese Wordnet with Shared Semantic Hierarchy 91
Key-Sun Choi, Hee-Sook Bae (KORTERM, Republic of Korea)

Evaluating the Contribution of EuroWordNet and Word Sense Disambiguation to Cross-language Information Retrieval 97
Paul Clough, Mark Stevenson (University of Sheffield, United Kingdom)

The Topology of WordNet: Some Metrics 106
Ann Devitt, Carl Vogel (Trinity College, Dublin, Ireland)

Comparing Lexical Chain-based Summarisation Approaches Using an Extrinsic Evaluation 112
William Doran, Nicola Stokes, Joe Carthy, John Dunnion (University College Dublin, Ireland)

Use of Wordnet for Retrieving Words from Their Meanings 118
İlknur Durgar El-Kahlout, Kemal Oflazer (Sabanci University, Istanbul, Turkey)

Grounding the Ontology on the Semantic Interpretation Algorithm 124
Fernando Gomez (University of Central Florida, Orlando, USA)

Using a Lemmatizer to Support the Development and Validation of the Greek WordNet 130
Harry Kornilakis, Maria Grigoriadou (University of Athens, Greece), Eleni Galiotou (University of Athens and Technological Educational Institute of Athens, Greece), Evangelos Papakitsos (University of Athens, Greece)

VisDic – Wordnet Browsing and Editing Tool 136
Aleš Horák, Pavel Smrž (Masaryk University in Brno, Czech Republic)

A Corpus Based Approach to Near Synonymy of German Multi-Word Expressions . . . 142
Christiane Hümmer (Berlin-Brandenburg Academy of Sciences, Germany)

Using WordNets in Teaching Virtual Courses of Computational Linguistics 150
Lothar Lemnitzer, Claudia Kunze (Universität Tübingen, Germany)

A Current Resource and Future Perspectives for Enriching WordNets with Metaphor Information 157
Birte Lönneker, Carina Eilts (University of Hamburg, Germany)

Sociopolitical Domain As a Bridge from General Words to Terms of Specifi c Domains 163
Natalia Loukachevitch, Boris Dobrov (Moscow State University, Russia)

Using WordNet Predicates for Multilingual Named Entity Recognition 169
Matteo Negri, Bernardo Magnini (ITC-Irst, Trento, Italy)

Results and Evaluation of Hungarian Nominal WordNet v1.0 175
Márton Miháltz, Gábor Prószéky (MorphoLogic, Budapest, Hungary)

Corpus Based Validation of WordNet Using Frequency Parameters 181
Ivan Obradović, Cvetana Krstev, Gordana Pavlović-Lažetić, Duško Vitas (University of Belgrade, Serbia and Montenegro)

Language to Logic Translation with PhraseBank 187
Adam Pease (Articulate Software Inc, Mountain View, USA), Christiane Fellbaum (Princeton University, USA)

Extending the Italian WordNet with the Specialized Language of the Maritime Domain 193
Adriana Roventini, Rita Marinelli (Istituto di Linguistica Computazionale, CNR, Pisa, Italy)

Word Association Thesaurus As a Resource for Building WordNet 199
Anna Sinopalnikova (Masaryk University in Brno, Czech Republic and Saint-Petersburg State University, Russia)

Quality Control for Wordnet Development 206
Pavel Smrž (Masaryk University in Brno, Czech Republic)

Extension of the SpanishWordNet 213
Clara Soler (Universitat Ramon Llull, Barcelona, Spain)

Pathways to Creativity in Lexical Ontologies 220
Tony Veale (University College Dublin, Ireland)

Automatic Lexicon Generation through WordNet 226
Nitin Verma, Pushpak Bhattacharyya (Indian Institute of Technology, Bombay, India)

Fighting Arbitrariness in WordNet-like Lexical Databases– A Natural Language Motivated Remedy 234
Shun Ha Sylvia Wong (Aston University, Birmingham, UK)

Finding High-Frequent Synonyms of A Domain-Specifi c Verb in English Sub-Language of MEDLINE Abstracts Using WordNet 242
Chun Xiao, Dietmar Rösner (Universität Magdeburg, Germany)

III Posters

Adjectives in RussNet 251
Irina Azarova (Saint-Petersburg State University, Russia), Anna Sinopalnikova (Saint-Petersburg State University, Russia and Masaryk University, Brno, Czech Republic)

Towards Binding Spanish Senses to Wordnet Senses through Taxonomy Alignment . . . 259
Javier Farreres, Karina Gibert, Horacio Rodríguez (Universitat Politècnica de Catalunya, Barcelona, Spain)

WordNet Exploitation through a Distributed Network of Servers 265
I. D. Koutsoubos (Patras University and Research Academic Computer Technology Institute, Patras, Greece), Vassilis Andrikopoulos (Patras University, Greece), Dimitris Christodoulakis (Patras University and Research Academic Computer Technology Institute, Patras, Greece)

WordNet Applications 270
Jorge Morato, Miguel Ángel Marzal, Juan Lloréns, José Moreiro (University Carlos III, Madrid, Spain)

Extending and Enriching WordNet with OntoLearn 279
Roberto Navigli, Paola Velardi (Università di Roma "La Sapienza", Italy), Alessandro Cucchiarelli, Francesca Neri (Università Politecnica delle Marche, Ancona, Italy)

Concerning the Difference Between a Conception and its Application in the Case of the Estonian WordNet 285
Heili Orav, Kadri Vider (University of Tartu, Estonia)

Soft Word Sense Disambiguation 291
Ganesh Ramakrishnan, B. P. Prithviraj, A. Deepa, Pushpak Bhattacharya, Soumen Chakrabarti (Indian Institute of Technology, Bombay, India)

Text Categorization and Information Retrieval Using WordNet Senses 299
Paolo Rosso (Polytechnic University of Valencia, Spain), Edgardo Ferretti (National University of San Luis, Argentina), Daniel Jiménez, Vicente Vidal (Polytechnic University of Valencia, Spain)

Jur-WordNet 305
Maria Teresa Sagri, Daniela Tiscornia (Institute for Theory and Techniques for Legal Information, CNR, Firenze, Italy), Francesca Bertagna (Istituto di Linguistica Computazionale, CNR, Pisa, Italy)

WordNet Has No 'Recycle Bin' 311
B. A. Sharada, P. M. Girish (Central Institute of Indian Languages, Mysore, India)

Automatic Word Sense Clustering Using Collocation for Sense Adaptation 320
Sa-Im Shin, Key-Sun Choi (KORTERM, KAIST, Daejeon, Korea)

WordNet for Lexical Cohesion Analysis 326
Elke Teich (Darmstadt University of Technology, Germany), Peter Fankhauser (Fraunhofer IPSI, Darmstadt, Germany)

Cross-Lingual Validation of Multilingual Wordnets 332
Dan Tufiş, Radu Ion, Eduard Barbu, Verginica Barbu (Institute for Artificial Intelligence, Bucharest, Romania)

Roles: One Dead Armadillo on WordNet's Speedway to Ontology 341
Martin Trautwein, Pierre Grenon (University of Leipzig, Germany)

Two Kinds of Hypernymy Faults in WordNet: the Cases of Ring and Isolator 347
Yang Liu, Jiangsheng Yu, Zhengshan Wen, Shiwen Yu (Peking University, China)

Statistical Overview of WordNet from 1.6 to 2.0 352
Jiangsheng Yu, Zhenshan Wen, Yang Liu, Zhihui Jin (Peking University, China)

Author Index 359

Part I

Figurative Language in WordNets and other Lexical Resources, and their Applications (Panel)

Metaphors in the (Mental) Lexicon

Christiane Fellbaum

Berlin-Brandenburg Academy of Sciences, Berlin, Germany
Princeton University, Princeton, New Jersey, USA
Email: fellbaum@princeton.edu

In this presentation, metaphors are defined as simple lexemes rather than phrases, specifically verbs and nouns. Dictionaries can treat these straightforwardly as cases of polysemy. For example, the entry for "tiger" may contain two senses, one referring to the wild cat, the other to a fierce person. The metaphoricity of the second sense need to not be noted, thus making entries for words like "tiger" indistinguishable from the entries for other polysemous words like "bank." Because of its particular design, WordNet makes it possible to detect many – though not all – cases of metaphoric extensions and to distinguish them from ordinary polysemy [1].

Dictionaries contain conventionalized metaphors (like "tiger" in the sense of fierce person), but cannot include spontaneously generated ad-hoc metaphors, such as when someone refers to her place of work as a "jail" ([2], inter alia). These metaphors are not only created by language users on the fly but also present no comprehension problems despite the fact that they are not represented in speakers' mental lexicons.

Both conventionalized and ad-hoc metaphors depend crucially on the exploitation of semantic similarity and analogy. I discuss the nature of metaphors in terms of semantic similarity as represented in WordNet, and argue that WordNet has the potential to account successfully for the phenomenon of ad-hoc metaphor. Relevant preliminary results of an empirical study of association and evocation among WordNet lexemes will be presented.

References

1. Fellbaum, C.: Towards a Representation of Idioms in WordNet. In Harabagiu, S., ed.: Proceedings of the Workshop on Usage of WordNet in Natural Language Processing Systems, COLING/ACL 1998, Montreal, CA. (1998) 52–57.
2. Glucksberg, S., Keysar, B.: Understanding metaphorical comparisons: Beyond similarity. Psychological Review **97** (1990) 3–18.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 3–3.

Clustering of Word Senses

Eneko Agirre

University of the Basque Country, Donostia 20.080, Spain,
Email: eneko@si.ehu.es, WWW: http://ixa.si.ehu.es

WordNet does not provide any information about the relation among the word senses of a given word, that is, the word senses are given as a flat list. Some dictionaries provide an abstraction hierarchy, and previous work has tried to find systematic polysemy relations [3] using the hierarchies in WordNet.

In [1,2] we apply distributional similarity methods to word senses, in order to build hierarchical clusters for the word senses of a word. The method uses the information in WordNet (monosemous relatives) in order to collect examples of word senses from the web. In the absence of hand-tagged data, those examples constitute the context of each word sense. The contexts are modeled into vectors using different weighting functions, e.g. χ^2 or $tf \cdot idf$. The similarity between the word senses can thus be obtained using any similarity function, e.g. the cosine. Once we have a similarity matrix for the word senses of a given word, clustering techniques are applied in order to obtain a hierarchical cluster.

The evaluation shows that our hierarchical clusters are able to approximate the manual sense groupings for the nouns in Senseval 2 with purity values of 84%, comparing favorably to using directly the hand-tagged data available in Senseval 2 (purity of 80%). The results are better than those attained by other techniques like confusion matrixes from Senseval 2 participating systems or multilingual similarity.

The primary goal of our work is to tackle the fine-grainedness and lack of structure of WordNet word senses, and we will be using the clusters to improve Word Sense Disambiguation results. We plan to make this resource publicly available for all WordNet nominal word senses, and we expect for the similarity measure to be valuable in better acquiring the explicit relations among WordNet word senses, including specialization, systematic polysemy and metaphorical relations.

References

1. E. Agirre and O. Lopez de Lacalle. Clustering wordnet word senses. In *Proceedings of the Conference on Recent Advances on Natural Language Processing (RANLP '03)*. Borovets, Bulgary. 2003.
2. E. Agirre, E. Alfonseca and O. Lopez de Lacalle Approximating hierarchy-based similarity for WordNet nominal synsets using Topic Signatures. In: P. Sojka et al. (Eds.): *Proceedings of the 2nd Global WordNet Conference*. Brno, Czech Republic. 2004.
3. W. Peters and I. Peters. Lexicalized systematic polysemy in WordNet. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*. Athens, Greece. 2000.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 4–4.

Sense Proximity versus Sense Relations

Julio Gonzalo

Departamento de Lenguajes y Sistemas Informáticos
Universidad Nacional de Educación a Distancia
E.T.S.I Industriales, Ciudad Universitaria s/n, 28040 Madrid, SPAIN
Email: julio@lsi.uned.es WWW: http://sensei.lsi.uned.es/~julio/

It has been widely assumed that sense distinctions in WordNet are often too fine-grained for applications such as Machine Translation, Information Retrieval, Text Classification, Document clustering, Question Answering, etc. This has led to a number of studies in sense clustering, i.e., collapsing sense distinctions in WordNet that can be ignored for most practical applications [1,5,6]. At the UNED NLP group, we have also conducted a few experiments in sense clustering with the goal of improving WordNet for Information Retrieval and related applications [4,3,2].

Our experiments led us to the conclusion that annotating WordNet with a typology of polysemy relations is more helpful than forming sense clusters based on a notion of sense proximity. The reason is that sense proximity depends on the application, and in many cases can be derived from the type of relation between two senses. In the case of metaphors, senses often belong to different semantic fields, and therefore a metaphor can be a relevant distinction for Information Retrieval or Question & Answer systems. For Machine Translation applications, however, the metaphoric sense extensions might be kept across languages, and therefore the distinction might not be necessary to achieve a proper translation.

In the panel presentation, we will summarize the experiments that led us to hold this position:

- In [2] we compared two clustering criteria: the first criterion, meant for Information Retrieval applications, consists of grouping senses that tend to co-occur in Semcor documents. The second criterion, inspired by [7], groups senses that tend to receive the same translation in several target languages via the EuroWordNet Interlingual Index (*parallel polysemy*). The overlapping of both criteria was between 55% and 60%, which reveals a correlation between both criteria but leaves doubts about the usefulness of the clusters. However, a classification of the sense groupings according to the type of polysemy relation clarifies the data: all homonym and metaphor pairs satisfying the parallel polysemy criterion did not satisfy the co-occurrence criterion; all generalization/specialization pairs did satisfy the co-occurrence criterion; finally, metonymy pairs were evenly distributed between valid and invalid co-occurrence clusters. Further inspection revealed that the type of metonymic relation could be used to predict sense clusters for Information Retrieval.
- In [3] we applied Resnik & Yarowsky measure to evaluate the Senseval-2 WordNet subset for sense granularity. We found that the average proximity was similar to the Senseval-1 sense inventory (Hector), questioning the idea that WordNet sense distinctions are finer than in other resources built by lexicographers. We also found that Resnik & Yarowsky proximity measure provides valuable information, but should

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 5–6.

be complemented with information about polysemy relations. There are, for instance, a significant fraction of homonyms that receive a non-null proximity measure, and the average proximity for metaphoric pairs is higher than would be expected for sense pairs belonging to different semantic fields. We believe that the classification of such pairs as homonyms is more valuable and has more predictive power than the quantitative measure of proximity.

In WordNet, the different senses of a word can be implicitly connected through the semantic relations between synsets. But these connections are too vague to understand the relations holding between senses: for instance, it is hard to decide when two senses of a word are homonyms, an information that is essential for Language Engineering applications, and can be found in other, more conventional lexicographic resources. We believe that, to achieve the full potential of wordnets as a de facto standard for lexical resources in computational applications, the relations between senses of polysemous words should be explicitly annotated. In the panel discussion, we will briefly discuss a proposal for a simple typology of polysemy relations, and the exploratory annotation of the senses for a thousand nouns in WordNet using this typology.

References

1. E. Agirre and O. Lopez de Lacalle. Clustering Wordnet word senses. In *Proceedings RANLP 2003*, 2003.
2. I. Chugur, J. Gonzalo, and F. Verdejo. A study of sense clustering criteria for information retrieval applications. In *Proceedings of Ontolex 2000*, 2000.
3. I. Chugur, J. Gonzalo, and F. Verdejo. Polysemy and sense proximity in the senseval-2 test suite. In *Proceedings of the ACL-2002 Workshop on "Word Sense Disambiguation: recent successes and future directions"*, 2002.
4. Julio Gonzalo, Irina Chugur, and Felisa Verdejo. Sense clustering for information retrieval: evidence from Semcor and the EWN InterLingual Index. In *Proceedings of the ACL'00 Workshop on Word Senses and Multilinguality*, 2000.
5. R. Mihalcea and D. Moldovan. Automatic generation of a coarse grained wornet. In *Proceedings of NAACL Worskhop on WordNet and Other LExical Resources*, 2001.
6. W. Peters and I. Peters. Automatic sense clustering in EuroWordnet. In *Proceedings of LREC'2000*, 2000.
7. P. Resnik and D. Yarowsky. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering*, 1999.

Implications of an AI Metaphor Understanding Project

John Barnden

School of Computer Science, University of Birmingham, UK
Email: J.A.Barnden@cs.bham.ac.uk

I shall explore the implications for lexical resources of my work on ATT-Meta, a reasoning system designed to work out the significance of a broad class of metaphorical utterances. This class includes "map-transcending" utterances, resting on familiar, general conceptual metaphors but go beyond them by including source- domain elements that are not handled by the mappings in those metaphors.

The system relies heavily on doing reasoning within the terms of the source domain rather than trying to construct new mapping relationships to handle the unmapped source-domain elements. The approach would therefore favour the use of WordNet-like resources that facilitate rich within-domain reasoning and the retrieval of known cross-domain mappings without being constrained to facilitate the creation of new mappings. The approach also seeks to get by with a small number of very general mappings per conceptual metaphor.

The research has also led me to a radical language-user-relative view of metaphor. The question of whether an utterance is metaphorical, what conceptual metaphors it involves, what mappings those metaphors involve, what word-senses are recorded in a lexicon, etc. are all relative to specific language users and shouldn't be regarded as something we have to make objective decisions about. This favours a practical approach where natural language applications can differ widely on how they handle the same potentially metaphorical utterance because of differences in lexical resources used.

The user-relativity is also friendly to a view where the presence of a word-sense in a lexicon has little to do with whether that sense is figurative or not. This stance is related to, Patrick Hanks' view that we should focus on norms and exploitations rather than on figurativity.

The research has furthermore led me to a deep scepticism about the ability to rely in definitions of metaphor on qualitative differences between domains. Scepticism about domains then causes additional difficulty in distinguishing between metaphor and metonymy. At the panel I will outline a particular view of the distinction.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 7–7.

Building and Extending Knowledge Fragments

Wim Peters

University of Sheffield
United Kingdom
Email: w.peters@dcs.shef.ac.uk

In this panel presentation I will contend that it is possible to extract knowledge fragments from WordNet [1] and EuroWordNet [2] that combine explicit knowledge structures already provided by the thesauri such as synonymy, hypernymy and thematic relations, and implicit information from the (Euro)WordNet's hierarchical structure and the glosses that are associated with each WordNet synset.

The initial emphasis of the work lies on the detection of patterns of figurative language use, more particularly cases of regular polysemy [3].

The work consists of three phases. First, an automatic selection process identifies candidates for instantiations of regular polysemy [4,5] in WordNet on the basis of systematic sense distributions of nouns. These systematic distributions can be characterized by a pair of hypernyms taken from the WordNet hierarchies that subsume the senses. For instance, in two of its senses 'law' falls under the pattern *profession* (an occupation requiring special education) and *discipline* (a branch of knowledge). This set of conventionalised/lexicalised figurative language use forms the basis of the building of knowledge fragments.

In the second stage, the underspecified relations that exist between the word senses that participate in patterns are further specified in an automatic fashion. This additional information is obtained by analyzing the glosses that are associated with the synsets of the word senses involved and their hypernyms. For example, the extracted pattern **person** (a human being; "there was too much for one person to do") and **language** (a systematic means of communicating by the use of sounds or conventional symbols) subsumes sense pairs of 257 words in WordNet such as *Tatar*, *Assyrian*, *Hopi*, and *Punjabi*. The analysis of the WordNet glosses yields 'speak' as a significant relation.

This explicit knowledge that can be gleaned from information implicit in glosses enriches the already existing knowledge structures of WordNet, thereby expanding its coverage as a knowledge base. Also, it forms the start of the explicit encoding of metonymic potential of words where they do not yet participate in the patterns.

In the third phase, increasingly larger knowledge frames are built up on the basis of these sense pairs. The relation triples extracted in the second stage (e.g. **person-speak-language**) form the basic building blocks of the frames.

Extension of these rudimentary frames takes place in two ways. First, the concept with which hypernyms from the regular polysemy patterns co-occur can be regarded as additional slots in a topical frame that characterizes a hypernym. For instance, the pattern **music-dance** covers words such as tango and bolero. **Music** in its turn co-occurs with a number of other concepts within the hypernym pairs that characterize the regular polysemic patterns. These concepts and the relations that have been extracted between these hypernyms form a further extension of the **music** frame.

A further extension takes the semantic context of EuroWordNet into account. From the superset of all concepts and relations that are linked to MUSIC in all eight language specific

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 8–9.

wordnets the MUSIC frame is extended with this new knowledge. The resulting structure is an extended knowledge frame that, amongst others, contains the following slot fillers and relations: **person-make/accomplish-music; musician isa person; musician play music; music-accompany-activity; dancing isa activity**.

These knowledge frames can be extended with information from other resources, and be used in a variety of applications.

References

1. Fellbaum, C., ed.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge, Mass. (1998).
2. Vossen, P.: Introduction to EuroWordNet. Computers and the Humanities (Special Issue on EuroWordNet) **32** (1998) 73–89.
3. Apresjan, J.: Regular Polysemy. Linguistics **142** (1973) 5–32.
4. Peters, W., Peters, I.: Lexicalised Systematic Polysemy in WordNet. In: Proceedings of the Second International Conference on Language Resources and Evaluation, Athens, Greece. (2000).
5. Peters, W., Wilks, Y.: Distribution-oriented Extension of WordNet's Ontological Framework. In: Proceedings RANLP-2001, Tzigov Chark, Bulgaria. (2001).

The *Heart* of the Problem: How Shall We Represent Metaphors in Wordnets?

Antonietta Alonge[1] and Birte Lönneker[2]

[1] Sezione di Linguistica, Facoltà di Lettere e Filosofia, Università di Perugia
Piazza Morlacchi, 11, Perugia 06100, Italy
Email: anto.alonge@unipg.it
[2] Institute for Romance Languages, University of Hamburg
Von-Melle-Park 6, 20146 Hamburg, Germany
Email: birte.loenneker@uni-hamburg.de

Motivated by the limits of EWN with respect to the treatment of metaphor and the consequences on the use of the database for WSD, we address the issue of the encoding of information on metaphors in wordnets. We assume as a starting point the theory of metaphor as a cognitive rather than a linguistic phenomenon, as proposed by [1] and [2]. According to this theory, metaphoric linguistic expressions are manifestations of 'conceptual metaphors', i.e. metaphorical structures which are present in our minds and relate a concrete *source domain* with a more abstract *target domain*. The adoption of this theoretical framework allows us to envisage devices to encode data both on conventional, well-established metaphoric expressions and on potential, novel metaphoric uses of words. We state that **1)** more information has to be encoded at the synset level, with the aims of confronting the lack of consistency and completeness of the database, and of adding data on sense relatedness, by means of a specifically defined new internal-relation (i.e., a new relation linking synsets within each language-specific wordnet); **2)** at a higher level, language-specific wordnets have to be linked to the ILI in a way that accounts for mappings between conceptual domains resulting in potential new metaphoric expressions. We thus propose to add an EQ_METAPHOR relation, pointing to new composite ILI units to account for regular metaphoric extensions of senses in EWN. Via the ILI links, the connection between specific synsets in a language would also be shown at the Top Ontology (TO) level as a connection (mapping) between top concepts (linked to different conceptual domains). On the other hand, the composite ILIs and the mappings at the TO level could be used to infer which words might *potentially* display a certain metaphorical sense extension, as this information can be derived through inheritance along taxonomies. Taking as a starting point domain-centered data from the Hamburg Metaphor Database, we discuss also non-taxonomic ways of "spreading" the information about potential metaphorical senses.

References

1. Lakoff, G., Johnson, M.: Metaphors we live by. UCP, Chicago/London (1980).
2. Lakoff, G.: The contemporary theory of metaphor. In Ortony, A., ed.: Metaphor and Thought. CUP, Cambridge (1993) 202–251.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 10–10.

Why WordNet Should Not Include Figurative Language, and What Would Be Done Instead

Patrick Hanks

[1] Berlin-Brandenburg Academy of Sciences
[2] Brandeis University
Email: hanks@bbaw.de

I shall argue that figurative language has no place in WordNets, FrameNets, MRDs, or any other lexical resource. Lexical resources should list norms of language use, not dynamic exploitations of norms. Interpreting figurative language should be achieved by other means. However, first we have to be quite clear about what we mean by 'figurative language'.

Confusion arises because so many norms of language use are of figurative origin. For example, *object* (in all literal senses of the modern word) originated as a Latin metaphor: 'something thrown in the way'. The literal meaning of *subject* in Latin is 'something thrown under'. *Ardent* feelings are literally burning feelings. Ouch! Lakoff and Johnson have shown other ways in which many of our most literal uses of language have metaphorical origins or associations.

The commonly made distinction between figurative and literal meaning is a red herring. It is not useful. Much more important is the distinction that can be made between conventional language use and dynamic language use, i.e. between norms and exploitations.

I shall discuss, using corpus evidence, examples such as the following:

- *keep one's head above water* (literally and figuratively);
- *a geometrical proof is a mousetrap* (Schopenhauer);
- *the "mousetrap" in American football*;
- *hazard a guess, hazard a destination*;
- *worm, virus*.

I shall look at the treatment of these words and expressions in WordNet and suggest possible improvements. I argue that language in use consists of uses of words that are either norms or exploitations. Until we know how to recognize a norm (astonishingly, we don't), there is not much point in talking about how to process exploitations, such as figurative language. The first priority, therefore, is to provide recognition criteria for norms of word usage. The norms can, in principle, be associated with WordNet entries, but a great deal of corpus pattern analysis is needed.

Why don't we know how to recognize a norm? In part because we still yearn for necessary conditions. In lexical semantics there are no necessary conditions. It is time to take seriously the proposal of Fillmore (1975), that the meaning of a word in a text should be interpreted by measuring similarity to a prototype. Fillmore's proposal is currently being implemented as FrameNet, with its focus on semantic frames. The Theory of Norms and Exploitations (TNE; Hanks, forthcoming) differs from FrameNet in that it provides syntagmatic criteria for normal uses of individual words (to which meanings, synsets, translations, frame roles, etc. can be attached).

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 11–11.

Part II

Papers

Approximating Hierarchy-Based Similarity for WordNet Nominal Synsets using Topic Signatures

Eneko Agirre[1], Enrique Alfonseca[2], and Oier Lopez de Lacalle[1]

[1] University of the Basque Country, Donostia 20.080, Spain,
Email: eneko@si.ehu.es, jibloleo@si.ehu.es, WWW: http://ixa.si.ehu.es
[2] Universidad Autonoma de Madrid,
Email: enrique.alfonseca@ii.uam.es, WWW: http://www.ii.uam.es/~ealfon

Abstract. Topic signatures are context vectors built for concepts. They can be automatically acquired for any concept hierarchy using simple methods. This paper explores the correlation between a distributional-based semantic similarity based on topic signatures and several hierarchy-based similarities. We show that topic signatures can be used to approximate link distance in WordNet (0.88 correlation), which allows for various applications, e.g. classifying new concepts in existing hierarchies. We have evaluated two methods for building topic sigantures (monosemous relatives vs. all relatives) and explore a number of different parameters for both methods.

1 Introduction

Knowledge acquisition is a long-standing problem in both Artificial Intelligence and Natural Language Processing (NLP). Huge efforts and investments have been made to manually build repositories with semantic and pragmatic knowledge but with unclear results. Complementary to this, methods to induce and enrich existing repositories have been explored (see [1] for a recent review).

In previous work we have shown that it is possible to enrich WordNet synsets [2] with topic signatures. Topic signatures try to associate a topical vector to each word sense. The dimensions of this topical vector are the words in the vocabulary and the weights try to capture the relatedness of the words to the the target word sense. In other words, each word sense is associated with a set of related words with associated weights. Figure 1 shows sample topic signatures for the word senses of church. Several of the topic signatures used in this paper can be found in http://ixa3.si.ehu.es/cgi-bin/signatureak/signaturecgi.cgi in its full version.

Topic signatures for words have been successfully used in summarisation tasks [3]. Regarding topic signatures for word senses, [4,5] show that it is possible to obtain good quality topic signatures for word senses automatically. [6] show that topic signatures for word senses can be used for extending WordNet's taxonomy, and [7] show that they are effective for clustering WordNet word senses.

In this paper we compare similarity measures for WordNet concepts based on topic signatures with measures based on the hierarchy WordNet (see [8,9] for recent references). The advantage of topic signatures over the similarity measures based on the hiearchy of

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 15–22.

1st. sense: **church, Christian_church, Christianity** *"a group of Christians; any group professing Christian doctrine or belief;"*
size church(1177.83) catholic(700.28) orthodox(462.17) roman(353.04) religion(252.61) byzantine(229.15) protestant(214.35) rome(212.15) western(169.71) established(161.26) coptic(148.83) jewish(146.82) order(133.23) sect(127.85) old(86.11) greek(68.65) century(61.99) history(50.36) pentecostal(50.18) england(44.77) saint(40.23) america(40.14) holy(35.98) pope(32.87) priest(29.76) russian(29.75) culture(28.43) christianity(27.87) religious(27.10) reformation(25.39) ukrainian(23.20) mary(22.86) belong(21.83) bishop(21.57) anglican(18.19) rite(18.16) teaching(16.50) christian(15.57) diocese(15.44)

2nd. sense: **church, church_building** *"a place for public (especially Christian) worship;"*
house(1733.29) worship(1079.19) building(620.77) mosque(529.07) place(507.32) synagogue(428.20) god(408.52) kirk(368.82) build(93.17) construction(47.62) street(47.18) nation(41.16) road(40.12) congregation(39.74) muslim(37.17) list(34.19) construct(31.74) welcome(29.23) new(28.94) prayer(24.48) temple(24.40) design(24.25) brick(24.24) erect(23.85) door(20.07) heaven(19.72) plan(18.26) call(17.99) renovation(17.78) mile(17.63) gate(17.09) architect(16.86) conservative(16.46) situate(16.46) site(16.37) demolition(16.16) quaker(15.99) fort(14.59) arson(12.93) sultan(12.93) community(12.88) hill(12.62)

3rd. sense: **church_service, church** *"a service conducted in a church;"*
service(5225.65) chapel(1058.77) divine(718.75) prayer(543.96) hold(288.08) cemetery(284.48) meeting(271.04) funeral(266.05) sunday(256.46) morning(169.38) attend(143.64) pm(133.56) meet(115.86) conduct(98.96) wednesday(90.13) religious(89.19) evening(75.01) day(74.45) friday(73.17) eve(70.01) monday(67.96) cremation(64.73) saturday(60.46) thursday(60.46) june(57.78) tuesday(56.08) crematorium(55.53) weekly(53.36) procession(50.53) burial(48.60) december(48.46) ceremony(46.47) september(46.10) interment(42.31) lead(38.79) family(34.19) deceased(31.73) visitation(31.44)

Fig. 1. Fragment of the topic signatures for the three senses of church built with the monosemous relatives method to extract examples from the Web. The values in parenthesis correspond to χ^2 values. Only the top scoring terms are shown.

WordNet is that they can be applied to unknown concepts, and thus allow for classifying new concepts. We also compare the impact of different ways of acquiring and modeling topic signatures.

The paper is structured as follows. Section 2 presents the method to construct topic signatures, alongside some parameters for the construction of them. In section 3 we review different methods to compute the similarity between topic signatures. Section 4 presents the experimental setting and Section 5 presents the results. Finally, Section 6 presents the conclusions and future work.

2 Construction of Topic Signatures

Two main alternatives for the construction of topic signatures have been presented in [4,5,6], which will be presented briefly in this section. Please refer to those papers for further details. The first step consists on acquiring examples for the target word senses. The idea is to use the information in WordNet in order to build appropriate queries, which are used to search in the Internet those texts related to the given word sense. The second step organizes the examples thus retrieved in document collections, one collection for each word sense. In the third step, we extract the words in each of the collections and their frequencies, and compare them with the data in the other collections. Finally, The words that have a distinctive frequency for one of the collections are collected in a list, which constitutes the topic signature for each word sense. The steps are further explained below.

2.1 Acquiring the Examples and Building the Document Collections

In order to retrieve documents that are associated to a word sense, [4,5,6] present different strategies to build the queries. Some of the methods that were used have problems to scale-up, as they require certain amount of hand correction, so we propose to use two simple methods to build queries:

1. use all relatives (synonyms, hyponyms, children, siblings) of the target word sense;
2. use only those relatives of the target word sense that are monosemous.

One can argue that the first method, due to the polysemy of the relatives, can gather examples of relatives which are not really related to the target word sense. In principle, the second method avoids this problem and should provide better examples.

In the current implementation 1) was performed retrieving up to 100 documents from Altavista, and extracting from them the sentences which contain any of the synset words; and 2) was performed retrieving up to 1000 sentences for each monosemous relative from Google snippets.

2.2 Representing Context

In order to model the retrieved examples we can treat the context as a bag of words, that is, all the words in the context are used in flat vector. In this case we build a vector of V dimensions (where V is the size of the vocabulary), where the words occurring in the contexts are the keys and their frequency the values. All the words are first lemmatized.

2.3 Weighting the Words in Context

Frequencies are not good indicators of relevancy, so different functions can be used in order to measure the relevance of each term appearing in the vector corresponding to one sense in contrast to the others. That is, terms occurring frequently with one sense, but not with others, are assigned large weights for the associated word sense, and low values for the rest of word senses. Terms occurring evenly among all word senses are also assigned low weights for all

the word senses. We have currently implemented five measures: two versions of tf.idf[1], χ^2, mutual information and t-score.

The topic signatures are vectors where the words have weights corresponding to the relevancy functions thus computed.

2.4 Filtering

In [5] it is shown that weighting functions can assign very high weights to rare terms appearing in the context of one of the word senses by chance. This effect can be reduced in the following way: we collect contexts of occurrences for the target *word* from a large corpus, and select the words that are highly related to the word. This list of words related to the target word is used in order to filter all topic signatures corresponding to the target word, that is, context terms which are not relevant for the target word are deleted from the topic signature. We have tested both filtered and unfiltered settings.

3 Similarity Measures

Once we have constructed the topic signatures, it is possible to calculate the similarity between word senses using their topic signatures. If every word which can appear in a topic signature is considered a dimension in a Euclidean space, the similarity between two topic signatures can be calculated using the cosine of the angle of the vectors, or the Euclidean distance between them[2].

In order to evaluate the quality of our similarity measures we have taken two similarity metrics based on the WordNet hierarchy and used them as gold standards.

- Resnik's distance metric based on the Information Content of the synset [10].
- The inverse of the minimal number of hypernymy links between the two synsets in the WordNet hierarchy, also called *Conceptual Distance*.

Besides, we have also taken the manually defined coarse-grained senses used in the Word Sense Disambiguation exercise Senseval-2. In order to define a similarity matrix based on this resource, we have considered two synsets similar if they are in the same coarse-grained sense (similarity 1), and dissimilar otherwise (similarity 0).

4 Experimental Setting

Each experiment has been performed with a different choice of parameters in the construction of topic signatures:

1. Building the collections (monosemous vs. all relatives);

[1] (a) $\frac{tf_t}{max_t tf_t} \times \log \frac{N}{df_t}$ (b) $\left(0.5 + \frac{0.5 \times tf_t}{max_t tf_t}\right) \log \frac{N}{df_t}$

[2] We have calculated the Euclidean distance of the unnormalized vectors because, in our experiments, a normalization produced that all the distances between the signatures became very similar and there was not much difference between the different weight functions used.

Table 1. Similarity values, using the monosemous relatives queries, the cosine similarity for comparing the signatures, and correlation to the link distance in WordNet as gold standard.

Weight:	**Chi2**		**Tf·idf$_1$**		**Tf·idf$_2$**		**MI**		**t-score**	
Filtering:	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**
art	0.85	0.67	0.71	0.83	0.82	0.82	**0.99**	**0.99**	0.58	0.59
authority	0.18	0.17	0.72	0.82	0.73	0.74	0.83	**0.85**	0.53	0.42
bar	0.3	0.31	0.53	0.45	0.64	0.66	**0.79**	0.74	0.15	0.05
bum	0.79	0.77	0.92	0.74	0.85	0.41	**1**	0.99	0.72	0.75
chair	0.48	0.41	0.78	0.71	0.78	0.8	**0.98**	0.91	0.61	0.67
channel	0.44	0.44	0.62	0.64	0.83	0.87	**0.92**	**0.92**	0.41	0.66
child	0.28	0.26	0.84	0.85	0.79	0.81	0.62	0.64	0.87	**0.89**
church	0.7	0.7	0.97	0.89	0.9	0.88	0.98	**1**	**1**	**1**
circuit	0.6	0.53	0.62	0.61	0.79	0.81	**0.97**	0.96	0.58	0.42
day	0.5	0.54	0.5	0.52	0.7	0.77	0.91	**0.92**	-0.02	0.04
dike	0	0	0	0	0	0	1	1	1	1
facility	0.41	0.54	0.68	0.64	0.7	0.72	**0.91**	0.78	0.47	0.39
fatigue	0.82	0.81	**0.92**	0.43	0.52	0.48	0.68	0.56	0.43	0.37
feeling	0.26	0.27	0.31	0.35	0.82	**0.87**	0.83	**0.87**	0.62	0.68
grip	0.29	0.24	0.8	0.66	0.7	0.76	0.86	**0.89**	0.43	0.45
hearth	0.8	0.68	0.75	0.92	0.71	0.8	0.96	**1**	0.89	0.89
MEAN	0.47	0.45	0.65	0.6	0.69	0.72	**0.88**	0.87	0.61	0.62

2. Weight function (χ^2, tf·idf, MI or t-score);
3. With or without filtering;
4. Similarity metric between the topic signatures: cosine or Euclidean.

The evaluation has been done with sixteen nouns from the Senseval 2 exercise that were also used in [7] (WordNet version 1.7).

The correlation between our proposed similarity measures and the three gold standard similarity measures was used as a quality measure. The correlation was computed in the following way. First, for every noun, a symmetric similarity matrix is calculated containing the gold standard similarity between each pair of senses, and another matrix is calculated using the topic signatures. The correlation between the two matrices has been calculated transforming the matrices into vectors (after removing the diagonal and the values which are duplicated because of its symmetry) and using the cosine between the vectors. A measure of 1 will give us perfect similarity, in contrast to a measure of 0.

5 Results

Table 1 shows the results for the sixteen words separately and overall, using monosemous relatives for collecting the documents, the cosine similarity between the topic signatures, and the link distance in WordNet as gold standard. In the case of the word *dike*, the similarities are always 0 or 1. This is due to the fact that it only has two senses in WordNet. Therefore, there is only one similarity value between the two senses, and the cosine similarity between using a theoretical metric and using the topic signatures is 1 when both values are non-zero,

Table 2. Results for the signatures obtained with the monosemous relatives procedure, given as correlation measures against each of the three gold standards

	Weight	*Chi2*		*Tf·idf*$_1$		*Tf·idf*$_2$		*MI*		*t-score*	
	Filtering	*No*	*Yes*	*No*	*Yes*	*No*	*Yes*	*No*	*Yes*	*No*	*Yes*
Gold std.	**Metric**										
Coarse-grained	*Euclidean*	0.14	0.12	0.25	0.23	0.19	0.08	0.3	**0.33**	**0.33**	0.29
senses	*cosine*	0.22	0.21	0.38	**0.47**	0.34	0.37	0.37	0.39	0.17	0.2
Resnik	*Euclidean*	0.31	0.28	0.35	0.44	0.28	0.39	**0.56**	**0.56**	0.55	0.51
	cosine	0.39	0.38	0.35	0.26	0.35	0.37	**0.52**	0.49	0.31	0.35
links	*Euclidean*	0.63	0.61	0.63	0.7	0.48	0.51	0.81	**0.87**	**0.87**	0.8
	cosine	0.47	0.45	0.65	0.6	0.69	0.72	**0.88**	0.87	0.61	0.62

Table 3. Results for the signatures obtained with the all relatives procedure, given as correlation measures against each of the three gold standards

	Weight	*Chi2*		*Tf·idf*$_1$		*Tf·idf*$_2$		*MI*		*t-score*	
	Filtering	*No*	*Yes*	*No*	*Yes*	*No*	*Yes*	*No*	*Yes*	*No*	*Yes*
Gold std.	**Metric**										
Coarse-grained	*Euclidean*	0.17	0.16	0.18	0.18	0.18	0.18	0.33	0.33	**0.35**	0.32
senses	*cosine*	0.33	0.3	0.33	**0.39**	0.34	**0.39**	0.32	0.34	0.03	0.04
Resnik	*Euclidean*	0.38	0.37	0.3	0.3	0.3	0.3	0.48	**0.49**	**0.49**	0.47
	cosine	0.44	0.28	0.47	0.46	0.51	0.48	**0.65**	0.61	0.42	0.3
links	*Euclidean*	0.65	0.65	0.57	0.57	0.57	0.57	0.82	0.84	**0.85**	0.82
	cosine	0.49	0.43	0.62	0.62	0.68	0.69	0.81	**0.84**	0.44	0.43

and 0 when one of the values happens to be 0 (as it is the case when both topic signatures have no word in common). The best results are obtained for unfiltered topic signatures where MI is used as the weighting function.

Tables 2 and 3 list the results obtained for each possible configuration. The results show that it is possible to approximate very accurately the similarity metric based on link distance, as it is possible to attain a similarity of 0.88 with monosemous relatives and the MI or the t-score weight functions. The similarity between Resnik's function and the signatures was somewhat lower, with a cosine similarity of 0.65 (again with the MI weight function, but with the all relatives signature). Finally, it was more difficult to approximate the similarity based on the coarse grained senses, as it does not provide similarity values in $\Re$ but binary values. Nonetheless, it was possible to obtain a cosine similarity of 0.47 with a tf·idf function.

Regarding the parameters of topic signature construction, the monosemous relative method obtains the best correlation when compared to the link distance gold standard. As this method uses a larger amount of examples than the all relatives method, we cannot be conclusive on this. Previous experiments [4] already showed that short contexts of larger amount of examples were prefereable rather that larger context windows and fewer examples. On the same gold standard, MI and t-score attain much better correlation scores that the rest of weighting functions. Filtering the topic signature does not improve the results, and both Euclidean distance and the cosine yield the same scores.

6 Conclusion and Future Work

The experiments show that it is possible to approximate accurately the link distance between synsets (a semantic distance based on the internal structure of WordNet) with topic signatures. However, Resnik's metric [10] has not been as easily captured by the topic signatures, so more work is needed to be able to approximate it with distributional procedures. The main source of the difference is that Resnik's metric gives a similarity of 0 to two synsets if they are located in different sub-taxonomies (with a different root node), such as *church* as a group, an entity or an act. On the other hand, there will probably be some similarity between the topic signatures of two such synsets. Finally, the gold standard metric based on the coarse-grained senses was the one that produced the lowest results. This is in clear contradiction with our word sense clustering experiments [7], where the clusters constructed using topic signatures replicated very well the coarse-grained senses. We think that the correlation metric is not a very appropriate evaluation method in this case, as any similarity metric will yield low correlation when compared to a boolean similarity metric.

Regarding the parameters for the construction of topic signatures, using monosemous relatives allows for better results. Contrary to our intuitions, filtering did not improve performance, and both Euclidean distance and the cosine yielded similar results. It has been a surprise that Mutual Information and t-score have provided much better results than other metrics, such as χ^2 and tf·idf, which have been used extensively for generating topic signatures in the past.

The next step in these experiments consists in ascertaining whether the settings for which the similarity has been better are also more useful when applied to the classification of new concepts, word sense disambiguation or word sense clustering.

Some other ideas for future work are:

- Compare to other similarity measured using WordNet [8].
- Repeat the experiment with other kinds of topic signatures, such as modeling the syntactic dependences between the synset considered and the context words [6].
- Explore further parameters in topic signature construction.

References

1. Maedche, A., Staab, S.: Ontology learning. In: Staab, S., Studer, R., eds.: Handbook of Ontologies in Information Systems. Springer Verlag (Forthcoming).
2. Fellbaum, C.: Wordnet: An Electronic Lexical Database. Cambridge: MIT Press (1998).
3. Lin, C. Y., Hovy, E.: The automated acquisition of topic signatures for text summarization. In: Proc. of the COLING Conference, Strasbourg, France (2000).
4. Agirre, E., Ansa, O., Hovy, E., Martinez, D.: Enriching very large ontologies using the WWW. In: Ontology Learning Workshop, ECAI, Berlin, Germany (2000).
5. Agirre, E., Ansa, O., Martínez, D., Hovy, E.: Enriching wordnet concepts with topic signatures. In: Proceedings of the SIGLEX workshop on WordNet and Other lexical Resources: Applications, Extensions and Customizations, in conjunction with NAACL, Pittsburg (2001).
6. Alfonseca, E., Manandhar, S.: Extending a lexical ontology by a combination of distributional semantics signatures. In: Proceedings of EKAW'02, Siguenza, Spain (2002) Also published in *Knowledge Engineering and Knowledge Management*. Lecture Notes in Artificial Intelligence 2473. Springer Verlag.

7. Agirre, E., de Lacalle Lekuona, O. L.: Clustering wordnet word senses. In: Proceedings of the Conference on Recent Advances on Natural Language Processing (RANLP '03), Bulgary (2003).
8. Banerjee, P., Pedersen, T.: Using measures of semantic relatedness for word sense disambiguation. In: Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City (2003).
9. Budanitsky, A., Hirst, G.: Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. In: Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics, Pittsburgh (2001).
10. Resnik, P. S.: Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research **11** (1999) 95–130.

The MEANING Multilingual Central Repository

J. Atserias[1], L. Villarejo[1], G. Rigau[2], E. Agirre[2], J. Carroll[3], B. Magnini[4], P. Vossen[5]

[1] TALP Research center, Universitat Politécnica de Catalunya. Catalonia
Email: batalla@talp.upc.es, luisv@talp.upc.es
WWW: http://www.lsi.upc.es/~nlp
[2] IXA Group, University of the Basque Country, Computer Languages and Systems
Email: rigau@si.ehu.es, eneko@si.ehu.es WWW: http://ixa.si.ehu.es/Ixa
[3] University of Sussex, Cognitive and Computing Sciences. UK
Email: J.A.Carroll@sussex.ac.uk WWW: http://www.cogs.susx.ac.uk/lab/nlp/
[4] ITC-IRST Italy
Email: magnini@itc.it WWW: http://tcc.itc.it
[5] Irion Technologies B.V. The Netherlands
Email: Piek.Vossen@irion.nl WWW: http://www.irion.nl

Abstract. This paper describes the first version of the Multilingual Central Repository, a lexical knowledge base developed in the framework of the MEANING project. Currently the MCR integrates into the EuroWordNet framework five local wordnets (including four versions of the English WordNet from Princeton), an upgraded version of the EuroWordNet Top Concept ontology, the MultiWordNet Domains, the Suggested Upper Merged Ontology (SUMO) and hundreds of thousand of new semantic relations and properties automatically acquired from corpora. We believe that the resulting MCR will be the largest and richest Multilingual Lexical Knowledge Base in existence.

1 Introduction

Building large and rich knowledge bases takes a great deal of expensive manual effort; this has severely hampered Knowledge-Technologies and HLT application development. For example, dozens of person-years have been invested into the development of wordnets (WNs) [1] for various languages [2,3], but the data in these resources is still not sufficiently rich to support advanced multilingual concept-based HLT applications directly. Furthermore, resources produced by introspection usually fail to register what really occurs in texts.

The MEANING project [4][6] identifies two complementary and intermediate tasks which are crucial in order to enable the next generation of intelligent open domain HLT application systems: Word Sense Disambiguation (WSD) and large-scale enrichment of Lexical Knowledge Bases (LKBs). Advances in these two areas will allow large-scale acquisition of shallow meaning from texts, in the form of relations between concepts.

However, progress is difficult due to the following interdependence: (i) in order to achieve accurate WSD, we need far more linguistic and semantic knowledge than is available in current LKBs (e.g. current WNs); (ii) in order to enrich existing LKBs we need to acquire information from corpora accurately tagged with word senses.

[6] http://www.lsi.upc.es/~nlp/meaning/meaning.html

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 23–30.

MEANING proposes an innovative bootstrapping process to deal with this interdependency between WSD and knowledge acquisition exploiting a multilingual architecture based on EuroWordNet (EWN) [2]. The project plans to perform three consecutive cycles of large-scale WSD and acquisition processes in five European languages including Basque, Catalan, English, Italian and Spanish. As languages realize meanings in different ways, some semantic relations that can be difficult to acquire in one language can be easy to capture in other languages. The knowledge acquired for each language during the three consecutive cycles will be consistently upload and integrated into the respective local WNs, and then ported and distributed across the rest of WNs, balancing resources and technological advances across languages.

This paper describes the first version of the Multilingual Central Repository produced after the first MEANING cycle. Section2 presents the MCR structure, content and associated software tools. Section 3 describes the first uploading process, and section 4 the porting process. Section 5 and 6 conclude and discusse directions for future work.

2 Multilingual Central Repository

The Multilingual Central Repository (MCR) ensures the consistency and integrity of all the semantic knowledge produced by MEANING. It acts as a multilingual interface for integrating and distributing all the knowledge acquired in the project. The MCR follows the model proposed by the EWN project, whose architecture includes the **Inter-Lingual-Index** (ILI), a **Domain ontology** and a **Top Concept ontology** [2].

The first version of the MCR includes only conceptual knowledge. This means that only semantic relations among synsets have been acquired, uploaded and ported across local WNs. The current MCR integrates: (i) the ILI based in WN1.6, includes EWN Base Concepts, EWN Top Concept ontology, MultiWordNet Domains (MWND), Suggested Upper Merged Ontology (SUMO); (ii) Local WNs connected to the ILI, including English WN 1.5, 1.6, 1.7, 1.7.1, Basque, Catalan, Italian and Spanish WN; (iii) Large collections of semantic preferences, acquired both from SemCor and from BNC; Instances, including named entities.

The MCR provides a web interface to the database based on Web EuroWordNet Interface[7]. Three different APIs have been also developed to provide flexible access to the MCR: first, a SOAP API to allow users to interact with the MCR, an extension of the WNQUERY Perl API to the MCR and a C++ API for high performance software.

3 Uploading Process

Uploading consists of the correct integration of every piece of information into the MCR. That is, linking correctly all this knowledge to the ILI. This process involves a complex cross-checking validation process and usually a complex expansion/inference of large amounts of semantic properties and relations through the WN semantic structure (see [5] fot further details).

[7] `http://nipadio.lsi.upc.es/wei.html`

3.1 Uploading WNs

To date, most of the knowledge uploaded into the MCR has been derived from WN1.6 (or SemCor); the Italian WN and the MWND, both use WN1.6 as ILI. However, the ILI for Spanish, Catalan and Basque WNs was WN1.5, as well as the EWN Top Concept ontology and the associated Base Concepts. To deal with the gaps between versions and to minimize side effects with other international initiatives (Balkanet, EuroTerm, eXtended WN) and WN developements around Global WordNet Association, we used a set of improved mappings between all involved resources[8].

3.2 Uploading Base Concepts

The original set of **Base Concepts** from EWN based on WN1.5 contained a total of 1,030 ILI-records. Now, the Base Concepts from WN1.5 have been mapped to WN1.6. After a manual revision and expansion to all WN1.6 top nodes, the resulting Base Concepts for WN1.6 total 1,535 ILI-records. In this way, the new version of Base Concepts covers the complete hierarchy of ILI-records (only nouns and verbs).

3.3 Uploading the Top Ontology

The purpose of the EWN **Top Concept ontology** was to enforce more uniformity and compatibility of the different WN developments. The EWN project only performed a complete validation of the consistency of the **Top Concept ontology** of the Base Concepts.

Although the classifi cation of WN is not always consistent with the **Top Concept ontology**, we performed an automatic expansion of the **Top Concept** properties assigned to the Base Concepts. That is, we enriched the complete ILI structure with features coming from the Base Concepts by inheriting the Top Concept features following the hyponymy relationship. The **Top Concept ontology** has been uploaded in three steps:

1. Properties are assigned to WN1.6 synsets through the mapping.
2. For those WN1.6 Tops (synsets without any parent) that do not have any property assigned through the mapping, we assigned to them the Top Concept ontology properties by hand.
3. The properties are propagated top-down through the WN hierarchy.

The following incompatibilities inside the **Top Concept ontology** have been used to block the top-down propagation of the **Top Concept** properties:

- *1stOrderEntity – 2ndOrderEntity – 3rdOrderEntity*;
- *substance – object*;
- *plant – animal – human – creature*;
- *natural – artifact*;
- *solid – liquid – gas*.

Thus, when detecting that any of the current **Top Concept ontology** properties of a synset is incompatible with other inherited (due possibly to multiple inheritance), this property is not assigned to the synset and the propagation to the synset's descendants stops.

[8] `http://www.lsi.upc.es/~nlp/tools/mapping.html`

3.4 Uploading SUMO

The Suggested Upper Merged Ontology (SUMO) [6] is an upper ontology created at Teknowledge Corporation and proposed as starting point for the IEEE Standard Upper Ontology Working group.

SUMO provides definitions for general purpose terms and is the result of merging different free upper ontologies (e.g. Sowa's upper ontology, Allen's temporal axioms, Guarino's formal mereotopology, etc.) with WN1.6. Currently only the SUMO labels and the SUMO ontology hyperonym relations are loaded into the MCR. We plan to cross-check the **Top Concept ontology** expansion and the **Domain ontology** with the SUMO ontology.

3.5 Uploading Selectional Preferences

A total of 390,549 weighted Selectional Preferences (SPs) obtained from two different corpora and using different approaches have been uploaded into the MCR. The first set [7] of weighted SPs was obtained by computing probability distributions over the WN1.6 noun hierarchy derived from the result of parsing the BNC. The second set [8] was obtained from generalizations of grammatical relations extracted from Semcor.

The SPs are included in the MCR as ROLE noun–verb relations[9]. Although we can distinguish subjects and objects, all of them have been included as a more general ROLE relation.

4 Porting Process

In the first porting process all the knowledge integrated into the MCR has been ported (distributed) directly to the local WNs (no extra semantic knowledge has been inferred in this process). Table 1 summarises the main results before (UPLOAD0) and after the whole porting process (PORT0) for Spanish, English and Italian. In this table, relations do not consider hypo/hyperonym relations and *links* stands for total number of Domains or Top Concept ontology properties ported (before application of the top-down expansion process).

4.1 An Example

When uploading coherently all this knowledge into the MCR, we added consistently a large set of explicit knowledge about each sense which can be used to differentiate and characterize better their particular meanings. We will illustrate the current content of the MCR, after porting, with a simple example: the Spanish noun *pasta*.

The word *pasta* (see table 2) illustrates how all the different classification schemes uploaded into the MCR: Semantic File, MWND, Top Concept ontology, etc. are consistent and makes clear semantic distinctions between the money sense (pasta_6), the general/chemistry sense (pasta_7) and the food senses (all the rest). The food senses of *Pasta* can now be further differentiate by means of explicit EWN Top Concept ontology properties. All the food senses are descendants of *substance_1* and *food_1* and inherits the Top Concept attributes *Substance* and *Comestible* respectively.

[9] In EWN, INVOLVED and ROLE relationships are defined symmetrically.

Table 1. PORT0 Main figures for Spanish, English and Italian

	Spanish		English		Italian	
Relations	**UPLOAD**	**PORT0**	**UPLOAD**	**PORT0**	**UPLOAD**	**PORT0**
be_in_state	1,302	=	1,300	+2	364	+2
causes	240	=	224	+19	117	+15
near_antonym	7,444	=	7,449	+221	3,266	=
near_synonym	10,965	=	21,858	+19	4,887	+54
role	106	=	0	+106	0	+46
role_agent	516	=	0	+516	0	+227
role_instrument	291	=	0	+291	0	+151
role_location	83	=	0	+83	0	+39
role_patient	6	=	0	+6	0	+3
xpos_fuzzynym	37	=	0	+37	0	+23
xpos_near_synonym	319	=	0	+319	0	+181
Other relations	31,644	=	29,120	+2,627	9,541	+22
Total	53,272	=	59,951	+4,246	18,175	+763
role_agent-semcor	0	+52,394	69,840	=	0	+41,910
role_agent-bnc	0	+67,109	95,065	=	0	+40,853
role_patient-semcor	0	+80,378	110,102	=	0	+41,910
role_patient-bnc	0	+79,443	115,102	=	0	+50,264
Role	0	+279,324	390,109	=	0	+174,937
Instances	0	+1,599	0	+2,198	791	=
Proper Nouns	1,806	=	17,842	=	2,161	=
Base Concepts	1,169	=	1,535	=	0	+935
Domains Links	0	+55,239	109,621	=	35,174	=
Domains Synsets	0	+48,053	96,067	=	30,607	=
Top Ontology Links	3,438	=	0	+4,148	0	+2,544
Top Ontology Synsets	1,290	=	0	+1,554	0	+946

Selectional Preferences can also help to distinguish between senses, e.g only the money sense has the following preferences as object: *1.44 01576902-v {raise#4}, 0.45 01518840-v {take_in#5, collect#2}* or *0.23 01565625-v {earn#2, garner#1}*.

We will investigate new inference facilities to enhance the uploading process. After full expansion (**Realization**) of the EWN Top Concept ontology properties, we will perform a full expansion through the noun part of the hierarchy of the selectional preferences acquired from SemCor and BNC (and possibly other implicit semantic knowledge currently available in WN such as meronymy information).

We plan further investigation to perform full bottom-up expansion (**Generalization**), rather than merely expanding knowledge and properties top-down. In this case, different knowledge and properties can collapse on particular Base Concepts, Semantic Files, Domains and/or Top Concepts.

Table 2. Food senses for the Spanish word *pasta*

Domain: chemistry-pure_science
Semantic File: 27-Substance
SUMO:
Substance-SelfConnectedObject-Object-Physical-Entity

Top Concept ontology
Natural-Origin-1stOrderEntity
Substance-Form-1stOrderEntity

pasta#n#7 10541786-n
paste#1
gloss: any mixture of a soft and malleable consistency

Domain: money-economy-soc.science
Semantic File: 21-MONEY
SUMO:
CurrencyMeasure-ConstantQuantity-PhysicalQuantity-Quantity-Abstract-Entity
Top Concept ontology
Artifact-Origin-1stOrderEntity
Function-1stOrderEntity
MoneyRepresentation-Representation-Function-1stOrderEntity

pasta#n#6 09640280-n
dough#2,bread#2,loot#2,...
gloss: informal terms for money

Domain: gastronomy-alimentation-applied_science
Semantic File: 13-FOOD
Top concept ontology
Comestible-Function-1stOrderEntity
Substance-Form-1stOrderEntity

Top Concept ontology
Natural-Origin-1stOrderEntity

Top Concept ontology
Part-composition-1stOrderEntity

pasta#n#4 05886080-n
spread#5,paste#3
gloss: a tasty mixture to be spread on bread or crackers

pasta#n#1 05671312-n
pastry#1,pastry_dough#1
gloss: a dough of fbur and water and shortening

pasta#n#3 05739733-n
pasta#1,alimentary_paste#1
gloss: shaped and dried dough made from fbur and water & sometimes egg

pasta#n#5 05889686-n *dough#1*
gloss: a dough of fbur and water and shortenings

Top Concept ontology
Artifact-Origin-1stOrderEntity
Group-Composition-1stOrderEntity

pasta#n#2 05671439-n
pie_crust#1,pie_shell#1
gloss: pastry used to hold pie fillings

5 Future Work

Having all these types of different knowledge and properties coming from different sources, methods, and completely expanded through the whole MCR, a new set of inference mechanisms can be devised to further infer new relations and knowledge inside the MCR. For instance, new relations could be generated when detecting particular *semantic patterns* occurring for some synsets having certain ontological properties, for a particular Domain, etc. That is, new relations could be generated when combining different methods and knowledge. For instance, creating new explicit relations (regular polysemy, nominalizations, etc.) when several relations derived in the integration process have confidence scores greater than certain thresholds, occurring between certain ontological properties, etc.

Obviously, new research is also needed for porting the various types of knowledge across languages. For instance, new ways to validate the ported knowledge in the target languages.

6 Conclusions

The first version of the MCR integrates into the same EWN framework (using an upgraded release of Base Concepts and Top Concept ontology and MWND) five local WNs (with four English WN versions) with hundreds of thousands of new semantic relations, instances and properties fully expanded. All WNs have gained some kind of knowledge coming from other WNs by means of the first porting process. We believe that the resulting MCR is the largest and richest multilingual LKB in existence.

We intend this version of the MCR to be a natural multilingual large-scale knowledge resource for a number of semantic processes that need large amounts of linguistic knowledge to be effective tools (e.g. Semantic Web ontologies).

When uploading coherently all this knowledge into the MCR a full range of new possibilities appears for improving both Acquisition and WSD tasks in the next two MEANING rounds.

Future versions of the MCR may include language dependent data, such as syntactic information, subcategorization frames, diathesis alternations, Dorr's Lexical Conceptual Structures, complex semantic relations [9], etc. The information will be represented following current standards (e.g. EAGLES), where these exist.

Regarding the *porting process*, we will investigate inference mechanisms to infer new explicit relations and knowledge (regular polysemy, nominalizations, etc.). Finally, more research is needed to verify the correctness of the various types of semantic knowledge ported across languages.

Acknowledgments

This research has been partially funded by the Spanish Research Department (HERMES TIC2000-0335-C03-02) and by the European Commission (MEANING IST-2001-34460).

References

1. Fellbaum, C., ed.: WordNet. An Electronic Lexical Database. The MIT Press (1998).

2. Vossen, P., ed.: EuroWordNet: A Multilingual Database with Lexical Semantic Networks . Kluwer Academic Publishers (1998).
3. Bentivogli, L., Pianta, E., Girardi, C.: Multiwordnet: developing an aligned multilingual database. In: First International Conference on Global WordNet, Mysore, India (2002).
4. Rigau, G., Magnini, B., Agirre, E., Vossen, P., Carroll, J.: Meaning: A roadmap to knowledge technologies. In: Proceedings of COLLING Workshop 'A Roadmap for Computational Linguistics', Taipei, Taiwan (2002).
5. Atserias, J., Villarejo, L., Rigau, G.: Integrating and porting knowledge across languages. In: Procceeding of Recent Advances in Natural Language Processing (RANLP'03), Bulgaria (2003) 31–37.
6. Niles, I., Pease, A.: Towards a standard upper ontology. In: In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001), Chris Welty and Barry Smith, eds (2001) 17–19.
7. McCarthy, D.: Lexical Acqusition at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences. PhD thesis, University of Sussex (2001).
8. Agirre, E., Martinez, D.: Integrating selectional preferences in wordnet. In: Proceedings of the first International WordNet Conference in Mysore, India (2002).
9. Lin, D., Pantel, P.: Discovery of inference rules for question answering. Natural Language Engineering **7** (2001) 343–360.

Russian WordNet

From UML-notation to Internet/Intranet Database Implementation

Valentina Balkova[2], Andrey Sukhonogov[1], and Sergey Yablonsky[1,2]

[1] Petersburg Transport University, Moscow av., 9, St.-Petersburg, 190031, Russia,
Email: ASukhonogov@rambler.ru
[2] Russicon Company, Kazanskaya str., 56, ap.2, 190000, Russia
Email: v_balk@front.ru,serge_yablonsky@hotmail.com,root@russicon.ru

Abstract. This paper deals with development of the first public web version of Russian WordNet and future parallel English-Russian and multiligual web versions of WordNet. It describes usage of Russian and English-Russian lexical language resources and software to process WordNet for Russian language and design of a database management systems for efficient storage and retrieval of various kinds of lexical information needed to process WordNet. Relevant aspects of the UML data models, XML format and related technologies are surveyed. The pilot Internet/Intranet version of described system based on Oracle 9i DBMS and Java technology is published at: http://www.pgups.ru/WebWN/wordnet.uix.

1 Introduction

This paper attempts to introduce results of an ongoing project of developing of the first public web version of Russian WordNet and future parallel English-Russian and multiligual web versions of WordNet. English-Russian parallel WordNet resources and software implementation for building parallel multilingual lexical database based on Princeton WordNet are introduced. The goal of database management system development is to build a multilingual (monolingual Russian now and bilingual English-Russian and multilingual in future) lexical database of wordnets for Russian language (WordNet.ru), which are structured along the same lines as the Princeton WordNet for English language. WordNet.ru contains information about nouns, verbs, adjectives and adverbs in Russian and is organized around the notion of a synset. The WordNet.ru represents basic resources for content-based language-technologies within and across the Russian and English languages. It will enable a form of multilingual text indexing and retrieval, a direct benefit from the multilingual semantic resource in:

- information-acquisition tools;
- authoring tools;
- language-learning tools;
- translation-tools;
- summarizers;
- semantic web.

The objectives of this project are not unique. Several analogous projects have been carried out to different stages (EuroWordNet, BalkanNet etc.) but there is no public web realization of Russian WordNet yet.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 31–38.

We have been implementing a combination of manual and automatic techniques. Today there are several WordNet viewers: the Princeton viewer and EuroWordNet viewer/editor (VisDic). The limitations of these popular WordNet tools for Russian WordNet design stimulate our development of Russian WordNet editor and Multilingual WordNet editor based on Oracle database management system.

The paper discusses the complete process of building and managing of monolingual Russian and parallel English-Russian version of WordNet database management system, including the development of UML/ER-specifi cations, architecture and examples of actual implementations of DBMS tools. The system is implemented using DBMS Oracle9i Release 2 and Java technology.

2 Lexical Resources for Russian WordNet

We use several Russian lexical resources. Russicon company has such main counterparts (Yablonsky S. A., 1998, 2003) for English-Russian and Russian WordNet development:

- *The General Russicon Russian lexicon* which is formed from the intersection of the perfect set of Russicon Russian grammatical dictionaries with inflection paradigms (200,000 paradigms that produce more then 6,000,000 inflection word forms). Lexicon consists of:
  - Russian basic grammatical dictionary;
  - Computer dictionary;
  - Geographical names dictionary;
  - Russian personal names, patronymics and surnames dictionary;
  - Business dictionary;
  - Juridical dictionary;
  - Jargon dictionary etc.
- *The Russicon Russian explanatory dictionary.* The dictionary gives the broad lexical representation of the Russian language of the end of the XX century. More then 100,000 contemporary entries include new words, idioms and their meanings from the language of the Eighties-Nineties. The dictionary is distinguished by its complete set of entry word characteristics, clear understandable defi nitions, its guidance on usage. All dictionary information for entries is structured in more then 60 attributes:
  - entry word;
  - multiple word entries;
  - usage notes;
  - precise, contemporary defi nitions;
  - derivations;
  - example sentences/citations;
  - idioms etc.
- *The Russicon Russian thesaurus* (set of 14,000 Russian synsets). Synonym list plus word list containing approximately 30,000 normalized entry words with inflection paradigms.
- *The Russicon Russian Orthographic dictionary.*

All dictionaries are implemented as text-files and as compressed linguistic databases connected to the Russicon language processor. *Text-files* of grammatical dictionaries contain normalized entry words (lemmas) with hyphenation and inflexion paradigm plus grammatical tags for each word of paradigm. The set of language tags consists of part of speech, case, gender, number, tense, person, degree of comparison, voice, aspect, mood, form, type, transitiveness, reflexive, animation. For thesaurus and explanatory dictionary we have two or more text – files, one always containing inflexion paradigms of all words of the dictionary. Formats of files are plain text and HTML.

We also use several print Russian dictionaries:

- a version of the new monoligual *Russian Explanatory Dictionary* (Efremova T. F., 2001 – 136.000 entry words) for improvement of the Russian WordNet structure;
- *The Russian Semantic Dictionary* (ed. Shvedova N. Y.,1998,2000, vol.1,2 – 39.000 + 40.000 entry words) and *The Explanatory Ideographical Dictionary of Russian Verbs* (Babenko L. G., 1999 – 25000 entry words) for improvement of the Russian WordNet hyponomy/hyperonymy and meronomy/holonymy relations;
- *The Russian Language Antonyms Dictionary* (L'vov M. R., 2002 – 3200 entry words) for improvement of the Russian WordNet antonomy relations.

3 English-Russian WordNet

Two complementary approaches were devised in EuroWordNet to build local wordnets from scratch:

- The merge approach: building taxonomies from monolingual lexical resources and then, making a mapping process using bilingual dictionaries;
- The expand approach: mapping directly local words to English synsets using bilingual dictionaries.

The merge approach is present in our Russian WordNet construction process from the beginning. We are really building taxonomies using Russian lexical resources mentioned above. After our first version will be finished we plan making mapping using bilingual dictionaries.

At the same time we use the expand approach for direct mapping of many words from English WordNet to Russian and vise verse. This approach is used for some English proper and geographical names.

4 The Current Status of the Russian WordNet

The statistics of synsets in the first version of WordNet.ru are displayed in Table 1.

We plan to include additionally 10,000 Russian local proper and geographic names in the first version.

The list of semantic relations in WordNet.ru is based mostly on Princeton WordNet Lexical and Conceptual Relations, and EuroWordNet Language-Internal Relations.

Main relations between synsets: hyponymy/hyperonymy, antonymy, meronomy/holonymy. Main relations between members of synsets: synonymy, antonymy, derivation synonymy,

Table 1. Statistics of synsets in the first version of WordNet.ru

Russian WordNet Word Report					
Total	Noun	Verb	Adj	Adv	Other
111749	44751	27997	20736	4997	13268

Synset report						
WordCnt	Total	Noun	Verb	Adj	Adv	Other
1	120549	53137	29351	25299	4976	7786
2	12825	3355	7077	1635	188	570
3	3637	1011	1675	378	121	452
4	2193	574	920	253	89	357
5	1424	351	581	186	78	228
6	1121	258	428	148	67	220
7	791	184	311	89	45	162
8	565	128	239	58	37	103
9	443	72	186	62	26	97
10	305	55	124	45	16	65
…	…	…	…	…	…	…
68	2	0	0	0	1	1
Total	144980	59294	41403	28316	5718	10249

derivation hyponymy. Two last relations are relations between aspect pairs and between neutral words and their expressive derivatives etc.

We produce inflection paradigm for every input word. The number of all inflections is approximately 5,000,000. This gives us possibility to output Russian WordNet synsets not only for lemma of input word, but for any inflection form of input word. It is important because Russian is highly inflection language.

5 Language Software

For many linguistic tasks of WordNet development we use such parts of language processor Russicon (Yablonsky S. A. 1998, 1999, 2003): system for construction and support of machine dictionaries and morphological analyzer and normalyzer.

6 WordNet Conceptual Model

6.1 UML Model Design

Today Unified Modeling Language (UML) defines a standard notation for object-oriented systems (Booch G., Rumbaugh J., and Jacobson I., 1998). Using UML enhances communication between linguistic experts, workflow specialists, software designers and other professionals with different backgrounds. At the same time UML diagrams are widely used for realation data base design (for example in Rational Rose).

The core part of Russian WordNet UML model includes **SYNSET, WORD, IDIOM, EXPLANATION** entities (Figure 1). For **SYNSET** entity such attributes are defined:

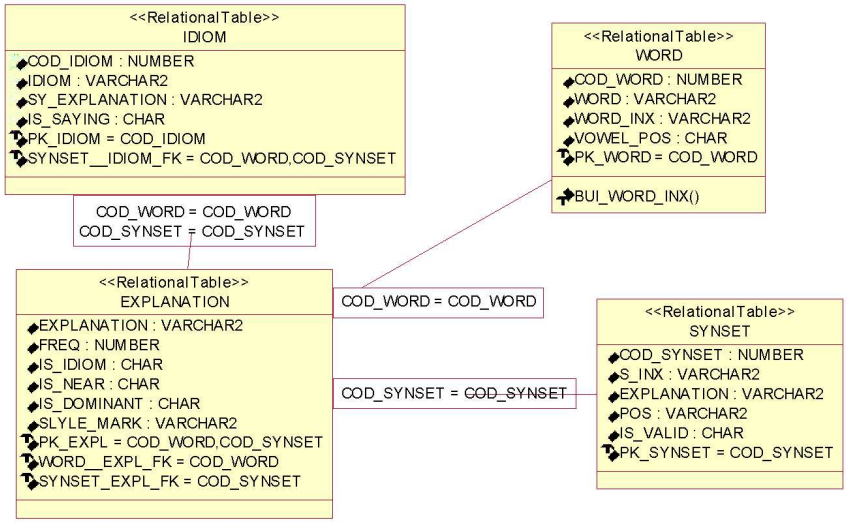


Fig. 1. Core part of Russian WordNet UML model

- COD_SYNSET (internal database synset identifier);
- S_INX (index) – unique synset identifier; it could be defined by user while working with thesaurus;
- EXPLANATION – synset explanation;
- POS – grammatical information;
- IS_VALID – validation flag.

For **WORD** entity such attributes are defined:

- COD_WORD (word identifire) – internally used database primary key;
- WORD (word code);
- WORD_INX (word index) – internally used database key for word search;
- VOWEL_POS (stress vowel) – stress position information string; up to 4 stresses in one word could be fixed;

Synset includes one or more words (lemmas) and one word could be included in more then one synset.

Entity **EXPLANATION** is used for storing information about meaning of the word. For it such attributes are defined:

- EXPLANATION (word meaning) – natural language word meaning description;
- IS_IDIOM – idiom identification flag (is true for idiom);
- IS_NEAR – near word identification flag;
- IS_DOMINANT – synset dominant word identification flag;
- STYLE_MARK.

For entity **IDIOM** such attributes are defined:

- COD_ IDIOM – internally used database primary key;
- IDIOM;
- SY_EXPLANATION – natural language idiom meaning description;
- IS_SAYING – saying identification flag.

In Russian WordNet model all types of WordNet relations between synsets are realized. Even more, there are no limitations on the type of relations between synsets. The semantics and number of relations is user defined. For that purpose user is given the so-called *sematic/type relation constructor*. Types of relation are devided into two main groups: *hierarchic (symmetric) and not hierarchic (symmetric and not symmetric).* In Russian WordNet model we plan to develop domain WordNets.

6.2 ER Model Design

At the same time ER (Entity-Relation) models are also very popular in relational data base design. Figure 2 presents the whole ER model of Russian WordNet.

7 Main Steps of Russian WordNet Development

The development process of Russian and English-Russian WordNet development could be devided into two main steps.

The first step ends by production of the first version of Russian WordNet with the number of word inputs more then 100,000. The exact numbers could be found in Section 4. For construction of Russian WordNet we developed Russian WordNet editor.

The second step ends by development of English-Russian version of WordNet. For that purpose we developed Multilingual WordNet Editor.

7.1 Russian WordNet Editor

Russian WordNet editor was developed to help production of Russian WordNet from above mentioned linguistic resources. It allows

- to join sysnsets from thesaurus, explanatory and other dictionaries;
- proceed relations between synsets and words of synsets.

It is a database management system in which users (linguist or knowledge engineer) can create, edit and view Russian WordNet. From a monolingual point of view they can work with any monolingual WordNet (for us – Russian) with its internal semantic relationships.

7.2 Multilingual WordNet Editor

We designed multiligual WordNet editor (beta version) that includes definition of the relations, the common data structure, the shared ontology, the Inter-Lingual-Index and the comparison option, Russian so-called Base Concepts (the Base Concepts are the major building blocks on which the other word meanings in the wordnets depend).

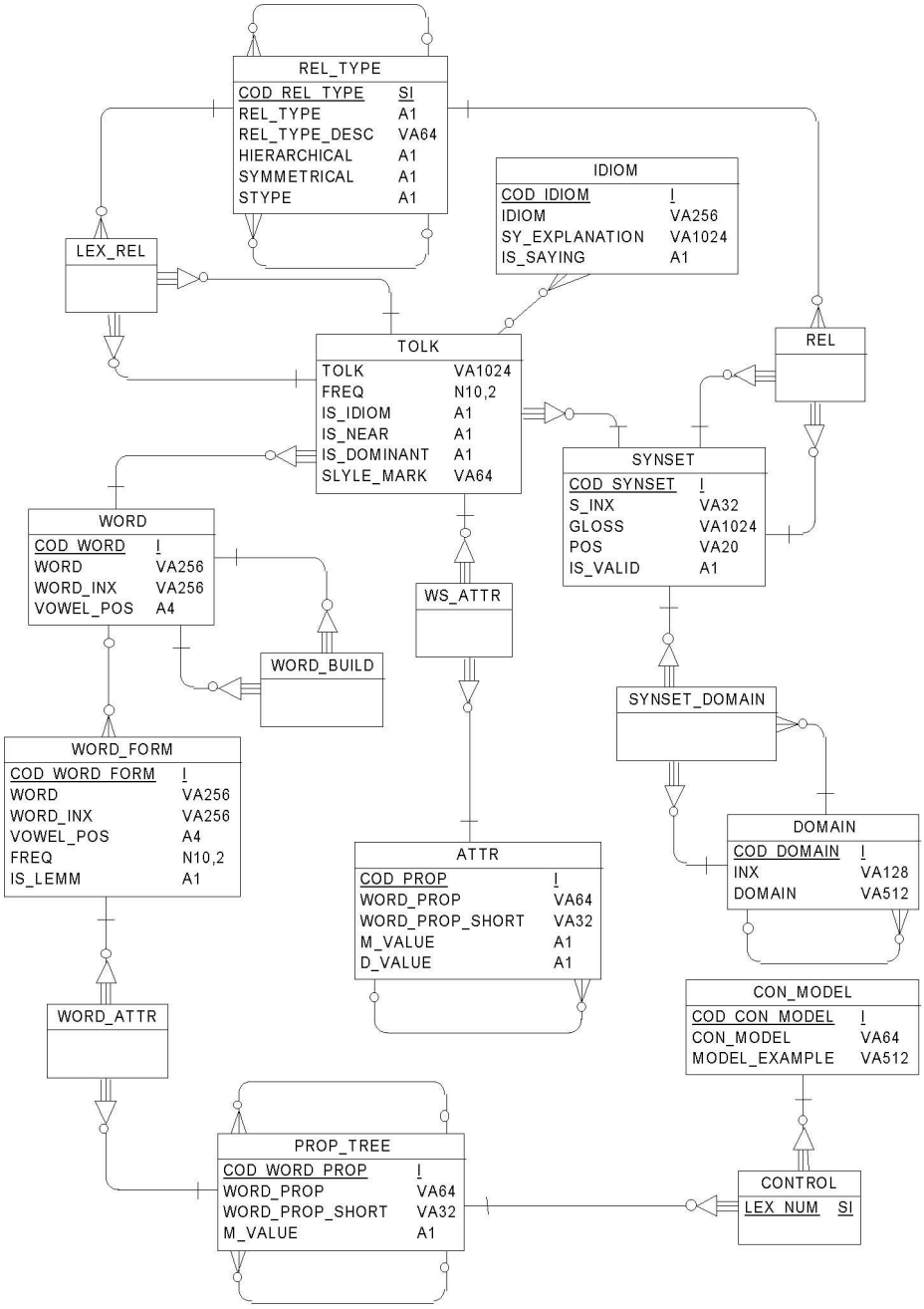


Fig. 2. Russian WordNet ER model

8 Internet Viewer

The pilot Internet version of described system based on Oracle 9i DBMS and Java technology is published at: `http://www.pgups.ru/WebWN/wordnet.uix`.

Our Internet/Intranet WordNet viewer is a database management system in which users (linguist or knowledge engineer) can look at the Russian and English WordNet databases.

9 Conclusion

We present the open UML-specifi cation and new pilot database management system on Oracle 9i DBMS for effi cient storage and retrieval of various kinds of lexical information needed to process English-Russian WordNet. Relevant aspects of the UML/ER data models and related technologies are surveyed. Bilingual WordNet system could be easily expanded in a real multiligual system.

References

1. Babenko L. G. – ed., 1999. Explanatory Ideographical Dictionary of Russian Verbs. – Moscow, Ast-Press.
2. Booch, G., Rumbaugh, J., and Jacobson, I., 1998. The Unified Modeling Language user guide, Addison-Wesley.
3. Efremova T. F., 2001. Novij Slovar Russkogo Yazika [Новый словарь русского языка], v.1,2. – Russian Language.
4. Fellbaum C. WordNet: an Electronic Lexical Database. MIT Press, Cambridge, MA,1998.
5. Lyons J. Semantics. (2 vol.) London and New York, 1977.
6. L'vov M. R., 2002. The Russian Language Antonyms Dictionary. – Moscow, Ast-Press.
7. Miller G. et al. Five Papers on WordNet. CSL-Report, vol.43. Princeton University, 1990.
8. `ftp://ftp.cogsci.priceton.edu/pub/wordnet/5papers.ps`.
9. Prószéky, Gábor & Márton Miháltz, 2002. Semi-automatic Development of the Hungarian WordNet. LREC-2002, Las Palmas, Spain.
10. Shvedova N. Y. – ed., 1998. Russian Semantic Dictionary, vol.1. – Moscow, Azbukovnik.
11. Shvedova N. Y. – ed., 2000. Russian Semantic Dictionary, vol.2. – Moscow, Azbukovnik.
12. Vossen, P. EuroWordNet: A Multilingual Database with Lexical Semantic Network. Dodrecht.
13. Kluwer, 1998.
14. Yablonsky S. A., 1998. Russicon Slavonic Language Resources and Software. In: A. Rubio, N. Gallardo, R. Castro & A. Tejada (eds.) Proceedings First International Conference on Language Resources & Evaluation, Granada, Spain.
15. Yablonsky S. A. (1999). Russian Morphological Analyses. In: Proceedings of the International ConferenceVEXTAL, November 22–24 1999, (pp. 83–90), Venezia, Italia.
16. Yablonsky S. A. (2003). Russian Morphology: Resources and Java Software Applications. In: Proceedings EACL03 Workshop Morphological Processing of Slavic Languages, Budapest, Hungary.

ArchiWordNet: Integrating WordNet with Domain-Specific Knowledge

Luisa Bentivogli[1], Andrea Bocco[2], and Emanuele Pianta[1]

[1] ITC-irst, Via Sommarive 18
38050 Povo – Trento, Italy
Email: bentivo@itc.it, pianta@itc.it
[2] Politecnico di Torino – Dipartimento di Casa Città, viale Mattioli 39
10125 Torino, Italy
Email: andrea.bocco@polito.it

Abstract. Linguistic resources with domain-specific coverage are crucial for the development of concrete application systems, especially when integrated with domain-independent resources. In this paper we present our experience in the creation of ArchiWordNet, a specialized WordNet for the architecture and construction domain which is being created according to the WordNet model and integrated with WordNet itself. Problematic issues related to the creation of a domain-specific wordnet and its integration with a general language resource are discussed, and practical solutions adopted are described.

1 Introduction

The ArchiWordNet (ArchiWN) project is a joint effort between ITC-irst and the Turin Polytechnic aiming at building a thesaurus for the architecture domain to be used within Still Image Server (SIS), an architecture image archive available at the Polytechnic.

SIS was created for educational purposes, with the aim of making accessible to Architecture students and researchers the huge iconographic heritage available in different departments, thus contributing to the preservation and development of the heritage itself. The digitized images are catalogued and organized in a database that can be queried through a web interface accessible within the Polytechnic Intranet. During the cataloguing phase, several keywords are assigned to each image. To make the use of the keywords more systematic and facilitate the retrieval of the images, it is necessary to constrain the keywords used by both indexers and end users through a thesaurus. However, up to now an exhaustive thesaurus for the architecture domain able to meet the needs of the image archive has not been available and thus we decided to create ArchiWN, a bilingual WordNet-like English/Italian thesaurus to be integrated into WordNet itself.

In this paper our experience in the creation of ArchiWN is presented. Section 2 describes the motivations behind the decision of building a WordNet-like thesaurus and its distinguishing features. In sections 3 and 3 some problematic issues related to the creation of a domain-specifi c WordNet and its integration with a general language resource are discussed, and the practical solutions adopted are presented. Finally, Section 5 outlines ArchiWN future enhancements and new application fi elds.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 39–46.

2 ArchiWordNet: a WordNet-like Thesaurus

The main characteristic of ArchiWN is that, while exploiting as much as possible information from already existing architecture thesauri and other specialized sources, it is structured according to the WordNet model [4] and fully integrated into it. More specifically, as we aim at creating a bilingual English/Italian resource, we decided to work within the MultiWordNet (MultiWN) framework. MultiWN [7] is a multilingual lexical database in which the Italian WordNet is strictly aligned with Princeton's English WordNet.

ArchiWN will differ from traditional thesauri with respect to both concepts and relations [2]. Thesauri usually represent concepts using a controlled vocabulary where many synonyms are missed. Also, they include few relations (such as "broader term", "narrower term", "used for", and "related to") whose semantics is rather informal. On the contrary, concepts in WordNet are represented by sets of synonymous words actually occurring in the real language, and WordNet relations are explicit and encoded in a homogeneous way, enabling transitivity and thus inheritance. Given these differences, we decided to adopt the WordNet model for a number of reasons. On the one side, the more rigorous structure of WordNet allows for a more powerful and expressive retrieval mechanism. On the other side, it makes ArchiWN more suitable for educational purposes, as it provides conceptual frameworks which can support learning: its well-structured hierarchies can be browsed to form both a general idea of the architecture domain and a structured knowledge of specific topics.

ArchiWN will differ from traditional thesauri not only in its structure but also in the fact that it is fully integrated with MultiWN. From a theoretical point of view, MultiWN offers a general and multilingual framework for the specialized knowledge contained in ArchiWN. From a practical point of view, the possibility of integrated access allows more flexible retrieval of the information. Moreover, given the huge cost in terms of human effort involved in the construction of such a resource, the integration is particularly useful as information already existing in the generic WordNet can be exploited in the creation of the specialized one.

Throughout the ArchiWN creation phase, we have been faced with the tension between the diverging aims of two different disciplines such as computational linguistics and architecture. More specifically, we had to find a trade off between the necessity of creating a linguistically motivated formalized resource, suitable also for Natural Language Processing applications, and building an application-oriented tool geared to meet the practical needs of specialists in the field. This interdisciplinary cooperation turned out to be an added value. In fact, with respect to other specialized thesauri, ArchiWN has the advantage of having a formalized structure and of inheriting linguistic oriented information from the generic WordNet; with respect to other lexical resources, it has the advantage that many synsets will be associated with images representing the concept.

Another distinguishing characteristic of ArchiWN with respect to other existing WordNet-like lexical resources is the fact that the synonyms will be ordered on the basis of their representativeness with respect to the concept they express: given a synset, the first synonym will be the word which is most commonly used by domain experts to express that concept.

In the creation of ArchiWN we had to face a number of problematic issues related both to the adoption of the MultiWN model and to the integration with MultiWN itself. In the

following Sections we discuss the different steps that have to be undertaken in order to build such a resource.

3 Adopting and Adapting the MultiWordNet Model

Two basic criteria have been followed in the construction of ArchiWN. First, we referred as much as possible to already existing and widely accepted specialized sources for the architecture and construction domain. Second, MultiWN information is exploited whenever possible to create those hierarchies for which a complete and well structured domain-specific terminology is not available.

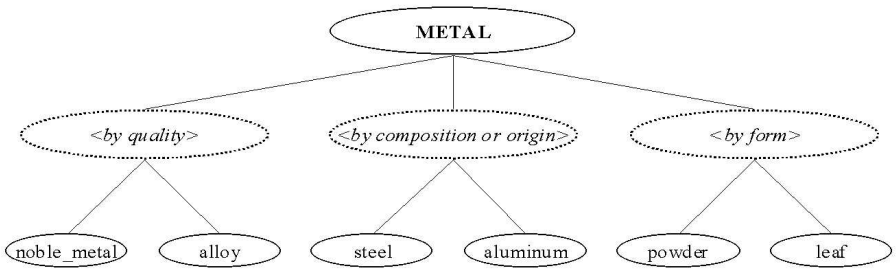
With regard to domain-specific sources, various specialized materials have been used to create both the synsets and the hierarchies of ArchiWN, among which the *Art and Architecture Thesaurus* (AAT) [6], the *Construction Indexing Manual* of CI|SfB [8], the international and national standardization rules (ISO, CEN, UNI), the *Lessico per la descrizione delle alterazioni e degradazioni macroscopiche dei materiali lapidei* created by the NORMAL commission, and other scientific literature in the area, technical dictionaries included. Both English and Italian sources are being used and correspondences between the two languages have to be found to create the bilingual synsets of ArchiWN. From the analysis of these sources, it turned out that very often they are not compatible with the MultiWN model. Either they are not structured on the basis of the ISA relation or they present mixed hierarchies where different levels are not homogeneous and relations between concepts are underspecified and ambiguous. On the contrary, relations in WordNet are explicit and information is encoded in a homogeneous way. Thus, it is necessary to reorganize these sources to make them compatible with the WordNet model. An example is given by the reorganization of the AAT hierarchy for the term “metal”, an excerpt of which is shown in Figure 1.

To make AAT compatible with the ArchiWN model, we had to interpret its spurious relations by disambiguating the type of relation connecting superordinate and subordinate concepts and by deciding how to manage intermediate “artificial” nodes which are not relevant from the point of view of the ISA hierarchy. As it can be seen in Figure 1, the artificial nodes have been eliminated and only the ISA relations have been maintained. The concepts previously connected to “metal” by a “form” relation have been modified, put in their appropriate ISA hierarchy, and connected to “metal” with the HAS-SUBSTANCE WordNet relation.

The second main source for the creation of ArchiWN, mainly used when a complete and structured domain-specific terminology is not available, is MultiWN itself. Synsets already existing in MultiWN which are considered appropriate by the domain experts are included into ArchiWN. However, this methodology cannot always be applied straightforwardly. In fact, as MultiWN synsets represent general language while ArchiWN must represent a specialized language, it is possible that both MultiWN synsets and relations are not always completely suitable for representing the architecture and construction domain. When included into ArchiWN, MultiWN synsets can undergo three different kinds of modification.

First, in those cases where the criterion for synonymy suitable for MultiWN is inadequate for ArchiWN, it is possible to add or delete synonyms to MultiWN synsets. This can happen as words that are considered synonyms in everyday usage may not be synonyms in the

AAT hierarchy



ArchiWN hierarchy

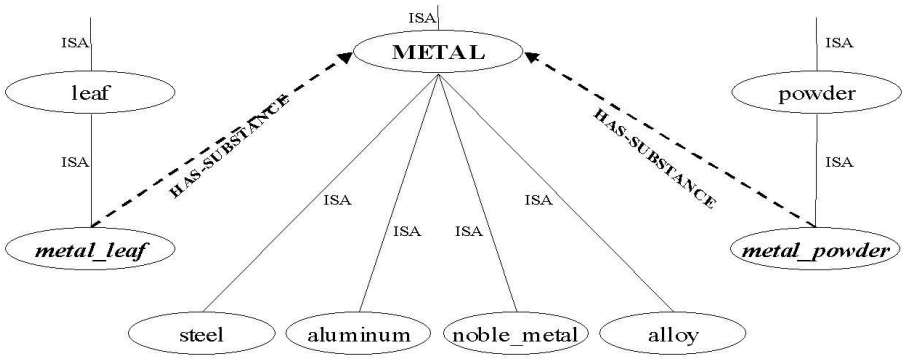


Fig. 1. Reorganization of the AAT hierarchy for "metal" according to the WordNet model

architecture domain. Second, when general language definitions are not compatible with a technical definition, it is possible to modify MultiWN definitions of the synsets. Third, it must be possible to delete and add relations between synsets. When included into ArchiWN, a synset can maintain all or some or none of its original MultiWN relations, depending on their appropriateness to the architecture domain. Moreover, other relations can be added to encode further information relevant to the specialized domain.

Finally, three new semantic relations, missing in MultiWN but useful to define concepts in the architecture domain, have been introduced in ArchiWN:

- HAS FORM (n/n) {tympanum} HAS FORM {triangle, trigon, trilateral};
- HAS ROLE (n/n) {metal section} HAS ROLE {upright, vertical};
- HAS FUNCTION[3] (n/v) {beam} HAS FUNCTION {to hold, to support, ...}.

[3] The HAS ROLE and HAS FUNCTION relations can be compared to the EuroWordNet [10] (EuroWN) INVOLVED/ROLE relation which connects second-order entities (i.e. nouns and verbs expressing properties, acts, processes, states, events) to first-order entities (i.e. concrete nouns referring to physical things). However, in EuroWN, the INVOLVED/ROLE relation is used for encoding information on arguments/adjuncts that are strongly implied in the meaning of a second-

4 Integrating ArchiWordNet with MultiWordNet

To integrate ArchiWN with MultiWN, a first list of 5,000 terms has been created relying on the specialized sources described above and on the direct experience of the domain experts. Then, the majority of such terms have been grouped in 13 semantic areas, as shown in Table 1. These semantic areas correspond to the main hierarchies to be represented in ArchiWN.

After the identification of the MultiWN nodes where to insert the ArchiWN hierarchies, the integration procedure requires (i) the actual inclusion of ArchiWN hierarchies in MultiWN, and (ii) the handling of the overlapping between terms present in both MultiWN and ArchiWN. This latter requirement is due to the fact that, unlike other domains characterized by a very specialized terminology, the architecture domain includes a significant amount of terms commonly used in the general language.

In the literature, different approaches are presented to address the problem of linking existing lexical/semantic hierarchies [3] and of integrating the information of a generic lexical resource with domain-specific information [9,5,1]. The methodology we developed to realize the integrated wordnet takes as a basis the "plug-in" approach proposed in [5] with some basic differences and extentions. In [5] two existing, independently created, wordnets are connected whereas ArchiWN is created so as to maximize the integration with MultiWN. Thus, to meet our needs, some existing procedures were extended and new procedures were created, especially for maximizing the exploitation of MultiWN information.

Our methodology consists of *basic operations* that can be performed on single MultiWN synsets and that constitute the basis of *complex procedures (plug-in)* which apply to entire hierarchies. The basic operations allow us to:

a) eclipse a synset;
b) tag a synset with the "architecture and construction" domain label;
c) add or delete relations to a synset;
d) add or delete synonyms in a synset;
e) modify the synset definition.

The eclipsing operation (a) removes a certain MultiWN synset and all relations originating from that synset. It is used to avoid overlappings when a specialized synset has been created in ArchiWN and a similar synset already exists in MultiWN but it is not considered suitable to be included into ArchiWN. The labeling operation (b) has the effect of including a MultiWN synset in ArchiWN, when this is considered suitable for the architecture and construction domain. It is used to avoid overlappings exploiting MultiWN information. Removing and adding relations to synsets (c) are the fundamental integration operations. Merging ArchiWN and MultiWN always requires adding one or more new relations to a synset (the root of the hierarchy in the case of complex procedures) and sometimes removing all or some of its original relations.

order verb/noun. For example, "to hammer" INVOLVED "hammer" and "hammer" ROLE "to hammer". On the contrary, given the specialized nature of ArchiWN, we are more interested in adding *encyclopaedic* information, concerning the usage of concrete entities within the architecture field. The HAS ROLE and HAS FUNCTION relations are used to encode the function of an entity; such function is not necessarily inherent in the semantics of the word designating the entity.

Finally, to customize MultiWN synsets to the architecture and construction domain operations of type (d) and (e) can be carried out (see Section 3).

To operate on ArchiWN and MultiWN hierarchies, we devised four complex procedures, able to cope with different integration requirements:

- *Substitutive plug-in.* A hierarchy from ArchiWN substitutes a MultiWN sub-hierarchy. This procedure, involving the eclipsing of all synsets in the MultiWN hierarchy, is used when an ArchiWN hierarchy is rich and well structured while the corresponding MultiWN one is not.
- *Integrative plug-in*. The two hierarchies are merged. The root of the ArchiWN sub-hierarchy substitutes the MultiWN one and the MultiWN hyponyms relevant to the architecture domain are included in ArchiWN through a labeling operation. This plug-in procedure is used when MultiWN has a well structured hierarchy and thus it is useful to integrate this information with the specialized one.
- *Hyponymic plug-in*. An ArchiWN hierarchy is connected as a hyponym of a MultiWN synset.
- *Inverse plug-in*. A MultiWN sub-hierarchy (possibly part of an eclipsed sub-hierarchy) is moved from MultiWN and connected to ArchiWN as a hyponym of an ArchiWN synset. This procedure is mainly used to exploit portions of MultiWN hierarchies which are considered relevant to the architecture and construction domain but are not in a correct position in MultiWN.

Given this methodology, we identified for each ArchiWN hierarchy one or more plug-in nodes in MultiWN and the complex procedures to be applied. As summarized in Table 1, some hierarchies can be directly plugged in MultiWN, while others required reorganizing MultiWN hierarchies. The results obtained in the integration phase are quite encouraging, showing not only that it is possible to integrate ArchiWN with MultiWN, but also that MultiWN can be widely exploited in the creation of ArchiWN hierarchies. In fact, for eight ArchiWN hierarchies we could exploit an integrative plug-in, while a substitutive plug-in was necessary for only three ArchiWN hierarchies. Finally, two ArchiWN hierarchies ("components of buildings" and "single buildings and buildings complexes") required a reorganization of some MultiWN sub-hierarchies, involving some plug-hyponymies, large synset eclipsing, but also a number of inverse plug-ins, which means the reuse of some MultiWN sub-hierarchies.

As regards the population of ArchiWN, up to now the "Simple buildings and building complexes" sub-hierarchy has been populated with about 900 synsets, containing in most cases both Italian and English synonyms along with an accurate definition.

This work has been done manually, using the MultiWN graphical interface which allows the user both to modify existing synsets and relations and to create new synsets.

During the creation of the bilingual synsets, we had to deal with the issue of lexical gaps, i.e. cases in which a language expresses a concept with a lexical unit whereas the other language does not. For example, the English synset for the word "kirk" (a Scottish church) has not an Italian correspondent and, viceversa, the Italian synset for "trullo" (a typical rural construction from Apulia, Italy) has not an English correspondent. However, this kind of idiosyncrasy does not represent a significant problem as it does not involve mismatches in the hierarchies. Moreover, as the specialized architecture lexicon mainly refers to objects and

Table 1. Integration of ArchiWN hierarchies with MultiWN

ArchiWN hierarchies	**MultiWN Plug-in nodes** (lemma/sense number)	**Type of plug-in**
Architectural styles	architectural_style/1	substitutive
Materials	material/1, substance/1	substitutive
Construction products	building_material/1	substitutive
Techniques	technique/1	integrative
Tools	tool/1	integrative
Components of buildings	structure/1, component/3, region/1	hyponymic
Single buildings and building complexes	structure/ArchiWN building/1, building_complex/1	hyponymic inverse
Physical properties	physical_property/1	integrative
Conditions	condition/1	integrative
Disciplines	discipline/1	integrative
People	person/1	integrative
Documents	document/1	integrative
Drawings and representations	drawing/2, representation/2	integrative

physical phenomena, in general we think that also for the remaining ArchiWN hierarchies we will not be faced with particularly problematic cross-linguistic idiosyncrasies.

5 Conclusion and Future Work

In this paper we have presented our experience in the creation of ArchiWN. The analysis of the problematic issues that arose, and the development and integration work carried out up to now show both that it is possible to integrate ArchiWN with MultiWN and that MultiWN itself can be considered a useful resource to be exploited in the creation of ArchiWN hierarchies.

With regard to future work, we will go on enriching the “Simple buildings and building complexes” hierarchy and populating the remaining hierarchies.

Moreover, we received a request from the Italian Architectural aluminium and steel manufacturers association (UNCSAAL) to create a multilingual specialized lexicon of approximately 1,000 synsets specifi cally referring to the window and curtain wall industry. In order to meet the needs of this industrial application, a further development of some of the hierarchies is planned, together with the extension of the resource to other languages such as German, French, and possibly Spanish.

ArchiWN’s range of applications will be twofold: it will be a thesaurus for cataloguing images within the SIS archive, and a useful integrated resource for Natural Language Processing applications. Moreover, an important achievement is represented by an agreement which is under way for the future usage of ArchiWN by the institutions in charge of cataloguing the architectural cultural heritage of the Piemonte region.

Acknowledgements

This project is being carried out under the precious scientific direction of Professor Gianfranco Cavaglià (Turin Polytechnic) and Fabio Pianesi (ITC-irst). Enrica Bodrato and Antonella Perin of the Turin Polytechnic have being compiling the thesaurus with competence and accuracy.

References

1. Buitelaar, P. and Sacaleanu, B.: Extending Synsets with Medical Terms. In: Proceedings of the First International Conference on Global WordNet, Mysore, India (2002).
2. Clark, P., Thompson, J., Holmback, H.and Duncan, L.: Exploiting a Thesaurus-Based Semantic Net for Knowledge-Based Search. In: Proceedings of AAAI/IAAI 2000, Austin, Texas (2000).
3. Daudé, J., Padro, L. and Rigau, G.: Mapping WordNets Using Structural Information. In: Proceedings of ACL 2000, Hong Kong (2000).
4. Fellbaum, C. (ed.): WordNet: an Electronic Lexical Database, The MIT Press, Cambridge (1998).
5. Magnini, B. and Speranza, M.: Integrating Generic and Specialized Wordnets. In: Proceedings of the Euroconference RANLP 2001, Tzigov Chark, Bulgaria (2001).
6. Petersen, T. (director): Art and Architecture Thesaurus, Oxford University Press, New York-Oxford (1994). `http://www.getty.edu/research/tools/vocabulary/aat/`.
7. Pianta, E. Bentivogli, L. and Girardi, C.: MultiWordNet: Developing an Aligned Multilingual Database. In Proceedings of the First International Conference on Global WordNet, Mysore, India (2002).
8. Ray-Jones, A. and Clegg, D.: CI|SfB. Construction Indexing Manual 1976, RIBA Publications, London (1991).
9. Turcato, D., Popowich, F., Toole, J., Fass, D., Nicholson, D., and Tisher, G.: Adapting a Synonym Database to Specific Domains. In: Proceedings of ACL 2000 Workshop on Information Retrieval and Natural Language Processing, Hong Kong (2000).
10. Vossen, P. (ed.): Computers and the Humanities, Special Issue on EuroWordNet, Volume 32, Nos. 2–3 (1998) 1 Buitelaar, P. and Sacaleanu, B.: Extending Synsets with Medical Terms. In: *Proceedings of the First International Conference on Global WordNet*, Mysore, India (2002).

Extending WordNet with Syntagmatic Information

Luisa Bentivogli and Emanuele Pianta

ITC-irst, Via Sommarive 18
38050 Povo – Trento, Italy
Email: bentivo@itc.it, pianta@itc.it

Abstract. In this paper we present a proposal to extend WordNet-like lexical databases by adding information about the co-occurrence of word meanings in texts. More specifically we propose to add *phrasets*, i.e. sets of free combinations of words which are recurrently used to express a concept (let's call them *Recurrent Free Phrases*). Phrasets are a useful source of information for different NLP tasks, and particularly in a multilingual environment to manage lexical gaps. At least a part of recurrent free phrases can also be represented through a new set of *syntagmantic* (lexical and semantic) WordNet relations.

1 Introduction

Most lexical information encoded in WordNet has a paradigmatic nature, that is if we take a word from a sentence in a real text, and consider which semantic and lexical relations are coded in WordNet with regard to that word, we will see that all relations hold between that word and other words that most probably do not occur in the same sentence or text. In Saussurean terms [5], paradigmatic relations occur *in absentia*, i.e. they hold with words that could in principle *substitute* each other rather than co-occur. On the other side, syntagmatic relations are *in praesentia*: they hold between words co-occurring in the same text. Syntactic relations are the best known kind of syntagmatic relations between words, whereas selectional restrictions between a verb and its arguments are a typical example of semantic syntagmatic relations [4].

As a matter of fact, information about the co-occurrence of words is not completely missing in WordNet. One can find such information in synonyms formed by more than one word and in gloss examples. However, WordNet includes only more-than-one-word synonyms that are elementary lexical meaning units, so they give information about the co-occurrence of words but not about the co-occurrence of meanings (as a more-than-one-word synonym involves only one meaning). On the other side, the information about the co-occurrence of words encoded in examples is not made explicit, and is out of the WordNet relational model.

In spite of the secondary role that syntagmatic relations play in WordNet, they are as relevant as paradigmatic relations both from a lexicographic and computational point of view. To have an idea of their lexicographic relevance, one only need to have a look at the space that examples of usage take in any dictionary entry, and it is every language learner's experience that an example of usage can be more useful than any definition to the comprehension of a word meaning. From a computational point of view, information about the co-occurrence of words is the most crucial, and in many cases, the only kind of information which is used

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 47–53.

for many NLP tasks. This is more and more true given the increasing role of statistics oriented, corpus based methods. In fact, co-occurrence is the most simple and effective kind of information that can be extracted from texts. A distinction needs to be done here between co-occurrence of words and co-occurrence of meanings. Whereas the former kind of information is indeed easily available in texts, the latter is much harder to be extracted, as it requires the disambiguation of texts. For this reason the encoding of information about the co-occurrence of meanings in a lexical resource as WordNet could be highly beneficial to the NLP community.

In the rest of this paper we will constrain the type of meaning co-occurrence information that we think should be encoded in WordNet. More specifically, in Section 2 we will concentrate on a set of expressions that we call Recurrent Free Phrases (RFPs). Then, in Sections 3 and 4 we will present two strategies to encode RFPs in WordNet. The first is based on a new data structure called *phraset*; the second is based on a new set of lexical and semantic relations. Finally, in Section 5 we will see that both dictionaries and corpora are useful sources of RFPs.

2 Recurrent Free Phrases

Following the Princeton WordNet model, synsets can include both single words and *multiword expressions*. See [3,10] for a recent discussion on the linguistic status of multiword expressions. More specifically WordNet includes *idioms*, that is relatively frozen combinations of words whose meaning cannot be built compositionally, and *restricted collocation*, that is combinations of words that combine compositionally but show a kind of semantic cohesion which considerably limit the substitution of the component words with synonyms. Multiword expressions must be distinguished from free combinations of words [2,7]. A *free combination* is a combination of words following only the general rules of syntax: the word meanings combine compositionally and can be substituted by synonyms. Whereas multiword expressions, along with single words, are elementary lexical units [4], free combinations do not belong to the lexicon and thus cannot compose synsets in WordNet.

However, as the boundaries between idioms, restricted collocations, and free combinations are not clear-cut, it is sometimes very difficult to properly distinguish a restricted collocation from a free combination of words. Moreover, applying this distinction in a rigorous manner leads to the consequence that a considerable number of expressions which are recurrently used to express a concept are excluded from wordnets as they are not lexical units.

For example, the English verb “to bike” is always translated in Italian with “andare in bicicletta” but the Italian translation equivalent seems to be a free combination of the word “andare” in one of its regular senses (dictionary definition: to move by walking or using a means of locomotion) with the restricted collocation “in bicicletta” (by bike). Expressions like “andare in bicicletta” contain relevant information about the co-occurrence of word meanings such as “andare” and “bicicletta, which should be independently coded in any Italian wordnet. We call these expressions Recurrent Free Phrases (RFPs). The main characteristics of RFPs are the following (some of them refer to the native speaker intuition, others are more corpus oriented):

i. RFPs are free combinations of words, which means that they fail the linguistic and semantic tests usually carried out to identify multiword expressions.

ii. RFPs are phrases, i.e. syntactic constituents whose head is either a noun or a verb or an adjective or a preposition. For instance, “eats the” is not an RFP.
iii. High frequency. E.g. “legge elettorale” (electoral law) is found at position 38 on a total of 2,108,000 bigrams extracted from an Italian reference corpus.
iv. High degree of association between the component words. For example, calculating association measures on the reference corpus, we found expressions like “paese europeo” (European country) which score very high.
v. Salience. It refers to the intuition of the native speaker lexicographer that a certain expression picks up a peculiar concept. The concept of salience is not necessarily related to frequency and word association. For example, our lexicographers think that “coscia destra” (right thigh) is less salient than “vertice internazionale” (international summit) whereas it has both a higher frequency and association score.

We are aware that, whereas characteristics from (i) to (iv) are all relatively well defined, the notion of salience is a little vague and needs more investigation. We make the hypothesis that salience is related to the amount of world knowledge that is attached to a certain expression and that cannot be simply derived from the composition of the meanings of the words making up the expression. To see this point consider the difference between “right thigh” and “right hand”. Both are fully compositional, but we feel that “right hand” is more salient than “right thigh”. The “right hand” is not only the hand that is attached to the right arm. This is also the hand we use to write, to swear etc. Note also that high frequency, high degree of association, and salience are all typical but not defining characteristics of RFPs.

RFPs can provide useful information for various kinds of NLP tasks, both in a mono- and multi-lingual environment. For instance, RFPs can be useful for knowledge-based word alignment of parallel corpora, to find correspondences when one language has a lexical unit for a concept whereas the other language uses a free combination of words. Another task which could take advantage of RFPs is word sense disambiguation. RFPs are free combinations of possibly ambiguous words, which are used in one of the regular senses recorded in WordNet. Take for instance the Italian expression “campo di grano” (cornfield). Its component words are highly ambiguous: “campo” has 12 different senses and “grano” 9, but in this expression they are used in just one of their usual senses. Now, suppose that when encoding RFPs, we annotate the component words with the WordNet sense they have in the expression; then, when performing word sense disambiguation, we only need to recognize the occurrence of the expression in a text to automatically disambiguate its component words.

Some RFPs are particularly relevant to the purposes of NLP tasks and we think they should be given priority for inclusion in any wordnet:

– RFPs expressing a concept which is not lexicalized in one language but is lexicalized in another language (i.e. in correspondence with a lexical gap).
– RFPs which are synonymous with a lexical unit in the same language.
– RFPs whose components are highly polysemous. This is meant to facilitate Word Sense Disambiguation algorithms.
– RFPs that are frequent, cohese and salient within a particular corpus considered as a reference corpus.

In the following two sections we will propose two ways of encoding in WordNet the co-occurrence information contained in RFPs, depending on their characteristics.

3 Extending WordNet with Phrasets

The first way to encode collocability information in wordnets is through the introduction of a new data structure called *phraset*, as proposed by [1]. A phraset is a set of RFPs (as opposed to lexical units) which have the same meaning. Phrasets can be added in correspondence with empty or non-empty synsets. We are currently studying the integration of phrasets in the framework of MultiWordNet [8], a multilingual lexical database in which an Italian wordnet has been created in strict alignment with the Princeton WordNet [6].

In a multilingual perspective, phrasets are very useful to manage *lexical gaps*, i.e. cases in which a language expresses a concept with a lexical unit whereas the other language does not. In MultiWordNet lexical gaps are represented by adding an empty synset aligned with a non-empty synset of the other language. Previously, the free combination of words expressing the non lexicalized concept was added to the gloss of the empty synset, where it was not distinguished from definitions and examples. With the introduction of phrasets, the translation equivalents expressing the lexical gaps have a different status, as it is shown in Example 1 below.

Phrasets are also useful in connection with non-empty synsets to give further information about alternative ways to express/translate a concept (Example 2).

Finally, it is important to stress that phrasets contain only free combinations which are recurrently used, and not definitions of concepts, which must be included in the gloss of the synset (Example 3). When the synset in the target language is empty and no expression is found in the phraset, this means that the target language lacks a synonym translation equivalent. The definition allows to understand the concept, but it is unlikely to be used to translate it.

Example 1

Eng-synset	{ toilet_roll }
Ita-synset	{ GAP }
Ita-phraset	{ rotolo_di_carta_igienica}

Example 2

Eng-synset	{dishcloth}
Ita-synset	{canovaccio}
Ita-phraset	{strofinaccio_dei_piatti, strofinaccio_da_cucina}

Example 3

Eng-synset	{straphanger}
Ita-synset	{GAP – chi viaggia in piedi su mezzi pubblici reggendosi ad un sostegno}
Ita-phraset	{ – }

Up to now 1,216 phrasets have been created in MultiWordNet, containing a total of 1,233 RFPs.

4 Extending WordNet with Syntagmatic Relations

In some cases word meaning co-occurrence information can be coded through semantic or lexical relations. Some steps in this direction have already been done in the framework of the MEANING project [9], an EU funded project aiming at enriching wordnets with semantic information useful for disambiguation purposes. One of the relations which is being added is the "envolve" semantic relation which encodes deep selectional restriction information, by relating verbal concepts with other concepts that typically occur as arguments (or participants) of the verb.

On the contrary, in our approach we deliberately focus on the kind of co-occurrence information that is not explained by selectional restriction phenomena. Consider for instance the RFP "campagna antifumo" (campaign against smoking). This expression is quite frequent in Italian newspapers, and shows a good degree of log-likelyhood association. Also the noun "campagna" in Italian is ambiguous between the meanings "campaing" and "country-side", but is monosemous in the above RFP, so it is worth including it in WordNet. If we choose to encode the co-occurrence of "campagna" and "antifumo" through a phraset, we need to create a new empty synset which is hyponym of "campagna" in the "campaign" sense, to add a phraset containing "campagna antifumo" in correspondence with such empty synset, and to annotate "campagna" and "antifumo" with their meanings in WordNet.

In principle we could follow a simpler strategy. If WordNet had a "has_constraint" relation relating nominal concepts with adjectival concepts that typically constrain the former, then all we would need to do is add an instance of such relation between the correct synsets for the noun "campagna" and the adjective "antifumo". The use of relations looks like as a concise and smart way of encoding meaning co-occurrence information. This has however a number of limitations:

- It is more suitable for representing bigrams than higher order n-grams. For instance we could somehow represent the fact that "campo" and "grano" co-occur in the RFP "campo di grano", but in this way we would lack the possibility of representing the fact that the two words are connected through the "di" (of) preposition. Also, using relations to represent RFP with more than two content words is completely impossible.
- It is not possible to represent the fact that two RFPs are synonyms.
- It is not possible to represent the fact that a certain RFP is the translation equivalent of a lexical unit in another language.
- It is not possible to represent restrictions on the order or the morphological features of the words of the RFP.

The solution currently adopted in MultiWordNet to represent syntagmatic relations tries to get the best of both phrasets and explicit relations. RFPs are indeed explicitly represented in phrasets, but a new lexical relation (composes/composed-of) between phrasets and synsets is used to annotate the senses of the words in the RFPs. Figure 1 shows how the RFP "campo di grano" is represented in MultiWordNet.

5 Recurrent Free Phrases in Dictionaries and Corpora

In [1] a study is presented to verify the possibility of acquiring RFPs from both dictionaries and corpora. First, we studied all the Translation Equivalents (TEs) of the Collins En-

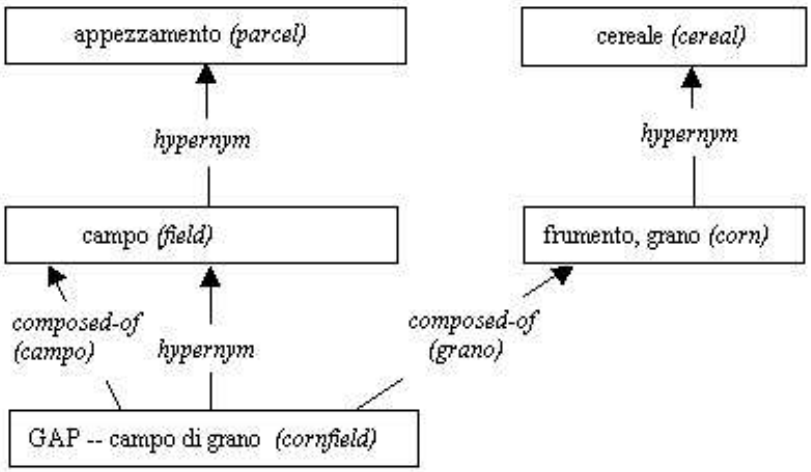


Fig. 1. Representing syntagmatic relations in MultiWordNet

glish/Italian dictionary corresponding to English to Italian gaps (7.8% of the total). By manually checking 300 Italian lexical gaps, a lexicographer found out that in 67% of the cases the TEs include a RFP. In the remaining cases the TEs are definitions. We used the result of this experiment to infer that more than half of the synsets which are gaps in any Italian wordnet potentially have an associated phraset.

In Section 3 we saw that phrasets can be associated also to regular (non empty) synsets. To assess the extension of this phenomenon, we first looked for cases in which the Collins dictionary presents an Italian TE composed of a single word, together with at least a TE composed of a complex expression. This happens in 2,004 cases (12% of the total). A lexicographer manually checked 300 of these complex expressions and determined that in 52% of the cases at least one complex expression is a RFP. In the remaining cases the complex expressions provided as TEs are either lexical units or definitions.

A second experiment has been carried out on an Italian corpus to compare multiword expressions and RPFs from a frequency point of view, and thus to assess the possibility of extracting RFPs from corpora with techniques similar to those used for collocation extraction. More specifically, we considered contiguous bigrams and trigrams with frequency higher then 3, and excluding stopwords. The results of the experiment show that, as expected, the number of bigrams that are lexical units decreases regularly along with the rank of the frequency, whereas non lexical units increase complementary. However, within non-lexical units the number of RFPs seems not to be correlated with the rank of the bigrams, fluctuating irregularly between a minimum of 3% and a maximum of 15%.

6 Conclusions and Future Work

We presented a proposal to extend the WordNet model with syntagmatic information about the co-occurrence of word meanings. This information is contained in RFP, that is free combinations of words characterized by high frequency, cohesion, and salience. Such expressions can be listed in phrasets (sets of synonymous RFPs), which are useful to handle lexical gaps in multilingual databases, and to provide alternative ways to express a concept in correspondence with regular synsets. The information contained in phrasets can be used to enhance word sense disambiguation algorithms, provided that each expression of the phraset is annotated with the specific meaning that its component words assume in the expression. The annotation of RFPs is implemented through a new lexical relation (composes/composed-of) relating phrasets and synsets. Evidence has been provided that RFPs can be extracted from both bilingual dictionaries and corpora with techniques similar to those used for collocation extraction. A lot of work need still to be done to better understand the lexicographic status of RFPs, and the practical implications of their inclusion in wordnets.

References

1. Bentivogli, L. and Pianta, E.: Beyond Lexical Units: Enriching WordNets with Phrasets. In: Proceedings of EACL-03, Budapest, Hungary (2003).
2. Benson, M., Benson, E., Ilson, R.: The BBI combinatory dictionary of English: a guide to word combinations. John Benjamins Publishing Company, Philadelphia (1986).
3. Calzolari, N., Fillmore, C., Grishman, R., Ide, N., Lenci, A., MacLeod, C., Zampolli, A.: Towards Best Practice for Multiword Expressions in Computational Lexicons. In: Proceedings of LREC 2002, Las Palmas, Canary Islands (2002).
4. Cruse, D. A.: Lexical semantics. Cambridge University Press, Cambridge (1986).
5. de Saussure, F.: Cours de linguistique générale. Payot, Paris (1916).
6. Fellbaum, C. (editor): WordNet: An electronic lexical database. The MIT Press, Cambridge, Mass. (1998).
7. Heid, U.: On ways words work together: research topics in lexical combinatorics. In: Proceedings of Euralex-94, Amsterdam, Holland (1994).
8. Pianta, E., Bentivogli, L., Girardi, C.: MultiWordNet: Developing an Aligned Multilingual Database. In Proceedings of the First International Conference on Global WordNet, Mysore, India (2002).
9. Rigau, G., Magnini, B., Agirre, E., Vossen, P., Carroll, J.: A Roadmap to Knowledge Technologies. In: Proceedings of COLING Workshop "A Roadmap for Computational Linguistics". Taipei, Taiwan (2002).
10. Sag, I., Baldwin, T., Bond, F., Copestake, A., Flickinger, D.: Multiword Expressions: a Pain in the Neck for NLP. In: Proceedings of CICLING 2002, Mexico City, Mexico (2002).

Exploiting ItalWordNet Taxonomies in a Question Classification Task

Francesca Bertagna

Istituto di Linguistica Computazionale, National Research Council,
Via G. Moruzzi 1, I-56100 Pisa, Italy
Email: francesca.bertagna@ilc.cnr.it

Abstract. The paper presents a case-study about the exploitation of ItalWordNet for Question Answering. In particular, we will explore the access to ItalWordNet when trying to derive the information that is crucial for singling out the answers to Italian Wh-questions introduced by the interrogative elements Quale and Che.

1 Introduction

The paper describes some aspects arised during the first phase of the work carried out for a Ph.D. research[1] dedicated to the exploration of the role of linguistic resources (from now on LRs) in a Question Answering (QA) application. The leading idea of the thesis is that the testing activity can highlight potentialities, together with problems and limitations, of the bulk of information collected during the last two decades by linguists and computational linguists. Altought LRs are not conceived to meet the requirements of a specific task (but rather to represent a sort of repository of information of general interest), they are significant sources of knowledge that should allow systems to automatically perform inferences, retrieve information, summarize texts, translate words in context from a language to another etc.. Computational lexicons storing semantic information, in particular, are supposed to provide a description of the *meaning* of the lexical units they collect. It is interesting to evaluate what is the *heuristic value* of such description and to what extent it is exploitable and useful to perfom specific tasks (e.g. in matching question and answer). Tons of papers have been written about the use of WordNet in IR and in QA and the time is mature to test also resources dedicated to languages other than English, such as, for instance, the Italian component of the EuroWordNet project (i.e. ItalWordNet). The first two sections of the paper will be devoted to briefly introduce the IWN project and the preliminar steps for question analysis. The core of the paper is represented by a sort of case-study dedicate to the description of the way the QA system can access the semantic information in IWN with the goal to derive what we call the Question Focus, the information crucial to match question and answer. Unfortunately, we are not able to provide validated results yet. We are in the process of assembling the available components of the QA downstream (the search engine, the chunker and the dependency parser, as well as the LRs) and we hope to be able to provide the first results soon. The current research is not collocated within a funded project but we hope to find occasion of fundings in the future.

[1] The Ph.D is carried out within a collaboration between Pisa University (Italy) and Istituto di Linguistica Computazionale of the National Council of Research. The grant is funded by the Italian National Council of Research.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 54–59.

2 ItalWordNet

The EuroWordNet (EWN) [11] project, retaining the basic underlying design of WordNet [7], tried to improve it in order to answer the needs of research in the computational field, in particular extending the set of lexical relations to be encoded between word meanings. In the last years an extension of the Italian component of EWN was realized[2] with the name of ItalWordNet (IWN) [10], inserting adjectives and adverbs, but also nouns and verbs which had not been taken into consideration yet in EWN. IWN follows exactly the same linguistic design of EWN (with which shares the Interlingual Index and the Top Ontology as well as the large set of semantic relation[3]) and consists now of about 70,000 word senses organized in 50,000 synsets.

3 Analysis of Italian Wh-Questions and Applicability for QA

Aiming at building a benchmark for Question Answering applications, we will concentrate our attention on factoid *Wh*-questions, which are supposed to be the forms more probably submitted by a user as a query. The corpus for QA consists now of about 800 Italian factoid *Wh*-questions, the majority of which obtained translating the TREC-9 question collection. We had also the opportunity to use the question collection from the first CLEF2003 (CL and monolingual) QA track [6]. The quality of the parser output can plays an important role in a QA application so a specific set of rules for the IDEAL Italian dependency parser [1] has been written[4]. On the other hand, a shallow parser (chunker) for Italian (CHUNK-IT) [4] provides us with the possibility to individuate information crucial for the task of question classification on the basis of the expected answer (i.e. what the user is looking for with his/her question). This information is the Question Stem (QS) and the Answer Type Term (ATT) [9]. The QS is the interrogative element we find in the first chunk of the sentence, while the ATT is the element modified by the QS (e.g. *Quanto costa un kg di pane?* or *Che vestito indossava Hillary Clinton in occasione di...?*)[5]. The convergence between these two information allows us to get closer to the expected answer type and to the text portion plausibly containing the answer. Some QSs (for example, *Quando* and *Dove*) allow the system to establish univocal correspondences between them and specific QFs. The relation between QF and QS is not bidirectional: to the same type of question can correspond different QFs (e.g. *Come si chiamava la moglie di JFK?* Vs *Come morì Janice Joplin?*[6]), and the same QF can be looked for via different QSs (e.g. *Quale poeta ha scritto la Divina Commedia?* Vs *Chi ha scritto la Divina Commedia?*[7]). We talked about *multi-strategies QA* because each QS has to be dealt with in its specificity. In what follows we will concentrate our attenction only on the interrogative elements of the Italian *Wh*-questions for handling which we have to explore information stored in LRs: the Question Stems *Che* and *Quale*.

[2] Within the SI-TAL project.

[3] For a complete list of the available semantic relations cf. [10]

[4] A detailed description of this phase and the results are in [2]

[5] *How much does a kg of bread cost?* Or *Which dress did Hillary Clinton wear when.....?*

[6] *What is JFK's wife name?* Vs *How did Janice Joplin die?*

[7] *Which poet wrote the Divina Commedia?* Vs *Who wrote the Divina Commedia?*

3.1 (Che|Quale)-questions

In capacity as interrogative adjective, *Che* is ambiguous between an interpretation selecting individuals and classes: when it is used to ask about an individual to be chosen among a group it overlaps, especially in North Italy, to the interrogative element *Quale*. For both, it is true the same consideration: generally, the QF refers to the entity belonging to the type of the noun modified by the interrogative adjective. For example, the answer of a question like: *Quale mammifero vive in mare?*[8] can be extracted from sentences like: *la balena vive nell'Oceano Atlantico*[9] where the informative links allowing the recognition of the answer are:

{Balena 1} –HAS_HYPERNYM → {cetaceo 1} –HAS_HYPERNYM → {mammifero 1};
{Atlantico 1} –BELONGS_TO_CLASS → {oceano 1} –HAS_HYPERNYM → {acque 1} –HAS_HYPONYM → {mare 1};

In this case we can lexically single out the QF searching among the hyponyms of the noun. This type of question is one of the most complex since the system has to resort to an additional lexical-semantics analysis module and the exploitation of language resources can make the difference. The need of an information stored in a lexical-semantics resource is also evident when we find questions like: *Quale stretto separa il Nord America dall'Asia?*[10] and *Quale parco nazionale si trova nello Utah*?[11].

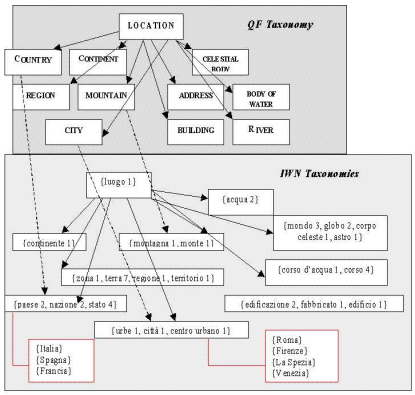


Fig. 1. Mapping the node Location of the QfTaxonomy on IWN

The semantic type of the noun modified by the interrogative adjective is the only thing able to tell us that we have to look for a named entity of the type *location* in the candidate

[8] *Which mammal lives in the sea?*
[9] *Whales live in Atlantic Ocean.*
[10] *Which strait separates North America and Asia?*
[11] *Which national park is in Utah?*

answer. These questions are not introduced by the interrogative adverb *Dove* (*Where*), but they are indeed used to ask about a location. But how do we derive the information that maps the *stretto* or the *parco nazionale* of the questions into the QF Location? In IWN, {parco nazionale 1} is a hyponym of {territorio 1, regione 1, zona 1, terra 7} while {stretto 1} is a hyponym of {sito 1, località 1, posto 1, luogo 2} and these areas of the IWN taxonomies can easily be mapped onto the Question Focus Location. The problem is that, when we want to project the QF Location on the IWN taxonomies, we have to address it on scattered and different portions of the semantic net. The node Location of the Question Focus taxonomy is mappable on the synset {luogo 1 – *parte dello spazio occupata o occupabile materialemte o idealmente*}, that has 52 first level hyponyms and that can be further organized in other sub-nodes, such as: country, river, region, etc. The major part of these taxonomies is leaded by the same synset {luogo 1}, which circumscribes a large taxonomical portion that can be exploit in the QF identification. To this area we also have to add other four sub-hierarchis {corso d'acqua 1, corso 4 – *l'insieme delle acque in movimento* }, {mondo 3, globo 2, corpo_celeste 1, astro 1}, {acqua 2 – *raccolta di acqua*}, {edificazione 2, fabbricato 1, edificio 1 – *costruzione architettonica*}. Figure1 gives an idea of this situation: the circumscribed taxonomical portion includes the nodes directly mapped on the QFs, all their hyponyms (of all levels) and all the synsets linked to the hyerarchy by means of the BELONGS_TO_CLASS/HAS_INSTANCE relation. A different way to group the IWN lexical items together is recurring to the EWN Top Ontology (TO). The EWN architecture allows us to select and circumscribe wide lexicon portions, kept together by: i) the links between the monolingual database and the ILI portion hosting the Base Concepts, ii) the links between the Base concepts and the TO, iii) the ISA relations linking the synset corresponding to the Base Concept with its conceptual subordinates of n level, from the top to its leaf nodes. In the case of QF Location, for example, we can extract all the synsets belonging to the Top Concept PLACE. The problem is that *River*, *Celestial_Body* and *Building* belong to other ontological portions ((*River* and *Celestial_Body* are classified as *Object/Natural* while *Building* as *Artifact*/*Building/ Object*) (see Figure 2). The Top Concepts *Object* and *Artifact* are too generic and not discriminating in the selection of the lexical area pertinent to the respective QFs. Thus the exploitation of the Top Ontology nodes can not be the default methodology for indivituating the relevant synsets[12]. The case of Location is only an example of the necessity to (manually) link the highest and most pertinent nodes of the lexical resources to the QFtaxonomy. We are now in the process[13] of adding a new module containing the almost 50 nodes of the QFTaxonomy to the IWN data structure, specifying, when possible, the subsumption links between the synsets and the type of expected answer. The internal ontological structure of ItalWordNet is obviously very different from the QFTaxonomy and it seems that the above mentioned strategy is much more practicable when working with concrete entities than with abstract entities. In *Quali conseguenze ha la pioggia acida?*[14], the candidate answer *L'impoverimento del terreno deriva dalle pioggie acide*[15] contains the answer element *impoverimento*, which is a

[12] The hypothesis of a hybrid strategy which uses both the Top Concepts and the lexical nodes has to be evaluated.

[13] Using the ItalWordNet tool.

[14] *Which are the consequences of the acid rain?*

[15] *the impoverishment of the soil derives from acid rain*

direct hyponym of the abstract noun *conseguenza*.[16] But in the question-answer pair: *Quale funzione ha la milza? La milza produce linfociti*[17] there is no hyponymy relation between *funzione* and *produrre*. In this case we should be able to resort to more complex inferences, as we see in Figure 3.

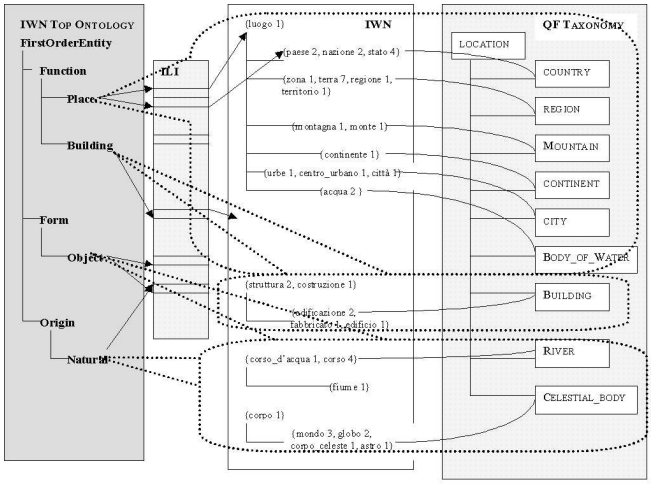


Fig. 2. Projection of the nodes of the QF Location on the EWN TO

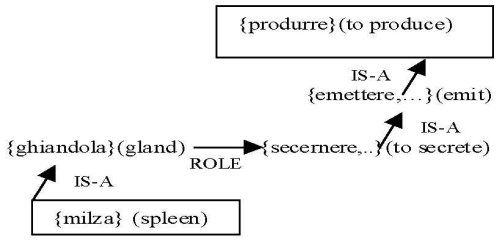


Fig. 3. An inferential path through the IWN synsets

[16] Another informative link is the semantic relatedness between the verb *derivare* (*to derive*) and the noun *conseguenza* (*consequence*), expressed in IWN by mean of a XPOS_NEAR_SYNONYM link between the synsets {derivare 1, conseguire 3,. . . , risultare 1} and {risultato 1, esito,. . . ., conseguenza 1}.

[17] *Which is the function of the spleen? The spleen products lymphocytes*

4 Future work

In the next step of our work we will try to provide a systematic analysis of the types of inference needed in the task of matching question and answer (very insightful in this sense is the work on *lexical chains* by [8]). We will verify whether it is possible to derive such inferences from the connections already stated in IWN by mean of the large set of semantic relations. It has to be avaluated also the impact of dynamic extraction of paraphrasis and inferential rules from texts [3,5], which constitutes a bottom-up approach leading to a notion of *meaning* inspired by distributional criteria. The idea is that dynamically boosting the "inferential" potentialities of static, hand-generated LRs can plays an important role in fi lling the gap between question and answer and, more generally, that the interplay between static lexical information and dynamic information acquired from text via processing is one of the way LRs could be improved and renewed in the future.

References

1. Bartolini R., Lenci A., Montemagni S., Pirrelli V., *Grammar and Lexicon in the Robust Parsing of Italian: Towards a Non-Naïve Interplay*, in Proceedings of COLING 2002 Workshop on Grammar Engineering and Evaluation, Taipei, Taiwan (2002).
2. Bertagna F., *Parsing Italian Wh-Questions*, ILC Internal Report, in prep.
3. Hermjakob U., Echihabi A., Marcu D., *Natural Language Based Reformulation Resource and Web Exploitation for Question Answering*, Proceeding of TREC-2002, (2002).
4. Lenci A., Montemagni S., Pirrelli V., *CHUNK-IT. An Italian Shallow Parser for Robust Syntactic Annotation*, in Linguistica Computazionale, Istituti Editoriali e Poligrafici Internazionali, Pisa-Roma, ISSN 0392–6907 (2001).
5. Lin D., Pantel P., *Discovery of Inference Rules for Question Answering*. In Natural Language Engineering 7(4):343–360 (2001).
6. Magnini B., Romagnoli S., Vallin A., Herrera J., Penas A., Peinado V., Verdejo F., de Rijke M., *The Multiple Language Question Answering Track at CLEF2003*, Working Notes for the CLEF2003 Workshop, Norwey (2003).
7. Miller, G., Beckwith R., Fellbaum C., Gross D., Miller K. J., *Introduction to WordNet: An On-line Lexical Database*. In International Journal of Lexicography, Vol.3, No.4 (1990) 235–244.
8. Moldovan D., Harabagiu S., Girju R., Morarescu P., Lacatusu F., Novischi A., Badulescu A., Bolohan O., *LCC Tools for Question Answering*, Proceeding of TREC-2002 (2002).
9. Paşca M., *Open-Domain Question Answering from Large Text Collections*, CSLI Studies in Computational Linguistics, USA (2003).
10. Roventini A., Alonge A., Bertagna F., Calzolari N., Girardi C., Magnini B., Marinelli R., Speranza M., Zampolli A., *ItalWordNet: Building a Large Semantic Database for the Automatic Treatment of Italian*. In Zampolli A., Calzolari N., Cignoni L. (eds.), Computational Linguistics in Pisa, Special Issue of Linguistica Computazionale, Vol. XVIII-XIX, Istituto Editoriale e Poligrafico Internazionale, Pisa-Roma (2003).
11. Vossen, P. (ed.), *EuroWordNet General Document*, 1999.

Morphosemantic Relations In and Across Wordnets

A Study Based on Turkish

Orhan Bilgin, Özlem Çetinoğlu, and Kemal Oflazer

Sabanci University, Human Language and Speech Technologies Laboratory
Istanbul, Turkey
Email: orhanb@sabanciuniv.edu, ozlemc@sabanciuniv.edu,
oflazer@sabanciuniv.edu

Abstract. Morphological processes in a language can be effectively used to enrich individual wordnets with semantic relations. More importantly, morphological processes in a language can be used to discover less explicit semantic relations in other languages. This will both improve the internal connectivity of individual wordnets and also the overlap across different wordnets. Using morphology to improve the quality of wordnets and to automatically prepare synset glosses are two other possible applications.

1 Introduction

Over the recent years, wordnets have become important resources for natural language processing. The success of Princeton WordNet (PWN) [1] has motivated the development of several other wordnets for numerous other languages[1].

Wordnets are lexical semantic networks built around the concept of a 'synset', a set of lexical items which are synonymous in a certain context. Semantic relations such as hyperonymy, meronymy and antonymy link synsets to each other and it is these semantic relations that give wordnets their essential value.

The number of semantic relations among synsets is an important criterion of a wordnet's quality and functionality. Thus, any method that would facilitate the encoding of semantic relations will be greatly helpful for wordnet builders. Furthermore, the recent proliferation of wordnets opened up the possibility of cross-linking across wordnets.

In this paper, we claim, with special emphasis on Turkish, that morphological processes in individual languages offer a good starting point for building wordnets and enriching them with semantic information encoded in other wordnets[2].

The present paper is structured as follows: Section 2 describes possible applications of the proposed method in monolingual and multilingual contexts. Section 3 provides an overview of the methodology in language-independent terms. Section 4 clarifies this methodology further by providing a case study for Turkish morphology and the possibility of exporting

[1] See "Wordnets in the World" at http://www.globalwordnet.org/

[2] The exchange of semantic relations across languages requires that the importing wordnet and the exporting wordnet are linked to each other in some way. The EuroWordNet project [2] and the BalkaNet project [3] solved this by introducing the concept of an 'Interlingual Index' (ILI), a common repository of language-independent concepts to which all other languages would be linked.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 60–66.

semantic relations from Turkish into English. Section 5 draws conclusions and provides insights regarding possible future work.

2 Areas of Application

Possible applications of the methodology proposed in this paper can be more formally described as follows:

Simple morphological derivation processes in a certain Language A can be used (i) to extract explicit semantic relations in Language A and use these to enrich Wordnet A; (ii) to detect mistakes in Wordnet B; (iii) to automatically prepare machine-tractable synset glosses in Wordnet A and/or Wordnet B; and most importantly (iv) to discover implicit semantic relations in Language B and use these to enrich Wordnet B.

The following three subsections clarify these applications in monolingual and multilingual contexts.

2.1 Monolingual Context: Single, Isolated Wordnet

Using morphologically-related word pairs to discover semantic relations is by far faster and more reliable than building them from scratch. Morphology is a relatively regular and predictable surface phenomenon. It is a simple task to extract from a wordlist all instances which contain a certain affix, using regular expressions. Using morphological relations to discover semantic relations is a good way to start a wordnet from scratch or enrich an existing one.

2.2 Multilingual Context: Several Wordnets Linked to Each Other

The more interesting application of the method is the sharing of semantic information across wordnets. There are two cases: i. Semantically-related lexical items in both the exporting and the importing language are morphologically related to each other.

In this case, the importing language (Turkish) could have discovered the semantic relation between “deli” (mad) and “deli**lik**” (mad**ness**), for instance, by using its own morphology. So, the benefit of importing the relation from English is quite limited. Still, importation can serve as a very useful quality-control tool for the importing wordnet, and this has indeed been the case for Turkish WordNet:

While building a wordnet for Turkish, the so-called “expand model” [4, p. 52] was used and synsets were constructed by providing translation equivalents for PWN synset members. Following the translation phase, a series of relations, e.g. STATE_OF relations, were imported from PWN. Since Turkish employs a morphological process to encode STATE_OF relations, the list of Turkish translation equivalents contained several morphologically-related pairs like “deli-deli**lik**” (mad-mad**ness**), “garip-garip**lik**” (weird-weird**ness**), etc. Pairs that violated this pattern probably involved mistranslations or some other problem, and the translation method provided a way to detect such mistakes.

ii. Semantically-related lexical items in the importing language are not morphologically related to each other.

In this more interesting case, the semantic relation is morphologically generated in the exporting language (Turkish) but not in the importing language (English)[3]. The causation relation between the lexical items “yıkmak” and “yık**ıl**mak”, for instance, is obvious to any native speaker (and morphological analyzer) of Turkish, while the corresponding causation relation between “tear down” and “collapse” is relatively more opaque and harder to discover for a native speaker of English and impossible for a morphological analyzer of English. Our method thus provides a way of enriching a wordnet with semantic information imported from another wordnet. Furthermore, the proposed method improves overlap among different wordnets as they borrow semantic links from each other.

2.3 Monolingual and/or Multilingual Context

A possible application in a monolingual and/or multilingual context is to automate the preparation of formal and thus machine-tractable synset glosses, based on the information imported from another language’s wordnet. Equipped with the information that the Turkish synsets for “yıkmak” (tear down) and “yık**ıl**mak” (collapse) are linked to each other via a “CAUSES” relation, one can safely claim that the English synset “tear down” can be glossed as “cause to collapse”. Similarly, the builders of a Turkish wordnet can safely claim that their synset for “yıkmak” can be glossed as “yık**ıl**masına neden olmak”, which is the Turkish equivalent of “cause to collapse”.

3 Methodology

The methodology that will enable the above-described applications involves the following language-independent steps:

3.1 Determining the Derivational Affixes

All derivational affixes in the exporting language are potential candidates. Some of these have a perfectly regular and predictable semantics, while some others do not. Affixes can also be ranked according to their productivity. An affix that can be attached to almost any root in the language in question is regarded as a productive affix. Thus, two criteria have to be taken into consideration while deciding to include an affix in the list: (i) the regularity of its semantics; and (ii) its productivity.

3.2 Constructing Morphosemantically-Related Pairs

Using a wordlist available to the exporting language, we extract all instances containing the affix we are interested in. Simple regular expressions are sufficient for this task. We then feed all of these instances to a morphological analyzer. If there is at least one morphological analysis that suggests the expected derivation process, this instance is included in the list of potential pairs.

[3] This phenomenon has also been discussed in [5, p. 11]

The morphological analysis also provides us with the root involved in the derivation process. Thus, we obtain a list of pairs such as “teach-teach**er**” or “hang-hang**er**”.

Almost all candidates which seem to, but do not actually, contain the relevant affix (such as moth-moth**er**) can be automatically eliminated by using morphological analysis results. In the case of the pair “moth-mother”, the morphological analysis of “mother” does not contain the analysis “moth+Agent” and this pair can thus be safely eliminated from the list.

3.3 Linking the Right Synsets via the Right Relation Type

The pairs generated in the last step are merely word forms and not word senses. For the correct assignment of a semantic link, we need to assign the correct sense to both members of the pair.

Faced with the ambiguous pair “regulate-regulator” the lexicographer has to decide: (i) that the verb ‘regulate’ in this pair is ‘regulate (sense 2)’ (“bring into conformity with rules or principles or usage; impose regulations”) and not ‘regulate (sense 5)’ (“check the emission of (sound)”); (ii) that the noun ‘regulator’ in this pair is ‘regulator (sense 2)’ (“an official responsible for control and supervision of an activity or area of public interest”) and not ‘regulator (sense 1)’ (“any of various controls or devices for regulating or controlling fluid flow, pressure, temperature, etc.”);(iii) that the resulting semantic relation involves “the second semantic effect of the suffix -or”. (“the person who regulates” and not “the device that regulates”).

4 Application of the Methodology to Turkish

Turkish, an agglutinative language with productive morphological derivation processes, employs several affixes which change the meaning of the root in a regular and predictable way [6]. There are some others which have a more complex semantics and change word meaning in more than one way. It is usually possible to specify most semantic effects of an affix and conclude, for instance, that the Turkish agentive suffix -CH[4] basically has four separate effects. Obviously, there are some fuzzy cases where it is difficult to specify the exact semantic effect. These cases usually involve semantic shifts and lexicalizations.

Table 1 illustrates Turkish suffixes we have identified as useful candidates[5].

Table 2 provides examples of morphosemantically-related pairs of Turkish words and the corresponding semantically-related pairs in English. This table clearly shows that productive and predictable morphological derivation processes in Turkish allow us to discover morphologically unrelated English words which are semantically related to each other.

The current wordlist for Turkish contains substantial numbers of words involving the suffixes listed in Table 1. We have identified the following number of instances for each suffix in Table3

[4] Throughout the following discussion of Turkish suffixes, H represents a meta-character denoting the high vowels ‘ı , i, u, ü’; A the vowels ‘a, e’; D the consonants ‘d, t’; and C the consonants ‘c, ç’. Thus each morpheme here actually stands for a set of allomorphs.

[5] We have used the semantic relation tags defined in Princeton WordNet and EuroWordNet whenever possible. These have been indicated in boldface type throughout this paper.

Table 1. List of Turkish suffixes and their semantic effects (* n = noun, v = verb, a = adjective, b = adverb)

SUFFIX	POS*	SEMANTIC EFFECT
-lAş	n-v, a-v	BECOME
-lAn	n-v	ACQUIRE
-lHk	a-n, n-n	**BE_IN_STATE**
-lH	n-a	1) SOMEONE_WITH 2) SOMETHING_WITH 3) SOMEONE_FROM
-sHz	n-a	1) SOMEONE_WITHOUT 2) SOMETHING_WITHOUT
-sAl	n-a	**PERTAINS_TO**
-(y)lA	n-b	WITH
-Hş	v-v	RECIPROCAL
-(H)l	v-v	**CAUSES**
-(H)t, DHr, -(H)r, -(A)r	v-v	**IS_CAUSED_BY**
-Hş	v-n	ACT_OF
-CA	a-b, n-b	MANNER

Table 2. Examples of Turkish-English Pairs

taş	taş**laş**mak	**INVOLVED_RESULT**
stone	petrify	
iyi	iyi**leş**mek	BECOME
good	improve	
hasta	hasta**lı k**	**STATE_OF**
sick	disease	
din	din**siz**	SOMEONE_WITHOUT
religion	infidel	
ölmek	öl**dür**mek	**IS_CAUSED_BY**
die	kill	
omurga	omurga**lı**	SOMEONE_WITH
spine	vertebrate	

A detailed analysis of two Turkish suffixes produced the results summmarized in Table4.

The two suffixes we have investigated are -DHr and -lA¸s, encoding CAUSES and BECOME relations, respectively.

Despite the fact that Turkish wordnet is a small-sized resource (10.000 synsets), it contains a significant number of synsets involving these morphosemantic relations.

In only a few cases does PWN 2.0 indicate a CAUSES relation between the respective synsets. In the case of the BECOME pairs, PWN 2.0 provides the underspecified relation called "ENG DERIVATIVE".

Some of the new links proposed involve morphologically unrelated lexical items which cannot be possibly linked to each other automatically or semi-automatically. Interesting examples in the case of the BECOME relation include pairs such as soap-saponify, good-improve, young-rejuvenate, weak-languish, lime-calcify, globular-conglobate, cheese-caseate, silent-hush, sparse-thin out, stone petrify. Interesting examples in the case of the CAUSE relation include pairs such as dress-wear, dissuade-give up, abrade-wear away, encourage-take heart, vitrify-glaze.

Table 3. Number of derived words for each Turkish suffix

SUFFIX	# OF PAIRS	POSSIBLE RELATIONS
-lHk	4,078	**BE_IN_STATE**
-lH	2,725	WITH
-sHz	1,001	WITHOUT
-Hş	991	ACT_OF
-lAn	758	ACQUIRE
-lAş	763	BECOME
-DHr	782	**CAUSES**
-CA	710	MANNER
-sAl	115	**PERTAINS_TO**
TOTAL	11,923	

Table 4. Statistics for twoTurkish suffixes

RELATION	# IN WORDLIST	# IN TWN	# IN PWN	% OF NEW LINKS
CAUSES	1511	80	18	77.5%
BECOME	763	83	11	86.7%

5 Conclusions and Future Work

We have tried to demonstrate that morphology offers a good starting point for enriching wordnets with semantic relations. More importantly, we have claimed that sharing morphosemantic relations across languages is an efficient way of enriching wordnets with semantic relations that are hard to discover. We have shown, at least for the case of Turkish, that there are

a large number of instances involving such predictable morphological phenomena that can be fruitfully exploited for semantic relation discovery.

Future research could concentrate on automating the decision task mentioned in Sect. 3.3. The outcome of a morphological derivation process is mutually determined by the semantics of the root and the affix. Thus, there is no real "decision" involved in steps (ii) and (iii) described in Sect. 3.3. For instance, the agentive suffix -CH in Turkish is capable of producing: (i) "commodity – seller/manufacturer" pairs if the root is a marketable artefact; (ii) "person – adherent" pairs if the root is a proper noun; (iii) "instrument – musician" pairs if the root is a musical instrument, etc.

As soon as we decide that the agentive suffix -CH is attached to the root "keman" (violin) in its, say, second sense (violin as a musical instrument), we are forced to conclude that the "musician" effect OR the "seller/manufacturer" effect and NOT the "adherent" effect of the suffix is at play here. Although we cannot fully disambiguate in the absence of additional contextual and pragmatic information, we can at least rule out the possibility that the "adherent" effect might be involved.

Using the hierarchy, and more fruitfully the top ontology, of a wordnet, we can obtain additional semantic information regarding the root and predict the semantic effect the affix will have when applied to this root. The success of such a study remains to be seen.

References

1. C. Fellbaum (ed.), WordNet: An electronic lexical database, Cambridge, MIT Press, 1998.
2. P. Vossen (ed.), EuroWordNet: A Multilingual Database with Lexical Semantic Networks, Dordrecht, Kluwer Academic Publishers, 1998.
3. "BalkaNet: A Multilingual Semantic Network for the Balkan Languages", [online], `http://www.ceid.upatras.gr/Balkanet` (Accessed: August 23, 2003).
4. P. Vossen, "EuroWordNet General Document", [online], `http://www.illc.uva.nl/EuroWordNet/docs.html` (Accessed: August 23, 2003).
5. C. Fellbaum and G. A. Miller, "Morphosemantic Links in Wordnet", in press, Traitement Automatique de Langues.
6. K. Oflazer, "Two-level Description of Turkish Morphology", Literary and Linguistic Computing, Vol. 9, No. 2, 1994.

A Prototype English-Arabic Dictionary Based on WordNet

William J. Black and Sabri El-Kateb

UMIST, Department of Computation, Manchester, M60 1QD, UK
Email: wjb@co.umist.ac.uk, Sabri.El-Kateb-2@student.umist.ac.uk

Abstract. We report on the design and partial implementation of a bilingual English-Arabic dictionary based on WordNet. A relational database is employed to store the lexical and conceptual relations, giving the database extensibility in either language. The data model is extended beyond an Arabic replication of the word↔sense relation to include the morphological roots and patterns of Arabic. The editing interface also deals with Arabic script (without requiring a localized operating system).

1 Introduction

Our goal is the development of an expandable computer-based lexical and terminological resource to aid the working translator or information scientist working with technical terminology in Arabic. [3] The plan has been to use a relational database representation of the Wordnet as a backbone on which to hang translation equivalents and information about domain-specifi c technical terminology. We are therefore concerned with the potential for the WordNet data model to be extensible. Accounts of earlier versions of the design are given in [2,1]. The present paper gives an up-to-date picture of the data model and design, together with information on implementation and on the lexicographer's user interface.

The EuroWordNet [7,8] approach to multilingual resource development has emphasized the separate integrity of the dictionaries in the different languages, and provided an additional bilingual index to support the search for translations. The effort reported here is on an altogether more limited scale, and stores the data for the different languages in the tables of a single database. In keeping with this small scale, the bilingual dictionary does not currently maintain either glosses or examples in the second language, although there is nothing to prevent the data model being so augmented in the future.

When considering languages more closely related to English, developing a multilingual wordnet can be as simple as providing the mapping of foreign words to synsets. Arabic has an extensive system of derivational morphology that embodies important semantic relations, which ought to be reflected in any conceptual dictionary. The prototype dictionary described here embodies these kinds of lexical relation as well as those present in the WordNet. It also supports Arabic script rather than relying on a transliteration.

The remaining sections discuss Arabic morphology; the data model used and its practical realization in a DBMS; the encoding of Arabic morphological information; the facilities of the current user interface for editing and updating the data; how lexical mismatches are handled.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 67–74.

2 Arabic Morphology

Arabic morphology is described as “non-concatenative”, not because of any absence of prefi xes and suffi xes, but because affi xation is not the only morphological process supporting inflection and derivation.

Arabic [4] has a word structure whereby related forms share a sequence of three or four consonants, following each of which are different vowels, according to the form. That is, words have a basic structure CVCVCV or CVCVCVCV. Prefi xes and suffi xes also contribute to the differentiation of forms. There are only three distinct vowels /a/, /ι/ and /u/, but these also come in long variants, indicated in transliterations by a following colon.

2.1 Arabic Script

Mst ltrt nglsh spkrs cn dcd txt n whch thr r nly cnsnts, thanks to the redundancy in the script. Arabic readers do this all the time, because most vowels are suppressed from the written language, including dictionary citation forms. The vowels can be indicated by diacritics placed above or below the consonant that precedes them, when necessary for expository purposes.

In addition to the three vowels, there are 25 consonants in the script, and as Arabic is a cursive script, the letters take different forms according to whether they occur in initial, medial or fi nal position in the written word.

Table 1 illustrates the way that semantically related forms are derived from a common root, with a set of words sharing the consonant sequence /w/ /l/ /d/. (The Arabic script letters for these consonants are و, ل and د respectively.)

Table 1. Words derived from a common root

Word	Translit.	Pattern	Pattern translit.	English
وِلَادة	wila:dah	فِعَالَه	fi’a:lah	delivery
توليد	tawli:d	تَفعِيل	taf’i:l	generation
تَوَالد	tawa:lud	تَفَاعُل	tafa:’ul	reproduction
وَالد	wa:lid	فَاعِل	fa:’il	male parent
مولُود	mawlu:d	مَفعُول	maf’u:l	new born baby
مولد	mawlid	مَفعِل	maf’il	birth

2.2 Inflection and Derivation

The same kinds of word change are used to inflect as well as derive forms in Arabic. Inflected forms do not customarily occur in printed dictionaries, and are therefore not of interest to the dictionary compiler. Whilst an on-line dictionary like the WordNet can allow users to enter

queries with inflected forms, if there is a morphological analyser or lemmatizer component, dictionary users know that it is the base or citation form they should expect to use.

Derivational morphology is another matter. In conventional dictionaries, it is customary for some derived forms to be made completely subsidiary to the headword, rather than having a separate entry. In WordNet 2.0, derivational relations between nouns and verbs can be traced, and these relations ought to be traceable in any other dictionary based on conceptual principles. Arabic dictionaries (mono- or bi-lingual) are sometimes ordered according to morphological roots, with large numbers of forms (possibly out of alphabetic sequence) being listed subsidiary to them.

In Arabic, speakers are much more conscious of derivational morphology, since the bulk of the vocabulary has a systematically encoded derivation from a few thousand roots (which are all verbs). In table 1, we see for example, that the vowels in the word transliterated as `wa:lid` are a long /a:/, an /ɪ/ and a null vowel. Words with different roots share this pattern, which has been transliterated `fa:'il`.[1] Seeing the words that share a pattern, one can be tempted to try to encode the meaning of the form as a semantic feature. However, such features are difficult to encode and not always productive.

Derivation and Borrowings The process of derivation has proved to be flexible enough to derive from non-native words. Arab linguists stress the need to make borrowed terms concordant with the phonological and morphological structure of Arabic, to allow acceptable derivatives. For example, the English term *oxide* is pronounced `oksa:yid` in Arabic but it is modified to `uksi:d` in order to generate the derivatives shown in table2.

Table 2. Derivations from a borrowed word

Arabic Word	pattern	English Word
aksada	fa'lala	oxidize
muaksad	mufa'lal	oxidized
aksadah	fa'lalah	oxidation
taaksud	tafa'lul	oxidation

Morphology in the Bilingual Wordnet We conclude that in an Arabic-English bilingual wordnet, the derivational root and form of each content word should be stored, since this way of semantically linking words is a basic expectation of a literate Arabic speaker. However, it is not considered appropriate to attempt to 'decode' the patterns as semantic features or named relations.

[1] All patterns are written by convention with the same consonants /f/ /'/ and /l/ (and short vowels are written as diacritics). Textbooks often refer to the patterns by number or mnemonic rather than using these consonants as a skeleton.

3 Strategy for Building the Arabic-English Wordnet

One way to construct a bilingual wordnet would be to write lexicographers' files and compile a database with the grinder. However, the data for the English and Euro WordNets are available in alternative formats, including XML and Prolog. Persistently stored in a relational database, the data can be readily extended or modified in real time without a compilation step. New tables have been constructed to encode translations between synsets and Arabic words, roots and patterns.

We used Prolog clauses, edited to turn them into database tables via the comma-separated file format, as described in [2]. For efficient hyponymy navigation, we store with each synset, the path to it from the top of the tree and all its immediate hyponyms. On-demand selective tree display is acceptably fast.

3.1 Adding Data for Other Languages

There are several alternative ways to add a second and subsequent language to a sense enumerative lexicon [9], who discuss ways to link the senses in separate language-specific conceptual lexicons. It is equally possible to extend the data model to create a single multilingual repository. In our design, there is a single set of conceptual relations shared by the two (or more) languages. To make the database multilingual, the basic need is to provide the word↔sense table[2] for the additional language(s). Three possible extensions to the data model are:

1. Label the *word* column *English*, and add columns for each language.
2. Add a column encoding the language of the table row.
3. Reproduce a word↔sense table for each language.

Alternative (i) is not very attractive, as it implies a change to the database structure whenever an additional language is added to the database, although it is reasonably space-efficient if most words have equivalents in the various languages. Between alternatives (b) and (c), although the former is the more language-independent, we actually adopted the latter despite the language identity's embodiment in the table name. This was because of additional columns of attributes (described below) needed for Arabic, but not for other languages.

4 Words, Roots and Patterns in the WN_S_ARABIC Table

The Arabic equivalent of the WN_S table has the root and pattern of each word as additional columns. This allows the system to support queries based on words, roots or patterns, as well as via synonymy, hyponymy and the other Wordnet relations, and by English translation. Figure 1 shows the result of a query based on a shared root with the query word. In the database as presently constituted, words are written as cited in conventional dictionaries, without diactritics, although patterns are, of necessity, written with diacritics.

[2] This table has attributes synset_id, word, part of speech, and integers indicating the relative frequency of word within synset and of the sense of the word. A join of the table with itself finds either the synonyms of a word or its alternative senses.

English	Arabic
skill	علم
learning	تعلم
acquisition	تعلم
information	معلومات
instructions	تعليمات
information	معلومة
intelligence	معلوماتية
enquiry	استعلام
world	عالم
scientist	عالم
man_of_science	عالم
teacher	معلم
know	علم
educated	متعلم

Fig. 1. Query result with derivationally related Arabic words

With a morphological analyzer, it should be possible to dispense with the *word* column in the database, deriving it on demand from the root-pattern combination, and also to provide the diacritic form and/or transliterations for the benefi t of learners of Arabic.

5 Editing Functionality and the User Interface

Fig. 2. Simulated Arabic keyboard

Users and editors of a wordnet have different needs. A read-only interface can use formatted displays of synset lists, hyponymy trees etc. For an editor, there has also to be the

possibility of making a single word or sense from those retrieved or browsed *current*. Overall, the editor must support similar user operations to the EuroWordNet Polaris editor [9]. New items added to the database are then linked into sense relations like hyponymy, relative to the *current* synset. The information displays treat each element as a distinct object rather than as text. Figure 3 shows the current version of the interface and examples of the controls necessary to support updating. All updates are made relative to an item previously retrieved,

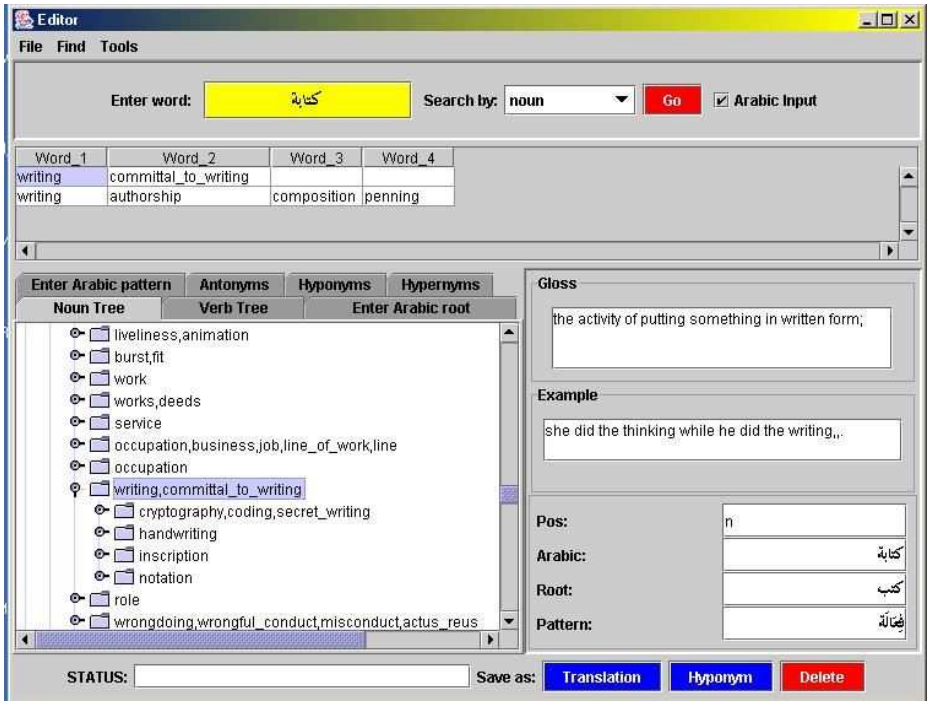


Fig. 3. Editor user's interface

so the interface has a query facility (the top panel in figure 3). This allows words to be entered in either English or Arabic (and additionally Arabic roots and patterns), and a number of alternative queries invoked (via the pull-down menu). Since words typically have multiple senses, the initial response to a query is to display a word↔sense matrix, as a table that allows cells, rows or columns to be selected (shown in the upper part of figure 3). Selecting a cell or a row makes a particular synset current. This in turn enables the tree-view to be generated and focused around the selected sense. At the same time, the gloss and examples (if any) for the selected sense are also retrieved and displayed. Any updates are made relative to the synset currently shown as selected.

Updates are confined to the entry of Arabic words equivalent to or related to the selected displayed synset. The editor enters the corresponding Arabic word, root and pattern in the fields in the panel towards the bottom right of Figure 3, pressing the button labeled

"Translation" to save the new word's details. This creates an entry in the WN_S_ARABIC table, with the same synset as the current one. Deletions from that table can be accomplished after retrieval of the item directly or via its English translation sysnet becoming current during browsing.

When a Direct Translation is not Possible There are numerous well-known conceptual difficulties in translating between languages. Both English and Arabic have many vocabulary items with no direct equivalent in the other language. Some of the fields in which these occur are religion, politics, food, clothing, etc. A small selection of Arabic words, all to do with Ramadan, and with no direct English equivalent is given in table 3.

Table 3. Words derived from a common root

Word	Transliteration	Meaning
سحور	suhu:r	light meal taken before starting a new day of Ramadan
مسحرَاتي	musahara:ti	man who beats a drum in the streets (before dawn) to wake people up to eat before they start a new day of fasting
إفطار	ifta:r	meal at the end of daily fasting during Ramadan
مدفع افطار	midfa' ifta:r	gun announcing the end of daily fasting during Ramadan
عمرة	umra	visit to the holy shrines in Mecca and Madina out of the time of the Pilgrimage

Where a word-root-pattern is entered having no English translation, a new Synset_id is allocated. Then this must be linked to its nearest hypernym (by adding a new row to the English table), and a new row to the Arabic version of the word↔sense table. An English gloss should also be added. What the user has to do in such a case is to find a suitable hypernym by search or browsing, prior to pressing the (save as) Hyponym button.

6 Conclusions and Further Work

We have described the design and partial implementation of a bilingual WordNet-based resource for English and Arabic, supported by a software framework built round a relational database. This enables us to store interesting conceptual relations additional to those in the original WordNet, and for the database to be extensible, particularly in the second language. To support the needs of end users, we will also need to incorporate a treatment of morphology. The original plan had been to adaopt the implementation by Ramsay and Mansur [5], although we are actively seeking alternatives that do not require multiple computer languages in the implementation. Other end-user-oriented features will be to widen the types of query supported, including free text queries of the glossary and example entries [6]. As computational linguists working on text mining applications, we are keen to experiment with the indirect use of the Arabic lexicon in revealing semantic relations useful to tasks such as WSD.

References

1. Black, W. J., El-Kateb, S.: Towards the design of an English-Arabic termonological and lexical knowledge base. In Proceedings of the Workshop on Arabic Natural Language Processing, ACL-2001, Toulouse, France (2001).
2. Denness, S. M.: A Design of a Structure for a Multilingual Conceptual Dictionary. MSc dissertation, UMIST, Manchester, UK (1996).
3. El-Kateb, Sabri: Translating Scientific and Technical Information from English into Arabic, MSc dissertation, University of Salford, UK (1991).
4. Holes, C. Modern Arabic. London: Longman (1995).
5. Ramsay, A. and H. Mansur.: Arabic Morphology: a categorial approach. Proceedings of the ACL2001 Workshop on Arabic Language Processing: Status and Prospects, Toulouse, July 6th, 2001.
6. Sierra, G. and McNaught, J.: Design of an onomasiological search system: A concept-oriented tool for terminology. Terminology **6**(1), 1–34. (2000).
7. Vossen, P.: Introduction to EuroWordNet. In: Nancy Ide, DanielGreenstein, Piek Vossen (eds), Special Issue on EuroWordNet. Computers and the Humanities, **32** (2–3) (1998), 73–89.
8. Vossen, P.: EuroWordNet: A Multilingual Database with Lexical SemanticNetworks, Dordrecht: Kluwer Academic Publishers (1998).
9. Vossen, P., Dez-Orzas, P., Peters, W.: The Multilingual Design of EuroWordNet. In: P. Vossen, N. Calzolari, G.Adriaens, A. Sanfilippo, Y. Wilks (eds.) Proceedings of the ACL/EACL-97 workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, Madrid, July 12th, 1997.

Automatic Assignment of Domain Labels to WordNet*

Mauro Castillo[1], Francis Real[1], and German Rigau[2]

[1] Universitat Politècnica de Catalunya
Jordi Girona Salgado 13, 08034 Barcelona, Spain
Email: castillo@lsi.upc.es, fjreal@lsi.upc.es
[2] Basque Country University
649 Posta Kutxa, 20080 Donostia, Spain
Email: rigau@si.ehu.es

Abstract. This paper describes a process to automatically assign domain labels to WordNet glosses. One of the main goals of this work is to show different ways to enrich sistematically and automatically dictionary definitions (or gloses of new WordNet versions) with MultiWordNet domains. Finally, we show how this technique can be used to verify the consistency of the current version of MultiWordNet Domains.

1 Introduction

Although the importance of WordNet (WN) has widely exceeded the purpose of its creation [12], and it has become an essential semantic resource for many applications [11,1], at the moment is not rich enough to directly support advanced semantic processing [6].

The development of wordnets large and rich enough to semantically process non-restricted text keeps on being a complicated work that may only be carried out by large research groups during long periods of time [4,2,3].

One of the main motivations of this work is to semantically enrich WN (or other lexic resources like dictionaries, etc.) with the semantic domain labels of *MultiWordNet Domains (*MWND*)* [8]. This resource has proved his utility in word domain disambiguation [7].

The work presented in this paper explores the automatic and sistematic assignment of domain labels to glosses and dictionary defi nitions.

This methodology may be also used to correct and verify the suggested labeling. It may also provide new cues to assign domain labels in dictionary defi nitions or in free texts.

This paper is organized as follows: section 2 introduces MWND, section 3 summarizes the experimental work carried out, section 4 is devoted to the the evaluation and results of the experiments and section 5 provides an in deep analisys of the experimental results. Finally, in section 6 some concluding remarks and future work are presented.

2 Semantic Resources

MWND [7] is a lexical resource developed in ITC-IRST where the *synsets* have been annotated semiautomatically with one or more domain labels. These domain labels are

* This paper had partially financed by the European Comition (MEANING IST-2001-34460), Generalitat de Catalunya (2002FI 00648) y UTEM-Chile.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 75–82.

organized hierarchically. These labels group meanings in terms of topics or scripts, e.g. Transport, Sports, Medicine, Gastronomy. which were partially derived from the Dewey Decimal Classification[3]. The version we used in these experiments is a hierarchy of 165 Domain Labels associated to WN1.6. Information brought by Domain Labels is complementary to what is already in WN. First of all Domain Labels may include synsets of different syntactic categories: for instance MEDICINE groups together senses from nouns, such as *doctor* and *hospital*, and from verbs such as *to operate*. Second, a Domain Label may also contain senses from different WN subhierarchies. For example, SPORT contains senses such as **athlete**, deriving from *person*, **game equipment**, from *artifact*, **sport** from *act*, and **playing field**, from *location*.

The table 1 shows the distribution of the number of domain labels per *synset*. This table also shows that most of the *synsets* have only one domain label.

Table 1. Distribution of domain labels for *synset* and distribution of *synset* with and without the domain label factotum in WN

domain labels for synset					
#	**noun**	**verb**	**adj**	**adv**	**%**
1	56458	11287	16681	3460	88.202
2	8104	743	1113	109	10.105
3	1251	88	113	6	1.4632
4	210	8	8	0	0.2268
5	2	1	0	0	0.0030

distribution synsets with CF and without SF factotum			
POS	**CF**	**SF**	**%fact**
noun	66025	58252	11.77
verb	12127	4425	63.51
adj	17915	6910	61.42
adv	3575	1039	70.93

On average, a noun synset has 1.170 domain labels assigned, a verbal synset 1.078, an adjectival synset 1.076 and and adverb synset 1.033.

When building MWND, any labels were assigned in high levels of the WN hierarchy and were automaticaly spread across the hypernym and troponym hierarchy. To our knowledge, a complete verification has not been made, neither automatic nor manual, of the whole set of assignments of domains to synsets.

The domain label *Factotum* includes two types of *synsets*:

Generic synsets: Used to mark the senses of WN that do not have a specific domain, for instance: person, dates, etc.

Stop Senses: The *synsets* that appear frequently in different contexts, for instance: numbers, colours, etc...

Table 1 shows the percentage of factotum labels for nouns, verbs, adjectives and adverbs in WN1.6. There is a high percentage of *synsets* labeled as factotum, except in nouns.

Recently, Domain information has been proven to be useful in many semantic applications. For instance, in Word Sense Disambiguation task (WSD), [5] emphasize the rol of domains. [9] introduce Word Domain Disambiguation (WDD) as a variant of WSD where for each word in a text the domain label is selected instead of the sense label (synset). In addition, MWND have been also used [10] in tasks such as “Text Categorization” (TC).

[3] `http://www.oclc.org/dewey`

3 Experiments

Even though MWND is a useful resource, it was semiautomatically constructed and it needs to be either manually or automatically validated. This validation would allow to study the domain label assignments to synsets of WN1.6 and acquire the implicit models of the domain assignment to glosses. With these models others resources as dictionaries or other WN versions without domains may be labeled. The main goals of the experiments described in this paper were:

- To study new automatic, consistent and robust procedures to assign domains labels to the WN1.6 glosses (or other versiones of WN), or to other definitions of generic dictionaries.
- To study new validation procedures of the consistency of the domain assignment in WN1.6, and especially, the automatic assigment of the factotum labels.

For the experiments, an small set of synsets (around 1%) was randomly selected as a test set and the other synsets were used as a training set (647 noun with 11.9% factotum and 121 verb with 60.33% factotum)

3.1 Labeling methodology

As a first attempt, we studied the performance of the automatic labeling metodology described in [13]. Rigau et al. used WN and a Spanish/English bilingual dictionary to automatically label a Spanish monololingual dictionary with WN Semantic Fields (or Lexicographic files).

We can use different similarity measures to obtain the importance (or saliency) of each word with respect each domain.

Using the salient words per domain gathered in the previous step, we can label each gloss again. When any of the salient words of a domain appears in a gloss, there is evidence that the gloss belongs to a particular domain. If several of these words appear, the evidence for that domain grows. Adding together their weights, over all words in a gloss, a program can determines the domain for which the sum is greatest. Thus, this automatic process depends on:

- The **similarity measure** used to assign domain weights to words 3.2. The words that form the synsets of the training data (variants, synonyms and gloss) are used to decide the frecuency of each word with respect to the domain labels that the synset has. Using different similarity measures, a weighted vector of Domains is generated for each word. For instance, table 2 shows a part of a weighted array for the nouns *soccer* (monosemous) and *orange* (polysemous).
- The **parameter filtering** applied in the experiment. Among others, the different weights for each part of information considered: *variants* (70%), words in the gloss (30%). The vectors obtained for each synset were normalized and only labels in the top 15% were considered (range [0.85..1]).

3.2 Measures

To estimate the weights of the words assigned to the domains 3 different formulas have been studied:

Table 2. Weighted array for nouns with factotum (CF)

word	weight	label	weight	label	word
soccer	2.826	soccer	8.181	botany	orange
soccer	2.183	play	5.129	gastronomy	orange
soccer	1.987	football	3.019	color	orange
soccer	1.917	sport	1.594	entomology	orange
soccer	0.998	rugby	1.205	jewellery	orange
...	...	...	...	...	...

M1: Square root formula	M2: Association Ratio	M3: Logarithm formula
$\frac{count(w,D)-\frac{1}{Ncount(w)count(D)}}{\sqrt{count(w,D)}}$	$Pr(w/D)log_2(\frac{Pr(w/D)}{Pr(w)})$	$log_2(\frac{Ncount(w,D)}{count(w)count(D)})$

4 Evaluation and Results

We studied the performance of the different labelling procedures by means of the following evaluation measures:

MiA measures the success of each formula (M1, M2 or M3) when the first proposed label is a correct one.

MiD measures the success of each formula (M1, M2 or M3) when the first proposed label is a correct one (or subsumed by a correct one in the domain hierarchy). For instance, if the proposed label is *Zoology* and the correct answer is *Biologogy* it is considered a correct answer.

Accuracy for the first proposed label	*Accuracy* for all the proposed labels
$AP = \frac{\text{success of the first label}}{\text{total of synsets}}$	$AT = \frac{\text{success of all the labels}}{\text{total of synsets}}$
Precision	*Recall*
$P = \frac{\text{(proposed and correct labels)}}{\text{(total proposed labels)}}$	$R = \frac{\text{(proposed and correct labels)}}{\text{total correct labels}}$

For nouns, different experiments were carried out. On average, the method assigns 1.23 domain labels per nominal synset and 1.20 domain labels per verbal synset.

The results when training with factotum and testing with factotum are shown in table 3; and presents the results when making the training and test without factotum. The best average results were obtained with the M1 measure. It must be emphasized that more than 70% of the first labels agree with MWND.

Table 4 presents the results obtained when training and testing for verbs with factotum, and shows the results obtained when training and testing verbs without factotum. In both cases the results are worst than the results obtained for the nouns. One of the reasons may be the high number of verbal *synsets* labeled with factotum domain(see table 1). However, in the case of verbs without factotum, the correct labeling at first proposal are fairly close to 70%.

Table 3. Results for nouns with (CF) and without factotum (SF)

CF					
N	**AP**	**AT**	**P**	**R**	**F1**
M1A	70.94	79.75	64.74	68.25	66.45
M1D	74.50	84.85	68.88	72.62	70.70
M2A	45.75	50.39	42.73	43.12	42.92
M2D	52.09	57.50	48.75	49.21	48.98
M3A	66.77	74.50	60.86	63.76	62.27
M3D	71.56	81.45	66.54	69.71	68.09

SF					
N	**AP**	**AT**	**P**	**R**	**F1**
M1A	73.95	81.82	66.81	68.68	67.73
M1D	78.50	87.24	71.24	73.24	72.23
M2A	52.45	57.52	49.32	48.24	48.77
M2D	59.44	65.21	55.94	54.71	55.32
M3A	74.48	82.69	68.41	69.41	68.91
M3D	78.85	88.64	73.33	74.41	73.87

Table 4. Results for verbs with (CF) and without factotum (SF)

CF					
V	**AP**	**AT**	**P**	**R**	**F1**
M1A	51.24	57.02	47.26	50.74	48.94
M1D	51.24	57.02	47.26	50.74	48.94
M2A	13.22	14.88	12.68	13.24	12.95
M2D	16.53	19.83	16.90	17.65	17.27
M3A	23.14	28.10	21.94	25.00	23.37
M3D	24.79	29.75	23.23	26.47	24.74

SF					
V	**AP**	**AT**	**P**	**R**	**F1**
M1A	69.77	76.74	64.71	55.93	60.00
M1D	74.72	83.72	69.23	61.02	64.86
M2A	20.93	25.58	19.64	18.64	19.13
M2D	41.86	51.16	38.60	37.29	37.93
M3A	41.86	55.81	39.34	40.68	40.00
M3D	53.49	67.44	46.77	49.15	47.93

From these tables, we can also observe that, M1 measure has better F1 than M2 and M3 and the behaviour of M1 and M3 is similar for nouns (CF and SF).

As espected, the method performs better for nouns than for verbs, because nouns have more and (maybe) more clear domain assigments.

For nouns, using the domain hierarchy, the performance increases, achieving 70.94% accuracy when assigning the first domain. However, using the domain hiearchy, it seems that for verbs only increases consistently when testing without factotum. In this case, for verbs the method obtains 51.24% accuracy when assigning the first domain.

Table 5. Training with factotum for nouns using the M1 measure

	Train CF			
	Test CF		Test SF	
	P	R	P	R
M1A	64.74	68.25	86.15	82.35
M1D	68.88	72.62	89.23	85.29

On table 5 there is a comparison for nouns using measure M1 and training with factotum and testing with (CF) and without factotum (SF).

For nouns, the best results are obtained training with factotum and testing without factotum, achieving a 86.15% of precision in the first assignment. One possible reason could be that labels, different than factotum, seems to be better assigned.

5 Discussion

Although the results are quite good, a more accurate analysis of the errors in the automatic assignments will show that the proposed labels are quite similars. It suggests a lack of systematicity in the semi-automatic assignment.

To illustrate possible errors, we show different examples where the proposed label has been considered a mistake in the evaluation.

1. **Monosemic words.** These words may help to find the correct domain.
   credit_application#n#1 (an application for a line of credit)
   Labeled with SCHOOL; proposal 1: Banking and proposal 2: Economy
   OBS: line_of_credit#n#1 is monosemous and is labeled as Banking.
   plague_spot#n#1 (a spot on the skin characteristic of the plague)
   Labeled with ARCHITECTURE; proposal 1: Physiology and proposal 2: Medicine
   OBS: plague#n#1 is monosemic and is labeled as Physiology-Medicine. In addition, skin#n has 6 senses as noun labeled with Anatomy, Transport and Factotum.
2. **Relations between labels.** Exists a direct relation in the domain hierarchy between the proposed labels and correct labels.
   academic_program#n#1 (a program of education in liberal arts and sciences (usually in preparation for higher education))
   Labeled with PEDAGOGY; proposal 1: School and proposal 2: University
   OBS: Pedagogy is the father of School and University.
   shopping#n#1 (searching for or buying goods or services)
   Labeled with ECONOMY; proposal 1: Commerce
   OBS: In the domain hierarchy, Commerce and Economy depend directly on Social_science.
   fire_control_radar#n#1 (radar that controls the delivery of fire on a military target)
   Labeled with MERCHANT_NAVY; proposal 1: Military
   OBS: Merchant_navy depends on Transport and Military and Transport depends on Social_science.
3. **Relations in** WN**.** Sometimes the *synsets* are related to words in the gloss.
   bowling#n#2 (a game in which balls are rolled at an object or group of objects with the aim of knocking them over play)
   Labeled with BOWLING; proposal 1: Play
   OBS: game#n#2 is hypernym and is labeled as Play. In addition, play#n#16 labeled as Play-Sport is related with holonym with game#n#2. In the domain hierarchy, Play and Sport are sibling and Bowling depends on Sport.
   cost_analysis#n#1 (breaking down the costs of som e operation and reporting on each factor separately)
   Labeled with FACTOTUM; proposal 1: Economy
   OBS: The word “cost” of the gloss have 3 senses labeled with Economy, Money and Quality.

4. **Uncertain cases.** There are cases where the proposed label is not represented by any pattern, but they may be considered as a correct label.
   birthmark#n#1 (a blemish on the skin formed before birth)
   Labeled with QUALITY; proposal 1: Medicine
   bardolatry#n#1 (idolization of William Shakespeare)
   Labeled with RELIGION; proposal 1: history and proposal 2: literature

Further analysis of these cases can help to obtain a validation method of the semi–automatic assigment of domains to synsets. A complete methodology should consider the addition, the removal or substitution of domains.

6 Conclusions and Further Work

The procedure to assign domain labels to WN gloss is very promising, especially because it is a difficult problem for the polysemy of WN and the semi-automatic process to generate the domain labels, using the WN hierarchy.

The proposal process is very reliable with the first proposal labels, reaching more that 70% on accuracy when testing without factotum.

We provided also an study of the typology of the errors. This suggest that in certain cases it is possible to add new correct labels or validate the old ones. In addition, other suggestion is that a lot of words labeled as factotum may be labeled with concrete domain label.

As future work we consider to make improvements, adaptations in the algorithm and test new methods to label other versions of WN.

References

1. E. Agirre and D. Martinez. Integrating selectional preferences in wordnet. In *Proceedings of the first International WordNet Conference in Mysore*, India, 21–25 January 2002.
2. J. Atserias, S. Climent, X. Farreres, G. Rigau, and H. Rodríguez. Combining multiple methods for the automatic construction of multilingual wordnets. In *Procceeding of RANLP'97*, pages 143–149, Bulgaria, 1997. Also to appear in a Book.
3. L. Bentivogli, E. Pianta, and C. Girardi. Multiwordnet: developing an aligned multilingual database. In *First International Conference on Global WordNet*, Mysore, India, 2002.
4. C. Fellbaum, editor. *WordNet. An Electronic Lexical Database*. The MIT Press, 1998.
5. J. Gonzalo, F. Verdejo, C. Peters, and N. Calzolari. Applying eurowordnet to cross-language text retrieval. *Computers and Humanities*, 1998.
6. S. Harabagiu, M. Pasca, and S. Maiorano. Experiments with open-domain textual question answering. In *Proceedings of COLING-2000*, Saarbruken Germany, 2000.
7. B. Magnini and G. Cavagliá. Integrating subject field codes into wordnet. In *In Proceedings of the Second Internatgional Conference on Language Resources and Evaluation LREC'2000*, Athens. Greece, 2000.
8. B. Magnini and C. Strapparava. User modelling for news web sites with content-based techniques. In *Proceedings WWW-2002, the Eleventh International World Wide Web Conference, Poster session*, Honululu, Hawaii, USA,, 2002.
9. B. Magnini, C. Strapparava, G. Pezzulo, and A. Gliozzo. Using domain information for word sense disambiguation. In *Proceedings of* 2^{nd} *International Workshop "Evaluating Word Sense Disambiguation Systems", SENSEVAL-2*, Toulouse, France, 2001.

10. B. Magnini, C. Strapparava, G. Pezzulo, and A. Gliozzo. Comparing Ontology-Based and Corpus-Based Domain Annotations in WordNet. *In Proceedings of First International WordNet Conference*, 2002.
11. D. McCarthy. *Lexical Acqusition at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences*. PhD thesis, University of Sussex, 2001.
12. G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. Five Papers on WordNet. *Special Issue of International Journal of Lexicography*, 3(4), 1990.
13. G. Rigau, H. Rodríguez M., and E. Agirre. Building accurate semantic taxonomies from monolingual mrds. In *In Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics COLING-ACL '98*, Montreal, Canada, 1998.

Creation of English and Hindi Verb Hierarchies and their Application to Hindi WordNet Building and English-Hindi MT*

Debasri Chakrabarti and Pushpak Bhattacharyya

Department of Computer Science and Engineering
Indian Institute of Technology, Mumbai, India
Email: debasri@cse.iitb.ac.in, pb@cse.iitb.ac.in

Abstract. Verbs form the pivots of sentences. However, they have not received as much attention as nouns did in the ontology and lexical semantics research. The classification of verbs and placing them in a structure according to their selectional preference and other semantic properties seem essential in most text information processing tasks like machine translation, information extraction *etc*. The present paper describes the construction of a verb hierarchy using Beth Levin's verb classes for English, the hypernymy hierarchy of the WordNet and the constructs and the knowledge base of the Universal Networking Language (UNL) which is a recently proposed interlingua. These ideas have been translated into the building of a verb hierarchy for Hindi. The application of this hierarchy to the construction of the Hindi WordNet is discussed. The overall motivation for this work is the task of machine translation between English and Hindi.

1 Introduction

The verb is the binding agent in a sentence. The nouns in a clause link to the main verb of the clause according to the verb's selectional preferences. However, verbs have not received as much attention as they deserve, when it comes to creating lexical networks and ontologies. Ancient Sanskrit treatises on ontology like the *Amarkosha* [1] deal meticulously with nouns, but not with verbs. The present day ontologies and lexical knowledge bases like *CYC [2], IEEE SUMO [3], WordNet [4,5], EuroWordNet [6], Hindi WordNet [7], Framenet [8] etc.* build deep and elaborate hierarchies for nouns, but the verb hierarchies are either not present or if present are shallow. The *Verbnet* project [9] is concerned exclusively with verbs and builds a very useful structure, but does not concern itself with building a *hierarchical structure.*

The classifi cation of verbs and placing them in a structure according to their selectional preference and other semantic properties seem essential in most text information processing tasks [9,10] like machine translation, information extraction *etc*. Additionally, *property inheritance* (*e.g. walk* inherits the properties of *move*) facilitates lexical knowledge building, for example, in a rule based natural language analysis system [11].

* Editor's note: This version of the paper may have not been typeset correctly due to the author's own formatting and non-availability of all fonts needed for retypesetting. The author's version is to be found on the accompanying CD.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 83–90.

The present paper describes the creation of a hierarchical verb knowledge base for an interlingua based machine translation system based on *Universal networking Language (UNL)* [12] and its integration to the Hindi WordNet. Use is made of (i) English verb classes and their alternation [10], (ii) the hypernymy hierarchy of WordNet [4,5] and the specifications and the knowledge base of the UNL system [12].

The organization of the paper is as follows. Section 2 deals with Levin's classification of English verbs. Section 3 is a brief introduction to the UNL system and the verb knowledge base therein. The creation of the verb hierarchy is explained in Section 4 with focus on the Hindi verbs. Section 5is on verbs and the Hindi WordNet. Section 6 concludes the paper and gives *future directions.*

2 Levin's Class of English Verbs

The key assumption underlying Levin's work is that the *syntactic behavior of a verb is semantically determined* [10]. Levin investigated and exploited this hypothesis for a large set of English verbs (about 3200). The syntactic behavior of different verbs was described through one or more alternations. *Alternation describes a change in the realization of the argument structure of a verb, e.g. middle alternation, passive alternation, transitive alternation etc*. Each verb is associated with the set of alternations it undergoes. A preliminary investigation showed that there is a considerable correlation between some facets of the semantics of verbs and their syntactic behavior so as to allow formation of classes. About 200 verb semantic classes are defined in Levin's system. In each class, there are verbs that share a number of alternations. Some example of these classes are the classes of the *verbs of putting*, which include *put verbs, funnel verbs, verbs of putting in a specified direction, pour verbs, coil verbs, etc*.

3 The Universal Networking Language (UNL)

The Universal Networking Language (UNL) [12] is an electronic language for computers to express and exchange information. UNL system consists of *Universal words (UW)* (explained below), *relations, attributes*, and the *UNL knowledge base (KB)*. The UWs constitute the vocabulary of the UNL, relations and attributes constitute the syntax and the UNL KB constitutes the semantics. The KB defines possible relationships between UWs.

UNL represents information sentence-by-sentence as a hyper-graph with concepts as nodes and relations as arcs. The representation of the sentence is a hyper-graph because a node in the structure can itself be a graph, in which case the node is called a *compound word (CW)*. Figure 1 represents the sentence *John eats rice with a spoon*.

In this figure, the arcs labeled with *agt* (agent), *obj* (object) and *ins* (instrument) are the relation labels. The nodes *eat(icl>do), John(iof >person), rice(icl>food)* and *spoon(icl>artifact)* are the *Universal Words* (*UW*). These are language words with *restrictions* in parentheses. *icl* stands for *inclusion* and *iof* stands for *instance of*. UWs can be annotated with attributes like *number, tense etc.* which provide further information about how the concept is being used in the specific sentence. Any of the three restriction labels- *icl, iof* and *equ-* can be attached to an UW for restricting its sense.

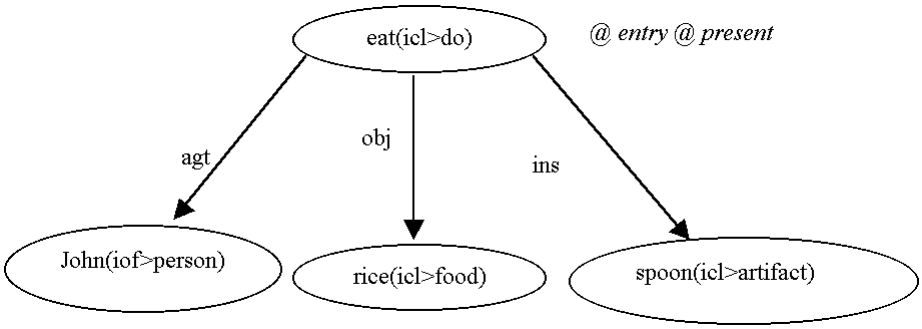


Fig. 1. UNL graph of *John eats rice with a spoon*

3.1 Verbal Concepts in UNL

The verbal concepts in the UNL system are organized in three categories. These are:

(icl>do) for defi ning the concept of an event which is caused by something or someone.
e.g., change(icl>do): as in *She changed the dress.*
(icl>occur) for defi ning the concept of an event that happens of its own accord.
e.g., change(icl>occur): as in *The weather will change.*
(icl>be) for defi ning the concept of a *state verb.*
e.g., remember(icl>be): as in *Do you remember me?*

The fi rst two categories correspond to the *action and* the *event verbs* respectively of the *nonstative class* and the third corresponds to *stative* [13]. A part of the hierarchy for the top concept *do* is shown in Figure 2.

```
do(agt>thing{,^gol>thing,icl>do,^obj>thing,^ptn>thing,^src>thing})
     → do(agt>volitional thing{,icl>do(agt>thing)})
         → do(agt>living thing{,icl>do(agt>volitional thing)})
            → do(agt>human{>living thing,icl>do(agt>living thing)})
     → do(agt>thing,gol>thing{,icl>do,^obj>thing,^ptn>thing,^src>thing})
```

Fig. 2. Partial hierarchical structure for *do*

The semantic hierarchy of the *do* tree is shown in Figure 3.

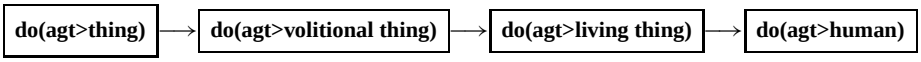


Fig. 3. Semantic hierarchy for *do*

```
"move" 'We should move ahead in this matter.' (to follow a procedure
       or take a course)
   (icl>act(agt>person))"
   [VINTRANS,VOA-ACT]
   → "move" 'How fast does your new car move?'
      (to change location)
      (icl>motion(>act(agt>thing))
      [VINTRANS,VOA-MOTN,VOA-ACT]
      → "move" 'Due to rain the cows were moving fast.'
      (to change the place or position of your body or a part of your body)
      (icl>motion{>act}(agt>volitional thing))
      [VINTRANS,VOA-MOTN,VOA-ACT]
   → "move" 'She cannot move her fingers.'
       (to cause to change the place or position of your body
       or a part of your body)
       (icl>motion(>act}agt>thing,obj>thing))
       [VTRANS,VOA-MOTN,VOA-ACT]

"move" She's made up her mind and nothing will move her.
(to change one's attitude or make sb change their attitude)
(icl>affect{>change}(agt>thing,obj>thing))
[VTRANS,VOO-CHNG]
```

Fig. 4. A part of *move* hierarchy

The specified relations for the *do* category are *agent*, *object*, *goal*, *partner* and *source*. It is stated that *agent* is the compulsory relation for this category. The *do* verb appearing in the hierarchy with only *agt* relation is the top node. In Figure 2, the symbol "^" specifies the *not* relation. It states that the top node of *do* does not take *gol (goal), obj (object), ptn (partner)* and *src (source)* relations. The second node in the figure shows that *do* appearing with *agt* and *gol* relation is the child of the top node. This hierarchy is set up using the argument structure of the verb. In the hierarchy the symbol '→' stands for the *parent-child* relationship.

4 Creation of the Verb Hierarchy

Levin's verb classes form the starting point. All the classes and the sub-classes are then categorized according to the UNL format (*vide* the previous section). Generally, to select the *top node*, the WordNet hypernymy hierarchy is used. However, when the WordNet hierarchy is not deep enough, dictionaries are used to arrive at the top node based on the perceived meaning hierarchy. Figure 5 shows a part of the hierarchy for the verb *put* (Similar partial tree for *move* appears in Figure 4). Everywhere, we first give the name of the verb, followed by an example sentence, the WordNet gloss, the UNL KB representation, the syntax frame and finally the grammatical and semantic categories (VTRANS, VOA-ACT etc.).

This example shows two types of sentence frames for the *put* class: one with the locative preposition (*in, around, into* etc.) and the other with the place adverb frame (*here/ there*). *hang* is the child node of *put*.

"put"
'Put your clothes in the cupboard'.
(to put something into a certain place)
(icl>move(agt>person,obj>concrete thing,gol>place)
(loc_prep{in/on/into/under/over)
[VTRANS, VOA-ACT]
→ **"hang"**
'He hanged the wallpaper on the wall'.
(to suspend or fasten something so that it is held up
from above and not supported from below)
(icl>put{>move}(agt>person,obj>concrete thing,gol>place)
(loc_prep{from/on)
[VTRANS, VOA-ACT]
"put"
'Put your things here'.
(to put something into a certain place)
icl>move(agt>person,obj>concrete thing,gol>place)
adv_plc{here/there)
[VTRANS, VOA-ACT]

Fig. 5. Hierarchy of the *put class*

4.1 Verb Hierarchy in Hindi

We elucidate the ideas with the example hierarchy for the Hindi verb रखना **; rəkʰna** (rakhanaa, meaning *put*) shown in Figure 6. In this figure, the name of the verb in Hindi is first mentioned, followed by the IPA transcription and the English transliteration. Then the corresponding English verb is given followed by the gloss from the English WordNet. After this comes the UNL representation with the example Hindi sentence (in IPA and English transliteration) and the sentence frame.

It is evident that there is a difference in the syntax frame with respect to English. For example, for the adverbial-place frame in English, the Hindi frame contains a locative postposition. This is due to the fact that case markers are obligatory features in the syntax of Hindi which is an inflectional language.

There are two different syntax frames specified for the *put class* in English [10], *viz., adv_plc* and *loc_prep*. Hindi has an extra frame for the same class. Thus, the syntax frames for the रखना (*put*) *class* are:

a. adv_man;
b. adv_plc + adv_man;
c. loc_postp + adv_man.

This leads to the discussion on the difference in the representations for *troponyms* in the two languages. *In English, the troponyms of a verb are usually different lexical terms. In Hindi, generally the verb itself with different syntax frames represents the troponyms.* It can thus be inferred that troponyms are lexically specified in English and syntactically in Hindi. The example of *arrange* in figure 7 makes this point clear.

A summary of the syntax frames for the verb *arrange* in the two languages is shown in Table 1.

रखना; **rəkʰna; rakhanaa**
'put' 'Put your things here.' (to put something into a certain place)
(icl>act(agt>person,obj>concrete thing,gol>place)
अपना समान यहाँ पर रखो; (əpna səman yəhã pər rəkʰo); apanaa samaan yahaa par rakho
{(adv_plc (यहाँ / वहाँ, 'yəhã / vəhã') + loc_postp (पर, 'pər')}

→ रखना, सजाना; **rəkʰna, səjana; rakhanaa, sajaanaa;**
'arrange' 'He arranged the books here.' (to put something in a particular order; to put into a proper or systematic manner)
(icl>put{>act}(agt>person,obj>thing)
उसने किताबों को यहाँ पर सजाकर रखा। usne kitabõ ko yəhã pər səjakər rəkʰa. usne kitabo ko yahaa par sajaakar rakhaa.
{(adv_man (सजाकर, 'səjakər'; क्रम से, 'krəm se') + (adv_plc (यहाँ / वहाँ 'yəhã / vəhã') + loc_postp(पर, 'pər')}

→ ढेर लगाना, इकट्ठा करना; **ḍʰer ləgana, ikəʈʰa kərna; Dhera lagaanaa, ikaTThaa karanaa**
'heap' 'He heaped woods here.' (to arrange in stacks)
(icl>arrange{>put}(agt>person,obj>functional thing,gol>functional thing)
उसने यहाँ पर लकड़ियां इकट्ठा की। usne yəhã pər ləkḍiyã ikəʈʰa kı. usne yahaa par lakdiya ikatthaa kii.
{(adv_plc (यहाँ / वहाँ, 'yəhã / vəhã + loc_postp (पर, 'pər')}

Figure 6 Hierarchy for *रखना 'put'*

रखना, सजाना; **rəkʰna, səjana; rakhanaa, sajaanaa;** 'arrange'
a. Sentence: उसने किताबों को सजाकर रखा। usne kitabõ ko səjakər rəkʰa. usne kitabo ko sajaakar rakhaa.
उसने किताबों को क्रम से सजाया। usne kitabõ ko krəm se səjayi. usne kitabo ko kram se sajaayaa.
'He arranged the books.'
Frame: adv_man (सजाकर, 'səjakər'; क्रम से , 'krəm se')

b. Sentence: उसने यहाँ पर किताबें क्रम से सजायीं / सजाकर रखीं।
usne yəhã pər kitabẽ krəm se səjayĩ/səjakər rəkʰĩ. usne yahaa par kitabe kram se sajaayii/sajaakar rakhii.
'He arranged the books here.'
Frame: adv_plc (यहाँ / वहाँ, 'yəhã / vəhã') + loc_postp (पर, 'pər') + adv_man (सजाकर, 'səjakər'; क्रम से, 'krəm se')

c. Sentence: उसने मेज के ऊपर किताबें क्रम से सजायीं / सजाकर रखीं।
usne mej ke upər kitabẽ krəm se səjayĩ/səjakər rəkʰĩ. usne mej ke uupar kitabe kram se sajaayii/sajaakar rakhi.
'He arranged the books on the table.'
Frame: loc_postp (के ऊपर, 'ke upər'; के नीचे, 'ke nice') + adv_man (सजाकर, 'səjakər'; क्रम से, 'krəm se')

Figure 7 Sentence frames for *arrange*

Table 1 Sentence frames for *arrange*

English	Hindi
1. adv_plc (here / there)	1. adv_man (सजाकर,'səjakər'; क्रम से,'krəm se' *etc.*)
2. loc_prep (in, inside, on *etc.*)	2. adv_plc (यहाँ / वहाँ, 'yəhã / vəhã') + loc_postp (पर, 'pər') + adv_man (सजाकर, 'səjakər'; क्रम से, 'krəm se' *etc.*)
	3. loc_postp (के ऊपर, 'ke upər'; के नीचे, 'ke nice' *etc.*) + adv_man (सजाकर, 'səjakər'; क्रम से, 'krəm se' *etc.*)

5 Verbs and the Hindi WordNet

The differences between Hindi and English verbs give rise to *language divergences* in machine translating one language to the other [14]. In English almost all the nouns can occur as verbs. But in Hindi verbalization of nominals is effected by combining two lexical items — noun/adjective/adverb and a *simple* verb. For instance,

noun and verb	आरंभ करना	'arəmb^{h} kərna / aarambha karanaa'	'to start';
adjective and verb	शांत करना	'ʃant kərna / shaanta karanaa'	'to calm down';
adverb and verb	उठाकर रखना	'd^{h}ıre kərna / uThaakara rakhanaa'	'to lift'.

According to traditional [15] and structural grammars [16], these verbs are classified as *conjunct verbs* with three sub-classes as shown above. From the syntax frames it is clear that the noun-verb combination is a true conjunct, as it gives a unique sense which is not decipherable from any other sources like sentence frames or semantic relations. On the other hand, the other two sub-classes can be deduced from sentence frames or through semantic relations. It is to be noted that the *compound verbs* in Hindi, *i.e*, a combination of a polar and a vector verb are dealt with in the manner of morphological processing. An instance of such verb is *गिर पड़ना, gir pəḍna, gira paRanaa* 'to fall down'.

In the Hindi WordNet, the *conjuct verbs* are stored through *conjuct-with* links between the first component (N/Adv/Adv) and the second (a simple verb). The verb hierarchy helps in optimizing the number of such links. The module for processing the compound verbs is a front end to the Hindi WordNet, just like the morphology module, and is table driven.

6 Results, Conclusions and Future Work

The work reported here started with English verbs. But these verbal concepts can be considered universal expressed using English alphabets. A hierarchy of English verbs has been created for the purpose of English Hindi machine translation. This hierarchy contains 5500 nodes (i.e. verbal concepts) corresponding to about 2000 unique English verbs. The principles behind organizing this hierarchy have been translated to Hindi, and a Hindi verb hierarchy too has been created. The top nodes in this hierarchy correspond to *act, move* and *put* classes in English. The verb hierarchy lends a structure to the organization of the verbs knowledge base in the Hindi WordNet. The coverage of both English and Hindi verbs is increasing everyday. A visualizer and an application programming interface for the verb knowledge bases in both the languages are under construction.

References

1. Jha Vishwanath, *Amarkosha by Amarsingha*, Motilal Banarasidas Publications, Varanasi, 1975.
2. Lenat D. B. and Guha R. V., *Building Large Knowledge Based System, Representation and Inference in the CYC Project*. Reading, Mass: Addison Wesley, 1990. `http://www.cyc.com`
3. `http://ontology.teknowledge.com/`
4. Miller, G., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. *Five Papers on WordNet.* CSL Report 43, Cognitive Science Laboratory, Princeton University, Princeton, 1990. `http://www.cogsci.princeton.edu/~wn`
5. Fellbaum, C. (ed.), *WordNet: An Electronic Lexical Database.* The MIT Press, 1998.
6. Vossen Piek (ed.), *EuroWordNet: A Multilingual Database with Lexical Semantic Networks.* Dodrecht. *Kluwer Academic Publishers*, 1998.
7. Chakrabarti Debasri, Narayan Dipak Kumar, Pandey Prabhakar, Bhattacharyya Pushpak, *Experiences in Building the Indo WordNet: A WordNet for Hindi*. Proceedings of the First Global WordNet Conference, 2002. `http://www.cfilt.iitb.ac.in/webhwn`
8. `http://framenet.icsi.berkeley.edu/~framenet`
9. `http://www.cis.upenn.edu/verbnet/`
10. Levin Beth, *English Verb Classes and Alternations A Preliminary Investigation.* The University of Chicago Press, 1993.
11. Dave Shachi and Bhattacharyya Pushpak, *Knowledge Extraction from Hindi Texts.* Journal of Institution of Electronic and Telecommunication Engineers, vol. 18, no. 4, July, 2001.
12. *The Universal Networking Language (UNL) Specifi cations*, Version 3.0, UNL center, UNDL Foundation, 2001. `http://www.unl.ias.edu/unlsys/unl/UNL%205specifications.html`.
13. Dowty, D., *Word Meaning and Montague Grammar*, Synthesis Language Library, Boston, 1979.
14. Dave Shachi, Parikh Jignashu and Bhattacharyya Pushpak, 2002, *Interlingua Based English Hindi Machine Translation and Language Divergence*, Journal of Machine Translation, Volume 17, September, 2002.
15. Bahari Hardev, *Vyavaharik Hindi Vyakaran Tatha Rachna.* Lokbharti Prakashan, Allahabad, India, 1997.
16. Singh Suraj Bhan, *Hindi ka Vakyatmak Vyakaran.* Sahitya Sahakar, Delhi, India, 1985.

Procedures and Problems in Korean-Chinese-Japanese Wordnet with Shared Semantic Hierarchy

Key-Sun Choi and Hee-Sook Bae

KORTERM, KAIST
373-1 Guseong-dong, Yuseong-gu, Daejeon, Republic of Korea
Email: {kschoi,elle}@world.kaist.ac.kr

Abstract. This paper introduces a Korean-Chinese-Japanese wordnet for nouns, verbs and adjectives. This wordnet is constructed based on a hierarchy of shared semantic categories originated from NTT Goidaikei (Hierarchical Lexical System). The Korean wordnet has been constructed by mapping a semantic category to each Korean word sense in a way that maps the same semantic hierarchy to the meanings of nouns, verbs, and adjectives. The meaning of each verb searched in the corpus is compared with its Japanese equivalent. The Chinese wordnet has been also constructed based on the same semantic hierarchy in comparison with the Korean wordnet. In terms of the argument structure, there is a semantic correspondence between Korean, Japanese and Chinese verbs.

1 Introduction

A Korean-Chinese-Japanese wordnet named CoreNet has been developed using a shared semantic hierarchy since 1994. This semantic hierarchy is originated in NTT Goidaikei[1], which consists of 2,710 hierarchical semantic categories. For the purpose of this paper, the term "wordnet" refers to a network of words, the term "concept" to the semantic category, and the term "sense" to the different meaning of word. In CoreNet, a total of 2,954 concepts are specifi ed. An increase in the number of concepts specifi ed in CoreNet is attributable to the necessity for reflecting the concepts found only in the Korean language. On the one hand, the same semantic hierarchy applied to both nouns and predicates in CoreNet, while different concept systems are applied to nouns and predicates in NTT Goidaikei.

Mapping the same semantic hierarchy to both nouns and predicates results in some advantages: fi rst, there are pattern similarities between nouns and predicates, especially in Chinese-derived words (that is N in the following example). For example, "N-hada and "N+suru" are the Korean and Japanese version of a basic pattern "do + N" in English; second, the language generation based on a conceptual structure takes freer phrase patterns regardless of either the noun or verb. This computational work has been accompanied by heuristics and trial-and-errors as well as semi-automatic approaches. Several linguistic resources have been used for building CoreNet. Among them, [2] and [3] have been primarily used as a basis for the meanings of Korean words. Most of the Chinese vocabulary is based on [5].

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 91–96.

2 Principles

CoreNet has been constructed according to the following principles: multiple mapping between the word sense and the concept, corpus-based, multilingualism, and application of a single concept system.

2.1 Mapping between Word Sense and Concept

The purpose of CoreNet is mainly to resolve semantic ambiguities using the following two functionalities. Firstly, every possible meaning of a word in the dictionary [3] is mapped to one or more concepts. For example, each meaning of the word "*school*" is mapped into three concepts; PLACE, ORGANIZATION, and BUILDING. In the second place, a syntactic-semantic structure is mapped to the predicate-argument structure. For example, a Korean verb "*gada*" has a set of 17 senses in the dictionary [3]; these word senses are mapped into the concepts such as GOING, LEARNING, SERVICE, DELIVERY, PROGRESS, CONTINUATION, ENTHUSIASM, SWEEP, and so on. This set of predicate concepts is identical to nouns'. On the other hand, each predicate has its unique argument structure. For example, "*gada*" is mapped into seven concepts (e.g., GOING, LEARNING) whose argument structures are different. Each argument is represented by the set of possible concept filler (e.g., [HUMAN]) and syntactic role (e.g., subject, dative, and object) while its Japanese equivalents (e.g., "*iku*") are addressed by the followings:

1. GOING([HUMAN,MAMMAL,VEHICLE]=subject), "*iku*"
2. LEARNING([HUMAN]=subject,[TEACHER]=dative), "*iku*"
3. DELIVERY([INFORMATION]=subject,[HUMAN]=dative), "*tutawaru*"
4. PROGRESS([TIME]=subject), "*sugiru*"
5. CONTINUATION([RELATION]=subject,[YEAR]=object), "*tuduku*"
6. ENTHUSIASM([GAZE]=subject,[GIRL]=dative), "*iku*"
7. SWEEP([EMOTION]=subj), "*kieru*"

2.2 Corpus-based usage

A set of vocabularies and their meanings are extracted from KAIST corpus [2]. The following shows what the argument structure of "*gada*" described in the section 2.1 is like when extracted from the corpus: GOING ([*horse*/MAMMAL, *bus*/VEHICLE]=subject)

Horse and *bus* are the terms extracted from the corpus while MAMMAL and VEHICLE are the concept names respectively mapped to the words *horse* and *bus*. This results in more specified categorization for the meaning of words than in dictionaries.

2.3 Multilingualism

All concepts are aligned with three languages: Japanese, Korean and Chinese. Among these three languages, all words that are nouns or predicates are categorized into a single concept hierarchy. Based on the meanings of words as well as concepts, verbs among three languages are also linked each other. The following is part of a list of concepts for the Chinese verb [qù]. Note that the *italicized* words are Korean equivalents. A sample list is shown in Figure 1.

1. GOING - *gada*
2. DELIVERY – *bonaeda*
3. EXCLUSION - *eobsaeda*

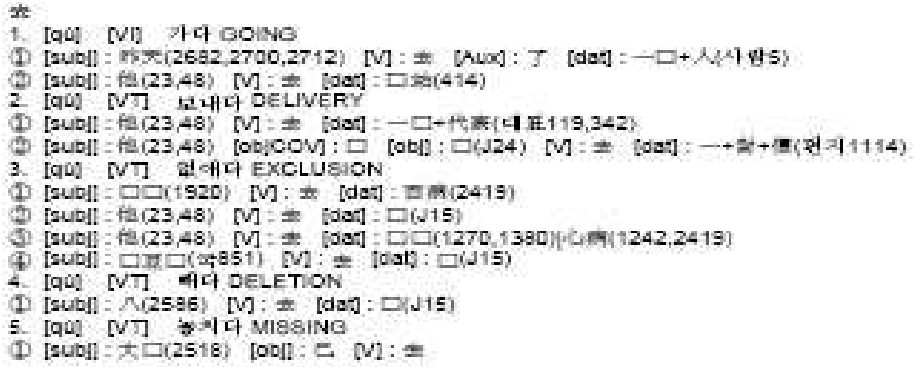


Fig. 1. An Entry in Chinese-Korean Verb CoreNet

2.4 Single Concept System

In general, concept systems and word nets are constructed for nouns. In CoreNet, however, a single concept system is shared by nouns, verbs, and adjectives. To this respect updates are continuously made for sharing of single concept system among three languages.

3 Procedures

3.1 Selection of Word Entry

A set of basic words is selected from the frequency-based vocabulary list of corpora compared with an existing set of basic Korean words. About 50,000 general vocabularies are selected for CoreNet word entries.

3.2 Bootstrapping for Initial Semantic Category Assignment

Using a Japanese-Korean electronic dictionary, we translated all Japanese words in the NTT Goidaikei into their Korean equivalents based on word meanings. Manual correction by experts of the results of automatic translation is followed for erroneous assignments between the two languages. This process also poses many problems. The most diffi cult problem issues from the difference in concept division systems. In Japanese, for example, concepts like GOING or SORTING have more subordinates than in Korean language, and vice versa for ROOT. In addition, FURNITURE has subordinate concepts like DESK, CHAIR, and FIREPLACE,

while in Korean, FIREPLACE is dealt with as part of KITCHEN. These problems arise from the difference in the way of thinking and culture. Then, we assign a semantic category by matching Korean words with their equivalent list for the semantic category in the NTT Goidaikei. No equivalent can be found in the translated word list and some errors can be found in a translation version. In the former case, a genus term for the word is extracted from descriptive statements of a machine-readable dictionary. In the latter case, manual correction is performed by experts.

3.3 Semantic Category Assignment Based on Word Sense Definitions [4]

Assuming that meanings falling under a concept are defined by similar words in the dictionary, we collected the definitions of the word senses that were mapped into one concept incorporating them into the concept's definition. This resulted in the creation of a chunk of definitions per concept. That is, the definition of a concept is indirectly represented by the chunk of definition of word senses that has already been assigned to the concept. For a given new word sense, its appropriate concept assignment is to be solved by how much the definition of the word sense is similar with the definition of concept. Assignment of proper concepts to the word sense can be viewed as retrieving a relevant definition chunk (of concept) for the given word sense. Each concept's definition is incrementally upgraded whenever the definition for a new word sense is assigned to the concept.

Our structured version of the Korean dictionary [3] includes such lexical relation information as synonyms, abbreviations, antonyms, *etc*. It is reasonable that the two senses linked by this lexical relation information (except for antonyms) fall under the same concept.

3.4 Manual Correction

The process of resolving the meaning of a word (i.e. word sense disambiguation) was manually performed in order to assign proper semantic categories to every possible meaning of a word, as well as translation errors were removed. The same manual correction was independently performed by two researchers. After comparative review over the results, only identically mapped sets were selected as final semantic categories with the purpose of ensuring highest accuracy. In the final stage, a third party examined different parts of the results to choose the proper ones. Despite this manual correction, it remains still some embarrassing cases. For example, 출입(出入) is a word having a concept combined with two concepts GO OUT and ENTER. In this case, we selected the concept of superior node when the latter contains all of concept elements as following: 出入 [GO OUT-ENTER, 2183].

4 Considerations

This section describes what we had to consider and decide about the underspecified sense, multiple concept mapping, verbal noun, and concept splitting.

4.1 Underspecified Sense and Multiple Concept Mapping

A word is mapped into several concepts that comprise respective meanings of the word. For example, *school* is an "institution for the instruction of students". The word *school* is mapped

into three concepts such as LOCATION, ORGANIZATION, and FACILITY. Unless the meanings of a word are fully specified in the mother dictionary [3], however, one meaning of the word must be mapped into several concepts. The word *school* is a good example of underspecified meanings.

4.2 Verbal Noun

A verb is assigned to concepts after it is transformed to a noun. For example, "*write*" is transformed to its noun form "*writing*" that is mapped into a concept WRITING falling under EVENT. An adjective "*be wise*" is transformed to "*wisdom*" that is mapped into PROPERTY under CAPABILITY, which falls under ATTRIBUTE. Consider an adjective "*be wide*". A sense is mapped respectively to POSITIVE PERSONALITY, EXTENT/LIMITS, and WIDTH (under the concept UNIT OF QUANTITY).

4.3 Concept Splitting

Every time inconsistency among nodes of concepts is discovered, a node may be added. For example, BODY has three sub-concepts in the NTT concept system: ARM, LEG, and HEAD. But, a word "*back*" cannot be assigned to any sub-concepts. At least, OTHERBODY should be added to the fourth sub-concepts under BODY. In the course of constructing verbs and adjectives wordnets, the concept splitting was performed by reclassifying the word senses. For example, ARRIVAL is subdivided into SITUATION ARRIVAL, TIME ARRIVAL, EXTENT ARRIVAL, and POSITION ARRIVAL.

5 Example

Figure 2 shows a screenshot of the Korean-Japanese noun wordnet. The screen has four windows. The upper left side of the window shows a correspondence between Japanese and Korean words and concept numbers. The lower left side of the window contains word senses and definitions in the dictionary [3]. The upper right side of the window shows all words under a concept QUANTITY numbered 2588. The lower right side of the window shows a part of the list of concept hierarchy.

6 Conclusion

CoreNet has been constructed in combination with its necessary corpora and lexical database. To begin with, the keynote system of the NTT Goidaikei [1] was used, which was followed by the development of a Korean version of noun systems. Despite the different semantic categories applied to predicates in the NTT Goidaikei, we have aggressively applied the same semantic categories to predicate systems in CoreNet. Further, what differs between CoreNet and NTT Goidaikei is that CoreNet features mapping between word senses (not just words) and concepts. Multilinguality is another feature of CoreNet designed to deliver a single concept system for different languages.

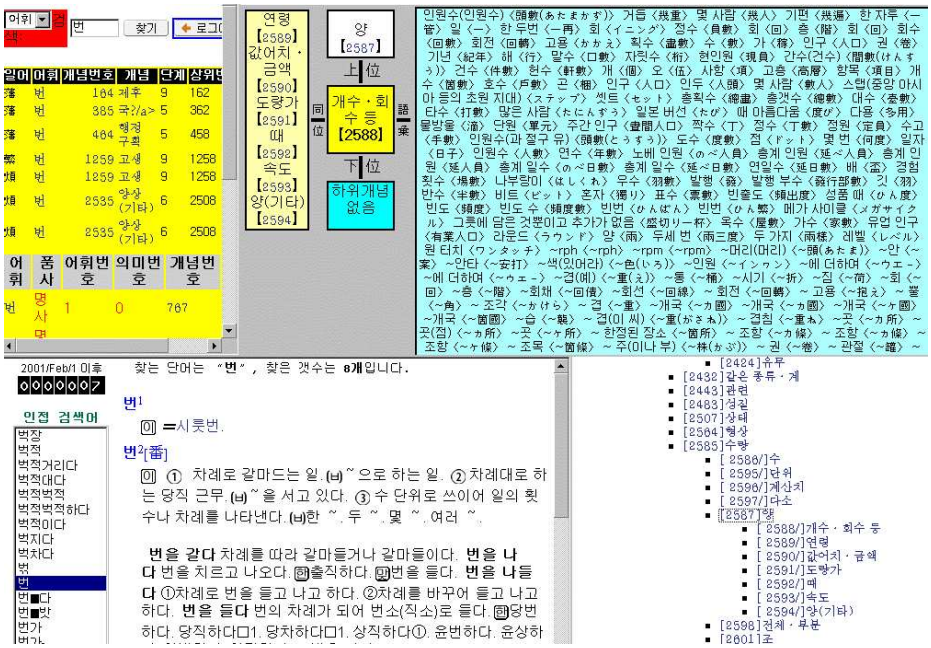


Fig. 2. A Screenshot of Korean-Japanese Noun Wordnet

References

1. Ikehara, S. *et al*: The Semantic System, volume 1 of Goidaikei – A Japanese Lexicon, Iwanami Shoten, Tokyo (1997).
2. KAIST Corpus, `http://morph.kaist.ac.kr/kcp/`, (in Korean), 1999–2003.
3. Hangeul Society, ed.: Urimal Korean Unabridged Dictionary, Eomungag (1997).
4. Lee, J.-H. *et al*.: Semi-Automatic Construction of Korean Noun Thesaurus by Utilizing Monolingual MRD and an Existing Thesaurus, Proceedings of the 16th PACLIC, Jeju (2002).
5. Yu, S.: Modern Chinese Grammar Information Dictionary, Peking University Press (1999).

Evaluating the Contribution of EuroWordNet and Word Sense Disambiguation to Cross-language Information Retrieval

Paul Clough and Mark Stevenson

University of Sheffield
Regent Court, 211 Portobello Street,
Sheffield, S1 4DP
United Kingdom
Email: p.d.clough@sheffield.ac.uk, marks@dcs.shef.ac.uk

Abstract. One of the aims of EuroWordNet (EWN) was to provide a resource for Cross-Language Information Retrieval (CLIR). In this paper we present experiments which test the usefulness of EWN for this purpose via a formal evaluation using the Spanish queries from the TREC6 CLIR test set. All CLIR systems using bilingual dictionaries must find a way of dealing with multiple translations and we employ a WSD algorithm for this purpose. It was found that this algorithm achieved only around 50% correct disambiguation when compared with manual judgment, however, retrieval performance using the senses it returned was 90% of that recorded using manually disambiguated queries.

1 Introduction

Cross-language information retrieval (CLIR) is the process of providing queries in one language and returning documents relevant to that query which are written in a different language. This is useful in cases when the user has enough knowledge of the language in which the documents are returned to understand them but does not possess the linguistic skill to formulate useful queries in that language. An example is e-commerce where a consumer may be interested in purchasing some computer equipment from another country but does not know how to describe what they want in the relevant language.

A popular approach to CLIR is to translate the query into the language of the documents being retrieved. Methods involving the use of machine translation, parallel corpora and machine readable bilingual dictionaries have all been tested, each with varying degrees of success [1,2]. One of the simplest and most effective methods for query translation is to perform dictionary lookup based on a bilingual dictionary. However, the mapping between words in different languages is not one-to-one, for example the English word "bank" is translated to French as "banque" when it is used in the 'financial institution' sense but as "rive" when it means 'edge of river'. Choosing the correct translation is important for retrieval since French documents about finance are far more likely to contain the word "banque" than "rive". A CLIR system which employs a bilingual dictionary must find a way of coping with this translation ambiguity.

The process of identifying the meanings of words in text is known as word sense disambiguation (WSD) and has been extensively studied in language processing. WSD is

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 97–105.

normally carried out by selecting the appropriate sense for a context from a lexical resource such as a dictionary or thesaurus but for CLIR it is more appropriate to consider the set of senses as the possible translations of a term between the source and target languages. For example, in an English-to-French CLIR system the word "bank" would have (at least) two possible senses (the translations "banque" and "rive"). By considering the problem of translation selection as a form of WSD allows us to make use of the extensive research which has been carried out in that area.

EuroWordNet (EWN) [3] is a lexical database which contains possible translations of words between several European languages and was designed for use in CLIR [4]. Section 2 describes the WSD algorithm we use to resolve ambiguity in the retrieval queries. In Section 3 we describe the experiments which were used to determine the improvement in performance which may be gained from using WSD for CLIR the results of which are presented in Section 4. Section 5 described an evaluation of the WSD algorithm used. The implications and conclusions which can be drawn from this work are presented in Sections 6 and 7.

2 Word Sense Disambiguation

One of the main challenges in using a resource such as EWN is discovering which of the synsets are appropriate for a particular use of a word. In order to do this we adapted a WSD algorithm for WordNet originally developed by Resnik [5]. The algorithm is designed to take a set of nouns as context and determine the meaning of each which is most appropriate given the rest of the nouns in the set. This algorithm was thought to be suitable for disambiguating the nouns in retrieval queries.

The algorithm is fully described in [5] and we shall provide only a brief description here. The algorithm makes use of the fact that WordNet synsets are organised into a hierarchy with more general concepts at the top and more specific ones below them. So, for example, `motor vehicle` is less informative than `taxi`. A numerical value is computed for each synset in the hierarchy by counting the frequency of occurrence of its members in a large corpus[1]. This value is dubbed the *Information Content* and is calculated as $\textit{Information Content}(\textit{synset}) = -\log \Pr(\textit{synset})$.

The similarity of two synsets can be found by choosing the synset which is above both in the hierarchy with the highest information content value (i.e. the most specific). By extension of this idea, sets of nouns can be disambiguated by choosing the synsets which return the highest possible total information content value. For each sense a value is returned indicating the likelihood that the sense being the appropriate one given the group of nouns.

3 Experimental Setup

3.1 Test Collection

Evaluation was carried out using past results from the cross-lingual track of TREC6 [6]. We used only TREC6 runs that retrieved from an English language collection, which was the 242,918 documents of the Associated Press (AP), 1988 to 1990. NIST supplied 25 English

[1] We used the British National Corpus which contains 100 million words.

CLIR topics, although four of these (topics 3, 8, 15 and 25) were not supplied with any relevance judgements and were not used for this evaluation.

The topics were translated into four languages (Spanish, German, French and Dutch) by native speakers who attempted to produce suitable queries from the English version. For this evaluation the Spanish queries were used to evaluate the cross-lingual retrieval and the English queries to provide a monolingual baseline. Spanish was chosen since it provides the most complete and accurate translation resource from the EWN languages. In addition the EWN entries for Spanish tend to have more senses than several of the other languages and is therefore a language for which WSD is likely to be beneficial.

In order to evaluate the contribution of the WSD algorithm and EWN separately the English and Spanish queries were manually disambiguated by the authors. The possible synsets were identified for each query (for the Spanish queries these were mapped from the Spanish synsets onto the equivalent English ones which would be used for retrieval). A single sense from this set was then chosen for each term in the query.

3.2 CLIR System

Our CLIR system employs 3 stages: term identification, term translation and document retrieval. The term identification phase aims to find the nouns and proper names in the query. The XEROX part of speech tagger [7] is used to identify nouns in the queries. Those are then lemmatised and all potential synsets identified in EWN.[2] For English queries this set of possible synsets were passed onto the WSD algorithm to allow the appropriate one to be chosen. Once this has been identified the terms it contains are added to the final query. (In the next Section we describe experiments in which different synset elements are used as query terms.) For Spanish queries the EWN Inter-Lingual-Index [3] was used to identify the set of English WordNet synsets for each term which is equivalent to to the set of possible translations. For each word this set of synsets was considered to be the set of possible senses and passed to the WSD algorithm which chooses the most appropriate. Non-translatable terms were included in the final translated query because these often include proper names which tend to be good topic discriminators.

Document retrieval was carried out using our own implementation of a probabilistic search engine based on the BM25 similarity measure (see, e.g. [8]). The BM25 function estimates term frequency as Poisson in distribution, and takes into account inverse document frequency and document length. Based on this weighting function, queries are matched to documents using a similarity measure based upon term co-occurrence. Any document containing at least one or more terms from the query is retrieved from the index and a similarity score computed for that document:query pair. Documents containing any number of query terms are retrieved (creating an OR'ing effect) and ranked in descending order of similarity under the assumption that those nearer the top of the ranked list are more relevant to the query than those nearer the bottom.

[2] For these experiments the Spanish lemmatisation was manually verified and altered when appropriate. This manual intervention could be omitted given an accurate Spanish lemmatiser.

3.3 Evaluation Method

We experimented with various methods for selecting synsets from the query terms: all synsets, the first synset and the synset selected by the WSD algorithm. It is worth mentioning here that WordNet synsets are ordered by frequency of occurrence in text and consequently the first synset represents the most likely prior sense. We also varied the number of synset members selected: either the headword (first member of the synset), or all synset terms. In the case of all synset terms, we selected only distinct terms between different synsets for the same word (note this still allows the same word to be repeated within a topic). This was done to reduce the effects of term frequency on retrieval, thereby making it harder to determine how retrieval effectiveness is affected by WSD alone. Preliminary experiments showed retrieval to be higher using distinct words alone. We also experimented with longer queries composed of the TREC6 title and description fields, as well as shorter queries based on the title only to compare the effects of query length with WSD.

Retrieval effectiveness is measured using the `trec_eval` program as supplied by NIST. With this program and the set of relevance documents as supplied with the TREC6 topics, we are able to determine how many relevant documents are returned in the top 1000 rank positions, and the position at which they occur. We use two measures of retrieval effectiveness computed across all 25 topics. The first is *recall* which measures the number of relevant documents retrieved. The second measure, *mean uninterpolated average precision* (MAP), is calculated as the average precision figures obtained after each new relevant document is seen [9].

4 CLIR Evaluation

The results of cross-lingual retrieval can be placed in context by comparing them against those from the monolingual retrieval using the English version of the title and description as the query. (EuroWordNet was not used here and no query expansion was carried out.) It was found that 979 documents were recalled with a MAP score of 0.3512. These results form a reasonable goal for the cross-lingual retrieval to aim towards.

Table 1. Results for Spanish retrieval with title and description

synset selection	synset members	recall	MAP
gold	all	890	0.2823
	1st	676	0.2459
all	all	760	0.2203
	1st	698	0.2215
1st	all	707	0.2158
	1st	550	0.1994
WSD	all	765	0.2534
	1st	579	0.2073

Table 1 shows retrieval results after translating the title and description. The first column ("synset selection") lists the methods used to choose the EWN synset from the set of possibilities. "gold" is the manually chosen sense, "all" and "1st" are the two baselines of choosing all possible synsets and the first while "auto" is the senses chosen by the WSD algorithm. The next column ("synset members") lists the synset members which are chosen for query expansion, either all synset members or the first one.

The best retrieval scores for manually disambiguated queries is recorded when all synset members are used in the query expansion which yields a MAP score of 0.2823 (see Table 1 row "gold", "all"). This is around 80% of the monolingual retrieval score of 0.3512. When WSD is applied the highest MAP score of 0.2534 is achieved when all synset members are selected (Table 1 row "WSD", "all"). This represents 72% of the MAP score from monolingual retrieval and 90% of the best score derived from the manually disambiguated queries.

In the majority of cases choosing all synset members leads to a noticeably higher MAP score than retrieval using the first synset member. This is probably because the greater number of query terms gives the retrieval engine a greater chance of finding the relevant document. The exception is when all synsets have been selected (see Table 1). In this case the retrieval engine already has a large number of query terms thorough the combination of the first member from all synsets and adding more makes only a slight difference to retrieval performance.

When translating queries, it would appear that using Resnik's algorithm to disambiguate query terms improves retrieval performance when compared against choosing all possible senses or the first (most likely) senses to disambiguate.

Table 2. Results for Spanish retrieval with title only

synset selection	synset members	recall	MAP
gold	all	828	0.2712
	1st	685	0.2192
all	all	735	0.2346
	1st	640	0.1943
1st	all	658	0.2072
	1st	511	0.1689
WSD	all	758	0.2361
	1st	650	0.2007

The experiments were repeated, this time using just the title from the TREC query which represents a shorter query. The results from these experiments are shown in Table 2. The manually annotated queries produces the highest MAP of 0.2712 (77% of monolingual). When the WSD algorithm is used the highest MAP is also recorded when all synset members were chosen. This score was 0.2361 (67% of monolingual). However, when the shorter queries are used the difference between WSD the two naive approaches (choosing the most frequent sense and choosing all senses) is much smaller. This is probably because the reduced

amount of context makes it difficult for the WSD algorithm to make a decision and it often returns all senses.

Table 2 also shows that choosing all synset members is a more effective strategy than choosing just the first member. We already noted this with reference to the results form the longer queries (Table 1) although the difference is more pronounced than when the longer queries were used. In fact it can be seen that when the short queries are used choosing all members for each possible synset (i.e. no disambiguation whatsoever) scores higher than choosing just the first member of the manually selected best sense. This shows that these shorter queries benefit far more from greater query expansion and that even correct meanings which are not expanded much do not provide enough information for correct retrieval.

5 Evaluation of WSD

It is important to measure the effectiveness of the WSD more directly than examining CLIR results. Others, such as [10,11], have found that WSD only has a positive effect on monolingual retrieval when the disambiguation is accurate. The manually disambiguated queries were used as a gold-standard aginst which the WSD algorithm we used could be evaluated. Two measures of agreement were computed: strict and relaxed. Assume that a word, w, has n senses denoted as $senses(w)(= w_1, w_2, ...w_n)$ and that one of these senses, w_{corr} (where $1 \leq corr \leq n$), was identified as correct by the human annotators. The WSD algorithm chooses a set of m senses, $wsd(w)$, where $1 \leq m \leq n$. The strict evaluation score for w takes into account the number of senses assigned by the WSD algorithm and if $w_{corr} \in wsd(w)$ the word is scored as $\frac{1}{m}$ (and 0 if $w_{corr} \notin wsd(w)$). The relaxed score is a simple measure of whether the WSD identified the correct senses regardless of the total it assigned and is scored as 1 if $w_{corr} \in wsd(w)$. The WSD accuracy for an entire query is calculated as the mean score for each term it contains.

The two evaluation metrics have quite different interpretations. The strict evaluation measures the degree to which the senses identified by the WSD algorithm match those identified by the human annotators. The relaxed score can be interpreted as the ratio of query words in which the sense identified as correct was not ruled out by the WSD algorithm. In fact simply returning all possible senses for a word would guarantee a score of 1 for the relaxed evaluation, although the score for the strict evaluation would probably be very low. Since it is important not to discard the correct sense for retrieval purposes the relaxed evaluation may be more relevant for this task.

Table 3. Results of WSD algorithm and first sense baseline compared against manually annotated queries

		Score	
Language	**Method**	Strict	Relaxed
English	WSD	0.410	0.546
	1st synset	0.474	
Spanish	WSD	0.441	0.550
	1st synset	0.482	

Table 3 shows the results of the evaluation of the WSD algorithm and baseline method of choosing the first sense against the manually annotated text for both the Spanish and English queries. The baseline scores are identical for each metric since it assigns exactly one sense for each word (the first) and the two metrics only return different scores when the technique assigns more than one sense.

We can see that the evaluation is similar across both languages. The baseline method actually outperforms automatic WSD according to the strict evaluation measure but scores less than it when the relaxed measure is used. We can also see that neither of the approaches are particularly accurate and often rule out the sense that was marked as correct by the human annotator.

However the results from the cross-language retrieval experiments earlier in this Section show that there is generally an improvement in retrieval performance when the WSD algorithm is used. This implies that the relaxed evaluation may be a more appropriate way to judge the usefulness of a WSD algorithm for this task. This idea has some intuitive plausibility it seems likely that for retrieval performance it is less important to identify the sense which was marked correct by an annotator than to try not to remove the senses which are useful for retrieval. It should also be borne in mind that the human annotation task was a forced choice in which the annotator had to choose exactly one sense for each ambiguous query term. In some cases it was very difficult to choose between some of the senses and there were cases where none of the EWN synsets seemed completely appropriate. On the other hand our WSD algorithm tended to choose several senses when there was insufficient contextual evidence to decide on the correct sense.

6 Discussion

The WSD algorithm's approach of only choosing senses when there is sufficient evidence suits this task well. However, the WSD results also highlight a serious limitation of EWN for CLIR. EWN's semantics are based on ontological semantics using the hyponymy relationship. That is, the EWN synset hierarchy contains information about the type of thing something is. So, for example, it tells us that "car" is a type of "motor vehicle". However, many types of useful semantic information are missing. One example is discourse and topic information. For example, "tennis player" (a hyponym of person) is not closely related to "racket", "balls" or "net" (hyponyms of artifact). Motivated by this example, Fellbaum [12] dubbed this the "tennis problem". This information is potentially valuable for retrieval where one aim is to identify terms which model the topic of the query.

Others, including [1,13,14], have used word co-occurrence statistics to identify the most likely translations and this could be considered a form of translation. This approach seems promising for CLIR since it returns words which occur together in text and these are likely to be topically related. This approach has potential to be developed into a WSD algorithm which could be applied to EWN.

There has been some disagreement over the usefulness of WSD for monolingual retrieval (see, for example, [11,15]). In particular [10,11] showed that WSD had to be accurate to be useful for monolingual retrieval. However, the results presented here imply that this is not the case for CLIR since the WSD methods were hindered by a lack of context and were not particularly accurate. The reason for this difference may be that retrieval algorithms

actually perform a similar purpose to WSD algorithms in the sense that they attempt to identify instances of words being used with the relevant meanings. WSD algorithms therefore need to be accurate to provide any improvement. The situation is different for CLIR where identifying the correct translation of words in the query is unavoidable. This can only be carried out using some disambiguation method and the results presented here suggest that some disambiguation is better than none for CLIR.

7 Conclusions

The results presented in this paper show that WSD is useful when CLIR was being carried out using EWN. The WSD algorithm used was not highly accurate on this particular task however it was able to outperformed two simple baselines and did not appear to adversely effect the retrieval results.

In future work we plan to experiment with different languages which are supported by EWN to test whether the differences in lexical coverage of the various EWNs have any effect on retrieval performance. One of the authors has already shown that combining WSD algorithms can be a useful way of improving their effectiveness for ontology construction [16]. We plan to test whether similar techniques could be employed to improve the automatic disambiguation of queries.

Acknowledgments

The authors are grateful for advice from Mark Sanderson and Wim Peters of Sheffi eld University. The work described here was supported by the EPSRC-funded Eurovision project at Sheffi eld University (GR/R56778/01).

References

1. Ballesteros, L., Croft, W.: Resolving ambiguity for cross-language retrieval. In: Research and Development in Information Retrieval. (1998) 64–71.
2. Jang, M., Myaeng, S., Park, S.: Using mutual information to resolve query translation ambiguities and query term weighting. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99), College Park, MA (1999) 223–229.
3. Vossen, P.: Introduction to EuroWordNet. Computers and the Humanities **32** (1998) 73–89 Special Issue on EuroWordNet.
4. Gilarranz, J., Gonzalo, J., Verdejo, F.: Language-independent text retireval with the EuroWordNet Multilingual Semantic Database. In: Proceedings of the Second Workshop on Multilinguality in the Software Industry: the AI contribution, Nagoya, Japan (1997) 9–16.
5. Resnik, P.: Disambiguating Noun Groupings with Respect to WordNet senses. In Armstrong, S., Church, K., Isabelle, P., Manzi, S., Tzoukermann, E., Yarowsky, D., eds.: Natural Language Processing using Very Large Corpora. Kluwer Academic Press (1999) 77–98.
6. Schaüble, P., Sheridan, P.: Cross-Language Information Retrieval (CLIR) Track Overview. In Voorhees, E., Harman, D., eds.: The Sixth Text REtrieval Conference (TREC-6), Gaithersburg, MA (1997) 31–44.
7. Cutting, D., Kupiec, J., Pedersen, J., Sibun, P.: A practical part-of-speech tagger. In: Proceedings of the Third Conference on Applied Natural Language Processing, Trento, Italy (1992) 133–140.

8. Robertson, S., Walker, S., Beaulieu, M.: Okapi at TREC-7: automatic ad hoc, filtering VLC and interactive track. In: NIST Special Publication 500-242: The Seventh Text REtrieval Conference (TREC-7), Gaithersburg, MA (1998) 253–264.
9. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison Wesley Longman Limited, Essex (1999).
10. Krovetz, R., Croft, B.: Lexical ambiguity and information retrieval. ACM Transactions on Information Systems **10** (1992) 115–141.
11. Sanderson, M.: Word sense disambiguation and information retrieval. In: Proceedings of the 17th ACM SIGIR Conference, Dublin, Ireland (1994) 142–151.
12. Fellbaum, C., ed.: WordNet: An Electronic Lexical Database and some of its Applications. MIT Press, Cambridge, MA (1998).
13. Gao, J., Nie, J., He, H., Chen, W., Zhou, M.: Resolving query translation ambiguity using a decaying co-occurence model and syntactic dependence relations. In: Proceedings of the 25th International ACM SIGIR Conference on Research and Development in Information Retreival, Tampere, Finalnd (2002) 183–190.
14. Qu, Y., Grefenstette, G., Evans, D.: Resolving translation ambiguity using monolingual corpora. In: Cross Language Evaluation Forum 2002, Rome, Italy (2002).
15. Jing, H., Tzoukermann, E.: Information retrieval based on context distance and morphology. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99), Seattle, WA (1999) 90–96.
16. Stevenson, M.: Augmenting Noun Taxonomies by Combining Lexical Similarity Metrics. In: Proceedings of the 19th International Conference on Computational Linguistics (COLING-02), Taipei, Taiwan (2002) 953–959.

The Topology of WordNet: Some Metrics

Ann Devitt and Carl Vogel

Computational Linguistics Group, Trinity College, Dublin, Ireland,
Email: devitta@cs.tcd.ie, vogel@cs.tcd.ie

Abstract. This paper outlines some different metrics intended for measuring node specificity in WordNet. Statistics are used to characterise topological properties of the overall network.

1 Introduction

Much work has been done on the notion of semantic relatedness between nodes in WordNet, (see [1] for a comprehensive survey of relatedness measures). This paper addresses a similar question – how comparable are two synsets in the WordNet network, not in terms of their content but in terms of the level or granularity or specifi ty they represent.

Although WordNet is a substantial knowledge base, it is not comprehensive. We do not know of work that records comparisons with arguably comparable resources like that supplied by CYC [3], however we expect that variant sparseness of coverage is endemic to all comparable knowledge bases. The level of detail in certain domains is essentially an accident of production dependent on the day, on the lexicographer, on the level of interest, etc. (for a case in point, note the recent addition of numerous concepts related to terrorism in WordNet 2, given the current political climate). Applications that use the data in WordNet to carry out some NLP task may themselves be subject to its vagaries. For example, two towns of comparable size in Ireland, Limerick and Drogheda, Limerick is encoded as both a port city and a type of poem where as Drogheda is encoded as a battle, being the site of a 16th century battle. A topic identifi er using WordNet as its knowledge base might identify texts about Drogheda to be historical or military, without the second possibility of the topic relating to modern day Ireland.

The aim of this paper is to record statistics about version 1.1.7. that are relevant to our ongoing work in defi ning a notion of specifi city that is determined by the topology of WordNet, and sensitive to variance in coverage across topic areas in WordNet. The measures are applicable to any knowledge source that has a defi nable topology. The results here are based on an amalgamation of link types assumed in WordNet but, a clear generalization is to factor in link types among nodes. Topological defi nitions in networks of heterogeneous links have been proposed before [6]. However, it is not yet clear whether any are fully appropriate to the sort of reasoning one would wish to do with WordNet.

The paper is divided into sections each detailing some basic measures for WordNet that characterize its overall topology: graph and node type §2 taxonomic distribution §3, parentage §4, node degree §5, depth and height §6 and clustering coeffi cients §7. Section 8 sets out some conclusions regarding what information has been gained on how these measures may be combined in an effort to determine node specifi city in WordNet.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 106–111.

2 Some Basic Measures

WordNet [2] version 1.1.7 contains 74488 noun synsets. As this paper deals with the *structure* of WordNet rather than its content, we refer to WordNet and its synsets in terms of a graph, a directed acyclic graph and not a tree as it allows multiple inheritance. Henceforth, we use "node" and "synset" interchangeably. Of these synsets or nodes, 58586 or 78.65% are leaf nodes, leaving 15902 internal nodes. Analysis of particular measures across WordNet, such as height and branching factor, must take account of the fact the almost 60,000 leaf nodes may and often do skew results.

3 Dimensional Distribution

There are nine designated most general root nodes to dimensions of the taxonomy, namely:

1 Entity
2 Abstraction
3 Group
4 Act, human action, human activity
5 Psychological feature
6 State
7 Phenomenon
8 Event
9 Possession

The node distribution in these hierarchies is set out in bar chart 1.[1]. As we can see from the chart, the Entity hierarchy is by far the largest and as such merits some investigation as a separate unit. This is concrete evidence of an aspect of the variance mentioned in §2.

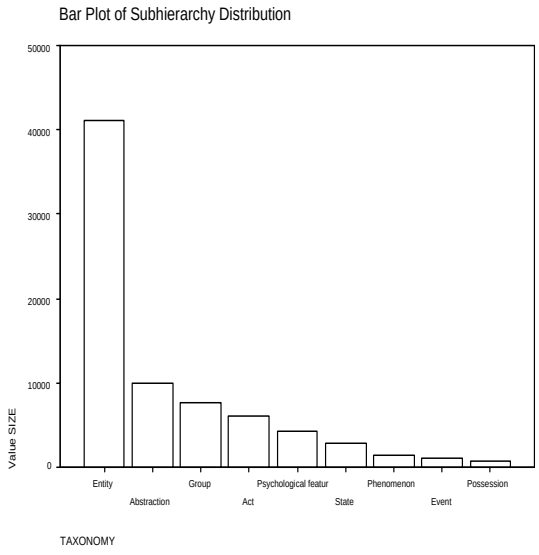


Fig. 1. Bar chart of synset distribution in top hierarchies

[1] The numbers refer to the numbers in the above list

4 Multiple Inheritance Quantified

As noted above, the taxonomy does allow multiple inheritance.

Example 1. The node referring to the multi-talented "Harley Granville-Barker" inherits from the more general nodes: "actor", "critic", "theatre producer", "director" and "playwright"

Example 2. Similarly here, the more general "sphere" and "model" nodes are parents of the synset representing "globe"

In all, these multiple inheritance nodes amount to just 2.28% of the total taxonomy. The histogram in Figure 2 shows the distribution of nodes with more than one parent according to their depth in the hierarchy. The histogram would strongly suggest that these multiple inheritance nodes are normally distributed throughout the depth of WordNet and, thence, their effects propogate down the hierarchy.

However, according to a χ^2 test of independence the distribution of multiple parent nodes in the hierarchy is significantly different within different subhierarchies, χ^2 (8, N=75180)=324.27, $p \leq 0.001$. Thus multiple inheritance is significantly more prevalent in certain sub-hierarchies.

One would expect that multiple parentage would imply a more specific concept node, from a content point of view. One might also posit that nodes deeper in the hierarchy are more specific. In this case, synsets in the right tail should be of comparable high specificity. Content inspection reveals the following as a sample of the highly-specific concepts in the right-tail of the distribution.

Example 3. sea bass, cytology, self-condemnation, bombardon

While nodes in the left tail, though with multiple parents, are less specific due to their position in the hierarchy

Example 4. person, artefact

It should be noted that multiple inheritance does not entail an overlap across sub-hierarchies. Only 689 synsets inherit from two distinct subhierarchies and of these only 6 inherit from more than two.

We hope to combine these topological measures to give a dependable measure of content specificity.

5 Branching Factor

The measure of node degree or branching factor here assumes the notion of dominance. Hence,

BranchingFactor= NoOfDescendants + 1 (the node itself).

This is to avoid problems with zero values in subsequent metrics and corresponds to the normal definition of dominance as a reflexive relation [4].

Branching factor (BF) in WordNet ranges from 1 to 573 with an average value of 2.023. Excluding leaf nodes (i.e., BF=1), however, the average branching factor value rises to 5.793.

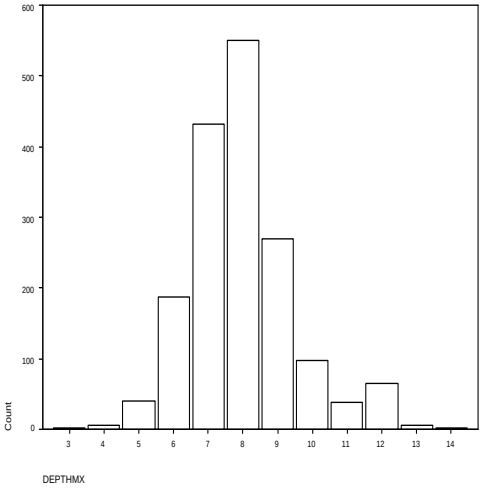


Fig. 2. Histogram of depth for nodes with multiple parentage

Indeed, 75.8% of the remaining 15902 nodes have a branching factor of less than 5 and almost 97% a value of less than 20.

A χ^2 test for BF>4, shows a significant difference in distribution in the phenomenon sub-hierarchy, $\chi^2(1,N=16406) = 11.23$, $p \leq 0.001$ alone.

This suggests that overall, in all subhierarchies, the structure is not shallow: small branching with a large number of total nodes suggests greater overall depth in paths. In the following section, we explore the notion of depth further.

6 Depth and Height

As each node may be parent to or descendant of several lineages, nodes may have several possible values for both height and depth. The values discussed here are

- Maximum depth: longest path from node to a top taxonomy node,
- Minimum depth: shortest path from node to a top taxonomy node,
- Maximum height: longest path from node to a leaf node, and
- Minimum height: shortest path from node to a leaf node.

The distribution of depth values in WordNet whether maxima or minima is normal (see figure 3). The data excluding leaf nodes is not substantially different. The means differ by 0.5 (7.1 with leaf nodes, 6.6 without) but the distribution is comparable.

The data for height, however, displays the effects of the preponderance of leaf nodes in the taxonomy.[2] The maximum distance from any node to a leaf node is 5. Two-thirds of all internal nodes are a single node from the bottom of the taxonomy and 93.6% of

[2] Both the data including and the data excluding leaf nodes display the same characteristics. Therefore we confine the discussion to maximum and minimum heights over all of WordNet

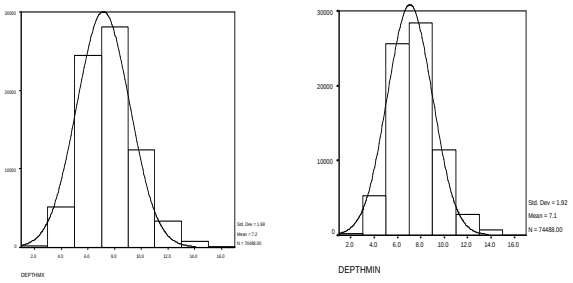


Fig. 3. Histogram of Maximum and Minimum Depth

nodes are a mere 1 or 2 nodes from a leaf node. In fact, for all values of the minimum height variable, the distribution of the depth variable is normal. Figure 4 shows that for both maximum and minimum height values, the distribution is common in natural language: a Zipfi an distribution, decreasing at an exponential rate.

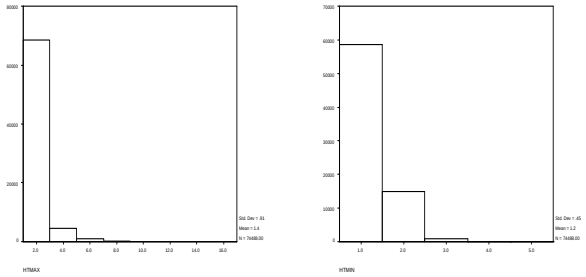


Fig. 4. Maximum and Minimum Height from a leaf node

Given the distribution of measures of height in WordNet, it would seem that depth may be a better measure of specifi city. A minimum height value of 2 does little to suggest how precise a concept may be, for within this selection are the following sample nouns: *production, voodoo, group, refracting telescope, citizenry and floor*.

It should be noted that the distribution of these measures is similar within the nine sub-taxonomies of WordNet.

7 Clustering Coeffi cients

Clustering coeffi cients as a fi ne-grained measure of graph topology and connectivity have been posited in [7]. It measures the relative number of connections between neighbouring nodes in a network, hence, how clustered an area of a network may be. The formula to calculate the clustering coeffi cient C_i of a node i is as follows, where k_i is the number of

connections to its neighbouring nodes and E_i is the number of connections between those k_i nodes.

$$\frac{2\Sigma_i}{k_i(k_i - 1)}$$

Higher-order coeffi cients measure connectivity between a node's immediate and more distant neighbours to a specifi c distance. The coeffi cient gives a normalized measure of connectivity across a whole graph.

A fi rst point to note is that the basic cluster coeffi cient is not useful for a graph such as WordNet. Only 62 synsets have a coeffi cient higher than zero. This would indicate that the nodes in WordNet do not form strong clusters readily. This is clearly due to the hierarchical rather than network structure of the taxonomy.

The higher order measure, taking immediate neighbours and nodes at one extra remove, is a more useful value, particularly for internal nodes, where the distribution is normal and the mean is 0.337.

This would suggest that although WordNet is not tightly clustered, its nodes may form clusters of wider diameter.

8 Some Conclusions on Node Specifi city Measures

The measures set out in the previous sections go some way to outlining the topology of WordNet. We have looked at the contrasting distributions of depth and height, the related concepts of branching factor and cluster coeffi cients, the notion of multiple inheritance and its signifi cance within the taxonomy.

A model of the topology of WordNet would be useful in guiding interpretation of its content, particularly for non-humans, somewhat in the same way as Sperber and Wilson's relevance theory [5] requires a specifi c logic to guide inference steps. The more information we have about the shape of the structure in abstract, the more we way be able to extract from the knowledge base in particular.

We are currently working on a qualitative evaluation of various composite measures, combinations of the metrics discussed here using Principal Components Analysis and heuristics, in order to determine specifi city of nodes in WordNet.

References

1. Alexander Budanitsky: Lexical Semantic Relatedness and its Application in Natural Language Processing Tech. Rep. CSRG-390 Department of Computer Science, University of Toronto (1999).
2. Christiane Fellbaum: WordNet,an electronic lexical database The MIT Press (1990).
3. D. B. Lenat, R. V. Guha: Building Large Knowledge Based Systems, Reading, Massachusetts: Addison Wesley (1990).
4. Barbara Partee, Alice ter Meulen and Robert Wall: Mathematical Methods in Linguistics, Kluwer Academic Publishers (1993).
5. Dan Sperber and Deirdre Wilson: Relevance: Communication and cognition (2^{nd} ed.) Oxford: Blackwell, (1995).
6. David Touretzky: The Mathematics of Inheritance Systems, Los Altos, CA: Morgan Kaufman (1986).
7. D. J. Watts and S. H. Strogatz: Collective dynamics of small world networks, Nature **401**, 130 (1999).

Comparing Lexical Chain-based Summarisation Approaches Using an Extrinsic Evaluation

William Doran, Nicola Stokes, Joe Carthy, and John Dunnion

Department of Computer Science,
University College Dublin, Ireland.
Email: William.Doran@ucd.ie, Nicola.Stokes@ucd.ie, Joe.Carthy@ucd.ie,
John.Dunnion@ucd.ie

Abstract. We present a comparative study of lexical chain-based summarisation techniques. The aim of this paper is to highlight the effect of lexical chain scoring metrics and sentence extraction techniques on summary generation. We present our own lexical chain-based summarisation system and compare it to other chain-based summarisation systems. We also compare the chain scoring and extraction techniques of our system to those of several other baseline systems, including a random summarizer and one based on tf.idf statistics. We use a task-orientated summarisation evaluation scheme that determines summary quality based on TDT story link detection performance.

1 Introduction

Summarisation is a reductive transformation of a source text into a summary text by extraction or generation [13]. It is generally agreed that automating the summarisation procedure should be based on text understanding that mimics the cognitive processes of humans. However, this is a sub-problem of Natural Language Processing (NLP) and is a very difficult problem to solve at present. It may take some time to reach a level where machines can fully understand documents, in the interim we must utilise other properties of text, such as lexical cohesion analysis, that do not rely on full comprehension of the text.

Lexical cohesion is the textual property responsible for making the sentences of a text seem to "hang together", indicated by the use of semantically related vocabulary [10]. Cohesion is thus a surface indicator of the discourse structure of a document. One method of representing this type of discourse structure is through the use of a linguistic technique called lexical chaining. Lexical chains are defined as clusters of semantically related words. For example, {*house, loft, home, cabin*} is a chain, where *house* and *home* are synonyms, *loft* is part of a *house* and *cabin* is a specialisation of *house*. The lexical chaining algorithms discussed in this paper identifies such lexical cohesive relationships between words using the WordNet taxonomy [9].

Since lexical chains were first proposed by Morris and Hirst [10], they have been used to address a variety of Information Retrieval (IR) and NLP applications, such as term weighting for IR tasks [15], malapropism detection [14], hypertext generation [6] and topic detection in broadcast news streams [16], to name but a few. More importantly however, in the context of this paper, lexical chains have been successfully used as an intermediate source text

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 112–117.

representation for document summarisation. This application of lexical chaining was first implemented by Barzilay and Elhadad [3]. They used lexical chains to weight the contribution of a sentence to the main topic of a document, where sentences with high numbers of chain words are extracted and presented as a summary of that document.

In this paper, we put forward a novel method of building extractive summaries of single documents using lexical chains. However, unlike other attempts to improve upon Barzilay and Elhadad's work [1,4,12], we evaluate our weighting and extraction schemes directly with theirs using an extrinsic or task-based evaluation technique. An intrinsic evaluation is the preferred method of evaluating summary quality used by most summarisation researchers. This type of evaluation requires a set of human judges to either create a set of gold standard summaries or score summary quality compared to the original text. However, this evaluation method is time consuming, expensive and quite often subjective and hence is inappropriate for estimating the effect of different schemes on summary performance. Therefore in this paper we propose a more efficient evaluation alternative based on the TDT story-link detection task [2], where summary quality is evaluated with respect to how well a story link detection system can determine if a pair of document summaries are similar (on-topic) or dissimilar (off-topic). We are also interested in finding out whether this type of evaluation is sensitive enough to pick up differences in the summary extraction techniques discussed in this paper. In the remainder of the paper, we explain in more detail how lexical chaining based summarisation works and how our work differs from Barzilay and Elhadad's. We also present our experimental methodology and results, the final section gives our conclusions and some future work.

2 Lexical Chaining and Text Summarisation

The basic chaining algorithm follows the following steps. First, we select a set of candidate words, generally nouns. Then search through the list of chains and if a word satisfies the relatedness criteria with a chain word then the word is added to the chain, otherwise a new chain is created.

The relatedness criteria are the relationships outlined by St.Onge [14]. St. Onge used WordNet [9] as the knowledge source for lexical chaining. He devised three different relationships between candidate words: extra-strong, strong and medium-strong. Extra-strong relations are lexical repetitions of a word and strong relations are synonyms or near-synonyms. Strong relations can also indicate a shared hypernym/hyponym or meronym/holonym, such that one word is a parent-node or child-node of the other in the WordNet topology. Medium-strength relations follow sets of rules laid out by St. Onge. These rules govern the shape of the paths that are allowable in the WordNet structure. St. Onge's algorithm uses a greedy disambiguation procedure where a word's sense is determined only by the senses of words that occur before it in the text. In contrast, a non-greedy approach waits until all words in the document are processed and then calculates the appropriate senses of all the words.

In general, most lexical chain based summarizers follow the same approach by firstly generating lexical chains, then the 'strongest' of these chains are used to weight and extract key sentences in the text. Barzilay and Elhadad [3] form chains using a non-greedy disambiguation procedure. To score chains they calculate the product of two chain characteristics: the length of the chain, which is the total number of words in the chain plus

repetitions and, the homogeneity of the chain, which is equal to 1 minus the number of distinct words divided by the length of the chain. Chain scores that exceed an average chain score plus twice the standard deviation are considered 'strong' chains. Barzilay et al. then select the first sentence that contains a 'representative' word from a 'strong' chain, where a 'representative' word has a frequency greater than or equal to the average frequency of words in that chain.

Most other researchers use this approach to building extractive summaries using lexical chains [1,12], with the exception of Brunn et al. [4] who calculate chain scores as the pairwise sum of the chain word relationship strengths in the chain. In the latter, sentences are ranked based on the number of 'strong' chain words they contain.

3 The LexSum System

Our chaining algorithm LexSum is based on [14,16] and uses a greedy lexical chaining approach. The first step in our chain formation process is to assign parts-of-speech to an incoming document. The algorithm then identifies all noun, proper nouns and compound noun phrases by searching for patterns of tags corresponding to these types of phrases e.g. presidential/JJ campaign/NN, or U.S/NN President/NN Bush/NP where /NN is a noun tag and /NP is a proper noun tag.

The nouns and compound nouns are chained by searching for lexical cohesive relationships between words in the text by following constrained paths in WordNet similar to those described in [14] using lexicographical relationships such as synonymy (*car, automobile*), specialisation/generalisation (*horse, stallion*), part-whole/whole-part (*politicians, government*). However, unlike previous chaining approaches our algorithm produces two disjoint sets of chains: noun chains and proper noun chains. Finding relationships between proper nouns is an essential element of modelling the topical content of any news story. Unfortunately, WordNet's coverage of proper nouns is limited to historical figures (e.g. Marco Polo, John Glenn) and so our algorithm uses a fuzzy string matching function to find repetition relationships between proper nouns phrases like George_*Bush* ? President_*Bush.*

Unlike Barzilay et al.'s approach, our algorithm calculates chain scores based on the number of repetitions and the type of WordNet relations between chain members. More specifically, as shown in equation1, the chain score is the sum of each score assigned to each word pair in the chain. Each word pair's score is calculated as the sum of the frequencies of the two words, multiplied by the relationship score between them,

$$chain\_score(chain) = \sum (reps_i + reps_j) * rel(i, j) \quad (1)$$

where $reps_i$ is the frequency of word i in the text, and *rel(i,j)* is a score assigned based on the strength of the relationship between word i and j, where a synonym relationship gets assigned a value of 0.9, specialisation/generalisation and part-whole/whole-part 0.7. Proper nouns chain scores are calculated depending on the type of match, 1.0 for an exact match, 0.8 for a partial match and 0.7 for a fuzzy match.

The next step in the algorithm ranks sentences based on the sum of the scores of the words in each sentence, where a word's score is a scaled version of its chain's score. The scaling factor is the minimum distance between a word and its predecessor or its successor in the chain. This idea is based on the fact that general topics tend to span large sections of

a discourse whereas subtopics tend to populate smaller areas. [7]. Therefore, the score of a word will be increased if semantically similar words are close by it in the text i.e. the topic is in the focus of the reader.

$$word\_score(word_i) = \alpha * chain\_score(chain(word_i)) \tag{2}$$

$$\alpha = 1 - \left(\min\left[dist(w_{i-1}, w_i), dist(w_i, w_{i+1})\right]/dist(w_1, w_n)\right) \tag{3}$$

Where $dist(w_i, w_j)$ is the number of words that separate two words in the text and $chain(word_i)$ is the chain $word_i$ belongs to. As explained earlier the sentence score is the sum of these word scores normalized with respect to the length of the sentence and the number of chain words it contains.

4 Experimental Methodology and Results

As explained above, we use a task-oriented evaluation methodology to determine the performance of our lexical chain based summarizer, as this type of evaluation can be automated and hence more efficient than an intrinsic evaluation that involves the time and efforts of a set of human judges. It also provides us with a means of evaluating summary performance on a larger than normal data set of news stories used in the DUC evaluation, i.e. 326 TDT documents and 298 TREC documents [5]. While intrinsic evaluation gauges summary quality directly by rating summary informativeness and coherency, extrinsic evaluation gauges the impact the summary generation procedure has on some task, thus indirectly determining summary quality. Several such tasks have been outlined as useful by TIPSTER [8], such as ad-hoc retrieval, categorization and question answering tasks.

In this paper we use the TDT Story Link Detection Task [2]. TDT is a research initiative that investigates the event-based organisation of news stories in a broadcast news stream. Story Link Detection (SLD) is the pair-wise comparison of stories to establish whether they discuss the same event. Thus for each distinct set of summaries generated (by each system), we evaluate summary quality by observing whether the SLD system can distinguish between on-topic and off-topic document summary pairs. Hence, the hypothesis underlying this type of summary evaluation is that an SLD system will perform well on summaries that have retained the core message of each news story, while it will perform poorly on summaries that in general failed to recognise the central theme of the documents in the data set. Our SLD system is based on an IR vector space model where document similarity is determined using the cosine similarity function [17]. As in the TDT initiative, we evaluate story link detection performance using two error metrics: percentage misses (document pairs that are incorrectly tagged as off-topic) and false alarms (document pairs that are incorrectly tagged as on-topic). A Detection Error Trade-off (DET) graph is then plotted for misses and false alarms rates at various similarity thresholds (ranging from 0 to 1) where a DET curve is produced for each set of generated summaries. Optimal SLD performance can then be determined by observing which of these curves lies closest to the origin, i.e. has the lowest miss and false alarm rates.

We evaluated three baseline systems LEAD, TF-IDF, and RANDOM, together with our own system, LexSum, using this evaluation strategy. The LEAD system creates summaries from the lead paragraph of each document, since news stories tend to contain a summary

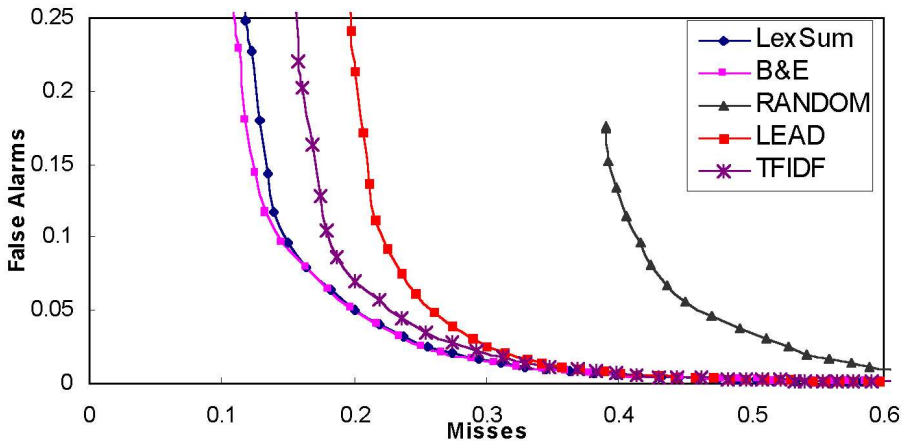


Fig. 1. This DET graph shows the Story Link Detection results of summaries (at a compression rate of 50%)

of the article in the first paragraph. The TF-IDF system extracts sentences which have high *tf-idf* weights values, where *tf-idf* is a term weighting scheme that is commonly used in IR research [17]. The final baseline extracts sentences at random from the source document and uses these as a summary. We also created a system, B&E that replicates Barzilay and Elhadad's scoring metric. We modified the B&E extraction technique to enable us to generate summaries of different lengths.

We generated summaries for all summarisers at summary compression rates of 10, 20, 30, 40, 50 and 60 percent (of the top ranked sentences in the text). Each of these summary sets was given as input to the SLD system and DET graphs were produced. Figure 1 is a DET graph illustrating the results for each summarisation system running at 50% compression. This graph is indicative of the general trend for all the compression rates. Both lexical chain systems outperform the baseline systems for all percentages except at 10% where the LEAD performs better. As expected the RANDOM summariser has the worst performance. The fact that lexical chain based summarisers outperform TFIDF, suggests that observing patterns of lexical cohesion is a more accurate means of identifying core themes in documents than using corpus statistics like *tf.idf*. Another observation from these experiments is that B&E's weighting scheme marginally outperforms ours at high false alarm and low miss rates; however this result is not statistically significant.

5 Conclusions and Future Work

In this paper, we have analysed some of the factors that affect lexical chain based summarisation using an extrinsic evaluation methodology. We found that the effect of the weighting scheme has little effect on the summaries. It is likely that both lexical chain based systems are selecting the same sentences, the extent of this trend warrants further

investigation. Both chaining systems perform better than the TF.IDF and LEAD systems, justifying the extra computation involved in lexical chaining. Also, the evaluation method proved to be sensitive enough to show the differences between the baseline systems and the lexical chain based systems. It is our intention to carry out an intrinsic evaluation of the summarisation systems described in this paper and compare these human-deduced summary quality ratings with the results of the automated evaluation presented above.

References

1. Alemany, L. and Fuentes M., 2003, *Integrating Cohesion and Coherence for Text Summarization.* In the Proceedings of the EACL Student Workshop, 2003.
2. Allan J., 2002, Introduction to Topic Detection and Tracking, In *Topic Detection and Tracking: Event-based Information Organization*, Kluwer Academic Publishers, pp. 1–16.
3. Barzilay R. and Elhadad M., 1997, Using Lexical Chains for Summarisation. In *ACL/EACL-97 summarisation workshop*. Pp 10–18, Madrid.
4. Brunn M., Chali Y., and Pinchak C. 2001, Text summarisation using lexical chains, In *Workshop on Text Summarisation in conjunction with the ACM SIGIR Conference 2001*, New Orleans, Louisiana, 2001.
5. DUC 2003 `http://www-nlpir.nist.gov/projects/duc/`.
6. Green, S. 1997, *Automatically generating hypertext by computing semantic similarity*, PhD thesis, University of Toronto.
7. Hearst, M. 1994, Multi-paragraph segmentation of expository text, In *Proceedings of the 32th Annual Meeting of the Association for Computational Linguistics*, 9–16. Las Cruces, New Mexico: Association for Computational Linguistics.
8. Mani I., House, D., Klein, G., Hirschman, L., Obrst, L., Firmin, T., Chrzanowski, M. and Sundheim, B. 1998, The TIPSTER SUMMAC text summarisation evaluation: Final report. MITRE Technical Report MTR 98w0000138, MITRE.
9. Miller G. A., Beckwith R., Fellbaum C., Gross, D., and Miller, K. 1990, *Five papers on WordNet*. Technical Report, Cognitive Science Laboratory, 1990.
10. Morris, J. and Hirst, G. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text, *Computational Linguistics* 17(1): 21–43.
11. Salton, G., Singhal, A., Mitra, M., and Buckley, C. 1997, Automatic text structuring and summarisation. *Information Processing and Management* 33(2):193–208.
12. Silber, G. and McCoy, K.. 2000, Efficient Text Summarisation Using Lexical Chains, In *Proceedings of the ACM Conference on Intelligent User Interfaces (IUI'2000)*.
13. Spark-Jones, K. 2001, Factorial Summary Evaluation, In *Workshop on Text Summarisation in conjunction with the ACM SIGIR Conference 2001*. New Orleans, Louisiana.
14. St. Onge, D. 1995, *Detection and Correcting Malapropisms with Lexical Chains*, M.Sc Thesis, University of Toronto, Canada.
15. Stairmand, M. 1996, *A Computational Analysis of Lexical Cohesion with Applications in Information Retrieva*,. Ph.D. Dissertation, Center for Computational Linguistics, UMIST, Manchester.
16. Stokes, N., J. Carthy, A.F Smeaton, SeLeCT: A Lexical Chain-based News Story Segmentation System. To appear in the AI Communications Journal.
17. van Rijsbergen, C. J., Information Retrieval, Butterworths, 1979.

Use of Wordnet for Retrieving Words from Their Meanings

İlknur Durgar El-Kahlout and Kemal Oflazer

Sabancı University
Faculty of Science and Nature
34956 Orhanlı -Tuzla İstanbul, Turkey
Email: ilknurdurgar@su.sabanciuniv.edu, oflazer@sabanciuniv.edu

Abstract. This paper presents a Meaning to Word System (MTW) for Turkish Language, that finds a set of words, closely matching the definition entered by the user. The approach of extracting words from "meaning"s is based on checking the similarity between the user's definition and each entry of the Turkish database without considering any semantics or grammatical information. Results on unseen user queries indicate that in 66% of the queries the correct responses were in the first 50 of the words returned, while for queries selected from the word definitions in a different dictionary in 92% of the queries correct responses were in the first 50 of the words returned. Our system make extensive uses of various linguistics resources including Turkish WordNet.

1 Introduction

Suppose one can not remember a word but knows a variety of contextual phrases that approximate his or her understanding of the word and wants to find the appropriate word (or words) that has similar meaning with his/her definition. For this problem, it will be of no use to attempt searching in a traditional dictionary to find the word. Traditional dictionaries are helpful for finding the meaning of a word but we need an application that works in the opposite direction.

Some examples of definitions taken from users and the corresponding meanings of those words taken from dictionary [1] are listed below.

- akımölçer (ammeter)
  - **User Definition:** *akımı ölçmek için kullanılan alet* (a device that is used to measure the current).
  - **Dictionary Definition:** *elektrik akımının şiddetini ölçmeye yarayan araç, amperölçer* (a device that measures the intensity of electrical current, amperemeter).
- istifa (resignation)
  - **User Definition:** *çalıştığı işten kendi isteğiyle ayrılmak* (leaving one's job voluntarily).
  - **Dictionary Definition:** *kendi isteğiyle görevden ayrılma* (leaving voluntarily, of a position).

The definitions collected from a set of users showed us that users usually define the words very similar to the actual dictionary definitions in terms of meaning. By using this similarity, we implement a system called Meaning to Word (MTW) for Turkish to find the appropriate words whose definitions match the given definition.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 118–123.

2 Meaning to Word

While finding the appropriate words, MTW deals with two challenging problems: (i) locating a number of candidate words whose definitions are "similar" to the definition in some sense, (ii) ranking these candidate words using a variety of ways to return a list sorted in terms of similarity. Our approach for extracting words from meanings is based on checking the similarity between the user definition and each entry of the dictionary by making a number analyses without taking into consideration the semantics or the context.

MTW works as follows: A user definition is given as an input to the system. The user definition is processed to construct a query. With this query, the database is searched and a list of candidate words is generated. The candidate words are sorted in terms of similarity and the list is returned to the user as a result. It should be noted that all the processing steps are fully automated, no human intervation or manual encoding is required. We use NLP techniques to enhance the effectiveness of term-based information retrieval.

2.1 Databases and Other Sources of Information

We use two resources in retrieving appropriate words for the user request. These sources are the explatonary Turkish Dictionary and Turkish WordNet.

MTW uses the Turkish dictionary to search in and match the corresponding meanings to the user's request. Dictionary has alphabetically sorted words and their meanings with 89,019 entries, 82,489 unique words and 21,653 unique stems.

Also, MTW uses Turkish WordNet to find the relations between words. Turkish WordNet [2] is structured in a similar way as the WordNet [3] around the notion of a synset. Synsets are linked across basic semantic relations such as hyponymy/hypernymy,antonymy and meronymy.

2.2 Query Generation

MTW does not use the user definition as it is; a set of useful information from the definition is selected with simple NLP techniques to form a query [4]. The steps are as follows:

Tokenization: We divide the symbols into two parts: Word symbols and non-word symbols. Characters other than letters and digits are treated as non-word symbols (e.g. punctuation marks)and eliminated from the definition because they are unnecessary for further retrieval.

Stemming: Because of the structure of Turkish, the words of the user's definition and the corresponding definition in the dictionary may have the same stem but different affixes. Stemming enables matching different morphological variants of the original definition's words.

Stop Word Removal: Stop words are words that contribute nothing or very little meaning; they should be removed from the query and dictionary definitions. If a word occurs frequently in a dictionary or has little meaning conceptually (such as prepositions, determiners), then it is not an informative word. The top $200 - 300$ frequent words in the dictionary and conceptually little meaning words are selected as stop words and removed from dictionary definitions and the query.

Stemming process takes place before the stop word removal because of the structure of Turkish. For example, the words *bir* (one), *biri* (one of them), *birileri* (some people) have the frequencies 19901, 12 and 2, respectively. Although all of the words have the same stem *bir* (one), it is possible to eliminate only the word *bir* (one) with the given frequencies.

2.3 Query Processing

While searching for the appropriately matching meaning, rarely all of the query words match the relevant meaning. For this reason, an approximate match is more suitable than the exact match of user's request with the dictionary meanings. In MTW, sub-queries are generated by using different combinations of words from the original query. Then, MTW sorts the sub-queries in order to their informativeness.

Subset Generation: MTW generates all $2^n - 1$ sub-queries for a n word query, where n is the number of words remaining in the query after stop word removal. Table 1 shows sub-queries generated from the query *yazlık büyük ev* (large house for summer). **Subset**

Table 1. Subset generation table for query *yazlık büyük ev* (large house for summer)

Subset number	yazlı k	büyük	ev	Generated subset
1	1	1	1	yazlı k büyük ev (large house for summer)
2	1	1	0	yazlı k büyük (large for summer)
3	1	0	1	yazlı k ev (house for summer)
4	1	0	0	yazlı k (for summer)
5	0	1	1	büyük ev (large house)
6	0	1	0	büyük (large)
7	0	0	1	ev (house)

Sorting: Searching the meanings with an unordered sub-query list is not efficient as we can not estimate which sub-query can give the correct meaning. For this reason, we should start from the most informative sub-query first. The sub-queries are sorted in order to the number of words that they contain. This lets the system to find the meanings matching maximum number of words before the others. If there are two meanings that match the same number of words then the system decides which of the sub-query is more informative than the other.

Table 2. Frequencies of each word of the query *yazlık büyük ev* (large house for summer)

Word	Word Occurrence	Stem Occurrence
yazlı k (for summer)	9	12
büyük (large)	931	1168
ev (house)	157	734

From Table 2, the word *yazlık* (for summer) is more informative than the words *büyük* (large) and *ev* (house), and the word *ev* (house) is more informative than the word *büyük* (large). For multi-word sub-queries, the logarithms of word frequencies are added and the result is used to define the information measure of the subset. We use the sum of word frequency logarithms as the frequencies of words are too small and directly multiplying the frequencies will cause information loss. The sorting formula is:

$$relevance\_of\_subset(j) = \frac{\sum_{i \in subset_j} \log(freq_i)}{N_j}. \quad (1)$$

where, $freq_i$ is the frequency of i^{th} word in dictionary and N_j is the number of words of the j^{th} subset.

2.4 Searching for 'Meaning'

Simplest idea for finding the similarity between two phrases is to match the common words of both, and return the best matching meaning. But, user's definition and actual meaning of a word generally have same concepts with different words [5]. For example:

- **User Definition:** *daha önce hiç evlenmemiş olan kişi* (a person who has never been married).
  - **Generated Query:** *daha, {önce, ön}, hiç, evlen, ol, kişi* (yet, {before, front}, never, marry, be, person).
- **Actual Definition:** *evlenmemiş kimse* (unmarried person).
  - **MTW Representation:** *evlen kimse* (marry, person).

At first sight, only the word *evlen* (marry) is matching with the actual meaning. But the words *kişi* (person) and *kimse* (someone) are similar words. Standart matching algorithm using only stems will fail to find this similarity. But for the efficiency of the retrieval, these words should also be counted as "matched". Use of Turkish WordNet helped us to find the possible matching words. In our method, we use the synonym words from the Turkish WordNet and expand the query. In Turkish WordNet there is a synset *{kişi (person), kimse (someone), şahıs, birey (individual), insan (human)}* containing both of the words. The original query contains only *kişi* (person) but the extended one will contain all the synset members including *kimse* (someone). The method is applied to all the words in the original query. The enhanced query will retrieve dictionary definitions with higher ranks.

2.5 Ranking

MTW uses three criteria to rank the candidate definitions: (i) the number of matched words is calculated. If any definition has more common words with the query than others, then this definition is more relevant; (ii) the length of the candidate definition is determined. If two candidates have the same number of matches with the user definition, the shorter candidate is ranked before the longer one; (iii) the longest common subsequence of the candidate definition and user definition is calculated. The definition that have longer common subsequence is ranked before the shorter ones.

Table 3. Results of MTW with all stems included

Rank	train_set	test_set	dict_train_set	dict_test_set
1 – 10	14 (28%)	24 (48%)	45 (90%)	41 (82%)
11 – 50	9 (18%)	9 (18%)	2 (4%)	5 (10%)
51 – 100	3 (6%)	3 (6%)	1 (2%)	2 (4%)
101 – 300	7 (14%)	2 (4%)	2 (4%)	1 (2%)
301 – 500	0 (0%)	1 (2%)	0 (0%)	1 (2%)
501 – 1000	4 (8%)	5 (10%)	0 (0%)	0 (0%)
over 1000	4 (8%)	1 (2%)	0 (0%)	0 (0%)
not found	9 (18%)	5 (10%)	0 (0%)	0 (0%)

3 Performance Evaluation

3.1 Setup

The experiments were carried out on two different test sets: test_set and dict_test_set. In addition, two train sets are used: train_set and dict_train_set. In the experiments 50 user definitions are used for each set. Queries for test_set and train_set are taken from real users. Users are given 50 different words and asked to define these words. Definitions for dict_test_set and dict_train_set are taken from a dictionary [2]. The dictionary definitions of the same 50 words that are given to the users are used as definitions.

3.2 Results

Sometimes stemming algorithms can produce different meaning stems for a word, such as for the query *en yüksek yer* (most highest place), the stemmer gives two different stems *yük* (load) and *yüksek* (high) for the word *yüksek*(high, if we are load) but only the word *yüksek* (high) is the correct stem. We test our method with two different approaches. In the first test, all of the stems returned from the stemmer (i.e., *yük* (load) and *yüksek*(high)) are included in the query. In the second test, a simple heuristic approach is used. We assume that the longest stem (i.e., *yüksek* (high)) returned from the stemmer is the correct stem and include only this stem to the query.

Tables 3 and 4 show the results of the method with all stems and only longest stems, respectively.

With our method, we can match the 66% of the user definitions and 92% of the dictionary definitions by using all stems, and 68% of the dictionary definitions and 90% of the dictionary definitions by using longest stem in the first 50 results. There is a decrease when we select the longest stem because the longest stem may not be the correct stem for every word. Although there is a little increase in the first 50 rank in the test_set, the performance decreases in the first 10 rank.

4 Conclusions

In this paper, we presented the design and implementation of a Meaning to Word system that locates a Turkish word that most closely matches the appropriate one, based on a

Table 4. Results of MTW with only longest stem included

Rank	train_set	test_set	dict_train_set	dict_test_set
1 – 10	15 (30%)	22 (44%)	45 (90%)	40(80%)
11 – 50	8 (16%)	12 (24%)	1 (2%)	5(10%)
51 – 100	3 (6%)	2 (4%)	0 (0%)	4(8%)
101 – 300	6 (12%)	2 (4%)	3 (6%)	1(2%)
301 – 500	1 (2%)	2 (4%)	0 (0%)	0(0%)
501 – 1000	2 (4%)	3 (6%)	0 (0%)	0(0%)
over 1000	6 (12%)	2 (4%)	0 (0%)	0(0%)
not found	9 (18%)	5 (10%)	0 (0%)	0(0%)

definition entered by the user. Using only simple and symbolic methods, the performance results of MTW on unseen data from real users are rather satisfactory. The results on unseen queries from a different dictionary shows that the methods used while implementing MTW are reasonable. MTW has the advantage of free stemming and expansion that gives a great flexibility to retrieval. By stemming and query expansion in MTW, the user's definition can match the correct word(s) even if the terms of the dictionary definition does not contain the same words with same affixes. Besides MTW has the disadvantage of false matches. Because of the noise from the wrong stems and unrelevant synonyms, MTW can produce many irrelevant candidates. MTW works best if the request is typed similar to the actual definition. MTW can be used in various application areas such as computer-assisted language learning, crossword puzzle solving, or as a reverse dictionary.

References

1. Püsküllüoğlu, A.: Arkadaş Türkçe Sözlük. Arkadaş Yayı nevi (1995).
2. Turkish WordNet, [online]: `http://fens.sabanciuniv.edu/TL`. (Accessed: September 5, 2003).
3. Miller, G.: WordNet: A Lexical Database for English. Communications of the ACM 38(11) pp 39–41(1995).
4. Strzalkowski, T.: Natural language information retrieval. Information Processing & and Management 31(3), 397–417 (1995).
5. Voorhees, E. M.: Using WordNet for Text Retrieval. In: Fellbaum, C. (ed.): WordNet – An Electronic and Lexical Database, MIT Press. Cambridge, Mass., (1998), pp 285–303.

Grounding the Ontology on the Semantic Interpretation Algorithm

Fernando Gomez

Dept. of Computer Science
University of Central Florida, Orlando, FL 32816
Email: gomez@cs.ucf.edu

Abstract. Some reorganizations and modifications to the WordNet ontology are explained. These changes have been suggested by extensive testing of the ontological categories with an algorithm for semantic interpretation The algorithm is based on predicates that have been defined for WordNet verb classes. The selectional restrictions of the predicates are WordNet ontological categories.

1 Introduction

This paper, a much shorter version of CS-TR-01-01 with the same title, provides a sample of our reorganizations and changes to the WordNet noun ontology (WordNet 1.6) [6]. These changes have been dictated by a semantic interpretation algorithm reported in [3]. The algorithm is based on predicates, or verbal concepts, that have been defi ned for WordNet verb classes [2]. The semantic roles of the predicates have been linked to the noun ontology and to syntactic relations. After the initial set up, the defi nition of new predicates has been followed by testing them using the algorithm. As of this writing, 3000 predicates have been defi ned and 95% of WordNet verb classes have been mapped into these predicates. In contrast to other ontologies for natural language [1,5], or to efforts to induce a concise set of ontological categories from WordNet [4], the principles guiding our changes have been the selectional restrictions in the semantic roles of the 3000 predicates. Hence, the failure of interpreting a sentence has been the clue for redefi ning some ontological categories. For instance, the concept *written-communication*, which has many subconcepts, is categorized in Wordnet 1.6 only as an *abstraction*. Thus, the interpreter failed to interpret such simple sentences as "She burned the letter/She put the letter on the table," because "letter" does not have *physical-thing* as one of its hypernyms (superconcepts). In "The fi sh frequently hides in a crevice," the interpreter failed to assign meaning to "hides" because "crevice" is categorizedin WordNet 1.6 only as an *abstraction*. In "Blood poured from the wound," the interpreter fails to assign meaning to "poured" because "wound" and its hypernym, "injury," are not as a *physical thing* in WN. The examples are many. This paper is organized as follows. Section 2 and its subsections discuss the concept of *physical-thing* and a few of its main subconcepts. Section 3 and subsections explain a few of the subconcepts of *abstraction6*, and section 4 gives our conclusions.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 124–129.

2 Physical-Thing

The concept of *physical-thing* corresponds to the WordNet 1.6 (henceforth WN) concept of entity1. Most subconcepts of entiy1 are physical things. Those few concepts which are not, such as the synset variable1 have been extracted from entity1. The concept of *physical-thing* is not the same as the concept of *physical-object*(*object1*) in WN. *Physical-objects* are countable while *physical-thing* includes concepts which are not countable such as the concept of *substance*, and concepts which are not physical objects such as the concepts of *physical-process* and *natural-phenomenon*. The latter two are tangled to *process* and *phenomenon*, respectively. The major subconcepts of *physical-thing* that have undergone some reclassifi cation as a result of our analysis are listed next. (We have used the star (*) and indentation to indicate the subconcepts of a given concept. Besides, we have used the arrow to indicate that a concept is also tangled to another concept. If a WN synset corresponding to our concept exists, it is listed in parentheses next to the concept. We have used the expression concept *a* goes to concept *b* in WN, in order to mean that concept *b* is a hypernym, or superconcept, of concept *a*.)

```
Physical-Thing
 * physical-object (object1)
 * location (location1)
 * substance (substance1)
 * physical-group
 * physical-process -> process
 * natural-phenomenon -> phenomenon
```

2.1 Physical-Object

Physical-object has everything in *object1* except *substance1* and *location1*, which have become subconcepts of *physical-thing*. These are the major subconcepts of *physical-object* that have undergone some reclassifi cation.

```
Physical-Object
 * physical-part (part7)
 * animate       (life-form1)
 * artifact      (artifact1)
```

The concept of *part7*, which in our modifi ed WN ontology (henceforth referred as “our ontology”) has been called *physical-part*, has two subconcepts *plant-part*, which in WN goes just to entity1, and *animal-body-part* (body-part1) which in WN goes to part7. In our ontology, *plant-part* and *animal-body-part* have been tangled to the concept *animate* (life-form1 in WN). Thus, we have:

```
physical-part(part7)
 * plant-part(plant-part1) -> animate
 * animal-body-part(body-part1) -> animate
```

The concept of *animate* (life-form1) has undergone few additions, one being *body-cell* (cell2) which in WN goes directly to *entity1*.

2.2 Artifact (Artifact1)

This concept has not undergone much change. However, many of the hyponyms of *structure1*, a hyponym of *artifact1*, have been tangled to *location* because most of its subconcepts (*hospital, building, area*, etc.) are used as locations. They fill the roles *to-loc* or *from-loc* of change of location verbs. More importantly, some of the hyponyms of *structure1* have also been tangled to *organization* because they are used as agents. Most of the subconcepts of *building1*, which is a subconcept of *structure1*, are also used as agents. Some of these concepts are: *tavern, library, hotel, restaurant,*.... This was discovered by failing to interpret sentences such as "The restaurant hired a new chef," and similar ones.

2.3 Location (Location1)

Location1 is directly a subconcept of *physical-object* (object1) in WN. In our ontology, it is a subconcept of *physical-thing*. It seems that the concept *location* is not as much a *physical-object* as the concept, say, *pencil*. One finds the sentences "Peter threw/kicked the pencil" acceptable, but not "Peter threw/kicked Europe" unless one is using them in a figurative sense. That sense is what the distinction between *physical-object* and *physical-thing* tries to grasp. These comments apply strongly to *substance* because this concept is not a countable entity. Some subconcepts of *location* in WordNet have been tangled to *organization* because they are used as such. For instance, the sentence "France invaded Italy during the Napoleonic wars" and many other similar sentences could not be interpreted because "France" was just as a *location* in WordNet. Below are some of these concepts:

```
location
* district ((district1)(territory2))
* state-or-province (state2)
* country  (country1) (state3)
* continent (continent1)
* residential-district
   (residential-district1)
```

State3 contains some few concepts such as *reich, carthage, holy roman empire*. Some subconcepts of *workplace1*, which in WN go to *location*, have been also tangled to *organization*. Some of these are: *farm* and its subconcepts, as well as *fishery, brokerage house* and a few others.

2.4 Physical-Group, Physical-Process, Natural-Phenomenon

WN distinguishes three senses of "group." The first sense of "group," group1, is a unique class containing many concepts. The problem with this is that group1 needs to be linked to the hierarchy, and one needs to decide if group1 must be made a subconcept of *abstraction* or of *physical-thing*. It seems obvious that the concept *group* is an abstraction, meaning a collection of abstract or physical things. However, many subconcepts of group1 or of some of their subconcepts are collections of physical things, e.g., "fleet," "flora," "fauna," "masses," etc. which are all subconcepts of group1 in WN. In the sentence "The hurricane pushed the fleet into the rocks," "push" is used in its physical sense: an inanimate cause causing a change

of location of physical things, namely ships. Thus, we have created the concept *physical-group* that contains as subconcepts all those concepts under group1 which are collections of physical things.

In WN, an important immediate subconcept of group1 is social-group1, which contains many subconcepts. Because social groups are frequently used as agents, in our ontology *social-group* has become a subconcept of *human-agent*, which includes individual humans and social groups. The concepts of "people," "citizenry," "multitude," and others have become subconcepts of *social-group*. Another subconcept of group1, animal-group1, has become a subconcept of *animal*. Animal-group1 contains such concepts as "pride," "flock," "swarm," "herd," etc. which are used as referring to the members of the group rather than to the group itself.

3 Abstraction

Next we discuss the following subconcepts of *abstraction* (abstraction6), namely: *possession2*, which is not a subconcept of *abstraction6* in WordNet, but a unique class. We also discuss the following concepts: *communication* and *space*, which are subconcepts of *abstraction6* in WN.

3.1 Possession (Possession2)

Possession2 (anything owned or possessed) is a unique class in WN, however in our ontology is a subconcept of *abstraction* (abstraction6). A major subconcept of *possession* that is not classified as a subconcept of *possession* in WN is *debt-instrument1*. In WN, *debt-instrument1* is a subconcept of *document3*. In our ontology, it is both a subconcept of *written-communication1* and *possession2*. *Debt-instrument1* contains many subconcepts such as *junk bond, note receivable*, etc. Another subconcept of *document3* which has also become a subconcept of *possession* is *letter of credit*.

One of the hyponyms of *possession2*, *territory2, dominion, territorial dominion, province, mandate, colony*, has been extracted from possession2 and made a subconcept of *location*. Another subconcept of *possession2*, *real-property1*, which contains such concepts as *hacienda, plantation*, etc. has been also extracted and made a subconcept of *location*. Some concepts of *possession2* have been tangled to *physical-thing* and *possession*. The major ones are: *property1, belongings, holding, material possession* which include such concepts as *personal effects, public property* and others. Besides, *currency1* ("the metal or paper medium of exchange that is presently used") and some of the senses of "treasure" have been also tangled to *physical-thing*. The main point to emphasize is that most of the concepts that have remained as subconcepts of *possession* express an abstract relation of ownership, debt, value, liability, etc., although some subconcepts have been tangled to *physical-thing*.

3.2 Communication

The major restructuring in the category *relation* (relation1) has been the subconcept of *communication*. This is the final hierarchy:

```
communication
 * act-of-communicating
         (communication1)
 * something-communicated
         (communication2)
    * written-communication  -> physical-thing
       (written_communication1)
    * print-media (print-media1)
```

In WN, *communication1* goes to *act2, human action, human activity* and *communication2* goes to social-relation1, which goes to *relation1*. Our analysis for these concepts is similar to the ones we have been just discussing, namely creating the concept *communication* to which we have not mapped any WN synset, and making *communication1* and *communication2* subconcepts of *communication*. A major concept under communication2 is that of *written-communication*. In WN, this concept is a subconcept of *communication2*. In our ontology, *written-communication* is also tangled to *physical-thing*. The interpreter was failing to interpret many sentences such as "He burned the prescription/letter..." because "prescription," "letter" were not subconcepts of *physical-thing*.

We have also made *print-media1*, which includes *newspaper* and its subconcepts (a total of 20 concepts), a subconcept of *written-communication*. In WN, *print-media1* is a subconcept of *artifact*. We have also mentioned that *debt-instrument* has become a subconcept of *written-communication* and *possession*.

3.3 Space

The first three senses of "space" in WN have undergone some reorganization. The first sense, space1, has no subconcepts, and has *abstraction6* as its immediate superconcept. *Space2, topological-space1* is mathematical space and has a few mathematical subconcepts. The immediate super-concepts of *space2* are: *set2* (an abstract collection of numbers or symbols) $\Rightarrow$ *abstraction6*. Space3 ("an empty area usually bounded in some way between things") has many subconcepts such as *crack, rip, hole, crevice, fault,...* The superconcepts of *space3* are *amorphous-shape1* $\Rightarrow$ *shape2* $\Rightarrow$ *attribute2* $\Rightarrow$ *abstraction6*. Our reorganization is:

```
space (space1)
 * mathematical-space (space2)
 * empty-area (space3)  -> location.
 * outer-space (space5) -> location
```

The other senses of "space" in WN remain as they are. We have made *mathematical-space (space2))* and *empty-area(space3)* subconcepts of *space (space1)*. More importantly, we have tangled *space3* to *location*, because space3 and its subconcepts are used most times as *location*. Note that *location* is a *physical-thing*, and we need a *physical-thing* as the selectional restriction of *change-of-location* and *cause-to-change-location* predicates. In fact, if *space3* were just a subconcept of *abstraction*, the interpreter would not be able to assign meaning to the PPs ("in a crevice," "in the space," "into the space") in the sentences: "The fish frequently hides in a crevice," "Pleural effusion is an accumulation of excessive amounts of liquid in the space between the two parts of the pleural membrane," "Peridural anesthesia

is caused by injecting the anesthetic into the space just outside the covering of the spinal cord."

In WN, space5 (outer-space) is a subconcept of *location* while in our ontology is also a subconcept of *space*. Basically, our representation is capturing the duality of the concept *space* as an *abstraction* and as a *location*. Most times, however, "space," is used as a *location* in ordinary language, e.g., "Some neutron stars, called pulsars, give off beams of radiation into space."

4 Conclusions

We have explained some reorganizations and changes to the WN noun ontology. These changes have been pointed out by a semantic interpretation algorithm which is based on predicates linked to the WN noun ontology. Space limitations have prevented us from discussing other important concepts in the WN upper-ontology (See CS-TR-01-01 with the same title.). These changes are very much within the principles that have been guiding Wordnet, and can be easily integrated into the Wordnet ontology. As our testing of the predicates continues, we expect to make additional changes although we do not think that they will be major ones.

References

1. Bateman, J. A. and Kasper, R. T. and Moore, J. D. and Whitney, R. A.: A General Organization of Knowledge for Natural Language Processing: the PENMAN upper model. (1990).
2. Fellbaum C.: English Verbs as a Semantic Net. In *WordNet: An electronic Lexical Database and some of its applications.* Fellbaum, C. (ed.), MIT Press, (1998).
3. Gomez, F: An Algorithm for Aspects of Semantic Interpretation Using an Enhanced WordNet. In the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics NAACL-2001. pp. 87–94 (2001).
4. Buitelaar, P.: CoreLex: systematic polysemy and underspecification. PhD thesis. Dept. of Computer Science, Brandeis University (1998).
5. Mahesh, K. and Niremburg, S.: A situated Ontology for Practical NLP. IJCAI Workshop on basic ontological issues in knowledge sharing. Montreal (1995).
6. Miller, G.A: Nouns in WordNet. In *WordNet: An electronic Lexical Database and some of its applications.* Fellbaum, C. (ed.), MIT Press, (1998).

Using a Lemmatizer to Support the Development and Validation of the Greek WordNet

Harry Kornilakis[1], Maria Grigoriadou[1], Eleni Galiotou[1,2], and Evangelos Papakitsos[1]

[1] Department of Informatics and Telecommunications, University of Athens, Panepistimiopolis, GR-157 84, Athens, Greece
Email: harryk@di.uoa.gr, gregor@di.uoa.gr, egali@di.uoa.gr, papakitsev@vip.gr
[2] Department of Informatics, Technological Educational Institute of Athens, Athens, Greece

Abstract. In this paper we aim to give a description of the computational tools that have been designed and implemented to support the development and validation process of the Greek WordNet, which is currently being developed in the framework of the BalkaNet project. In particular, we focus on the description of a lemmatizer for the Greek language, which has been used as the basis for a number of tools supporting the linguists in their work of developing and validating the Greek WordNet.

1 Introduction

The software infrastructure needed in view of building the Greek WordNet was developed during two consecutive projects. The DiaLexico project [3] which aimed at the construction of a lexical database with semantic relations for the Greek language and the BalkaNet project [9], which aims at the development of a multilingual lexical database with semantic relations for each of the following languages: Bulgarian, Czech, Greek, Romanian, Serbian and Turkish. The deployment of computational tools has been proved to be of major importance in the course of the aforementioned projects. The tools and resources used for the development of the monolingual Greek WordNet had to take into account the peculiarities of the Greek language, which is considered as a lesser-studied one.

In this paper we focus on the description of a lemmatizer, which has been used as the basis of a number of tools supporting the linguists in their work of extracting and processing the necessary linguistic information from dictionaries and corpora. Up to now, lemmatizers have been developed for the Greek language, mainly as tools to support specific applications, or as part of systems that support full morphological processing and require a large number of lexical resources. Examples of such systems are [5] and [8] which utilize the two-level morphology model [7] which uses a morpheme based lexicon, grammatical rules and a finite-state automaton and [6] where a lazy tagging method with functional decomposition is implemented. In our approach the lemmatizer was designed so as to be useful for a number of different tools, to require as few lexical resources as possible and to be computationally efficient.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 130–135.

2 Aspects of Greek Inflectional Morphology

Since Greek is a lesser-studied language and without the wealth of resources available for other languages, in the development of tools for the monolingual Greek WordNet we had to take into account the peculiarities of the Greek language. In this section a very brief presentation of the morphology and inflection of the Greek language that is necessary for the understanding of the rest of the paper, is given. For a more detailed description of the Greek language the reader is referred to a grammar of the Modern Greek language such as [4].

The Greek alphabet consist of 24 letters, 17 consonants (β, γ, δ, ζ, θ, κ, λ, μ, ν, ξ, π, ρ, σ, τ, ϕ, χ, ψ) and 7 vowels which may appear either unstressed (α, ε, η, ι, o, υ, ω) or stressed ($\acute{\alpha}$, $\acute{\varepsilon}$, $\acute{\iota}$, $\acute{\eta}$, $\acute{o}$, $\acute{\upsilon}$, $\acute{\omega}$). Each word of two or more syllables has a stressed syllable that is pronounced the loudest, and in written script it is denoted by a stress mark (') over the nuclear vowel of the syllable. Each word may carry only one stress mark and according to a phonologic rule the stress may fall only upon the ultimate, penultimate or antepenultimate syllable. Word stress in Greek is *distinguishing* (e.g. $\nu\acute{o}\mu o\varsigma$ ('nomos – law) is different from $\nu o\mu\acute{o}\varsigma$ (no'mos – administrative region). Furthermore, word stress is *moving* i.e. the stress may change its position within the inflectional paradigm of the same word. For example, the word $\theta\acute{\alpha}\lambda\lambda\alpha\sigma\alpha$ ('θalasa – sea) in the genitive plural case becomes $\theta\alpha\lambda\lambda\alpha\sigma\acute{\omega}\nu$ (θala'son – of the seas).

Articles, nouns, adjectives, pronouns, verbs and participles are declinable. Nouns decline for number (singular, plural) and case (nominative, genitive, accusative, vocative), adjective decline for number, case, gender (male, female and neuter) and degree, while verbs conjugate for voice (active, passive), mood (indicative, subjunctive, imperative), tense (past, non-past), aspect (momentary, continuous), number (singular, plural) and person (1^{st}, 2^{nd}, 3^{rd}) leading up to almost sixty different forms for each verb. From the above, it is easy to see that Greek is highly inflected and having to deal with each inflectional type of a word separately, would be an unnecessary burden to a linguist developing the Greek WordNet, since the citation form of each word is all that is required. Therefore, we have developed a lemmatizer for the Greek language, which can find the citation form of inflected Greek words.

3 A Lemmatizer for the Greek Language

The function of the lemmatizer is, when given as input a word in Greek, to analyze the word and to find its dictionary citation form. The lemmatizer can deal with the inflection of nouns, adjectives and verbs that do not alter their stem (which includes all derived verbs and verbs of the 2^{nd} conjugation [4]) and can also deal with cases of irregular inflection. Furthermore it can handle stress movement. In order to achieve these, the lemmatizer keeps an amount of lexical information, which is kept in three lists: a list of words, a list of inflectional information and a list of irregular forms.

- List of words: A list containing the citation form of all the words in a dictionary.
- List of inflectional information: A list containing information about how words are inflected in Greek. Each entry in the list is of the form [*inflected_ending, citation_ending1, stress_movement1, citation_ending2, stress_movement2... citation_endingN, stress_movementN*] where each *stress_movement* is a possible ending of

the citation form of an inflected word ending in *inflected_ending*. Each *stress_movement* is a number that defi nes how the stress of the word moves when going from the inflected form to the citation form. Each *stress_movement* takes values between -2 and 2 that represent the following:

-2: the stress moves two syllables to the left;
-1: the stress moves one syllable to the left;
0: no stress movement;
1: the stress moves one syllable to the right;
2: the stress moves two syllables to the right.

– List of irregular forms: A list of pairs in the form [*irregular_inflected_form, citation_form]*, one pair for each irregular inflected form in the language. e.g. [*εἶδα, βλέπω*] where *εἶδα* ('iδa) is an irregular form (past tense, 1st singular, indicative, active voice) of the verb *βλέπω* ('vlepo) (see).

The algorithm for lemmatizing the input word is as follows:

```
1. Search for the input word in the wordlist
If it is found
    Return the word and exit.
else
    Go to step 2
2. Search for the input word in the list of irregulars
If a pair [inflected_form, citation_form] is found
    Return citation_form and exit.
else
    Go to step 3
3. Search in the list of inflectional endings for the ending of
the input word. Find the longest possible ending that matches the
word.
If a list [inflected_ending, citation_ending1,
citation_ending2,...] is found
    Go to step 4
else
    The input word could not be lemmatized so return the input
    word and exit.
4. For each citation_ending in [citation_ending1,
citation_ending2...] do
    Remove inflected_ending from the input word
    Append citation_ending to the word
    Make the appropriate adjustment to the position of the stress
    mark on the word (See description of list of inflections above).
    Search for the new word in the wordlist.
    If it is found
        Return the word and exit.
    else
        Continue with the next citation_ending
```

```
5. If no word was found in step 4
   The input word could not be lemmatized so return the input
   word and exit.
```

4 Tools That Use the Lemmatizer

The lemmatizer has been used for three different tools whose purpose is to support the linguistic team in the development of the Greek WordNet. These tools are: A tool that counts the frequency of lemmatized word forms in text corpora, a tool that given a Greek word finds the English translation of that word and a part of speech tagger used in the annotation of corpora.

4.1 Lemmatized Word-frequency Counter

Calculating the frequency of appearance of words in corpora is useful in determining some of the base concepts. For this purpose the ECI corpus has been used. ECI is a medium-sized corpus (around 2 million words) of Modern Greek, compiled by the Universities of Edinburgh and Geneva as part of the European Corpus Initiative Multilingual Corpus. When determining base concepts it is often useful to be aware of the frequency of words in corpora, so as to avoid using as base concepts words which might be frequent in English but infrequent in Greek due to different lexical patterns between English and Greek.

The computational tool that was developed is a tool that counts the occurrences of words in corpora, in all their inflected forms. Given a number of texts in Greek the tool creates a list giving the frequency of total occurrences of each word in the texts, regardless of the inflection type in which this word appears.

In Table 1 we present an example of the results given by the word-frequency counter considering the appearances of the word $\acute{\alpha}\nu\theta\rho\omega\pi o\varsigma$ ('anθropos – man) in the ECI corpus. The frequency of each inflectional type is given separately, and in the bottom row the total occurrences of the word are given.

4.2 Translator of Words from Greek to English

The function of the word translator tool is, given a Greek word, to find the English translation of that word. The lemmatizer is a necessary component of this tool because Greek is a highly inflected language and different inflected forms of the same word may correspond to only one word form in a language with a limited inflectional system, such as English. When given a word as input this tool initially runs the lemmatizer on that word, so as to find the citation form of this word and then by looking up that word in a bilingual Greek to English dictionary we find the English translation of that word.

In the framework of WordNet development the translation is used to find the correspondence of words appearing in Greek corpora to their Inter-Lingual-Index (ILI) numbers [10]. The ILI is an unstructured list of Princeton WordNet 1.5 & 1.7 [2] synsets, with each synset in a monolingual WordNet having at least one equivalence relation with a record in this ILI. Since in the Princeton WordNet the literals of the synsets are in English, translating a Greek word to English will easily allow one to find the corresponding ILI numbers of that word.

Inflectional type	Word	Frequency
Nominative Singular	άνθρωπος	749
Genitive Singular	ανθρώπου	474
Accusative Singular	άνθρωπο	419
Vocative Singular	άνθρωπε	1
Nominative Plural	άνθρωποι	430
Genitive Plural	ανθρώπων	163
Accusative Plural	ανθρώπους	219
Total Occurrences		2455

Table 1. The count for the various inflected forms of the word "άνθρωπος"

4.3 Part of Speech Tagger

Given the lemmatizer and some information about the part of speech of words extracted from a dictionary of the Greek language, it was easy to extend the lemmatizer into a part of speech tagger for Greek texts. The wordlist was extended with part of speech information for each word, i.e. each entry in the list took the form [*word, part-of-speech1, part-of-speech2...*] allowing for each word to belong to multiple parts of speech. Therefore, once the lemmatization of a word into its citation form has been performed, we can assign a part of speech to the input word.

The extraction of the part of speech of each word was performed using the Triantafyllidis electronic dictionary of the Greek language as input and the tools developed by Galiotou et al. for the extraction of linguistic information from the definitions of electronic dictionaries [3].

This part of speech tagger is used for annotation of a Greek language corpus that is to be used as a resource for the validation of the Greek WordNet in the framework of the BalkaNet project. In particular the Greek text of George Orwell's 1984 is being annotated so as to be used for producing comparative coverage statistics for the WordNets developed as part of the project. For the rest of the languages participating in the project (except Turkish) an aligned and annotated version has already been developed as part of the Multext-East project [1], and an aligned and annotated version of the Greek text is required for acquiring reliable statistics.

5 Conclusions

In this paper, we dealt with the computational infrastructure which was developed for supporting the work of the linguists in building the Greek WordNet. In particular, we focused on the description of a lemmatizer which was used in a number of computational tools for extracting and processing linguistic information. We argued that a lemmatizer is indispensable to the processing of a highly inflected language like Greek and we described the use of the lemmatizer by other tools such a part-of-speech tagger, a word-frequency counter in corpora and a tool used for the retrieval of English translations of Greek inflected forms in a bilingual dictionary. Future work concerns the development of new tools and the enhancement of existing ones for the processing of morphosemantic information in dictionaries and corpora taking into account the particularities of the Greek language.

References

1. Erjavec, T., Ide, N., Petkevic, V., Veronis, J.: Multext-East: Multilingual Text, Tools and Corpora for Central and Eastern European Languages. Corpora Proceedings of the First TELRI European Seminar (1996) 87–98.
2. Fellbaum C. (ed.) WordNet: An Electronic Lexical Database. MIT Press (1998).
3. Galiotou E., G. Giannoulopoulou, M. Grigoriadou, A. Ralli, C. Brewster, A. Arhakis, E. Papakitsos, A. Pantelidou: Semantic Tests and Supporting Tools for the Greek WordNet, Proceedings of the NAACL Workshop on WordNet and Other Applications, Carnegie Mellon, Pittsburgh, PA, (2001) 183–185.
4. Mackridge P.: The Modern Greek Language. Oxford University Press (1985).
5. Markopoulos G.: A Two-Level Description of the Greek Noun Morphology with a Unification-Based Word Grammar. In Ralli A., Grigoriadou M., Philokyprou G., Christodoulakis D., Galiotou E. (eds.): Working Papers in NLP, Diaulos, Athens (1997).
6. Papakitsos E., Grigoriadou M., Ralli A.: Lazy Tagging with Functional Decomposition And Matrix Lexica: An Implementation in Modern Greek. Literary and Linguistic Computing, 13(4) (1998) 187–194.
7. Ralli A., Galiotou E.: Affixation in Modern Greek: a Computational Treatment. Proceedings of EURISCON '91 (1991).
8. Sgarbas K., Fakotakis N., Kokkinakis G.: A PC-KIMMO Based Morphological Description of Modern Greek.. Literary and Linguistic Computing, 10 (1995) 189–201.
9. Stamou S., Oflazer K., Pala K., Christoudoulakis D., Cristea D., Tufis D., Koeva S., Totkov G., Dutoit D., Grigoriadou M.: BalkaNet: A Multilingual Semantic Network for Balkan Languages. Proceedings of the First International WordNet Conference, Mysore, India (2002).
10. Vossen P. (ed.): EuroWordNet: A Multilingual Database with lexical Semantic Networks. Kluwer Academic Publishers (1998).

VisDic – Wordnet Browsing and Editing Tool

Aleš Horák and Pavel Smrž

Faculty of Informatics, Masaryk University Brno
Botanická 68a, 602 00 Brno, Czech Republic
Email: hales@fi.muni.cz, smrz@fi.muni.cz

Abstract. This paper deals with wordnet development tools. It presents a designed and developed system for lexical database editing, which is currently employed in many national wordnet building projects. We discuss basic features of the tool as well as more elaborate functions that facilitate linguistic work in multilingual environment.

1 Introduction

Princeton WordNet became one of the most popular language resources. It is currently used in many areas of natural language processing such as information retrieval, automatic summarization, document categorization, question answering, machine translation etc. To integrate into the applications, many researchers work with the Princeton database and transform data to their own proprietary formats.

The Princeton team also developed a data browser for WordNet which can be downloaded [1] together with English data both for Windows and UNIX platform. No WordNet editing tools are provided as the only instruments for majority of the lexicographic work in Princeton are standard text editors. The consistency of data is not therefore checked during the editing process itself, it is postponed to later phases.

Year by year the number of Princeton WordNet clones and WordNet-inspired initiatives increased. In 1998–1999 the EU project EuroWordNet 1 and 2 [2] took place, in which multilingual approach has dominated and WordNets for 8 European languages, particularly for English, Dutch, Italian, Spanish, French, German, Czech and Estonian, have been developed. The Interlingual Index (ILI), Top Ontology, set of Base Concepts and set of Internal Language Relations have been introduced as well [3]. These changes also led to the design and development of the new database engine for EuroWordNet and it resulted in the editing and browsing tool called Polaris [4].

In 2001 the EU project Balkanet [5] has been launched which can be viewed as a continuation of EuroWordNet project. It has been conceived as a multilingual as well and within its framework WordNets for 6 languages are being presently developed, particularly for Greek, Turkish, Romanian, Bulgarian, Serbian and Czech. Before Balkanet has started it had already been obvious that Polaris tool had no future because its development had been closed and as a licensed software product (by Lernout and Hauspie) it had been rather expensive for most of the research institutions involved (typically universities). Moreover, the system had been provided only for MS Windows platform.

As the developers of Czech WordNet within EuroWordNet 2 project we came to the conclusion that a new tool for WordNet browsing and editing has to be developed rather

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 136–141.

quickly. At the same time we realized that it was necessary to look for the solution that would also support establishing the necessary standards for WordNet like lexical (knowledge) databases. Thus we decided to develop a new tool for WordNets based on XML data format, which can be used for lexical databases of various sorts. The tool is called VisDic and it has been implemented recently in Natural Language Processing Laboratory at Faculty of Informatics, Masaryk University for both Windows and Linux platform.

2 Basic Functionality

VisDic was developed as a tool for presentation and editing (primarily WordNet-like) dictionary databases stored in XML format. Most of the program behaviour and the dictionary design can be confi gured. With these capabilities, we can adopt VisDic to various dictionary types—monolingual, translational, thesaurus or generally linked wordnet lexicons.

2.1 Multiple Views of Multiple Wordnets

The main working window is divided into several dictionary panels. Each panel represents a place for entering queries and browsing context of one specifi ed wordnet dictionary. The panels can display different wordnets as well as multiple contexts of the same dictionary.

The contents of a panel offers, besides the query input and matching results list, a set of overlapping notebooks tabs each of which represents one kind of view of the same entry from the list of results. The order, the type and even the content of each notebook tab is specifi ed by the user in the confi guration fi les (see3.6). The main types of views are described in the following sections.

2.2 Freely Defi ned Text Views

The content of the Text View notebook tab is entirely built from the user defi nition that follows the XML structure of the wordnet entry. The editor can thus present an easily readable view of the entry with highlighting important parts of the entry content (see the Figure 1).

POS: n ID: ENG171-12836307-n
Synonyms: sunset:1, sundown:1
Definition: the time in the evening at which the sun begins to fall below the horizon
-->> [hypernym] *[n] hour:2, time of day:1
-->> [holo_part] *[n] evening:1, eve:4, eventide:1
-->> [near_antonym] [n] dawn:1, dawning:1, morning:3, aurora:1, first light:1, daybreak:1, break of day:1, break of the day:1, dayspring:1, sunrise:1, sunup:1, cockcrow:1
<<-- [near_antonym] [n] dawn:1, dawning:1, morning:3, aurora:1, first light:1, daybreak:1, break of day:1, break of the day:1, dayspring:1, sunrise:1, sunup:1, cockcrow:1

Fig. 1. An example of freely defi ned text view of wordnet entry

2.3 Edit

The editing capabilities allow to give the user a full control over the content and linking of each entry in the wordnet hierarchy. To prevent the user from moving the entry as an object in the spider web of the linkage relations, the linguist rather specifies all the links in a textual dialog, where all the bindings are displayed in one place with consistency checks after each change request.

The actual contents of the Edit notebook tab is also entirely driven by the user instructions in the configuration, where each editing field is named and assigned to an XML tag in the entry.

2.4 Tree and RevTree

The wordnet dictionaries are specific by a heavy network of various kinds of relations between the dictionary entries with the function to capture the ontology relations on the underlying natural language.

The navigation in such environment is thus a crucial point of a successful linguistic work with wordnet data. Since the linkage relations generally do not need to obey any rules, that could make the resulting structure to be an arbitrary directed acyclic graph, or DAG. VisDic implements a browsing mechanism for general graphs. The navigation process works with two interconnected notebook tabs, which always both start at the same dictionary entry and display its position in the graph represented as a breadth-first path trees of all the linkage relations that lead from the entry to other entries in the dictionary. Each of the notebook tabs displays mutually opposite linkage relations, allowing the user to choose the direction of graph navigation in every step.

To facilitate the orientation and to help to position the entry in the wordnet hierarchy, the navigation also displays the path from the entry to its top in the hyper-hyponymical relation tree (see the Figure 2). For more advanced navigation the linguist may also use advanced tree browsing techniques (described in 3.3).

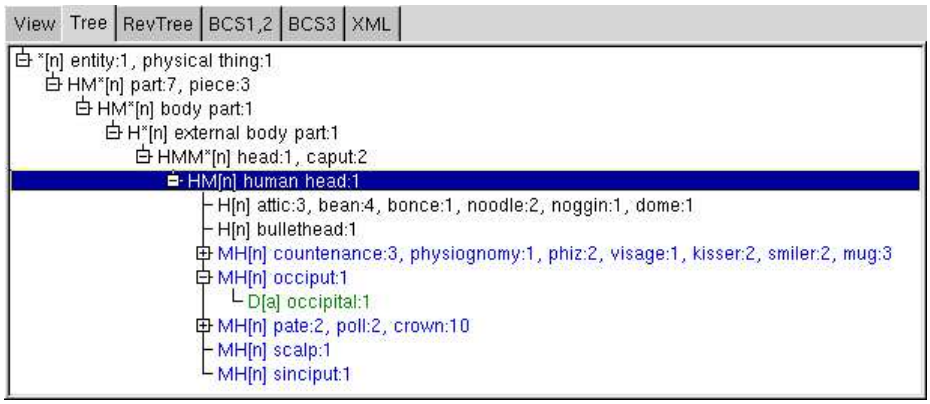


Fig. 2. The tree-like navigation in the wordnet linkage relations graph

2.5 Query Result and External File Lists

Common actions in the wordnet creation and editing often include processing of a subset of entries based on certain criteria. VisDic offers a suitable kind of views for this situation, which allow to prepare a notebook tab with a list of entries matching any user specified query or a list of entries identified by entry-IDs gathered in a plain text file.

2.6 Plain XML View

Sometimes users need a thorough view into the data contained in the dictionary entry. XML View notebook tab offers this possibility. In this view, the user can see a graphically structured XML text, which represents the entry structure as it is stored in the dictionary.

3 Advanced Functionality

The basic functionality described in the previous section generally conforms to any XML based dictionary. However, linguistic work specialized to wordnet creation and editing requires some more specific and more sophisticated functions in the editor.

3.1 Synchronization

Within the creation of a national (e.g. Czech) wordnet, which would correspond to the English wordnet as a primary reference, one of the most frequent operation is a lookup of a dictionary entry (synset) from one wordnet in another dictionary. Such lookup uses either the SYNSET.ID tag (as a direct equivalent) or one of the, so called, equivalence tags (or attributes) defined in the configuration. An example of such tag may be REVMAP or MAPHINT used to help the linguist to process ambiguous link references between various versions of English wordnet.

The lookup function in VisDic can work in two modes: as an instant (one time) lookup — the *Show (by)* operation, and also as a firmly established link between two notebook tabs called the *AutoLookUp (by)*. In case of AutoLookUp, any move to another dictionary entry in the source notebook tab leads to an automatic lookup of the new entry in the destination tab. VisDic allows to have any acceptable combination of autolookups among all the notebook tabs.

3.2 Editing Support

The efforts of unifying national wordnets based on the English wordnet in many cases lead to copying of synset information between different language dictionaries. Such functionality in VisDic is splitted into two common situation — either the SYNSET.ID of an existing synset is to be unified with the ID of the English synset (*Take key from* operation) or a whole new entry is to be copied to another dictionary (*Copy entry to*).

3.3 Tree Browsing

The basic navigation in related synsets (in some cases reduced to the hyper- and hyponymical relations tree) is supplemented with two important wordnet operations — *Topmost entries* and *Full expansion*.

The Topmost entries operation identifies all synsets, which are (in the tree subset of linkage relations) found as the roots of relational hierarchy, i.e. are not hung below some other synset. This helps the linguist to identify the level 1 entries as well as so far unfiled entries.

The Full expansion allows the user to see all possible descendants of a selected synset in the linkage relations graph. During the operation cycle detection techniques check the violations of tree properties in the graph. Some relations can be also configured to be left out from the full expansion process.

3.4 Consistency Checks

Semi-automatic processing, which often takes part in the national wordnets creation, as well as common human processing of the data inevitably brings in the possibility of mistakes. The inconsistencies, which may be revealed as a duplicity, are controlled by VisDic consistency checks, which contain

- check duplicate IDs;
- check duplicate literals and senses;
- check duplicate synset literals;
- check duplicate synset links.

These checks allow the linguist to identify the most common errors e.g. after merging data from various sources.

3.5 Journaling

The work on a large and representative national wordnet usually employs more than one linguist working on the data. The synchronization of the resulting dictionary is made possible in VisDic with the usage of *journaling*.

During the work with VisDic, any changed to the data is marked in a journal file. Each journal file is specific to one dictionary and one user at a time. Such journal file can then be "applied" to the dictionary data and merged with the original. In this way, the simultaneous work of several linguists can be easily interchanged with a common data source.

3.6 XML Configuration

Most of the functionality in the VisDic wordnet editor can be adopted to the local needs by means of its configuration files. All settings for the VisDic application are stored in several XML files.

The main configuration file (`visdic.cfg`) serves for global application data storage such as the list of dictionaries, the list of views, fonts, colors, histories, etc.

Besides this, each wordnet dictionary has its special configuration file (*dictionary*`.cfg`), which enables the linguist to set up most of the texts displayed in the application as well as the content of notebook tabs specific to the particular dictionary with respect to the XML structure of the entries.

4 Conclusions and Future Directions

VisDic, during its rather short history, has already proved its suitability for lexical database creation. The main power of VisDic manifests itself especially in development of highly interlinked databases such as wordnet. Its unique features have assured VisDic the leading role in many wordnet editing projects.

The development of such tool is never really closed. The future directions of our work will concentrate at specific support for linguists, improvements in the customization and user interface and team cooperation functionality. Entirely new horizons appear in the ongoing development of VisDic successor, the client-server lexical database editor DEB [6].

Acknowledgements

This work was supported by Ministry of Education of the Czech Republic Research Intent CEZ:J07/98:143300003 and by EU IST-2000-29388.

References

1. WordNet Project Website, `http://www.cogsci.princeton.edu/~wn/`.
2. Eurowordnet Project Website, `http://www.illc.uva.nl/EuroWordNet/`.
3. Vossen, P., ed.: EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Kluwer Academic Publishers, Dordrecht (1998).
4. Louw, M.: Polaris User's Guide. Technical report, Lernout & Hauspie – Antwerp, Belgium (1998).
5. Balkanet project website, `http://www.ceid.upatras.gr/Balkanet/`.
6. Smrž, P., Povolný, M.: DEB – Dictionary Editing and Browsing. In: Proceedings of the EACL03 Workshop on Language Technology and the Semantic Web: The 3rd Workshop on NLP and XML (NLPXML-2003), Budapest, Hungary (2003), 49–55.

A Corpus Based Approach to Near Synonymy of German Multi-Word Expressions*

Christiane Hümmer

Berlin-Brandenburgische Akademie der Wissenschaften, Jägerstr. 22/23,
10117 Berlin, Germany
Email: huemmer@bbaw.de

Abstract. The core of this paper is a detailed corpus-based analysis of the two nearly synonymous German idioms etw. liegt jmdm. im Blut and etw. ist jmdm. in die Wiege gelegt. The central conclusions drawn from this analysis are: On the basis of the behaviour of the semantic arguments of the two idioms – their presence or absence as well as certain semantic properties – clear statements can be made about the context conditions under which the two idioms are interchangeable and those allowing the realisation of one of them while excluding the other one. Furthermore, it is stated that even in the contexts that allow both idioms, the choice of one or the other makes a subtle difference. This difference has to do with the metaphorical image encoded in the idiom. The prominent degree of prototypicality of certain traits demonstrates that speakers actively use these subtle differences. The paper constitutes thus an investigation on the level below WordNet synsets discussing the concept of synonymy underlying WordNet organisation.

1 Introduction

In this paper, the results of a corpus-based analysis of the following two nearly synonymous idioms are reported:

etw. ist jmdm. in die Wiege gelegt	*etw. liegt jmdm. im Blut*
'sth. was put in sb.'s cradle'	'sth. lies in sb.'s blood'

They were taken from a large collection of semantically closely related multi-word expressions (MWE), collected from an onomasiologically sorted [1] and a synonym dictionary of German [2]. The collection of MWEs serves as the data base for a larger (PhD-) project about synonymy of MWEs in German. This paper exemplifi es the methodology used and the results that can be expected in this project.

* I would like to thank Patrick Hanks for very patiently helping me work out and edit this paper as well as Kerstin Krell and Ekaterini Stathi for comments on earlier versions. I am also greatly indepted to Christiane Fellbaum for constantly accompanying my work with her advice and for always being open for discussions on any linguistic subject. Work on my PhD dissertation has been made possible by the Wolfgang-Paul-Award of the Alexander-von-Humboldt-foundation, imparted to Christiane Fellbaum. Thanks also to my colleagues Alexander Geyken, Alexej Sokirko und Gerald Neumann, who facilitate the access to the DWDS corpus and provide the computational tools for corpus search and to Ralf Wolz for editing advice.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 142–149.

The PhD dissertation is part of the project "Kollokationen im Wörterbuch" ('Collocations in the Dictionary') at the Berlin-Brandenburg Academy of Sciences. This project aims at investigating syntactic, semantic and morphological properties of German idioms on the basis of the 980 M word DWDS corpus, which was compiled from texts representing a wide variety of genres and covering the entire 20th century [3]. The same corpus was used as a source of empirical evidence for the investigation presented here. For each of the MWEs examined, a subcorpus containing all the occurrences of that expression in the corpus has been extracted using the linguistic query tools developed by members of the project "Kollokationen im Wörterbuch". All the conclusions drawn and all the quantitative statements made in this paper are based on a manual analysis of the whole subcorpora.

Since idioms are not generally found in WordNet and constitute a number of problems for codification (see Fellbaum 1998 [4] and Fellbaum 2002 [5]), the two idioms examined are not WordNet entries. Nevertheless, they are semantically close enough to be assumed to be candidates for membership in one common synset once they can be enconded. The lexicological case study presented in this paper is therefore a suitable example for investigating the concept of synonymy WordNet makes use of. As Miller et al. [6] point out, the WordNet organisation principle of synonymy is based on the idea of substitutability without change of truth values. Since the authors doubt the existence of absolute synonyms, substitutability in some contexts is assumed as a sufficient prerequisite for making two lexical units become members of the same synset. This paper shows the results of basing this intuition on a corpus-based research.

2 Results

Conclusions that have been drawn from the corpus data focus on two main questions:

1. What are the context conditions that make the two idioms converge or diverge semantically?
2. Assuming that even interchangeable expressions are not absolutely synonymous, what governs the choice between the two idioms in contexts where they are interchangeable? And how can this be identified in the corpus data?

Concerning the first question, it can be said that the conditions of semantic convergence and divergence can be formulated quite clearly in terms of the behaviour of the semantic arguments of the idioms.

As for the external valency of the idiom *etw. ist jmdm. in die Wiege gelegt*, its maximal realisation is achieved when the idiom is realised in the active voice. Although this is not its prototypical[1] syntactic form, it occurs in a considerable amount of cases (150 out of 609, see below), as corpus evidence proves. This maximal realisation can be classified as an instance of the semantic frame

DONOR – RECIPIENT – THEME[2],

[1] The notion of prototypicality of meaning and form, very important for the present investigation, is also found in many previous publications, e.g. Hanks 1994 [7] and Hanks 1997 [8].

[2] The names of the semantic roles are taken from the FrameNet specifications of Frame Elements ('Giving Frame') [9]

syntactically realised as subject – indirect object – direct object.
The idiom *etw. liegt jmdm. im Blut* takes arguments that may be described as

PROPERTY and PROTAGONIST[3],

syntactically realised as subject and possessive dative. A SOURCE is also often mentioned but not as part of the argument structure of the idiom.

In order to gain an overview of corpus evidence, for each idiom a table was constructed with columns as slots for the arguments and the rows containing the particular lexical items filling the semantic argument positions in the corpus (see Table1 for a very small part of this table).

Table 1. Extract from the table containig realisations of semantic arguments of *etw. liegt jmdm. im Blut*

	PROTAGONIST	PROPERTY	SOURCE
1.	Jason Gebert (poss. dat.)	eine Freude an schönen Farben … (Subj) (‘taking pleasure in beautiful colours…’)	vom Vater her (Adjunct/VP) (‘from his father’)
2.	ihr (poss. dat.) (‘to her’)	Die Fliegerei (Subj) (‘flying’)	Vater war Flugkapitän … (context) (‘father was an aircraft captain’)
3.	denen (poss. dat.) (‘to them’)	das Bedürfnis nach Bewegungsfreiheit… (Subj) (‘a want for freedom of movement’)	als Britensprösslingen (Adjunct/Dat) (‘as offsprings of British’)
4.	den Clintons (poss. dat.) (‘to the Clintons’)	Wahlkämpfe (Subj) (‘election campaigns’)	[den Clintons (Dat]
5.	den Katholiken (poss. dat.) (‘to the catholic’)	das Lügen (Subj) (‘lying’)	[den Katholiken (Dat)]
6.	euch Bienen (poss. dat.) (‘to you bees’)	Fliegen und immer fliegen (Subj) (‘flying and always flying’)	[euch Bienen (Dat)]
7.	ihm (poss. dat.) (‘to him’)	Das unsittliche Leben (Subj) (‘the immoral life’)	
8.	[deutschen (Adj modifying Blut)] (‘German’)	Die Angriffslust (Subj) (‘aggressiveness’)	deutschen (Adj modifying Blut)

These tables show that basic semantic frames are frequently modified considerably in actual use:

In the context of the expression *jmdm. liegt etw. im Blut*, very frequently expressions can be found that encode the SOURCE of the PROPERTY attributed to the PROTAGONIST. This SOURCE is, in most cases, the family or a group (often genetically specified) that the

[3] FrameNet roles from the ‘Mental_property’ frame [9]

PROTAGONIST belongs to. It appears as merged with the entity denoting the PROTAGONIST (rows 4,5,6,8), or as an independent expression in the closer (rows 1,3) or wider context (row 2).

This SOURCE argument is very similar to the DONOR semantic role of *etw. ist jmdm. in die Wiege gelegt*. The difference between the two lies in the emphasis that can be given to this argument: the DONOR in *etw. ist jmdm. in die Wiege gelegt* can be expressed as an Agent taking the subject position, whereas SOURCE can only appear in less prominent positions.[4] In addition, the fact that SOURCE is often explicitly expressed in the context of *etw. liegt jmdm. im Blut* assigns a role to the PROTAGONIST to whom a PROPERTY is attributed, similar to the RECIPIENT role of the idiom *etw. ist jmdm. in die Wiege gelegt*.

The expression *etw. ist jmdm. in die Wiege gelegt* is realised in 460 cases out of 609 in a combination of passive voice and past tense or in a special German passive form called 'Zustandspassiv'. As a result, in many cases DONOR is not expressed in the context. Instead of an activity, the predicate denotes then a state in which THEME is a PROPERTY of the RECIPIENT. In such cases THEME and PROPERTY are very similar to RECIPIENT and PROTAGONIST in *etw. liegt jmdm. im Blut*.

> from: Frankfurter Rundschau 09.03.2000, S. 12
> Hilfsbereitschaft und der Blick für Missstände und Ungerechtigkeit scheinen der tatkräftigen Frau *in die Wiege gelegt.*
> ('helpfulness and an eye for problems and injustice seem to *have been put in the cradle* of this energetic woman')

The elements that normally fill the argument positions for THEME in *etw. ist jmdm. in die Wiege gelegt* and for PROPERTY in *etw. liegt jmdm. im Blut* to a large extent stem from the same semantic class. As can be seen in the following examples, they are often very similar to each other:

Table 2. Examples of similar lexical items filling the Property argument of *etw. liegt jmdm. im Blut* and the Theme argument of *etw. ist jmdm. in die Wiege gelegt*

PROPERTY argument of *etw. liegt jmdm. im Blut*	THEME argument of *etw. ist jmdm. in die Wiege gelegt*
Das Verkaufen (Faculty) ('selling')	Millionen von Kuscheltieren in alle Welt zu verkaufen (Faculty) ('to sell millions of cuddly toys to the whole world')
die Liebe zu alter Technik (Inclination) ('a love for old technology')	die Liebe zur Musik (Inclination) ('a love for music')
Opposition (Attitude) ('opposition')	Widerstand (Attitude) ('resistance')

[4] With Grimshaw [10] it is assumed that the syntactic function the arguments of a predicate fulfill is determined by an hierarchy of thematic roles and a salience hierarchy of aspectual prominence of arguments. In particular, Grimshaw assumes that the thematically and aspectually most prominent argument is always realised as the subject.

From these observations, general contextual conditions can be formulated under which the two idioms converge semantically.

Etw. ist jmdm. in die Wiege gelegt converges maximally with *etw. liegt jmdm. im Blut* when it is realised in the 'Zustandspassiv', therefore leaving out the DONOR semantic argument and expressing a state instead of an activity. In addition, the THEME argument is filled by a lexical item that can be categorised as belonging to the semantic class of Faculties, Inclinations and Attitudes.

Etw. liegt jmdm. im Blut converges with *etw. ist jmdm. in die Wiege gelegt* when it takes an additional semantic argument similar to the DONOR of *etw. ist jmdm. in die Wiege gelegt.*

In such cases, one expressions can be substituted for the other salva veritate.

The opposite case, where the context only allows one of the idioms, can be described as the complement of what was said above. Basically, *etw. liegt jmdm. im Blut* cannot be substituted for *etw. ist jmdm. in die Wiege gelegt* when *etw. ist jmdm. in die Wiege gelegt* is realised in the active voice as an activity of some Agent that takes the semantic role of a DONOR. Another condition that makes the two idioms diverge is fulfilled when the filler for the THEME position belongs to a semantic class that is not compatible with the PROPERTY position and vice versa. For example, something that has been put in somebody's cradle has to be interpreted as a condition capable of affecting the whole life of that person from the beginning on:

> from: Frankfurter Rundschau 20.10.1997, S. 18
> Der Sohn eines Gummisohlenfabrikanten, dem *weder Geld noch Kunstwerke in die Wiege gelegt worden waren*, hatte sich als junger Mann in österreichischen Revuen als Werber für Schuhcreme verdingt ...
> ('The son of a maker of rubber soles, *into whose cradle neither money nor works of art had been put*, hired himself out, as a young man, to Austrian revues as an advertiser for shoe polish...')

In this context, it is impossible to use the idiom *etw. liegt jmdm. im Blut*.

In other words, presence or absence of certain semantic arguments as well as the semantic and syntactic role they play in the sentence and the semantic class of the lexical items that realise them determines closeness or distance of the two nearly synonymous expressions.

Concerning the question what makes the two idioms different in contexts where they are maximally synonymous, the focus of the investigation was placed on the influence of the metaphoric images and connotations associated with them.

As a starting point, it can be said that *blood* is a much stronger and more drastic image than *cradle*. When *etw. liegt jmdm. im Blut* is used, the speaker usually makes either a very strong statement about the deep-rootedness of a PROPERTY in a PROTAGONIST or it serves to express his (ironic) distancing himself from what he says. This happens above all in cases where a cliché is expressed, which is very frequently the case with *etw. liegt jmdm. im Blut* (see below):

> From: Frankfurter Rundschau (Jahresausgabe 1998)
> Wenn sie nur nicht so aggressiv wären, die Rothäute mit ihren Hakennasen und der Kriegsbemalung. Dabei steht ihnen doch der Federschmuck, den sie stets tragen, so gut. Und wenn sie erst ihre berühmten Tänze aufführen. Großartig. Wir alle wissen

doch: *Negern und Indianern liegt der Rhythmus im Blut*. Das ist einfach angeboren bei den schwarzen Perlen und roten Kriegern. Wie – Sie finden diese Sätze rassistisch, dumm und unerträglich? Wir auch! Wir fragen uns nur, warum immer öfter als Werbe-Gag vielerorts mannshohe Abbilder jener Menschen herumstehen, die die doch so zivilisierten Weißen vor noch nicht langer Zeit versklavt, verfolgt und ermordet haben.
('if only they wouldn't be so aggressive, those redskins with their hooknoses and their war paint. And the feathered headdress they wear fits them so well. And when they perform their famous dances. Grandiose. We all know of course: Rhythm is in the blood of the Negroes and the Indians. It's simply innate in those black pearls and red warriors. What? You find these sentences racist, stupid, and unbearable? So do we! We just ask ourselves why more and more often, as a commercial gimmick, in some places a life-sized picture of those people, who have been enslaved, persecuted and murdered by those remarkably civilised whites not long ago, stands around.')

This assumption is strongly supported by some observations from the corpus.

For example, typical modifications taken by the idioms can give a hint of their characteristic semantic traits. Under this view, the fact that *tief* ('deeply') is a typical modification that appears with the idiom *etw. liegt jmdm. im Blut* (8/326; MI: ∼2.74) highlights the profound rootedness of the PROPERTY in the PROTAGONIST expressed by the idiom. In contrast to this, a typical modification of *etw. ist jmdm. in die Wiege gelegt* is *bereits/schon* ('already') (80/609; MI: ∼1.9), emphasising the early age of the RECIPIENT when receiving THEME.

Another fact related to the meaning of the image is the prototypicality of having a genetic group or an individual representing a genetic group in the PROTAGONIST position of *etw. liegt jmdm. im Blut*. This happens in 78 out of 326 cases, not counting those cases where the PROTAGONIST position is filled by a pronoun whose reference cannot be recovered from the context of one sentence. Some examples are: *die Deutschen* ('Germans') (10 times), *die Schweizer* ('Swiss') (2 times), *Neger* ('Negroes') (2 times) *Indianer* ('Indians'), *Latinos* (2 times), *Juden* ('Jews') (3 times), *Briten* ('British') (2 times) *Südländer* ('southerners') etc.

Altogether, prototypicality effects seem to support the intuitive insight that the image carried by *etw. liegt jmdm. im Blut* favours the interpretation of the PROPERTY argument as something innate while the THEME argument of *etw. ist jmdm. in die Wiege gelegt* tends to be interpreted as something determined by social circumstances or education.

3 Conclusion

The discussion of the corpus-based analysis of two nearly synonymous idioms in this paper shows three main points:

Semantic convergence and divergence of the two idioms is proportional to the behaviour of their semantic arguments. The idiom *etw. ist jmdm. in die Wiege gelegt* was basically analysed as belonging to the frame DONOR (Subject) – RECIPIENT (IO) – THEME (DO), but it converges in its semantic interpretation with *etw. liegt jmdm. im Blut* under the following conditions:

– the DONOR is not present (basically when the idiom is realised in the passive);

- the 'Zustandspassiv' changes the interpretation of the idiom: It encodes a state instead of an activity and the THEME arguments can be seen as a PROPERTY of a PROTAGONIST (otherwise known as the RECIPIENT);
- the lexical material filling the THEME argument position can be interpreted as belonging to the semantic class of faculties, inclinations or attitudes.

It diverges most strongly from *etw. liegt jmdm. im Blut* when

- the subject is present and interpreted as an agent that carries out a giving action or when
- the lexical material in the THEME position is to be interpreted as an starting condition for some individual (the PROTAGONIST) from the beginning of his life.

The idiom *etw. liegt jmdm. im Blut* was basically analysed as belonging to the frame

PROPERTY (subject) – PROTAGONIST (IO).

It converges with and diverges from *etw. ist jmdm. in die Wiege gelegt* mainly with the presence or absence of an additional semantic argument. This argument contains information about the source (or DONOR) of the PROPERTY.

Even in contexts where both idioms should be equally possible the choice of one or the other makes a subtle difference that has to do with the idiomatic image associated with the idiom.

Blut ('blood') is a much stronger image than *Wiege* ('cradle'). In consequence, the use of *etw. ist jmdm. in die Wiege gelegt* is, in most cases, more neutral than *etw. liegt jmdm. im Blut*. When *etw. liegt jmdm. im Blut* is used, it frequently implies either a much more radical statement about the deep-rootedness of a PROPERTY in some PROTAGONIST or, to the other extreme, serves as a way of marking an ironic distancing of the speaker.

The fact that in the corpus the PROTAGONIST can denote a genetic group in both *etw. ist jmdm. in die Wiege gelegt* and *etw. liegt jmdm. im Blut*, but is realised as such very significantly more frequently with *etw. liegt jmdm. im Blut* is only one sign that shows how speakers make use of this distinction.

In summary, from the fact that passive and 'Zustandspassiv' are prototypical syntactic forms for *etw. ist jmdm. in die Wiege gelegt* and that a genetic group in the PROTAGONIST position is prototypical for *etw. liegt jmdm. im Blut* the conclusion can be drawn that the two idioms converge significantly in the language use.

Still, corpus evidence demonstrates that speakers agree on a subtle semantic difference between the two.

With respect to synonymy, the case study at hand supports the claim that the intuitive notion of substitutability should be grounded on corpus evidence. Generalisations over corpus data allow insight on the degree of synonymy in terms of shared or mutually exclusive context conditions as well as about preferred or prototypical contexts for the realisation of the two synonym candidates. Such statements are very important for tasks such as fine-grained lexical choice for Natural Language Generation. Parallel to what Edmonds [11] and Edmonds and Hirst [12] show for synonym words on the basis of dictionary definitions, it can be said that for those tasks a much more fined-grained distinction between lexical units is needed than the one provided by WordNet synsets.

References

1. Hessky, R., Ettinger, S. (eds.): Deutsche Redewendungen. Ein Wörter- und Übungsbuch für Fortgeschrittene. Narr, Tübingen (1997).
2. Schemann, H. (ed.): Synonymenwörterbuch der deutschen Redensarten. Unter Mitarbeit von von Renate Birkenhauer. Straelener manuskripte, Straelen (1989).
3. Cavar, D., Geyken, A., Neumann, G. Digital Dictionary of the 20th Century German Language. In: Erjavec, T., Gros, J. (eds.): Proceedings of the Language Technologies Conference 17–18 october 2000 Ljubljana. On-line proceedings (2000) `http://nl.ijs.si/isjt00/index-en.html`.
4. Fellbaum, C.: Towards a Representation of Idioms in WordNet. In: Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems. University of Montréal, Montréal, Canada (1998) 52–57.
5. Fellbaum, C.: VP Idioms in the Lexicon: Topics for Research using a Very Large Corpus. In: Busemann, S. (ed.): Proceedings of KONVENS 2002 30. september–2. october 2002 Saarbrücken. On-line Proceedings (2002) `http://konvens2002.dfki.de/cd/index.html`.
6. Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K. J.: Introduction to WordNet: An On-Line Lexical Database. Journal of Lexicography 3(4) (1990) 235–244.
7. Hanks, P.: Linguistic Norms and Pragmatic Exploitations. Or, why Lexicographers need Prototype Theory, and Vice Versa. In: Kiefer, F., Kiss, T., Pajzs, J. (eds.): Papers in Computational Lexicography: Complex 1994. Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest (1994) 89–113.
8. Hanks, P.: Lexical Sets: Relevance and Probability. In: Lewandowska-Tomaszczyk, B., Thelen, M. (eds.): Translation and Meaning. Part 4. Euroterm, Maastricht (1997).
9. FrameNet: `http://www.icsi.berkeley.edu/~framenet/`.
10. Grimshaw, J.: Argument Structure. MIT Press, Cambridge, Mass London, England (1990).
11. Edmonds, P.: Semantic Representation of Near-Synonyms for Automatic Lexical Choice. University of Toronto, Toronto (1999).
12. Edmonds, P., Hirst, G.: Near-Synonymy and Lexical Choice. Computational Linguistics 28(2) (2002) 105–144.

Using WordNets in Teaching Virtual Courses of Computational Linguistics

Lothar Lemnitzer and Claudia Kunze

Seminar für Sprachwissenschaft
Universität Tübingen
Wilhelmstr. 19, 72074 Tübingen, Germany
E-Mail: lothar@sfs.uni-tuebingen.de, kunze@sfs.uni-tuebingen.de

Abstract. This paper focuses on wordnets, especially GermaNet, as topics of teaching and learning in the field of Computational Linguistics. We are aiming at two major goals: to use wordnets for the design of tasks in core modules of the Computational Linguistics curriculum on the one hand, and, on the other hand, to enhance the wordnet structure and its accessibility by the different student projects that have been defined and accomplished. These projects, coping with various structural and content-oriented issues of wordnets, have evolved from three virtual courses taught in Tübingen and Osnabrück. They will be presented in this paper. By establishing wordnets as teaching and learning contents, advanced students should be attracted to join the international wordnet research community.

1 Introduction

In this paper, we will outline how lexical semantic wordnets like GermaNet [1] can be useful subjects of teaching and learning in the fi eld of Computational Linguistics. GermaNet currently forms part of three virtual courses within the framework of a national E-Learning project, MiLCA[1]: *Computational Lexicography* and *Applied Computational Linguistics*, held in Tübingen, and *NLP tools for Intelligent Computer Aided Language Learning*, held in Osnabrück. These are virtual courses open to students of different universities in Germany and Switzerland, which shall yield core modules of Computational Linguistics curricula. The students gather in a virtual classroom with shared work spaces, a whiteboard and communication facilities for collaboration on various exercises. In the *Computational Lexicography* course, GermaNet fi gure/table as a prototype of a lexical database. Within *Applied Computational Linguistics*, a course which is centered around a tool providing for intelligent access to dictionaries, the GermaNet data structure constitutes one of the underlying dictionary sources. The *I-CALL* course has developed, among other features, a vocabulary trainer on the basis of GermaNet data.

From our teaching experience, we have learned that students enjoy working with GermaNet. Lexical semantic wordnets seem to be appealing for the clarity and simplicity of

[1] The project on which this paper reports – MiLCA (media intensive learning modules for the practical training of computational linguists) - is being funded by the German Federal Ministery of Education and Research (project ID: 01NM167). The authors of this paper are, however, fully responsible for its content.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 150–156.

their structures, the richness of their contents and the variety of natural language processing tasks in which they may play a role.

The projects which we will describe in the following section focus on various aspects of wordnets:

- their linguistic contents,
- their data structure and presentation,
- tools for accessing and visualizing wordnet structures,
- issues of evaluation and
- wordnets as lexical resources for NLP applications.

This division may also be regarded as a proposal for an appropriate agenda in view of research topics[2].

Some student projects have already been completed, others are still under development and some of them are planned for teaching future courses. The outcome of the accomplished projects turned out to be quite encouraging.

2 The Student Projects in Detail

This section will present the student assignments w.r.t. the linguistic structure of wordnets (2.1), the data structure (2.2) and the development of tools (2.3). In section 2.4, evaluation tasks will be outlined, whereas section 2.5 deals with projects that are using GermaNet as a lexical resource. For each assignment, the title, a short description and the intended outcome of the work will be given.

2.1 Linguistic Aspects

Analysis of the meronymy / holonymy relation and its encoding in GermaNet. While Princeton WordNet encodes three different types of meronymy relations, and EuroWordNet realizes one generic meronymy pointer as well as five subtypes, in GermaNet only a unique pointer covering all instances of meronymy has been realized so far. Concept pairs which are encoded as meronyms should be checked under the following aspects: a) Is the application of three meronymy pointers feasible for GermaNet? b) Will a subdivision into different meronymy pointers yield transitivity for these relations or are there still examples in which transitivity is blocked? c) Are there examples for pairs of concepts where the meronymy relation is not symmetric?

The investigation is based on the subdivision of meronymy from WordNet as well as on the classification proposed by Chaffin [2, p. 274ff]. This project, which is currently under way, aims at refining the meronymy / holonymy relation in GermaNet.

Analysis of the antonymy relation. Similarily to the case of meronymy, GermaNet implements a unique pointer for encoding the antonymy relation between lexical units. Different types of opposites, like ‘man’ vs. ‘woman’, ‘busy’ vs. ‘lazy’, ‘warm’ vs. ‘cold’ or ‘arrive’ vs. ‘leave’ are, thus, subsumed and uniformly treated under the label of antonymy.

[2] The agenda encompasses topics which may have been discussed in the wordnet community rather than entirely new items

The student exercise consists in developing an adequate subclassification of antonymy, dividing the data into appropriate subclasses, which should account for complementary opposites as well as for scalar and gradable opposites. Furthermore, a set of relevant features should be defined, which captures opposites of, e.g., sexus or directionality for nouns and verbs. The empirical analysis of the GermaNet antonyms should consider the categories being proposed in the descriptive approaches of [3,4].

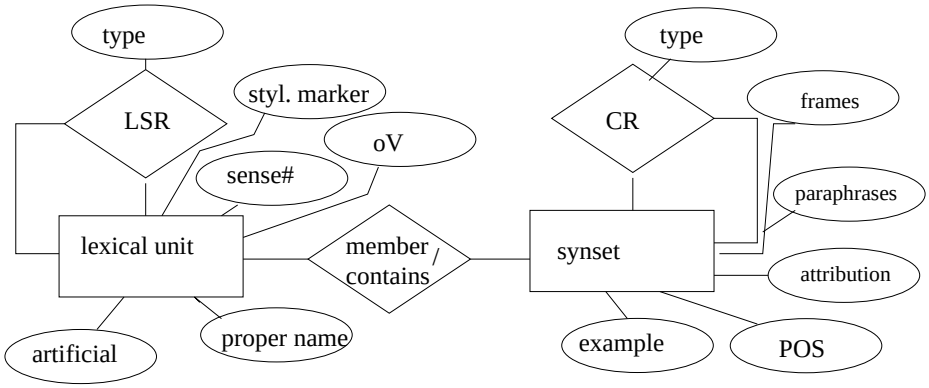
Applicability of regular polysemy in wordnets. Pustejovsky et al. criticize WordNet for ignoring existing regularities between senses [5]. It is, however, still unclear whether wordnets should implement regular sense relations or not, and which should be the appropriate hierarchical level for the application of such rules. The analysis, therefore, concentrates on lexical (sub-)fields which are in the scope of a regular sense extension, e.g. instances of the type 'building-institution-staff' or instances of the type 'tree-wood-fruit'. It will be checked whether generic rules are feasible or not, and, if so, on which hierarchical level they should apply or when blocking of these rules would be necessary.

The interaction of verb concepts, verbal argument structure and Aktionsart / Aspect. Aspectual properties of verbs have recently become a major topic in semantic research. Therefore, it would be interesting to prove or disprove the necessity and feasibility of encoding further information on argument structure and Aktionsart / Aspect in the existing hierarchy of verbal concepts. Some preliminary investigation reveals that aspectual hierarchies cannot be assumed straightforwardly; otherwise, the representation of causative and inchoative variants of a verb within one synset has to be abandoned. A closer examination shall focus on a specific verb field, e.g. verbs of locomotion, and test the possibility of creating an inheritance hierarchy (with overwriting) for aspectual features in interaction with argument structure. This project will be assigned in the near future.

2.2 The Data Model and Data Structure of Wordnets

Conversion of the lexicographers' files into an XML format. Neither the GermaNet lexicographers' files nor the compiled database yield an ideal format for data exchange, presentation and integration into NLP tasks. XML is more convenient for these purposes. Based on the data model of GermaNet, which is captured by an Entity-Relationship graph (cf. figure 1), several students developed programs which convert the Lexicographers' Files of GermaNet into an XML representation. The respective DTDs have been created collaboratively. The outcome of this project is documented in [6] and [7]). This task is designed to contribute to the ongioing discussion on the standardization of wordnet formats (e.g. the BalkaNet approach, cf. [8]).

Integration of the GermaNet objects and relations into the Resource Description Framework. Some work has already been done with WordNet within the framework of the Semantic Web initiative [9], but the resulting files encompass only a part of the Princeton WordNet. Before starting to convert GermaNet accordingly, and even more exhaustively, we would like to understand how well wordnet structures fit into the structures of full-fledged knowledge representation languages like DAML and OIL, which are built on top of RDF. An examination of these languages with wordnet structures in mind shall prove or disprove the usefulness of GermaNet objects and relations for the RDF and the other knowledge representation languages mentioned above. The work is under way.



CR=conceptual relation; LSR=lexical−semantic relation; oV=orthographic variant

Fig. 1. An entity relationship graph of the GermaNet data model

GermaNet representation as Scalable Vector Graphics. SVG (cf. [10]) might turn out to be a reliable standard as well as a handy tool for the visualization of wordnet objects and relations. A wordnet can be conceived as a large map where one wants to zoom in at a particular synset and see the data and relations that are associated with it. The project, which is not yet assigned, intends to explore the feasibility of data conversion into the SVG format and the functionality of existing visualization tools.

2.3 Tools

Development of tools for the extraction of the lexical and conceptual neighborhood of a lexical unit or a synset. The tools currently being developed are based on the XML representation of the data. The assignment in question addresses a user need for extracting data which are neighboring a particular synset or lexical unit. Currently, there are two projects devoted to this task: one employs a relational database for the intermediate representation of the data, the other accesses the data in their original format, using XSL Transformations to generate the output. Both methods will be evaluated in terms of their processing speed and flexibility. The GUI of one of the tools is shown in figure 2.

Visualization of the wordnet. Within another student project, a visualization tool which operates over the whole wordnet structure has been developed[3]. The XML representation of the wordnet is used as data base. The visualization of the data is very flexible, yet, too slow for realistic user scenarios. Results of the project have originally been presented in [7]. The outcome of the project has motivated our search for representation alternatives, e.g. Scalable Vector Graphics (see 2.2).

[3] Another promising approach to the visualiszation of wordnet structure has been launched by the Czech wordnet group: Visdic, cf. [11]

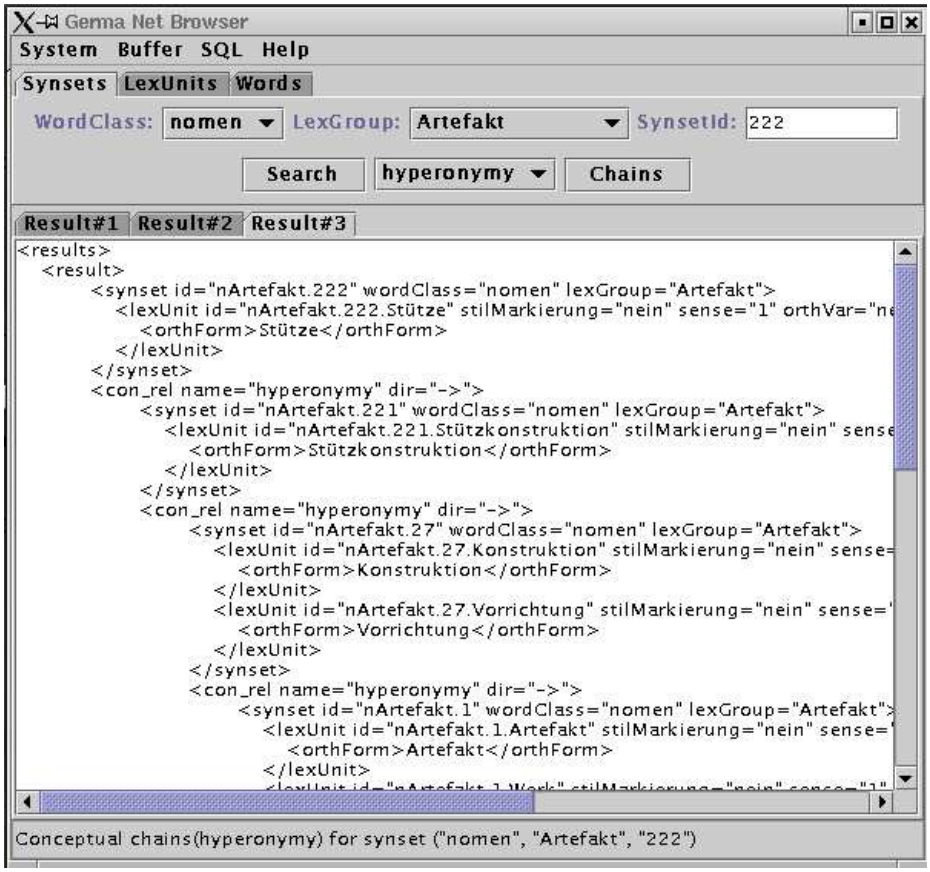


Fig. 2. GUI of a GermaNet extraction tool

2.4 Evaluation of Wordnet Data

Evaluation of the synset approach for IR and MT applications. 1. With a perspective on Machine (Aided) Translation, the EuroWordNet [12] ILI links, which are established between whole synsets (instead of lexical units), will be explored and compared with corresponding equivalence relations in a bilingual dictionary. The following questions are of interest: Does the majority of equivalence pairs between lexical units of the two languages, which are established indirectly through the relations between the synsets they are members of, really constitute pairs that can be used in substitution operations in MT? What is the relative share of mismatches for a particular language pair? 2. For IR applications, user tests will be performed on a search engine front-end, which expands query terms with their direct lexical and conceptual neighbors. The task to be developed could benefi t from results of cross-language IR evaluation (cf. [13]). *Evaluation of the feasibility of the "sense clustering" approach.* On the basis of a corpus of citations, some words which are highly polysemous shall be examined and the GermaNet senses mapped onto the corpus citations. The manual encoding of these

data within different sense division scenarios serves to prepare an experiment with automatic classifiers, which will be trained and tested on the different versions of the sense encoded corpus data.

2.5 Use of GermaNet As a Lexical Resource

LSI generated lexical semantic relations compared to GermaNet relations. In a larger project, which will lead to a diploma thesis, clusters of lexical units with alleged nearness in semantic space have been extracted from a large German newspaper corpus using Latent Semantic Indexing. In this project, the conceptual and lexical relations which are used for the construction of wordnets serve to evaluate the quality of the automatically generated "sense clusters". The aim of the evaluation is to investigate whether the lexical clusters yielded by LSI are really semantic, as the supporters of this approach claim.

GermaNet as a lexical basis for a vocabulary trainer. A group of students in Osnabrück, Edinburgh and Tübingen have developed a network-like platform for collaborative work. Within this framework, GermaNet as the central source of lexical knowledge supports a vocabulary trainer. The outcome of the project will be reported on the GLDV-Workshop on "Applications of the German Wordnet in Theory and Practice" in October 2003 (cf. [14]).

3 Conclusion

We have presented various examples of student projects that focus on GermaNet, and wordnets in general, as subjects of teaching and learning. With these examples, we have demonstrated how stimulating research and development projects can be in the teaching of advanced students. Wordnets are highly attractive for students for reasons of their simplicity and clarity and the richness of information they provide. With this paper, we want to claim the usefulness of wordnets for teaching in a broader range of subjects, including Computational Linguistics, Computational Lexicography, General Linguistics, Cognitive Science and Language Teaching. We would like to establish, within Global Wordnet Association, a repository which should encompass:

- a list of small to medium-sized research and development tasks for advanced undergraduate and graduate students (including task descriptions, methods, resources needed, possible outcomes);
- a list of results of student assignments (students might be further motivated by this prospect of publishing their work);
- a forum for discussing didactic issues.

References

1. Kunze, C., Naumann, K.: Germa Net Homepage. `http://www.sfs.uni-tuebingen.de/lsd`.
2. Chaffin, R.: The concept of a semantic relation. In Lehrer, A., Kittay, E. F., eds.: Frames, Fields, and Contrasts. New Essays in Semantic and Lexical Organization. Lawrence Erlbaum, Hillsdale (1992) 253–288.
3. Cruse, D. A.: Lexical Semantics. Cambridge University Press, Cambridge (1986).

4. Agricola, C., Agricola, E.: Wörter und Gegenwörter. Antonyme der deutschen Sprache, Leipzig: VEB Bibliographisches Institut (1987).
5. Pustejovsky, J., Boguraev, B., Verhagen, M., Buitelaar, P., Johnston, M.: Semantic Indexing and Typed Hyperlinking. In: Proceedings of the AAAI '97. (1997) `http://www.cs.brandeis.edu/~llc/publications/aaai97.ps`.
6. Kunze, C., Lemnitzer, L.: Standardizing Wordnets in a Web-compliant Format: The Case of GermaNet. In: Proc. LREC 2002 Workshop on Wordnet Structures and Standardization, a nd how these affect Wordnet Applications and Evaluation. (2002) 24–29.
7. Kunze, C., Lemnitzer, L.: GermaNet – representation, visualization, application. In: Proc. LREC 2002, Gran Canaria, May/June. (2002) 1485–1491.
8. Smrz, P.: Storing and retrieving Wordnet database (and other structured dictionaries) in XML lexical database management system. In: Proc. of First International WordNet Conference, Mysore, India (2002).
9. Melnik, S., Decker, S.: Wordnet RDF Representation. `http://www.semanticweb.org/library/` (2001).
10. Ferraiolo, J., Fujisawa, J., Jackson, D.: Scalable Vector Graphics (SVG) 1.1 Specification. `http://www.w3.org/TR/SVG11/` (2003).
11. Pavelek, T., Pala, K.: Visdic – a new tool for wordnet editing. In: Proc. of First International WordNet Conference, Mysore, India (2002).
12. Vossen, P.: Eurowordnet homepage. (URL: `http://www.illc.uva.nl/EuroWordNet/`).
13. Peters, C., Braschler, M., Gonzalo, J., Kluck, M., eds.: Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Darmstadt, Germany, September 3–4, 2001, Revised Papers. In Peters, C., Braschler, M., Gonzalo, J., Kluck, M., eds.: CLEF. Volume 2406 of Lecture Notes in Computer Science., Springer (2002).
14. Beck, K.: Ein Vokabeltrainer auf der Grundlage von GermaNet und Mapa (Mapping Architecture for People's Association). In Kunze, C., Lemnitzer, L., Wagner, A., eds.: Proceedings of the First GermaNet User Conference, Tübingen, 9–10 October 2003, GLDV (2003).

A Current Resource and Future Perspectives for Enriching WordNets with Metaphor Information

Birte Lönneker and Carina Eilts

Institute for Romance Languages, University of Hamburg
Von-Melle-Park 6, D-20146 Hamburg, Germany
Email: birte.loenneker@uni-hamburg.de, carina.eilts@uni-hamburg.de

Abstract. This article deals with the question whether metaphors might be integrated into WordNets in a more systematic way. After outlining the advantages of having more information on metaphors in WordNets, it presents the Hamburg Metaphor Database and a possible method for integrating metaphors and corresponding equivalence relations into monolingual WordNets. Finally, problems are discussed that will have to be faced before more metaphor information could be included in WordNets.

1 Introduction

This article confronts the problem of how information on metaphors might be integrated into WordNets in a more systematic way. In order to decide what this means, certain theoretical assumptions have to be made. We adopt the viewpoint that in most cases, "basic" or "literal" senses of a word can be identified. We then assume that a literal sense can be the basis for different kinds of – attested or hypothetical – metaphorical senses. As pointed out by [1], current WordNets do not display information on the relationship between these different word senses in a systematic way. We furthermore follow a cognitive framework introduced by [2], according to which individual metaphorical word senses illustrate the mapping from a more *concrete* conceptual "*source domain*", in which the corresponding literal sense is situated, to a more *abstract* conceptual "*target domain*", in which the metaphorical sense is situated. Several other theoretical viewpoints could be adopted when dealing with metaphors; however, for practical tasks, it is necessary to choose one (main) theoretical framework as a starting point.

The practical task envisioned here consists in adding metaphor information to WordNets. Why and for whom this kind of information would be useful is outlined in Section 2. A resource that will facilitate the enrichment of WordNets with systematic information on metaphors is the Hamburg Metaphor Database, containing metaphorical example sentences in French and German and their annotations with EuroWordNet and conceptual domain data (Section 3). While building and using this resource, we developed ideas of how the actual WordNet enrichment could be performed, but we also detected some points that require clarification before this work can start (Section4). Accordingly, Section 5 presents directions for future work.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 157–162.

2 Motivation

The fine sense distinctions made in WordNets have sometimes been criticized. However, in the case of metaphors, there are several reasons why a literal-figurative distinction is useful. Especially if this distinction is not only reflected in different synsets, but also documented by a relation between them and by information on the underlying domain mapping, metaphor information in WordNets can enhance a number of applications, for example:

- **Information Retrieval.** Information Retrieval would gain a lot from clearly identified metaphorical senses, because these senses are not valid for the parallel polysemy criterion (cf. [3]).
- **Word Sense Disambiguation.** Word Sense Disambiguation could be improved if lexical resources like WordNets provided senses and glosses for metaphors, enabling the automatic creation of semantically tagged corpora for machine learning (cf. [4]).
- **Language teaching.** Language teaching benefits from a domain-oriented view of metaphors; conceptually structured word/metaphor lists have proved to increase vocabulary retention (cf. [5]).

3 The Hamburg Metaphor Database

In view of the applications mentioned in Section 2 and inspired by work by Alonge and Castelli [1], the Hamburg Metaphor Database[1] (HMD) is being created in order to support studies of metaphors and WordNets. Based on domain-centered corpora, HMD provides both synset-oriented and domain-centered views on French and German metaphors, reachable online through a query interface.

The creation process of entries for the database can be briefly summarized as follows: Sentences or parts of sentences containing metaphors are extracted from a corpus and entered as "examples" into the database. The metaphorically used lexemes are identified in the examples and entered as "lexemes". Each lexeme is looked up in the respective part of EuroWordNet (EWN) [6]. If the intended metaphorical sense is already encoded in EWN, the corresponding synset is entered into the "metaphorical synset" field, as in the French example in Table 1: For *naissance* 'birth', the synset `naissance:3` (glossed as "the time when something begins [...]; 'they divorced after the birth of the child' or 'his election signaled the birth of a new age' ") allows a metaphorical reading. Synsets might also display an exclusively metaphorical sense of a lexeme, e.g. `father:5` 'a person who holds an important or distinguished position [...]'. However, if a lexeme can only be located in a synset which is interpreted as showing its basic sense, the synset is entered as "literal synset". Consider the German sentence in Table 1: The verb *verdunkeln* 'to darken' appears only in literal synsets; the selected transitive one, `vernebeln:1 verdunkeln:2` is glossed as "make less visible or unclear; 'The stars are obscured by the clouds' ". Finally, in case the lexeme does not appear in any EWN synset, no synset information is encoded in HMD.

The next step consists in finding conceptual domain information for the metaphorical mapping that is documented by the metaphor, as outlined in Section 1. The "source" domain underlies the literal sense of the lexeme (for instance, BIRTHING for the lexeme *naissance*

[1] `http://www.rrz.uni-hamburg.de/metaphern` [30.08.2003]

Table 1. Selected data from the metaphor table in HMD

Language	Example	Lexeme	Metaphorical synset	Literal synset	Source (Berkeley terms)	Target (Berkeley terms)
fr	A l'approche du conseil des 15 et 16 décembre à Madrid [...] Yves-Thibault de Silguy explique [...] que cette réunion doit constituer l'acte de naissance de la monnaie unique	naissance	naissance:3		BIRTHING	CREATING
de	Ein Aufklärer, der selber verdunkelt, ist unglaubwürdig.	verdunkeln		vernebeln:1 verdunkeln:2	DARK	BAD

and DARK for the lexeme *verdunkeln*, cf. Table 1), while the "target" domain is the one in which the metaphorical sense is situated (e.g. CREATING for *naissance* and BAD for *verdunkeln*). Two different naming systems for conceptual domains are used in HMD: The one of the Berkeley Master Metaphor List [7], and a proprietary German naming system, in which we add domain names missing from the Berkeley list.

User interfaces to the database allow for a query according to synsets, languages, domains, and corpora. The different corpora can be accessed by selecting one of the Master theses, in which the corpora were collected and documented. The Institute for Romance Languages in Hamburg currently disposes of 15 theses treating figurative language use in a cognitive linguistics framework.[2] Metaphors from six of these theses have been filed in HMD by August, 2003.

At the time of this writing, the database contains 394 corpus examples, documenting metaphorical uses of 300 distinct lexemes (138 in German, 162 in French). The French lexemes appear in 125 distinct synsets, 66 of them having a metaphorical meaning in EWN, and 59 displaying a literal meaning. The German lexemes appear in much less synsets; one of the reasons for this is that compounds were not split up into their parts. The database contains German synset annotations for 12 metaphorical and 29 literal synsets.

Although there is a large domain overlap in the French and German parts of HMD, the diversity of covered source and target domains is higher in the French part: 49 distinct source domains and 37 target domains have been identified for the French metaphors, while the German ones have been annotated as illustrating 22 source domains and 16 target domains. Metaphorical mappings "highlight" only certain aspects of the target domain which are seen in terms of the source domain [2]; therefore, several source domains might coexist for the same target domain and highlight different aspects: For instance, POLITICS (target) can be seen in terms of FIGHT, SPORTS, THEATER, or STUDY [8].

[2] Supervisor: Prof. Dr. Wolfgang Settekorn, French Linguistics/Media Science.

Several other databases and searchable lists of metaphors exist on the World Wide Web. For example, the ATT-Meta Project Databank[3] developed by John Barnden contains examples of usage of metaphors of mind. The Berkeley MetaNet database MetaDB also includes domain information.[4] However, to our knowledge, no other metaphor database explicitly includes WordNet information.

4 Towards a Systematic Metaphor Representation in WordNets

The current structure of EWN does not include a relationship which would allow the linking of metaphorical synsets to literal synsets. We therefore envision a method of adding new **eq_metaphor** relations at the level of a composite index, following [8]: A study of HMD example sentences and lexemes taken from several source and target domains led to the conclusion that a domain-centered view with a "central synset" referring to the overall source domain (an event like BIRTHING, in most cases), could be used as a starting point to semi-automatically add metaphorical synsets to existing WordNets. After manually connecting the central synset to its "parallel" metaphorical synset (containing identical literals with different indexes), parallel metaphorical synsets can automatically be created for all synsets that are connected to the central synset by a hyperonym, holonym, role – or possibly other – sense relationship. Glosses [9] for the new synsets could be created using templates to be filled with information like source synset and parts of glosses from the "central" source and target synsets (cf. Figure 1). A computer-assisted manual cleaning should be performed with special attention to those lexemes for which metaphorical senses already exist as synsets in EWN. These, as well as others actually *attested* in HMD, can be specially marked, in order to distinguish them from the remaining automatically created *potential* metaphors.

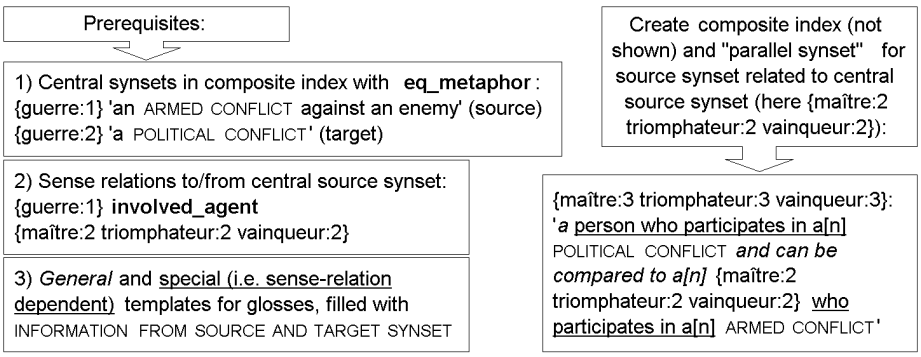


Fig. 1. Automatic creation of metaphorical synsets

[3] `http://www.cs.bham.ac.uk/~jab/ATT-Meta/Databank/` [30.08.2003]
[4] Personal communication from Michael Meisel, ICSI, Berkeley [4 September 2003].

In that way, gaps in EWN corresponding to empty synset fields in HMD would be filled. Still, other problems detected while building HMD need separate consideration and are summarized in what follows.

- **Missing glosses and scarsity of relationships.** Synsets in EuroWordNet do not always have glosses. If glosses are missing or incomplete, only the relations to other synsets might tell which sense is to be attributed to a synset. Given the small range of semantico-conceptual relations actually used in the French EWN – apart from hyperonymy, it contains only some antonymy and meronymy relations –, interpreting a synset is sometimes close to guessing.
- **Incorrect and incomplete synsets.** Incorrect synsets are rare, but they occur: e. g. French {père:2 parent:3 mère:2} ({father parent mother}). If 'father' and 'mother' were synonyms, they should be interchangeable in the same context without changing its meaning, which is not the case. Incomplete synsets are those from which at least one "literal" seems to be missing, as for instance the French synset {magazine:1 périodique:3} 'magazine'; there is no obvious reason why the literal *revue* has been omitted.
- **Literal-figurative inconsistencies.** Sometimes, HMD encoders detect a synset with an apparently metaphorical meaning, showing semantico-conceptual relationships to clearly literal synsets (cf. also [8]). As long as metaphors are only documented in the database as explained above, this is not a crucial problem; however, as soon as one would like to create domain views and treat metaphorical mappings using a more or less automated procedure, these inconsistencies will result in errors.
- **Collocations and compounds.** Problems arise when (parts of) collocations or compounds bear a metaphorical meaning. Idioms (as a special case of collocations) are rarely represented in WordNets [10]; it is also difficult to individuate one single constituent in them displaying metaphorical usage, like in German *den Weg freimachen* 'to clear the way', French *mettre sur les rails* 'to put on the rails' – the whole idiom has a metaphorical meaning. For highly compounding languages like German, some compounds are represented as literals in EWN synsets, others not. Apart from the fact that the searched items might not be found in EWN, the ascription of domain mappings to whole compounds is problematic, because in general only one of the constituents is used figuratively (cf. German *Lügensumpf* 'swamp of lies', *Spendensumpf* 'swamp of donations' – only 'swamp' is metaphorical).

A more in-depth study on metaphors and WordNets, aiming at adding structured information on metaphors to WordNets using well-established EWN-means (composite index, synsets, relations and glosses), will thus have to take into consideration much more topics and issues than those directly related to metaphor.

5 Conclusion and Future Work

The encoding of systematic information on literal-metaphorical-relationships in WordNets necessitates careful analysis of the problems encountered, and step-by-step solutions. We hope to continue our work in two parallel lines:

1. Process those additional interpreted corpora that are available to the Hamburg Metaphor Database, in order to provide more material on metaphors, involved synsets and domain mappings.
2. For some selected source domains, create add-ons to the monolingual parts of EWN. If necessary, we will correct synsets of the source domain and complete the source domain structure by adding semantic relations, with the help of a tool for WordNet editing like VisDic [11] and taking into account further developed resources like GermaNet. Using the semi-automatic methods described above, metaphorical synsets and glosses will then be created. Finally, a script could integrate the add-ons into existing EWN-fi les.

References

1. Alonge, A., Castelli, M.: Which way should we go? Metaphoric expressions in lexical resources. In: Proceedings LREC-02, Las Palmas, Gran Canaria, 29–31 May. Volume VI. (2002) 1948–1952.
2. Lakoff, G., Johnson, M.: Metaphors we live by. University of Chicago Press, Chicago/London (1980).
3. Chugur, I., Gonzalo, J., Verdejo, F.: A study of sense clustering criteria for information retrieval applications. In: Proceedings OntoLex 2000, Sozopol, Bulgaria, 8–10 September. (2000).
4. Mihalcea, R., Moldovan, D.: An automatic method for generating sense tagged corpora. In: Proceedings AAAI '99, Orlando, Florida, 18–22 July. (1999) 461–466.
5. Boers, F.: Metaphor awareness and vocabulary retention. Applied Linguistics **21** (2000) 553–571.
6. Vossen, P.: EuroWordNet general document Version 3, Final (1999) `http://www.illc.uva.nl/EuroWordNet/docs.html`.
7. Lakoff, G., Espenson, J., Schwartz, A.: Master metaphor list, Second draft copy (1991) `http://cogsci.berkeley.edu`.
8. Lönneker, B.: Is there a way to represent metaphors in WordNets? Insights from the Hamburg Metaphor Database. In: Proceedings of the ACL-03 Workshop on the Lexicon and Figurative Language, Sapporo, Japan, July 11. (2003) 18–26.
9. Pala, K., Smrž, P.: Glosses in WordNet 1.5 and their standardization/consistency. In: Proceedings of the LREC-02 Workshop on WordNet Structures and Standardisation, and How These Affect WordNet Applications and Evaluation, Las Palmas, Gran Canaria, 28 May. (2002) 20–23.
10. Fellbaum, C.: Towards a representation of idioms in WordNet. In: Proceedings of the COLING/ACL-98 Workshop on Usage of WordNet in Natural Language Processing Systems, Montreal, 16 August. (1998) 52–57.
11. Horák, A., Pala, K., Smrž, P.: Lexical semantic networks and ontologies in XML, their viewing and authoring. In: Proceedings of the CIC-03 Workshop on WWW Based Communities For Knowledge Presentation, Sharing, Mining and Protection, Las Vegas, Nevada, 25 June. (2003).

Sociopolitical Domain As a Bridge from General Words to Terms of Specific Domains

Natalia Loukachevitch and Boris Dobrov

Research Computing Center of Moscow State University,
Leninskie Gory, Moscow, 119992, Russia
Email: louk@mail.cir.ru, dobroff@mail.cir.ru

Abstract. In the paper we argue that there exists a polythematic domain which is situated in an intermediate area between senses of a general language area and specific domains. The concepts of this domain can be naturally added to general wordnets together with publicly known technical terms. Such enhanced wordnets can provide much more considerable preliminary coverage of domain specific texts, improve efficiency of word sense disambiguation procedures.

1 Introduction

Majority of the texts in electronic collections contain as general words as terms from specific domains. To effectively organize automatic text processing, knowledge resources have to include descriptions of both types of language expressions. However for years general words and domain terms were studied by different research communities. Lexicology and lexicography studied meanings of general words, technical terms were considered by terminologists in the general theory of terminology. Wuster wrote that the main difference in consideration of general words by lexicologists and terms by terminologists was as follows: terminologists begin consideration from a concept, but lexicologists from a form of a linguistic expression [15]. He wrote that terminological research starts from the concept which has to be precisely delimited and that in terminology concepts are considered to be independent from their designations. This explains the fact that terminologists talk about 'concepts' while linguists talk about 'word meanings'.

But now when linguists began to develop wordnets for various languages, the situation is changing. Creating wordnets linguists construct hierarchical semantic networks, try to find similar "synsets" for different languages, build the top ontology of language-independent concepts [2]. These directions of lexical research are much closer to the study of concepts, therefore the distinction between approaches seems to be considerably less serious.

Recently researchers began development of wordnets for specific domains [1,14]. From this point of view it is very important to understand how a general wordnet and domain specific wordnets interact with each other, how development of domain specific wordnets correlates with terminology research, if it is possible to combine lexical and terminological knowledge in the same linguistic resource.

In this paper we argue that there exists a polythematic domain which is situated in an intermediate area between senses of general language and concepts of specific domains and partially intersects with both ones. The concepts of this domain can be naturally added to

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 163–168.

general wordnets together with the most known technical terms. Such enhanced wordnets can provide much more considerable preliminary coverage of domain-specific texts, to serve as a reliable source for development of domain-specific ontologies.

2 Features of Terms

There are a lot of definitions of a term given by terminologists. Most of them consider a term as a word or expression designating a concept in a special domain. A specific feature of a term is that its relations with other terms of the domain is described by a definition [11].

The whole set of terms of a domain is comprised by the terminology of the domain. This system of terms during the process of its development usually undergoes procedures of standardization and normalization to be understandable for all specialists in the domain.

For choice of appropriate terms in the standardization process terminologists consider the following features of an ideal term [12]:

- the term must relate directly to the concept. It must express the concept clearly;
- there should be no synonyms where absolute, relative or apparent;
- the contents of terms should be precise and not overlap in meaning with other terms;
- the meaning of the term should be independent of context.

According to [5] "the objective of term-concept assignment in a given special language is to ensure that a given term is assigned to only one concept is represented by only one term".

This means that in ideal cases there must be a biunivocal relationship between concepts and terms in each special field of knowledge. For a terminology nothing could be better than that: no synonymy, no homonymy and no polysemy.

Though this ideal situation only happens in a few well structured fields and does not happen for the rest, this terminologists' point of view stresses how considerable is difference between a term and a word of general language. However, in reality the gap a word – a term is not so broad.

3 Term Formation and Words of General Language

An impregnable barrier between words of a general language and terminologies does not exist. A lot of terms (for example, terms in technical domains) appeared in specific domains become elements of a general language. On the other hand a general language word can change its meaning and become an element of a terminology.

Among possible transitions from a general language to a terminology it is important to distinguish the following cases:

1. a sense of a general word and a sense of the same wordform as a technical term are really different. For example, a new sense of a term can result from metaphoric shift or domain specification of a general sense. So there is a general sense of word "function", there is term "function" in biology, there is term "function" in mathematics and so on. A usage of word "function" can never have all or several of these meanings, that is an important rule of distinguishing of different senses of a word fulfills: " senses of a lexical form are antagonistic to one another; that is to say, they can not be brought into play simultaneously without oddness" [3].

2. a sense of a term in a domain-specific terminology is only slightly refined in comparison to a sense of the same word as a general language expression. Let us consider several terms from criminal law that also exist as words of general language. In this cases dictionaries often use terminological definitions as glosses such as ***arson**-Law. the malicious burning of another's house or property, or in some statutes, the burning of one's own house or property, as to collect insurance [10]*.

If one supposes that there exist two senses of such legislative terms as *arson, murder* or *bail*, then one have to agree that for too many usages it is impossible to distinguish the general usage of a word from the terminological use, especially in media texts. So news reports can be understood by ordinary people and at the same time such texts can contain a lot of domain-specific terms.

In such situations we should not distinguish two senses of such words. In fact, the same sense "works" in a general language and in domain-specific language.

One can argue that terminological definitions delimit domain concepts stricter than definitions of explanatory dictionaries. Indeed, the borders of a general language sense can be very vague. Using a general language word we distinguish typical cases and can mistake or doubt in complicated cases (as previously one could think that a whale is a fish). A terminology tries to provide a concept with more definite boundaries, for example, legislators use a page long definition to distinguish "new construction" from "repair" for taxation needs. However we think that if there is an agreement in typical cases the problem vague vs. strict boundaries of a sense is not a reason to separate senses. We suppose that people do not think about concept boundaries because of lack of necessity. If necessary they readily use domain definitions as a support. So for the most known legislative terms general dictionaries use law definitions.

4 Notion of Sociopolitical Domain

It is important to understand how many senses of general language words practically coincide with senses in specific domains. A scope of such senses is not restricted with the legal domain. Let us take word "Building" as a noun in sense 1: *a relatively permanent enclosed construction over a plot of land, having a roof and usually windows and often more than one level, used for any of a wide variety of activities, as living, entertaining, or manufacturing* [10].

Terms with similar senses are necessary at least in two fields of public activity such as the construction trade and the field of public utilities. It means that majority of artifact senses of general language words coexist as terms in two fields of business activity: a field of industrial production of the artifact and a field using the artifact.

Main classes of such "dual" concepts include transportation means, job positions, technical devices, food, agricultural plants and animals, other natural objects, social, political and economic processes, art work and so on. These concepts are very important in everyday life, therefore people need language expressions to speak about them. At the same time fields of social activities, social sciences include them in their special languages. We estimate that almost 40 percents of general language word senses are used in various social subdomains. (For all estimations the lexical and terminological resource of Russian language RuThes containing more than 105 thousand words, collocations and terms [7] is used).

Thus we can distinguish a large specific domain, incorporating all these concepts – a domain of political, economic and social life, a domain that comprises general language senses coinciding with concepts of various domains of social activities. We call this polythematic domain "sociopolitical domain".

The sociopolitical domain has very interesting properties. These properties do it very useful to distinguish a sociopolitical zone in wordnets conceptual systems for automatic text processing goals.

5 Properties of Sociopolitical Domain

The sociopolitical domain has the following properties**.**

Property 1. Location of senses of the sociopolitical zone in general wordnets. Synsets belonging to the sociopolitical zone are situated mainly in the lower levels of the wordnet's conceptual system. Therefore the senses are the most thematically definite. The consequences of the fact are as follows: if such a general word as "creation" is used in a text, it can relate to different entities, different elements of the text structure. If such a "sociopolitical" word as "transportation" is mentioned several times in a text it is possible to suppose that all usages of the word are elements of the same topic structure and use this fact, for example, for construction of lexical chains and identification of the thematic structure of a text [9].

Property 2. Lexical ambiguity within the sociopolitical zone of the general language conceptual structure is much lower. For instance, in the current version of RuThes the ratio, denoting amount of second, third and other senses of expressions,

N = (number of relations "word-concept" – number of different words)
/ number of different words

is more than 4 times lower in the sociopolitical zone than in the whole resource.

Property 3. Lexical disambiguation for synsets within the sociopolitical zone is much easier, because different senses are often situated in different social subdomains and have rather different contexts of their usage in texts. For information-retrieval purposes synsets of the sociopolitical zone are much more important. Therefore it is possible to divide word sense disambiguation into three parts:

- disambiguation within the sociopolitical zone;
- disambiguation of term senses belonging the sociopolitical zone and general levels of the language conceptual system, to decide if a sociopolitical sense was applied;
- work with undisambiguated words out of the sociopolitical domain.

This combined approach to lexical disambiguation can diminish problems of incorrect disambiguation in automatic text processing in wordnet-based information-retrieval systems.

Property 4. Besides linguistic expressions having dual functions as general language means and terminological means there are a lot of terms (usually multiword terms) in domains of public affairs which can be understood by majority of the native speakers such as *aircraft industry, crime prevention, military assistance, internal migration.* The existence of such a polythematic terminological level, its importance for various information needs was recognized by developers of information-retrieval thesauri. Several general sociopolitical

thesauri [6,13] have been created and are used for indexing and retrieval of such important types of documents as governmental, parliamentary, international documents.

This set of terms can be naturally added to the sociopolitical zone of a wordnet. The inclusion of multiword expressions gives additional information for disambiguation. Such an enhanced wordnet becomes a valuable initial source for development of domain-specific ontologies. So for development of Avia-Ontology, describing interaction of an operator (air crew) and board equipment in various flight situations (1200 concepts, 3400 terms), almost a third part of the ontology was taken from thesaurus RuThes [4], comprising a lot of Russian sociopolitical terminology.

6 Related Work

Broadly speaking, the sociopolitical domain can be compared with an aggregate of all subject fields proposed in [8], except the Factotum field. The main differences are as follows:

- Systems of subject fields can be quite different. We propose not to work with any given system but analyze if a synset belongs a set of possible domains of social activity.
- The main point here is not to find such domains for maximal number of synsets but provide real analysis of domains otherwise multiple overgeneration of subject field codes can arise.
- It is important not only to mark "sociopolitical" synsets but recognize the existence of a broad layer of synsets belonging to as the general language system as to upper levels of various specific domains' hierarchies.

7 Conclusion

A border between a general language lexicon and terminologies of specific domains is not sharp and abrupt. It looks more as a broad strip and contains general language senses practically coinciding with concepts of social subdomains and concepts of specific domains understandable for native speakers.

Detailed description of concepts, terms, words from this "transition area", called "sociopolitical domain", can be naturally added to a wordnet' semantic network and facilitate solution of such problems as lexical disambiguation and identification of the text structure, enhance coverage of domain-specific texts by wordnets' synsets, improve effectiveness of the wordnets use in various automatic text processing applications.

Acknowledgements

Partial support for this work is provided by the Russian Foundation for Basic Research through grant # 030100472.

References

1. Buitellar, P., Sacalenau, B.: Extending Synsets with Medical Terms In proceedings of the NAACL workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations, Pittsburg, USA (2001).
2. Climent, S., Rodriguez, H., Gonzalo, J.: Definitions of the links and subsets for nouns of the EuroWordNet project. – Deliverable D005, WP3.1, EuroWordNet, LE24003 (1996).
3. Cruse: Lexical Semantics. – Cambrige (1986).
4. Dobrov, B., Loukachevitch, N., Nevzorova, O.: The Technology of New Domains' Ontologies Development. Proceedings of X-th Intern. Conf. KDS 2003 "Knowledge-Dialogue-Solution". June 16–26, Varna, Bulgaria. pp. 283–290. (2003).
5. ISO/DIS 704: Terminology work – Principles and methods. Geneva, ISO (Revision of second edition 704:1987) (1999).
6. LIV: *Legislative Indexing Vocabulary*. Congressional Research Service. The Library of Congress. Twenty first Edition (1994).
7. Loukachevitch, N. V., Dobrov, B. V.: Development and Use of Thesaurus of Russian Language RuThes. Proceedings of workshop on WordNet Structures and Standartisation, and How These Affect WordNet Applications and Evaluation. (LREC 2002) / Dimitris N. Christodoulakis, Gran Canaria, Spain – p. 65–70 (2002).
8. Magnini, B., Cavaglia, G.: Integrating Subject Field Codes into WordNet. – Proceedings of LREC 2000, Athens (2000).
9. Morris, J., Hirst G.: Lexical cohesion computed by thesaural relations as an indicator of the structure of a text. Computational linguistics, 17(1), pp. 21–48 (1991).
10. Random House Unabriged dictionary: Random house, Inc. (1999).
11. Rondeau, G.: Introduction a a terminologie. Quebec (1980).
12. Sager, J.C.: A Practical Course in Terminology Processing. Amsterdam: J. Benjamins (1990).
13. Thesaurus EUROVOC: Vol. 1–3 / European Communities. – Luxembourg: Office for Official Publications of the European Communities, Ed. 3. – English Language (1995).
14. Vossen, P.: Extending, Trimming and Fusing WordNet for Technical Documents. In: Proceedings of WordNet and Other Lexical Resources: Applications, Extensions and Customizations, Pittsburg, USA (2001).
15. Wuster, E.: Einfurung in die Allgemeine Terminologielehre and terminologishe Lexicographie. – Wien; N.Y., 1979/Bd 1–2.

Using WordNet Predicates for Multilingual Named Entity Recognition

Matteo Negri and Bernardo Magnini

ITC-Irst, Centro per la Ricerca Scientifica e Tecnologica,
Via Sommarive, 38050 Povo (TN), Italy
Email: negri@itc.it, magnini@itc.it

Abstract. *WordNet predicates* (WN-PREDS) establish relations between words in a certain language and concepts of a language independent ontology. In this paper we show how WN-PREDS can be profitably used in the context of multilingual tasks where two or more wordnets are aligned. Specifically, we report about the extension to Italian of a previously developed Named Entity Recognition (NER) system for written English. Experimental results demonstrate the validity of the approach and confirm the suitability of WN-PREDS for a number of different NLP tasks.

1 Introduction

WORDNET predicates (WN-PREDS) are defined over a set of WORDNET synsets which express a certain concept. A WN-PRED takes as input a word *w* and a language *L* in which the word is expressed, and returns TRUE if at least one sense of *w* in *L* is subsumed by at least one of the synsets defining the predicate, and FALSE otherwise. As an example, a WN-PRED "*location-p*" can be defined over the high-level synsets `location#1`, `mandate#2`, `road#1`, `solid_ground#1`, `body_of_water#1`, `geological_formation#1`, and `celestial_body#1`[1]. According to the previous definition:

location-p [<capital>, <English>]

returns `capital#3` (i.e.TRUE) since this sense of "capital" in the English WORDNET is subsumed by at least one of the synsets defining the predicate (*i.e.* `location#1`). On the other hand:

location-p [<computer>, <English>]

returns FALSE since none of the senses of "computer" is subsumed by one of the synsets defining the concept of location.

WORDNET predicates establish relations between a single word in a language and a general concept in a language independent ontology. However, WORDNET predicates are context independent *i.e.* they produce the same result for the same word, independently of the context in which the word occurs. As a consequence, their practical use is limited to applications (such as the one proposed in this paper) in which predicates are coupled with contextual information.

[1] Throughout the paper WORDNET word senses are reported with this `typeface#1`, where #1 is the corresponding sense number in WORDNET 1.6, while Named Entity categories are indicated with this TYPEFACE.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 169–174.

While the use of WORDNET predicates has been proposed in several NLP tasks, including Named Entity Recognition (NER) [3] and Question Answering (QA) [6], this paper addresses their more specific use in a multilingual scenario, where two or more wordnets are aligned. Starting from the WORDNET predicates used in an NER system for written English (overviewed in Section 2), we experimented the portability of the approach building an Italian system without any change in the predicates (Section 3). Results (Section 4) are highly encouraging, and demonstrate the suitability of the proposed methodology both in term of performance and in term of time required for system development.

2 Using WORDNET Predicates for NER

NER is the task of identifying and categorizing entity names (such as persons, organizations, and locations names), temporal expressions (dates and times), and certain types of numerical expressions (monetary values and percentages) in a written text. Knowledge-based approaches, which represent a possible solution to the NER problem, usually rely on the combination of a wide range of knowledge sources (for example, lexical, syntactic, and semantic features of the input text as well as world knowledge and discourse level information) and higher level techniques (*e.g.* co-reference resolution). In this framework, dictionaries and extensive gazetteer lists of first names, company names, and corporate suffixes are often claimed to be a useful resource. Nevertheless, several works (see, for example, [5]) pointed out some drawbacks related to the pure list lookup approach, which mainly depend on the required dimensions of reliable gazetteers, on the difficulty of maintenance of this kind of resource, and on the possibility of overlaps among the lists. Moreover, their availability for languages other than English is rather limited.

An effective solution to these problems has been recently proposed in [3], and relies on a rule-based approach which avoids the difficulties related to the construction and maintenance of reliable gazetteers by making the most of the information stored in the WORDNET hierarchy. The starting point, as also suggested by [4], is that the identification and classification of a candidate named entity can be tackled by considering two kinds of information, namely *internal* and *external* evidence. The former is provided by the candidate string itself, while the latter is provided by the context in which the string appears. As an example, in the sentence, "Judge Pasco Bowman II, who was appointed by President Ronald Reagan ...", the candidate proper names "Pasco Bowman II" and "Ronald Reagan" can be correctly marked with the tag PERSON either by accessing a database of person names (*i.e.* considering their internal evidence) or by considering the appositives "Judge", "II" and "President", or the pronoun "who" as external evidence for disambiguation.

While internal evidence is mostly conveyed by proper nouns, external evidence can be conveyed by the presence in the text of *trigger words*, *i.e.* predicates and constructions providing sufficient contextual information to determine the class of candidate proper nouns in their proximity [9]. For instance, systems designed to deal with this kind of information usually access more or less complete hand-crafted word lists containing expressions like "director", "corporation", and "island" in order to recognize respectively person, organization, and location names into a given text.

In light of these considerations, the basic assumption underlying the approach suggested by [3] is that the huge number of possible trigger words that can be extracted from WORD-

NET compensates for the relatively limited availability of proper nouns, thus forming a reliable basis to accomplish NER without the further use of gazetteer lists. In this framework, they propose a semi-automatic procedure to extract trigger words from WORDNET, and to separate them from proper nouns bringing internal evidence. This procedure exploits the IS-A relation to distinguish between *Word_Classes* (*i.e.* concepts bringing external evidence, such as `river#1`) and *Word_Instances* (*i.e.* particular instances of those concepts, such as `Mississippi#1`, which can be marked as entity words also without any contextual information) present in WORDNET. For instance, as for the NE category LOCATION, starting from the high level synsets already listed in Section 1, and considering as proper nouns their capitalized hyponyms, they obtain 1591 English Word_Classes and 2173 Word_Instances. Once the relevant high level synsets have been selected, and the corresponding Word_Classes and Word_Instances have been mined from the WORDNET hierarchy, WORDNET predicates relevant to each NE category (*e.g.* "person-p", "person-name-p" "location-p", "location-name-p", "organization-p", etc.) are used to access this information in the NER process. The task is accomplished by means of simple rules that check for different features of the input text, detecting the presence of particular word senses satisfying the WORDNET predicates, as well as word lemmas, parts of speech or symbols.

3 Porting to Italian

The construction of an NER system for written Italian represented an ideal opportunity to test the portability of the above outlined approach, which [3] has claimed to be well-suited to multilingual extensions. In fact, in addition to its effectiveness in the NER task, mining information from WORDNET also offers a practicable way to address multilinguality. This is due to the recent spread of multilingual semantic networks aligned with WORDNET, a necessary condition for the complete reusability of the predicates defined on the English taxonomy.

Our extension to Italian takes advantage of MULTIWORDNET [8], a multilingual lexical database developed at ITC-Irst which includes information about English and Italian words. MULTIWORDNET is an extension of the English Princeton WORDNET, keeping as much as possible of the original semantic relations. Italian synsets have been created in correspondence with English synsets, whenever possible, by importing lexical and semantic relations from the corresponding English synsets. The Italian part of MULTIWORDNET currently covers about 43,000 lemmas, completely aligned with English WORDNET 1.6.

Exploiting the alignment between the two languages, Italian Word_Classes and Word_Instances have been mined from MULTIWORDNET starting from the high-level synsets defined on the English taxonomy and collecting their Italian equivalents as well as their hyponyms. Table 1 shows their distribution with respect to the NE categories we used in our experiments (namely PERSON, LOCATION, and ORGANIZATION), compared to the distribution of the English words. It's worth noting that, in order to improve the system performance, all the English Word_Instances have been also used in our extension since most of them (*e.g.* proper nouns like "William Shakespeare", "Beverly Hills", and "UNESCO") usually are not translated into Italian. The same holds for some of the English Word_Classes (*e.g.* "anchorman", "checkpoint", and "corporation"), which can be considered as trigger words also when they are encountered within an Italian text. This way, even though the over-

Table 1. Distribution of Word_Classes and Word_Instances in MULTIWORDNET

	#ENG Classes	*#ENG Instances*	*#ITA Classes*	*#ITA Instances*
PERSON	6775	1202	5982	348
LOCATION	1591	2173	979	950
ORGANIZ.	1405	498	890	297
TOTAL	9771	3873	7851	1595

all number of Italian words is lower, both internal and external evidence are still effectively captured by the system.

Using the information mined from the MULTIWORDNET hierarchy, and taking advantage of the complete reusability of the English WORDNET predicates, the process of recognition and identification of NEs is carried out in three phases.

Preprocessing. In the first phase, the input text is tokenized and words are disambiguated with their lexical category by means of a statistical part of speech tagger developed at ITC-Irst. Also, multiwords recognition is carried out in this phase: about seven thousand Italian multiwords (*i.e.* collocations, compounds, and complex terms) have been automatically extracted from MULTIWORDNET and are recognized by pattern matching rules.

Basic rules application. In the second phase, a set of approximately 400 *basic rules* is in charge of finding and tagging all the possible NEs present in the input text. Most of these rules capture internal and external evidence by means of the WORDNET predicates used to mine the Italian taxonomy. As an example, Table 2 describes a rule containing the WORDNET predicate "location-p", which is satisfied by any of the 979 Italian Word_Classes of the category LOCATION extracted from MULTIWORDNET. This rule captures contextual evidence matching with sentences formed by a capitalized noun followed by a verb whose lemma is "essere" (*i.e.* "to be"), a determiner, and any of those trigger words, like "capitale" in "Roma e' la capitale italiana" (*i.e.* "Rome is the Italian capital").

Table 2. A rule matching with "Roma e' la capitale italiana"

PATTERN	*t1 t2 t3 t4*
t1	[pos = "NP"] [ort = Cap]
t2	[lemma = "essere"]
t3	[pos = "DT"]
t4	[sense = (location-p *t4* Italian)]
OUTPUT	<LOCATION>t1<\LOCATION>

Composition rules application. Besides the application of the basic rules, a correct NER procedure requires the application of higher level rules in charge of resolving co-references between recognized entities and proper names not yet disambiguated, as well as handling tagging ambiguities, tag overlaps and inclusions. For instance, considering the start/end position of the tags, the content, and the tag type of the candidate entities, these rules handle inclusions which may occur when a recognized entity contains other more specific entities, as in "Università di Napoli" (*i.e.* "Naples University"), where a proper noun belonging to the

category LOCATION (*i.e.* "Napoli") is included into an entity belonging to the more general category ORGANIZATION.

4 Results and Conclusion

System performance was evaluated using the scoring software provided in the framework of the DARPA/NIST HUB4 evaluation exercise [1]. Scores (i.e. F-measure, Precision and Recall) have been computed by comparing a 77 Kb reference tagged corpus[2] with an automatically tagged corpus according to *type*, *content* and *extension* of the NE categories PERSON, LOCATION, and ORGANIZATION. Table 3 illustrates the results achieved by our system, compared with the performance of the English version described by [3].

Table 3. Overall Precision, Recall and F-Measure scores

	Recall	*Precision*	*F-Measure*
PERSON	91.48 (87.29)	85.08 (88.38)	88.16 (87.83)
LOCATION	97.27 (92.16)	80.45 (81.17)	88.07 (86.32)
ORGANIZATION	83.88 (82.71)	72.70 (83.02)	77.89 (82.87)
All categories	91.32 (87.28)	74.75 (82.99)	82.21 (84.12)

As can be seen from Table 3, even though MULTIWORDNET is smaller than WORDNET, our results compare well with the ones achieved by the English version. For instance, considering the category LOCATION, even if for WORDNET 1.6 provides about 600 Word_Classes more than the Italian part of MULTIWORDNET, the difference between the two F-Measure scores is rather small (*i.e.* 0.67). The suitability and the portability to other languages of the WORDNET-based approach to NER are also confirmed by the relatively limited amount of time required for system development. In fact, since the WORDNET predicates defined on the English taxonomy were reused without any change, all the effort was concentrated on the creation of the Italian rules, which took approximately one person month.

As a final remark, it's worth noting that while in the present work WORDNET predicates have been defined according to the concepts that are relevant for the NER task (*i.e.* PERSON, LOCATION, and ORGANIZATION), a wider set of such predicates can be easily realized by taking advantage of the concepts defined in already available upper-level ontologies and their mappings to WORDNET. Among these ontologies, an important role in the framework of approaches similar to the one described in this paper could be played by the SUMO ontology [7], with about 1100 concepts completely mapped against WORDNET, and the DOLCE ontology [2], whose mapping to WORDNET is, however, still under development.

[2] Reference transcripts of two Italian broadcast news shows, including a total of about 7,000 words and 322 tagged named entities, were manually produced for evaluation purposes

References

1. Chinchor, N., Robinson, P., Brown, E.: Hub-4 Named Entity Task Definition (version 4.8). Technical Report, SAIC. `http://www.nist.gov/speech/hub4_98` (1998).
2. Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., Schneider, L.: Sweetening Ontologies with DOLCE. Proceedings of EKAW 2002. Siguenza, Spain (2002).
3. Magnini, B., Negri M., Prevete R., Tanev H.: A WORDNET-Based Approach to Named Entities Recognition. Proceedings of SemaNet '02: Building and Using Semantic Networks Taipei, Taiwan (2002) 38–44.
4. McDonald, D.: Internal and external evidence in the identification and semantic categorization of proper names. In: Boguraev, I., Pustejovsky, J. (eds.): Corpus Processing for Lexical Acquisition, Chapter 2. The MIT Press, Cambridge, MA (1996).
5. Mikheev, A., Moens, M., Grover, C.: Named Entity recognition without gazetteers. Proceedings of EACL-99, Bergen, Norway (1999).
6. Moldovan, D., Harabagiu, S., Girju, R., Morarescu, P., Lacatusu, F., Novischi, A., Badulescu, A., Bolohan, O.: LCC Tools for Question Answering. Proceedings the TREC-2002 Conference, NIST, Gaithersburg, MD (2002), Bergen, Norway (1999).
7. Niles, I., Pease, A.: Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. Proceedings of the 2003 International Conference on Information and Knowledge Engineering (IKE '03), Las Vegas, Nevada, (2003).
8. Pianta, E., Bentivogli, L., Girardi, C.: MultiWordNet: Developing an Aligned Multilingual Database. Proceedings of the 1st International Global WordNet Conference, Mysore, India (2002).
9. Wakao, T., Gaizauskas, R., Wilks, Y.: Evaluation of an Algorithm for the Recognition and Classification of Proper Names. Proceedings of the 16th Conference on Computational Linguistics (COLING '96), Copenhagen, Denmark (1996).

Results and Evaluation of Hungarian Nominal WordNet v1.0

Márton Miháltz and Gábor Prószéky

MorphoLogic
Késmárki u. 8, Budapest, 1118 Hungary
Email: mihaltz@morphologic.hu, proszeky@morphologic.hu

Abstract. This paper presents recent results of the ongoing project aimed at creating the nominal database of the Hungarian WordNet. We present 9 different automatic methods, developed for linking Hungarian nouns to WN 1.6 synsets. Nominal entries are obtained from two different machine-readable dictionaries, a bilingual English-Hungarian and an explanatory monolingual (Hungarian). The results are evaluated against a manually disambiguated test set. The final version of the nominal database is produced by combining the verified result sets and their intersections when confidence scores exceeded certain threshold values.

1 Introduction

The project started in 2000, with the aim of creating a Hungarian nominal WordNet ontology with semi-automatic methods [6]. Our basic strategy was to attach Hungarian entries of a bilingual English-Hungarian dictionary to the nominal synsets of Princeton WordNet, version 1.6 (WN, [4]), following the so-called expand approach [7]. This way, the synsets formed by the Hungarian nouns can inherit the WN semantic relations. In order to achieve this, we used heuristic methods, developed partly by previous similar projects [1,2], and partly by us, which rely on information extracted from several machine-readable dictionaries (MRDs). This approach relies on the assumption that nominal conceptual hierarchies, which describe the world, would be similar across English and Hungarian languages to a degree which is sufficient for producing a preliminary version of our WordNet.

2 Machine-Readable Dictionaries Used

We used two different MRDs to assist the heuristics which disambiguate the Hungarian nouns against Princeton WordNet synsets. The *MoBiDic* bilingual English-Hungarian electronic dictionary contains 17,700 Hungarian nominal entries, corresponding to 12,400 English equivalents covered in WordNet 1.6. These Hungarian nouns serve as the basis of the attachment procedure.

The other MRD we used is an electronic version of *the Magyar Értelmező Kéziszótár* (EKSz, [3]) monolingual explanatory dictionary. It covers over 42,000 nominal headwords, whose different senses correspond to over 64,000 different definitions. We used these definitions to gain semantic information in order to assist the heuristics that disambiguate Hungarian nouns against WN synsets via their English translations in the bilingual dictionary.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 175–180.

3 Methods

The bilingual dictionary provides 1.71 English translations on average for each Hungarian nominal headword. These English translations correspond to 2.16 WordNet synsets on average. We implemented several heuristic methods in order to accomplish the automatic disambiguation of Hungarian nouns against the candidate WN synsets.

3.1 Methods Relying on the Bilingual Dictionary

The first group of heuristics was developed by Atserias et al for the Spanish WordNet project [1]. These heuristics rely on information found in the connections between Hungarian and English words in the bilingual dictionary, and between English headwords and corresponding synsets in WN.

- MONOSEMIC METHOD: if an English headword is monosemous with respect to WN (belongs to only one synset), then the corresponding Hungarian headword is linked to the synset.
- VARIANT METHOD: if a WN synset contains two or more English words that each has only one translation to the same Hungarian word, it is linked to this synset.
- INTERSECTION METHOD: links a Hungarian headword to all synsets sharing at least two of its English translations.

A fourth kind of heuristic depends on morpho-semantic information found in the Hungarian side of the bilingual dictionary. A number of Hungarian headwords in the bilingual dictionary are endocentric (noun + noun) compounds, which have the property that the second segment of the compound defines the semantic domain of the whole word. For example, the compound *hangversenyzongora* (‘grand piano’) can be analysed as *hangverseny+zongora* (‘concert’+‘piano’), where the second segment, *zongora* serves as the DERIVATIONAL HYPERNYM noun of the compound. This piece of semantic information can be used with the modified conceptual distance formula (Section 3.2) in order to select a synset from the candidate ones.

3.2 Methods Relying on the Monolingual Explanatory Dictionary

The nominal definitions of the EKSz monolingual explanatory dictionary were POS-tagged and morphologically analyzed using the Humor analyzer [5].Using this information to recognize morpho-syntactic patterns, we were able to identify genuses, or hypernym words in 53,500 definitions, synonyms (10,500 definitions), plus holonyms (826 definitions) and meronyms (584 definitions).

Part of the acquired semantic information was used for the attachment of Hungarian nouns in the following way:

- SYNONYMS: the synset is chosen from the ones available for all the translations of the headword, which contains the greatest number of the synonyms’ English translations.
- HYPERNYMS: for those cases where both the headword and the corresponding acquired hypernym have English translations, the headword is disambiguated against WordNet using a modified version of the conceptual distance formula, developed by Atserias et al. [1], shown in Figure 1.

$$dist'(w_1, w_2) = \min_{\substack{c_{1i} \in w_1 \\ c_{2j} \in w_2 \\ depth(c_{1i}) < depth(c_{2j})}} |path(c_{1i}, c_{2j})|$$

Fig. 1. The simplified conceptual distance formula is applied to the pairs of English translations of a Hungarian noun and its hypernym. The formula returns two concepts (WN synsets) representing words which are closest to each other in the WN hypernym hierarchy

A third heuristic depends on the LATIN equivalents available for about 1,500 EKSz headwords, mostly covering various animal or plant species, taxonomic groups, diseases etc. Since WN also contains most of these Latin words in different synsets, these could be used to attach the EKSz headwords in a straightforward way.

Performance of all the individual methods relying on the bilingual and monolingual dictionaries is shown in Table 1.

Table 1. Performance of each method: number of Hungarian nouns and WN synsets covered, and number Hungarian noun-WN synset connections

Method	Hungarian nouns	WN 1.6 synsets	Connections
Mono	8 387	5 369	9 917
Intersection	2 258	2 335	3 590
Variant	164	180	180
DerivHyp + CD	1 869	1 857	2 119
EKSz synonyms	927	707	995
EKSz hypernyms + CD	5 432	6 294	9 724
EKSz Latin equivalents	1 697	838	848

3.3 Methods for Increasing Coverage

In those cases where the identified hypernyms or synonyms had no English translations, we used two methods to gain a related hypernym word that has a translation and hence can be used to disambiguate with the aid of the modified conceptual distance formula.

The first method was to look for derivational hypernyms of the synonyms or hypernyms, using the methods described above. Since hypernymy is transitive, the hypernym of the headword's hypernym (or synonym) will also be a hypernym.

The other method looks up the hypernym (or synonym) word as an EKSz an entry, and if it corresponds to only one definition (eliminating the need for sense disambiguation), then the hypernym word identified there is used, if it is available (and has English equivalents). These two methods provided a 9.2% increase in the coverage of the monolingual methods.

Table 2 summarizes the results of all the automatic methods used on different sources in the automatic attachment procedure.

Table 2. Total figures for the different types of methods

Type of Methods	Hungarian nouns	WN 1.6 synsets	Connections
Bilingual only	10 003	7 611	13 554
Monolingual	7 643	7 380	10 901
Monoling. + incr. cov. 1–2	8 343	8 199	12 185
Total	**13 948**	**12 085**	**22 169**

4 Validation and Combination of Results

In order to validate the performance of the automatic methods, we constructed a manual evaluation set consisting of 400 randomly selected Hungarian nouns from the bilingual dictionary, corresponding to 2 201 possible WN synsets through their English translations. Two annotators manually disambiguated these 400 words, which meant answering 2 201 yes-no questions asking whether a Hungarian word should be linked to a WN synset or not. Inter-annotator agreement was 84.73%. In the cases where the two annotators disagreed, a third annotator made the final verdict.

We first validated the different individual methods against the evaluation set. The results are shown in Table 3.

Table 3. Precision and recall on the evaluation set, plus coverage of all Hungarian entries in the bilingual dictionary for the individual attachment methods, in descending order of precision. The Latin method is not included, because for the most part it covers terminology not covered by the general vocabulary of the evaluation set

Method	Precision	Recall	Coverage
Variant	92.01%	50.00%	0.50%
Synonym	80.00%	39.44%	8.00%
DerivHyp	70.31%	69.09%	17.50%
Incr. cov. 1.	67.65%	46.94%	7.50%
Mono	65.15%	55.49%	69.25%
Intersection	58.56%	35.33%	17.50%
Incr. cov. 2.	58.06%	28.57%	6.00%
Hypernym	48.55%	41.71%	49.25%

Atserias et al [1] and Farreres et al [2] describe a method of manually checking the intersections of results obtained from different sources. They determined a threshold (85%) that served as an indication of which results to include in their preliminary WN. Then drawing upon the intuition that information discarded in the previous step might be valuable if it was confirmed by several sources, they checked the intersections of all pairs of the discarded result sets. This way, they were able to further increase the coverage of their WNs without decreasing the previously established confidence score of the entire set.

We used a similar approach. We decided to use two thresholds, 70% and 65%, creating the bases for two versions of the final nominal WN (*min65* and *min70*). The first set included results from the VARIANT, SYNONYM and DERIVHYP methods, the second contained these

plus the results from the INC. COV. 1 method and MONO methods. Both sets also included the results from the LATIN methods, as manual inspections estimated its precision to be fairly high (over 80%). Table 5 shows the figures for the base sets.

The next step was to validate the intersections of all the pairs of results not included in the previous step. The scores for the best-performing combinations are presented in Table 4.

Table 4. Precision, recall and coverage of intersections of sets not included in the base sets

Intersections of methods	Precision	Recall	Coverage
Inc. cov. 2. & Hypernym	95.78%	50.00%	1.50%
DerivHyp & Inc. cov. 2.	94.64%	80.03%	1.00%
DerivHyp & Intersection	92.20%	75.10%	0.75%
Inc. cov. 2. & Intersection	88.14%	90.00%	0.50%
Inc. cov. 2. & Mono	87.50%	70.00%	2.00%
DerivHyp & Mono	84.38%	87.10%	8.00%
Hypernym & Mono	71.91%	52.46%	21.00%
DerivHyp & Hypernym	70.97%	66.67%	7.25%
Hypernym & Intersection	67.86%	30.16%	6.25%

For the two final versions of the Hungarian nominal WN 1.0, we combined the min70 and min65 base sets with intersection sets having precision score over 70% and 67%, respectively (Tables 5 and 4).

Table 5. Overall results for the two versions of Hungarian nominal WordNet v1.0, with their constituting base and intersection sets

Result set	#Words	#Synsets	#Connections	Precision
min70 base	2 445	2 170	2 722	76.14%
min70 additional intersections	7 183	6 142	8 579	76.70%
min70 final set	**7 927**	**6 551**	**9 635**	**75.38%**
min65 base	12 275	11 597	20 439	65.11%
min65 additional intersections	3 110	2 698	3 431	66.91%
min65 final set	**12 839**	**12 004**	**22 169**	**63.35%**

5 Conclusions, Further Work

We used several automatic methods to attach Hungarian nominal headwords of a bilingual dictionary to WN 1.6 synsets. The various heuristics were validated against a manually disambiguated set, and from their combinations we produced two versions of the nominal database, having estimated precisions of 63 and 75 percent, with different numbers of words covered.

There are two ways to further enrich our initial nominal WN. On the one hand, to increase its coverage, we will apply the methods which proved to be most successful (VARIANT, SYNONYM, DERIVHYP) on new sources—additional bilingual dictionary modules, dictionaries with multi-word phrases, thesauri etc.

On the other hand, in order to increase the confidence of the existing result sets, a completely manual checking of the links between WN 1.6 synsets and Hungarian nouns will be necessary. This will have to rely on strict guidelines, which will be based on the pilot work disambiguating the entries in the evaluation set.

We have also applied for funding to support work on the further extension of our core Hungarian WN. This would include: revising the entire WN from a point of view independent of the English Princeton WN, adding databases for remaining other parts of speech, and connecting our WN to the EuroWordNet [8] framework.

References

1. Atserias, J., S., Climent, X., Farreres, G., Rigau, H., Rodríguez: Combining multiple methods for the automatic construction of multilingual WordNets. Proc. of Int. Conf. on Recent Advances in Natural Language Processing, Tzigov Chark (1997).
2. Farreres, X., G., Rigau, H., Rodriguez: Using WordNet for building Wordnets. Proc. of COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems, Montreal (1998).
3. Juhász, J., I., Szöke, G. O. Nagy, M. Kovalovszky (eds.): Magyar Értelmező Kéziszótár. Akadémiai Kiadó, Budapest: (1972).
4. Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, K. J. Miller: Introduction to WordNet: an on-line lexical database. Int. J. of Lexicography 3 (1990) 235–244.
5. Prószéky, Gábor: Humor: a Morphological System for Corpus Analysis. Language Resources and Language Technology, Tihany (1996) 149–158.
6. Prószéky, G. M. Miháltz: Automatism and User Interaction: Building a Hungarian WordNet. Proc. of the Third International Conference on Language Resources and Evaluation, Las Pa mas de Gran Canaria, Spain (2002).
7. Vossen, P.: Right or Wrong. Combining lexical resources in the EuroWordNet project. Proceedings of Euralex-96, Goetheborg (1996).
8. Vossen, P. (eds): EuroWordNet: A Multilingual Database with Lexical Semantic Networks, Kluwer Academic Publishers, Dordrecht (1998).

Corpus Based Validation of WordNet Using Frequency Parameters

Ivan Obradović[1], Cvetana Krstev[2], Gordana Pavlović-Lažetić[3], and Duško Vitas[3]

[1] Faculty of Mining and Geology, Email: ivano@afrodita.rcub.bg.ac.yu
[2] Faculty of Philology, Email: cvetana@matf.bg.ac.yu
[3] Faculty of Mathematics, Email: gordana@matf.bg.ac.yu, vitas@matf.bg.ac.yu
University of Belgrade

Abstract. In this paper we define a set of frequency parameters to be used in synset validation based on corpora. These parameters indicate the coverage of the corpus by wordnet literals, the importance of one sense of a literal in comparison to the others, as well as the importance of one literal in a synset in comparison to other literals in the same synset. The obtained results can be used in synset refinement, as well as in information retrieval tasks.

1 Introduction

The main goal of BalkaNet, the Balkan wordnet project (BWN) is the development of a multilingual database with wordnets for a Bulgarian, Czech, Greek, Romanian, Serbian and Turkish [5]. In its initial phase, Balkanet followed the approach similar to that of EuroWordNet (EWN) developing monolingual wordnets interconnected through an interlingual index (ILI). The development of databases started with a translation of a common set of concepts named Base Concepts in EWN, using the Princeton WordNet (PWN) as the source.

The Serbian WordNet (SWN) has been developed according to this common approach. In the absence of both an explanatory dictionary and an English/Serbian dictionary in electronic form, the translation of English synsets from PWN was done manually, while preserving the PWN semantic structure. The fact that a Serbian dictionary of synonyms does not exist even in paper form made this task even more difficult. In order to establish a relation with the reference six volume explanatory Serbian dictionary of Matica Srpska (RMS), the senses attributed in SWN to literals, or words representing sysnset lemmas in general correspond to the ones in this dictionary. Since the RMS dictionary was published in 1971, new senses had to be attributed in SWN to some of the existing literals but also new literals had to be added. Another reason for refinement of senses defined by RMS is due to the fact that concepts, and hence literal senses in PWN are far more fine grained that the ones in RMS.

The conditions under which SWN has been developed brought up the question of validation of Serbian synsets on corpora. The idea to semantically tag corpora with senses from WordNet has first been realized within the SemCor project [3]. The use of monolingual and multilingual corpora for synset validation leading to the introduction of new literals or removal of existing ones from a synset has already been tackled in [2,4]. Further refinement of this approach is presented in this paper. In order to establish more precise criteria for synset

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 181–186.

validation a set of numerical parameters related to literal-sense pair frequency in corpora has been developed.

2 Frequency Parameters

In order to evaluate the quality of a synset in terms of the comprehensiveness and adequacy of literals used for the lexicalization of a particular concept on one hand, and to establish an ordering among literals within a synset which may be used in information retrieval tasks, on the other, we define a set of indices as numerical measures of relevance of particular literals to synsets they are used in.

Let **S** be the finite set of all synsets within a wordnet: $\mathbf{S} = \{S_i | S_i$ is a synset describing a specific concept, $i = 1, 2, \ldots, N_S\}$; N_S is the total number of synsets within a wordnet. Let **L** be the finite set of all literals used as lexicalizations of one or more concepts: $\mathbf{L} = \{L_k | L_k$ is a literal used in at least one synset of the wordnet, $k = 1, 2, \ldots, N_L\}$; N_L the total number of different literals used within the wordnet. When a literal $L_k \in \mathbf{L}$ is used as a lexicalization of a specific concept described by the synset S_i, it is used in a specific sense (a sense tag is attached to the literal). Omitting the index k of the literal we shall mark all literal-sense pairs within a nonempty synset $S_i \in \mathbf{S}$ in a sequence as LS_{ij} $(j = 1, 2, \ldots, n_i)$, where $n_i \geq 1$ is the total number of literals within the nonempty synset S_i.

We shall define the indices for literals within the wordnet, with the aim to determine the relevance of a particular literal to a synset it is used in. In order to determine these indices for a literal a search is performed on a corpus and all occurrences of the selected literal as well as its inflectional forms are identified within a context of a predefined length. We shall first denote the total number of occurrences of a literal L_k within the corpus, regardless of its sense, as L_k^C. The next step is a time-consuming one since it requires manual identification of the sense in which the literal has been used in every particular concordance line identified in the corpus. When this task is completed then the number of occurrences of a literal within the corpus in each specific sense is established. For the senses covered by the wordnet, the appropriate synset S_i the literal belongs to can then be identified. We then proceed taking into account only these senses, and denote the number of times the literal L_k has been used for lexicalization of a concept described by the synset S_i as LS_{ij}^C. The sum of these numbers obtained for all possible senses of a literal covered by the wordnet yields L_k^{WN}, namely, the number of cases when a literal has been used within the corpus as a lexicalization of a concept represented in the wordnet. It is clear that $L_k^{WN} \leq L_k^C$, and that the target of each wordnet should be that for all literals, ideally, $L_k^{WN} = L_k^C$ holds. This would mean that all possible sense usages of a literal identified within the corpus have been covered by wordnet synsets.

If we want to express the relevance of a particular literal L_k to a particular synset S_i within a corpus, then we should compare the number of occurrences of this literal in the corpus denoting the concept represented by the synset S_i, that is LS_{ij}^C, to the total number of occurrences of this literal within the corpus, namely L_k^C. Thus we define the *overall synset relevance index* of a literal as the ratio of the number of times this literal has been used in a specific sense and the total number of occurrences of this literal in the corpus, namely: $I_{ik}^C = LS_{ij}^C / L_k^C$ where the literal from LS_{ij} equals the literal L_k. The index range is $0 < I_{ik}^C \leq 1$, where $I_{ik}^C = 1$ means that the literal L_k is used in one and only one sense, and that is to lexicalize the concept described by the synset S_i.

Since the wordnet coverage of the senses of a literal does not always have to be complete, we define the *wordnet synset relevance index* as the relevance of a particular literal L_k to a particular synset S_i within a more restricted part of the corpus, that is, the part already covered by the wordnet. This index is defined as the ratio of the number of times this literal has been used in a specific sense and the total number of occurrences of a literal within the corpus denoting concepts represented in the wordnet (L_k^{WN}), namely: $I_{ik}^{WN} = LS_{ij}^C / L_k^{WN}$, where the literal from LS_{ij} is the literal L_k. As is the case with I_{ik}^C, the index range is $0 < I_{ik}^{WN} \leq 1$, where $I_{ik}^{WN} = 1$ means that the literal L_k is used in one and only one sense. Since $L_k^{WN} \leq L_k^C$, then $I_{ik}^{WN} \geq I_{ik}^C$. As, ideally, $L_k^{WN} = L_k^C$ should hold for every literal, in an ideal case $I_{ik}^{WN} = I_{ik}^C$ should also be true.

In order to evaluate how close a particular literal L_k is to the ideal case, namely when all its possible senses are covered by the wordnet, we should compare the number of occurrences of a literal within the corpus denoting concepts represented in the wordnet L_k^{WN} to the total number occurrences of the literal within the corpus L_k^C. We therefore define the *wordnet coverage index* of a literal L_k, namely $I_k^{WNC} = L_k^{WN} / L_k^C$. The index ranges between 0 and 1, and in case of full coverage is equal to 1.

All previous indices evaluated the relevance of a literal to a synset regardless of possible other literals within that synset. In order to compare the relevance of a literal within a synset in comparison to other literals denoting the same concept we define the *local synset relevance index* of the literal L_k as the ratio of the number of occurrences of this literal in the corpus denoting the concept represented by the synset S_i, that is LS_{ij}^C, and S_i^C, the number of occurrences of all literals denoting this concept (i.e. belonging to synset S_i): $I_{ik}^L = LS_{ij}^C / S_i^C$, $S_i^C = \Sigma_{j=1}^{n_i} LS_{ij}^C$. It should be noted that the range of the index is $0 < I_{ik}^L \leq 1$ where $I_{ik}^L = 1$, holds when either the synset has only one literal, or other literals from that synset have not appeared in the corpus.

3 The Validation Procedure

In order to test this approach a subset of literal strings, that we called *main strings* has been chosen among those nouns and verbs that have the most senses in Serbian wordnet. Next, a subcorpus has been compiled consisting of contemporary newspaper texts comprising 1.7MW. Concordances were produced for all the inflectional forms of these nouns and verbs. In the next step all the synsets in which the main strings appear have been identified, as well as literal strings, that we called *supporting strings*, that occur beside them in these synsets. For these supporting strings concordances have also been produced. The main and supporting literal strings form the "lexical sample" as defined by the SENSEVAL project [1].

The produced concordances (around 10.000) have than been manually analyzed by lexicographers. In the first step the concordance lines containing the homograph forms have been rejected. In the remaining lines the senses have been identified according to the RMS dictionary and SWN, and marked using the same sense labels.

On the basis of the obtained results tables have been produced and the indices introduced in the section 2 calculated. These data for the noun *lice* and the verb *proizvesti* are given in Tables 1 and 2. For each of the main strings only the senses that are present in SWN are represented. The frequency of occurrence of the these senses in the corpus is given in column

Table 1. The frequency parameters for the lemma *lice* obtained on newspaper corpus

Synset	**lice**	LS_{ij}^C	uloga:1a	lik:3	strana:1b	S_i^C	I_{ik}^C	I_{ik}^{WN}	I_{ik}^L
face, human face	1a	33	*	*	*	33	0.063	0.085	1.000
face:6	2a	353	*	*	*	353	0.675	0.912	1.000
character:4, role:2, theatrical role:1,. . .,	2b	1	34	3	*	38	0.002	0.003	0.026
face:14	3	0	*	*	*	0	0.000	0.000	0.000
side:5,	5a	0	*	*	5	5	0.000	0.000	0.000
	L_k^{WN}	387							
	other	136					0.260	*	*
	L_k^C	523	298	20	861				
	I_k^{WNC}	0.740	I_{ik}^C 0.114	I_{ik}^C 0.150	I_{ik}^C 0.006				
			I_{ik}^L 0.895	I_{ik}^L 0.079	I_{ik}^L 1.000				

LS_{ij}^C. The row L_k^{WN} represents the frequency of the occurrence of all the senses of a string that are covered by SWN, while the row **other** represents the frequency of the occurrence of those senses that are not yet covered. L_k^C is the sum of these two data, and represents the total frequency of the occurrence of the main string, while the index I_k^{WNC} represents their ratio. Among 12 main strings that have been analyzed, three had the value of this index 1, which means that for these strings all the senses identifi ed in RMS dictionary (and perhaps some more) have been included in SWN. For all analyzed literals this index ranges from 0.246 to 1.

Table 2. The frequency indices for the lemma *proizvesti* obtained on newspaper corpus

Synset	**proizvesti**	LS_{ij}^C	prouzrokovati:1	potaknuti:2x	iznedriti:1	proizvoditi:3	napraviti:1a	S_i^C	I_{ik}^C	I_{ik}^{WN}	I_{ik}^L
produce:3,. . .	1a	6	31	1	*	*	*	38	0.090	0.091	0.158
yield:1, give:2,. . .	1b	1	*	*	0	*	*	1	0.015	0.015	1.000
produce:2, make:6,. . .	3	59	*	*	*	106	21	186	0.881	0.894	0.317
	L_k^{WN}	66									
	other	1							0.015	*	*
	L_k^C	67	31	1	99	114	159				
	I_k^{WNC}	0.985	I_{ik}^C 1.000	I_{ik}^C 1.000	I_{ik}^C 0.000	I_{ik}^L 0.930	I_{ik}^L 0.132				
			I_{ik}^L 0.816	I_{ik}^L 0.026	I_{ik}^L 0.000	I_{ik}^L 0.570	I_{ik}^L 0.113				

The parameter S_i^C gives the overall occurrence of the synset, that is all its literals, in the corpus. The indices I_{ik}^C, I_{ik}^{WN}, and I_{ik}^L in the upper part of the table refer to the main string, while the same indices in the lower part refer to the appropriate supporting strings. The first one is the ratio LS_{ij}^C/L_k^C: for instance, for the sense 1a of the main string *lice*, this index is 0.063, which means that this sense represents 6.3% of all the occurrences of this string in corpus. The second index is the ratio LS_{ij}^C/L_k^{WN}. For the same sense of the string *lice* its value is 0.085 meaning that it covers 8.5% of all the occurrences that represent senses from SWN. Finally, the third index is the ratio LS_{ij}^C/S_i^C. For the sense 2a of the string *lice* the value of this index is 0.026, meaning that of all occurrences of this synset, 2.6% were represented by this particular literal.

If for some string the value of its index I_{ik}^L is close to 0 it can indicate that it has been misplaced in the synset, especially in the cases when both indices L_k^C and S_i^C are considerably greater than 0. For instance, that is the case for the string *napraviti:1a* (Table 2). The string *napraviti* has a considerable frequency on corpus ($L_k^C = 159$), and the synset to which the literal string *napraviti:1a* belongs also has a considerable frequency ($S_i^C = 186$). However, its local synset relevance index is relatively low ($I_{ik}^L = 0.113$), and the synonymy of the literal string *napraviti:1a* with the main string *proizvesti* should be reconsidered.

The calculated indices enable the ordering of the literal strings in a synset. This can be useful for information retrieval (IR) tasks that are seen as one of the most interesting applications of BWN. Especially, strings that have a low value of I^L and a high value of I^C and which are not necessarily misplaced in a synset, can be neglected in IR tasks, thus reducing the recall but improving the precision.

Table 3. The frequency parameters for the lemma *lice* obtained on literary corpus

Synset	**lice**	LS_{ij}^C	uloga:1a	lik:3	strana:1b	S_i^C	I_{ik}^C	I_{ik}^{WN}	I_{ik}^L
face, human face	1a	380	*	*	*	380	0.936	0.979	1.000
face:6	2a	3	*	*	*	3	0.007	0.008	1.000
character:4, role:2,	2b	3	6	1	*	10	0.007	0.008	0.300
face:14	3	0	*	*	*	0	0.000	0.000	0.000
side:5,	5a	2	*	*	4	6	0.005	0.005	0.333
	L_k^{WN}	388							
	other	18					0.044	*	*
	L_k^C	406	22	25	287				
	I_k^{WNC}	0.956	I_{ik}^C 0.273	I_{ik}^C 0.040	I_{ik}^C 0.014				
			I_{ik}^L 0.600	I_{ik}^L 0.100	I_{ik}^L 0.667				

In order to test the impact of the nature of the corpus to index values the validation procedure was performed on a small literary corpus of 0.5 MW for aselected number of literals. The results obtained show that the index values can be largely affected by the nature of the corpus. Thus, for example, the values of both I_{ik}^C and I_{ik}^{WN} have dramatically changed for senses 1a and 2a of the noun *lice* (Table 3). This does not come as too much of a surprise

since meaning 2a ("A part of a person that is used to refer to a person") is more used in newspaper texts whereas the meaning 1a ("The front of the human head…") in literature. The changes seem to be far less dramatic for the indices I_{ik}^{L}, but in order to draw some final conclusions the literals should be tested on a larger corpus.

4 Conclusion

The applied procedure confirmed the importance of the validation of synsets on a corpus. The adequacy of placement of each literal and its sense in a synset can not be fully assessed without analyzing its appearances in the concordance lines. The frequency indices can serve as useful numerical indicators in this assessment procedure. However, to get a fair estimate of a literal in terms of these parameters, the procedure needs to be applied on a large and balanced corpus. To that end automatic or/and semi-automatic procedures need to be developed in order to alleviate the time-consuming task of manual concordance analysis.

References

1. Kilgarriff, A. and Rosenzweig, J.: English SENSEVAL: Report and Results. In: *Proc. of LREC*, Athens, May–June (2000).
2. Krstev, C., et al.: Corpora Issues in Validation of Serbian Wordnet. In: *Proc. of the Conference "Text, Speech, and Dialogue"*, Springer LNCS, 138–145 (2003).
3. Miller, G. A., et al.: Using a semantic concordance for sense identification. In: *Proc. of the ARPA Human Language Technology Workshop*, 240–243 (1994).
4. Obradović, I., et al.: Application of Intex in Refinement and Validation of Serbian Wordnet. *6th Intex Workshop*, 28–30th May, Sofia (2003).
5. Stamou, S., et al.: BALKANET: A Multilingual Semantic Network for Balkan Languages, *Proc of 1st Global WordNet Conference*, Mysore, India (2002).

Language to Logic Translation with PhraseBank

Adam Pease[1] and Christiane Fellbaum[2]

[1] Articulate Software Inc
278 Monroe Dr. #30, Mountain View, CA 94040
Email: adampease@earthlink.net
[2] Princeton University
Department of Psychology, Green Hall, Princeton, NJ 08544
Email: fellbaum@princeton.edu

Abstract. We discuss a restricted natural language understanding system and a proposed extension to it, which is a corpus of phrases. The Controlled English to Logic Translation (CELT) system allows users to make statements in a domain-independent, restricted English grammar that have a clear formal semantics and that are amenable to machine processing. CELT needs a large amount of linguistic and semantic knowledge. It is currently coupled with the Suggested Upper Merged Ontology, which has been mapped by hand to WordNet 1.6. We propose work on a new corpus of phrases (called PhraseBank) to be added to WordNet and linked to SUMO, which will catalog common English phrase forms, and their deep meaning in terms of the formal ontology. This addition should significantly expand the coverage and usefulness of CELT.

1 Introduction

We first discuss the existing components which make up the Controlled English to Logic Translation system, including its formal ontology and lexicon. We then describe CELT itself. The body of the paper discusses the PhraseBank effort and how it should improve the utility of CELT.

1.1 Upper Ontology

The Suggested Upper Merged Ontology (SUMO) (Niles&Pease, 2001) is a free, formal ontology of about 1000 terms and 4000 definitional statements. It is provided in first order logic, and also translated into the DAML semantic web language. It is now in its 56th version; having undergone three years of development, review by a community of hundreds of people, and application in expert reasoning and linguistics. SUMO has been subjected to formal verification with an automated theorem prover. It has also been mapped to all 100,000 noun, verb, adjective and adverb word senses in WordNet, which not only acts as a check on coverage and completeness, but also provides a basis for application to natural language understanding tasks. SUMO covers areas of knowledge such as temporal and spatial representation, units and measures, processes, events, actions, and obligations. Domain specific ontologies have been created that extend and reuse SUMO in the areas of finance and investment, country almanac information, terrain modeling, distributed computing, endangered languages description, biological viruses, engineering devices, weather and a

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 187–192.

number of military applications including terrorist events, army battlefield planning and air force mission planning. It is important to note that each of these ontologies employs rules. These formal descriptions make explicit the meaning of each of the terms in the ontology, unlike a simple taxonomy, or controlled keyword list.

SUMO has natural language generation templates and a multi-lingual lexicon that allows statements in KIF and SUMO to be expressed in multiple natural languages (Sevcenko, 2002). These include English, German, Czech, Italian, Hindi (Western character set) and Chinese (traditional characters and pinyin). A Tagalog lexicon is under development. Automatic translations can be viewed on line at `http://virtual.cvut.cz/kifb/en/`.

1.2 Restricted Natural Language

The Controlled English to Logic Translation (CELT) (Pease&Murray, 2003) (Murray et al, 2003) system performs syntactic and semantic analysis on restricted natural language input, and transforms it first order logic in Knowledge Interchange Format (KIF) syntax (Genesereth, 1991). The terms in the resulting KIF expressions come from the SUMO. This mapping of WordNet synsets to the ontology provides a deeper semantic analysis of the terms than what can be provided by a lexicon alone. A lexicon provides basic information, much like a dictionary. SUMO provides information about the term's concepts, attributes, and relationships.

CELT can perform active reasoning (via its associate inference engine) to derive answers that are not explicitly stated in the knowledge base. The knowledge is represented in domain knowledge bases (specified domain information), and a mid-level (more general domain information) and upper-level ontology (common sense concepts, world knowledge). The advantage of a tiered, modular knowledge structure is that it is efficient and reusable.

The user asks queries and makes assertions to CELT in a specified grammatical format. This subset of English grammar is still quite extensive and expressive. The advantage of the controlled English is that when the grammar and interpretation rules are restricted, then every sentence in the grammar has a unique parse. This eliminates the problems of ambiguity with other parsing approaches that would result in retrieving non-appropriate answers. For further discussion of controlled English grammars and applications, see Sowa (1999).

To overcome some of the limitations of CELT syntax, such as only handling indicative verbs and singular nouns, we developed other methods to extend its coverage. We use morphological processing rules, derived from the "Morphy" code of WordNet, to transform other verb tenses and plural verbs into the various tenses and numbers required. Discourse Representation Structures (DRSs) (Kamp & Reyle, 1993) handle context to resolve anaphoric references, implications, and conjunctions.

CELT does not limit the parts of speech or the number of word senses a word can have. Nor is the number of words limited. More importantly, CELT is not a domain specific system. It is a completely general language which can be specialized and extended for particular domains along with domain specific vocabulary. WordNet is being leveraged to provide core coverage of common English words. Currently we have about 100,000 words senses in our system. Individual words are identified based on the parse and lexicon.

2 Phrases in Language Understanding

Much of current NLP work, including part of speech and semantic tagging, focuses on language at the word level. But statistics show that speakers do not compose messages by freely combining words according to the rules of syntax and morphology. Much of language is composed of chunks or phrases, where specific lexical items co-occur in set patterns (Mul'cuk, 1998) The most frequent verbs in English (based on the Brown Corpus statistics) include "have," "do," "make," "take," and "give." These verbs also are among the most polysemous and their meanings are represented by dozens of distinct senses in lexical resources, including WordNet. Clearly, they represent a challenge for any natural language processing application. One type of phrase are verb-noun chunks involves so-called "light" or "support" verbs (Church&Hanks, 1990), such as "have a shock," "do the laundry," "make a face," and "give birth (to)." Thus, "take" occurs most frequently not in what might be called its primary sense, roughly paraphrasable as "get hold of with one's hands," but in a collocations like "take walk" or "take a hit." Other examples are "have a shock," "do the laundry," "do lunch," "make a face," "make progress," "give birth (to)," "give grief (to)." These phrases are characterizable by two properties. First, the noun carries most of the semantic weight, with the verb providing relatively little information. Second, the verb phrase is often roughly synonymous with a simple verb that is morphologically related to the noun: "do/have lunch-lunch," "take a walk-walk," "make progress-progress," etc.

Other examples are verb phrases like "pay attention/heed/homage," which require the particular choice of a verb in a sense that is specific to these phrases. English has hundreds or perhaps thousands of such phrases. The author of a large-scale study of the uses of "take" (Church&Hanks, 1989) estimates that there are at least 10 000 phrases that follow the pattern "support verb plus noun". The focus of our proposed work is on such phrases and phrase patterns. We believe that the automatic processing of natural language queries and answers will be greatly enhanced in an approach that considers chunks and phrases.

CELT first classifies phrases and identifies the patterns according to which they are composed and which define their meanings. In the current system, the corpus of phrases is quite limited, numbering only a few dozen. After having been parsed, the words in the frame-slot representation can be disambiguated against WordNet. Currently, the disambiguation of a polysemous word is performed by selecting the first sense of that word in WordNet, which displays the senses in the order determined by the frequency with which they were annotated to tokens in the Brown Corpus (Francis and Kucera, 1964) Miller et al. found that selecting the most frequent sense yields an accuracy rate of 65% (Miller et al, 1993). This method is clearly not good enough for reliable disambiguation. Moreover, the tagging effort was limited to a small number of words, covering a thematically unbalanced subset of the Brown Corpus. A reliable system must include more accurate lexical disambiguation.

By classifying phrases and establishing phrase patterns according to their semantics, we can match the component words of the phrases to WordNet entries with a very high degree of accuracy. For example, our classification will permit us to state with high degree of confidence that the sense of the verb "make" in a context where the parser has identified the word "trouble" as its direct object must be assigned sense 3 in WordNet: verb.creation: make, create (make or cause to be or become; "make a mess in one's office"; "create a furor"). The phrases will also be matched to template logical forms, allowing CELT to output a range

of logic statements that more precisely capture the semantics of the sentence than would be otherwise possible by looking only at word senses and the syntactic parse.

One possible straightforward solution for the automatic processing of such phrases would be to ignore the light verb and treat the noun as the related verb. Thus, “take a walk” would be interpreted as “walk,” and “give birth” as “birthe.” But this turns out not to be an acceptable approach. First of all, the verbs are often polysemous, and the system would have to decide which sense to associate with the noun in such phrases. Second, to understand a text, a system needs to analyze the syntactic relations among sentence constituents, to, to put it simply, to understand “who does what to whom.” While the subject in both the phrases “take a walk,” “have lunch,” and “give birth” and in the corresponding verbs “walk,” “lunch,” and “birthe” is the Agent of the event, this is not the case in superfi cially similar phrases like “take a hit” and “have a shock” where the subject is the Undergoer, or Patient, in the event, and does not play the same semantic role (Agent, Stimulus) as the subject of “hit” and “shock.” A system that ignores the light verb and equates the noun with the related verb would seriously misinterpret the text in such cases.

Moreover, some phrases include the the same noun, but different verbs: “do lunch/have/take lunch,” “take/give a break (to).” In the fi rst case, the meaning difference is subtle (“do” implying a social event), whereas in the second, the meaning of the two phrases is entirely unrelated.

A second solution would be to treat the entire phrase as a lexical unit. In fact, the lexical status of phrases like “take a walk” is unclear. On the one hand, they are partly compositional; one might argue that “take” in “take a vacation,” “take a walk,” and “take lunch” has an independent meaning and denotes the participation in an event. On the other hand, the phrases are idiosyncratic collocations: why do we say “make a decision” and not “take a decision” and why is it “take a photo” and not “make a photo” (as in French)? The restrictions on such phrases have to learned and stored in speaker’s mental lexicons.

But treating these phrases as a unit is not unproblematic for language processing. First, the lexicon would have to be augmented with a very large number of phrases; some of the patterns are in fact productive. More seriously, the parser would need to recognize the verb and the noun as a unit in all and only all the relevant cases so as to match it against the lexicon entry. This can be diffi cult in cases where the verb and the noun are not adjacent and do not conform to the lexicon entry, as in “take a long walk” or “inappropriate remarks were made.”

Instead, we propose an approach that avoids these problems. We classify light verb phrases and light verb phrase patterns semantically. For example, we collect phrases like “have a shock” and “have a surprise,” distinguishing them from superfi cially similar phrases like “have dinner” and “have a nap.” In the fi rst case, the verb means “experience” (currently WordNet sense 11) and selects for a mental or emotional state. The subject is an Experiencer, and the event is a punctual achievement (Vendler 1967, Dowty 1991) In the second case, the phrases denote activities or processes and “have” here means roughly “partake of” or “engage in” (there’s currently no corresponding WordNet sense).

Actually, there is some kind of mutual selection of specifi c senses, (or co-composition, in Pustejovsky’s sense). Not only the verb, but the noun, too, is polysemous. For example, nouns like “dinner” and “nap” exhibit systematic polysemy between a process/activity and a result/product reading. (Cf: dinner lasted 3 hours=activity; dinner was on the table=product.)

So the question is, for each of the phrases, which noun reading do we get with which verb? In other words, the goal is not only to disambiguate the verb but also the noun.

WordNet generally does not include collocations or phrasemes like “make a remark” and “take a walk,” because the lexemes in WordNet’s synsets should be treatable as units by NLP systems. But a system that considers “make a remark” as internally unmodifiable will have problems dealing with tokens like “make a nasty remark” or “remarks were made.”

We first plan to collect a large number of phrasemes like “make a remark,” “take a walk,” and “have a surprise.” Next, we classify the expressions in terms of their semantics. For example, in “make a remark/comment/point/joke,” the object nouns denote a linguistic expression, whereas in “make a mistake/blunder/error/faux pas” the noun denotes a kind behavior. The verbs in these phraseme classes have a different semantics, too. In “make a comment/joke” etc. the verb means “create mentally,” whereas in “make a mistake/blunder” etc. “make” means “commit” or “perform.” In phrases like “have a surprise/shock/...,” the verb means “suffer” or “undergo,” and the noun denotes a mental state or feeling. The full semantics of the phrase will be expressed in a template logical expression in KIF and using SUMO terms. Spaces in the template will be left to fill in with the contents of slots in the parse frame. As a simplified example, “John takes a walk.” would be parsed into a frame like [John, subject][takes a walk, VP template 547] which would be keyed to a logical template below left, which would be filled in with the results of the parse and combined with the logical output of word-level interpretation to yield the logic expression at below right

```
(exists (?walk <subject>)
  (and
    (instance ?walk Walking)
    (agent ?walk <subject>)))
```

```
(exists (?walk ?john)
  (and
    (instance ?walk Walking)
    (instance ?john Human)
    (names ‘‘John’’ ?john)
    (agent ?walk ?john)))
```

References

1. Church, K. W. and Hanks, P., (1990). Word association norms, mutual information and lexicography. Computational Linguistics, 16(1):22–29.
2. Dowty, D., 1991: Thematic Proto-Roles and Argument Selection, Language 67, 547–619.
3. Fellbaum, C. (Ed.). (1998). WordNet: An electronic lexical database (language, speech, and communication). Cambridge, MA: MIT Press.
4. Francis, W., and Kucera, H., (1964). Brown Corpus Manual. Revised 1979. Available at `http://www.hit.uib.no/icame/brown/bcm.html`.
5. Genesereth, M., (1991). “Knowledge Interchange Format”, In Proceedings of the Second International Conference on the Principles of Knowledge Representation and Reasoning, Allen, J., Fikes, R., Sandewall, E. (eds), Morgan Kaufman Publishers, pp. 238–249.
6. Kamp, H. & Reyle, U. (1993). From discourse to logic. New York: Kluwer Academic Publishers.
7. Melc’uk, I., (1998). Collocations and Lexical Functions. In: Cowie, Ed. 23–53.
8. Miller, G., (1995). WordNet: A Lexical Database for English. Communications of the ACM, Vol. 38 No. 11, 39–41.
9. Miller, G., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. (1993). “Introduction to WordNet: An On-line Lexical Database.”.

10. Murray, W. R., Pease, A., and Sams, M. (2003). Applying Formal Methods and Representations in a Natural Language Tutor to Teach Tactical Reasoning. 11th International Conference on Artificial Intelligence in Education (AIED) conference in Sydney. pp 349–356. IOS Publications.
11. Niles, I. & Pease A., (2001). "Towards A Standard Upper Ontology." In Proceedings of Formal Ontology in Information Systems (FOIS 2001), October 17–19, Ogunquit, Maine, USA, pp. 2–9. See also `http://ontology.teknowledge.com`.
12. Niles, I., & Pease, A., (2003). Mapping WordNet to the SUMO Ontology. Proceedings of the IEEE International Knowledge Engineering conference, Las Vegas, NV, June 23–26.
13. Pease, A., and Murray, W., (2003). An English to Logic Translator for Ontology-based Knowledge Representation Languages. In Proceedings of the 2003 IEEE International Conference on Natural Language Processing and Knowledge Engineering, Beijing, China.
14. Sevcenko, M.: Online Presentation of an Upper Ontology, In: Proceedings of Znalosti 2003, 19–21 February 2003, Ostrava, Czech Republic.
15. Sowa, J. F.. (1999). Controlled English.
Available at `http://users.bestweb.net/~sowa/misc/ace.htm`.
16. Vendler, Z., 1967: Verbs and Times, in Linguistics in Philosophy, Cornell University Press, Ithaca, NY.

Extending the Italian WordNet with the Specialized Language of the Maritime Domain

Adriana Roventini and Rita Marinelli

Istituto di Linguistica Computazionale, Consiglio Nazionale delle Ricerche,
via Moruzzi 1, Pisa, Italy
Email: adriana.roventini@ilc.cnr.it, rita.marinelli@ilc.cnr.it

Abstract. In this paper we describe the creation, we are carrying out of a specialized lexicon belonging to the maritime domain (including the technical and commercial/maritime transport domain) and the link of this lexicon to the generic one of the ItalWordNet lexical database. The main characteristics of the lexical semantic database and the specific features of the specialized language are described together with the coding performed according to the ItalWordNet semantic relations model and the approach adopted to connect the terminological database to the generic one. Some of the problems encountered and a few expected advantages are also considered.

1 Introduction

The growing amount of non-structured information, stored in natural language, requires the availability of computational instruments able to handle this type of information. In this context, the extension of the ItalWordNet (henceforth IWN) database with the navigation and the shipping terminology, constitutes an important enrichment, given the remarkable incidence of this lexical domain in many contexts of everyday and business life; in its turn, the specialized lexicon gains semantic information automatically manageable, as well as the link to WordNet 1.5.

The globalisation of trade, business and travel, alongside technology development are producing changes also within the maritime activity and the related terminology; consequently the techniques of communication, translation and diffusion of terms have also changed. Historical reasons and, most of all, the introduction of industrial techniques and logistic procedures, originated and developed in Anglo-Saxon countries, in the field of transport have led to a kind of 'monopole' of the English language in this sector of economy. Furthermore, the great importance of transports, together with the continuous technical progress, have determined the need – for the countries involved in the transportation network – to introduce reliable tools to manage the ever-increasing new English technical terminology, in an attempt to avoid the far too easy attitude to simply introduce new English terms as neologisms in the national languages.

The Italian lexical-semantic database IWN (Roventini et al., 2002), contains encoded detailed information of a semantic and conceptual type according to a multidimensional model of meaning which is particularly useful for applications dealing with textual content. Within the IWN database, lexical information is represented in such a way as to be used by different computational systems in many types of applications. Therefore, we have

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 193–198.

considered it useful to take advantage of the IWN linguistic model to build and structure the specialized language of navigation and maritime transport, aware that "Adopting the perspective of linguistics to account for terms, requires their description by means of the same models that we use for other lexical units. " (Cabré, 1998/99).

In the following sections we describe: the main features of the IWN database (Section 2), the construction of the terminological subset (Section 3), the foreseen advantages and improvements (Section 4).

2 The Italian WordNet

IWN is a lexical-semantic database developed within the framework of two different research projects: EuroWordNet (Vossen 1999) and SI-TAL (Integrated System for the Automatic Treatment of Language) a National Project devoted to the creation of large linguistic resources and software tools for the processing of written and spoken Italian. During the SI-TAL project the Italian WordNet was improved and extended by the insertion of adjectives, adverbs and a set of proper names belonging to both the geographic and human domains. Moreover, a terminological wordnet was added for the economic and financial domain, in such a way that it was possible to access both the generic lexicon in the database and the specialized one, or also both lexicons at the same time (Roventini et al., 2000, Magnini & Speranza 2001).

IWN inherited the EWN linguistic model (Alonge et al., 1998) which provides a rich set of semantic relations, and the first nucleus of data (verbs and nouns). The wordnet was structured in the same way as the Princeton WordNet (Miller et al., 1990, Fellbaum 1998) around the notion of synset (i.e. a set of synonymous word meanings), but many other semantic relations between the synsets were identified and extensively (e.g. the hyponymy or IS-A relation) or partially encoded; among these the cross-Part of Speech (PoS) relations between words referring to similar concepts and belonging to the same semantic order: for example the noun *ricerca* (research) and the verb *ricercare* (to research), which indicate the same situation or eventuality, are linked by a xpos_near_synonym relation.

IWN has also inherited from EWN the distinction between language-internal relations and equivalence relations and the Top Ontology. The language internal relations apply between synsets of the Italian wordnet, among which the hyperonymy/hyponymy relation is the most important relation encoded for nouns and verbs together with synonymy and xpos_near_synonym. This is due to the possibility it provides to identify classes of words for which one can draw generalizations and inferences. The equivalence relations between the IWN synsets and the Inter-Lingual Index (ILI)[1] are defined similarly to the internal relations. Thus, for instance, synonymy and eq_synonymy can be defined in a similar way, the only difference being that the latter holds between a synset in the Italian wordnet and a synset in the ILI. The Top Ontology (TO) is a hierarchy of language-independent concepts, reflecting fundamental semantic distinctions, built within EWN to provide a common framework for the most important concepts and partially modified in IWN to account for adjectives and adverbs. Via the ILI, all the concepts in the generic and specific wordnet are directly or indirectly linked to the TO.

[1] The ILI is a separate language independent module containing all WN1.5 synsets but not the relations among them.

3 Construction of the Terminological Wordnet

The maritime terminological lexicon has been structured according to the design principles of the generic wordnet, i.e. applying the same semantic relations model and exploiting the possibility – available in IWN through the ILI – of linking the specialized terms to the corresponding closest concepts in English and, consequently, to the EuroWordNet multilingual lexical database.

First of all, with the suggestions of a domain expert and consulting various sources[2] we started to design the terminological data base top level, identifying the most relevant and representative domain concepts or basic concepts (henceforth BCs). The choice of these BCs was carried out following various criteria, but in particular we selected the concepts that in both the generic database and the specialized dictionaries show a large number of hyponyms, and/or that are more frequently used in this particular domain of maritime navigation and transport (Marinelli et al., 2003).

A first nucleus of over 150 BCs was identified, such as *nave* (ship), *vela* (sail), *porto* (harbour) *ormeggio* (mooring), *carico* (cargo), *spedizione* (shipment), *navigazione* (navigation), *trasporto* (transport), *tariffa* (tariff), *nolo* (freight) and so on, which are sufficiently general and constitute the root nodes of the specialized database we are developing. Most of these BCs were exported from the generic database and then imported in the terminological one exploiting the export/import capabilities of the IWN management tool. It is possible, in fact, to import or export one or more concepts as XML files. As a next step all these BCs were linked to the generic wordnet by means of the plug_in relations (see the following paragraph). Other BCs were included "ex novo", because they were not present with their maritime senses in the generic database, but very frequently used and representative of this specific domain, e.g.: *classe* (class), *fanale* (light), *armare* (to equip), *agente marittimo* (shipping agent), *punto* (position), *destino* (destination).

Starting from this first nucleus the database has then been increased, by coding the hyponyms and codifying other important semantic relations.

Most BCs are the root of a terminological sub-hierarchy and their hyponyms are often constituted by the base concept term itself followed by an adjective or a prepositional phrase which narrows and at same time specifies the meaning, a typical new-words formation that is particularly frequent in specialized languages. For instance considering the BCs *carico* (cargo), *tariffa* (tariff), *nolo* (freight) the following compounds or multiwords were encoded: *carico completo* (full cargo), *carico di merci varie* (general cargo), *carico in coperta* (deck cargo), *carico parziale* (part load cargo), *tariffa doganale* (custom tariff), *tariffa di trasporto* (transport tariff), *tariffa forfettaria* (flat-rate tariff), *nolo anticipato* (freight prepaid), *nolo intero* (full freight), *nolo secondo il valore* (ad valorem freight), *nolo a destino* (freight payable at destination).

Terms belonging to all the different grammatical categories of nouns, verbs, adjectives, adverbs and a small set of proper names are being codified in the terminological data base

[2] Several information sources have been used to select the BC: the "Dizionario Globale dei termini marinareschi", edited by the *Capitaneria del Porto di Livorno*, online on the Web; the "Dizionario di marina", edited by Barberi Squarotti G., Gallinaro I, (2002); the "Glossario dello spedizioniere" (Annuario Federspedi 1988); the "Dizionario di termini marittimi mercatili", compiled by P. R. Brodie and translated by E. Vincenzini, Lloyd's of London Press, Legal Publishing and Conferences Division, 1988.

(until now 2000 lemmas), using the many types of IWN semantic relations. The BC *porto* (harbour), for instance, is linked to *luogo* (place), by means of the hyperonymy relation; it is also connected to *imbarco* (shipment) and *sbarco* (unloading) by a role_location relation, to *avamporto* (outer harbour) by a has_mero_location relation, to the adjective *portuale* (harbour) by the has_pertained relation, to a set of proper names by the has_instance relation.

Each term is connected with the ILI by an equivalence relation: when possible an eq_synonym or eq_near_synonym relation is used, otherwise an eq_has_hyperonym relation is coded, e.g. *porto* eq_synonym harbour, *carico parziale* eq_has_hyperonym cargo; by these links to the ILI, the terms are also connected to the TO.

When the English synonym of the term was not found in the ILI and the term was linked to its hyperonym, the English synonym of the term was recorded in a list by which the ILI should be updated and enlarged. A feasibility study is envisaged with this aim.

The English term or multiword (or its acronym) is often known and used much more than the Italian one in the maritime transport activity: for instance the abbreviation RO-RO (Roll On/Roll Off) usually indicates *nave traghetto per automezzi* (ferry for vehicles transport), the abbreviation FOB (Free On Board) is used to say *con le spese pagate fino a bordo*, (loading costs paid up to ship's broadside), CIF (Cost Insurance and Freight) to say *costi fino a bordo più assicurazione e nolo mare pagati* (loading costs, insurance and sea-freight prepaid). In these and in many similar cases, we included in the synset both the English term (or multiword or acronym) and the Italian one as variants.

3.1 The Link Structure

As said before, the BCs identified for this terminological lexicon constitute the top level and are the root nodes for the plug-in operation which allows linking between the generic and specialized wordnets.

The database management tool has the following main functions: i) a simultaneous parallel consultation of the two databases to facilitate insertion of the relations; ii) three types of plug_in relations can link synsets of the two different databases: the eq-plug-in relation, as equivalence synonymy relation, the hyp-plug-in relation, as equivalence hyperonymy or hyponymy relation; iii) an integrated research between the two databases in such a way that if the synset is found in both the databases and there is an eq-plug-in relation between the synsets, the synset belonging to the specific domain partially eclipses the generic one.

As a matter of fact, once defined, a 'plug-in' relation connects a terminological sub-hierarchy (represented by its root node) to a node of the generic wordnet, so that all downward (hyponymy and istances) and horizontal (such as part-of relations, role relations, cause relations, derivation, etc.) relations are taken from the terminological wordnet, while all upward (hyperonymy) relations are taken from the generic one.

If the lemma is retrieved in both databases and there is not a eq-plug-in relation between the synsets, the synset belonging to the specific domain does not eclipse the other one and the results of the research are presented all together.

4 Final Remarks

Our choice to perform this type of study was determined by the fact that nowadays maritime terminology is object of great interest in a marine nation like Italy; furthermore, maritime

terminology dictionaries are rare and sometimes it is very difficult to find the English translation of these terms or, on the contrary, the English terms prevail over the Italian synonyms, in particular as far as maritime transport is concerned.

The availability of definitions and translations of specific terms is a useful tool for work (export-import companies, maritime agencies, etc.), for school and for didactic activities of various types (nautical Institutes, professional training, etc.) and, in general, whenever a reference to terms of this specific domain is needed.

The sea transport field is managed by English terminology, but in everyday life a constantly updated translation is necessary, on many particular occasions. From a 'commercial' point of view, the English language prevails over all other languages: contracts, negotiations, chartering and operation documents of cargo ships (bills of lading, etc.) are in English, and so are a great number of reference books. From the point of view of 'usefulness', there are circumstances in which it is necessary to refer to a translation of technical terms that is correct, abreast and absolutely unambiguous. This is for example the case of legal actions, when a judge is faced with English terminology, the Italian translation is very often difficult or unknown, and, at the same time, he is forced to refer strictly to the Italian Navigation Code written in Italian.

In this context, we think it would be desirable to carry on with this work, increasing the number of terms and starting a cooperation with the concerned organizations[3] in order to enrich and refine this maritime navigation and transport lexicon and reach a definite version officially recognized and validated, which could be greatly useful in many future activities. Furthermore, we believe that the link between the specialized wordnet and WN1.5, through the IWN generic lexicon, is essential both to face globalization and to maintain our linguistic identity.

References

1. Alonge, A., Calzolari, N., Vosse, P., Bloksma, L., Castellon, I., Marti, T., Peters, W.: The Linguistic Design of the EuroWordNet Database, Special Issue on EuroWordNet, in: N. Ide, D. Greenstein, P. Vossen (eds.), "Computers and the Humanities", XXXII (1998), 2–3, 91–115.
2. Cabré, M. T., Do we need an autonomous theory of terms?, in: "Terminology", vol. 5, n. 1, (1998/1999), pp. 5–19.
3. Dizionario Globale dei termini marinareschi, edited by the "Capitaneria del Porto di Livorno", online `http://www.capitanerialivorno.portnet.it/Dizionario/`.
4. Dizionario di Marina medievale e moderno della Reale Accademia d'Italia, Roma, 1937.
5. Fellbaum, C. ed.: WordNet: An Electronic Lexical Database, MIT Press, Cambridge, MA, (1998).
6. Layton, C. W. T. Dictionary of nautical words and terms, Glasgow, 1958.
7. Magnini, B., Speranza, M. Integrating Generic and Specialized Wordnets, in: Proceedings of Recent Advances in Natural Language Processing, RANLP-2001, Tzigov Chark, Bulgaria, 2001, pp. 149–153.
8. Marinelli, R., Roventini, A., Spadoni, G.: Linking a subset of Maritime Terminology to the Italian WordNet in: Proceedings of the Third International Conference on Maritime Terminology, Lisbon, 2003.

[3] For example organizations such as Confitarma/Associazione Armatori Italiani, Federagenti/Federazione Agenti Marittimi, Federspedi/Federazione Spedizionieri, Assologistica/Associazione dei Terminal e Imprese portuali, Assoporti/Associazione delle Autorità Portuali Italiane.

9. Miller, G., Beckwith, R., Fellbaum, C., Gross, D., Miller, K. J. (1990) Introduction to WordNet: An On-Line Database, in: "International Journal of Lexicography", 3(4), pp. 235–244.
10. Roventini, A., Alonge, A., Bertagna, F., Calzolari, N., Magnini, B., Marinelli, R., Speranza, M., Zampolli, A.: ItalWordNet: a large semantic database for the Automatic Treatment of the Italian Language in: Proceedings of the First Global WordNet Conference, Central Institute of Indian Languages, Mysore, India, 2002, pp. 1–11.
11. Roventini, A., Alonge, A., Bertagna, F., Calzolari, N., Cancila, J., Girardi, C., Magnini, B., Marinelli, R., Speranza, M., Zampolli, A.: ItalWordNet: Building a Large Semantic Database for the Automatic Treatment of Italian, in: "Linguistica Computazionale", vol. XVI–XVII (2003), pp. 745–791, Giardini, Pisa.
12. Vossen, P. (ed.): EuroWordNet General Document, 1999. `http://www.hum.uva.nl/~EWN`.

Word Association Thesaurus As a Resource for Building WordNet

Anna Sinopalnikova

Masaryk University, Brno, Czech Republic
Saint-Petersburg State University, Russia
Email: anna@fi.muni.cz

Abstract. The goal of the present paper is to report on the on-going research for applying psycholinguistic resources to building a WordNet-like lexicon of the Russian language. We are to survey different kinds of the linguistic data that can be extracted from a Word Association Thesaurus, a resource representing the results of a large-scaled free association test. In addition, we will give a comparison of Word Association Thesaurus and other language resources applied to wordnet constructing (e.g. text corpora, explanatory dictionaries) from the viewpoint of the quality and quantity of information they supply the researcher with.

1 Introduction

Since 1985 methodology of wordnet building has undergone significant changes. Starting with the primarily psycholinguistic techniques adopted in the Princeton WordNet (PWN), it switched to the entirely different methodology of the EuroWordNet (EWN) project based on the usage of existing resources, either the PWN itself within the expand model, or available national language resources within the merge model.

In this article we will introduce a connecting link between those two methodologies and present a resource, which, on the one hand, contains psycholinguistic data, but on the other hand, in a well-structured form that makes it computer-processable and, thus susceptible of both PWN and EWN methods.

In the second part of the paper we define some basic notions of psycholinguistics, necessary for the further discussion. Section 3 is dedicated to observation of different types of the empirical linguistic data derived from WAT and applied to wordnet constructing. In the last section we will compare the results of WAT usage with that of text corpora from the viewpoint of their coverage.

2 Basic Concepts

Originally the term '**association**' was used in psycholinguistics to refer to the connection or relation between ideas, concepts, or words, which exists in the human mind and manifests in a following way: an appearance of one entity entails the appearance of the other in the mind; thus '**word association**' being an association between words. In modern studies this term is often expanded to the scope of corpus linguistics and lexicography, but we will use it in its traditional sense.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 199–205.

The simplest experimental technique to reveal the association mechanism is a '**free association test**' (FAT). Generally, a list of words (**stimuli**) is presented to subjects (either in writing or orally), which are asked to respond with the first word that comes into their mind (**responses**). As opposed to other, more sophisticated forms of association experiments (e.g. controlled association test, priming etc.), FAT gives the broadest information on the way knowledge is structured in the human mind.

The results of FAT series carried out with several hundreds stimuli and a few thousand subjects, reported in a form of tables, were given the name '**Word Association Norms**' (WAN). The body of WAN constitutes the list of stimuli, lists of responses with their absolute frequencies for each stimulus word. Along with the response distribution, frequency of response is considered to be an essential index, reflecting the strength of semantic relations between words.

The first WAN were collected by Kent and Rosanoff [1] on the base of the list of 100 stimulus words including common nouns and adjectives, and 1000 subjects being involved. Since then, numerous WAN for many European and Asian languages (monolingual, as well as bilingual and trilingual) were published using mostly Kent and Rosanoff list of stimuli and expanding their experience to other languages, e.g. [2,3,4].

Word Association Thesaurus (WAT) is quite similar to WAN, but it excels significantly in size (it includes several thousands of stimuli). Also the procedure of data collection is much more complicated: a small set of stimuli is used as a starting point of the experiment, responses obtained for them are used as stimuli in the next stage, the cycle being repeated at least 3 times. In so doing, WAT is expected to be a 'thesaurus', i.e. to cover 'all' the vocabulary and reflect the basic structure of a particular language. As opposed to WAN, so far WATs are available for two languages only: English (by [5, Kiss et al]): 8400 stimuli – 54000 words – 1000 subjects, (by [6, Nelson et al]): 5000 stimuli – 75000 responses – 6000 subjects; and Russian (by [7, Karaulov et al]): about 8000 stimuli – 23000 words – 1000 subjects.

3 What Kind of Linguistic Information Could Be Extracted from WAT

It is usually questioned what FATs actually show? They do indicate that certain words are related in some way, but do not specify how. Although full of valuable information, the results of word association tests should be interpreted with great care [8].

The first who made an attempt of linguistic interpretation of word associations was Deese [9] who applied word associations to measure a semantic similarity of different words. His main assumption was that similar words must evoke similar responses. Thus, counting the stimulus word itself as a response by each subject, he computed the index of correlation between pairs of words as the intersection of the two distributions of responses and interpreted it as a measure of semantic similarity.

In the following subsections we demonstrate how WATs could help to solve the problems of the wordnet coverage and its appropriate structuring.

3.1 The Core Concepts of the Language

Experiments [10] show that in every language there is a limited number of words those appear as responses in WAT more frequently than other words. Such a set of words has much in

common with frequency lists (according to corpora-driven data) – they are among the most frequently used ones, and sets of top concepts (according to existing ontologies) – they have above-average number of relations to other words. This set is quite stable:

- it does not change much with time;
- it does not depends on the starting circumstances, e.g. on words that were chosen as the starting set of stimuli, or the number of subjects.

E.g., the Russian WAT [7] contains 295 words with more then 100 relations, among them are человек ('man'), дом ('house'), любовь ('love'), жизнь ('life'), есть ('be/eat'), думать ('think'), жить ('live'), идти ('go'), большой ('big/large'), хорошо ('good'), плохо ('bad'), нет (не) ('no/not') ..., while Edinburgh WAT [5] includes 586 such words: *man, sex, no (not), love, house; work, eat, think, go, live; good, old, small. . .*

These words determine the fundamental concepts of a particular language, and thus should be incorporated into lexical database as its core components (e.g., EWN Base Concepts [11]). Representing the most general concepts, these words are associated to most other (more specific) words by means of hyponymy relations. Extracting this set of basic concepts we are to tackle the problem of wordnet structuring.

3.2 Syntagmatic Relations

According to the law of contiguity, through life we learn "what goes together" and reproduce it together. Therefore, if a stimulus word is a verb, responses are expected to be all its co-occurring words: its right and left micro-contexts; nouns, adjectives and adverbs that could function in a sentence as its arguments.

This data could be incorporated into a wordnet both as surface context patterns for words (e.g. selectional restrictions/preferences, valency frames for verbs, etc.), and as deep semantic relations between words (e.g. ROLE/INVOLVED relations). Moreover, each pattern may be accompanied by the probabilistic index reflecting frequency of its occurrence in WAT (and, as a hypothesis, its probability in texts).

Also this data is useful for performing other tasks of wordnet constructing. It provides an empirical basis for distinguishing different senses of a word, establishing relations of synonymy, hyponymy, and antonymy.

3.3 Paradigmatic Relations

The law of contiguity may also explain the co-occurrence of paradigmatically related words in WAT. As synonyms, hyponyms/hyperonyms, meronyms/holonyms, or antonyms regularly go together in macro-contexts, they often appear together as pairs 'stimulus – response' in WAT.

Explicitly presented paradigmatic relations are a distinctive feature of WAT that differs it from other language resources (there is no such explicit information in explanatory dictionaries, and to extract it from corpora one needs to apply some sophisticated techniques).

This information may be included directly in terms of semantic relations between wordnet entries; also it helps us to enrich and to check out the set of relations encoded earlier.

3.4 Domain Information

Apart from the data on conventional set of semantic relations such as synonymy, hyponymy, meronymy etc., WAT provides more subtle information concerning domain structuring of knowledge. E.g., *hospital* → nurse, doctor, pain, ill, injury, load… This type of data is not so easy to extract from corpora, in explanatory dictionaries it is presented partly (generally covers special terminology only) and mostly based on the lexicographers' intuitions. E.g., *Syringe* – (medicine) *a tube with a nozzle and piston or bulb for sucking in and ejecting liquid in a thin stream*[1]. As opposed to conventional language resources (LRs), WAT explicitly presents the way common words are grouped together according to the fragments of reality they describe.

Domain relations may be attributed to each word in a wordnet; that give us broader (in comparison with context patterns, see 'Syntagmatic relations') knowledge of the possible contexts for each wordnet entry. The necessity of such an expansion becomes obvious if we take into account that domain information becomes crucial while we approach wordnet usage in IR systems.

3.5 Relevance of Word Senses for Native Speakers

The fact is that about 80% of associations of a word in WAT [12], as well as 90% of occurrences of a word in a corpus [13], are related to 1–3 of its senses. That allows us to measure the relevance of a particular word sense for native speakers, and, hence, to find an appropriate place for it in the hierarchy of senses. E.g., if we consider the word *lap* and its associations, we could find that 3 senses (lap_1 – 'the flat area between the waist and the knee of a seated person', lap_2 – 'one circuit of a track or racetrack' and lap_3 – 'take up with the tongue in order to drink') account for 61% of its word associations (cf. lap_1 → *knee, sit, sit on*, etc. lap_2 → *circuit, race, run*, etc. lap_3 → *cat, milk, pap* etc.). Those could be regarded as the most important from the viewpoint of native speakers. Other senses, such as 'polish (a gem, or metal or a glass surface)' obviously constitute the periphery (∼2%). And there is no hint of the sense 'a part of an item of clothing' while it is presented in the explanatory dictionaries (cf. [13]).

These empirical evidences also help us to define the necessary level of sense granularity: to include into the wordnet no more and no less senses of each word than native speakers do differentiate. Thus, the problem of unnecessarily over-multiplying of sense entries (usually mentioned regarding PWN 1.5.) could be avoided.

3.6 Relevance of Relations for Native Speakers

It is clear that in a WN words must have at least a hyperonym and desirably a synonym. But what concerns relations other than Hyponymy and Synonymy, how could we ensure that we include all the necessary relations, and that what we include is necessary? Relations are not the same for different PoS, but also they are not the same for different words within the same PoS. E.g., according to [5] for English native speakers the most relevant relation of *buy* is that to its conversive *sell*, while for *cry* the most important relation would be INVOLVED_AGENT *baby*.

[1] This definition as well as the ones below was taken from New Oxford Dictionary of English. Oxford University Press (1998).

3.7 Semantic Classification of Words Obtained by Using Formal Criteria Only

Within the same PoS the proportion of syntagmatic and paradigmatic associations varies considerably. E.g. for Russian verbs the number of syntagmatic associations can vary from 35% to 90%. This ratio correlates with syntagmatic features of verbs, such as a number of valencies, strength of valencies, and their character (obligatory/optional), which in turn correlate with semantic features of the verb. This hypothesis is proved while building semantic classifications of verbs on the basis of formal criteria (e.g. the number of syntagmatic associations). The resulted classes turned to have much in common with semantic classes acquired by means of logic or componential analyses (cf. [14,15]).

This data supply us with empirical basis for appropriate structuring of lexical database: grouping the words into semantic classes, etc.

4 WAT vs. Corpus

It is unanimously recognized that to build an adequate and reliable lexical database (e.g. wordnet), reflecting all the potentialities of a language, it is not enough to rely upon information produced by 'experts' (i.e. linguists, lexicographers) and stored in conventional LRs, whatever advantages for machine usage they offer [16]. One should rather explore the raw data, and extract information from language in its actual (i.e. written and spoken texts), and its potential use (i.e. native speakers' knowledge of language), that could be examine by means of psycholinguistic techniques.

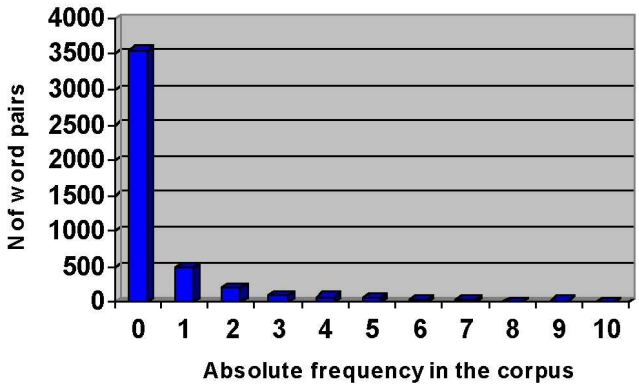


Fig. 1. Overlap between RWAT and the corpus.

Several researchers [17,18,19] performed statistical analysis and comparison of such 'raw' LRs, namely, text corpora and word associations, in order to confirm the correlation between frequency of XY co-occurrence in a corpus and the strength of association X-Y in WAN. Those experiments successfully demonstrated that corpora could be used to obtain the same measures of association strength as WAN, at least for the most frequent words. In our research we made a comparison in the opposite direction, and were to show that a

WAT covers more language phenomena than a corpus. For that purpose the Russian WAT [7] and a balanced text corpus of about 16 mln words were used. 6000 'stimulus-response' pairs e.g. бояться – темноты ('be afraid of – darkness') were extracted from RWAT in random order, and then searched in the corpus. The window span was fixed to $-10;+10$ words.

The most interesting result of our experiment was that about 64% word pairs obtained from subjects do not occur in the corpus (see the first column on Figure 1).

By excluding all unique associations (that with absolute frequency = 1) from the query list, the proportion of absent pairs may be reduced to 42%, which is still higher than expected. The distribution of the non-unique associations that were not found in the corpus could be seen in Table 1.

Table 1. Distribution of word associations that do not occur in the corpus.

N of occurrences in the corpus	N of occurrences in RWAT	% of all word pairs missed
0	2	48
0	3	22
0	4	14
0	5	8
0	6–10	5
0	11–15	<1
0	15–20	<1
0	>20	0

Looking for explanation we assumed that paradigmatically related words frequently appear as 'stimulus-respond' and less frequently co-occur in texts. But more detailed observation of the word pairs chosen revealed unexpectedly high ratio of syntagmatic word pairs to be absent. For verbs this number was about 84% of total amount of absent pairs. Whereas paradigmatically related words were regularly presented in the corpus.

Thus, we are to conclude that the experiment performed proves the value of WAT as a LR, which could supply the researcher with data otherwise inaccessible.

5 Conclusion

The advantages of using WAT in wordnet constructing may be stated as follows:

1. **Simplicity** of data acquisition.
2. Great **variety** of semantic information extracted.
   As it was shown in Sections 3 and 4, WAT is equal to or excels other LRs in several respects.
3. **Empirical nature** of data extracted (as opposed to theoretical one, cf. conventional dictionaries, that supposes the researcher's introspection and intuition to be involved, and hence, leads to over- and under-estimation of the language phenomena).
   As it was shown in Section 4, WAT may function as a source of 'raw' linguistic data, comparable to a balanced text corpus, and could supply all the necessary empirical information in case of absence of the latter.

4. **Probabilistic** nature of data presented (data reflects the relative rather then absolute relevance of language phenomena).

To sum up we may add, that the parallel usage of WAT and other LR is an effi cient way of conducting constant checking-out of wordnet construction, its refi ning and expanding. Thus, we believe the high consistency and coverage of wordnets could be achieved.

References

1. Kent, G. H., Rosanoff, A. J.: A Study of Association in Insanity. American Journal of Insanity, 67 (1910) 37–96.
2. Kurcz, I.: Polskie normy powszechnosci skojarzen swobodnych na 100 slow z listy Kent-Rosanoffa. Studia psychologiczne, tom 8. Warszawa (1967).
3. Novák, Z.: Volné slovní párové asociace v češtině. Praha (1988).
4. Rosenzweig, M. R.: Etudes sur l'association des mots. Année psychol. (1957).
5. Kiss, G. R., Armstrong, G., Milroy, R.: The Associative Thesaurus of English. Edinburgh (1972).
6. Nelson, D. L., McEvoy, C. L., Schreiber, T. A.: The University of South Florida word association, rhyme, and word fragment norms (1998) `http://www.usf.edu/FreeAssociation/`.
7. Karaulov, Ju. N. et al.: Russian Associative Thesaurus. Moscow (1994, 1996, 1998).
8. Clark, H. H.: Word associations and linguistic theory. In: J. Lyons (ed.). New horizons in linguistics. Harmondsworth: Penguin (1970) 271–286.
9. Deese, J.: The Structure of Associations in Language and Thought. Baltimore (1965).
10. Ufimtseva, N. V.: The core of the Russian mental lexicon (on the basis of large-scaled association tests). In: Proceeding of the Conference on Corpus Linguistics and Linguistic Databases. St-Petersburg, (2002) (in Russian).
11. Vossen, P. (ed.): EuroWordNet: A Multilingual Database with Lexical Semantic Network. Dodrecht, Kluwer (1998).
12. Ovchinnikova, I. G., Shtern, A. S.: Associative strength of the Russian words. In: Psycholinguistic problems of phonetics and semantics. Kalinin (1989) (in Russian).
13. Hanks, P.: Immediate context analysis: distinguishing meanings by studying usage. In: Heffer, Ch., Sauntson, H. (eds.) Words in Context. CD. Birmingham (2000).
14. Sinopalnikova, A. A. Classifying Russian Verbs according to their syntagmatic word associations. Diploma thesis, Saint-Petersburg State University (2000) (in Russian).
15. Ushakova, A. A. Classifying Russian Verbs: Componential and Definition analysis. Diploma thesis, Saint-Petersburg State University (2000) (in Russian).
16. Calzolari, N.: Lexicons and Corpora: between Theory and Practice. In: Proceedings of the 8þ International Symposium on Social Communication. Santiago de Cuba, (2003). 461–469.
17. Church, K. W., Hanks, P.: Word Association Norms, Mutual Information, and Lexicography. Computational Linguistics 16(1) (1990) 22–29.
18. Wettler, M., Rapp R.: Computation of Word Associations Based on the Co-Occurrences of Words in Large Corpora. In Proceedings of the 1st Workshop on Very Large Corpora: Academic and Industrial Perspectives. Columbus, Ohio (1993) 84–93.
19. Willners, C.: Antonyms in context: A corpus-based semantic analysis of Swedish descriptive adjectives. PhD thesis. Lund University Press (2001).

Quality Control for Wordnet Development

Pavel Smrž

Faculty of Informatics, Masaryk University in Brno
Botanická 68a, 602 00 Brno, Czech Republic
Email: smrz@fi.muni.cz

Abstract. This paper deals with quality assurance procedures for general-purpose language resources. Special attention is paid to quality control in wordnet development. General issues of quality management are tackled; technical as well as methodological aspects are discussed. As a case study, the application of the described procedures is demonstrated on the quality evaluation techniques in the context of the BalkaNet project.

1 Introduction

The BalkaNet project [1] aims at the development of wordnet-like lexical semantic networks for Czech and 5 Balkan languages – Bulgarian, Greek, Romanian, Serbian, and Turkish. As it shares many fundamental principles with the EuroWordNet project [2], it has been expected to employ the same procedures, policy, structure and tools as the previous project. However, discovered limitations of the EuroWordNet approach brought us to the decision to change data format, to design and implement new applications, and also to propose a modified perspective of the future development of the lexical semantic databases. Our conception, structure and tools are currently applied not only by members of the BalkaNet consortium but also by many other teams developing lexical databases all over the world.

There are many application-specific language resources developed with the goal to be directly integrated in a particular environment. On the other hand, there are resources that have been used or aim at their application in various NLP tasks. WordNet is the most prominent example. Though created to model human mental lexicon it has been employed in many domains from information retrieval to cultural linguistics, from text classification to language teaching, word-sense disambiguation, machine translation, etc.

Many well-established methods are available to evaluate the quality and contribution of language resources for specific application tasks. For example, the standard precision/recall graphs or F-measures are the most popular in the information retrieval. The fields of evaluation machine translation or information extraction systems pay also traditionally a strong attention to the quality assurance.

The procedures of quality control for general-purpose language resources are much less known. Moreover, the results of our research clearly show that this area has been strongly underestimated in many previous projects. Another finding suggests that if quality assurance policy has not been applied the results could differ considerably from that what was declared.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 206–212.

2 General Considerations

The most obvious requirement for a resource that aims at general usage is the availability of documentation of the process of its development and the final state of data. Resource documentation should be comprehensive but at the same time concise to allow quick scan. Unfortunately, many language resources resulting from various research projects account the role to a set of the standard project deliverables. In addition to the fact that these documents are often longer than necessary and do not describe all aspects of the resource, this approach does not reflect the process of development. Deliverables correspond to the state of knowledge and development of the resource at a particular time. Decisions and views can change during the project. The best strategy is therefore to summarize the description of resources in the end of such projects and check validity of information in all documents that will be part of the documentation.

The terminology used in the resource description should be also explicitly defined. Even the meaning of terms that seem to be basic in the context should be tackled. For example, synonymic set – synset – is the fundamental building block of wordnets but still it should be precisely described what kinds of variants (typographic, regional, register ...) will be contained in a synset. The Princeton WordNet itself is not entirely consistent in this respect – lake, loch and lough – as regional variants of the same concept – form 3 different synsets, lake is the hypernym of the two others.

The description of the data format in which the resource is provided plays also a crucial role. As XML has become de facto standard for data interchange, it is natural to make data available in XML and release the relevant DTD description. Data types of XML entities and other constraints on the tag content should be also specified. Elaborate standards from "the XML family", e. g. XML Schema [3] can be used to formally capture these definitions.

Along with the description of the data format it is appropriate to publish quantitative characteristics of the created data. A special attention should be paid to empty tags in the case of XML representation as it may signalize data inconsistency.

Our experience in previous projects aiming at development of language resources clearly showed that one of the most successful procedures to control the quality of linguistic output is to implement a set of validation checks and regularly publish their results. It holds especially for projects with many participants that are not under the same supervision. Validation check reports together with the quantitative assessment can serve as development synchronization points too.

3 Case Study of Quality Control in BalkaNet

The BalkaNet project will run till August 2004. Thus, we are not able to present the final documentation of all decisions that have been made in the course of the multilingual wordnet development. However, we present the current state of the project which reflects the refined quality control policy the BalkaNet consortium has adopted.

All partners agreed to prepare and update "resource description sheet" for the wordnet they develop. Such a specification should contain at least:

- description of the content of synset records and constraints on data types;

- types of relations included together with examples;
- degree of checking relations borrowed from PWN (see the note about the expand model below);
- numbering scheme of different senses (random, according to their frequency in a balanced corpus, from a particular dictionary, etc.)
- source of definitions and usage examples;
- order of literals in synsets (corpus frequency, familiarity, register or style characteristics).

One of the main characteristics that holds from very beginning of BalkaNet is the focus on large-scale overlap between national wordnets. The goal of this approach is to maximize the possibility of future applicability of the created database as a whole. A special set of synsets – BCS (BalkaNet Common Synsets) has been chosen and all partners agreed on the schedule of the gradual development. Several criteria have been adopted in the BCS selection process, which has taken the following steps:

1. All synsets contained in EuroWordNet base concepts have been included to maximize the overlap between the two projects.
2. The set has been extended based on the proposals of all partners who added synsets corresponding to the most frequent words in corpora and in various dictionary definitions for their particular languages.
3. As an additional criterion, several noun synsets that had many semantic relations in the Princeton WordNet database have been added.
4. All the selected synsets based on PWN 1.5 have been automatically mapped to PWN 1.7.1, which is currently the version BalkaNet is connected to. The synsetss that found one-to-one correspondence in the new version have been finally chosen.
5. All the hypernyms and holonyms of the chosen synsets have been added to BCS as it was decided to close the set in this respect.

All the steps (except the second for the proposer) imply the adoption of expand model for building a substantial part of the national wordnets. However, there is still room for the merge model, e. g. a significant portion of verb synsets in the Czech wordnet originated that way.

Synsets are formed by true context synonyms as well as variants (typographic, regional, style, register ...) in the BalkaNet wordnets. Moreover, verb synsets contain literals linked by a rich set of relations, e. g. aspect opposition and iteratives.

All the data should be linked to PWN till the end of the project. BalkaNet started with the idea to provide correspondence with PWN 1.5 and thus be compatible with EuroWordNet. However, the discovered limitations of PWN 1.5 let to the switch to PWN 1.7.1 which is much more consistent. As the new PWN 2.0 has been released in the last months the possibility of automatic re-linking of BalkaNet data to this version will be investigated too.

All national wordnets share the same data structure in XML. A synset described in this notation could look like:

```
<SYNSET>
  <ID>ENG171-08299742-n</ID> <POS>n</POS>
  <SYNONYM>
    <LITERAL>front man<SENSE>1</SENSE></LITERAL>
```

```
    <LITERAL>front<SENSE>8</SENSE></LITERAL>
    <LITERAL>figurehead<SENSE>1</SENSE></LITERAL>
    <LITERAL>nominal head<SENSE>1</SENSE></LITERAL>
    <LITERAL>straw man<SENSE>1</SENSE></LITERAL>
  </SYNONYM>
<ILR><TYPE>hypernym</TYPE>ENG171-08207586-n</ILR>
<DEF>a person used as a cover for some questionable activity</DEF>
</SYNSET>
```

The corresponding DTD for all BalkaNet wordnets then looks like:

```
<!ELEMENT WORDNET - - (SYNSET*) >
<!ELEMENT SYNSET  - - (ID, POS, SYNONYM, ILR*, ELR*, BCS?,
                       DEF?, USAGE*, SNOTE*, STAMP?) >

<!ELEMENT SYNONYM - - (LITERAL+) >
<!ELEMENT LITERAL - - (#PCDATA, SENSE, LNOTE?) >
<!ELEMENT SENSE   - - (#PCDATA) >
<!ELEMENT LNOTE   - - (#PCDATA) >

<!ELEMENT ILR  - - (TYPE, #PCDATA) >
<!ELEMENT ELR  - - (TYPE, #PCDATA) >
<!ELEMENT TYPE - - (#PCDATA) >

<!ELEMENT ID    - - (#PCDATA) >
<!ELEMENT POS   - - (#PCDATA) >
<!ELEMENT BCS   - - (#PCDATA) >
<!ELEMENT DEF   - - (#PCDATA) >
<!ELEMENT USAGE - - (#PCDATA) >
<!ELEMENT SNOTE - - (#PCDATA) >
<!ELEMENT STAMP - - (#PCDATA) >
```

The ID tag acts as the primary key of the entries and is also used in links where it substitutes the verbosity of proper XML linking mechanisms [4,5,6]. Identifiers are found in two slightly different forms:

1. Synsets connected to PWN are identified by three-part strings – the first is the version identifier (e. g. ENG15 for PWN version 1.5), the second is the offset in the PWN files for nouns, adjectives, verbs, or adverbs, and the third one is the concrete POS.
2. Synsets added by the consortium partners start with the three-letter language identifiers that correspond to the international standard ISO 639-2. The following number is generated sequentially to ensure uniqueness.

The second mentioned group is just a matter of the progressive development of national wordnets. Most of the synsets will be linked to their English equivalents till the end of the project. It means they will get IDs from PWN. The rest will form the core of what is called BalkaNet ILI (Inter-Language Index), or BILI. The prefix will be BWN10 and English

definition will be provided. The most discussed examples of this type so far are the names of meals served in the Balkan region.

A special mechanism has been adopted to signalize lexical gaps – concepts that are not lexicalized in a language. Such entries are labeled <NL/> in the BalkaNet database and they should be ignored when working with a particular wordnet as a monolingual resource.

The current DTD complies with the needs of the development process (BCS tags for synchronization, STAMP tag for management purposes, etc.). The final version will probably eliminate these tags and maybe adds others to facilitate linking to other resources.

Simple scripts using standard utilities like sort or diff tools have been implemented to compute quantitative characteristics. All the XML files are first normalized to eliminate effects of the different structure. The following frequency values are then computed:

- tag frequencies;
- ratio of the number of literals in the national wordnet and in PWN;
- ID prefix frequencies;
- frequency of link types;
- frequency of POS;
- coverage of BCS;
- number-of-senses distribution;
- number of "multi-parent" synsets;
- number of leaves, inner nodes, roots, free nodes in hyper-hyponymic "trees";
- path-length distribution.

Table 1 captures the most interesting statistics that reflect the state of Balkanet development in the end of the second year of the project.

Table 1. Current statistics on wordnets developed in BalkaNet

Wordnet	Bulgarian	Czech	Greek	Romanian	Serbian	Turkish	Princeton
Synsets	13,425	25,453	13,523	11,698	4,557	9,509	111,223
Literals	24,118	37,883	17,759	23,571	7,891	14,382	195,817
Lit/Syn	1.80	1.49	1.31	2.01	1.73	1.51	1.76
BCS	8,496	7,525	5,427	6,744	4,307	7,391	8,496

4 Automatic and Semi-automatic Quality Checking

The quality control has been one of the priorities of the BalkaNet project. As our evaluation proves even the actual data from the second year of the project are more consistent than the results of previous wordnet-development projects. Part of the success story definitely lies in the implementation of strict quality control and data consistency policy.

Data consistency checks can be considered from various points of view. They can be fully automatic or need less or more manual effort. Even if supported by software tools, manual checks present tedious work that moreover needs qualified experts. Another criterion for

applicability of checks is whether they can be applied to all languages or they are language-specific (e. g. constraints on characters from a particular codepage). An important issue is also the need for additional resources and/or tools (e. g. annotated monolingual or parallel corpora, spell-checkers, explanatory or bilingual dictionaries, encyclopedias, lemmatizers, morphological analyzers).

Similarly to the scripts for quantitative characteristics we have developed a set of checks that validate wordnet data in the XML format. The following inconsistencies are regularly examined on all BalkaNet data:

- empty ID, POS, SYNONYM, SENSE (XML validation);
- XML tag data types for POS, SENSE, TYPE (of relation), characters from a defined character set in DEF and USAGE;
- duplicate IDs;
- duplicate triplets (POS, literal, sense);
- duplicate literals in one synset;
- not corresponding POS in the relevant tag and in the ID postfix;
- hypernym and holonym links (uplinks) to a synset with different POS;
- dangling links (dangling uplinks);
- cycles in uplinks (conflicting with PWN, e. g. "goalpost:1" is a kind of post is a kind of "upright:1; vertical:2" which is a part of "goalpost:1");
- cycles in other relations;
- top-most synset not from the defined set (unique beginners) – missing hypernym or holonym of a synset (see BCS selecting procedure above);
- non-compatible links to the same synset;
- non-continuous numbering where declared (possibility of automatic renumbering).

The results of the checks are also regularly sent to the developers that are responsible for corrections. The current practice will be probably even further simplified when a new tool for consistency checking with a user-friendly graphical interface will be developed.

Semi-automatic checks that need additional language resources to be integrated are usually performed by each partner depending on the availability of the resources:

- spell-checking of literals, definitions, usage examples and notes;
- coverage of the most frequent words from monolingual corpora;
- coverage of translations (bilingual dictionaries, parallel corpora);
- incompatibility with relations extracted from corpora, dictionaries, or encyclopedias.

In addition to the above-mentioned checks, BalkaNet developers often work with outputs of various pre-defined queries retrieving "suspicious" synsets or cases that could indicate mistakes of lexicographers. For examples, these queries can list:

- nonlexicalized literals;
- literals with many senses;
- multi-parent relations;
- autohyponymy, automeronymy and other relations between synsets containing the same literal;
- longest paths in hyper-hyponymic graphs;

- similar definitions;
- incorrect occurrences of defined literals in definitions;
- presence of literals in usage examples;
- dependencies between relations (e. g. near antonyms differing in their hypernyms);
- structural difference from PWN and other wordnets.

Besides all the mentioned validation checks, quality of created resources is evaluated in their application. Several partners already used their data to annotate corpus text for WSD experiments. Such an experience usually shows missing senses or impossibility to choose between different senses. Another type of work that helps us to refine information in our wordnet was the comparison between the semantic classifications from the wordnet with the syntactic patterns based on computational grammar.

5 Conclusions and Future Directions

It is obvious that the effort aiming at the quality of developed resources paid already off in the form of consistent resulting data that can be successfully used in various applications. The BalkaNet project will follow the started approach and the set of consistency checks used to validate wordnets will be published in its end.

We will try to test and generalize the GUI tool for validation checking mentioned above. We will also continue to develop the XML based application that will employ XSLT and other XML standards to define the tests [7].

Acknowledgements

This work was supported by Ministry of Education of the Czech Republic Research Intent CEZ:J07/98:143300003 and by EU IST-2000-29388.

References

1. Balkanet project website, `http://www.ceid.upatras.gr/Balkanet/`.
2. Eurowordnet project website, `http://www.illc.uva.nl/EuroWordNet/`.
3. Fallside, D. C.: XML Schema Part 0: Primer (2001) `http://www.w3.org/TR/xmlschema-0/`.
4. DeRose, S., Maler, E., Orchard, D.: XML Linking Language (XLink) Version 1.0 (2001) `http://www.w3.org/TR/xlink`.
5. DeRose, S., Jr., R. D., Grosso, P., Maler, E., Marsh, J., Walsh, N.: XML Pointer Language (XPointer) W3C Working Draft (2002) `http://www.w3.org/TR/xptr`.
6. Clark, J., DeRose, S.: XML Path Language (XPath) Version 1.0 (1999) `http://www.w3.org/TR/xpath`.
7. Smrž, P., Povolný, M.: Deb – dictionary editing and browsing. In: Proceedings of the EACL03 Workshop on Language Technology and the Semantic Web: The 3rd Workshop on NLP and XML (NLPXML-2003), Budapest, Hungary (2003) 49–55.

Extension of the SpanishWordNet

Clara Soler

Universitat Ramon Llull, C. Padilla 326–332, 08025 Barcelona, Spain
Email: clarasp@blanquerna.url.es

Abstract. WordNet divides adjectives in descriptives and relationals basically and they are represented in an enumerative way. The category was not introduced in EuroWordNet and Spanish adjectives in the SpanishWordNet are the translation of the English synsets. This paper describes a proposal of organizing and incorporating adjectives into the SpanishWordNet in terms of representing its polymorphic behaviour. The new organization would be made according to the adjectives taxonomy of MikroKosmos ontology. It results that the ontological approach can be used to explain adjectives polysemy. In the end a new adjectival classification appears in EuroWordNet, in terms of the three types of entities of the Top Ontology.

1 Introduction

This paper describes a proposal for incorporating and organizing Spanish adjectives into the lexical database SpanishWordNet by means of the MikroKosmos Ontology[1]. Adjectives being currently displayed in the SpanishWordNet are a translation of the English adjectives contained in WordNet (version 1.5) into Spanish and Catalan languages. Thus, their semantic organization follows the model of the WordNet system.

It has been already suggested to extend EuroWordnet with language-neutral ontologies, such as CYC, MikroKosmos or Sensus [11]. In this case, in which adjectives are the focus, the procedure adapted to carry out the extension of the SpanishWordNet will be based on expanding the Top Concept Ontology of EuroWordNet with part of MikroKosmos ontology structure. The choice of MikroKosmos ontology is due to its lexical approach to represent a model containing information about types of things.

Section 2 outlines the classification and organization of adjectives in WordNet 1.5 and in the SpanishWordNet, as well as how adjective polysemy is considered in these databases. Section 3 proposes the extension of the SpanishWordNet by means of MikroKosmos; this will imply the incorporation and classification of adjectives according to ontological criteria, and will further imply a new classification of adjectives in terms of the three types of entities that constitute the Top Ontology of EuroWordNet, and finally the possibility of presenting the new approach to polysemy.

[1] *MikroKosmos Ontology* is one of the components of the MikroKosmos project on computational semantics, which is an automatic knowledge-based translation system. It is integrated by diverse microtheories whose objective is to describe the static meaning of all the lexical categories in different languages [10]

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 213–219.

2 Adjectives Polysemy in WordNet, EuroWordNet, and the SpanishWordNet

2.1 In WordNet

Antonymy is the lexical relation that expresses in WordNet synsets which are opposite in meaning. It also divides and organizes adjectives in two main classes: the class of Descriptives, which have antonyms, and the class of Relationals, without antonyms[2]. The first ones are organized into non-hierarchic synsets formed by one, or more, pairs of antonym adjectives. Relational adjectives are represented with pointers to the noun or verb from which they derive.

Apart from constituting a criterion to classify adjectives in WordNet, the relation of Antonymy helps to disambiguate polysemous nouns (as is stated in Fellbaum [5]). Descriptive adjectives express opposed values of attributes, most of which are bipolar. Some of these adjectives do not have direct antonyms but can acquire them indirectly via another semantic relation (Similar to). This relation distinguishes a peripheral o satellite adjective synset linked to the most central synset. This peripheral adjective (e.g.*moist*) may do not have a direct antonym, but via the Similar to Relation (*moist* is Similar to *wet*) acquires the indirect antonym *dry*. Now, as it is observed in Fellbaum [5] many of the less frequent and unusual adjectives (and less polysemous), are quite selective in relation to the noun they modify, and thus they constitute a class of adjectives that can be used to disambiguate the meaning of a polysemous noun. This verification suggests a division between a small set of highly polysemous common adjectives, such as *big*, *small*, *good*, *bad*, *new*, *old* etc., and a greater set of more discriminating, less interchangeable adjectives, like *academic* and *international*. A distinction can be made between those adjectives that can help to disambiguate a noun and those that cannot.

This difference is reflected between direct and indirect antonyms. Indirect antonyms are compatible with less nominal heads and are therefore less polysemous. They probably contribute to the disambiguation of the nouns that modify. The Relational ones are not organized in terms of sets of antonyms and are less polysemous.

2.2 In EuroWordNet

There are two main reasons why adjectives were not included in EuroWordNet. It is considered in Fellbaum [5] the information conveyed by an adjective, being a modifier, is less vital that the one expressed by nouns and verbs for understanding sentences in an NLP system. The other reason is the difficulty of its own semantics: adjectives are considered highly polysemous and that makes difficult to represent it in an enumerative lexicon like EuroWordNet, where is pretended to distinguish all senses of a word form.

However, if adjectives semantics was reconsidered, and perhaps their polysemy was not that high probably their inclusion in EuroWordNet would present less difficulty. In WordNet adjective polysemy is related to features such as its frequency, its compatibility with greater or smaller number of nominals, and noun disambiguation. This paper presents an approach to adjective polysemy based on completely different criterion.

[2] Apart from these types WordNet contains the participal adjectives file. These adjectives are considered a kind of Descriptives without antonyms, and are kept in a separated file.

2.3 In the SpanishWordNet

As previously stated, the Spanish and Catalan adjectives adopted by the SpanishWordNet are translations of the English word represented in WordNet 1.5. These adjectives are expressed in an enumerative and descriptive way, in correspondence with the structure of the semantic net. WordNet has already been noted for its excessive grain size, and sometimes the number of lexical entries exceeds those really necessary. This is especially the case of adjectives considered highly polysemous (Fellbaum [5]) e.g. *big*, *good*, *big*... This situation also occurs in the SpanishWordNet.

3 Extension of the SpanishWordNet

Adjective classifi cation in WordNet is extensive enough and it takes account of both semantic and syntactic information. However there are some questions that remain with no answer: the classifi cation does not give any account of the relation between different senses of the same adjectival form. What happens with Relationals? Are not they polysemous? Etc. We propose in this paper that the same classifi cation can be made from other criterion. The fact that the Antonymy relation is a lexical relation between word forms and not concepts makes diffi cult to give an explanation of adjective polysemy. The proposed new criterion come from within the framework of the MikroKosmos project [9]. In this model, the lexicon mediates between a language of meaning representation and an ontology. Adjective meaning is explained according to this conceptual ontology. The lexical entries are instances of ontological types, and each one of them indicates a lexical connection of these units of the language to ontological concepts. In this framework adjectives are divided into Scalars, which are based on ontological concepts of Property and into Non-Scalars (Relationals). These are then subdivided into Denominals, that are based on ontological concepts of Entity, and in Deverbals, that are based on ontological concepts of Event. The basic criterion to establish a class of adjective is its association to a certain ontological type. This representation reflects the semantic structure of the adjective. An adjective always has either a noun or a verb as a reference. Ontological criterion supplies extra information to lexical criterion of WordNet classifi cation and makes the analysis to become deeper and more comprehensive.

3.1 Treatment and Representation of Adjectives Polysemy

Adjective polysemy has become a subject increasingly studied within the area of NLP studies. The so called adjectival polymorphism is logically treated differently from different perspectives and points of view. It is not the objective of this paper to explain our own point of view about it, but we can outline the following:

– Polysemy always implies a change of meaning. It is possible to distinguish then between ambiguity and polysemy. In the case of ambiguity the adjective really does not suffer a change of meaning but acquires different shades. This is the case for instance of an adjective such as *good*, considered highly polysemous in several works (WordNet itself is an example). From our point of view in most of the cases is just ambiguity.

– Adjective polysemy can be explained within the framework of the lexical ontological semantics formulated in MikroKosmos. According to the ontological classifi cation argumented in [9], an adjective is polysemous when it is more of an ontological type. A casuistry then appears, examples of which are given next.

A Change of Ontological Type: from an Ontology of Entities or Events to an Ontology of Properties.

(1) Next Sunday I have a *familiar* meal.

The ontological nature of this adjective is of Entity.

(2) At work there is a very *familiar* atmosphere.

The ontological nature of this adjective is Evaluative (Scalar type).

A Change within the Same Ontological Type.

(3) El día *claro*[3]

Property: luminosity. Antonym: dark

(4) La crema *clara*[4]

Property: density. Antonym: thick

(5) *High* mountains

Property: height. Antonym: low

(6) *High* sea

In this case *high* is a synonym of *stormy*, which is a Denominal adjective. Antonym: calm.

The change takes place here within the Scalar adjectives. In (3), and (4) a change of scale takes place: from the property *brightness*, into the scale of the property *thickness*. It is a change that occurs at the same level, and apparently it seems diffi cult to predict which property comes fi rst and which one derives from the other. Nevertheless, the change in (5) and (6) is different. The jump occurs from a scale of an objective property, such as *height*, to a Non-Scalar adjective. *High*is obviously a Scalar adjective, but probably in the case of *High sea* it becomes part of a collocation of a semi-compositive nature. Its real meaning becomes *stormy* which is a Denominal.

3.2 Extension of the SpanishWordNet

Having established our proposal concerning the polysemy of an adjective the next step is to establish the procedure to implement a representation of the adjectives in the SpanishWordNet which could express the new classifi cation. This could be carried out by introducing some of the features of the MikroKosmos ontology to the Top Ontology of EuroWordNet. Let us outline the structure of both ontologies:

[3] *El día claro* means *the bright day*

[4] *La crema clara* means *the thin custard*

Top Ontology Structure EuroWordNet was founded upon two parts: the covering of a shared set of common Basic Concepts; and the extension of the lexical base from these Concepts using semiautomatic techniques. The Basic Concepts constitute a set of 1024, and the Top Ontology was created in order to classify them. It consist of 63 fundamental semantic distinctions used in various semantic theories and paradigms. These Top Concepts are organized by means of subtype and opposition relations. The ontology provides an independent structuring of the language to the Basic Concepts in terms of these semantic distinctions (which are considered to be more semantic features than common conceptual classes). These 63 semantic distinctions are classified by three types of entities: 1st-Order-Entities, 2nd-Order-Entities and 3rd-Order-Entities (following Lyons [6]). The 2nd-Order-Entities are those that can be denoted by any part of the speech: nouns, verbs, adjectives and adverbs. They represent any static or dynamic situation that cannot be grasped, seen, felt, or experienced as an independent physical thing. They are located in time and they can happen rather than exist. The Top Ontology is linked to the ILI (Inter-Lingual-Index), so are the word meanings int the local synsets (local wordnets such as the SpanishWordNet).

Mikrokosmos Structure MikroKosmos is organized according to a set of concepts. Each concept constitutes a collection of properties with partially specified values. The concepts are organized hierarchically. Semantically, the first difference between them occurs between 'free concepts' and 'bounded concepts'. The 'free concepts' represent classes of objects and classes of events that have their corresponding instances in a TMR (Text Meaning Representation). The 'bounded concepts' represent classes of properties that categorize the objects and the events and that normally do not have instances but appear as values of the objects and instantiated events. The Concept Root is *ALL*, and the subclasses are *Events*, *Objects* and *Properties*.

Properties are the conceptual basics of the ontology. They help to define the concepts and can appear in the ontology in two different ways: As defined types of concepts or as values of the definitions of the objects and events. A value is the basic mechanism that represents relations between concepts. It is the fundamental metaontological mechanism. One of the subtypes of values is 'relaxable-to', which indicates the point at which the ontology allows violations of the restrictive selections giving rise to non literal uses such as the metaphor or metonymy. A proposal to extend SpanishWordNet is given next:

1. In the Top Ontology The adjectives must be classified under the 2nd-Order-Entities. Some of the Base Concepts belonging to these type of entities already refer to situations which can be denoted by adjectives (e.g. Social, Physical). It is therefore necessary to make an exhaustive verification in order to know which of those Basic Concepts refer to adjectives. It will then be possible to establish which are the lacking concepts.

2. In the MikroKosmos Ontology MikroKosmos Ontology establishes nine scales/properties of numerical type and four of literal type[5] that identify Qualifying (descriptive) adjectives. It is therefore necessary to verify if all scalar adjectives are covered.

[5] These scales come to be the ontological correlate of the different adjective taxonomies proposed in the framework of semantic studies of adjectives, being the one proposed by Dixon [4] the most relevant.

3. Joining both ontologies

- Incorporation of the adjectival synsets into the ILI.
- Incorporation of the adjective ontological taxonomy of MikroKosmos into the Top Ontology. This will consist of the following: Incorporation of the new basic concepts that represent the adjective scales into the class of 2nd-Order-Entities. The scalar adjectives will be defined according to these concepts. Denominals will be defined according to the Basic Concepts (nominal) classified as 1st-Order-Entities and 3rd-Order-Entities, which are already introduced as they constitute EuroWordNet. Deverbals will also be defined according to basic concepts classified as 2nd-Order-Entities. Polysemy can be represented using the 'relaxable-to' relation, which can connect the different ontological concepts, from which the adjective derives its different meanings.

4. Later classification of adjectives The incorporation of these new basic concepts allows a double classification according to the three specified entity types. On the one hand, one classification of adjectives can be made according to the basic concepts understood as 2nd-Order-Entities, since any of them will be subsumed under these entities, and on the other hand they can be classified according to the ontological concepts from which they derive, that is to say, according to all three entity types, of first, second and third order.

4 Conclusions

We put forward a proposal to classify, to reorganize and to represent the semantic structure of the adjective in the SpanishWordNet. It can be proposed for EuroWordNet too. This will allow a more global understanding of its behaviour. This paper is focused on the paradigmatic aspect of the adjective and to study how to represent the sintagmatic aspect, which takes into account the adjective-noun combinations (of compositive, semicompositive and noncompositive character) constitutes one of the tasks to make next. Another one is to determinate the different polysemous types, and to study in depth the linguistic phenomenon of adjective polysemy.

References

1. Alonge, A., N. Calzolari, P. Vossen, L. Bloksma, I. Castellón, M.A. Martí and W. Peters: *The Linguistic Design of the EuroWordNet Database*. Kluwer Academic Publishers. Computers and the Humanities **32** 91–115(1998).
2. Apresjan, Ju. D.: *Regular Polysemy*. Linguistics **142** Mouton The Hague Paris (1974).
3. Bouillon, P. and E. Viegas: *The Description of Adjectives for Natural Language Processing: Theoretical and Applied Perspectives*. Atelier Thematique TALN, Cargèse, 12–17 juillet (1999).
4. Dixon, R. M. W.: *Where Have All the Adjectives Gone?* In: Robert M. W. Dixon, **Where Have All the Adjectives Gone? and Other Essays in Semantics and Syntax**. Berlin-Amsterdam-New-York: Mouton. 1–62 (1982).
5. Fellbaum, C.: *A Semantic Network of English: The Mother of All WordNets*. Computers and the Humanities **32** 209–220 Kluwer Academic Publishers (1998).
6. Lyons, J.: *Semantics*. Cambridge University Press. London (1977).

7. Miller, G., R. Beckwith, C. Fellbaum, D. Gross and K. Miller: *Five Papers on WordNet*. CSL Report **43**. Cognitive Science Laboratory. Princeton University (1990).
8. Nirenburg, S. and V. Raskin: *Ontological Semantics* (not published).
9. Raskin, V. and S. Nirenburg: *Lexical Semantics of Adjectives: A Microtheory of Adjectival Meaning*. NMSU CRL MCCS-95-288. Mahesh, K. and S. Nirenburg. Knowledge-Based Systems for Natural Language Processing. NMSU CRL MCCS-96-296 (1995).
10. Raskin, V. and S. Nirenburg:*Adjectival Modifi cation in Text Meaning Representation*. Proceedings of the 17th International Conference on Computational Linguistics (1996).
11. Rodríguez, H., S. Climent, P. Vossen, L. Bloksma, W. Peters, A. Alonge, F. Bertagna and A. Roventini: *The Top-DoWordNet Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology*. Computers and the Humanities **32** 117–152. Kluwer Academic Publishers (1998).
12. Vossen, P.: *Introduction to EuroWordNet*. Computers and the Humanities. Kluwer Academic Publishers (1998).

Pathways to Creativity in Lexical Ontologies

Tony Veale

Department of Computer Science,
University College Dublin, Belfield, Dublin 6, Ireland.
Email: Tony.veale@UCD.ie
WWW: http://www.cs.ucd.ie/staff/tveale/home/

Abstract. Language is a highly creative medium, and lexicalized ontologies like WordNet are rich in implicit evidence of the conceptual innovations underlying lexical inventiveness. We argue that WordNet's overt linguistic influences make it far more conducive to the development of creative thinking systems than other, more formalized conceptual ontologies like Cyc.

1 Introduction

Creativity is a vexing phenomenon to pin down formally [1], which is perhaps why we tend to think of it in largely metaphoric terms. For example, creativity is often conceived as a form of mental agility that allows gifted individuals to make astonishing mental leaps from one concept to another [2]. Alternately, it is popularly conceived as a form of lateral thinking that allows those who use it to insightfully cut sideways through the hierarchical rigidity of conventional categories [3]. Common to most of these metaphors is the idea that creativity involves recategorization, the ability to meaningfully move a concept from one category to another in a way that unlocks hidden value, perhaps by revealing a new and useful functional property of the concept. For example, psychometric tests such as the Torrance test of creative thinking [4] try to measure this ability with tasks that, e.g., ask a subject to list as many unusual and interesting uses of old tin cans as possible.

The ad-hoc nature of creativity is such that most ontologies, perhaps all ontologies, do not and can not provide the kinds of lateral linkages between concepts to allow this kind of inventive recategorization. Instead, ontologies tend to concentrate their representational energies on the hierarchical structures that, from the lateral thinking perspective, are as much a hindrance as an inducement to creativity. This is certainly true of WordNet [5], whose *isa* hierarchy is the most richly developed part of its lexical ontology, but it is also true of language-independent ontologies like Cyc [6], which are rich in non-hierarchical relations but not of the kind that capture deep similarity between superfi cially different concepts. It is connections like these that most readily fuel the recategorization process.

However, because WordNet is an ontology of lexicalized concepts, it necessarily captures much of the lexical creativity evident in everyday language. Often, this word-use is a reflection of deeper recategorization processes at the conceptual level. We argue that if we can identify and extract this evidence using automatic or semi-automatic means, we then have a basis for augmenting WordNet with the lateral connections from which novel creative pathways can be constructed.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 220–225.

2 Polysemy versus Homonymy

Polysemy is a form of lexical ambiguity in which a word has multiple related meanings. The form of polysemy that interests us most from a creativity perspective is function-transforming polysemy, which reflects at the lexical level the way concepts can be extended to fulfi ll new purposes. For instance, English has a variety of words that denote both animals and the meat derived from them (e.g., *chicken, lamb, cod*), and this polysemy reflects the transformation potential of animals to be used as meat.

If we can identify all such instances of function-transforming polysemy in WordNet, we can generalize from these a collection of pathways that allow a system to hypothesize creative uses for other concepts that are not so entrenched via polysemy. For example, WordNet defi nes several senses of *knife*, one as an *{edge-tool}* used for cutting and one as a *{weapon}* used for injuring. Each sense describes structurally similar objects (sharp flat objects with handles) with a common behavior (cutting) that differ primarily in function (i.e., slicing vs. stabbing). This polysemy suggests a generalization that captures the functional potential of any other *{edge-tool}*, such as *{scissors}* and *{shears}*, to also be used as a *{weapon}*. More formally, for every polysemous sense pairing $<\omega_1, \omega_2>$ with immediate hypernyms $<h_1, h_2>$, we can create a category subsumption entailment $h_1(x) \rightarrow h2(x)$ if h_2 is a broader category that h_1, which is to say, if h_2 has more descendent hyponyms than h_1. Since *{weapon}* is a broader category than *{edge-tool}*, we can infer that other edge-tools may be used as weapons too, but conversely, we do not infer that all weapons are potential edge-tools. In effect, the generalization represents an inductive hypothesis that it is the sharp edge in a tool that allows it to be used as a weapon.

3 Identifying Creativity-Supporting Polysemy in WordNet

It is crucial that our generalization process be able to distinguish polysemy from homonymy – another form of ambiguity in which the multiple senses of a word are not related – since WordNet's synset representation does not explicitly mark either phenomenon.

True polysemous relationships can be recognized using a variety of automatic approaches. In the top down approach, cousin relations [5,7] are manually established between concepts in the upper-ontology to explain the systematicity of polysemy at lower levels. For instance, once a connection between *{animal}* and *{food}* is established, it can be instantiated by words with both an animal and food sense. However, this approach is limited by the number of high-level connections that are manually added, and by the need to list often copious exceptions to the pattern (e.g., *mate* the animal partner, and *mate* the berry drink, are merely homonyms; the latter is not derived from the former). Conversely, in the bottom-up approach, systematic patterns are fi rst recognized in the lower ontology and then generalized to establish higher-level connections [8,9,10]. For instance, several words have senses that denote both a kind of music and a kind of dance (e.g., *waltz, tango, conga*), which suggests a polysemous relationship between *{music}* and *{dance}*.

Both of these approaches treat polysemy as a systematic phenomenon best described at the level of word families. However, while such a treatment reveals interesting macro-tendencies in the lexicon, it does little to dispel the possibility that homonymy might still operate on the micro-level of individual words (as demonstrated by the size of the exception

list needed for the first approach). We thus prefer to use an evidential case-by-case approach to detecting polysemy, connecting a pair of senses only when explicit local taxonomic evidence can be found to motivate a connection. This evidence can take many forms, so a patchwork of heuristic detectors is required. We describe here the three most interesting of these heuristics.

The coverage of each heuristic is estimated relative to that achieved by the *cousins* collection of 105 regular polysemy noun-sense groupings that are hand-coded in WordNet [7]. Over-generation is estimated relative to the overlap with the cousins exception list [7], which permits us to also estimate the accuracy of each heuristic.

Explicit Ontological Bridging: a sense pair $<\omega_1, \omega_2>$ for a word ω can be linked if ω_1 has a hypernym that can be lexicalized as M-H and ω_2 has a hypernym that can be lexicalized as M, the rationale being that ω_2 is the M of ω_1 and ω_1 is the H of ω_2. E.g., the word *olive* has a sense with a hypernym *{fruit-tree}*, and another with the hypernym *{fruit}*, therefore M = *fruit* and H = *tree*. *(Coverage: 12%, Accuracy: 94%)*.

Hierarchical Reinforcement: if $<\alpha_1, \alpha_2>$ and $<\beta_1, \beta_2>$ are sense pairs for two words α and β where α_1 is a hypernym of β_1 and α_2 is a hypernym of β_2, then $<\alpha_1, \alpha_2>$ reinforces the belief that $<\beta_1, \beta_2>$ is polysemous, and vice versa. For example, *herb* denotes both a plant and a foodstuff in WordNet, and each of these senses has a hyponym that can be lexicalized as *sage*. *(Coverage: 7%, Accuracy: 12%)*.

Cross-Reference: if $<\omega_1, \omega_2>$ is a sense pair for a word ω and the WordNet gloss for $\omega_2$explicitly mentions a hypernym of ω_1, then ω_2 can be seen as a conceptual extension of ω_1. For instance, the railway-compartment sense of *diner* mentions *restaurant* in its gloss, while another sense actually specifies *{restaurant}* as a hypernym. This suggests that the railway sense is an extension of the restaurant sense that uses the later as a ground for its definition. *(Coverage: 62%, Accuracy: 85%)*.

These heuristics are very effective at arguing for polysemy on the local merits of individual words. However, for every creatively-useful instance of polysemy like *knife* (*{weapon}* versus *{edge-tool}*), there is an unhelpful instance like *capsule* (*{space-vehicle}* versus *{medicine}*), for one cannot meaningfully reuse aspirin-capsules as spacecraft, and vice versa. At present, we manually filter those instances of polysemy (almost 50%) from the set produced by the above heuristics whenever structural and behavioral properties are not preserved between senses.

4 Types of Ontological Creativity

The polysemy relationships that can be extracted from WordNet are merely the residue of past creativity by the language community. However, new creative insights can be generated by generalizing from these entrenched precedents, to either broaden existing categories and admit new members not previously considered eligible, or to re-categorize members of existing categories under different branches of the ontology.

Category Broadening: Imagine we want to broaden the WordNet category *{weapon}*. The members of this category can be enumerated by recursively visiting every hyponym of the category, which will include *{knife}*, *{gun}*, *{artillery}*, *{pike}*, etc. But by traversing polysemy links as well as *isa* relations, additional prospective members can be reached and admitted on the basis of their functional potential. Thus, the polysemy of *knife* causes not

only *{dagger}* and *{bayonet}* but *{steak_knife}* and *{scalpel}* to be visited. Stretching category boundaries even further, the generalization *edge_tool(x)* → *weapon(x)* allows the category *{edge_tool}* to be subsumed in its entirety, thereby allowing *{scissors}, {axe, ax}, {razor}* and all other sharp-edged tools to be recognized as having weapon-like potential.

Category broadening is a very revealing process, not only about the functional potential of everyday objects, but also about the inevitable gaps in an ontology like WordNet. For instance, the category *{apparel, clothing, clothes}* can be broadened to admit baseball gloves, anklets, metal helmets, furs and animal skins, while the category *{medicine, medication}* can be broadened to admit toiletries and oleoresins, and the category *{food}* can be broadened to admit a variety of potentially edible substances, some too disgusting to list here.

Category Hopping: Imagine, following the Torrance test, we want to move the concept *{coffee_can}* to a new category that will offer a functional perspective on how to effectively reuse old tin cans. The existing WordNet categories that house *{coffee_can}* can be enumerated by recursively visiting each of its hypernyms in turn, which will include *{can, tin_can},{container}* and *{artifact}*. Now, each of these hypernyms is a potential point of departure to another category if, as well as traversing *isa* relations, we use polysemy relationships to slip from one rail of the ontology to another. WordNet defines *{coffee_can}* as a hyponym of *{can, tin_can}*, and from here a leap can be made to *{steel_drum, drum}*, since both are hyponyms of *{container}* whose glosses further specify them as kinds of *metal container*. From *{steel_drum, drum}* there exists a polysemy link to *{tympan, membranophone, drum}*, a non-container artifact which WordNet defines as a hyponym of *{percussion_instrument}*. This chain of reasoning, from *{coffee_can}* to *{tin_can}* to *{steel_drum}* to *{tympan, membranophone, drum}*, supports the creative insight that allows an old tin can to be used a musical drum, and central to this insight is the polysemy of *drum*. In general, polysemy supports creativity by providing just one very important link in the recategorization chain. A dog collar can be fashionably reused as a necklace because the polysemy of *collar* links *{collar}* to *{choker, collar}*. We can meaningfully think of jewelry as a piece of fine-art (and thus consider exhibiting it in a gallery) because of the polysemy of *gem* that links *{gem, jewel}* to *{gem, treasure}*. Likewise, we can think of photography as a fine art because photograph and art collide via the polysemy of *mosaic*, *vignette* and *scene*.

5 Creativity, Utility and Similarity

Some recategorizations will exhibit more creativity than others, largely because they represent more of a mental leap within the ontology. We can measure this distance using any of a variety of taxonomic metrics [11], and thus rank the creative outputs of our system. For instance, it is more creative to reuse a coffee can as a *{percussion_instrument}* than as a *{chamberpot, potty}*, since like *{tin_can}* the latter is already taxonomized in WordNet as a *{container}*. Any similarity metric (called σ, say) that measures the relative distance to the lowest common hypernym will thus attribute greater similarity to *{coffee_can}* and *{potty, chamberpot}* than to *{coffee_can}* and *{tympan, drum, membranophone}*. This allows us to measure the creative distance in a recategorization from α to γ as $1 - \sigma(\alpha, \gamma)$.

Of course, distance is not the only component of creativity, as any recategorization must also possess some utility to make it worthwhile (e.g., there is a greater distance still between tin cans and fish gills, but the former cannot be sensibly reused as the latter). In other words,

a creative product must be unfamiliar enough to be innovative but familiar enough to be judged relative to what we know already works. This is the paradox at the heart of ontological creativity: to be creative a recategorization must involve a significant mental leap in *function* but not in *form*, yet typically (e.g., in WordNet), both of these qualities are ontologically expressed in the same way, via taxonomic structure. This suggests that taxonomic similarity σ must be simultaneously maximized (to preserve structural compatibility) and minimized (to yield a creative leap).

Fortunately, polysemy offers a way to resolve this paradox. If a creative leap from α to γ is facilitated by a polysemous link from $<\beta, \gamma>$, the sensibility of the leap can be measured as $\sigma(\alpha, \beta)$ while the creativity of the leap can be measured as $1 - \sigma(\alpha, \gamma)$. The value of a creative product will be a function of both distance and sensibility, as the former without the latter is unusable, and the latter without the former is banal. The harmonic mean is one way of balancing this dependency on both measures:

$$value(\alpha, \gamma) = 2\sigma(\alpha, \beta)(1 - \sigma(\alpha, \gamma))/(1 + \sigma(\alpha, \beta) - \sigma(\alpha, \gamma))$$

Other variations on this formula can be used to give greater of lesser weight to the roles of sensibility and distance in determining the value of a creative insight.

6 Concluding Observations

The ideas in this paper have now been implemented in a computational system called *Kalos* (a Greek word connoting beauty through fitness of purpose [3]). A collection of 25 different polysemy detectors (of which 3 were described here) achieve 96% of the coverage offered by WordNet's own cousin relations, at a precision of 85%. In our pilot study, we focused on the subset of these polysemous relations that connect artifactual noun senses, where this subset is hand-filtered to yield 991 instances of behaviour-preserving, function transforming polysemy. Generalizing from these instances and performing a second phase of hand-checking to filter out spurious hypotheses, we are left with 454 inter-category subsumption hypotheses. These generalizations are a powerful addition to WordNet's upper and middle ontologies, facilitating a creative flexibility in determining category membership that is useful to a variety of applications, from creative writing tools to text understanding systems.

References

1. Wiggins, G. Categorizing Creative Systems, in: Proc of the 3rd Workshop on Creative Systems, IJCAI'03, Acapulco, Mexico. (2003).
2. Hutton, J.: Aristotle's Poetics. Norton, New York (1982).
3. de Bono, E. Parallel Thinking. Viking Press: London (1994).
4. Torrance, E. P. The Torrance Tests of Creative Thinking. Scholastic Testing Service. Bensonville, Illinois. (1990).
5. Miller, G. A.: WordNet: A Lexical Database for English. Communications of the ACM, Vol. 38 No. 11 (1995).
6. Lenat, D., Guha, R. V.: Building Large Knowledge-Based Systems. Addison Wesley (1990).
7. WordNet documentation. `http://www.princeton.edu/~wn/` (2003).

8. Peters, W., Peters, I., Vossen, P. Automatic sense clustering in EuroWordNet. In: Proc of the 1st international conference on Language Resources and Evaluation. Spain. (1998).
9. Peters, I., Peters, P. Extracting Regular Polysemy Patterns in WordNet. Technical Report, University of Sheffield, UK. (2000).
10. Peters, W., Peters, I. Lexicalized Systematic Polysemy in WordNet. In the proceedings of the 2nd international conference on Language Resources and Evaluation. Athens. (2000).
11. Budanitsky, A., Hirst, G. Semantic Distance in WordNet: An experimental, application-oriented evaluation of five measures. In: Proc of the Workshop on WordNet and Other Lexical Resources, North-American chapter of ACL. Pittsburgh. (2001).
12. Lakoff, G.: Women, Fire and Dangerous Things. Uni. of Chicago Press: Chicago (1987).

Automatic Lexicon Generation through WordNet

Nitin Verma and Pushpak Bhattacharyya

Department of Computer Science and Engineering,
Indian Institute of Technology Bombay – 400076, India
Email: nitinv@iitb.ac.in, pb@cse.iitb.ac.in

Abstract. A lexicon is the heart of any language processing system. Accurate words with grammatical and semantic attributes are essential or highly desirable for any application – be it machine translation, information extraction, various forms of tagging or text mining. However, good quality lexicons are difficult to construct requiring enormous amount of time and manpower. In this paper, we present a method for automatically generating the dictionary from an input document – making use of the *WordNet*. The dictionary entries are in the form of Universal Words (UWs) which are language words (primarily English) concatenated with disambiguation information. The entries are associated with syntactic and semantic properties – most of which too are generated automatically. In addition to the WordNet, the system uses a *word sense disambiguator*, an *inferencer* and the *knowledge base (KB)* of the *Universal Networking Language* which is a recently proposed interlingua. The lexicon so constructed is sufficiently accurate and reduces the manual labour substantially.

1 Introduction

Construction of good quality lexicons enriched with syntactic and semantic properties for the words is time consuming and manpower intensive. Also word sense disambiguation presents a challenge to any language processing application, which can be posed as the following question: *given a document D and a word W therein, which sense S of W should be picked up from the lexicon?*. It is, however, a redeeming observation that a particular W in a given D is mostly used in a single sense throughout the document. This motivates the following problem: *can the task of disambiguation be relegated to the background before the actual application starts? In particular, can one construct a* **Document Specifi c Dictionary** *wherein single senses of the words are stored*?

Such a problem is relevant, for example, in a machine translation context [2]. For the input document in the source language, if the *document specific dictionary* is available a-priori, the generation of the target language document reduces to essentially syntax planning and morphology processing for the pair of languages involved. The WSD problem has been solved before the MT process starts, by putting in place a lexicon with the document specifi c senses of the words.

In this paper we have addressed this problem by showing how the WordNet [5,3] can be used to construct a document specifi c dictionary. Section 2 briefly describes the UNL system and the Universal Words [4]. Format of UW Dictionary is described in Section 3. Section 4 narrates about the resources used for dictionary generation and Section 5 explains the methodology for dictionary generation. Section 6 gives the results obtained by performing experiments on the system and lists out the future directions for this work.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 226–233.

2 Universal Networking Language (UNL)

UNL [4] is an interlingua for machine translation [2] and is an attractive proposition for the multilingual context. In this scheme, a source language sentence is converted to the UNL form using a tool called the *EnConverter* [4]. Subsequently, the UNL representation is converted to the target language sentence by a tool called the *DeConverter* [4]. The sentential information in UNL is represented as a hyper-graph with concepts as nodes and relations as arcs. The UNL graph is a hyper-graph because the node itself can be a graph, in which case the node is called a *compound word* (CW). Figure 1 represents the sentence *John eats rice with a spoon*.

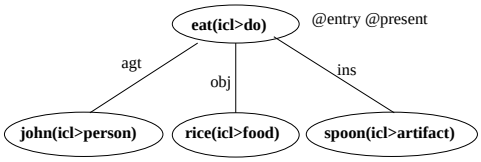


Fig. 1. UNL graph of *john eats rice with a spoon*

The UNL graph is represented as a set of directed binary relations between two concepts present in the sentence. The relation *agt* (figure 1) stands for *agent*, *obj* for *object* and *ins* for *instrument*. The binary relations are the basic building blocks of the UNL system, which are represented as strings of 3 characters or less each.

In the above figure the nodes such as *eat(icl>do), John(iof>person), and rice(icl>food)* are the *Universal Words (UW)*. These are language words with *restrictions* in parentheses. *icl* stands for *inclusion* and *iof* stands for *instance of*. UWs can be annotated with attributes which provide further information about how the concept is being used in the specific sentence. Any of the three restriction labels, *viz., icl, iof* and *equ*, is attached to an UW for restricting its sense. For example, two senses of *state* will be represented in the UNL system in the following way:

- *state(icl>express) to express something clearly and carefully*.
- *state(icl>country) a politically organized body of people under a single* government.

A UW is created using the *specifications* of the *UNL Knowledge Base (KB)*. UNL KB organizes the UWs in a *hierarchy*. A *part* of the UW hierarchy for *nouns* in the UNL KB is shown in figure 2 which is self-explanatory.

For verbs, the hierarchy is not so deep. All the verbs are organized under three categories, *viz., do, occur* and *be*. The first two are *aktionstat verbs* and the last one is the set of *stative verbs*. The adjective, adverb and preposition hierarchies too are quite shallow. The adjectives that are both *attributive* and *predicative* are given the restriction $(aoj > thing)$, where *aoj* is a semantic relation denoting *attribute of the object* and *thing* denotes a nominal concept. The adjectives which are only *predicative* are given the restriction $(mod > thing)$ where *mod* is the *modifier* relation. The adverbs are uniformly expressed through $(icl > how)$.

```
Depth   UW
-----   --------------
   0    thing
        |
   1    |__abstract thing{(icl>thing)}
   2    |  |___activity(icl>abstract thing)
   3    |  |   |___broadcasting(icl>activity{>abstract thing})
   3    |  |   |___defense(icl>activity{>abstract thing})
   3    |  |   |___development(icl>activity{>abstract thing})
   2    |  |___art(icl>abstract thing)
   3    |  |   |___fine arts(icl>art{>abstract thing})
   3    |  |   |___music(icl>art{>abstract thing})
   2    |  |___aspect(icl>abstract thing)
     ........
     ........
```

Fig. 2. Hierarchy of *noun* UWs in the UNL KB (a snapshot)

3 L-UW Dictionary

The dictionary maps the *words* of a natural language to the *universal words* of the UNL system [6]. For example

[dog] "dog(icl>mammal)" (. . . $attributes$ . . .)

[bark] "bark(icl>do)" (. . . $attributes$ . . .)

are the entries in an English-UW dictionary. When the sentence *The dog barks* is given to an UNL-based English-Hindi MT system, the UWs *dog(icl>mammal)* and *bark(icl>do)* are picked up. These are disambiguated concepts different from other senses of *dog* and *bark*, for example the *pursue* sense of *dog* ($dog(icl > do)$) and the *skin of the tree* sense of *bark* ($bark(icl > skin)$). *If the L-UW dictionary contains only document specific UWs, the analyser and the generator systems do not commit error on account of WSD.*

The *attributes* attached to each entry in the L-UW dictionary are the *lexical, grammatical,* and *semantic* properties of the language specific words (*NOT of the UWs*). The syntactic attributes include the word category – *noun, verb, adjectives, adverb etc.* and attributes like *person* and *number* for nouns and *tense* for verbs. The *Semantic Attributes* are derived from an *ontology*. Figure 3 shows a part of the *ontology* used for obtaining semantic attributes [6].

4 Resources for Dictionary Generation

For generating the document specific dictionary we use the *WordNet, a WSD System*, the *UNL KB* and an *inferencer*. The approach is *Knowledge Based* [12]. The UNL KB as shown in figure2 is stored as a *mysql* database. The table *UNL-KB-table* in figure4 shows a part of this storage structure for nouns.

The word sense disambiguator [1] works with an accuracy of about 70% for nouns. The essential idea is to use the *noun–verb* association – as given in a co-occurence dictionary – to obtain a set of semantic clusters for the noun in question. The densest cluster denotes the

```
Part of ontology for nouns              Part of ontology for verbs
==================================      ==============================

Animate (ANIMT)                         Verbs of Action (VOA)
  o Flora (FLORA)                         o Change (VOA,CHNG)
     =>Shrubs (ANIMT, FLORA, SHRB)        o Communication (VOA,COMM)
  o Fauna (FAUNA)                       Verbs of State (VOS)
     =>Mammals (MML)                      o Physical State (VOS,PHY,ST)
     =>Birds (ANIMT, FAUNA, BIRD)         o Mental State (VOS,MNTL,ST)
  .....                                   .....

Part of ontology for adjectives         Part of ontology for adverbs
==================================      ==============================

Descriptive (DES)                       Time (TIME)
  o Weight (DES,WT)                     Frequency (FREQ)
  o Shape (DES,SHP)                     Quantity (QUAN)
  o Quality (DES,QUAL)                  Manner (MAN)
Relational (REL)                        .....
.....
```

Fig. 3. Ontology and Semantic attributes

most likely sense of the word. Taking the example of *the crane flies* we get two semantic clusters involving the hypernyms and the hyponyms of the *bird* sense and the *machine sense*. Since the former has much larger association with *fly*, it becomes the winner.

For other parts of speech, the first sense as given in the WordNet is chosen, which as per the WordNet is the most frequently used sense.

The semantic attributes are generated from a rule-base linking the lexico-semantic relations of the WN with the semantic properties of the word senses. To take an example, if the hypermyny is *organism*, then the attribute *ANIMT* signifying *animate* is generated. We have more than 1000 such rules in the rule base.

5 Methodology for Dictionary Generation

As discussed so far, there are two parts to the dictionary entry generation, *viz.*, creating UWs and assigning the syntactic and semantic attributes. The following subsections discuss this.

5.1 POS Tagging and Sense Disambiguation

The document is passed to the word sense disambiguator [1]. This picks the correct sense of the word with about 70% accuracy. As a side effect the words are POS tagged too. The output of this step is a list of entries in the format **Word:POS:WSN**, where POS stands for part of speech and WSN indicates the WordNet sense number. The *syntactic* attributes are obtained at this stage.

5.2 Generation of UWs

The WN and UNL KB are used to generate the restriction for the word. If the word is a noun, the WN is queried for the hypernymy for the marked sense. All the Hypernymy ancestors $H_1, H_2, \ldots, H_n$ of W up-to the *unique beginner* are collected. If $W(icl > H_i)$ exists in the UNL KB, it is picked up and entered in the dictionary. If not, $W(icl > H_1)$ is asserted as the dictionary entry.

for example, for *crane* the *bird*-sense gives the hypernyms as *bird, fauna, animal, organism* and finally *living_thing*. $crane(icl > bird)$ becomes the dictionary entry in this case. Figure 4 illustrates this process.

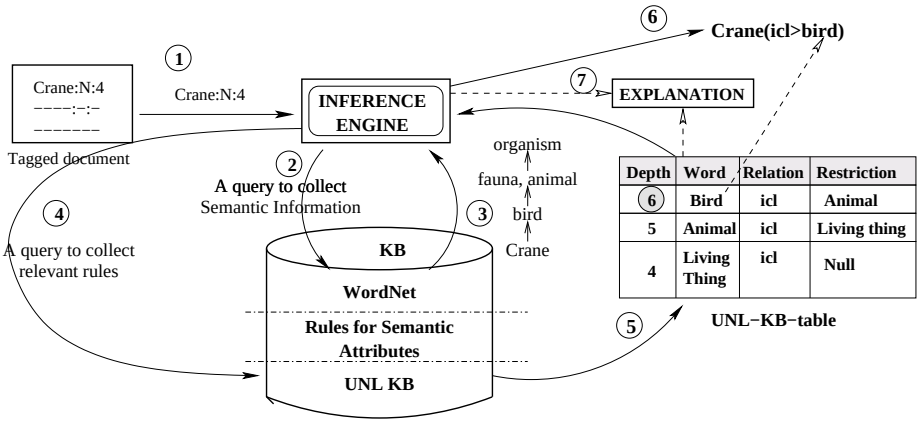


Fig. 4. Universal Word Creation: an example

For verbs, the hypernymy ancestors are collected from the WN. If these include concepts like *be, hold, continue etc.*, then we generate the restriction $(icl > be)$ (case of *be* verb). If not, the corresponding *nominal word* (for example, the nominal word for the verb *rain* is *rain* itself) of the verb is referred to in the WN. If the hypernyms of the nominal word include concepts like *phenomenon, natural_event etc.*, then we generate the restriction $(icl > occur)$ signifying an *occur* verb. If both these conditions are not satisfied, then the restriction $(icl > do)$ is generated.

For adjectives, use is made of the *is_a_value_of* semantic relation in the WN. For example, for the adjective *heavy* the above relation links it to *weight*. If this relation is present then the restriction (*aoj* > *thing*) is generated. Else we generate (*mod* > *thing*) (please refer back to section 3).

For adverbs, $(icl > how)$ is by default generated, as per the specifications of the UNL system.

5.3 Creation of Semantic Attributes

As explained in section 4, WN hypernymy information and the rule base is used to generate the semantic attributes of nouns. The tables in the figure 5 shows sample of such rules for all

the POS words. The first entry in the table 1 corresponds to the rule: IF hypernym = *organism* THEN generate *ANIMT* attribute. For example for the *bird* sense of *crane* (**crane:N:4**), the entry [$crane$]"$crane(icl > bird)$"(N, *ANIMT*, *FAUNA*, *BIRD*); is generated.

Table 1: Rules for noun

HYPERNYM	ATTRIBUTE
organism	ANIMT
flora	FLORA
fauna	FAUNA
beast	FAUNA
bird	BIRD

Table 2: Rules for verb

HYPERNYM	ATTRIBUTE
change	VOA,CHNG
communicate	VOA,COMM
move	VOA,MOTN
complete	VOA,CMPLT
finish	VOA,CMPLT

Table 3.1: Rules for adjectives

IS_VALUE_OF	ATTRIBUTE
weight	DES,WT
strength	DES,STRNGTH
qual	DES,QUAL

Table 3.2: Rules for adjectives

SYNONYMY OR ANTONYMY	ATTRIBUTE
bright	DES,APPR
deep	DES,DPTH
shallow	DES,DPTH

Table 4: Rules for adverbs

SYNONYMY	ATTRIBUTE
backward	DRCTN
always	FREQ
frequent	FREQ
beautifully	MAN

Fig. 5. Rules for generating Semantic attributes

6 Experiments and Results

We have tested our system on documents from various domains like agriculture, science, arts, sports *etc*. each containing about 800 words. We have *measured* the *performance* of this system by calculating its *precision* in every POS category. The precision is defined as

$$Precision = \frac{Number\ of\ entries\ correctly\ generated}{Total\ entries\ generated}$$

figure 6 shows the results. The average precision for nouns is **93.9%**, for *verbs* **84.4%**, for *adjectives* **72.4%** and for *adverbs* **58.1%**.

The dictionary generated by the above methodology performs well in case of nouns and verbs. The reason for low accuracy for adjectives and adverbs is the shallowness in the hierarchy and lack of many semantic relations for these parts of speech. The system is being routinely used in our work on machine translation in a tri-language setting (*English*, *Hindi* and *Marathi*) [7,8]. It has reduced the burden of lexicography considerably. The incorrect entries – which are not many – are corrected manually by the lexicon makers. Figure 7 shows the dictionary generated (the wrong entries are marked by a *) after running our system on a document containing the following paragraph.

Modern agriculture depends heavily on engineering and technology and on the biological and physical sciences. Irrigation, drainage, conservation, and sanitary engineering – each of which is important in successful farming – are some of the fields requiring the specialized knowledge of agricultural engineers.

The future work consists in generating restrictions involving *iof (instance-of), equ (equivalent to), pof (part of)* and such other constructs. Efforts are also on to migrate the system to WordNet 2.0 which has the very useful relations of *derived_from* and *domt* doing cross POS linkage in the WN. It is hoped that this will mitigate the problems arising from the low accuracy of the WSD system and the shallowness of the non-noun hierarchies.

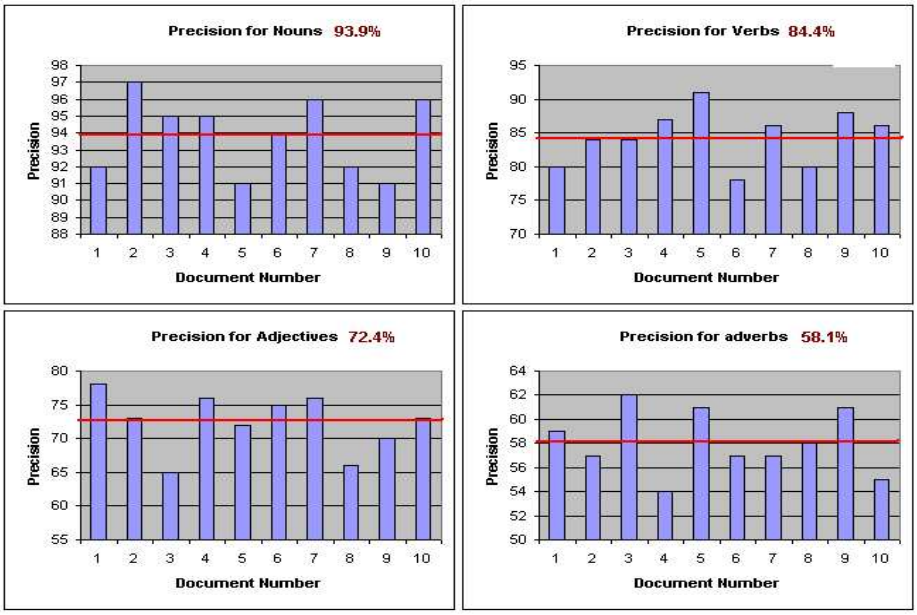


Fig. 6. Experiments and Results

```
  [Modern]{}"modern(aoj>thing)"(ADJ,DES,APPR)<E,0,0>
  [agriculture]{}"agriculture(icl>business)"(N,INANI,EVENT,ABS)<E,0,0>
  [depend]{}"depend(icl>be(aoj>thing))"(VRB,CONT,VOS-PHY-ST)<E,0,0>
  [heavily]{}"heavily"(ADV,QUAN)<E,0,0>
  [engineering]{}"engineering(icl>subject)"(N,INANI,PSYFTR,ABS)<E,0,0>
  [technology]{}"technology(icl>subject)"(N,INANI,PSYFTR,ABS)<E,0,0>
  [biological]{}"biological(mod<thing)"(ADJ,REL)<E,0,0>
  [physical]{}"physical(mod<thing)"(ADJ,DES,SHAPE)<E,0,0>
  [scienc]{}"science(icl>skill)"(N,INANI,PSYFTR,ABS)<E,0,0>
  [Irrigation]{}"irrigation(icl>act)"(N,INANI,EVENT,ABS)<E,0,0>
* [drainage]{}"drainage(icl>change)"(N,INANI,EVENT,ABS)<E,0,0>
  [conservation]{}"conservation(icl>improvement)"(N,INANI,EVENT,NAT,ABS)<E,0,0>
* [sanitary]{}"sanitary(aoj>thing)"(ADJ)<E,0,0>
  [important]{}"important(aoj>thing)"(ADJ,DES,NUM)<E,0,0>
  [successful]{}"successful(aoj>thing)"(ADJ,DES,SND)<E,0,0>
* [field]{}"fields(icl>person)"(N,ANIMT,FAUNA,MML,PRSN,PHSCL)<E,0,0>
  [requir]{}"require(icl>necessitate(agt>thing,gol>place,src>place))"
                                                (VRB,VOA-POSS)<E,0,0>
* [specialized]{}"specialized(mod<thing)"(ADJ)<E,0,0>
  [knowledge]{}"knowledge(icl>cognition)"(N,INANI,PSYFTR,ABS)<E,0,0>
  [agricultural]{}"agricultural(aoj>thing)"(ADJ,REL)<E,0,0>
  [engineer]{}"engineer(icl>person)"(N,ANIMT,FAUNA,MML,PRSN,PHSCL)<E,0,0>
```

Fig. 7. UW Dictionary generated after running the system on a sample document

References

1. Dipak K. Narayan and Pushpak Bhattacharyya.: *Using Verb-Noun association for Word Sense Disambiguation*. International Conference on Natural language processing, November 2002.

2. W. John Hutchins and Harold L. Somers.: *An Introduction to Machine Translation*. Academic Press, 1992.
3. Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.: *Five papers on WordNet*. Available at URL: `http://clarity.princeton.edu:80/~wn/`, 1993.
4. The Universal Networking Language (UNL) Specifications, United Nations University. Available at URL: `http://www.unl.ias.unu.edu/unlsys/`, July 2003.
5. Christiane Fellbaum.: *WordNet: An Electronic Lexical Database*. The MIT Press, 1998.
6. P. Bhattacharyya.: *Multilingual information processing using UNL*. in Indo UK workshop on Language Engineering for South Asian Languages LESAI, 2001.
7. Shachi Dave, Jignashu Parikh and Pushpak Bhattacharyya,: *Interlingua Based English Hindi Machine Translation and Language Divergence*, Journal of Machine Translation, Volume 17, September, 2002. (to appear).
8. Shachi Dave and Pushpak Bhattacharyya,: *Knowledge Extraction from Hindi Texts*, Journal of Electronic and Telecommunication Engineers, vol. 18, no. 4, July 2001.
9. Hiroshi Uchida and Meiying Zhu. *The Universal Networking Language beyond Machine Translation*. UNDL Foundation, September 2001.
10. James Martin, and Steven Oxman. *Building Expert Systems, A tutorial*. 1998.
11. Susan Lindsay. *Practical Applications of Expert Systems*. QUD information sciences, 1988.
12. Adrian A. Hopgood *Knowledge-Based Systems for Engineers and Scientists*. CRC Press LLC, 1992.

Fighting Arbitrariness in WordNet-like Lexical Databases – A Natural Language Motivated Remedy

Shun Ha Sylvia Wong

Computer Science, Aston University, Aston Triangle, Birmingham B4 7ET, U.K.
Email: s.h.s.wong@aston.ac.uk

Abstract. Motivated by doubts on how faithfully and accurately a lexical database models the complicated relations that exist naturally between real-world concepts, we have studied concept organisation in WordNet 1.5 and EuroWordNet 2. Based on the arbitrariness in concept classification observed in these wordnets, we argue that concept formation in natural languages is a plausible means to improve concept relatedness in lexical databases. We also illustrate that word formation in Chinese exhibits natural semantic relatedness amongst Chinese concepts which can be exploited to aid word sense disambiguation.

1 Introduction

Research has shown that lexical databases are good sources of lexical knowledge for various Natural Language Processing (NLP) tasks. Over the years, several lexical databases have been developed, e.g. HowNet [1], WordNet [2], EuroWordNet [3] and CCD [4]. These knowledge bases differ in their detailed organisation of real-world concepts and how the knowledge base is structured. However, they all share one common feature – they all aim to specify a hierarchy of language-independent concepts which, in the developers' view, characterises important semantic distinctions between the concepts. These concepts are interrelated through a set of relations. Wong & Fung [5] observed that many of these concepts and relations are in common.

While such formalised knowledge bases are known to be well-defined hierarchical systems, there remains doubt as to how faithfully and accurately such artificial constructs model the complicated relations that exist naturally between real-world concepts. Based on the observation done on WordNet 1.5 and EuroWordNet 2, we discuss some common weaknesses in WordNet-like lexical databases. Motivated by Wong & Pala's studies [6,7], we propose a means to alleviate these weaknesses. We have carried out an experiment on the potential applicability of the proposed means of alleviation has been carried out. This paper gives a brief account of the results.

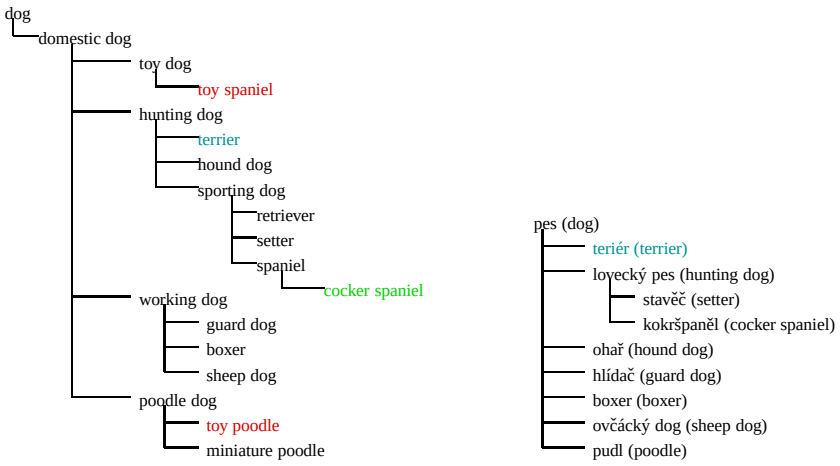
2 Some Common Weaknesses of WordNet-like Lexical Databases

In existing lexical databases, the classification of concepts is often based on hand-crafted guidelines and an individual's interpretation of the guidelines. Though exploiting existing electronic dictionary resources reduces the time involved in the manual classification process

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 234–241.

dramatically [3], by and large, given a set of relations and a set of concepts, to associate them with each other remains a subjective process.

Let us consider the concepts `toy poodle` and `toy spaniel` in Princeton WordNet 1.5 [2], i.e. "*the smallest poodle*" and "*a very small spaniel*", respectively. Both concepts are characterised by their smallness (in size) and they are also associated with the same set of top concepts in the 1stOrderEntity of the EuroWordNet 2 top ontology: `Animal`, `Form`, `Living`, `Natural`, `Object`, `Origin`. However, they are grouped under different hyperonyms (cf. Figure 1). While `toy dog` refers to "*any of several breeds of very small*



(a) WordNet 1.5 (b) Czech WordNet

Fig. 1. Extracts of `dog` concept hierarchies

dogs kept purely as pets" and poodles are also kept purely as pets, it is rather surprising that `toy poodle` is not classifi ed as a kind of `toy dog` and that `poodle dog` is not a hyponym of `domestic dog`. Furthermore, WordNet 1.5 specifi es that `toy spaniel` is "*a very small spaniel*" and `cocker spaniel` has hyperonym `spaniel`. While both `toy spaniel` and `cocker spaniel` are a kind of `spaniel`, this relation is not captured in WordNet 1.5. Imagine using such a concept hierarchy to aid a search on articles about various kinds of spaniels. Articles on toy spaniels would likely be ignored. As each system of concepts is defi ned according to the developers' view of the real-world, it is inevitable that the resulting ontology of concepts is fragmented and incoherent.

EuroWordNet was inspired by, and structured along the same line as, WordNet 1.5. WordNet 1.5 also serves as an interlingua within EuroWordNet. Real-world concepts and events exist regardless of the existence of natural languages. One would expect a concept to bear the same properties irrespective of its physical expression in different languages. However, while `terrier` in English is a `hunting dog`, its Czech counterpart `teriér`

is not[1] (Cf. Figure 1). The difference between the concept hierarchies in Figure 1 further exemplifies the existence of arbitrariness in concept classification.

Large-scale lexical databases are also prone to human errors. In EuroWordNet, the same synset in various European language wordnets are linked by Inter-Lingual-Index (ILI), which is in fact a list of meanings taken from WordNet 1.5 [3]. However, rather than relating `hunting dog` in English to `lovecký pes` (literally: *hunting dog*) in Czech, ILI incorrectly relates `lovecký pes` to `sporting dog`. This also explains why `teriér` and `ohař` (*hound dog*) are not considered as hunting dogs in Czech WordNet. If the association were to be done automatically based on the underlying component concepts, i.e. `lovecký` (*hunting*) and `pes` (*dog*), instead of relying on human classification, the mistake could have been avoided.

To attain a cohesive level of concept representation which is error-free from an human perception of the real-world is not an easy task. As a lexical knowledge base serves as the core foundation of various NLP tasks, a fragmented and incoherent knowledge base would, no doubt, hinder its effectiveness significantly.

3 A Natural Language Motivated Remedy

The aim of natural languages is to facilitate a concise communication of real-world concepts by means of sequences of symbols. This leads one to think whether the system for knowledge representation employed in a natural language could aid the development of a lexical database. Such a system is likely to be less subjective because, typically, it is a system developed, tested and agreed upon by millions of people over centuries. However, this system, though it exists, is hidden in most natural languages, especially those with phonetically-driven orthography.

Unlike most natural languages, the Chinese language displays a considerable amount of semantic information even at the character level. This distinctive feature suggests that the system of Chinese characters might contain a rich but concise system of inter-related concepts.

3.1 Chinese Characters

Chinese script has originated from picture-writing. Though over thousands of years of development, modern Chinese script is no longer dominated by pictographs [8,9], most Chinese characters continue to display some semantic information of the concept that it represents. Each Chinese character plays the role of a morpheme in the Chinese language. They all represent concepts that exist in the real-world.

According to Xu Shen's etymological dictionary, over 99% of the included Chinese characters display relevant semantic information to the concept that they represent [8,9]. The unique derivation of Chinese characters enables semantically related concepts to be grouped together naturally through their meaning component parts. For instance, concepts of psychological aspects like 怒 (*anger*), 恥 (*shame*), 想 (*think*) and 愛 (*love*) all possess the meaning component 心 (*heart / mind / feelings*) and concepts of trees like 橡 (*rubber tree*), 松 (*pine*), 杏 (*apricot*) and 樺 (*birch*) all share the component 木 (*tree /*

[1] Note that the words 'terrier' and 'teriér' are a pair of English-Czech cognates.

wood). Following this grouping, clusters of concepts displaying various semantic relations can be formed. While lexical databases often rely on subjective and even ad hoc judgement on concept classifi cation, the semantic relatedness displayed by such clusters of Chinese characters provides a means to concept classifi cation which is more objective, more explicit and, hence, easier to capture.

3.2 Chinese Concept Formation

There are over 50,000 characters in the Chinese script, but an average educated Chinese knows roughly about 6,000 characters [8]. Surprisingly, this rather limited knowledge of the Chinese script does not prohibit a Chinese from effective communication.

In English, the combination of letters to form words has little direct correlation with the meaning of words. With most Indo-European languages, it is possible to retrieve the composite meaning of a word by analysing its morphemic structure automatically [10] or semi-automatically [11]. However, with the presence of allomorphs and irregular morphology in words, to achieve reliable automatic analytical results is not an easy task.

Unlike Indo-European languages, Chinese words are typically composed of two Chinese characters. Each component character contributes part of the underlying meaning of a word, e.g. 噴射 (*jet*) = 噴 (*spurt*) + 射 (*shoot*). This characteristic holds even for words that are composed of more Chinese characters, e.g. 噴射式戰鬥機 (*fighter jet*) = 噴 (*spurt*) + 射 (*shoot*) + 式 (*model / style*) + 戰 (*battle / war*) + 機 (*machine / chance*). Thus, the knowledge of a few thousands characters allows a Chinese to deduce the meaning of words, even words which were previously unseen. Likewise, new words can also be formed by meaningful concatenation of characters.

Derivational morphology in Chinese is displayed naturally in Chinese word formation. Each Chinese character within a word corresponds to one morpheme. A study on the composite meaning of over 3,400 randomly selected Chinese words has been performed. This study revealed that the underlying meaning of over 99% of them correlates with the meaning of their component characters. Klimova & Pala [11] observed that morphemic structures of Czech words show suffi cient regularity to shed light on improving the relatedness of concepts (which are organised as synsets) within EuroWordNet by means of Internal Language Relations (ILRs). This leads us to investigate the potential for Chinese word formation in enriching sense relations in existing lexical database.

With our collection of Chinese words, we grouped them according to their component characters. We found that each cluster of Chinese words displays a high level of sense relatedness. For instance, 假髮 (*wig*), 長假髮 (*peruke*), 長髮 (*long hair*), 短髮 (*short hair*), 直髮 (*straight hair*) and 曲髮 (*curly hair*) all end with 髮 (*hair*[2]) and they all describe various appearances of a person's hair. The Chinese words 牙齒 (*tooth*[3]), 牙膏 (*toothpaste*), 牙刷 (*toothbrush*), 牙線 (*dental floss*) and 牙醫 (*dentist*) begin with the component character 牙 (*a canine tooth*) which reveals that these Chinese words are all related to teeth.

Although word formation based on concatenation of morphemes exists in many natural languages, e.g. teach and teacher in English, učit (*teach*) and učitel (*teacher*) in Czech,

[2] 髮 often refers to hair on a person's head because its component part 髟 means (*long hair*).
[3] 牙齒 is composed of 牙 (*a canine tooth*) and 齒 (*a tooth, the upper incisors*).

lehren (*teach*) and Lehrer (*teacher*) in German, due to evolution of natural languages, the morphemic structure of a word might not be traceable without considering other influential natural languages. Furthermore, the set of morpheme involved in a general use of any natural language is larger than that in the Chinese language. Thus, the set of relations observed in these languages is likely not to be sufficiently representative for improving knowledge representation in a large scale lexical database.

4 Exploiting Concept Relatedness in Chinese

Concept relatedness naturally displayed among Chinese words enables clusters of semantically related Chinese words to be formed. One might argue that typical concept relations like *hyponymy/hyperonymy* also enable concept clustering. At a glance, the Chinese data shown in Section 3.2 simply correspond to a typical case of hyperonyms in WordNet, EuroWordNet and HowNet, and attributes in HowNet and CCD. In our view, the Chinese data also display the nature of multiple inheritance in concept formation. For instance, the Chinese concept 戰車 (*chariot*) is composed of 戰 (*battle / war*) and 車 (*vehicle*). These two component concepts contribute equally to the well-formedness of meaning for 戰車 (*chariot*). Hence, rather than simply considering the concept 戰車 (*chariot*) as a hyponym of vehicle with the attribute 戰 (*battle / war*), we also view 戰 (*battle / war*) and 車 (*vehicle*) as two distinct contexts in which the concept 戰車 (*chariot*) are likely to appear. This concept, when used in a text in conjunction with other concepts, shapes the overall context of the text. This characteristic also has a potential to assist in topic detection [12].

Concept relatedness in Chinese provides a ready means to exploit conceptual density in word sense disambiguation. Consider the polysemous English word fight in Figure 2. Each sense forms a cluster with their semantically related concepts. For example, the fight sense "*to hit, punch and beat (a person)*" (打) has a proximity to beat to death (打死); whereas the sense "*contending for, by or as if by combat*" (戰鬥) relates to the concept `battle`. The senses "*to engage in a quarrel*" (爭執) and "*to strive vigorously and resolutely*" (爭取) are semantically closer to each other than the fight sense "*to hit, punch and beat (a person)*" (打) because they both comprise the `argue` (爭) component.

We have implemented a Java program to perform word sense disambiguation on English texts based on the Chinese representation of each English sense expressed by an English word. Our disambiguation process is based solely on the relatedness of concepts that are expressed in each sample text. It does not take into account any part-of-speech information of the source word forms. The disambiguation process comprises three tasks: sample text preprocessing, dictionary lookup and word sense selection. The test preprocessing and dictionary lookup processes seek to locate all available Chinese interpretations of an English lexical unit in our dictionary of 2566 English-Chinese word pairs. Typically, an identifiable lexical unit in our sample texts is associated with 4–5 Chinese concepts. In word sense selection, the dominating context of each sample text is determined by counting the occurrence of each Chinese character which exists in the Chinese interpretations of each English lexical unit. The interpretation(s) which fall(s) in the determined dominating context is selected to be the intended sense of a lexical unit. A paper reporting on the implementation of the word sense disambiguation method is in preparation.

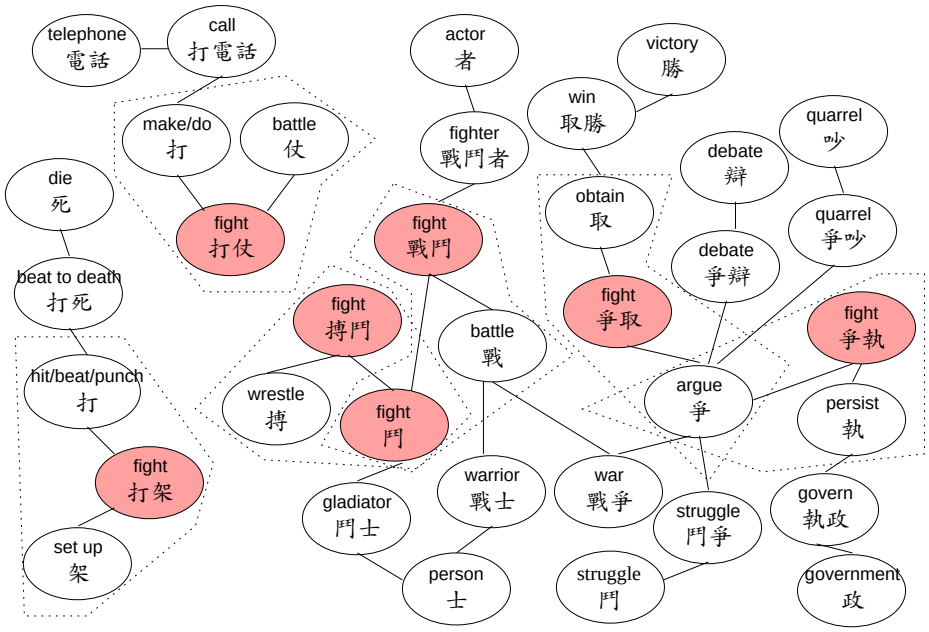


Fig. 2. Various senses of the English word 'fight' and their Chinese counterparts

5 Experimental Results

An experiment was carried out to test the effectiveness of the above approach to word sense disambiguation. Seven pieces of sample texts were taken from the NIV Bible [13], amounting a total of 62 sentences or 1313 words.

With regard to the target disambiguation lexical unit fight, the occurrences of fight in four pieces of the sample texts were referring to "*attempt to harm or gain power over an adversary by blows or with weapons*"[4]. The fight sense used in another two pieces of the sample texts refers to the sense of "*contending for, by or as if by combat*" or "*waging or carrying on (a battle)*.". The sense of fight used in the remaining piece of the sample texts can be interpreted as either of the above two main senses of fight. The disambiguation process correctly ruled out most of the inappropriate senses of fight in each case. In addition, the appropriate sense of lexical units like beat, blow, chariot, hit, loss, march, officer, plunder, strike are also correctly identified.

The disambiguation algorithm chooses the Chinese words 打, 打仗 and 打架 for intrepretting the sense "*to attempt to harm or gain power over an adversary by blows or with weapons*" of fight in Exodus 2:13, Joshua 10:2–7 and Exodus 21:20–22. The most frequently occurring character in the corresponding pieces of sample texts is 打 (*to hit*). In our approach, this character represents the prominent primary concept which dictates the context of the sample texts. Both 打 and 打仗 intrepret the sense of fight appropriately, but the interpretation 打架 is not desirable. This is chosen because this concept is also made

[4] Note that this sense of fight does not entail an organised military campaign.

up of the primary concept 打 (*to hit*) and, at present, the challenge poses by polysemous characters (e.g. 打) has been ignored.

In summary, taking note of the 45 lexical units whose interpretations were affected by the disambiguation process, 37 of them were appropriately interpreted within the context of our sample texts. Only 3 of them did not contain the best available interpretations.

Before the disambiguation, a total of 189 concepts were associated to the 45 lexical units; during the disambiguation, 125 of these concepts were ruled out. This means that, on average, our method reduced an ambiguous lexical unit of 4.2 interpretations to 1.4 interpretations even without considering part-of-speech information. Amongst the 64 remaining concepts, 56 of them appropriately interpreted the lexical units within the context of our sample texts[5]. Only 8 of them can be considered as inappropriate interpretations. Thus, by considering context information (as displayed by concept relatedness in Chinese) alone, our approach achieves 87.5% correctness in word sense disambiguation.

6 Conclusion

Based on the arbitrariness in concept classifi cation observed in WordNet 1.5 and EuroWordNet 2, we have argued that concept formation in natural languages is a plausible means to improve concept relatedness in lexical databases. We have illustrated that word formation in Chinese exhibits natural semantic relatedness amongst Chinese concepts.

Lexical databases are good sources of lexical knowledge for domain-independent word sense disambiguation. To achieve good results, it is therefore vital for a lexical database to be as complete and coherent as possible. We have demonstrated that a method which simply exploits sense relatedness displayed naturally amongst Chinese words can aid word sense disambiguation. We believe enriching concept relations within existing lexical databases using relations inspired by sense relatedness in Chinese is worth pursuing. We propose that such a sense relatedness should be included in enhancing WordNet-like lexical databases.

7 Acknowledgments

The initial ideas of this paper spring from discussions with Doc. Karel Pala in summer, 2002. The author would like to thank him for his invaluable advice and encouragement on this research work.

References

1. Dong, Z., Dong, Q.: HowNet. [Online] Available at: `http://www.keenage.com/zhiwang/e_zhiwang.html` [7 June, 2001] (1999).
2. Fellbaum, C., ed.: WordNet: An Electronic Lexical Database. MIT Press (1998).
3. Vossen, P., et al.: Final report on EuroWordNet 2. Technical report, University of Amsterdam, Amsterdam (1999) [CD ROM].

[5] Note that 46 of the 56 appropriate interpretations came from the 37 appropriately interpreted lexical units.

4. Yu, J., Liu, Y., Yu, S.: The specification of the Chinese Concept Dictionary. Journal of Chinese Language and Computing **13** (2003) 177–194 [In Chinese].
5. Wong, P. W., Fung, P.: Nouns in WordNet and HowNet: An analysis and comparison of semantic relations. In Singh, U.N., ed.: Proceedings of the First Global WordNet Conference 2002, Mysore, 21–25 January 2002, Mysore, Central Institute of Indian Languages, Central Institute of Indian Languages (2002) 319–322.
6. Wong, S. H. S., Pala, K.: Chinese Radicals and Top Ontology in WordNet. In: Text, Speech and Dialogue—Proceedings of the Fourth International Workshop, TSD 2001, Pilsen, 10–13 September 2001. Lecture Notes in Artificial Intelligence, Subseries of Lecture Notes in Computer Sciences, Berlin, Faculty of Applied Sciences, University of West Bohemia, Springer (2001).
7. Wong, S. H. S., Pala, K.: Chinese Characters and Top Ontology in EuroWordNet. In Singh, U.N., ed.: Proceedings of the First Global WordNet Conference 2002, Mysore, 21–25 January 2002, Mysore, Central Institute of Indian Languages, Central Institute of Indian Languages (2002) 224–233.
8. Harbaugh, R.: Zhongwen.com – Chinese Characters and Culture. [Online]. Available at: `http://www.zhongwen.com/x/faq6.htm` [21 August, 2003] (1996).
9. Lu, A. Y. C.: Phonetic Motivation – A Study of the Relationship between Form and Meaning. PhD thesis, Department of Philology, Ruhr University, Bochum (1998).
10. Goldsmith, J.: Unsupervised learning of the morphology of a natural language. Computational Linguistics **27** (2001).
11. Klímova, J., Pala, K.: Application of WordNet ILR in Czech word-formation. In M., G., et al., eds.: Proceedings of the Second International Conference on Language Resources & Evaluation (LREC 2000), Athens, 31 May – 2 JUNE 2002, ELRA – European Language Resources Association (2000) 987–992.
12. Alan, J., ed.: Topic Detection and Tracking – Event-based Information Organization. The Kluwer Internation Series on Information Retrieval. Kluwer Academic, Norwell, MA (2002).
13. International Bible Society, ed.: The NIV (New International Version) Bible. 2nd edn. Zondervan, Grand Rapids, MI (1983) [Online]. Available at: `http://bible.gospelcom.net/` [13 August, 2003].

Finding High-Frequent Synonyms of A Domain-Specific Verb in English Sub-Language of MEDLINE Abstracts Using WordNet

Chun Xiao and Dietmar Rösner

Institut für Wissens – und Sprachverarbeitung,
Universität Magdeburg
39106 Magdeburg, Germany
Email: xiao@iws.cs.uni-magdeburg.de, roesner@iws.cs.uni-magdeburg.de

Abstract. The task of binary relation extraction in IE [3] is based mainly on high-frequent verbs and patterns. During the extraction of a specific relation from MEDLINE[1] English abstracts, it is noticed that besides the high-frequent verb itself which represents the specific relation, some other word forms, such as the nominal and adjective forms of this verb, as well as its synonyms, also play a very important role. Because of the characteristics of the sub-language in MEDLINE abstracts, the synonym information of the verb can not be obtained directly from a lexicon such as WordNet[2] [1]. In this paper, an approach which makes use of both corpus information and WordNet synonym set (WN-synset) information is proposed to find out the synonyms of a domain-specific verb in a sub-language. Given a golden standard synonym list obtained from the test corpus, the recall of this approach achieved 60% under the condition that the precision is 100%. The verbs corresponding to the 60% recall cover 93.05% of all occurrences of verbs in the golden standard synonym list.

1 Introduction

The rapid growth of the size of digital databases inspired the research on automatic information extraction (IE) instead of the traditional manual IE. With the development of natural language processing techniques, more and more tools and resources are available, which leads to fruitful applications in the IE domain. Recent years the IE in biomedical domain has been also very well researched, particularly the task of named entity (NE) recognition. Moreover, relation extraction and event extraction have been also investigated.

Relation extraction is a main task of IE, as defined in the Message Understanding Conferences (MUCs) [3]. In recent years, the extraction of protein-protein interactions in biomedical articles and abstracts are reported in many works such as [2,4,5,6,7]. In this work, the relations to be extracted are binary ones, and the frequently occurring verbs as well as patterns are used in order to construct the template elements of the relations which will be extracted.

[1] PubMed offers free access to MEDLINE, with links to participating on-line journals and other related databases, available at http://www.ncbi.nlm.nih.gov/PubMed/

[2] http://www.cogsci.princeton.edu/~wn/index.shtml

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 242–247.

From the most frequent domain-specific verbs[3] in biomedical texts, we can learn the most frequent relations in this domain. From a test corpus with 800 MEDLINE abstracts extracted from the *GENIA Corpus V3.0p*[4], we can see that "induce", "mediate", "affect", and etc. are the most frequent domain-specific verbs in MEDLINE abstracts. Those high-frequent domain-specific verbs can be semantically categorized. For instance, the verbs such as "activate", "associate", and "interact" were used as the key verbs in extracting the protein-protein interactions in [2,4]. Theoretically, even given a complete lexicon which contains all the lexical entries, the categorization of the verbs in a corpus could still not be solved perfectly, if additional contextual cues are not available. Because many words are polysemous, i.e. they have more than one semantic interpretation, contextual information is necessary for disambiguation. In fact, we do not have such a perfect lexicon, even WordNet, therefore the situation is much more difficult.

In our experiment, we aimed to extract the inhibitory relation in MEDLINE abstracts, since this relation is one of the basic relations in the biomedical domain[5]. This work is based on some previous works such as NE recognition, part of speech tagging, even shallow or full parsing, etc. A very fundamental problem in this relation extraction task is how to choose the proper high-frequent verbs that represent an inhibitory relation.

Obviously the synonyms of the verb "inhibit" have to be taken into account, according to the synonym information provided by a lexicon such as WordNet. But the vocabulary of the sub-language of MEDLINE abstracts seems quite different compared to the general English[6]. Many of the synonyms of the verb "inhibit" provided by WordNet (Version 1.7.1) do not occur even once in the 800-abstract test corpus, such as "subdue", "conquer", etc. Some of these synonyms occur only with a very limited frequency, e.g. "confine" occurs only once in the test corpus. Instead, what can be found in the test corpus are verbs such as "block", "prevent", and so on, as example 1 shows. They are not in the synonym list of "inhibit" in WordNet but provide cues of an inhibitory relation.

Example 1. *Aspirin appeared to* **prevent** *VCAM-1 transcription, since it dose-dependently inhibited induction of VCAM-1 mRNA by TNF.*

Following shows the occurrences of some WordNet synonyms (WN-synonyms) of "inhibit", as well as some non-WordNet synonyms (nonWN-synonyms) in the 800-abstract test corpus.

- **WN-synonyms** suppress (69), limit (16), restrict (5)
- **nonWN-synonyms** block (124), reduce (119), prevent (53)

In addition, we found although the nominal forms of *"inhibit"* are more frequent than the verb forms, the verb "inhibit" occurs quite frequently in the test corpus. It is different from the familiarity description of "inhibit" in WordNet, which says *"inhibit used as a verb is*

[3] Actually, the domain-specific verbs should not include the general verbs independent of the domain in the scientific papers, such as "analyze", "indicate", "observe", and so on. Spasić et al. [8] also discussed this problem.

[4] GENIA project, available at `http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/`

[5] In some relation extraction works, inhibitory relation is treated as a kind of protein-protein interaction.

[6] WordNet is regarded as a semantic lexicon for general English, since its sources are quite broad [1].

rare". And we found that the "estimated frequency" in WordNet differs from that in the sub-language of MEDLINE abstracts. For instance, in WordNet, "restrain" is more frequent than "limit", but in the test MEDLINE abstract corpus, the situation is just reversed. This indicates that the expressions in the sub-language of MEDLINE abstracts are quite domain-specific.

This paper proposes an approach in order to find out these synonyms in the sub-language. It is constructed as follows: section 2 describes the approach of finding the synonyms of a verb in the sub-language of MEDLINE abstract, and, section 3 presents the result and discussion.

2 Finding Out Synonyms in Sub-language Corpus

Definition: Keyword, Core Word, and Language Unit In this experiment, let *keyword* denote a word whose base form is "inhibit", while *core word* denotes the verb "inhibit". For example, "inhibitory" and "inhibition" are both keywords in this experiment. A *language unit* may be a sentence, several sentences, or a paragraph, even several paragraphs, which expresses the same semantic topic.

In order to find out the synonyms of the core word, with the help of WordNet information, the corpus information is also considered. In this test the verbs which occur around a keyword in the text of an abstract are examined.

This idea comes from the assumption that the synonyms of a verb, which have very close semantic relation with its corresponding keyword, have a likelihood to co-occur in the same language unit with the keyword than with other words. Note that in our approach only the localization of all the verbs around the keyword is considered. Other information such as the sentence boundaries and sentence structures, are not considered yet, although they must be very useful in some other corpora. Because in MEDLINE abstract corpus, each abstract consists of only one paragraph, namely several sentences[7], and each abstract either has only one topic, or the topics in an abstract are dependent on each other, then the whole abstract can be treated as a language unit. The vocabulary of a language unit is limited heavily by the topic(s), which means it is very likely that the vocabulary consists of words that have close semantic relations to each other in a language unit. Namely, the vocabulary in the same language unit can be more probably grouped into fewer synonym or antonym sets. Moreover, with the localization of a keyword, the verbs around the keyword may be limited semantically to have semantic relations (synonyms or antonyms) with the keyword[8].

2.1 Method and Resources Used in The Experiment

Golden Standard List (S_G) for Evaluation At first a synonym list of the verb "inhibit" is obtained by counting the frequencies of each verb in a manually produced 50-synonym list in the test corpus, based on WN-synset information, and choosing the ones with more than 6 occurrences. By this process a 10-word synonym list is obtained, which is used in the following work as a golden standard list S_G. In S_G only 3 verbs come directly from the WN-synset of "inhibit", but the rest 7 verbs come from its hypernyms and the synonyms'

[7] For the 800-abstract test corpus, each abstract consists of 8.41 sentences in the average, excluding the title of each abstract.

[8] Because of the restriction of the pages, an example here is omitted.

synonyms. This golden standard list provides the standard to evaluate this approach.

Expansion of Synonym List (S_i): Learning Synonym Information from WordNet In order to make use of the WN-synset information, the synonyms of each word which is a synonym of "inhibit" are considered in order to improve the coverage of synonyms in the MEDLINE abstract corpus. Let S_i $(i > 0)$ be the expanded WN-synset word list, it can be obtained in the following way: at first the synonym list of "inhibit" is expanded by adding all synonyms of this verb, the list contains 16 items by then, which is symbolized as S_1. Furthermore, S_1 can be also expanded by adding all synonyms of each verb in the list, the list is then expanded to be a 94-item one, i.e. S_2. If we want to enhance the recall, we can just expand this synonym list by recursively adding the complete synonym list of each word in this list again, and go on. But at the same time the misleading information will grow in an exponential way.

Verb List (V_j) from the Test Corpus: Collecting Verb Candidates (S_g) We can get a set of verbs (V_j) which are chosen from the test corpus around a keyword in the window size of j $(j > 0)$, with the corresponding frequencies from the test corpus. The list provides the corpus information in our experiment. In the 800-abstract test corpus, for example, there are total 318 verbs around the keyword in a searching window of size 2. In these 318 verbs, the occurrences of 23 verbs are $\geq$ 26 times. It is quite surprising that in these 23 words, 9 of them are synonyms or antonyms of the verb "inhibit", including the verb itself. The expanded synset lists S_i $(i > 0)$ are used to give synonym information of the high-frequent verbs around a keyword. If a high-frequent verb around a keyword or one of the synonyms of this high-frequent verb is in this synonym list, it will be added to the learnt synonym candidate list S_g.

Expansion of Misleading Verb List ($STOP_k$): Learning Misleading Information from Genre Analysis of Corpus and WordNet Because the sub-language in MEDLINE abstracts quite often uses the verbs to construct the whole abstracts structure, such as "suggest", "indicate", "show", and so on, they should be excluded from S_g. An initialized stop-word list $STOP_0$ is given with 15 such verbs (including several antonyms of "inhibit") in this experiment. However, the necessary expansion of the stop-word list $STOP_k$ $(k \geq 0)$ is carried out also in a similar way as the expansion of S_i. If a verb v, $v \in V_j$ and $v \in STOP_k$, then $S_g = S_g - \{v\}$.

Balance between Recall and Precision This approach is a bidirectional one. That is, in one direction the positive synonym information is expanded according to WN-synsets, or the searching windows are enlarged, so that the recall will be improved but the precision will be impaired; in the other direction, the stop-word list is also expanded in order to improve the precision, meanwhile the recall will be impaired. Therefore, the balance between recall and precision is also very important. That means, the expansions of both the synonym list and the stop-word list are limited. For instance, in this experiment, the synonym list has been expanded for maximal 4 times (S_i, $i = 1...4$), whereas the stop-word list has been expanded only once ($STOP_1$). In addition, by only focusing on the relative high-frequent words in this experiment, the work of evaluating recall and precision is much simplified.

3 Result and Discussion

This approach makes use of three kind of sources. One is the synonyms information of the verb “inhibit” obtained independently from any corpus but from a lexicon (WordNet). The second is the frequencies of verbs around a keyword, which depends closely on the corpus. The last is the information of unlikely verbs, which depends partly on the verb “inhibit” itself, i.e. its antonyms, and partly also on the corpus, i.e. the verbs for the construction of MEDLINE abstracts.

Table 1. Recall (R_j) and precision (P_j) on synonym list S_i ($i = 1, ..., 4$), in searching window with window size j ($j = 1, ..., 5$). The word frequency limit in this table is ≥ 15 in the test corpus, with an expanded stop-word list of 256 items (first part of this table) and 1512 items (second part of this table), respectively.

256	R_1	P_1	R_2	P_2	R_3	P_3	R_4	P_4	R_5	P_5
S_1	20%	100%	40%	100%	40%	100%	40%	100%	40%	100%
S_2	30%	100%	**60%**	**100%**	60%	100%	60%	100%	60%	100%
S_3	30%	100%	60%	100%	60%	100%	60%	85.71%	60%	85.71%
S_4	30%	100%	60%	85.71%	60%	85.71%	60%	75%	60%	75%

1512	R_1	P_1	R_2	P_2	R_3	P_3	R_4	P_4	R_5	P_5
S_1	10%	100%	20%	100%	20%	100%	20%	100%	20%	100%
S_2	10%	100%	**30%**	**100%**	30%	100%	30%	100%	30%	100%
S_3	10%	100%	30%	100%	30%	100%	30%	100%	30%	100%
S_4	10%	100%	30%	75%	30%	75%	30%	75%	30%	75%

Look at the data with 256 stop words in Table 1, with the increase of expansion of both synonym and stop-word lists, the recall comes to 60% stably, in which only 33.4% comes directly from the WN-synset of “inhibit”. And in the test corpus, the verbs corresponding to the 60% recall cover 93.05% of all occurrences of verbs in the golden standard list, this means that this approach finds out the most frequent synonyms of “inhibit” in the test corpus. It also indicates that these high-frequent synonyms distribute mainly in ± 2 positions around a keyword. Note that here *position* refers to a verb chunk around a keyword. In comparison to the data with 1512 stop words, the data with 256 stop words indicate when the stop-list is too large, it causes the decrease of recall sharply. Then the stop-word list should not be expanded too much so that the intersection of $STOP_k (k > 0)$ and $S_i (i > 0)$ can be minimized.

By this approach, it should be possible to semantically classify the high-frequent domain-specific verbs in MEDLINE abstracts for further IE tasks. However, this approach is limited to be applied in MEDLINE abstract corpus. Second, the core word occurring in the test corpus should not be too sparse. In case that the core word occurs with a low frequency in the test corpus, its synonyms with high frequencies should be considered instead. Since this approach focuses only on the high-frequent verbs in the corpus, the recall is rather moderate. In future work it will be investigated how syntactic cues and information from phrase patterns could improve the recall.

References

1. Christiane Fellbaum: WordNet: An Electronic Lexical Database. The MIT Press. Cambridge. Massachusetts (1998) Foreword, xv–xxii, Chapter 1, 23–46.
2. T. Sekimizu, H. S. Park and J. Tsujii: Identifying the Interaction between Genes and Gene Products Based on Frequently Seen Verbs in Medline Abstracts. Genome Informatics, Universal Academy Press (1998).
3. Jim Cowie, Yorick Wilks: Information Extraction. Handbook of Natural Language Processing (2000) 241–160.
4. J. Thomas, D. Milward, C. Ouzounis, S. Pulman, M. Carroll: Automatic extraction of protein interactions fromscientific abstracts. In: The Pacific Symposium on Biocomputing'2000, Hawaii (2000) 541–551.
5. A. Yakushiji, Y. Tateisi, Y. Miyao, J. Tsujii: Event Extraction from Biomedical Papers Using a Full Parser. In: The Pacific Symposium on Biocomputing. (2001) 6:408–419.
6. R. Gaizauskas, K. Humphreys, G. Demetriou: Information Extraction from Biological Science Jounal Articles: Enzyme Interactions and Protein Structures. In: The Workshop Chemical Data Analysis in The Large: The Challange of The Automation Age. (2001).
7. T. Ono, H. Hishigaki, A. Tanigami, T. Takagi: Automated extraction of information on protein-protein interactions from the biological literature. Bioinformatics, 17(2)(2001) 155–161.
8. I. Spasić, G. Nenadić, and S. Ananiadou: Using Domain-Specific Verbs for Term Classification. In: The ACL 2003 Workshop on NLP in Biomedicine. Sapporo, Japan, (2003) 17–24.

Part III

Posters

Adjectives in RussNet

Irina Azarova[1] and Anna Sinopalnikova[1,2]

[1] Saint-Petersburg State University, Russia
Email: azic@bsr.spb.ru
[2] Masaryk University, Brno, Czech Republic
Email: anna@fi.muni.cz

Abstract. This paper deals with the problem of structuring adjectives in a wordnet. We will present several methods of dealing with this problem based on the usage of different language resources: frequency lists, text corpora, word association norms, and explanatory dictionaries. The work has been developed within the framework of the RussNet project aiming at building a wordnet for Russian. Three types of relations between descriptive adjectives are to be discussed in detail, and a technique for combining data from various resources to be introduced.

1 Introduction

Up to date presenting adjectives within a wordnet remains one of the most difficult and disputable matters of the lexical semantics.

Although there is no common solution for structuring adjectives in wordnets, some general considerations are adopted by most of the researchers. **Firstly**, it is generally accepted that being a 'satellite' words, adjectives posses very specific meaning (vague, highly dependent on the meaning of accompanying nouns). It is usually stressed that adjectives, descriptive ones, in particular, have no denotation scope of their own. **Secondly**, due to their specific semantic and syntactic properties, semantic organization of adjectives is entirely different from that of other open classes of words. Thus, **thirdly**, methods of revealing the semantic organization for nouns and verbs do not hold for the adjectives [1,2,3].

Adopting these statements as a base of our research, we are to describe the ways semantic organisation of Russian descriptive adjectives is examined. Although the facts discovered could not be expanded on all other languages, the methodology applied is of a scientific value and may contribute significantly to the standards of wordnet building.

2 Frequency List Study

Usually a wordnet building process starts with the analysis of most frequent words (extracted either from corpora [4], or explanatory dictionaries [4,5]) in order to obtain the list of general concepts representing the core structure of a language, so-called Base Concepts.

In addition to its main task performing, the frequency analysis yields many subsidiary results that are useful for the next stages of wordnet constructing. As far as frequency lists of Russian [6,7] concern, it appears that among more than 6500 adjectives given descriptive ones occupy most positions, including the 76% of the 50 top positions.

The following conclusions could be made:

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 251–258.

Table 1. Top frequent Russian adjectives in a large corpus (according to [7]).

Rank	Word	Eng	Ipm	Rank	Word	Eng	Ipm
62	большой	big, large	1630.96	150	последний	last	630.17
114	хороший	good	853.71	180	старый	old	528.25
116	новый	new	840.18	194	белый	white	493.36
128	конечный	final, last	732.33	203	главный	main	467.77
137	нужный	necessary	690.34	224	маленький	small	411.52

1. The fact discovered confirms the general view of descriptive adjectives as the 'most typical' representatives of this PoS.
2. High frequency of a certain adjective doesn't indicate whether it is caused by its numerous senses or by its preferential status, or by both simultaneously.
3. The adjective's frequency reveals which member of an antonym pair is marked, being more common. The detailed corpora analysis, e. g. usual position of some adjective after the negative particle *не* ('not'), allows us to define precisely which antonym is semantically marked. The positive value of some parameter is usually supposed to be prone to a markedness, e. g. an opposition between 'big' (*большой*) and 'small' (*маленький, малый*). The information of an antonym's 'markedness' is to be used while generating appropriate definitions for adjectives (see the last section).
4. Frequency data helps us to set order into the synsets, to establish the priority of synonyms from the viewpoint of their usage. Being a neutral term, dominant synonym is expected to occur in texts more often then other members of the corresponding synset.
5. Frequency data allow us to verify the hypothesis of the correlation between two modes of synset organization: from the most frequent synonym to less frequent ones, and from a neutral dominant synonym to expressive and terminological ones.

3 Distinguishing word senses

According to the data shown in Table 1, adjective *большой* ('big/large') is the most frequently used Russian adjective. The fact calls for an explanation, regarding that *большой* usually considered to denote so-called visual assessment of size, which is narrower than that of the adjective *хороший* ('good'), ordinarily said to indicate a general assessment of an object, event, or quality. This situation may be accounted for either by high ambiguity of the adjective *большой*, or by the more abstract nature of this adjective.

To specify and to distinguish between word senses of *большой*, we apply 2 language resources: text corpus[3], and association tests[4]. Extracting from both resources data on syntagmatic properties of the adjective, e. g. selectional restrictions, we base our case study on the general consideration: "Every distinction in a meaning is reflected by distinctions

[3] A balanced corpus of Russian texts for the study includes about 16 mln words. Texts belonging to different functional styles were taken in the following proportions: fiction –20%, newspapers and magazines – 40%, popular science texts – 30%, laws – 10%. The time boundaries are defined as 1985–2003.

[4] RWAT – The Russian Word Association Thesaurus by Karaulov et al. [8] and RWAN – Russian Word Association Norms by Leontiev et al. [9] were used.

in form" separately made by many of the linguists working in the area of corpus-based lexicography [10,11].

In our research we focus mainly on the lexical and semantic context markers, and partly domain ones. The analysis of noun collocations with the adjective *большой* is to assist to reach a decision regarding the number of word senses, which should be distinguished in the RussNet.

From RWAT we extract noun-responses of *большой* combining freely with the adjective in question (ignoring idioms like *Большой театр, большой палец*). Noun-responses may be organized into several groups:

(1) spatial artefacts (*house, town, shop*, etc.);
(2) three-dimensional natural objects (*forest, ball, mushroom*, etc.);
(3) animals (*bear, elephant*, etc.);
(4) two-dimensional objects (*sheet, circle*);
(5) persons (*man, boy, son*);
(6) personal characteristics (*friend, fool, coward*, etc.);
(7) parts of human body (*nose, mouth*);
(8) abstract nouns (*brain, experience, talent*, etc.).

By summing up associations in groups (including unique ones) we distinguish those three, which are the most numerically strong: 1, 6, 8. Checking these data across the corpus, we receive the same leading groups of nouns, the top frequent collocants of *большой* being: *money* (127), *man* (39), *eyes* (36), *problem* (22), *opportunity* (21), *hope* (20), *group* (18), *town* (13), *loss* (13), *difficulty* (12), *distance* (11), etc.

Thus, on the base of facts discovered we may draw a conclusion that the most frequent sense of the adjective *большой* (according to the corpus and RWAT data) is the 'indication to the above-average spatial characteristics of an object'. That holds for both natural objects (including animals) and artefacts, the last including objects with absolute above-average size, e.g. *дворец* 'palace', *город* 'city', *слон* 'elephant', *самолет* 'aeroplane', as well as with relative one, e.g. *капля крови* 'blood driblet', *прыщ* 'smirch', *гриб* 'mushroom', etc. It is in this particular sense *{большой$_1$}* is related to its augmentative hyponym *{огромный$_1$, громадный$_1$}* 'very big' and antonym *{маленький$_1$, малый$_1$}* 'of a minor, less than average size'.

First sense covers its usage with noun-groups (1), (2), (3), (4), (7). Other senses manifested are (ordered by frequency):

– With nouns from group (8) *большой$_2$* signalizes 'above-average level of quantifying features [intensity, duration, importance] of some event or state', e. g. *большая проблема, большие сложности.*
– With nouns from group (6) *большой$_3$* is used for indicating to 'high intensity of some human's trait' mentioned by a noun, e.g. *большой друг.*
– With several nouns from group (5) pointing to children *большой$_4$* refers to 'grown up from infancy', e.g. *большой мальчик.*

4 Establishing Relations

As we have shown in the previous section, both the RWAT and our corpus supply us with the evidences on the syntagmatic relations of the adjectives. But they also allow us to observe their paradigmatic relations as well.

Regarding the frequency of words from the same PoS (probably, paradigmatically related to adjectives under consideration), we may conclude that paradigmatic relations are highly relevant for adjectives: *большой –> маленький 47, огромный 15, малый 12, толстый 6, высокий, длинный, крупный 3,* etc. (the total amount of associations in RWAT counting 536); and *большой – маленький* 98 (MI = 6.072), *малый* 69 (MI = 7.728), *крупный* 15 (MI = 4.095), *мелкий* 15 (MI = 4.817) etc. out of total amount of 9762 lines in the corpus.

1. These lists of co-occurring words give us a hint on what adjectives could belong to the same semantic field, or to the same hyponymy tree. Thus, for example, we may conclude that *маленький, огромный, малый, толстый, высокий, длинный,* etc. probably belong to the same semantic field as *большой.*
2. Comparing the context patterns (see Section 3) for these adjectives, we are able to establish links between them and to organize them into tree structures.

The general approach to this task performance suppose the fulfilment of following conditions:

– To establish a **Hyponymy** link we need the evidences in favour of context inclusion, see Section 4.1.
– **Antonymy** relations are often characterised by the identical contexts. Antonymous adjectives also may co-occur in contrastive sentences ('and/or/but'), e.g. *большие и малые программы, нажимать большие или маленькие кнопки* or *план большой, а зарплата маленькая.* See Section 4.2.
– For **synonymous adjectives** identity of contexts is believed to be quite a rare phenomenon, rather we observe incompatible contexts (complementary distribution), e.g. *незамужняя женщина* and *неженатый мужчина.* As an additional criterion we may rely upon co-occurrence of synonyms in enumerating phrases (e.g. *большой, крупный нос*). See Section 4.3.

4.1 Adjectives and Hyponymy

Following the GermaNet proposal to "make use of hyponymy relations wherever it's possible" [12], in RussNet we adopt formal approach based on the adjective collocations with nouns. Empirical data proves that in Russian it's the adjective that predicts the noun (class of nouns) to collocate with, not vice versa, e. g. *долговязый (ланкы, страппинг)* involves the pointer to a human being, i. e. it can collocate with such nouns as *мальчик* (*a boy*), *человек* (*a man*).

Thus, the main idea underlying our work is that **hyponymy tree** for descriptive adjectives may be built according to that of nouns: i. e. if 2 adjectives from the same semantic field collocate with 2 nouns linked by the hyponymy, we are to build the hyponymy link for these adjectives [13].

We consider the procedure for retrieving the information about hyponyms using the above mentioned adjective *большой*. There are several multiple adjective responses in the RWAT: *огромный* 'huge', *толстый* 'thick', *круглый* 'round', *высокий* 'high', *длинный* 'long', *крупный* 'large-scale', *сильный* 'strong', *красивый* 'nice', *необъятный* 'immense'. The next step is to specify weather these responses are syntagmatic or paradigmatic. For that purpose we apply to the corpus-driven data on adjective co-occurrences. It appears, that some adjectives do collocate with *большой* in our corpus, e.g. *толстый* 'thick' and *круглый* 'round', however, *красивый* 'nice' occurs 4 times with rather high MI-score (8.063). Also syntagmatic relations are manifested by associations with a copulative conjunction *и* 'and' in RWAT, e.g. *и красивый, и круглый*. Thus, we could exclude adjectives *красивый* and *круглый* from paradigmatic associations, consider *огромный*, *высокий*, *длинный*, *крупный*, *сильный*, *необъятный* to be paradigmatic, and *толстый* – ambivalent.

Lists of word associations for *высокий*, *длинный*, *сильный* look nearly-identical: their leading responses are nouns (*путь* 55; *человек* 54), and antonymous adjectives (*низкий* 48; *короткий* 54; *слабый* 42), while for *огромный*, *крупный* and *необъятный* the leading responses compose *большой* and nouns. The former fact may evidence in favour of a hyponymy link, the latter one may count for synonymy or hyponymy. An ambivalent adjective *толстый* has a structure of the first type.

4.2 Adjectives and Antonymy

Although in Princeton WN antonymy is regarded as a relation between words rather than synsets, in RussNet antonymy is considered to be one of the semantic relations between **synsets**.

Yet we by no means are to reject the differentiation of **direct and indirect antonymy**. We suppose that setting order into a synset helps us to manage this problem adequately. As RWAT shows, in Russian it is usually synset representatives ('dominant literals') that are related by antonymy directly, all other members of synsets are opposed through this pair, i.e. indirectly. E. g. *большой* is strongly associated with *маленький, маленький* is associated with *большой*, while *малый* is associated first of all with *маленький*, its association with *большой* is rather weak. But there still is a possibility that several pairs of direct antonyms may appear in the frame of two synsets, like in English *ларге* ↔ *смалл, биг* ↔ *литтле*. However, our study of 533 most frequently used descriptive adjectives (on the basis of RWAT) proves this phenomenon is not that characteristic for Russian.

4.3 Adjectives and Synonymy

In its first and second senses *большой* is a dominant of synsets. As syntagmatic data driven from RWAT and the corpus show, these synsets may include an adjective *крупный* as well. **Firstly**, this adjective occurs regularly as a response to *большой* in the RWAT, it belongs to the 10 most frequent ones. Also regarding backward associations, we discover that *большой* is the first and hence, the most strong, response to *крупный*. The same observation holds for *огромный* and *громадный,* but as opposed to *крупный* both this adjectives fail the implicative synonymy test. E.g. *Большая сумма денег* ⇔ *Крупная сумма денег,* but

Огромная сумма денег ⇒ *Большая сумма денег*, and not vice versa. **Secondly**, comparing syntagmatic associations of *большой* and *крупный*, we observe a signifi cant overlap of the lists. Some responses (∼21%) literally coincide, e. g. *человек, город, нос, выигрыш, успех, специалист*, many others are semantically similar (i. e. belong to the same semantic fi eld) e.g. *разговор, план*, etc. So do the micro-contexts patterns for these adjectives. **Thirdly**, more detailed study of the corpus proves that *крупный* is used mainly in specifi c domains: commerce and fi nance texts, e.g. *крупный бизнес, крупный московский автоторговец, крупный производственный филиал, крупный "рынок"* и т. д. Thus, it is clear, that in the corpus the adjective *крупный* occurs far less frequent than *большой* (3882 lines against 19566). **Fourthly**, in most of the observed contexts *крупный* may be easily substituted by *большой*. **Fifthly**, analysis of defi nitions from Russian explanatory dictionaries [14,15] shows the signifi cant overlap in structure of several defi nitions given to *крупный* and *большой*.

As a side result of the analysis we also observe that the fi rst sense given in the dictionaries for *крупный* 'consisting of large particles or objects of above-average size' (*крупный песок, жемчуг*) includes an indication to an aggregate or collection of identical or similar units, that could not belong to the same semantic fi eld as *большой*$_1$. This is confi rmed by the substitution test: *крупный песок*, but **большой песок.* The priority of that sense is not supported by the actual data: in RWAT nouns illustrating this sense of *крупный* (*дождь, снег, град, виноград, корм, порошок, шрифт, слезы*) are obviously peripheral – their absolute frequency never exceeds 5, and their number gives only 2.7% of total amount of responses. Frequency data counts against the actual priority of the historically original 'aggregate' sense: *крупный* is used less frequent in this sense, so it should be treated within a wordnet as a secondary (*крупный*$_3$).

All the facts discovered – similar meanings, substitutability, similarity of responses in RWAT and contexts in the corpus, domain markedness of *крупный* and neutrality of *большой* – enable us to conclude that the adjective *крупный* belongs to the same synsets as *большой*$_1$ and *большой*$_2$. According to the data on usage, the synsets should be ordered as follows: {*большой*$_1$, *крупный*$_1$}; {*большой*$_2$, *крупный*$_2$}.

5 Generating Appropriate Defi nitions

As for the adequate representation of systemic relations of adjectives, defi nitions given in conventional dictionaries are considered to be inconsistent and insuffi cient. The possible explanation for that lies in the diffi culty of performing this task within the framework of traditional lexicography. Specifi c semantic features of adjectives, such as their mainly signifi cative meaning and absence of clear denotation, dependence on the modifi ed nouns etc. make the traditional methods quite an unreliable base for defi nition generation. In order to construct appropriate defi nitions for adjectives we rely upon their **relations** to each other and to nouns they co-occur with.

The relevance of relations may be rated from the viewpoint of the defi nition generation:

1. For descriptive adjectives **antonymy** is by no means one of the most important and rich in content relations [16,17,18]. Semantic markedness of opposition members determines the direction of the defi nition generation. Unmarked member is to be defi ned through

the marked one (e.g. *истинный* through *ложный*). Their definitions in Princeton WN are reversed: *true* – 'consistent with fact or reality; not false', *false* – 'not in accordance with the fact or reality or actuality'. In case of definition based on the antonymy relation special attention should be paid to cycles, when antonyms are defined through each other.
2. **Hyponymy** seem to be useful for definition construction in cases of augmentative/diminutive hyponyms. For most descriptive adjectives denote various assessments of gradable properties, intensity or mildness is among the most frequent components of their meanings. E.g. *невысокий* – 'not very low'.

The semantic structure of adjectives is considered to be dependent on and specified by the nouns they modify [1]. Thus another necessary contribution to definition generation concerns the coding of meanings of nouns, adjectives co-occur with. The relations within noun–adjective collocations may be divided into several types: goal-instrument e.g. *athletic equipment*, result-cause e.g. *healthy air*, feature-whole *big house*, etc. [3]. Each type of relations requires a specific model of definition (specification of how and to what extent meaning of a co-occurring noun modify an adjective's meaning): *healthy*$_3$ – *promoting health* e.g. *healthy air.*

6 Conclusions and Future Work

Diverse language resources – frequency lists, association norms, corpus analysis – affords us to establish a clear-cut adjective structure in the RussNet (a wordnet for Russian) [19]. The described technique aims at listing different senses of an adjective, differentiating synonymy and hyponymy links, defining antonym pairs, generating proper sense definition explaining the difference between co-hyponyms.

It is important now to apply it consistently to the whole stock of the descriptive adjectives in RussNet, verifying and correcting the method. Using it on the large scale may find difficulties due to the absence of association data, or an insufficient number of occurrences in the corpus for less frequent adjectives.

References

1. Gross D., Fellbaum C., Miller K.: Adjectives in WordNet. International Journal of Lexicography 3 (4) (1990) `ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.ps`.
2. Apresjan Ju. D.: Lexical semantics. Vol. 1–2. Moscow (1995).
3. Willners C.: Antonyms in context: A corpus-based semantic analysis of Swedish descriptive adjectives. PhD thesis. Lund University Press (2001) `http://www.ling.lu.se/education/gradstud/disputation/Omslag.doc`.
4. Vossen, P. (ed.): EuroWordNet: A Multilingual Database with Lexical Semantic Network. Dodrecht, Kluwer (1998).
5. Pala K., Ševeček P.: The Czech WordNet, EuroWordNet (LE-8928). Deliverable 2D014 (1999) `http://www.hum.uva.nl./~ewn/docs.html`.
6. Zasorina L. N. (ed.): Frequency Dictionary of Russian. Moscow (1977) (40.000 entries).
7. Sharoff S. A.: Frequency List of Russian (2000) (35.000 entries) URL: `www.artint.ru/projects/freqlist`.
8. Karaulov Ju. N. et al.: Russian Associative Thesaurus. Moscow (1994, 1996, 1998).

9. Leontiev A. A. (ed.) Norms of Russian Word Associations Moscow (1977) (about 100 entries).
10. Sinclair, J.: Corpus, concordance, collocation. Oxford: Oxford University Press (1991)
11. Apresjan Ju. D.: Systematic lexicography / translated by K. Windle. Oxford University Press. (2000).
12. Naumann K. Adjectives in GermaNet. (2000) http://www.sfs.uni-tuebingen.de/lsd/.
13. Azarova I. et al.: RussNet: Building a Lexical Database for the Russian Language. In: Proceedings of Workshop on Wordnet Structures and Standardisation and How this affect Wordnet Applications and Evaluation. Las Palmas (2002) 60–64.
14. Evgenjeva A. P. (ed.): Dictionary of Russian (vol. 1–4). Moscow (1985–88).
15. Ozhegov S. I., Shvedova N. I.: Explanatory Dictionary of Russian. Moscow (1992).
16. Charles W. G., Miller G. A.: Contexts of Antonymous Adjectives. Applied Psycholinguistics 10 (1989) 355–375.
17. Fellbaum C.: Co-occurrence and antonymy. International Journal of Lexicography 8(4) 281–303.
18. Justeson J. S., Katz S. M.: Co-occurrence of Antonymous Adjectives and Their Contexts. Computational Linguistics 17 (1991) 1–19.
19. RussNet: Wordnet for Russian: URL: http://www.phil.pu.ru/depts/12/RN/.

Towards Binding Spanish Senses to Wordnet Senses through Taxonomy Alignment

Javier Farreres[1], Karina Gibert[2], and Horacio Rodríguez[1]

[1] Computer Languages and Systems Department, Universitat Politècnica de Catalunya, Campus Nord, 08028 Barcelona, Spain, Email: farreres@lsi.upc.es
[2] Statistics and Operations Research Department, Universitat Politècnica de Catalunya

Abstract. This work tries to enrich the Spanish Wordnet using a Spanish taxonomy as a knowledge source. The Spanish taxonomy is composed by Spanish senses, while Wordnet is composed by synsets (English senses). A set of weighted associations between Spanish words and Wordnet synsets is used for inferring associations between both taxonomies.[3]

1 Introduction and Previous Work

This work continues a line of research directed to build Wordnets for languages in an automated way. Trying to delay human intervention as much as possible, a taxonomy alignment is performed. Using a set of associations previously obtained between Spanish words and Wordnet synsets, together with a logistic model that weights those associations, and a Spanish taxonomy of senses extracted with automatic processes, inference of associations from Spanish senses to Wordnet synsets is studied. This work uses results obtained in some previous works, introduced below.

In [Atse98] the interaction between different methods that link Spanish words with WordNet synsets was studied. Intersections between pairs of methods were proposed for maximizing the number of links together with the global accuracy.

In [Farr02] a methodology considering maximal intersections among all the methods in [Atse98] was proposed. A logistic model[4] was obtained for estimating the probability of correctness of a given link.

In [Rigau98] a method is offered for automatically generating a taxonomy of Spanish senses by means of a Word Sense Disambiguation process on the genus of the dictionary definitions. Even though the genus is detected with adequate precision, the sense discrimination has a much higher degree of error. This causes that, when building a complete branch of Spanish senses from the taxonomy, at some point some error will deem a chain of incorrect ancestors.

In [Farr98] some ideas were proposed related to taxonomy alignment. Studying simple geometrical configurations that tie the Spanish taxonomy of [Rigau98] with Wordnet, ways to increase probability of selected associations were proposed, as well as possibilities for inferring new associations.

[3] This research has been partially funded by Aliado (TIC2002-04447-c02-01).

[4] The logistic regression approximates the probability associated with a vector of booleans.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 259–264.

2 Basic Concepts

The terms defined below are used along this paper.

A *Spanish word* is a word covered by a Spanish monolingual dictionary. A *Spanish sense* is a sense of a Spanish word as defined by the Spanish monolingual dictionary, thus it is source dependent. Two kinds of associations are considered. A *WtS* is an association of a Spanish word to a Wordnet synset, with a probability of correctness, named in this paper as the *logistic probability*, calculated with the logistic model obtained in [Farr02]. An *StS* is an association of a Spanish sense to a Wordnet synset. Whenever a Spanish sense has no *StS*, it may always inherit the *WtS* of the word it belongs to. The *branch* starting at some sense is the sequence of ancestors of that sense up to the top, including the sense. A *gap* in a Spanish branch is a Spanish sub-branch that has no association and that separates Spanish sub-branches with associations.

The PRB Given an *StS* c, *PRB(c)* (*pair of related branches*) is defined as the pair of branches developed upward, on the one hand, from the Spanish sense till sixth level (as justified in section 5) and, on the other hand, from the corresponding Wordnet synset up to the top, together with all the associations connecting both branches. See figure 1 for a graphical example.

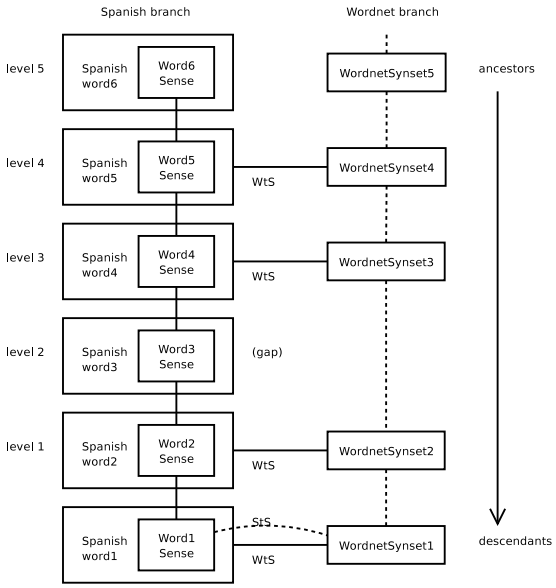


Fig. 1. The PRB

PRB is the concept managed in this work to allow study of the relationship between the Spanish taxonomy and Wordnet.

3 Towards the Induction of Upper Connections

After obtaining 66.000 associations in [Farr02], the only work left seemed to be a manual validation. But upon observing the data, many obviously wrong results were detected, frequently related to *WtS* of Spanish words without any Spanish sense supporting this association. If the Spanish senses could be contrasted with those *WtS*, the obvious errors could be deleted automatically. The natural resource to be applied, and the one that was at our reach, was a Spanish taxonomy of senses, even if generated automatically.

When the alignment of two taxonomies was considered, other useful applications arose as transforming *WtS* into *StS*, inferring *StS* without previous *WtS*, detecting erroneous *WtS* and knowing which Spanish senses remain uncovered.

4 Induction of Basic Connections

As a first stage of this research, monosemic Spanish words with only one *WtS* were considered. For this specific case, *StS* can be directly obtained from *WtS* to produce a starting kernel. Using the Spanish taxonomy as knowledge source, 1263 such monosemic Spanish words were detected, giving 1263 *StS* between 1263 Spanish senses and 1195 Wordnet synsets, that is, 1.06 Spanish senses per Wordnet synset.

From those, 685 *StS* were randomly chosen and manually evaluated, obtaining 559 correct evaluations and 19 incorrect evaluations giving a global accuracy of 96.7%.

5 The PRB Concept

Knowing that the Spanish taxonomy is not error free in the detection of the parent sense, the behavior of the Spanish ancestors of the senses taking part in the *StS* was studied for determining the distribution of errors.

For each *StS* the complete branch of the Spanish sense was built using [Rigau98], the complete branch of the synset was retrieved from Wordnet, and all the associations connecting both branches were identified.

Those parallel branches were classified on the basis of the level of the first association (a *WtS* or a previously identified *StS*) above the base *StS*, and the results are shown in table 1. The case where no ancestor has an association is the one that accumulates the highest number of errors, 14. In few cases the level is above five.

Upon these results, it was decided that further experiments would be carried out with Spanish branches of up to 5 ancestors, while the Wordnet branch would have no limit. The set of parallel branches within these parameters was named *PRB*, and the 685 *StS* were used to generate the corresponding *PRB*s.

6 PRB with Association on the First Level

298 of the *PRB*s generated have and immediate Spanish ancestor with an association above the base *StS*. For those *PRB*s, three parameters have been studied: the cardinality of the relationship between the branches in the *PRB* as defined by the set of associations, the number

Table 1. Level of first ancestor with an association

Level	Count	OK	KO	%
none	159	145	14	91
1	298	298		100
2	70	67	3	95
3	29	28	1	96
4	11	10	1	90
5	6	6		100
7	3	3		100
11	1	1		100
27	1	1		100

of senses with an association in the Spanish branch of the *PRB* and the existence of gaps in the Spanish branch.

The set of associations of each *PRB* defines a relationship between the two branches. The cases below were detected. Table 2:left shows the number of *PRB*s with structure in each of the cases.

***PRB* with crossings:** two Spanish senses have associations that cross each other (see cases *d), e)* in figure 2). There are only 11 cases, probably due to errors in the sense disambiguation process or to differences of lexicalization between the two languages. It was not studied further.

1:1: only one association links any Spanish sense, and only one association links any synset.

1:N: only one association links any Spanish sense, but several associations may link any synset (see cases *a)* to *c)* in figure 2).

N:1: several associations may link any Spanish sense, but only one association links any synset (see cases *f)* to *i)* in figure 2).

M:N: several associations may link any Spanish sense, and several associations may link any synset.

For every *PRB*(c) the logistic probability of c was obtained and table 2:left shows the average probability per group. Groups 1:1 and 1:N have a similar mean probability, higher than the other two groups. It seems, then, that the structure of *PRB*(c) may provide some useful information about the correctness of *StS* c.

Continuing with the study of the probability of c, the number of Spanish senses with an association in the *PRB* was proved to be related to the value. However, a third factor, the existence of gaps or not, demonstrated to affect the relation. In table 2:right the mean probability of *StS* c depending on the number of associations of *PRB*(c) and the existence of gaps or not is displayed. It can be seen how the mean probability increases with the number of associations, and also it is greater if no gaps exist.

The behavior of the mean probability taking into account the three factors together is displayed in table 2:center, which shows that the correctness of c in general increases with the number of Spanish senses with an association in the *PRB* without gaps and also if it presents an 1:1 or 1:N structure. However the behavior of *PRB*s with gaps is less clear in this context.

Fig. 2. PRB configurations for classes 1:N, N:1, with crossings
Solid lines mean direct relations, dotted lines mean indirect relations, arrows mean associations

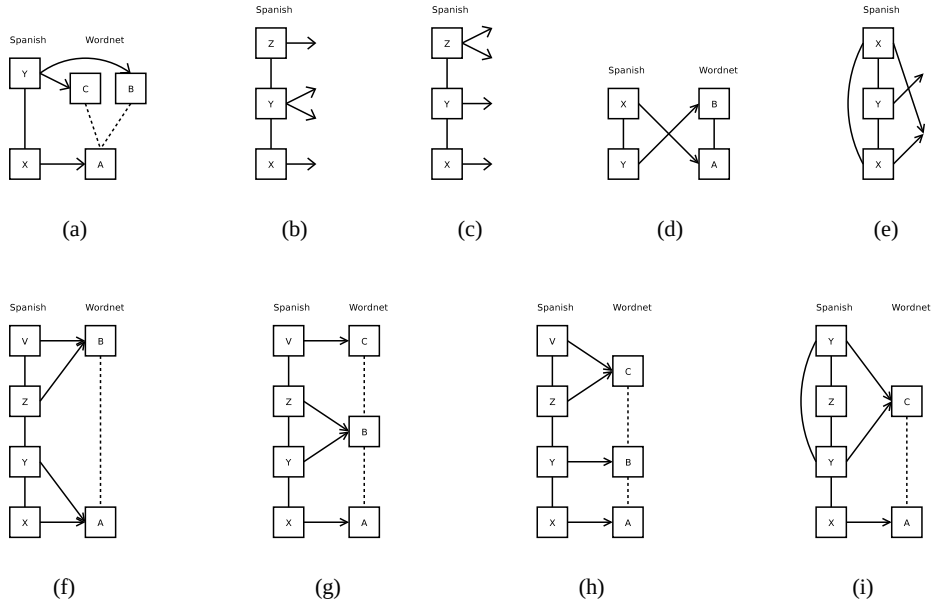
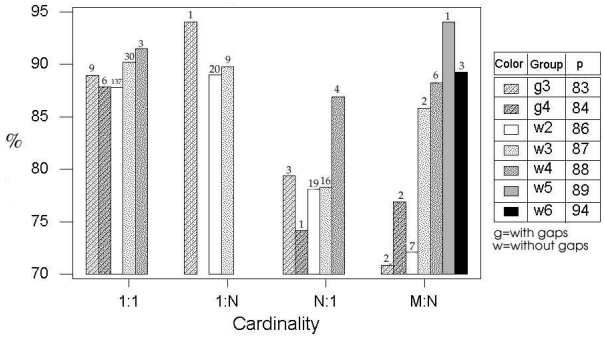


Table 2. Left: Cardinality sets. Right: Chains of cardinality sets.

Cardinality	Conn.	$\bar{p}$
1:1	184	0.88
1:N	30	0.89
N:1	49	0.79
M:N	23	0.81
crossing	11	



7 Results and Conclusions

A set of 1263 *StS* where induced from *WtS* following a simple heuristic, with a 96.7% estimated correctness percentage §4.

For the 1263 Spanish words, their branches were developed and their *StS* or *WtS* to Wordnet were identified. It was seen that developing chains with five ancestors is enough to get all the relevant information. So, given a *StS* c, the concept of *PRB*(c) was introduced in order to study the relationship between the Spanish taxonomy and Wordnet §5.

The internal structures of *PRB*s were studied. Depending on the level of the first ancestor with an association, different groups were obtained. *PRB*s with the first ancestor with an association at level 1 are faced in §6 where all possible patterns taking place in this family of *PRB*s are identified and displayed in figure2.

Finally the research extracted three factors that affect relationship between the structure of the *PRB* and the logistic probability of the base *StS* obtained with the model previously developed in [Farr02]: the cardinality of the relation between the branches of the *PRB*, the number of Spanish senses with an association in the *PRB* and the existence of gaps in the Spanish branch, summarizing the results in table 2.

The main result of this paper is that, indeed, the probability of the *StS* c used to generate the *PRB* tends to increase mainly with the number of Spanish senses with an association in the *PRB*. That is, in fact, a quite surprising and interesting result since the logistic model was based on the solution sets of methods which don't use the Spanish taxonomy at all.

8 Future Work

After analyzing the simplest case, some parameters that affect the probabilities of the *StS* used to generate *PRB*s have been identified. Research is now centered on monosemic Spanish words with several association in order to evaluate how these parameters help to choose the correct associations, and what other factors appear that were not detected during the present study. Plans are in progress for analyzing the cases of polysemic words with only one link.

When all the factors would have arisen after the preliminary studies pointed above, the work will be centered on how to take profit of the taxonomic relation, and how to infer data from upper levels of *PRB*s.

References

Atse98. J. Atserias, S. Climent, J. Farreres, G. Rigau, and H. Rodríguez. *Recent Advances in Natural Language Processing*, chapter Combining Multiple Methods for the Automatic Construction of Multilingual WordNets. Jon Benjamins Publishing Company, Amsterdam, The Netherlands, 1998.

Farr02. J. Farreres, K. Gibert, and H. Rodríguez. Semiautomatic creation of taxonomies. In G. Ngai *et al.*, editor, *Proceedings of SEMANET'02*, Taipei, 2002.

Farr98. J. Farreres, G. Rigau, and H. Rodríguez. Using wornet for building wordnets. In *Proceedings of COLING/ACL "Workshop on Usage of WordNet in Natural Language Processing Systems"*, Canada, 1998.

Rigau98. G. Rigau. *Automatic Acquisition of Lexical Knowledge from MRDs*. PhD thesis, Universitat Politècnica de Catalunya, 1998.

WordNet Exploitation through a Distributed Network of Servers

I. D. Koutsoubos[1,2], Vassilis Andrikopoulos[1], and Dimitris Christodoulakis[1,2]

[1] Computer Engineering and Informatics Department, Patras University,
26500 Patras, Greece
Email: andrikop@ceid.upatras.gr
[2] Research Academic Computer Technology Institute,
61 Riga Feraiou, 26221, Patras, Greece
Email: koutsoub@cti.gr, dxri@cti.gr

Abstract. The architecture of a lexical database in which multilingual semantic networks would be stored requires the incorporation of flexible mechanisms and services, which would enable the efficient navigation within and across lexical data. We report on WordNet Management System (WMS), a system that functions as the interconnection and communication link between a user and a number of interlinked WordNets. Semantic information is being accessed through a distributed network of servers, forming a large-scale multilingual semantic network.

1 Introduction

WordNet has been identified as an important resource in the human language technology and knowledge processing communities. Its applicability has been cited in many papers and systems have been implemented using WordNet. Almost every NLP application nowadays requires a certain level of semantic analysis. The most important part of this process is semantic tagging: the annotation of each content word with a semantic category. WordNet serves as a useful resources with respect to this task and has so far been used in various applications including Information Retrieval, Word Sense Disambiguation, Machine Translation, Conceptual Indexing, Text and Document Classification and many others.

There is an increasing amount of wordnet resources being made available for NLP researchers. These resources constitute the basic raw materials for building applications such as the abovementioned. Semantic networks standardization is of prime importance in the case of WordNets incorporation in real life applications. Towards a vision of next-generation tools and services that will enable the widespread development and use of wordnet resources we present a distributed WordNet server architecture in which WordNet servers, analogous to database servers, provide facilities for storing and accessing wordnet data via a common network API. Apart from distributing wordnets over multiple servers the system is capable of distributing wordnet-related services over multiple servers.

2 Advantages of Distributed Systems

We can summarize the motivations for adopting a distributed architecture for WordNet management-exploitation:

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 265–269.

Distributed Information Sources: WordNet resources may be scattered across multiple physical locations. Access to multiple resources may be mediated and rendered in a uniform way.

Sharing: Applications need to access several services or resources in an asynchronous manner in support of a variety of tasks. It would be wasteful to replicate problem-solving capabilities for each application. Instead it is desirable that the architecture supports shared access to agent capabilities and retrieved information.

Complexity Hiding: A distributed architecture allows specifying different independent problem-solving layers in which coordination details are hidden to more abstract layers.

Modularity and Reusability: A key issue in the development of robust analysis application is related to the enhancement and integration of existing stand-alone applications. Agent may encapsulate pre-existing linguistic applications, which may serve as components for the design of more complex systems. Inter-agent communication languages improve interoperability among heterogeneous services providers.

Flexibility: Software agents can interact in new configurations "on-demand", depending on the information flow or on the changing requirements of a particular decision making task.

Robustness: When information and control is distributed, the system is able to degrade gracefully even when some of the agents are not able to provide their services. This feature is of particular interest and has significant practical implications in natural language processing because of the inherent unpredictability of language phenomena.

Quality of Information: The existence of similar analysis modules able to provide multiple interpretation of the same input offers both 1) the possibility of ensuring the correctness of data through cross-validation and 2) a mean of negotiating the best interpretation(s).

3 Our Approach

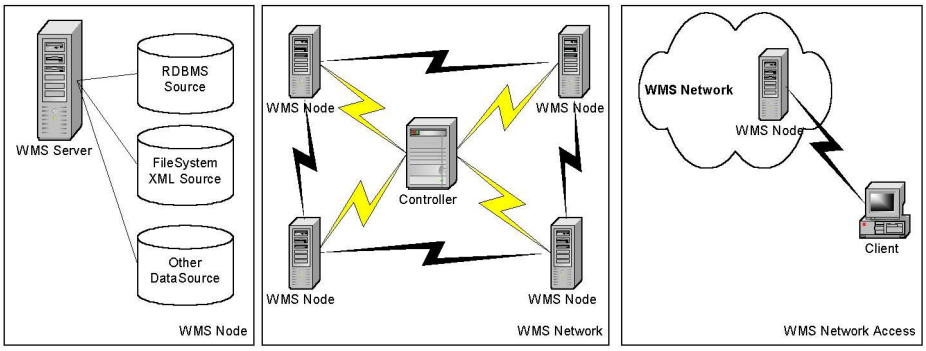
For the implementation of a flexible multilingual lexicographic database where navigation in the linguistics information would be facilitated there is an apparent need that flexible mechanisms and services are provided by a main technical infrastructure of the multilingual network. The WordNet Management System (WMS) is the system that acts as the interconnection and communication link between a user and any of the involved monolingual systems. As part of this communication someone should have the ability to submit requests for wordnet data contained in the WMS network. Moreover, keeping all the benefits of the Web, such as distributed work environment, concurrent access to the data and multiple views of the data will be achieved through the WMS.

From its definition, WMS falls into the Data Integration framework, being able to manage a distributed, dynamic network of homogeneous data. Previous systems built for this purpose [8,11,12] are often characterized by a centralized system that controls and manages interactions among distributed information sources in order to serve requests for data. As a consequence, in a distributed environment where no a priori knowledge of the location of specific data is possible, the traditional mediator and federated databases approaches are not appropriate. Furthermore, approaches such as [7,9,10] that provide a source- and query-independent mediator do not deal with decentralized systems with participants and

information sources location upredictability. As mentioned in [6] a P2P approach would be a more appropriate solution, since it lacks a centralized structure and promotes the equality among peers and their collaboration only for the time necessary to fulfi ll a request.

On the other hand, a feature that was considered very important during the design of the system, was the ability of the system to provide data to the wider possible set of data consumers, ranging from simple users to industrial solution-based applications. This requirement called for a variety of rendering mechanisms and interfaces with variable complexity for the communication of the system with its users. The ideal solution to this problem would be an API for wordnet access, as described in [4] or an extension of it, covering more recent achievements in the interface technologies like the Web Services technologies [`http://java.sun.com/webservices/`].

Taking both requirements into account, WMS was designed following a mixed approached, borrowing elements from both architectures to solve specifi c problems. Specifically, it was decided that the WordNet providers, i.e. the sources of WordNet data, should form a network of servers, using P2P techniques and thus creating a unifi ed semantic data resource which could be accessed from data consumers, linked as clients to the servers of the system, without taking into account resource-specifi c details which are hidden to them. The architecture of WMS is summarized in the following fi gures.



3.1 Network of Servers

Each WMS server hosts one (or more) wordnet data sources which are interconnected via the ILI structure [3]. WordNet data sources are identifi ed by language and version (creating a unique pair). A WMS server is considered a node in the P2P network and is treated equally by its peers. For each peer to acquire knowledge of the data available in the network (and additionally their location and how to access them), a super-node was added to the system. It serves as a yellow pages provider, or a directory service, registering WordNet hosts and distributing this information to the other nodes of the network. The super-node maintains all information about the servers of the network and the data hosted in each one. By communicating with the super-node, each node registers itself on the network and acquires

information concerning all the distributed WordNet data sources, which validates on the grounds of accessibility and availability.

Furthermore, the server operates on two modes. In the first mode, it provides the data exclusively for its hosted data sources and links to remote ones to the clients, with the client responsible for acquiring the remote data. In its second mode, the server is responsible for both local and remote data sources, providing remote data by executing remotely the requested operations and forwarding the results to the clients.

3.2 Clients

For the purpose of architecture, we consider any kind of semantic data consumer, either simple solutions as a site or more sophisticated ones as information brokering systems, possible clients to the system. In order to accommodate the multiple needs defined by such a variety of systems, each WordNet provider was decided to also act as a server for these clients. Using the standard client-server schema, the data consumer has to submit its requests to a WMS server in order to retrieve the necessary results. Additionally, a uniform API is provided to the interested parties in the form of a number of services provided by a Web Services mechanism, which add a level of abstraction between the clients and the data resources, facilitating the usage of the system for the implementation of different in their nature applications which use semantic data in very different ways.

3.3 Data Management in the WMS

For the internal communication of the nodes of the WMS network, a custom XML messaging schema is used. Provision was taken during the design of the schema to keep it as flexible and extendable as possible to accommodate possible future enhancements of wordnet data. For the same purpose, the API that describes the functions provided by WMS is also designed with openness in mind, allowing the extension of the available operations and the flexible incorporation of new ones.

As far as the communication with the clients is concerned, a variety of access methods are provided, ranging from simple HTTP requests and SOAP to RMI. The actual messaging uses XML to describe the requests and the data, but lacking a standardized WordNet protocol, describing data and functions, the system provides templating mechanisms for defining the requests and the responses.

WMS provides the developers with the capability to use and maintain different types of storage mechanisms for their respective WordNets, from simple solutions as text files to more sophisticated ones like binary structures. The requirements for such an abstraction are set by the system in the form of an API, which a developer that wants to use a specific medium has to implement. Currently, WMS provides by default mechanisms for access to the majority of commercial Relational Database Management Systems and to XML files that use the VisDic formalism [5].

4 Discussion and Future Enhancements

We have presented the architecture of WordNet Management System, a distributed network of servers that provides facilities for WordNet exploitation. In the future, it is envisaged to

incorporate other types of lexical resources besides wordnets and to provide the mechanisms for interaction with other NLP modules, such as a module for Semantic Indexing of documents.

Acknowledgements

This research is supported under project BalkaNet: "Design and Development of a Multilingual Balkan WordNet", funded by the European Commission under the framework of the IST Programme (IST-2000-29388). We would like to especially thank professor Thanos Skodras for his support and encouragement.

References

1. Fellbaum Ch. (ed.) (1998) WordNet: An Electronic Lexical Database, Cambridge, M.A: MIT Press.
2. Stamou S., Ofhzer K., Pala K., Christodoulakis D., Cristea D., Tufis D., Koeva S., Totkov G., Dutoit D., Grigoriadou M. (2002) "BALKANET: A Multilingual Semantic Network for Balkan Languages". In Proceedings of the GWA 1st International WordNet Conference, Mysore, India, Jan. 21–25, 2002.
3. Vossen P. (ed.) (1998) EuroWordNet: A Multilingual Database with Lexical Semantic Networks, Kluwer Academic Publishers, Dordrecht.
4. Miatidis M., Assimakopoulos D., Koutsoubos I.-D., Kourousias G., Tzagarakis M., Christodoulakis D. (2001) "Access to WordNets Through API".
5. Pavelek, T. and Pala K. (2002) "WordNet standardization from a practical point of view". In Proceedings of the LREC Workshop on WordNet Structures and Standardisation, and How These Affect WordNet Applications and Evaluation, Las Palmas, Gran Canaria. 30–34, May 28th, 2002.
6. Panti M., Penserini L., Spalazzi L. (2002) "A pure P2P approach to information integration". Tec. Report 2002-02-19, Istituto di Informatica, University of Ancona, 2002.
7. R. J. Bayardo, W. Bohrer, et al. (1997) "InfoSleuth: Agent-Based Semantic Integration of Information in Open and Dynamic Environments". In Proceedings of the ACM SIGMOD International Conference on Management of Data, vol. 26, No. 2, June 1997.
8. H. Garcia-Molina, Y. Papakonstantinou, et al. (1997) "The TSIMMIS approach to mediation: data models and languages". In Journals of Intelligent Information Systems, 8:117–132, 1997.
9. A. Levy, A. Rajaraman, J. Ordille (1996) "Querying Heterogeneous Information Sources Using Source Descriptions". In Proceedings of the 22nd VLDB Confernece, September, 1996.
10. M. Nodine, W. Bohrer, A. H. Hiong Ngu (1999) "Semantic Brokering over Dynamic Heterogeneous Data Sources in InfoSleuth". In Proceedings of the 15þ International Conference on Data Engineering, March 1999.
11. A. Sheth and J. Larson (1990) "Federated Database Systems for Managing Distributed, Heterogeneous and Anonymous Databases". In ACM Transaction on Database Systems, 22(3), 1990.
12. G. Wiedehold (1992) "Mediators in the Architecture of Future Information Systems". In IEEE Computer Magazine, 25:38–49, March 1992.

WordNet Applications

Jorge Morato[1], Miguel Ángel Marzal[2], Juan Lloréns[1], and José Moreiro[2]

[1] Dept. Computer Science, Universidad Carlos III, Madrid, Spain
Email: jorge@ie.inf.uc3m.es, llorens@ie.inf.uc3m.es
[2] Dept. Library Science, Universidad Carlos III, Madrid, Spain
Email: mmarzal@bib.uc3m.es, jamore@bib.uc3m.es

Abstract. This paper describes WordNet design and development, discussing its origins, the objectives it initially intended to reach and the subsequent use to which it has been put, the factor that has determined its structure and success. The emphasis in this description of the product is on its main applications, given the instrumental nature of WordNet, and on the improvements and upgrades of the tool itself, along with its use in natural language processing systems. The purpose of the paper is to identify the most significant recent trends with respect to this product, to provide a full and useful overview of WordNet for researchers working in the field of information retrieval. The existing literature is reviewed and present applications are classified to concur with the areas discussed at the First International WordNet Congress.

1 Introduction

WordNet, one of a series of manually compiled electronic dictionaries, is restricted to no specific domain and covers most English nouns, adjectives, verbs and adverbs. Although there are similar products, such as Roget's International Thesaurus, or CYC, Cycorp: Makers of the Cyc Knowledge Server for artificial intelligence-based Common Sense. CYC contains 100,000 concepts and thousands of relations. Each concept is assigned to certain terms related by prepositional logic. The present paper analyses the reasons for WordNet's success and, in particular, the main applications of the tool over the last ten years.

2 Wordnet Development

The origin of this tool is to build a lexical-conceptual model and database, consisting of both lexical units and the relations between such units, structured into a relational semantic network.

Originally intending to create a product that could combine the advantages of electronic dictionaries and on-line thesauri, an expert team of linguists and psycholinguists headed by G. A. Miller began research at Princeton University's Cognitive Science Laboratory in 1985 that would culminate in the appearance of WordNet in 1993.

WordNet offers researchers, many of which were not initially envisaged by the authors, along with its cost-free use and well-documented open code. The result has been the appearance of applications in different fields of research, making it an ideal tool for disambiguation of meaning, semantic tagging and information retrieval. Therefore, although four members manage, maintain and develop WordNet many other teams collaborate in driving implementation of the product, as attested by two facts:

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 270–278.

1. The speedy pace of release of new versions of WordNet.
2. Organised world-wide promotion of WordNet, through the creation of the *Global WordNet Association*, which, in conjunction with CIIL Mysore, IIT Bombay and IIIT Hyderabad, held the 1st International WordNet Conference in 2002. Primarily technical, the conference was structured under six areas of interest: Linguistics, WordNet architecture, WordNet as a lexical resource and component of NLP, Tools and Methods for WordNet Developme, nt., Standardisation., Applications (information extraction and retrieval, document structuring and categorisation, language teaching).

These six topics are still present in the 2nd International Conference of the Global WordNet Association (GWC 2004) held at Masaryk University, Brno.

3 Applications

The success of WordNet, as mentioned, is largely due to its accessibility, quality and potential in terms of NLP. Figure 1 below shows the results of a search run on the bibliographic database LISA, INSPEC, IEEE and ResearchIndex and on the Universidad Carlos III's OPAC. The documents were publised from 1994 till 2003. This search, while not necessarily exhaustive in respect of WordNet research, does nonetheless show how such research effort is distributed. It will be observed that the major use of this tool has been in the area of conceptual disambiguation.

Fig. 1. WordNet Applications

3.1 Improvements in WordNet

The data record of publications dealing with WordNet shows that there has been a tendency to improve the product in a number of respects. The objective is to make WordNet much more effective and relevant than any existing on-line dictionary by incorporating greater semantic wealth and taking a more contextual approach. Several possibilities have been explored to achieve this:

Studies geared to improving software There is a clear prevalence, in terms of volume, of papers geared to expanding and enriching the WordNet structure. One of such endeavours is VerbNet [1], designed to make up for shortcomings in the associative relations between verbs; another is the Lingua::WordNet interface [2], which furnishes an editable presentation of WordNet, with meronym categories never before implemented. Finally, substantial efforts have been made to standardise, integrate and distribute the tool.

Multilingual WordNet One of the most relevant endeavours has been the development of EuroWordNet, a project based on WordNet structure whose ultimate purpose is to develop multilingual databases with wordnets for several European languages. Each wordnet adopts an autonomous lexicalisation structure and all are interconnected through an interlinguistic index, for which relations have been added and modified and new levels identified in WordNet. For a multilingual description of EuroWordNet see [3,4].

This paper poses the possibility of automatically transferring a list of English verbs, classified by their syntactic characteristics, to WordNet synsets.

3.2 Improvements in Natural Language Processing Systems

Such improvements are found in a substantially larger number of papers on WordNet, regarded here to be a tool well suited to a series of applications such as discussed below:

Information retrieval and extraction These operations are closely related to organisation and representation of knowledge on the Internet. One of the lines of research pursued is the application of artificial intelligence to information retrieval, stressing the local components and inferential process of human reasoning in the design of automatic information retrieval systems. The method for incorporating logic and inference focused on WordNet shortly after it appeared [5]. WordNet has been used as a comprehensive semantic lexicon in a module for full text message retrieval in a communication aid, in which queries are expanded through keyword design [6]. WordNet has, then, started to be used as a linguistic knowledge tool to represent and interpret the meaning of, and provide the user with efficient and integrated access to, information; integration, indeed, has become an increasingly necessary feature with the development of multiple database access systems and one in which WordNet's identification and interpretation of semantic equivalents is extraordinarily useful [7].

Mandala [8] proposed the use of WordNet as a tool for the automatic construction of thesauri, based either on co-occurrence determined by automatic statistical identification of semantic relations, or on the predicate-argument association, in which the most significant words of an environment (predicate) and those with which they relate are identified to construct the argument. In another vein, Moldovan [9] opted to use WordNet in the development of a natural language interface to optimise the precision of Internet search engines by expanding queries.

Concept identification in natural language This operation is designed to detect the terms requested, not only for extraction, but to suit them to the full semantic richness and complexity of a given information need. WordNet applications have followed a dual course in such applications:

1. **Disambiguation** i.e., precision and relevance in response to a query via resolution of semantic inconsistencies. Moldovan [9] described schematically the semantic disambiguation as follows:
   (1) All the noun–verb pairs in the sentence are selected.
   (2) The most likely meaning of the term is chosen (subprocess that Moldovan calls terminological disambiguation). Internet is used with this goal.
   (3) Drawing from the most frequently appearing concepts (step 2), all the nouns are selected in the "glossaries" of each verb and its hierarchical subordinates.
   (4) Drawing from the most frequently appearing concepts, all the nouns are selected in the "glossaries" of each noun and its hierarchical subordinates.
   (5) A formula is applied to calculate the concepts common to the nouns in points 3 and 4.

   Disambiguation is unquestionably the most abundant and varied WordNet application. Indeed, there is a wide range of possibilities.
   WordNet has served as a support for the development of tools to enhance the efficiency of Internet resource searches. One example is the IWA/H project for building an ontological framework able to disambiguate search criteria via mixed knowledge representation technique systems (ARPA KRSL); others include tools such as Oingo and SimpliFind, two Internet products that avoid ambiguity in natural language searches by using the WordNet lexicon, duly expanded by creating millions of word associations to refine the search process.
   The use of WordNet for improving search engines is interesting the IWA/H project was based on the MORE technique developed by the RBSE project for more efficient retrieval of Internet resources, as discussed by Eichmann [10].
   WordNet has, naturally, been used for disambiguation in traditional models to enhance information retrieval efficiency: for the development of a classifier, implemented with WordNet, able to combine a neurone-like network to process subject contexts and a network to process local context; for the exploitation of a Bayesian network able to establish lexical relations with WordNet as a source of knowledge, integrating symbolic and statistical information [11]; for the development of a statistical classifier, implemented with WordNet lexical relations, able to identify the meaning of words, combining the context with local signs [12]; and as support for the development of a computational similarity model to add on-line semantic representation to the statistical corpus. WordNet has, therefore, proved its worth as an ideal methodological element to disambiguate the meaning of words in information extraction systems [13]. As a result, projects have been launched to disambiguate nouns in English language texts using specification marks deriving from WordNet taxonomies as a knowledge base, as well as to reduce polysemy in verbs, classified by their meanings via predicate associations, with a view to optimising information retrieval. Methods for nouns [14] and verbs [1,4] has already been analysed.
   At the same time, new disambiguation models have been tested in conjunction with WordNet by: generating ontological databases with a systematic classification of multiple meanings derived from WordNet [15]; or generating broad corpora to signify words on the grounds of WordNet synonymies or definitions in the wording of queries [16]. One result has been the appearance of GINGER II, an extensive dictionary semantically tagged using 45 WordNet categories and an algorithm for interpreting

semantic text by determining verb senses, identifying thematic roles and joining prepositional phrases [17]. More recently R. Mihalcea and D. Moldovan presented AutoASC, which automatically generates sense tagged corpora that prove to be very effective for disambiguation in information retrieval; this product incorporates the latest WordNet gloss definitions [18].

2. **Semantic distance** Three concepts recur in WordNet literature that entail a certain amount of ambiguity: terminological distance, semantic distance and conceptual distance. The terms semantic distance and conceptual distance are found to be used in several studies to pursue the same objective and deploy the same methodology for resolving the issue at hand. Terminological distance, by contrast, often appears to refer to the suitability of the word selected to express a given concept.
   Semantic distance is understood to mean the contextual factor of precision in meaning. In his particularly relevant papers, Resnik [19] computes class similarity, defining class to be the nouns of a synset plus the nouns in all the subordinate synsets. Although the concept of semantic similarity between classes was proposed by Resnick [19]. WordNet was quickly enlisted to build and operate with FEDDICT, a prototype on-line dictionary to develop an information retrieval technique based on the measurement of the conceptual distance between words, in which WordNet semantic relations proved to be highly useful [20]. A very interesting sequel to this endeavour was provided by McTavish [21] who used WordNet semantic domains to establish categories that could be used to analyse conceptual semantic distances in terms of social environments to better organise terms for retrieval.
   Computational linguistics is, however, the area that has placed the greatest emphasis on *relations* and *semantic distances* between lexemes, the measures of which were classified by A. Budanitsky [22]. This paper highlights the measures that use WordNet as a resource and for implementation of functions, in particular: Hist-St. And Leacock–Chodorow, in which similarity, albeit in the IS-A link only, rests on the shortest path between two synsets; and Resnik, Jiang, Conrath and Lin, for all of whom *information content* is a determining factor of similarity in their measures of distance.

Query expansion In 1994 Smeaton [23] proposed an expansion system based on calculating the tf-idf for the query terms and adding to it half the tf-idf for the WordNet synonyms for these terms. Gonzalo [24] later reported the benefits of applying WordNet to queries, using it as a WSD (Word Sense Disambiguator) able to enhance the search process by including semantically related terms and thus retrieve texts in which the query terms do not specifically appear.

Document structuring and categorisation Intellective efforts and operations in this area are geared to a new organisation and representation of knowledge. In this case the focus is on the aspects of the tool suited to document categorisation: extraction of semantic traits by grammatical categorisation of WordNet nouns, verbs and adjectives [25]; and categorisation of the relevance of the data in INFOS by predicting user interest on the basis of a hybrid model using keywords and WordNet conceptual representation of knowledge [26].

Further research along these lines came in the form of a computational method presented by S. M. Harabagiu [27] for recognising cohesive and coherent structures in texts, drawing on

WordNet lexical-semantic information, whose objective is to build designs for the association between sentences and coherence relations as well as to find lexical characteristics in coherence categories. WordNet became an ancillary tool for semantic ontology design geared to high quality information extraction from the web, and has prompted new endeavours such as the WebOntEx (Web Ontology Extraction) prototype developed by Keng Woei Tan [28] which is designed to create ontologies for the semantic description of data in the web.

Judith Klavans [29] devised an algorithm for automatically determining the genre of a paper on the grounds of the WordNet verb categories used. With their WN-Verber, they determined that some verbal synsets and their highest subordinates are less frequent in certain document typologies.

Audio and video retrieval This is a challenge in need of increasingly urgent attention in view of the burgeoning development of hypermedia and non-text information. The MultiMediaMiner [30], is a prototype to extract multimedia information and knowledge from the web that uses WordNet to generate conceptual hierarchies for interactive information retrieval and build multi-dimensional cubes for multi-media data. Finally, WordNet has been used in query expansion to index radio news programme transcriptions effected by a prototype designed to retrieve information from radio broadcasts [31].

Other WordNet applications

Parameterisable information systems While anecdotal, the J. Chai [32] proposal to create an information system (called Meaning Extraction System) that can be configured in terms of a specific user profile is appealing. The user chooses from a series of texts (training collection) the ones that appear to be of greatest interest. WordNet identifies the hierarchical (IS-A) synsets related to the terminology of the documents selected. This process generates rules that enable the system to identify, *a priori*, the documents that the user will find to be of interest.

Language teaching and translation applications As discussed in point 3.1.2, applications have been devised and tested to improve the composition of texts drafted by non-native English writers. However, yet another line of research has been addressed in international conferences on WordNet, namely, foreign language teaching. One example is the article by X. Hu and A. Graesser, which proposes using the WordNet vocabulary to evaluate pupils' command of a given language [33].

As a translation aid based on the application of semantic distance algorithms, WordNet has also been used to develop a potential error corrector for the positioning of words [34]. One very intuitive formula consists of using *conceptual interlingua* representation of texts and queries such as used in the CINDOR system, which accommodates WordNet-supported inter-linguistic combinations, obviating the need for an expert translation for retrieval. The CINDOR system was presented and tested at TREC-7 and seems to be useful for cross- or combined linguistic retrieval [35].

4 Trends

Trends are difficult to ascertain and evaluate in view of the clearly instrumental and application-based dimension that underlies WordNet's success. Nonetheless, a comparative analysis of the most recent publications provides some insight into a number of trends in WordNet use:

1. Development of interlinguistic indices for multilingual conceptual equivalence, without translation. Subsidiarily, this endeavour has also been geared to perfecting integrated access to information, driven by the rapid development of multiple database access systems.
2. Use as an ideal tool to optimise the retrieval capacity of existing systems: natural language interfaces for search engines; automatic generation of tools for semantic disambiguation of concepts (corpora, dictionaries, directories, thesauri) and the creation of knowledge summaries from expanded queries.
3. Support for the design of grammatical categorisations designed to classify information by aspects and traits, but in particular to design and classify semantic ontologies that organise web data – semantically, to be sure.
4. Basis for the development of audio-visual and multi-media information retrieval systems.
5. In the last 3 years ontologies construction have been one of the most dynamic areas and its applications to the semantic web [36].

5 Conclusions

Although WordNet applications are growing steadily and research may be expected to increase in the coming years as new versions are released, the tool has certain shortcomings that should be addressed in future studies.

Limitations and Problems are:

1. Although WordNet is an electronic resource, it was, after all, designed for manual consultation and not for automatic processing of natural language texts; as a result, no particular emphasis was placed on enabling the system to automatically differentiate between the various concepts involved.
2. Another problem is its multidisciplinarity, which prompts flawed operation in many NLP systems, due to which processing is usually conducted with sublanguages or specific records.
3. Classification was performed manually, which means that the reasons and depth of classification may not be consistent.
4. While the synset simplification affords obvious advantages, in the longer term it leads to shortcomings. These are particularly acute in semantic proximity calculations and may create insuperable situations whenever the context of the discourse in which the relation appears is not contained in the synset information.
5. The overabundance of nuance in the concepts calls, in nearly any NLP application, for prior calculation of the frequency of the concept in a given domain. Such calculation is one of the sources of system error, especially where WordNet glosses – extracted, as noted above, from the Brown Corpus – are used, due to the uneven coverage afforded to the different domains.

References

1. Palmer, M.: Consistent criteria for sense distinctions. Computers and the Humanities, 34 (1–2) (2000) 217–222.
2. Brian, Dan: Lingua::WordNet. The Perl Journal (2000) `http://www.brians.org/wordnet/article/`.
3. Vossen, P.: Introduction to EuroWordNet. Computers and the Humanities, 32(2–3) (1998) 73–89.
4. Green, Rebecca, Pearl, L., Dorr, B. J., and Resnik, P.: Mapping lexical entries in verbs database to WordNet senses. Proceedings of the 39þ Annual Meeting of the Association for Computational Linguistics (ACL 2001), Toulouse, France, July 9–11 (2001).
5. Nie, J. Y., and Brisebois, M.: An inferential approach to information retrieval and its implementation using a manual thesaurus. Artificial Intelligence Review, 10 (5–6) (1996) 409–439.
6. Van de Riet, R., Burg, H., and Dehne, F.: Linguistic instruments in information system design. FOIS. Proceedings of the 1st International Conference. Amsterdam: IOS Press (1998).
7. Jeong-Oog-Lee & Doo-Kwon-Baik: Semantic integration of databases using linguistic knowledge. Proceedings Advanced Topics in Artificial Intelligence. Berlin: Springer-Verlag, (1999).
8. Mandala, Rila, Tokunaga, T., Tanaka, Hozumi, O., Akitoshi, Satoh, K.: Ad Hoc Retrieval Experiments Using WordNet and Automatically Constructed Thesauri. TREC-7 (1998) 414–419.
9. Moldovan, D. I. and Mihalcea, R.: Using WordNet and lexical operators to improve Internet searchers. IEEE Internet Computing, 4 (1) (2000) 34–43.
10. Eichmann, David: Balancing the Need for Knowledge and Nimbleness in Transportable Agents. Position Paper for the Dartmouth Workshop on Transportable Agents. (1996). URL: `http://mingo.info-science.uiowa.edu/eichmann/DWTA96.html`, last hit on 2/3/02.
11. Wiebe, Janyce, O'Hara, Tom, and Bruce, Rebecca: Constructing Bayesian Networks from WordNet for Word-Sense Disambiguation: Representational and Processing Issues. Use of {W}ord{N}et in Natural Language Processing Systems: Proceedings of the Conference Association for Computational Linguistics, Somerset, New Jersey (1998). 23–30.
12. Towell, G. and Voorhees, E. M.: Disambiguating highly ambiguous words. Computational Linguistics, 24 (1) (1998) 125–145.
13. Chai, Joyce Y. and Biermann, Alan W.: A WordNet based rule generalization engine for meaning extraction, to appear at Tenth International Symposium On Methodologies For Intelligent Systems (1997).
14. Montoyo, A. and Palomar, M.: Word sense disambiguation with specification marks in unrestricted texts. Proceedings 11th International Workshop on Database and Expert Systems Applications. Los Alamitos (Ca): IEEE Press (2000) 103–107.
15. Buitelaar, P.: CORELEX: an ontology of systematic polysemous class. Proceedings FOIS'98. Amsterdam: IOS Press. (1998) 221–235.
16. Mihalcea, R. and Moldovan, D. I.: Automatic acquisition of sense tagged corpora. Proceedings of the 12th International Florida AI Research Society Conference. Menlo Park(Ca): AAAI Press (1999) 293–297 and 16th: 461–466.
17. Dini, L., Tomasso, V., and Segond, F.: GINGER II: an example-driven word sense disambiguator. Computers and the Humanities, 34 (1–2) (2000). 121–126.
18. Mihalcea, R. and Moldovan, D. I.: AutoASC, a system for automatic acquisition of sense tagged corpora. International Journal of Pattern Recognition and Artificial Intelligence, 14 (1) (2000) 3–17.
19. Resnick, P: Selection and Information: A class-based approach to lexical relationships. PhD dissertation. University of Pennsylvania (1993).
20. Richardson, R., Smeaton, A. F., and Murphy, J: Using WordNet for conceptual distance measurement. Proceedings of the 16th Research Colloquium of the British Computer Society Information Retrieval Specialist Group. London: Taylor Graham. (1996) 100–123.

21. McTavish, D. G., Litkowski, K. C. and Schrader, S: A computer content analysis approach to measuring social distance in residential organizations for older people. Social Science Computer Review, 15 (2) (1997) 170–180.
22. Budanitsky, A. and Hirst, G.: Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. (2001). URL: `http://www.cs.toronto.edu/pub/gh/`.
23. Smeaton, Alan F., Kelledy, Fergus, and O'Donell, Ruari: TREC-4 Experiments at Dublin City University: Thresholding Posting Lists, Query Expansion with WordNet and POS Tagging of Spanish. Proceedings of TREC-4. Gaithersburg (USA): D. Harman (Ed.) (1994).
24. Gonzalo, J., Verdejo, F., Chugur, I., and Cigarran, J.: Indexing with WordNet synsets can improve text retrieval. Proceedings of the COLING/ACL '98 Workshop on Usage of WordNet for NLP. Montreal (Canada) (1998) 38–44.
25. Scheler, G.: Extracting semantic features from unrestricted text. WCNN'96. Mahwah (NJ): L. Erlbaum. (1996).
26. Mock, K. J. and Vemuri, V. R: Information filtering via hill climbing, WordNet and index patterns. Information Processing and Management, 33 (5). (1997) 633–644.
27. Harabagiu, S. M.: WordNet-based inference of contextual cohesion and coherence. Proceedings of the 11th International Florida Artificial Intelligence Research Symposium Conference. Menlo Park (Ca): AAAI Press (1998) 265–269.
28. Keng Woei Tan, Hyoil-Han and Elmasri, R.: Web data cleansing and preparation of ontology extraction using WordNet. Proc. 1st International Conference on Web Information Systems Engineering. Los Alamitos (Ca): IEEE Computational Society, 2. (2000) 11–18.
29. Klavans, Judith and Kan, Min-Yen: Role of verbs in document analysis. Proceedings of the Conference, COLING-ACL. Canada: Université de Montreal. (1998).
30. Zaiane, O. R., Hagen, E., and Han, J.: Word taxonomy for online visual asset management and mining. Application of Natural Language to Information Systems. Proc. 4th Internat. Conference NLDB'99. Vienna: Osterreichische Comput. Gesseleschaft (1999) 271–275.
31. Federico, M.: A system for the retrieval of Italian broadcast news. Speech Communication, 32 (1–2) (2000). 37–47.
32. Chai, Joyce Y. and Biermann, Alan W.: The use of word sense disambiguation in an information extraction system. Proceedings 16th National Conference on Artificial Intelligence. Menlo Park (Ca): AAAI Press (1999) 850–855.
33. Hu, X & Graesser, A.: Using WordNet and latent semantic analysis to evaluate the conversational contributions of learners in tutorial dialogue. Proceedings of ICCE'98, 2. Beijing: China Higher Education Press (1998) 337–341.
34. Shei, C. C. and Pain, H.: An ESL writer's collocational aid. Computer Assisted Language Learning, 13 (2) (2000) 167–182.
35. Diekema, A., Oroumchian, F., Sheridan, P., and Liddy, E. D.: TREC-7 evaluation of Conceptual Interlingua Document Retrieval (CINDOR) in English and French. Gaitherburg (USA): TREC-7 National Institute of Standards & Technology (1999) 169–180.
36. Khan, L, Luo, F: Ontology construction for information selection. Proceedings of the 14 IEEE ICTAI 02 (2002).

Extending and Enriching WordNet with OntoLearn

Roberto Navigli[1], Paola Velardi[1], Alessandro Cucchiarelli[2], and Francesca Neri[2]

[1] Università di Roma "La Sapienza", Dipartimento di Informatica, Via Salaria 113
I-00198 Roma, Italy
Email: velardi@dsi.uniroma1.it, navigli@dsi.uniroma1.it
[2] Università Politecnica delle Marche, D.I.I.G.A., Via Brecce Bianche 12, I-60131 Ancona, Italy
Email: cucchiarelli@diiga.univpm.it, neri@diiga.univpm.it

Abstract. OntoLearn is a system for word sense disambiguation, used to automatically enrich WordNet with domain concepts and to disambiguate WordNet glosses. We summarize the WSD algorithm used by Ontolearn, called *structural semantic interconnection*, and its main applications.

1 The Structural Semantic Interconnection Algorithm

OntoLearn is a system for the automatic extraction of concepts from texts that has been developed over the past few years at the University of Roma "La Sapienza", with the contribution of several other researchers in Italy. The system has been used and is being enhanced in the context of European and national projects[1].

The key task performed by OntoLearn is semantic disambiguation, a task we applied to various problems, namely:

- associate complex domain terms (e.g. *local area networks*) with the appropriate WordNet synsets (e.g. respectively: {*local#2*} (adj.), {*area#1, country#4*}, {*network#2, communications network#1* }) in order to enrich WordNet with new domain concepts and learn domain-specific ontologies [2,3];
- disambiguate WordNet glosses [1];
- disambiguate words in a query for sense-based web query expansion [4].

Semantic disambiguation is performed using a method we have named *structural semantic interconnection (SSI)*, a structural approach to pattern recognition, that uses graphs to describe the objects to analyze (word senses) and a context free grammar to detect common semantic patterns between graphs. Sense classification is based on the number and type of detected interconnections.

In this paper we provide a high-level intuitive description of the SSI algorithm, which is rather complex. A thorough description is in [3], but a complete reformalization is in progress.

SSI is a kind of *structural pattern recognition*. Structural pattern recognition [5] has proven to be effective when the objects to be classified contain an inherent, identifiable organization, such as image data and time-series data. For these objects, a representation based on a "flat" vector of features causes a loss of information which negatively impacts on

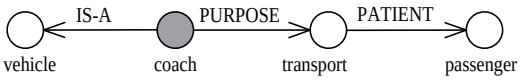
[1] Harmonise IST-13015 in the Tourism domain; WonderWeb IST-2001-33052 on ontology infrastructure for the semantic web, and the national MIUR-SP6 project on Web Learning.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 279–284.

classifi cation performances. The classifi cation task in a structural pattern recognition system is implemented through the use of grammars which embody precise criteria to discriminate among different classes. The drawback of this approach is that grammars are by their very nature application and domain-specifi c. However, machine learning techniques may be adopted to learn from available examples.

Word senses clearly fall under the category of objects which are better described through a set of structured features. Compare for example the following two feature-vector (a) and graph-based representations (b) of the WordNet 1.7 defi nition of *coach#5* (*a vehicle carrying many passengers, used for public transport*):

(vehicle, passenger, transport) (a)



(b)

The graph representation shows the semantic interrelationships among the words in the defi nition, in contrast with the "flat" feature vector representation.

Provided that a graph representation for alternative word senses in a context is available, *disambiguation can be seen as the task of detecting certain "meaningful" interconnecting patterns among such graphs*. We use a context free grammar to specify the type of patterns that are the best indicators of a semantic interrelationship and to select the appropriate sense confi gurations accordingly.

To automatically generate a graph representation of word senses, we use the information available in WordNet 1.7 augmented with other on-line lexical resources, such as semantically annotated corpora, list of domain labels, etc. Figure 1 is an example of the semantic graph generated for sense #2 of *bus*. In the fi gure, nodes are word senses, arcs are semantic relations. The following semantic relations are used: *hyperonymy* (car *is a kind of* vehicle, denoted with $\stackrel{kind-of}{\longrightarrow}$), *hyponymy* (its inverse, $\stackrel{has-kind}{\longrightarrow}$), *meronymy* (room *has-part* wall, $\stackrel{has-part}{\longrightarrow}$), *holonymy* (its inverse, $\stackrel{part-of}{\longrightarrow}$), *pertainymy* (dental *pertains-to tooth* $\stackrel{pert}{\longrightarrow}$), *attribute* (dry *value-of* wetness, $\stackrel{att}{\longrightarrow}$), *similarity* (beautiful *similar-to* pretty, $\stackrel{sim}{\longrightarrow}$), *gloss* ($\stackrel{gloss}{\longrightarrow}$), *topic* ($\stackrel{topic}{\longrightarrow}$), *domain* ($\stackrel{dl}{\longrightarrow}$). *Topic*, *gloss* and *domain* are extracted respectively from annotated corpora, sense defi nitions and domain labels. Every other relation is explicitly encoded in WordNet.

The basic *semantic disambiguation step* of the SSI algorithm is described hereafter. Let $C = \{w_0, w_1, \ldots, w_{n-1}\}$ be a list of co-occurring words. In a generic step i of the algorithm, let $D = \{S_j^a, S_i^b, \ldots, S_m^c\}$ be a list of semantic graphs, one for each of the words $W_D = \{w_a, w_b, \ldots, w_c\}$, $W_D \subseteq C$ already disambiguated in steps $1, 2, \ldots, i-1$. Let further $P = \{w_p, w_q, \ldots, w_z\}$ be the list of words in C that are still ambiguous, where $W_D \cup P = C$ and $W_D \cap P = \emptyset$. D is called the *semantic context* of P.

Until all words $w_r \in P$ have been analyzed, do:

- Let $S_{w_r} = \{S_1^r, S_2^r, \ldots, S_k^r\}$ be the set of senses of w_r, each represented by a semantic graph.

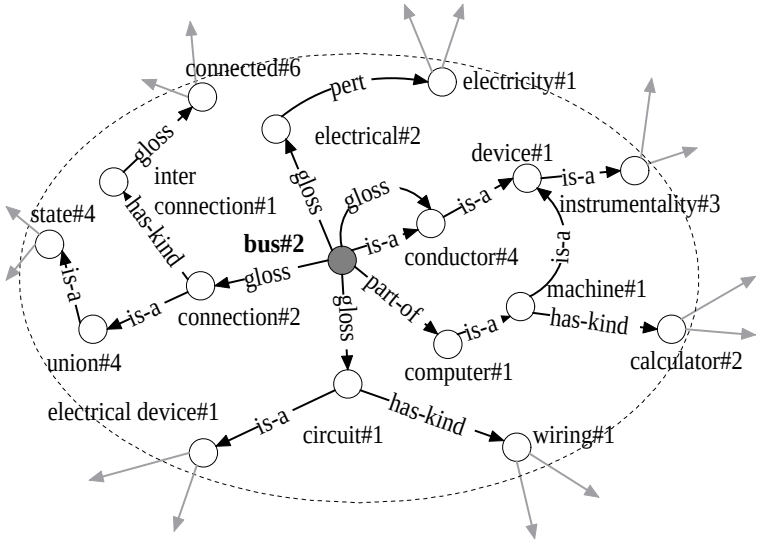


Fig. 1. Example of derived semantic graph for sense #2 of *bus* in WordNet

- Find the best sense $S_l^r \in S_{w_r}$, according to a classification criterion $\Im$. If $\Im$ is not met, skip to a subsequent word in P.
- Add S_l^r to D, delete w_r from P.

Repeat until either P is empty, or no new words are found that meet the classification criterion $\Im$. We now describe the classification criterion $\Im$.

Classification is based on searching specific interconnection patterns between some of the semantic graphs in D and the semantic graphs associated to senses of a word w_r. Each matching pattern increases the weight $w(S_k^r)$ of the correspondent word sense. The classification criterion assigns sense S_l^r to word w_r if $w(S_l^r) = argmax_k(w(S_k^r))$ and $w(S_l^r) \geq \beta$, where β is a fixed threshold.

Interconnection patterns are described by a context free grammar. For the sake of space we are unable to give here an account of the grammar. An intuitive example of an elementary pattern between two semantic graphs S_j^i S_k^h is informally described by the following sentence: "The graph S_j^i is connected to the graph of S_k^h through a holonymy path". For example: *window#7* $\stackrel{part-of}{\longrightarrow}$ *computer screen#1*. The grammar includes several complex patterns made of elementary ones, e.g. *holonymy-hyperonymy* sequences. We are now left with the problem of how to initialize the list D. Initialization depends upon the specific disambiguation task being considered. In OntoLearn, we experimented the SSI algorithm for three disambiguation tasks:

1. Disambiguation of the words in a WordNet gloss (e.g. *retrospective#1: "an exhibition of a representative selection of an artist's life work"*).

2. Disambiguation of words in a query (e.g queries from TREC web retrieval tasks: *"how we use statistics to aid our decision making?"*).
3. Disambiguation of complex terms (e.g. *connected bus network*).

In task 1, D is initialized with the sense described by the gloss under consideration, possibly augmented with the senses of all unambiguous words in the gloss, e.g. for the *retrospective* example, we have: *D={retrospective#1, statue#1, artist#1}* and *P={work, exhibition, life, selection, representative, art}*.

In task 1, we are sure that D in step 1 includes at least one semantic graph, that of the synset whose gloss we are disambiguating. In the other two tasks, either one of the words at least in set C is monosemous, or the algorithm begins with an initial guess, selecting the most probable sense of the less ambiguous word. If the total score is below a given threshold, the algorithm is then repeated with a different initial guess.

We now consider a complete example of the SSI algorithm for the complex term disambiguation task: *connected bus network*. As no word is monosemous, the algorithm makes a guess about the sense of the less ambiguous word, namely *network*. The only sense of *network* passing the threshold is #3, "an intersected or intersecting configuration or system of components". Initially we have $D = \{network\#3\}$ and $P = \{connected, bus\}$. At the first step, the following pattern involving the domain label relation is matched: *network#3* $\xrightarrow{dl}$ *connected#6* (i.e. the two concepts have the same domain label "computer_science"). So, $D = \{network\#3, connected\#6\}$ and $P = \{bus\}$. Finally, linguistic parallelism (i.e. the two concepts have a common ancestor) and domain label patterns provide the correct indication for the choice of the second sense of bus, "an electrical conductor that makes a common connection between several circuits". The final configuration is thus $D = \{network\#3, connected\#6, bus\#2\}$ and $P = \emptyset$.

2 Evaluation of SSI Algorithm

Each of the three tasks described in previous sections have been evaluated using standard (when available) and ad-hoc test bed. A summary evaluation for each task is shown in the three tables below. Details are provided in previously referenced papers. The baseline in Tables 1 and 3 is computed selecting the first WordNet sense (the most probable according to authors). In Table 3, in order to obtain a 100% recall, sense #1 is selected when no interconnections are found for appropriate sense selection. Furthermore, to increase the set D at step 1, we jointly disambiguate many terms having word strings in common (e.g. *public transport service, bus service, coach service,* etc.).

Table 1. Summary of experiments on gloss disambiguation

Domains	#Glosses	#Words	#Disamb. words	#Disamb. words ok	Recall	Precision	Baseline Precision
Tourism	305	1345	636	591	47.28%	92.92%	82.55%
Generic	100	421	173	166	41.09%	95.95%	67.05%

Table 2. Summary of experiments on sense-based query expansion

First 20 TREC 2002 web track queries	Without sense expansion (baseline)	With sense expansion (best expansion strategy)
Avg. n. of correct retrieved GOOGLE pages over first 10	5.12	6.29
% of increase over baseline	–	22.76%

Table 3. Summary of experiments on complex term disambiguation

# of complex terms (tourism domain)	Average words per term	Precision	Baseline Precision
650	2.2	84.56%	79.00%

As shown in Table 3 and in other papers, the performance of the SSI algorithm in the WordNet extension task is between 84% and 89% depending upon domains. Furthermore, the extended WordNet may include other types of errors (e.g. inappropriate terminology), therefore it needs to be inspected by domain experts for refinements. To facilitate the human task of evaluating new proposed concepts, we defined a grammar for each semantic relation type to compositionally create a gloss for new complex concepts in an automatic fashion.

Let $cc(h,k) = S_j^k \xrightarrow{sem\_rel} S_l^h$ be the complex concept associated to a complex term $w_h w_k$ (e.g. *coach service, or board of directors*), and let:

<**GNC**> be the gloss of the new complex concept $cc(h,k)$;
<**HYP**> the direct hyperonym of $cc(h,k)$ (e.g. respectively, *service#1* and *board#1*);
<**GHYP**> the gloss of HYP;
<**FPGM**> the main sentence of the correct gloss of the complex term modifier (e.g respectively: *coach, director*).

We provide here two examples of rules for generating GNC:

1. if sem_rel=*attribute*, <GNC>::=**a kind of** <HYP>, <GHYP>, <FPGM>
2. if sem_rel=*purpose*, <GNC>::=**a kind of** <HYP>, <GHYP>, **for**<FPGM>

The following are examples of generated definitions for rules 1 and 2.

COMPLEX TERM: Traditional garment *(tourism)*
<HYP>::=garment#1
<GHYP>::=an article of clothing
<FPGM>::=consisting of or derived from tradition
<GNC>::=**a kind of** garment, an article of clothing, consisting of or derived from tradition

COMPLEX TERM: Classification rule *(computer science)*
<HYP>::=rule#11
<GHYP>::=a standard procedure for solving a class of problems
<FPGM>::= the basic cognitive process of arranging into classes or categories
<GNC>::=**a kind of** rule, a standard procedure for solving a class of problems, **for** the basic cognitive process of arranging into classes or categories

3 Conclusion

Current research on OntoLearn follows two directions: on the theoretical side, we are trying to obtain a better formalization of the structural semantic interconnection methodology through the use of graph grammars. On the application side, we are extending the type of semantic information that is extracted by Ontolearn. Furthermore, we are augmenting the information represented in semantic graphs, using other semantic resources, such as FrameNet.

References

1. Gangemi, A., Navigli, R., Velardi, P.: Axiomatizing WordNet: a Hybrid Methodology. Workshop on Human Language Technology for the Semantic Web and Web Services at the 2003 International Semantic Web Conference, Sanibel Island, Florida, USA (2003).
2. Missikoff, M., Navigli, R., Velardi, P.: Integrated Approach for Web Ontology Learning and Engineering. IEEE Computer, November 2002.
3. Navigli, R., Velardi, P., Gangemi, A.: Corpus Driven Ontology Learning: a Method and its Application to Automated Terminology Translation. IEEE Intelligent Systems **18** (2003) 22–31.
4. Navigli, R., Velardi, P.: An Analysis of Ontology-based Query Expansion Strategies. Workshop on Adaptive Text Extraction and Mining at the 14þ European Conference on Machine Learning, Cavtat-Dubrovnik, Croatia (2003).
5. Olszewski, R. T.: Generalized Feature Extraction for Structural Pattern Recognition. In Time-Series Data, PhD dissertation, Carnagie Mellon University CMU-CS-01-108 (2001).

Concerning the Difference Between a Conception and its Application in the Case of the Estonian WordNet

Heili Orav and Kadri Vider

University of Tartu, Department of General Linguistics,
Liivi 2, 50409 Tartu, Estonia

Abstract. One source of Estonian WordNet have been corpora of Estonian. On the other hand, we get interested in word sense disambiguation, and about 100,000 words in corpora are manually disambiguated according to Estonian WordNet senses. The aim of this paper is to explain some theoretical problems that "do not work well in practice". These include the differentiation of word senses, metaphors, and conceptual word combinations.

1 Introduction

By now the research group of computational linguistics at the University of Tartu has worked six years on the thesaurus of Standard Estonian or the Estonian WordNet (EstWN)[1]

Although the thesaurus covers only about ten thousand concepts, experiments in the disambiguation of textual words show that thesaurus entries cover the majority of senses of Estonian core vocabulary [1].

When setting up the Estonian WordNet we followed the principles of Princeton WordNet and EuroWordnet. For a more detailed discussion see Kahusk and Vider [2].

The existing thesaurus was used as the Estonian basic lexicon for SENSEVAL-2 contest[2].

The aim of this paper is to point out some theoretical problems that 'do not work well in practice'. These include the differentiation of word senses, metaphors, and conceptual word combinations.

2 Estonian WordNet and Word Sense Disambiguation Task

Lexically the thesaurus is derived from the existing traditional dictionaries (mainly the "Explanatory Dictionary of Estonian") and a text corpus (providing information about usage).

At present the Estonian WordNet contains about ten thousand synsets: mostly noun (66 %) and verb concepts (27 %), but also a limited number of adjectives (2.6 %) and proper nouns (4.4 %). Each synset has more than two semantic relations; hyponymic and hyperonymic relations predominate.

[1] This paper is based on work supported in part by the Estonian Science Foundation under Grant 5534 and by Estonian State Target Financing R&D project number 0182541s03 "Eesti keele arvutimudelid ja keeleressursid: teoreetilised ja rakenduslikud aspektid."

[2] See `http://www.sle.sharp.co.uk/senseval2/`

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 285–290.

We got interested in word sense disambiguation (WSD) couple of years ago and at present time we have a corpus of about 100,000 manually disambiguated textual words. The texts were taken from the Corpus of Estonian Literary Language. The sense numbers of the Estonian thesaurus were used to disambiguate only nouns and verbs because the including of adjectives in the thesaurus began only recently.

At present we are adding new word senses to the EstWN on the basis of word sense disambiguation. These findings reveal some theoretical and practical drawbacks in setting up the thesaurus.

3 Too Broad or Too Narrow?

When looking up the meaning of a specific textual word in the thesaurus, it often seems that the meaning recorded in the thesaurus is either too specific or too general for the given context. The disambiguation of word senses in a text reveals quite clearly that a broader or narrower meaning of the word is synonymous with the senses of other words in a concrete usage but not in the conceptual system.

Let us take a look at the example sentence

Example 1. Laps läks kooli 'the child went to school',

where it is irrelevant whether the child went to school as an educational institution or a building, or actually both were meant. At the same time the sentence

Example 2. Linn on ehitanud sel aastal juba kolm kooli 'this year the town has already built three schools'

means that in this case only the school building is meant.

kool_1 [polysemic sense that applies both to the institution and the building]
⇒**kool_2** [school building]
⇒**kool_3** [educational institution]

Fig. 1. Hyponymic senses for *kool* ('school')

If the thesaurus provides the hyponymic and hyperonymic senses for the word *kool* 'school' (see Figure 1), there will be more than enough different senses of *kool*. The second and the third senses (narrower senses) are covered by sense 1 as a more general one. In the case of manual disambiguation the marking of the more general sense (sense 1) is usually justified. Sense 2 will be needed only for such cases as example sentence 2. However, if the synset including sense 1 has both the building and institution as its hyperonyms, then *kool* in sentence 2 could be disambiguated correctly by means of sense 1 as well.

In a semantically related thesaurus like WordNet each synset can have only a single hyperonymic relation. Therefore, it is highly inconvenient to present regular polysemy, and one tries to avoid polysemy by adding broader or narrower senses of the same word. This,

however, creates for the semantic disambiguator a disturbingly large number of senses that are rarely used and are difficult to distinguish from one another.

One way to decide whether the addition of a narrower or broader sense to the thesaurus is justified is to find translation equivalents for the meanings of textual words. For example, the Estonian verb *kuduma* has at least two clearly distinguishable senses that belong into different synsets in English Wordnet (see Figure 2).

kuduma_1 *weave, tissue* [of textiles; create a piece of cloth by interlacing strands, such as wool or cotton]
kuduma_2 *knit* [make textiles by knitting]

Fig. 2. Different senses of verb *kuduma* belong into different synsets, and have different literals in English ('weave' and 'knit')

The above-mentioned WordNet senses correspond to subdivisions 1.a. and 1.b. of entry *kuduma* in "Explanatory Dictionary of Estonian". It means that they are regarded as rather specific subsenses of the more general meaning of *kuduma* 1. in Estonian. However, it is difficult to find an example of the verb *kuduma* in the text, where it is not important whether one is weaving a fabric or knitting using knitting needles. It shows that the thesaurus has to introduce two clearly distinguishable senses of *kuduma* (in addition to senses 2 and 3 provided in the explanatory dictionary). For the same reason, one might omit the more general sense of *kuduma* (sense 1 in the explanatory dictionary).

Naturally it is difficult and perhaps even impossible to distinguish the meanings in one language from the perspective of many other languages, and there is no good reason for preferring a certain language for translation equivalents for the purpose of a monolingual thesaurus. However, one should consider the use of translation equivalents as a possibility if the thesaurus makers disagree on whether the senses in the thesaurus are too narrow or too broad.

4 What Should We Do with Metaphors?

Metaphors and metaphorical meanings of words are a topical issue in linguistics and lexicology. Even the well-known psycholinguist and founder of WordNet George A. Miller provided a thorough classification of metaphors in "Metaphor and Thought" [3].

Metaphors present an appropriate touchstone for a thesaurus. They raise the question whether the senses arising from the metaphorical use of words should be added as new meanings to the thesaurus or not. Their occurrence in text is really rather unpredictable and chaotic. And if we add the metaphorical uses to the thesaurus, then how should we explain them properly. As is known, the understanding of a metaphor depends on the context.

Below you will find an example from our semantically disambiguated corpus:

Example 3.

Loopis taas oma murruvahus latvu vastu kaldakivisid, peksis neid vanu vaenlasi, kes kuidagi ei tahtnud endid veerevate lainemägede teelt ära koristada. (tkt0034)

'it was once again hurling its foamy tops against the rocks, it was lashing its old enemies who wouldn't make way to the rolling mountainous waves'

The author has described a stormy sea. In the case of manual semantic disambiguation one would ask the question what do the words *latv* 'treetop', *vaenlane* 'enemy', *loopima* 'hurl', *peksma* 'beat, lash', *koristama* 'clean, clear' mean. One might presume that these words have specific meanings in the thesaurus that cannot be extended to the textual meanings without pointing out their metaphoricalness.

It is possible to distinguish between two main types of knowledge in the comprehension of a text [4]:

1. semantic knowledge is knowledge of extralinguistic reality;
2. pragmatic knowledge is knowledge regulating communication (social norms, conventions).

Because EstWN is based on the existing traditional dictionaries and a text corpus (providing usage information), one might suppose that the semantic information in the database reflects semantic knowledge.

The addition of metaphors to the thesaurus would make it a thesaurus that combines semantic and pragmatic combinations. It would increase the size of the thesaurus to a remarkable degree. For this reason until now we have tried to avoid the addition of metaphors, but problems are opened.

5 Conceptual Word Combinations

Conceptual word combinations present another problem in the disambiguation of word senses. The thesaurus includes 984 such combinations as entries, three quarters of them being phrasal and phraseological verbs. They are mostly two-word combinations, but there are also some three- and even four-word combinations as well.

Comparison with the database of Estonian collocations (multi-word units, see Kaalep & Muischnek [5])[3] shows that 635 items overlap as phrasal and phrsaeological verbs and only six as noun expressions.

Why do we call them conceptual word combinations and not phraseological units? Phraseology proceeds from language use, and a phraseological unit is a combination that is always used together but the meaning of which differs from the sum of the meanings of its constituents [6]. A large number of conceptual word combinations in the thesaurus are phraseological units as well (metaphorical phraseological verbs, for example). On the other hand, the thesaurus entries include many combinations constituting a conceptual whole. They cannot be regarded as phraseological units because their meaning arises from the meaning of their constituents, and they are not collocations in statistcal terms.

Conceptual word combinations became thesaurus entries as:

[3] See `http://www.cl.ut.ee/ee/ressursid/pysiyhendid.html`.

1. synonyms (e.g. *meenutama, meelde tuletama* ‘recall, remember’; *üllitama, välja andma* ‘publish’);
2. specifi c hierarchical nodes (e.g. *emotsionaalne seisund* ‘emotional state’, *ruumiline omadus* ‘spatial characteristic’, *üleloomulik olend* ‘supernatural creature’, *suuruse või koguse muutmine* ‘modifi cation of size or amount’;
3. technical terms (e.g. *ilmaütlev kääne* ‘abessive case’, *damaskuse teras* ‘Damascus steel’, *kreeka tähestik* ‘Greek alphabet’;
4. explanations (e.g. *kultiveerima, kultuurina kasvatama* ‘cultivate, grow as a culture’, *naer, naeru hääl* ‘laughter, sound of laughter’, *hääletaja, pöidlaküüdiga sõitja* ‘hitchhiker, a person thumbing a lift’.

Synonyms (1) and technical terms (3) are the only groups of word combinations that justify their inclusion in the thesaurus from the perspective of word sense disambiguation. From the same perspective one can only welcome the fact that two thirds of the word combinations included in the thesaurus can be also found in the database of multi-word units. The latter database is likely to serve in the future as a basis for morphological and syntactic recognition of word combinations in texts. Once the computational analysis of previous levels is able to recognize multi-word units in a text, it will be possible to fi nd the matching senses in the thesaurus. Because it is likely that the components of noun combinations occur close to each other in a text, formally it is easier to spot them fi rst automatically and then compare them against the word list of the thesaurus. The recognition of verb combinations, however, is still an unmanageable task for lemmatizers. Due to inadequate pre-processing at the present level of semantic disambiguation the conceptual word combinations are provided with wrong meanings both in the course of automatic tagging and sometimes also in manual tagging. On the other hand, the thesaurus includes as synonyms a certain number of (verb) combinations that are not included in the database of multi-word units because of their rare occurrence. However, these combinations are essential for the thesaurus (e.g. *arvamusele jõudma* ‘reach an opinion’, *keelele tulema* ‘come on the tip of one’s tongue’, *ühel meelel olema* ‘be of the same opinion’). Thus, these combinations should be included in the database of multi-word units in cooperation with the creators of this database.

Thus, the combinations in groups (2) and (4) seem useless from the perspective of word sense disambiguation. If we defi ne these groups on the basis of absence from the database of multi-word units, then it will be easy to fi nd a good reason for carrying out a semantic analysis by components once the fi xed combination recognition software is complete. There is strong likelihood that this is going to happen to the explanatory conceptual combinations of group (d). In addition, one should also consider their suitability as thesaurus entries. It would be reasonable to place such combinations in the explanation fi eld of a synonymous entry.

6 Conclusions

It appears that the creation of a concept-based thesaurus is not as easy as it seems at fi rst sight. The main problems in setting up a thesaurus include:

- under- or over-differentiation;
- metaphors;
- conceptual word combinations.

The practical use of the thesaurus in WSD task showed that the senses based on the traditional defi ning dictionary and the intuition of lexicographers may be either too narrow or too broad. This fact compels the thesaurus makers to order the word senses both in the thesaurus and to think about the reliability of the previous theoretical views.

At the same time semantic disambiguation experiments show that the meaning of the sentence and the meaning of the lexical words constituting the sentence are largely dependent on the functional words. Unfortunately, the latter are not included in the thesaurus, and the semantic tagging system that is based on the thesaurus does not take them into account. Prior recognition of conceptual word combinations would make at least one part of such meaning-differentiating units 'visible' for word sense disambiguation.

References

1. Kahusk, N., Orav, H., Õim, H.: Sensiting inflectionality: Estonian task for SENSEVAL-2. In: Proceedings of SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguating Systems, Toulouse, France, CNRS—Institut de Recerche en Informatique de Toulouse, and Univeresité des Sciences Sociales (2001) 25–28.
2. Kahusk, N., Vider, K. In: Estonian WordNet Benefits from Word Sense Disambiguation. Central Institute of Indian Languages, Mysore, India (2002) 26–31.
3. Miller, G. A. In: Images and models, similes and metaphors. 2nd edn. Cambridge University Press (1979).
4. Õim, H.: Семантика и теория понимания языка. Анализ лексики и текстов директивного общения эстонского языка. PhD thesis, University of Tartu (1983).
5. Kaalep, H.J., Muischnek, K. In: Inconsistent Selectional Criteria in Semi-automatic Multi-word Unit Extraction. Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest (2003) 27–36.
6. Õim, A.: Fraseoloogiasõnaraamat. Eesti Teaduste Akadeemia Keele ja Kirjanduse Instituut, Tallinn, Estonia (1993).

Soft Word Sense Disambiguation

Ganesh Ramakrishnan, B. P. Prithviraj, A. Deepa, Pushpak Bhattacharyya, and Soumen Chakrabarti

Department of Computer Science and Engineering, Indian Institute of Technology, Mumbai, India
Email: hare@cse.iitb.ac.in, prithvir@cse.iitb.ac.in, adeepa@cse.iitb.ac.in, pb@cse.iitb.ac.in, soumen@cse.iitb.ac.in

Abstract. Word sense disambiguation is a core problem in many tasks related to language processing. In this paper, we introduce the notion of *soft word sense disambiguation* which states that *given a word, the sense disambiguation system should not commit to a particular sense, but rather, to a set of senses which are not necessarily orthogonal or mutually exclusive*. The senses of a word are expressed by its WordNet synsets, arranged according to their relevance. The relevance of these senses are probabilistically determined through a Bayesian Belief Network. The main contribution of the work is a completely probabilistic framework for word-sense disambiguation with a semi-supervised learning technique utilising WordNet. WordNet can be customized to a domain using corpora from that domain. This idea applied to question answering has been evaluated on TREC data and the results are promising.

Keywords: Soft Sense Disambiguation, Synset-Ranking, Bayesian Belief Networks, Semi-supervised learning

1 Introduction

Word sense disambiguation is defined as the task of finding *the* sense of a word in a context. In this paper, we explore the idea that one should not commit to a particular sense of the word, but rather, to a *set of its senses* which are not necessarily orthogonal or mutually exclusive. Very often, WordNet gives for a word multiple senses which are related and which *help connect* other words in the text. We refer to this observation as the relevance of the sense in that context. Therefore, instead of picking a single sense, we rank the senses according to their relevance to the text. As an example, consider the usage of the word *bank* in fig. 1. In WordNet, *bank* has 10 noun senses. The senses which are relevant to the text are shown in figure 2.

A passage about some bank A Western Colorado bank with over $320 Million in assets, was formed in 1990 by combining the deposits of two of the largest and oldest financial institutions in Mesa County

Fig. 1. One possible usage of *bank* as a *financial_institution*

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 291–298.

Relevant senses

1. *depository financial institution, bank, banking concern, banking company:* a financial institution that accepts deposits and channels the money into lending activities; "he cashed a check at the bank"; "that bank holds the mortgage on my home"
2. *bank, bank building:* a building in which commercial banking is transacted; "the bank is on the corner of Nassau and Witherspoon"
3. *bank:* (a supply or stock held in reserve for future use (especially in emergencies))
4. *savings bank, coin bank, money box, bank:* (a container (usually with a slot in the top) for keeping money at home; "the coin bank was empty")

Fig. 2. Some relevant senses for *bank*

These senses are ordered according to their relevance in this context. It is apparent that the first two senses have equal relevance. The applicability of the senses tapers off as we move down the list. This example motivates soft sense disambiguation. We define *soft sense disambiguation* as the process of enumerating the senses of a word in a ranked order. This could be an end in itself or an interim process in an IR task like question answering.

1.1 Related Work

[Yarowsky 1992] proposes a solution to the problem of WSD using a thesaurus in a supervised learning setting. Word associations are recorded and for an unseen text, the senses of words are detected from the learnt associations. [Agirre and Rigau 1996] uses a measure based on the proximity of the text words in WordNet (*conceptual density*) to disambiguate the words. The idea that translation presupposes word sense disambiguation is leveraged by [Nancy 1999] to disambiguate words using bi-lingual corpora. The design of the well-known work-bench for sense disambiguation *WASP* is given in [Kilgarriff 1998]. The idea of constructing a BBN from WordNet has been proposed earlier by [Wiebe, Janyce, et al. 1998] and forms a motivation for the present work. However, unlike [Wiebe, Janyce, et al. 1998] we particularly emphasise the need for soft sense disambiguation, *i.e.* synsets are considered to probabilistically cause their constituent words to appear in the texts. Also we describe a comprehensive training methodology and integrate soft WSD into an interesting application, *viz.*, QA. Bayesian Balief Network (BBN) is used as the machine for this probabilistic framework. It is also demonstrated, how the BBN can be customized to a domain using corpora from that domain.

2 Our Approach to Soft WSD

We describe how to induce a Bayesian Belief Network (BBN) from a lexical network of relations. Specifically, we propose a semi-supervised learning mechanism which simultaneously trains the BBN and associates text tokens, which are words, to synsets in WordNet in a probabilistic manner ("soft WSD").

In general, there could be multiple words in the document that are caused to occur together by multiple hidden concepts. This scenario is depicted in figure 3. The causes themselves may have hidden causes.

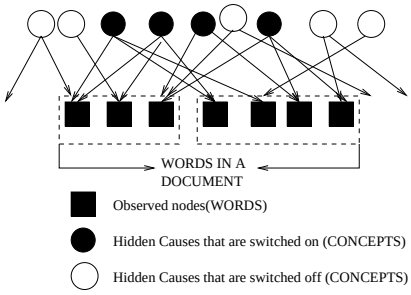


Fig. 3. Motivation

These causal relationships are represented in WordNet which encodes relations between words and concepts (synsets). For instance WordNet gives the *hypernymy* relation between the concepts { animal} and { bear}.

2.1 Inferencing on Lexical Relations

It is diffi cult to link words to appropriate synsets in a lexical network in a principled manner. On the example of *animal* and *bear*, the English WordNet has fi ve synsets on the path from *bear* to *animal*: {carnivore...}, {placental_mammal...}, {mammal...}, {vertebrate..}, {chordate...}. Some of these intervening synsets would be extremely unlikely to be associated with a corpus that is not about zoology; a common person would more naturally think of a *bear* as a kind of animal, skipping through the intervening nodes.

Clearly, any scoring algorithm that seeks to utilize WordNet link information must also *discriminate* between them based (at least) on usage statistics of the connected synsets. Also required is an estimate of the likelihood of instantiating a synset into a token because it was *activated* by a closely related synset. We fi nd a Bayesian belief network (BBN) a natural structure to encode such combined knowledge from WordNet and corpus (for training).

2.2 Building a BBN from WordNet

Our model of the BBN is that each synset from WordNet is a boolean *event* associated with a word. Textual tokens are also events. Each event is a node in the BBN. Events can *cause* other events to happen in a probabilistic manner, which is encoded in Conditional Probabiity Tabless. The specifi c form of CPT we use is the well-known *noisy-OR* for the words and *noisy-AND* for the synsets. This is because a word is *exclusively* instantiated by a cluster of parent synsets in the BBN, whereas a synset is compositionally instantiated by its parent synsets. The noisy-OR and noisy-AND models are described in [J. Pearl 1998].

We introduce a node in the BBN for each noun, verb, and adjective synset in WordNet. We also introduce a node for each token in the corpus. Hyponymy, meronymy, and attribute links are introduced from WordNet. *Sense links* are used to attach tokens to potentially matching synsets. For example, the string “flag” may be attached to synset nodes {sag, droop, swag, flag} and {a conspicuously marked or shaped tail}. (The purpose of probabilistic

disambiguation is to estimate the probability that the string "flag" was *caused* by each connected synset node.)

This process creates a hierarchy in which the parent-child relationship is defined by the semantic relations in WordNet. A is a parent of B iff A is the *hypernym* or *holonym* or *attribute-of* or A is a synset containing the word B. The process by which the BBN is built from WordNet graph of synsets and from the mapping between words and synsets is depicted in figure4. We define *going-up* the hierarchy as the traversal from child to parent.

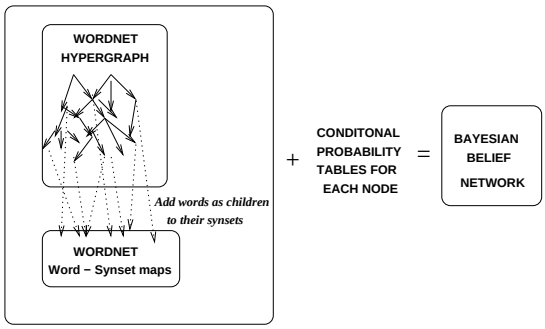


Fig. 4. Building a BBN from WordNet and associated text tokens.

2.3 Training the Belief Network

The figure5 describes the algorithm for training the BBN obtained from the WordNet. We initialize the CPTs as described in the previous section. The instances we use for training are windows of length M each from the untagged corpus. Since the corpus is not tagged with WordNet senses, all variables, other than the words observed in the window (i.e. the synset nodes in the BBN) are hidden or unobserved. Hence we use the Expectation Maximization algorithm [Dempster 1977] for parameter learning. For each instance, we find the expected values of the hidden variables, given the "present" state of each of the observed variables. These expected values are used after each pass through the corpus to update the CPT of each node. The iterations through the corpus are done till the sum of the squares of Kullback-Liebler divergences between CPTs in successive iterations do not differ more than a small threshold. In this way we customize the BBN CPTs to a particular corpus by learning the local CPTs.

3 The WSD Algorithm: Ranking Word Senses

Given a passage, we clamp the BBN nodes corresponding to words, to a state of 'present' and infer using the network, the score of each of its senses which is the probability of the corresponding synset node being in a state of "present". For each word, we rank its senses in decreasing order of its score. In other words, the synset given the highest rank (probability) by this algorithm becomes the most probable sense of the Word.

```
1: while CPTs do not converge do
2:    for each window of M words in the text do
3:       Clamp the word nodes in the Bayesian Network to a state of 'present'
4:       for each node in Bayesian network do
5:          find its joint probabilities with all configurations of its parent nodes (E Step)
6:       end for
7:    end for
8:    Update the conditional probability tables for all random variables (M Step)
9: end while
```

Fig. 5. Training the Bayesian Network for a corpus

```
1: Load the Bayesian Network parameters
2: for each passage p do
3:    clamp the variables (nodes) corresponding to the passage words (w1, w2...wn) in network to
      a state of 'present'
4:    Find the probability of each sense of each word, being in state 'present' i.e., Pr(s|w1, w2..wn)
5: end for
6: Report the word senses of each word, in decreasing order of ranks.
```

Fig. 6. Ranking word senses

4 Evaluation

We use documents from *Semcor 1.7.1 corpus* [Semcor] for disambiguation. Semcor corpus is a subset of the famous Brown corpus [Brown Corpus] sense-tagged with WordNet 1.7.1 synsets. Our soft WSD system produces rank ordered synsets on the semcor words (at most two senses). We show below in figure 7 the output of the system for the word *study*. Both semcor's tag and our system's first tag are correct, though they differ. The second tag from our system has low weightage and is wrong in this context. The synsets marked with ** represent the correct meaning.

Passage from Semcor It recommended that Fulton legislators act to have these laws *studied* and revised to the end of modernizing and improving them.

Semcor tag: [Synset: [Offset: 513626] [POS: verb] Words: analyze, analyse, study, examine, canvass – (consider in detail and subject to an analysis in order to discover essential features or meaning; "analyze a sonnet by Shakespeare"; "analyze the evidence in a criminal trial"; "analyze your real motives")]

soft WSD tags: **[Synset: study 0 consider 0 [Gloss =]: give careful consideration to; "consider the possibility of moving" [Score = 0.62514]]

[Synset: study 4 meditate 2 contemplate 0 [Gloss =]: think intently and at length, as for spiritual purposes; "He is meditating in his study" [Score = 0.621583]]

Fig. 7. Example of *first match* with Semcor's marking

Next we present an example of the second marking of the sense being correct. The word in question is the verb *urge* (figure 8).

Passage from Semcor It *urged* that the city take steps to remedy this problem.

Semcor tag: Synset: [Offset: 609547] [POS: verb] Words: urge, urge_on, press, exhort – (force or impel in an indicated direction; "I urged him to finish his studies")

soft WSD tags: [Synset: cheer 1 inspire 1 urge 1 barrack 1 urge_on 1 exhort 1 pep_up 0 [Gloss =]: urge on or encourage esp. by shouts; "The crowd cheered the demonstrating strikers" [Score = 0.652361]]

**[Synset: recommend 1 urge 3 advocate 0 [Gloss =]: push for something; "The travel agent recommended strongly that we not travel on Thanksgiving Day" [Score = 0.651725]]

Fig. 8. Example of the *second match* being correct

Table 1 summarizes soft WSD results obtained by us. If the first meaning given by the soft WSD system is correct then it is counted towards the *first match*; similarly for the *second match*.

Table 1. Results of soft WSD

Total ambiguous nouns	139
Nouns first match	66
Nouns second match	46
Total ambigous verbs	67
verbs first match	24
verbs second match	23

5 An Application: Question Answering

In this section, we mention our work on the extension of ideas presented in the previous sections to the problem of question answering, which inherently requires WSD to connect question words to answer words. The BBN is trained using the algorithm in figure 5 on the corpus to be queried. The trained BBN is used to rank passages (windows of N consecutive words) from the corpus using the algorithm presented in figure 9.

We performed QA experiments on the TREC-9 question-set and the corresponding corpus. The Mean Reciprocal Rank (MRR) figures for the different experiments are presented in table 2. Clearly, inferencing with trained BBN outperforms inferencing with untrained BBN while both inferencing procedures, outperform the baseline algorithm, the standard TFIDF retrieval system.

*The effect of WSD:*It is interesting to note that training does not substantially affect disambiguation accuracy (which stays at about 75%), and MRR improves *despite* this

```
1: Load the Bayesian Network parameters
2: for each question q do
3:    for each candidate passage p do
4:       clamp the variables (nodes) corresponding to the passage words in network to a state of
         'present'
5:       Find the joint probability of all question words being in state 'present' i.e., Pr(q|p)
6:    end for
7: end for
8: Report the passages in decreasing order of Pr(q|p)
```

Fig. 9. Ranking candidate answer passages for given question

Table 2. MRRs for baseline, untrained and trained BBNs

System	MRR
Asymmetric TFIDF	0.314
Untrained BBN	0.429
Trained BBN	0.467

fact. This seems to indicate that learning joint distributions between query and candidate answer keywords (via synset nodes, which are "bottleneck" variables in BBN parlance) is as important for QA as is WSD. Furthermore, we conjecture that "soft" WSD is key to maintaining QA MRR in the face of modest WSD accuracy.

6 Conclusions

In this paper a robust, semi-supervised method for sense disambiguation using WordNet (*soft sense disambiguation*) was described. The WordNet graph was exploited extensively. Also, the task of soft WSD was integrated into an application *viz.* question answering.

The future work consists in exploring the use of links others than the hypernymy-hyponymy. Also WordNet 2.0 provides derivational morphology links between verb and noun synsets, the use of which needs to be investigated. Adjectives and adverbs too have to be tackled in the system. The intervention of human experts at critical steps to improve accuracy is a very interesting issue meriting attention.

The paradigm of *active learning* is highly promising in such problems as are the concern of the present work. With human help the system can tune itself for sense disambiguation using a relatively small number of examples.

References

Fellbaum 1998. Fellbaum Christiane, ed. the WordNet: An Electronic Lexical Database. *MIT Press*, Map 1998.

Nancy 1999. Nancy Ide. Parallel Translations as Sense Discriminators. In *Proceedings of SIGLEX99, Washington D.C, USA, 1999.*

Kilgarriff 1998. Adam Kilgarriff. Gold Standard Data-sets for Evaluating Word Sense Disambiguation Programs. In *Computer Speech and Language 12 (4), Special Issue on Evaluation, 1998.*

Yarowsky 1992. David Yarowsky. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of the* 14^{th} *International Conference on Computational Linguistics (COLING-92)*, pages 454–460, Nantes, France, 1992.

Agirre and Rigau 1996. Agirre, E. and Rigau, G. Word sense disambiguation using conceptual density. In *Proceedings of COLING '96*.

J. Pearl 1998. J. Pearl. In *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* Morgan Kaufmann Publishers, Inc.

Wiebe, Janyce, et al. 1998. Wiebe, Janyce, O'Hara, Tom, Rebecca Bruce. Constructing bayesian networks from WordNet for word sense disambiguation: representation and processing issues In *Proc. COLING-ACL '98 Workshop on the Usage of WordNet in Natural Language Processing Systems*.

Dempster 1977. P. Dempster, N. M. Laird and D. B. Rubin. Maximum Likelihood from Incomplete Data via The EM Algorithm. In *Journal of Royal Statistical Society*, Vol. 39, pp. 1–38, 1977.

Semcor. `http://www.cs.unt.edu/~rada/downloads.html#semcor`.

Brown Corpus.
http://clwww.essex.ac.uk/w3c/corpus_ling/content/corpora/list/private/brown/brown.html

TREC. TREC `http://trec.nist.gov`.

Appendix I: Bayesian Belief Network

A Bayesian Network for a set of random variables $X = \{X_1, X_2, \ldots, X_n\}$ consists of a directed acyclic graph (DAG) that encodes a set of conditional independence assertions about variables in X and a set of local probability distributions associated with each variable. Let $\mathbf{Pa}_i$ denote the set of immediate parents of X_i in the DAG, and $\mathbf{pa}_i$ a specifi c instantiation of these random variables.

The BBN encodes the joint distribution $\Pr(x_1, x_2, \ldots, x_n)$ as

$$\Pr(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} \Pr(x_i | \mathbf{pa}_i) \tag{1}$$

Each node in the DAG encodes $\Pr(x_i|\mathbf{pa}_i)$ as a "conditional probability table" (CPT). Figure §10 shows a Bayesian belief network interpretation for a part of WordNet. The synset {*corgi, welsh_corgi*} has a causal relation from {*dog, domestic_dog, canis_familiaris*}. A possible conditional probability table for the network is shown to the right of the structure.

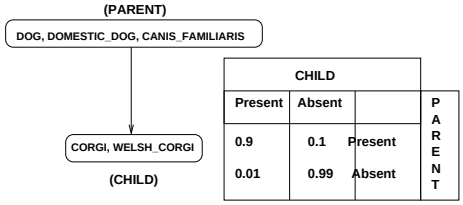


Fig. 10. Causal relations between two synsets.

Text Categorization and Information Retrieval Using WordNet Senses

Paolo Rosso[1], Edgardo Ferretti[2], Daniel Jiménez[1], and Vicente Vidal[1]

[1] Dept. of Computer Systems and Computation,
Polytechnic University of Valencia, Spain.
Email: prosso@dsic.upv.es, djimenez@dsic.upv.es, vvidal@dsic.upv.es
[2] LIDIC-Dept. of Computer Science,
National University of San Luis, Argentina.
Email: ferretti@unsl.edu.ar

Abstract. In this paper we study the influence of semantics in the Text Categorization (TC) and Information Retrieval (IR) tasks. The K Nearest Neighbours (K-NN) method was used to perform the text categorization. The experimental results were obtained taking into account for a relevant term of a document its corresponding WordNet synset. For the IR task, three techniques were investigated: the direct use of a weighted matrix, the Singular Value Decomposition (SVD) technique in the Latent Semantic Indexing (LSI) model, and the bisecting spherical k-means clustering technique. The experimental results we obtained taking into account the semantics of the documents, allowed for an improvement of the performance for the text categorization whereas they were not so promising for the IR task.

1 Introduction

Nowadays, nearly all kind of information is stored in electronic format: digital libraries, newspapers collections, etc. Internet itself can be considered as a great world database which everybody can access to from everywhere in the world. In order to provide inexperienced users with a flexible access to information, it is crucial to take into account the meaning expressed by the documents, that is, to relate different words but with the "same" information. The classical vector space model introduced by Salton [10] for IR was shown by Gonzalo et al. [4] to give better results if WordNet synsets are chosen as the indexing space instead of terms: up to 29% improvement in the experimental results was obtained for a manually disambiguated test collection derived from the SemCor corpus.

2 Document Codification: Vector of Terms and Vector of Synsets

In the present study, we used the vector space model for the codification of a document with a vector of terms. The vector space model was also used when WordNet synsets were chosen as the indexing space instead of word forms, in order to relate different terms with the same information. Due to the phenomenon of polysemy, it was important to identify the exact meaning of each term. The disambiguation of the meaning of the term was obtained through its context (i.e., the portion of the text in which it is embedded), the WordNet ontology [7]

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 299–304.

and a collection of sense-tagged samples, to train the supervised method for the Word Sense Disambiguation (WSD) [8]. In order to perform the WSD, each term of a document needed first to be tagged (as noun, verb, adjective or adverb) according to its morphological category. This Part-Of-Speech (POS) task was performed by the TnT POS-tagger [1]. The POS-tagged vector of each document was used as input data for the supervised sense-tagger. The final output was a *sense-tagged vector*, that is, a vector tagged with the disambiguated sense for each term of the document of the data sets. In the final vector of each document (and query of the IR task), those terms that were not sense-tagged were removed.

3 The Semantic K Nearest Neighbours Technique

The K Nearest Neighbours is one of the most used techniques for the text categorization task due to its good performance. Given a set of labelled prototypes (i.e., categories) and a test document, the K-NN method finds its k nearest neighbours among the training documents. The categories of the K neighbours are used to select the nearest category for the test document: each category gets the sum of votes of all the neighbours belonging to it and that one with the highest score is chosen. Other strategies calculate these scores taking into account the distances between the K neighbours and the test document or, alternatively, using a similarity measure like the scalar product. In this last strategy, which is the one that we used in our work, each document is represented through a vector of terms and each category gets a score equal to the sum of the similarities between the K neighbours and the test document.

The number of terms of any given collection of documents of medium size may be approximately ten of thousands. Therefore, it was very important to optimise the list of terms that identified the collection. This optimisation was focused to reduce the number of terms eliminating those with poor information. A list of stopwords was used to reduce the number of terms that identify the collection. It included terms which did not provide any relevant information: typically, words as prepositions, articles, etc. Some of these techniques help to improve the results of categorization in determined data sets, once noisy vocabulary is eliminated. There are several methods for selecting the terms to remove. In our work, we employed the *Information Gain* (IG) method [13]. IG measured the amount of information which contributed a term for the prediction of a category, as a function of its presence or absence in a given document. Once calculated the IG_i value for each term i, those terms with the highest value were selected being the most relevant.

4 The Techniques for Information Retrieval

The IR models used in this work are classified within the vector space model and are based in the well-known matrix of terms by documents. With the weighted matrix we modelled the IR system induced by the document collection. We also investigated the LSI model, which is based on the SVD technique, and a clustering model which uses the bisecting spherical k-means algorithm.

4.1 The LSI Technique

There are several techniques in the LSI model. Our approach is based on the SVD technique, in which a part of the spectrum of the singular values of the matrix is calculated [2]. Given a

partial SVD of an arbitrary matrix M, we must find p numbers $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_p \geq 0$ and p vectors $u_i \in \Re^m$ and $v_i \in \Re^n$ such that:

$$M \approx M_p = U_p \Sigma_p V_p^T = \sum_{i=1}^{p} u_i \sigma_i v_i^T \quad (1)$$

The evaluation of queries within the SVD technique is based on the calculation of te angle between the query vector with all the document vectors of the collection.

4.2 The Clustering Technique

When searching for a document, it is often useful (for speed, efficiency, or undestandability) to provide it with a structure. In an electronic document collection, such structure should be provided automatically, and may be based on several similarity criteria: by contained terms, by document structure, by document category, by meaning of content. A popular structure is provided by grouping, or *clustering*. The clustering technique used in this work to evaluate semantic lemmatisation (i.e, the expansion to synonyms) was the Bisecting-Spherical K-Means [5]. This algorithm tries to join the advantages of the Bisecting K-Means algorithm with the advantages of a modified version of the Spherical K-Means. The Bisecting-Spherical K-Means clustering algorithm tries to find k disjoint clusters $\{\pi_j\}_{j=1}^{k}$, from the document collection expressed by matrix M such that it maximizes the following objective function:

$$f\left(\{\pi_j\}_{j=1}^{k}\right) = \sum_{j=1}^{k} \sum_{m \in \pi_j} m^t c_j \quad (2)$$

where c_j is the normalised *centroid or concept vector of the cluster* π_j, which it is calculated given the following expression:

$$t_j = \frac{1}{n_j} \sum_{m \in \pi_j} m \; ; \; c_j = \frac{t_j}{\|t_j\|} \quad (3)$$

where n_j is the number of documents in the cluster π_j.

5 Experimental Results: The Influence of Semantics

5.1 The Text Categorization Task

Different experiments were carried out over the modified 20Newsgroups corpus [9] which was pre-processed taking into account for each relevant term its WordNet synset. For each document, its vectors of terms and WordNet synsets were obtained using the *Rainbow system* [6]. The text categorization task was performed employing the K-NN method, where K was set equal to 30. The 30-KNN classifier carried out the text categorization taking into account the semantics of each document. For this experiment, the vector of synsets of each document was used, instead of its vector of terms.

The goodness of the semantic K-NN classifier was measured determining the error percentage obtained classifying a set of test documents. Figure 1 shows the comparison of the error percentage obtained with (WordNet synsets) and without (terms) the introduction of the semantics with respect to the size of the vocabulary.

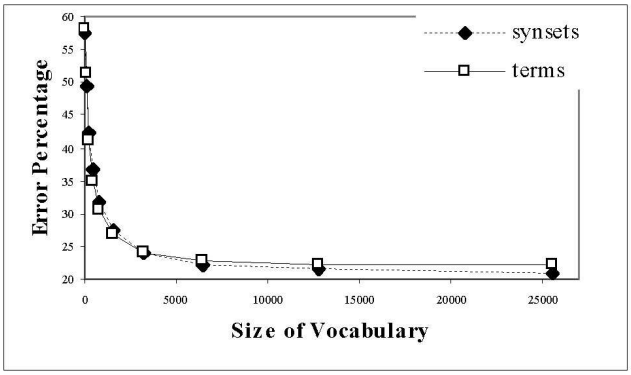


Fig. 1. Text categorization (20Newsgroups corpus): terms vs. WordNet synsets

5.2 The Information Retrieval Task

The criteria used to evaluate the IR experiments, was the average precision-recall ratio:

$$\bar{P}(r) = \sum_{i=1}^{N_q} \frac{P_i(r)}{N_q}$$

where $\bar{P}(r)$ is the average precision at the recall level r, N_q the number of queries used, and $P_i(r)$ the precision at recall level r for the i-th query. To get each $P_i(r)$, first we evaluated the i-th query obtaining a sorted document set ordered by relevance. Then we calculated the precision each time a relevant document appeared in the answering set. In this data set we interpolated 11 standard recall levels as follows: let $r_j \in \{0,...,10\}$, be a reference to the j-th standard recall level, then, $P(r_j) = \max_{r_j \le r \le r_j+1} P(r)$.

The collection used for the experiments contains articles from the 1963 Times Magazine [12]. Query statistics were also obtained for the query collection, formed by a total of 83 queries with an average of 15 words and one line per query. In Figure 2 the most representative results of the study are presented: concretely, the SVD and clustering comparisons between semantic lemmatisation and stemming, which associates words by the root. In fact, words usually have different morphological variants with similar semantic interpretations which would be considered as the same term in IR systems. Stemming algorithms (or stemmers) attempt to reduce a word to its stem or root form. The joining of words with the same information to a single term, also reduces the number of terms that identify the document collection. The experiments were carried out employing the Paice stemming algorithm [3]. In all the studied cases, the semantic lemmatisation had a worse performance than the stemmer. We can observe that the performance of the semantic lemmatisation with the SVD is slightly better than the semantic lemmatisation with the rest of the methods.

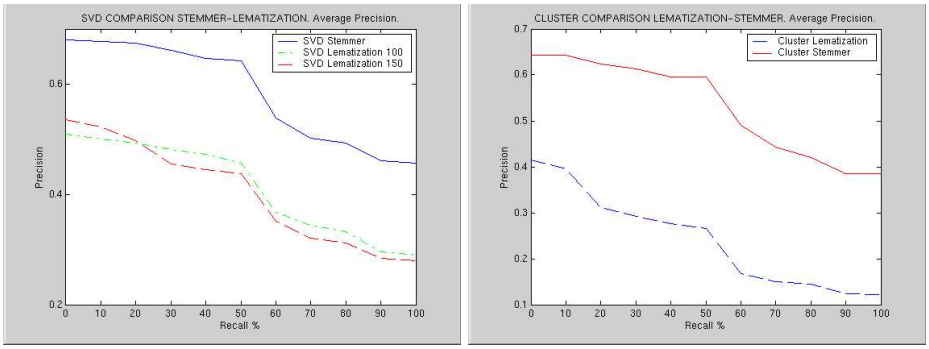


Fig. 2. Semantic lemmatization vs. stemming (Times Magazine corpus): SVD (left) and clustering (right) comparisons.

6 Conclusions and Further Work

In this paper, we investigated whether the introduction of semantic information could help to improve the tasks of TC and IR. With regard to the study of how the semantic 30-KNN performed, it can be remarked that when documents are indexed with WordNet synsets, the performance slightly improved. Therefore, the use of words which refer to the same concept is a research direction we plan to investigate further. As future work, it would be interesting to carry out some experiments using other data sets (e.g. the TREC document collection). In these experiments, the two vector representations should be also combined, in order to take into account with different weights, terms and WordNet synsets at the same time. With regard to the poor performance we obtained for the IR task, it could be due to mainly three reasons. First, the queries of 15 words were pretty long (normal queries are 1.5 words on average) and such long queries implicitly have a disambiguation effect. We should expect better effect of using WordNet for the normal 1 or 2 queries. Second, the semantic lemmatisation related synonyms when they are in the same morphologic group: it should be combined with standard morphological lemmatisation because they could complement each other. Moreover, also other relations could be exploited in the semantic lemmatisation, possibly including the contextual information of the glosses of all the hyponyms. Last, but not least, indexing by WordNet synsets can be very helpful for text retrieval tasks only if the error rate is below 30% [4] and, unfortunately, the state-of-the-art of WSD techniques perform with error rates ranging from 30% to 60% which cannot guarantee better results than standard word indexing.

Acknowledgements

The work of P. Rosso was partially supported by the TIC2000-0664-C02 and TIC2003-07158-C04-03 projects. The work of E. Ferretti was made possible by AECI. We are grateful to A. Molina and F. Pla for making their sense-tagger available.

References

1. Brants, T.: TnT – A Statistical Part-Of-Speech Tagger. In: http://www.coli.uni-sd.de/~thorsten/tnt.
2. Dumais, S., Furnas, G., and Landauer, T.: Using latent semantic analysis to improve access to textual information. In: Proc. of Computer Human Interaction (1988).
3. Fox, C., Fox, B.: Efficient Stemmer Generation Project. In: http://www.cs.jmu.edu/common/projects/Stemming/.
4. Gonzalo, J., Verdejo, F., Chugur, I., Cigarrán, J.: Indexing with WordNet Synsets can improve Text Retrieval. In: Proc. of the Workshop on Usage of WordNet for NLP (1998).
5. Jiménez, D., Ferretti, E., Vidal, V., Rosso, P., and Enguix, C. F.: The Influence of Semantics in IR using LSI and K-Means Clustering Techniques. In: Proc. of Workshop on Conceptual Information Retrieval and Clustering of Documents, ACM Int. Conf., Dublin, Ireland (2003): 286–291.
6. McCallum, A.: Bow: A Toolkit for Statistical Language Modelling, Text Retrieval, Classification and Clustering. In: http://www.cs.cmu.edu/~mccallum/bow/.
7. Miller, A.: WordNet: Lexical Database for English. In: Communications of the ACM, Vol. 38 (1995): 39–41.
8. Molina, A., Pla, F., Segarra. E.: A Hidden Markov Model Approach to Word Sense Disambiguation. In: Proc. of IBERAMIA2002, Seville, Spain, Lecture Notes in Computer Science (2002).
9. Rennie, J.: Original 20 Newsgroups Data Set. In: http://www.ai.mit.edu/~jrennie.
10. Salton, G., Buckley, C.: Term Weighting Approaches in Automatic Text Retrieval. In: Information Processing and Management, Vol. 24 (1998): 513–523.
11. Text Retrieval Conference (TREC) document collection. In: http://www.trec.nist.gov.
12. Times Magazine corpus. In: ftp://ftp.cs.cornell.edu/pub/smart/time/.
13. Yang, Y., Pedersen, O.: A Comparative Study on Feature Selection in Text Categorization. In Proc. of the Int. Conf. on Machine Learning, (1997): 412–420.

Jur-WordNet

Maria Teresa Sagri[1], Daniela Tiscornia[1], and Francesca Bertagna[2]

[1] ITTIG (Institute for Theory and Techniques for Legal Information)-
National Research Council, Via Pianciatichi 56/16, 50127 Firenze, Italy
Email: sagri@Ittig.cnr.it, tiscornia@Ittig.cnr.it
[2] ILC (Istituto di Linguistica Computazionale)-National Research Council,
Via Moruzzi 1, 56100 Pisa, Italy
Email: francesca.bertagna@ilc.cnr.it

Abstract. The paper describes Jur-Wordnet, an extension for legal domain of the Italian ItalWordNet database, aimed at providing a knowledge base for the multilingual access to sources of legal information. Motivations and aims are discussed, together with details concerning the linguistic architecture and construction methodology.

1 Introduction

The subject of this paper is a description of *Jur*-WordNet (*Jur*-WN), an extension for legal domain of the Italian *ItalWordNet* (IWN) database, aimed at providing a knowledge base for the multilingual access to sources of legal information. In the first section of the paper, we will introduce the application needs that are at the basis of the demand of such a lexical resource. A brief description of IWN will be introduced, focussing on the points of contacts between the Italian general wordnet and *jur*-WordNet. Then, the strategies followed during the *jur*-WordNet construction will be describe, with special attention to the handling of lexical polisemy and to the creation of an ontological layer of description.

2 Application Needs for the Legal Sector

The starting point was the *Norme in rete* (Law on the Net) project, launched in 1999 as part of the Italian *E-government Plan. Norme in rete* involves the most important Italian institutions with the goal to "create a portal which, through a single and simple user interface, allows research on all the documentation of normative interest published free on Internet, particularly by institutional sites." [12]. The portal allows free access to normative information through standard methods of editing, processing, and distributing data; the project provides codification standards for source types, identifiers (*urn*[3]), structure, links, and *metainformation.* System design, by now consolidated, consists of classes of XML DTDs[4] for structuring normative texts and of metadata, the most relevant part of which deals

[3] Uniform References Notation, which allows the identification of the partitions of legislative texts independently of the location

[4] See: http://www.normeinrete.it/standard/standard_xml.htm; http://www.lexml.de, http://www.legalxml.org/, http://lri.jur.uva.nl/METALex/.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 305–310.

with the formal/structural features of each type of source, and with *urns* for the identification of the partitions of texts. The aim of Jur-WN is providing the system with a knowledge-base able to supply:

- a source of metadata for the semantic tagging of legislative texts, both at the level of articles and of dispositions. It may also be used in the legislative drafting phase as an enrichment of the specialised XMLeditor now in the development phase [18], and of others legal sources.
- A support resource for information retrieval systems, for facilitating access to heterogeneous and multilingual data.
- An interface between the *common language* approach of citizen and the specific terminology of legal standard[5]. The greatest part of legal thesauri are primarily designed for the "professional" user and not for members of the public.
- A conceptual knowledge base, which can be used for a wide variety of applications and task, such as information extraction, question answering, automatic tagging, knowledge sharing, norm comparison, etc.

3 Overall Architecture of the IWN Database

The EuroWordNet (EWN) [16] project retains the basic underlying design of WordNet [11], trying to improve it in order to answer the needs of research in the computational field, in particular extending the set of lexical relations. In the last years, an extension of the Italian component of EWN was realized with the name of IWN [13]. IWN follows exactly the same linguistic design of EWN (with which shares the Interlingual Index -ILI- and the Top Ontology -TO- as well as the large set of semantic relation[6]) and consists now of about 70,000 word senses organized in about 50,000 *synsets*. Terminological wordnets dedicated to specific domains and linked to the generic module were envisaged, but at the moment only the eco-WordNet module[7] is publicly available, while we are still building the jur-WordNet plug-in. By means of the ILI, all the concepts in the generic and specific wordnets are directly or indirectly linked to the TO. In the EWN model a Domain Ontology was foreseen and in IWN a Domain Ontology was developed for the economic domain. An ontology dedicated to the legal domain is also in construction in *jur*-WN.

3.1 The Plug-in Mechanism

During the IWN project, an innovative methodology (the so-called Plug-in model) for linking domain-independent and domain-specific wordnets was defined. The plug-in relations in jur-

[5] The Proposal for a Directive of the European Parliament and of the Council on the re-use and commercial exploitation of public sector documents(14047/02) is aimed at encouraging the re-use of *Public Sector Information* by private operators *for commercial purposes.* Legal and regulatory information, as well as information on rights and duties are a relevant part of PSI**.** In the regulation of public/private relationship in the market place, the "added value" is a crucial point, dealing with the assessment of Intellectual Property Right and of pricing policies, where added value is mainly conceived as capacity to improve the accessibility for citizen of relevant information, both from a technical and a subjective (content-driven) perspective.

[6] For a complete list of the available semantic relations cf. [13]

[7] developed by Istituto per la Ricerca Scientifica e Tecnologica of Trento (IRST)

WN concern only nouns, which represent the vast majority of the db lexical entries. The plug-in model is realized by means of three plug-in relations defined in order to allow the integrated consultation of the two databases: i) PLUG_SYNONYMY (connecting IWN and domain-specific wordnet whenever it is possible to find an IWN synset having the same meaning of an domain-specific synset), ii) PLUG_NEAR_SYNONYMY (connecting synsets which have 'similar' meanings but are not interchangeable in contexts or whose lists of hyponyms are not compatible) and iii) PLUG_HYPONYMY (connecting an IWN synset and a domain-specific synset with a more specific meaning). The linking via plug-in relations has two effects: (i) the creation of one or more plug-in synsets, where the pairs of synsets involved in the connections are substituted by plug-in synsets and are therefore no longer accessible in the integrated consultation; (ii) the eclipsing of certain synsets, i.e those reachable from IWN through downward links (i.e. its hyponyms) and those reachable from the domain-specific wordnet through upward links (i.e. its hyperonyms). Eclipsed synsets are no longer accessible in the integrated consultation. For a more detailed description of the plug-in model and relations, cf. [13].

4 Jur-WN As a Lexical Resource and a Content Description Model

Jur-WN is a multi-layered lexical resource [14]. First of all, a large set of semantic relations (inherited from the linguistic design of the general IWN database) can be used to link synsets within the same domain-specific module. Then, the plug-in model provides the lexicographer with the possibility to exploit the information already available in the general wordnet, without the necessity to encode general lexical-semantic information from scratch. The latter, more conceptual and abstract layer is the "ontological" one, made up of the higher level of jur-WN, which becomes a core ontology for the legal domain. The first two layers are designed to improve legal information retrieval from heterogeneous (legislation, legal cases, policies) and multilingual sources. Providing a legal lexicon, allowing the handling of linguistic phenomena as polisemy and synonymy, means also to establish a bridge between the common language – often used from the non-jurist ones in order to place legal questions – and the technical language of the law. Under this viewpoint the *plug*-relations linking Jur-WN and Italwordnet allow a more precise definition of technical meanings of terms used in the common Italian, such as *autorizzazione* (*authorisation*), *alienazione* (*alienation*), and the specification of terms acquiring specific law meaning such as *alimenti* (*alimony*) and *mora* (*delay*). Moreover, plug-relations allow the insertion of domain-specific syntagms which ihnerit the "semantics" of their domain-independent head: for instance, the *accettazione delle prove* (*evidence acceptance*), *accettazione della testimonianza* (*witness acceptance),* of the legal domain are linked, trough a plug-hyponymy relation, to the synset *accettazione* (*acceptance*) of the IWN lexicon, by means of which is also linked to the Top-Ontology shared by all the Euro-WordNet databases.

As a source of *metadata for content description,* we need a standard of metadata based on the ontological nature of the entities of the legal domain: within *jur*-WN, an ongoing effort is dedicated to the creation of an ontological level [5]: from the 1500 synsets structured so far, the higher terms/concepts (about 40) have been organised selecting concepts that, acquiring a specific meaning in the legal domain and roughly matching the classical partitions of legal

theory[8], are organised in a *legal core ontology* [8], that takes into account both the new upper levels (DOLCE) [4], and the proposal in the field of legal ontologies [8,15]. For a detailed description of the results for the ontological level, cf. [5].

5 Method of Development of the Semantic Network

In the construction of Jur-WN the "citizens' perspective" was taken into account and a "bottom-up" approach from existing linguistic/terminological resources was followed, selecting as starting points the most frequent terms in user queries of the major legal information retrieval systems.[9] We have used:

- For identification of the relevant terms: the query strings of the Progetto N.I.R. and those of ITALGIURE; the lists of terms linked by AND in the queries provide about 13.000 syntagms; the lists of terms linked by OR in the queries provide the analogical chain and the identification of synonyms.
- For definition of the principal technical concepts: handbooks, dictionaries, legal encyclopedias, etc., [3,2,6,1,10,11] and the L.L.I. containing historical archive of Italian legislative language [18].
- For determination of the syntagms relative to the principal lemmas: the syntagms extrapolated by the ITALGIURE Information Service.

Each sense of the basic terms is then considered as a possible "root" of a sub-hierarchy of terms and syntagms. The general method, in part conducted using automated procedures, considers the syntagms as hyponyms every time their "head" is identical to that of the "basic terms." For example, we identify two different senses of provvedimento (ruling); that is, as public authority act and as disciplinary measure. Nine relative hyponyms are attached to sense 1 (e.g., provvedimento amministrativo -administrative ruling-, and provvedimento legislativo -legislative ruling-) while to sense 2 are linked five terms (e.g., ingiunzione -injunction-, sanzione -sanction-, arresto -arrest- and detenzione -detenction-), which are semantically more specific even if lexically different. Often, the syntagms are considered more interesting if they are linked to basic terms by different semantic relations; for example, verbale d'udienza (trial transcript) is linked to udienza (trial) as 'role-instrument' and to verbale (transcript) as hyponym. Where possible, synonym variants were also included. By the end of this phase, the terms collected are about 1500. The still ongoing phase consists of connecting Jur-WN with IWN and with the ILI (Inter-Lingual Index) in order to integrate the synsets with the networks of the Italian and the other European wordnets.

5.1 Polisemy Handling

Polysemy arises in legal terms both in relation to common language and within the specific context. For example, at legal level, the Italian term *canone* can refer to a payment (in money or goods) or to a legal norm of universal character. *Alimento* considered in the singular is "nutriment" while in the plural is a compulsory payment in the field of divorce

[8] Concepts as *licenza* (*license*), *autorizzazione* (*authorisation*), and *delega* (*delegation*).

[9] We will also evaluate the coverage of the synsets labelled with "law" in MultiWordNet

(*alimony*). The WordNet model permits handling multiple senses in an explicit manner and this allows us to establish conceptual correspondences among terms in different languages. It is especially effi cacious in the legal domain: in law we do not speak of the translation of a legislative text but rather of its multilingual versions. The issue concerning multilingual versions of legal texts is crucial in European Community, where a dual approach is taken: the semantic relations established *a priori* on a conceptual nucleus are integrated with the context comparison on which the Eurodicautom translator is based; for example, the term *prescrizione* corresponds to at least six English terms: *statute of limitations*, *requirement*, *inscription* etc..

Prescrizione1	**Prescrizione 2**	**Prescrizione 3**
synonym: norma, regola (norm, rule, prescription)		
has-hyperonym: diritto (law) *has-hyponym*: prescrizione medica	*has-hyperonym*: fatto giuridico (legal fact) *has-hyponym*: prescrizione speciale, prescrizione ordinaria *cause*: acquisition	*has-hyperonym*: Fatto giuridico (legal fact) *has-hyponym*: prescrizione della pena, prescrizione del reato *cause*: expiration *involved:* termini di prescrizione
equal to: **requirement**	*equal to*: **prescription**	*equal to*: **prescription of claims, limitation of action**

In the above example, we see that word sense discrimination takes into account the distinctions among common and technical meanings (between sense 1, 2 and 3), and among legal institutions (between senses 2 and 3), as well as the confusion between cause (*passage of time*) and effect (*extinction/acquisition*) and between *lapse of time* and *final term*. In other words, we need to manage "semantic overlapping" with more sophisticated linguistic and representational devices, devices that permit us to make distinctions concerning the ontological nature of the concepts. Terminological domains seem to offer a profi table test of the relations between ontology and lexicon: "it is possible that a lexicon with a semantic hierarchy might serve as the basis for a useful ontology, and an ontology may serve as a grounding for a lexicon. This may be so in particular in technical domains, in which vocabulary and ontology are more closely tied than in more-general domains." [7]

6 Future Work

The *jur*-IWN database is still under development: we expect to reach a satisfying coverage of the basic legal contents trough the defi nition of about 3000 synsets. The enrichment of the lexical database will probably act as a test of the ontological level, and allow refi nement and completion of the work done. The European Commission has recently approved, under the E-Content Program, the Project Lois (*Lexical Ontologies for Legal Information Sharing*), aimed at the localization of WordNets for legal domain to Italian, English, German, Czech, Portuguese and Dutch, in order to allow cross-lingual retrieval across different national collection of laws. Furthermore, it will enable cross-lingual access to legislative corpora by inexperienced users and better retrieval by experienced users.

References

1. De Mauro T., *Il Grande Dizionario italiano dell'uso*, UTET, Torino, Italy (2000).
2. *Enciclopedia del diritto*, Giuffrè, Varese, Italy (1989).
3. *Enciclopedia giuridica*,Treccani, Roma, Italy (1995).
4. Gangemi A., Guarino N., Masolo C., Oltramari, A., Schneider L., *Sweetening Ontologies with DOLCE*. in *Proceedings of EKAW 2002*, Siguenza, Spain (2002) 166–178.
5. Gangemi A., Sagri M. T., Tiscornia D., Metadata for Content Description in Legal Information, Workshop Legal Ontologies, ICAIL2003, Edinburgh. In press for *Artifi cial Intelligence and Law Journal*, Kluwer.
6. *Grande Dizionario enciclopedico del diritto*, Fratelli Fabbri Editore, Milano, Italy.
7. Hirst G., *Ontology and the Lexicon*. In Staab, Steffen and Studer, Rudi (eds) Handbook on Ontologies in Information Systems, Berlin: Springer, 2003.
8. `http://wonderweb.semanticweb.org/deliverables/D17.shtml`.
9. *Il Dizionario della lingua Italiana*, Garzanti, Milano, Italy (2002).
10. *Il Nuovo Zingarelli, Vocabolario della lingua italiana*, Zanichelli Ed. Milano, Italy (2002).
11. Miller, G., Beckwith R., Fellbaum C., Gross D., Miller K. J., *Introduction to WordNet: An On-line Lexical Database*. In International Journal of Lexicography, Vol.3, No.4, (1990) 235–244.
12. Report on *"Il progetto Norme in rete"*, Italy (2000) (`http://www.normeinrete.it/documenti`).
13. Roventini A., Alonge A., Bertagna F., Calzolari N., Girardi C., Magnini B., Marinelli R., Speranza M., Zampolli A., *ItalWordNet: Building a Large Semantic Database for the Automatic Treatment of Italian*, in "Linguistica Computazionale", Istituti Editoriali e Poligrafici Internazionali, Pisa-Roma, ISSN (2003).
14. Sagri M. T., *Progetto per lo sviluppo di una rete lessicale giuridica on line attraverso la specializzazione di ItalWornet*. In *Informatica e Diritto*, ESI, Napoli, (2003).
15. Visser P., Bench Capon T., *Ontologies in the Design of Legal Knowledge Systems, towards a Library of Legal Domain Ontologies*, in Proceedings of *Jurix 99*, Leuven, Belgique (1999).
16. Vossen P. (ed.), *EuroWordNet General Document*, 1999. `http://www.hum.uva.nl/~ewn`.
17. `http://www.ittig.cnr.it/banche/LLI/`.
18. `http://www.ittig.cnr.it/organizzazione/personale/biagioli/normeinrete`.

WordNet Has No 'Recycle Bin'

B. A. Sharada and P. M. Girish

Central Institute of Indian Languages
Manasagangotri, Mysore-570 006. India
Email: sharada@ciil.stpmy.soft.net, drsharada@sancharnet.in,
pm_girish@rediffmail.com

Abstract. This paper is conceived and prepared to provide an overview of the compound words in the WordNet, the miracle lexicon of the new millennium. Indeed meanings are not expressed by single words only such as noun, verb, etc., but also languages do have many ways to express content and the concept. Compound words are one among them. Wide range of words and expressions are included in the WordNet. They express a clear view on the existence of concepts in language and culture. After a keen verification, it is found that, some very frequent compound words are not included in the WordNet available online. This paper lists out some such frequent compound words in English. As far as WordNet is concerned – this study is more an application oriented than architecture. Algorithms followed in the development of Subject Heading list are suggested.

1 Introduction

A compound word is a combination of two or more words used to express a single concept. In English, words, particularly adjectives and nouns are combined to form compound words in a variety of ways. Two words will be joined together by a hyphen "fi re-fly" and then joined as one word "fi refly" [5]. Meys W J states that "Functionally, compounding is clearly a linguistic economy-mechanism allowing one to express in a concise way something which would otherwise have to be rendered by means of an – often much more elaborate – phrase" [12]. Many studies have been undertaken in evolving the theories of combining two or more words by Aronoff [1], Chomsky [4], Bauer [2], Bresnan [3], Marantz [11], Williams [16], Lieber [9], Roeper and Siegel [15],etc.

The combinations may be among two nouns, an adjective and a noun, a noun and a verb, etc. such as:

N N	postman	N – Noun
N A	color-blind	V – Verb
A N	high school	A – Adjective
A A	super-fi ne	P – Preposition
P N	under wears	
V N	pound-rice	
V A	diehard	
N V	spoon feed	
A V	deep-fry	
P V	incoming	
V V	drop-kick	
P P	within	

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 311–319.

In the above combinations, VA,VN and PA are predicted to be a rare possibility [10]. But few exceptions could be found for VN in the context of Indian English. For example: boiled rice.

2 Need for the Study

The need for the present study arose while doing linguistic analysis of some texts relating to language learning and information retrieval applications. The powerful online database "WordNet" was checked and it was found that many terms that we call compound words could not be located in the WordNet. Hence these terms were separately listed so that they could be included in the WordNet and make it more comprehensive.

3 Source

The words are collected from the articles that appeared in periodicals, newspapers and other mass media published in India. In order to make the study more wide some more words were collected and checked for which, intuitive knowledge was one of the criteria for the data collection.

4 Compound Words and Analysis

The treatment of compound words in WordNet was very insignifi cant in its earlier version 1.6. Some of the compound words got entered as one word in the later version. That is, the orthographic representation of a compound word will be entered as one term without giving any space or hyphen. The word "compound" itself has become a part of such words like – compound fraction, compound fracture, compound interest, compound word, compound eye, etc. When the search word "compound word" was entered in the WordNet for different senses, 'Sorry, no matches found.' was displayed.

There are three forms of compound words [5]:

a. Closed form: words joined together such as – keyword, textbook, lineup, newspaper, etc.
b. Hyphenated form: Words joined with a hyphen such as – World-wide, Indo-Aryan, Mother-in-law, brief-case, etc.
c. Open form: neither of the above such as – Compound word, Preview theater, Match box, etc.

All the three forms mentioned above are present in WordNet.

For closed form example – wildfi re, mailman, manhood, etc. That is, compound words are without hyphen and space (pre-nominal entered as prenominal). In such a case, words listed in the present study also could be treated in the similar way.

With regard to hyphenation, WordNet has stated, "the hyphenation presents special diffi culties when searching WordNet" [13]. But in the recent version [7] some of the compound words have hyphen in the middle [Cross-country].

In case of Open form, space is considered as a delimiter in WordNet [13]. Example: sky blue, white collar, etc.

In pursuance of the dictionary of compound words in the search-engine, Dictionary: compound [7] was located (updated up to July 23^{rd} 2003). This Hyper dictionary has English dictionary, Dream dictionary, domain specific dictionaries such as Computer and Medical and a thesaurus. The definition of the word along with its grammatical category in brackets is provided with link to each and every word used in the definition. Synonyms and 'See Also' entries are followed. Here also each word has a hyperlink.

The compound words are listed in **Appendix 1**, which were tested in the WordNet, as they are. Among the 180 compound words, 50 words that are in italics were located in the WordNet. Some of the usages of the compound words that were not found in the WordNet can be seen in the following ways:

1. In some cases, though the affixes such as 'co-', 'sub', 'super', 'pre', 'hood', etc., have semantic value, they cannot function as independent words, in their affix-meanings [12].
2. The semantic elements of compound words are different from what the words actually represent as primary meaning. A specific meaning is obtained only when they are used together. In this case, both semantics and pragmatics have wide role to play in dissecting the meaning.

   For example:

Operation flood	- Use and production of milk products in a large quantity.
Collective unconscious	- Is a Freudian terminology to express a sort of socio-mental attitude.
Fall gay	- A person who is punished for the wrong doing of another person.
Recycle bin	- To treat a computer file that has already been used so that it can be used again. It is a component in all the computers.
Lion hearted	- A person having hard nature

3. Some compound words have the thematic or connotative meaning which is completely different from the primary meaning. In the initial stages, it will have limitations in its frequency of usages.

   Limitation may be among – age, gender, profession and other social variables such as religion, education, etc.

   For example, the thematic or connotative meaning for:

Chief minister	- A person who takes a decision in a family
Central Government	- Parents
Tree cover	- Fresh look
Snake gourd	- Very thin person

4. It is a known fact that language is culture bound. So various culture specific words can also be seen.

   For example:

Auto rickshaw	- Auto rickshaw is a three-wheeled vehicle and an economic variety of transportation.
Mid day meal	- Mid day meal refers to the meal that is offered in the school for children free of cost to promote education in economically backward community. Though there is a word in English as 'Lunch' it is not used in this context for differentiating.
Regional language	- Regional language is the language that is in currency in a particular state or a part in the union and in totality of a region.
Panchayat raj	- Village administration

Like wise, terms like Snow-clearing may be in currency in the place where snow fall is a routine matter.

Let us look at the compound words such as: gang shooting, breast-feeding and food poisoning. Although shoot, feed and poison are typically used as transitive verbs, the meanings are not compatible with interpretation such as "to shoot gangs", "to feed breast" and "to poison food". Rather, gang shooting is a shooting incident somehow related to gang activities, breast-feeding is a way to feed babies, and food poisoning is a case of illness caused by unsanitary food [14]. The head word usually at the right side of the compound word gives clue to the description of the meaning. Though the latter two are added in the hyper Dictionary few terms like 'Gang shooting' are not found in the WordNet. The conventional meanings of some of the other terms are mentioned in **Appendix 2.**

5 Conclusion

As Lieber, Rochelle states that, a major goal of current linguistic research is to construct a theory of the lexicon which allows us to characterize the notion of possible word in a simple manner with a minimum of theoretical machinery. Such a theory would ideally predict the possibility of certain sorts of inflected or derived forms, compound, and reduplicated words, while ruling out others [10]. WordNet has mentioned in its third objective that "meanings are not just expressed by nouns and verbs or single words. Language uses a variety of ways to express content ..." [13]. In addition to this, WordNet has improved much within a span of two years. In 2002, WordNet hardly included compound words. It may be recalled here that in the GWN 2002 conference held at the Central Institute of Indian Languages, Mysore, India, it was discussed in the concluding session to include compound words in its lexicon. But now a hyper dictionary is available on the net and that is a tremendous development in WordNet.

For some words in general category and domain specifi c compound words, it is suggested that the algorithm followed in constructing List of Subject Headings (SH) could be followed. SH is a part of Indexing Language and is sharp and equal to summarized text. In SH the importance is given only to the concepts and not to the structure words. If the SH contains two words it will be the combination of an Adjective and a Noun. This order will be inverted to give importance to the Noun. For Example: 'pumping machinery' will be rendered as 'machinery, pumping'. Controlled vocabulary is used in forming the concepts [8]. A

controlled vocabulary contains a unique term for each meaning. Also this may not hold well in all compound words.

This study shows that there is a great potential for WordNet to deal with compound words appearing not only in different grammatical categories but also from all disciplines including interdisciplinary and multidisciplinary research.

References

1. Aronoff, M.: Word Formation in Generative Grammar. Linguistic Inquiry Monograph I. MIT Press, Cambridge, Mass (1976).
2. Bauer, Laurie.: English Word-formation. Cambridge University Press, Cambridge (1984).
3. Bresnan, J.: Passivization: Part II of a theory of Lexical Rules and Representations. MIT Press, Cambridge, Mass (1980).
4. Chomsky, N.: Lectures on Government and Binding, Dordrecht, Foris (1981).
5. Compound Words. `http://webster.commnet.edu/grammar/compounds.htm` (2003).
6. Fairbairn, G. J and Winch. C.: Reading, Writing and Reasoning: A guide for all students. SRHE and Open University Press, Buckingham (1991).
7. Hyperdictionary. `http://www.hyperdictionary.com/dictionary/compound` (2003).
8. LCSH.: Library of Congress Subject Headings. 17th Edition. Library of Congress, Washington D. C. (1993).
9. Lieber, Rochelle: On the organization of the lexicon. Doctoral dissertation. MIT Press, Cambridge, Mass (1980).
10. Lieber, Rochelle: Argument linking and compounds.Linguistic Inquiry,14/2 (1983) pp. 251–285.
11. Marantz, A.: On the nature of grammatical relations,Doctoral dissertation, MIT Press, Cambridge, Mass (1981).
12. Meys, W. J.: Compound Adjectives in English and the Ideal Speaker-listener. North – Holland Amsterdam (1975).
13. MORPHY(7WN): `http://www.cogsci.princeton.edu/~wn/man/morphy.7WN.html`.
14. Oshita, Hiroyuki.: A view from Suffixation and A-structure alteration. In: Booig, G., Marle, J. V. (eds): Yearbook of morphology 1994. Kluwer Academic Publishers, Dordrecht (1995) 179–205.
15. Roeper, T. and Siegel, M. E. A.: A Lexical Transformation for Verbal Compounds. Linguistic Inquiry, 9 (1978) 199–260.
16. Williams, E.: Argument Structure and Morphology. University of Massachusetts at Amherst (1980).

Appendix 1

Table 1. List of Compound Words(Words found in wordNet are in Italics)

Abundant promise	*Bill Collector*	Body spray
Alma mater	Bitter gourd	Boiled rice
Assistant master	Black lash	Branch-brown
Auto rickshaw	Black leg	*Breathtaking*
Bad shot	Black Master	*Broad sheet*
Benefi t show	Black money	Cable Network
Big bull	*Blue collar*	*Cell phone*

Central government
Chain Smoker
Cheque leaf
Chief Minister
Closed chapter
Co brother
Co-editor
Co sister
Collective-unconscious
Color-blind
Community-hall
Compound eye
Compound fraction
Compound fracture
Compound interest
Compound word
Contact programme
Cross border
Cross-country race
Door-leveler
Draw-sheet
Dry-clean
Dry ginger
Dying patient
E-Magazine
Earmarked
Eco feminism
Eco-linguistics
Evergreen-hits
Ever last
Fall gay
Fan mail
Fat cat
Fire-Fighter
Forest cover
Fresh-smell
Gang shooting
Giant killer
Girl-crazy
Glass palace
Golden opportunity
Green Rebellion
Green-crazy
Green signal
Group music
Hand-made
Handwriting
Hanging cot
Heart-breaking
Help line
Hercules task
Hidden agenda
Hidden cost
Hollywood
Home page
Hot drinks
House-top
Ice cream
Ill ommened
Inner politics
Jackpot
Kingmaker
Knock-out
Land mark
Leech gathers
Left-branching
Letter-writing
Long sight
Lower-house
Magic world
Mailman
Mail shot (Advertisement post)
Manhood
Many-sided
Mega-hit
Meta-analytical
Mid day
Mid day meal
Mid noon
Mixed Language
Morpho-thematic
New hand
North Indian
Off shot
One act play
Over whelm
Open book
Operation flood
Out look
Over-ground
Over whelm
Own house
Painstaking
Pan fried
Panchayat raj
Play act
Post-modify
Pound rice
Power delivery
Pre press
Press-button
Pressroom
Proof-read
Provident fund
Recycle bin
Re-do
Red army
Red carpet welcome
Red street
Red tape
Regional language
Right-branching
Rented house
Rough note
Rough tough
Scorching sun
Seafood
Search-box
Search-engine
Search term
See off
Self-respect
Short-circuit
Short sight
Sign-post
Silver screen
Sister concern
Sky blue
Slow dry
Snake gourd
Snow-clearing
Soft drinks
South Indian
Spider man

Spoon-feed
Stand-by
Stress-pattern
Sub-section
Sub urban
Super-hit
Superimpose
Tailor-made
Talk show
Teacher aspirant
Tell-tale
Test drive
Total starvation
Tree cover
Tree-diagram
Twelfth hour
Upper-house
Vacuum cleaner
Visiting time
Water-resistant
White collar
White money
White rebellion
Wildfire
Word formation
Yellow card

Appendix 2

Table 2. Meanings

Abundant promise	Excellent, great in number or quantity
Assistant master	Designation of teacher in a public school
Big-bull	A person who is important and highly influential
Black-log	A person who continues to work when his/her fellow workers are on strike: cheater: one who betrays his friend
Black-money	Money earned by illegal means
Body-spray	A spray used for body freshness: A pleasing personality
Boiled rice	A variety of rice where the paddy grains are boiled before making rice out of it
Branch down	Ruin: quite arrogant
Central government	Government of a country having a number of states and its governments: parent
Chief minister	A chief among state ministers: diplomat: a person who takes a decision in the family
Closed chapter	Broken friendship or relationship to a person or an establishment
Co-brother	Cousin brother
Co-sister	Cousin sister
Community-hall	Hall for a group of people of the same race.
Contact-program	A program for helping teaching program in person to students getting education through correspondence course
Cooked story	Gossip
Cross border	Frontier: keeping rivalry: annoyed relationship
Door-leveler	Give exposure to somebody/someone
Draw sheet	Lucky enough
Dying patient	A person who deserves sympathy
Fan mail	Letters from fans to the persons they admire.
Forest cover	Large area of land thickly covered with trees
Fresh smell	Innovative venture
Giant killer	Person who defeats another one stronger than him(Sports): Win over an unusually large person
Glass palace	Illusion

Golden opportunity	Most favorable situation
Green rebellion	Agricultural progress
Green signal	Sanction
Group music	Group song: Unique demand: uniform decision
Hand – made	Not so professional
Hanging cot	Alter position
Heart breaking	Shocking
Hercules task	Most diffi cult work
Hidden agenda	Mysterious political plans
Hidden cost	Black market price
House top	A parliament section: super
Inner politics	Under current play of an issue
Leech gather	Traditional doctor
Left over	Food remaining at the end of a meal
Left branching	Marxian terminology to denote progressive development in accordance with their theoretical applications
Magic world	Unreal world
Mail shot	An advertisement post
Mid day	Afternoon
Mid noon	Peak at the noon
Mixed language	Mixing two or more different language
New hand	New cover
North Indian	An Indian cultural as well as geographical sphere
One act play	No twist and turn
Open book	Plain and clean: clean image of a person
Over – ground	Unreal
Own house	Permanent place for living
Pound rice	A variety of rice where the paddy is soaked and pound to get a flat variety of rice
Press room	News room
Proof read	To read and correct a piece of written or printed work before publication.: be cleared before actions
Recycle bin	To treat a computer fi le that has already been used so that it can be used again
Re-do	To do again differently: to place a thing as it is again
Red army	Conspiracy wing
Red carpet welcome	Warm welcome: receive somebody with an open heart
Red street	Anti social place where prostitutes live together and treat customers
Right – branching	Proper development
Rented house	Not permanent place for living
Rough note	Not a fare copy: not justifi ed
Scorching sun	Doing hard work: great diffi culty
See off	Farewell
Short sight	A person who does not have future plans
Sister concern	Branch of an institution

Snake guard	A kind of vegetable: very thin person
Snow clearing	Route get cleared
South India	A cultural and geographic sphere in India
Stress pattern	Accent
Sub-urban	Partially urban
Tell- tale	Gossip on cinema actresses:
Total starvation	Too much of suffering
Tree cover	Greenish plants: fresh look
Twelfth hour	Last moment
Upper house	A division in an assembly/parliament
Visiting time	See a person in a proper time
White money	Authorized currency
White rebellion	A good progress in milk products
Word formation	New creation of the word
Yellow card	Punishment: convict.

Automatic Word Sense Clustering Using Collocation for Sense Adaptation

Sa-Im Shin and Key-Sun Choi

KORTERM, KAIST
373-1 Guseong-dong, Yuseong-gu, Daejeon, Republic of Korea
Email: miror@world.kaist.ac.kr, kschoi@world.kaist.ac.kr

Abstract. A specific sense of a word can be determined by collocation of the words gathered from the large corpus that includes context patterns. However, homonym collocation often causes semantic ambiguity. Therefore, the results extracted from corpus should be classified according to every meaning of a word in order to ensure correct collocation. In this paper, K-means clustering is used to solve this problem. This paper reports collocation conditions as well as normalized algorithms actually adopted to address this problem. As a result of applying the proposed method to selected homonyms, the optimal number of semantic clusters showed similarity to those in the dictionary. This approach can disambiguate the sense of homonyms optimally using extracted texts, thus resolving the ambiguity of homonyms arising from collocation.

1 Introduction

In a wide sense, collocation is a pattern of words that coexist in the fixed window or in the same sentence. According to the Yahoo monolingual dictionary, a Korean word *shinbu* has five senses: (1) believable or unbelievable work; (2) a certificate in old Korea – Chosun; (3) a Catholic priest; (4) an amulet; (5) a bride. So, the collocation with the word *shinbu* results in ambiguous context information due to the five senses.?? Through the collocations extracted from the corpus, we can seize some facts concerning the collocation. For example, 'Buddhist priest', 'Africa', 'Braman', 'discipline', and 'appointment', etc. are related to the third sense 'Catholic priest', while 'couple', 'beautiful', 'match', etc. are collocations for the fifth sense 'bride'.

This paper is organized as follows. Firstly, we introduce the sense clustering model of collocation and discuss how to decide the optimal number of cluster. Secondly, we suggest the similarity measure in terms of validity. Finally, we observe and discuss on the experimental results comparing them with other researches.

2 Representation of Collocation

The words in the collocation also have their collocations. A target word for collocation is called the 'central word', and a word in a collocation is referred to as the 'contextual word'. Upon the assumption that there are $w$words placed in the right and left of the central word x, contextual words $x_i^{\pm j}$ for the central word x are represented as follows:

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 320–325.

$\langle x_i^{-w}, \ldots, x_i^{-1}, x, x_i^{+1}, \ldots, x_i^{+w} \rangle$. If collocation patterns between contextual words are similar, it means that the contextual words are used in a similar context – where used and interrelated in same sense of the central word – in the sentence. If contextual words are clustered according to the similarity in collocations, contextual words for polysemous central words can be classifi ed according to the senses of the central words.

The following is a mathematical representation used in this paper. A collocation of the central word x, window size w and corpus c is expressed with function $f : V \times N \times C \rightarrow 2^{C/V}$. In this formula, V means a set of vocabulary, N is the size of the contextual window that is an integer, and C means a set of corpus. In this paper, vocabulary refers to all content words in the corpus. Function f shows all collocations. C/V means that C is limited to V as well as that all vocabularies are selected from a given corpus and $2^{C/V}$ is all sets of C/V. In the equation (1), the frequency of x is m in c. We can also express $m = |c/x|$. The window size of a collocation is $2w + 1$.

$$f(x, w, c) = \left\{ \begin{array}{c} \langle x_1^{-w}, \ldots, x_1^{-1}, x, x_1^{+1}, \ldots, x_1^{+w} \rangle \\ \ldots \\ \langle x_m^{-w}, \ldots, x_m^{-1}, x, x_m^{+1}, \ldots, x_m^{+w} \rangle \end{array} \right\} \tag{1}$$

$g(x) = \{(x, i), i \in I_x\}$ is a word sense assignment function that gives the word senses numbered i of the word x. I_x is the word sense indexing function of x that gives an index to each sense of the word x. All contextual words $x_i^{\pm j}$ of a central word x have their own contextual words in their collocation, and they also have multiple senses. This problem is expressed by the combination of g and f as follows:

$$g \circ f(x, w, c) = \left\{ \begin{array}{c} \langle g(x_1^{-w}), \ldots, g(x_1^{-1}), g(x), g(x_1^{+1}), \ldots, g(x_1^{+w}) \rangle \\ \ldots \\ \langle g(x_m^{-w}), \ldots, g(x_m^{-1}), g(x), g(x_m^{+1}), \ldots, g(x_m^{+w}) \rangle \end{array} \right\} \tag{2}$$

In this paper, the problem is that the collocation of the central word is ordered according to word senses.

3 Sense Clustering Model Using Collocation

This research applies K-means clustering [4] to the automatic clustering of collocations as introduced below. This method classifi es contextual words of the central word into K clusters. For this method, |Ix| refers to K. This approach has been used to extract collocations within a similar context and sense of the central word.

1. Choose K initial cluster centers $z_1(1), z_2(1), \ldots, z_K(1)$, where $k = 1$.
2. At the k-th iterative step, distribute the corpus $\{x\}$ among K clusters by the following condition, where $C_j(k)$ denotes a cluster whose center is $z_j(k)$:

$$x \in C_j(k) \text{ if } sim(x, z_j(k)) > sim(x, z_i(k)), \quad i = 1, 2, \ldots, K; \quad i \neq j \tag{3}$$

1. Compute a series of new cluster centers $z_j(k + 1)$, $j = 1, 2, \ldots, K$ in a way that minimize the sum of similarities from all points in $C_j(k)$ to the new cluster center.

2. If $||z_j(k+1)-z_j(k)||<\alpha$ $(j = 1, 2, \ldots, K)$, then terminate. Otherwise, go to step 2.

Corpus c is represented as $\{x_i\}$ $(1\leq i\leq q)$:q is the number of unique words in c). $(t_{i1}, \ldots, t_{iq})$ is a vector representation of each word x_i according to contextual words as well as co-occurrence frequency. $(t_{i1}, \ldots, t_{iq})/w$ is represented on account of restrictions by the fixed window size w. A cluster C_{aj} means the j-th cluster of the central word x_a. The center z_{aj} of each cluster C_{aj} and the contextual word x_a^i $(i = -w, \ldots, +w)$ for x_a is represented as follows:

$$\vec{z}_{aj} = \left(t_1^{z_{aj}}, \ldots, t_q^{z_{aj}}\right), \quad \vec{x}_a^i = \left(t_1^{x_a^i}, \ldots, t_q^{x_a^i}\right) \tag{4}$$

Each frequency is $t_a^b = \log(P_{ab}/P_a P_b)$ while a and b are targets for collocation. P_{ab} is the probability of co-occurrence between a and b. The cosine similarity between the center z_{aj} and the contextual word x_a^i is expressed as follows:

$$sim\left(\vec{z}_{aj}, \vec{x}_a^i\right) = \sum\nolimits_{m=1,q} t_m^{z_{aj}} t_m^{x_a^j} \Big/ \sqrt{\sum\nolimits_m (t_m^{z_{aj}})^2 \sum\nolimits_m (t_m^{x_a^j})^2} \tag{5}$$

During this process, each contextual word is classified into one cluster having the largest similarity value while repeating this process until the results remain unchanged.

4 Sense Clustering Algorithm: Similarity and Optimal Decision

During the k-th repetition, we update a new center $z_j(k+1)$ using an average of the newly generated j-th cluster $C_j(k)$ based on the center $z_j(k)$. Revised K-means clustering algorithm for sense clustering is addressed by the following subsections.

4.1 Determination of Cluster Centers

In the process of initial clustering, the centers of clusters are determined by randomly selected K contextual words. In each clustering cycle, their centers are adjusted by the average frequency of the contextual words in each cluster. Throughout the repetition process, the center of each cluster is converged to the real center, and similar contextual words are also clustered toward these centers. The equation (6) shows the center $z_j(k+1)$ of the j-th cluster for the next clustering cycle.

$$\vec{z}_j(k+1) = \left(\tfrac{1}{N_j}\textstyle\sum_{i=1,q} t_i^1, \ldots, \tfrac{1}{N_j}\sum_{i=1,q} t_i^q\right) \tag{6}$$

4.2 Termination Conditions for Clustering

The clustering algorithm cycle repeats until the clustering results become stable. It determines whether termination requirements are met. If clustering results meet termination requirements without any change in the clustering results, the clustering process is completed. In this paper,

termination conditions are determined by the rate of variations after each clustering cycle. The following equations indicate the validity in the p-th clustering cycle.

$$\begin{aligned} validity_p &= intra_p \big/ inter_p \\ intra_p &= 1/N \sum_{i=1,K} \sum_{x \in C_i} sim(\vec{x}, \vec{z}_i) \\ inter_p &= \max(sim(\vec{z}_i, \vec{z}_j)) \end{aligned} \tag{7}$$

Internal cohesion $intra_p$ is the average similarity between the center of each cluster and its members, which measures the cohesion of each cluster. External similarity $inter_p$ is the maximum similarity among the centers of clusters and this value is determined by the similarity with the most similar cluster. So, external similarity expresses the discrimination of the clusters. If variations of the validity are lower than that of the threshold, the clustering process is completed. Our experiments show the threshold is 10^{-6}.

5 Experiment and Analysis

5.1 Collocation Normalization

In order to apply correct collocation, it need to remove the noise and trivial collocation. This process is called normalization, and its process is specifi cally provided as follows: (1) remove noise in the tagging or the corpus; (2) remove the words of foreign origin – aimed at avoiding data sparseness; (3) remove one-syllable words; (4) remove statistically unrelated words. According to the Zipf's law, 80% of the words appear only once, while the rest forms 80% of the corpus [1]. Therefore, it can be said that the words with high frequency appear regardless of their semantic features [1]. High frequency words like this are called statistically unrelated words. The words with high frequency can be removed not only by these statistically unrelated words through the sorting of collocations of each contextual word but also by frequently appearing words.

5.2 Number of Clusters

In this research, the number of cluster K is not arbitrarily determined. K refers to the ambiguity of the central word. Therefore, it is important to determine K in reflecting the real ambiguity of the central word. In the process of performing repeated experiments, we selected the optimal number of clusters according to the ambiguity of the central word. The variance of K is determined by statistical analysis of existing dictionaries. The result of analysis of nouns, adjectives, verbs and adverbs in [8] shows that the word with the maximum number of meanings has 41 senses, except for frequently appearing one-syllable words that we already removed.

5.3 Experimental Results

We used KAIST corpus for the experiments [7]. We extracted collocations of about 10 million words from the KAIST corpus. In the experiments, K-means clustering was applied to some of the most famous homonyms. The clustering results are shown in Table 1. The "normalizing rate" means the rate of removing statistically unrelated words in a collocation

Table 1. Clustering Results by Each Normalizing Rate

Normalizing rate / Number of clusters			*A*	*B*	*C*	*D*
word	[8]	[9]				
shinbu	5	2	2	14	9	10
yuhag	5	3	2	2	6	6
buja	4	3	2	3	5	6
sudo	2	2	2	2	4	4
gong'gi	5	5	2	4	4	5

of each contextual word. A indicates the results without normalization, while B, C, and D indicate the results of clustering at a normalizing rate of 0%, 30%, and 50%, respectively.

The results of clustering show that the words are classified in a more specific way than in dictionaries. However, for unnormalized clusterings as in A, the number of clusters is smaller than that shown in other results. That's because correct clustering was interfered by the noise as well as most of frequently appearing words in the collocation. But B, C and D sometimes construct unsuitable clusters. If an initial center is determined by most frequently appearing words, many contextual words are clustered in this cluster. Because most frequently appearing words contain richer context, this center is most likely to match than less frequently appearing centers.

Fluctuations in the number of clusters are found similar to the word senses in dictionaries. It means that K-means clustering proposed in this paper ensures achieving the optimal number of clusters K. Research results show the distribution of practical senses appearing in the corpus.

6 Results and Discussion

This paper is intended to resolve sense ambiguity in collocations through the removal of the noise as well as the application of an automatic clustering method – K-means clustering. Since the clustering method proposed in this paper determines automatically the number of clusters based on the corpus, these numbers guarantee optimal clustering results that have led us to extract practical senses in conducting our research.

However, this research suggests some points to further improve. Firstly, the collocation of contextual words that is used to disambiguate also contains ambiguity. This recursive ambiguity is still reflected on the results. In the second place, the pattern of contextual words in Equation (1) evolves into sorting and clustering problems when contextual words are expressed with their specific senses like Equation (2). This problem is also clearly dependent on how to represent each sense set $g(x)$; the one definitions in dictionaries or logical forms, the other by semantic categories in thesaurus. We need to integrate this research into the research on thesaurus and sense definition. In the third place, there is a problem concerning the optimality of clustering. The method of calculating similarity is affected by clustering results. In this paper, similarity is determined by the relative frequency of occurrence of the collocation of contextual words. Since calculating similarity is accompanied by a judgment which contextual words can be effective in performing clustering, the ensuing experiments

must be performed. Specifi cally, the one is the clustering method using LSI (Latent Semantic Indexing), based on the frequency of occurrence of words, and the other approach to clustering is using the classical *tf-idf* method. This paper applies the latter method indirectly, but more direct application is needed.

Heyer, et al. [5,6] extracted collocations from a large-scaled corpus and constructed an online dictionary called 'Wortschatz'. Nonetheless, this work contains normalizing problems concerning the occurrence pattern of collocations. Lin [3] constructed an English thesaurus using an automatic clustering. But the thesaurus is not only locally and limitedly covered, but also are sense ambiguities in the words excluded. Park [2] constructed a collocation map using collocation and Bayesian network. Pantel [10] discovered senses from English text using CBC and proposed the evaluation measure of this result by comparing with WordNet [12]. This method considered limited contextual words in [10] – frequent nouns over the threshold, but we allow all content words in the same sentence. Ji [11] proposed the *clinique* and clustering methods for collocation clustering. [11] applied collocation clustering in the sense ambiguity in machine translation – selecting translated words and sense ambiguity in compound nouns.

References

1. Zipf G.: Human Behavior and the Principle of Last Effort. Cambridge (1949).
2. Young C. Park and Key-Sun Choi: Automatic Thesaurus Construction Using Bayesian Networks. Information Processing and Management (1996).
3. Lin D.: Automatic Retrieval and Clustering of Similar Words. Coling-ACL (1998).
4. Ray S., Turi R. H.: Determination of Number of Clusters in K-means Clustering and Application in Colour Image Segmentation. In Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques, India (1999).
5. Heyer G., Quasthoff U., Wolff C.: Information Extraction from Text Corpora. IEEE Intelligent Systems and Their Applications (2001).
6. Läuter M., Quasthoff U., Wittig T., Wolff C., Heyer G.: Learning Relations using Collocations. IJCAI (2001).
7. KAIST Corpus: `http://kibs.kaist.ac.kr/` (1999–2003).
8. Hangeul Society, ed.: Urimal Korean Unabridged Dictionary, Eomungag, (1997).
9. Yonsei Dictionary: `http://clid.yonsei.ac.kr:8000/dic/default.htm` (2003).
10. Patrick Pantel and Dekang Lin.: Discovering Word Senses from Text. In Proceedings of ACM Conference on Knowledge Discovery and Data Mining. Edmonton, Canada. (2002).
11. Hyungsuk Ji, Sabine Ploux and Eric Wehrli.: Lexical Knowledge Representation with Contexonyms. In Proceedings of the 9th Machine Translation. (2003).
12. WordNet: `http://www.cogsci.princeton.edu/~wn/`.

WordNet for Lexical Cohesion Analysis

Elke Teich[1] and Peter Fankhauser[2]

[1] Department of English Linguistics
Darmstadt University of Technology, Germany
Email: E.Teich@mx.uni-saarland.de
[2] Fraunhofer IPSI, Darmstadt, Germany
Email: fankhaus@ipsi.fraunhofer.de

Abstract. This paper describes an approach to the analysis of lexical cohesion using WordNet. The approach automatically annotates texts with potential cohesive ties, and supports various thesaurus based and text based search facilities as well as different views on the annotated texts. The purpose is to be able to investigate large amounts of text in order to get a clearer idea to what extent semantic relations are actually used to make texts lexically cohesive and which patterns of lexical cohesion can be detected.

1 Introduction

Using a thesaurus to annotate text with lexical cohesive ties is not a new idea. The original proposal is due to [1], who manually annotated a set of sample texts employing Roget's Thesaurus. With the development of WordNet [2,3], several proposals for automizing this process have been made. For the purpose of detecting central text chunks which can be used for summarization (e.g. [4,5]), this seems to work reasonably well. But how well does an automatic process perform in terms of linguistic-descriptive accuracy? It is well known from the linguistic literature that any two words between which there exists a semantic relation may or may not attract each other so as to form a cohesive tie (cf. [6]). So when do we interpret a semantic relation between two or more words instantiated in text as cohesive or not? First, certain parts-of-speech may be more likely to contract lexical cohesive ties than others, e.g., nouns may be more likely to participate in substantive cohesive ties than verbs. Another motivation may be the type of vocabulary: special purpose vocabulary may be more likely to contract cohesive ties than general vocabulary. Another possible hypothesis is that cohesive patterns differ due to the type of text (register, genre). While repetition generally appears to be the dominant means to establish lexical cohesion, the relative frequency of more complex relations, such as hyponymy or meronymy may depend on the type of text.

In order to investigate such issues, large amounts of data annotated for lexical cohesion are needed. Manual analyses are very time-consuming and may not reach a satisfactory intersubjective agreement. Completely automatic analysis may introduce signifi cant noise due to ambiguity [4]. Thus, in this paper we follow an approach in the middle ground. We use the sense-tagged version of the Brown Corpus, where nouns, verbs, adjectives, and adverbs are manually disambiguated w.r.t. WordNet, and use WordNet to annotate the corpus with potential lexical cohesive ties. (Section 2.1). We also describe facilities for fi ltering candidate ties and for generating different views on the annotated text. (Section 2.2). We discuss the results on an exemplary basis, comparing the automatic annotation with a manual annotation (Section 3). We conclude with a summary and issues for future work (Section 4).

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 326–331.

2 Lexical Cohesion Using WordNet

Lexical cohesion is commonly viewed as the central device for making texts hang together experientially, defining the aboutness of a text (field of discourse) (cf. [6, chapter 6]. Along with reference, ellipsis/substitution and conjunctive relations, lexical cohesion is said to formally realize the semantic coherence of texts, where lexical cohesion typically makes the most substantive contribution (according to [7], around fifty percent of a text's cohesive ties are lexical).

The simplest type of lexical cohesion is *repetition*, either simple string repetition or repetition by means of inflectional and derivational variants of the word contracting a cohesive tie. The more complex types of lexical cohesion rely on the systemic semantic relations between words, which are organized in terms of *sense relations* (cf. [6, 278–282]). Any occurrence of repetition or of relatedness by sense relation can potentially form a cohesive tie.

2.1 Determining Potential Cohesive Ties

Most of the standard sense relations are provided by WordNet, thus it can form a suitable basis for automatic analysis of lexical cohesion. As the corpus, we use the Semantic Concordance Version of the Brown Corpus, which comprises 352 texts (out of 500)[3] Each text is segmented into paragraphs, sentences, and words, which are lemmatized and part-of-speech (*pos*) tagged. For 185 texts, nouns, verbs, adjectives, and adverbs are in addition sense-tagged with respect to WordNet 1.6, i.e., with few exceptions, they can be unambiguously mapped to a synset in WordNet. For the other 167 texts, only verbs are sense-tagged.

Using these mappings, we determine potential cohesive ties as follows. For every sense-tagged word we compute its semantic neighborhood in WordNet, and for each synset in the semantic neighborhood we determine the first subsequent word that maps to the synset.

For the semantic neighborhood we take into account and distinguish between most of the available kinds of relations in WordNet: *synonyms*, *hyponyms*, *hypernyms*, *cohypernyms*, *cohyponyms*, *meronyms*, *holonyms*, *comeronyms*, *coholonyms*, *antonyms*, the *pos* specific relations *alsoSee*, *similarTo* for adjectives, *entails* and *causes* for verbs, and the (rather scarce) relations across parts-of-speech *attribute*, *participleOf*, and *pertainym*. Where appropriate, the relations are defined transitively, for example, hypernyms comprise all direct and indirect hypernyms, and meronyms comprise all direct, indirect, and inherited meronyms. In addition, we also take into account *lexical repetition* (same *pos* and lemma, but not necessarily same synset), and *proper noun repetition*.

For each potential cohesive tie we determine in addition the number of intervening sentences, the distance of the participating words from a root in the WordNet hypernymy hierarchy, as a very rough measure of their specificity, and the length and branching factor of the underlying semantic relation, as very rough measures of its strength.

[3] For the purpose of using off-the-shelf XML processing technology (XPath, XSLT, and an XML database), we have transformed the available SGML version of the corpus to XML; likewise we have transformed the WordNet 1.6 format to XML. Moreover, because some of the texts are compiled from multiple sources, we have enriched them with the bibliographic and segmenting information available from the ICAME version of the corpus. For details see [8].

Due to the excessive computation of transitive closures for the semantic neighborhood of each (distinct) word, this initial step is fairly demanding computationally. In the current implementation (which can be optimized), processing a text with about two thousand words takes about 15 minutes on an average PC. Thus we perform this annotation offline.

2.2 Constraints and Views on Lexical Cohesion

The automatically determined ties typically do not all contribute to lexical cohesion. Which ties are considered cohesive depends, among other things, on the type of text and on the purpose of the cohesion analysis. To facilitate a manual post analysis of the ties we can filter them by simple constraints on *pos*, the specificity of the participating synsets, the kind, distance, and branching factor of the underlying relation, and the number of intervening sentences. This allows, for example, to exclude very generic verbs such as "be", and to focus the analysis on particular relations, such as lexical repetition without synonymy (rare), or hypernyms only.

The remaining ties are combined to lexical chains in two passes. In the first pass, all transitively related (forward-)ties are combined. The resulting chains are not necessarily disjoint, because there may be words w_1, w_2, w_3, where w_1 and w_2 are tied to w_3, but w_1 is not tied to w_2. This results in a fairly complex data structure, which is difficult to reason about and to visualize. Thus in a second pass, all chains that share at least one word are combined into a single chain. The resulting chains are disjoint w.r.t. words, and may optionally be further combined if they meet in a specified number of sentences.

To further analyze lexical cohesion, we have realized three views. In the *text view*, each lexical chain is highlighted with an individual color, in such a way that colors of chains starting in succession are close. This view can give a quick grasp on the overall topic flow in the text to the extent it is represented by lexical cohesion. The *chain view* presents chains as a table with one row for each sentence, and a column for each chain ordered by the number of its words. This view also reflects the topical organization fairly well by grouping the dominant chains closely. Finally, the *tie view* displays for each word all its (direct) cohesive ties together with their properties (kind, distance, etc.). This view is mainly useful for checking the automatically determined ties in detail. In addition, all views provide hyperlinks to the thesaurus for each word in a chain to explore its semantic context. Moreover, some statistics, such as the number of sentence linking to and linked from a sentence, and the relative percentage of ties contributing to a chain are given.

Because filtering ties and combining them to chains can essentially be performed in two (linear) passes over the text, and the chains are rather small (between 2 and 200 words), producing these views takes about 2 seconds for the texts at hand, and thus can be performed online.

3 Discussion of Results

This section discusses the results of the automatic analysis on a sample basis, comparing the automatic analysis with a manual analysis of the first 20 sentences of three texts from the "learned" section of the Brown corpus (texts j32, j33 and j34). For the automatic analysis only nouns and verbs which are at least three relations away from a hypernym root, and adjectives

which are tied to a noun or a verb have been included. Following [9] for the manual analysis, whenever a choice had to be made on which type of relation to base the establishment of a tie, priority was given to repetition.

Table 1 shows the results for the automatic analysis, Table 2 gives the results for the manual analysis (strongest chains only). The chains are represented by the anchor word for simple repetition, and a subset of the participating words for the other types of relations. The number of words and the number of sentences are given in parentheses.

Table 1. Major lexical chains – automatic analysis

j32	j33	j34
form/stem/word (18;11) information/list/ spelling (17;13) dictionary/entry (17;11) text (11;9) store (2;2) storage (2;2)	sentence/subject/ word/... (26;14) stress (14;11)	tone/tonal (8;5) tone system/ consonant system (7;5) linguist (5;5) linguistics (2;2) field (6;6)

Table 2. Major lexical chains – manual analysis

j32	j33	j34
dictionary (16;11) form (14;8) information (11;9) text (11;9) store/storage (4;4)	stress (14;11) complement/predicator/ subject (13;9) sentence (4;4)	linguist/linguistics (13;7) tone/tonal (11;5) field (6,6)

As can be seen from the tables, there is a basic agreement between the automatic and the manual analysis in terms of the strongest chains (e.g., in j33, *stress* builds one of the major chains in both analyses). However, some ties are missed in the automatic analysis, e.g., *store/storage* in j32 or *linguist/linguistics* in j34. Also, there are some differences in the internal make-up of the established chains. For example, in j32, *dictionary* and *information* build major chains in both analyses, but the *information* chain includes a few questionable ties in the automatic analysis, e.g., *list* as an indirect hyponym. Also, the chain around *form* includes *word* and *stem* in the automatic analysis, which would be fine, but then words like *prefix*, *suffix*, *ending* at later stages of the text are not included. The chain built around *complement/predicator/subject* in j33 is separate from the *sentence* chain in the manual analysis, but in the automatic analysis the two are arranged in one chain due to meronymy, thus resulting in the strongest chain for this text in the automatic analysis.

The mismatches arise due to the following reasons. (1) *Missing relations*. Only some relations across parts-of-speech and derivational relations are accounted for in WordNet,

e.g., *linguistic/linguistics* but not *linguist/linguistics* or *store/storage*[4]. (2) *Spurious relations*. Without constraints on the length and/or branching factor of a transitive relation rather questionable ties are established, e.g. *alphabetic character* as a rather remote member of *list*. (3) *Sense proliferation*. In some instances the sense-tagging appears to be overly specific, e.g., in j34, *explanation* as 'a statement that explains' vs. 'a thought that makes sth. comprehensible'. Using synonymy without repetition these senses do not form a tie. On the other hand, with repetition, some questionable ties are established, e.g., for *linguistic* as 'linguistic' vs. 'lingual'. (4) *Compound terms*. In some instances the manual analysis did not agree with the automatic analysis w.r.t. compound terms. E.g. *tonal language* is sense-tagged as a compound term and thus not included in the chain around *tone/tonal*.

Generally, the unconstrained automatic annotation is too greedy, i.e., too many relations are interpreted as ties. Unsatisfactory precision is not so much of a problem, however, because the annotation can be made more restrictive, e.g., by including a list of stop words and/or by determining the appropriate maximal branching and maximal distance for each text to be analyzed. More serious is the problem of not getting all the relevant links, i.e., unsatisfactory recall, usually due to missing relations in the thesaurus. To a certain extent this can be overcome by combining chains that meet in some minimal number of sentences, as a very specific form of collocation.

4 Summary and Envoi

We have presented a method of analyzing lexical cohesion automatically. Even if the results of the automatic analysis do not match one-to-one with a manual cohesion analysis, the automatic analysis is not that far off. Some problems are inherent, others can be remedied (cf. Section 3). Even with an imperfect analysis result, we get a valuable source for the linguistic investigation of lexical cohesion. Knowing that not all words that are semantically related contract cohesive ties, we can set out to determine factors that constrain the deployment of sense relations for achieving cohesion comparing the automatic annotation with a manual analysis. Also, we can give tentative answers to the questions posed in Section 1. Looking at part-of-speech, we can confirm that the strongest chains are established along nouns, and the strongest chains are established along the special purpose vocabulary rather than the general vocabulary[5]. Moreover, although repetition (and synonymy) is the most-used cohesive device, the frequency of other relations taken together (in particular hyponymy and meronymy) about matches that of repetition for the texts at hand.

Future linguistic investigations will be dedicated to questions of this kind on a more principled basis. In order to get a more precise idea of the reliability of the automatic analysis, we are carrying out manual analyses on a principled selection of texts from the corpus (e.g., larger samples covering all registers in the corpus) and compare the results with those of the automatic analysis. Moreover, we plan to investigate cohesion patterns, such as the relative frequency of repetition vs. other types of relations, for the different registers. Ultimately, what we are after are cross-linguistic comparisons, including the comparison of translations with original texts in the same language as the target language [10].

[4] The version we have worked with is WordNet 1.6. WordNet 2.0 has been extended so as to handle such relations more comprehensively.

[5] This also holds when less specific words are taken into account by the automatic analysis.

References

1. Morris, J., Hirst, G.: Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Computational Linguistics **17** (1991) 21–48.
2. Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K. J.: Introduction to WordNet: An on-line lexical database. Journal of Lexicography **3** (1990) 235–244.
3. Fellbaum, C., ed.: WordNet: An electronic lexical database. MIT Press, Cambridge (1998).
4. Barzilay, R., Elhadad, M.: Using lexical chains for text summarization. In: Proceedings of ISTS 97, ACL, Madrid, Spain (1997).
5. Silber, H. G., McCoy, K. F.: Efficient text summarization using lexical chains. In: Proceedings of Intelligent User Interfaces 2000. (2000).
6. Halliday, M., Hasan, R.: Cohesion in English. Longman, London (1976).
7. Hasan, R.: Coherence and cohesive harmony. In Flood, J., ed.: Understanding Reading Comprehension. International Reading Association, Delaware (1984) 181–219.
8. Fankhauser, P., Klement, T.: XML for data warehousing – chances and challenges. In: Proceedings of DaWaK 03, LNCS 2737, Prague, CR, Springer (2003) 1–3.
9. Hoey, M.: Patterns of lexis in text. Oxford University Press, Oxford (1991).
10. Teich, E.: Cross-linguistic variation in system and text. A methodology for the investigation of translations and comparable texts. de Gruyter, Berlin and New York (2003).

Cross-Lingual Validation of Multilingual Wordnets

Dan Tufi ¸s, Radu Ion, Eduard Barbu, and Verginica Barbu

Institute for Artificial Intelligence, 13, Calea 13 Septembrie, 050711,
Bucharest 5, Romania
Email: {tufis, radu, eduard, vergi}@racai.ro

Abstract. Incorporating Wordnet or its monolingual followers in modern NLP-based systems already represents a general trend motivated by numerous reports showing significant improvements in the overall performances of these systems. Multilingual wordnets, such as EuroWordNet or BalkaNet, represent one step further with great promises in the domain of multilingual processing. The paper describes one possible way to check the quality (correctness and completeness) of the interlingual alignments of several wordnets and pinpoints the possible omissions or alignment errors.

1 Introduction

Semantic lexicons are one of the most valuable resources for a plethora of natural language applications. Incorporating Wordnet or its monolingual followers in modern NLP-based systems already represent a general trend motivated by numerous reports showing signifi cant improvements in the overall performances of these systems. Multilingual wordnets, such as EuroWordNet and the ongoing BalkaNet, which adopted the Princeton Wordnet [1] as an interlingual linking device, represent one step further with great promises in the domain of multilingual processing. A general presentation of the BalkaNet project is given in [2]. The detailed presentation of the Romanian wordnet, part of the BalkaNet multilingual lexical ontology, is given in [3,4]. The EuroWordNet is largely described in [5].

Depending on the approach in building the monolingual wordnets included into a multilingual lexical semantic network and on the idiosyncratic properties of each language, the semantic alignment of the wordnets may be pursued and validated in several ways. We distinguish among syntactic and semantic validation methods.

Syntactic validation methods are concerned with checking whether a wordnet is structurally well-formed with respect to a set of rigorously and formally described restrictions such as: all the literals in a synset should have a legal sense identifi er or, no literal with the same sense should appear in more than one synset or, there should be no dangling or unlinked synsets, and many others. Such kinds of errors are easy to spot, although not necessarily very easy to correct (especially when they are due to different granularity of the language resources used to build the wordnets). Semantic validation methods (in this context) rely on the notion of semantic equivalence between the word senses in two or more languages used to express the same concept.

2 Assumptions and the Basic Methodology

One fundamental assumption in the study of language is its compositional semantics. Compositionality is a feature of language by virtue of which the meaning of a sentence is

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 332–340.

a function of the meanings of its constituent parts (going down to the level of the constituent words). With this tarskian approach to meaning, our methodology assumes that the meaning building blocks (lexical items-single or multiple word units) in each language of a parallel text could be automatically paired (at least some of them) and as such, these lexical items should be aligned to closely related concepts at the ILI level. That is to say that if the lexical item W^i_{L1} in the first language is found to be translated in the second language by W^j_{L2}, common intuition says that it is reasonable to expect that at least one synset which the lemma of W^i_{L1} belongs to, and at least one synset which the lemma of W^j_{L2} belongs to, would be aligned to the same interlingual record or to two interlingual records semantically closely related.

As a test-bed, we use the wordnets developed within the BalkaNet European project and the "*Nineteen Eighty-Four*" parallel corpus [6] which currently includes four relevant languages for BalkaNet (with the prospects of extending the corpus to all the BalkaNet languages). This project aims at building, along the lines of EuroWordNet lexical ontology, wordnets for five new Balkan languages (Bulgarian, Greek, Serbian, Romanian and Turkish) and at improving the Czech wordnet developed in the EuroWordNet project. The methodology for semantic validation assumes the following basic steps:

A) given a bitext T_{L1L2} in languages L1 and L2 for which there are aligned wordnets, one extracts the pairs of lexical items that are reciprocal translations: $\{<W^i_{L1}W^j_{L2}>^+\}$;
B) for each lexical alignment of interest, $<W^i_{L1}W^j_{L2}>$, one extracts the synsets in each language that contain the lexical items of the current pair and respectively their ILI projections. For every lexical item recorded in the monolingual wordnets there will result two lists of ILI labels, one for each language, L^1_{ILI} and L^2_{ILI}. Based on the content evaluation of these two lists, several lines of reasoning might be followed highlighting various problems related to: the implementation of one or the other of the two wordnets, the alignment to the ILI; different sense granularity among wordnets; lexical gaps; wrong translation in the bitext, etc.

The first processing step is crucial and its accuracy is essential for the success of the validation method. A recent shared task evaluation (`http://www.cs.unt.edu/~rada/wpt`) of different word aligners, organized on the occasion of the Conference of the NAACL showed that step A) may be solved quite reliably. The best performing word alignment system [7] produced lexicons, relevant for wordnets evaluation, with an aggregated F-measure as high as 84.26%.

The content evaluation of L^1_{ILI} and L^2_{ILI} assumes a definition for the semantic distance between ILI records. Our system uses Siddharth Patwardhan and Ted Pedersen's WordNet-Similarity PERL module, a WN plug-in implementation of the five semantic measures described in [8].

3 Interlingual Validation Based on Parallel Corpus Evidence

If we take the position according to which word senses (language specific) represent language independent meanings, abstracted by ILI records, then the evaluation procedure of wordnets interlingual alignment becomes straightforward: in a parallel text, words which are used to

translate each other should have among their senses at least one pointing to the same ILI or to closely related ILIs. However, both in the EuroWordNet and in BalkaNet the ILI records are not structured, so we need to clarify what "closely related ILI" means. In the context of this research, we assume that the *hierarchy preservation* principle [4] is sound. This principle may be stated as follows:

if in the language L1 two synsets M_1^{L1} and M_2^{L1} are linked by a (transitive) hierarchical relation H, that is $M_1^{L1} H^n M_2^{L1}$ and if M_1^{L1} is aligned to the synset N_1^{L2} and $M_2^{L1}$is aligned to N_2^{L2} of the language L2 then $N_1^{L2} H^m N_2^{L2}$ even if n≠m (chains of the H relation in the two languages could be of different lengths). The difference in lengths could be induced by the existence of meanings in the chain of language L1 which are not lexicalized in language L2.

Under this assumption, we take the *relatedness* of two ILI records R1 and R2 as a measure for the *semantic-distance* between the synsets Syn1 and Syn2 in PWN that correspond to R1 and R2. One should note that every synset is linked (EQ-SYN) to exactly one ILI and that no two different synsets have the same ILI assigned to them. Furthermore, two ILI records R1 and R2 will be considered closely related if *relatedness*(R1, R2)=*semantic-distance* (Syn1, Syn2)≤ k, where k is an empirical threshold, depending on the monolingual wordnets and on the measure used for evaluating semantic distance.

Having a parallel corpus, containing texts in k+1 languages (T, L_1, $L_2 \ldots L_k$) and having monolingual wordnets for all of them, interlinked via an ILI-like structure, let us call the T language as the target language and L_1, $L_2 \ldots L_k$ as source languages. The parallel corpus is encoded as a sequence of *translation units* (TU). A translation unit contains aligned sentences from each language, with tokens tagged and lemmatized as exemplified in Figure 1 (for details on encoding see `http://nl.ijs.si/ME/V2/msd/html/`).

```
<tu id="Ozz.113">
 <seg lang="en">
    <s id="Oen.1.1.24.2"><w lemma="Winston" ana="Np">Winston</w>
        <w lemma="be" ana="Vais3s">was</w>      ... </s>
 </seg>
 <seg lang="ro">
    <s id="Oro.1.2.23.2"><w lemma="Winston" ana="Np">Winston</w>
     <w lemma="fi" ana="Vmii3s">era</w>     ...     </s>
 </seg>
 <seg lang="cs">
    <s id="Ocs.1.1.24.2"><w lemma="Winston" ana="Np">Winston</w>
    <w lemma="se" ana="Px---d--ypn--n">si</w>  ...   </s>
  </seg>
  . . .
</tu>
```

Fig. 1. A partial translation unit from the parallel corpus

We will refer to the wordnet for the target language as T-wordnet and to the one for the language L_i as the i-wordnet. We use the following notations:

T_word = a target word;

T_word_j = the j-th occurrence of the target word;
eq_{ij} = the translation equivalent (TE) in the source language L_i for T_word_j;
EQ = the matrix containing translations of the T_word (k languages, n occurrences):

Table 1. The translation equivalents matrix (EQ matrix)

	Occ #1	Occ #2	…	Occ #n
L_1	eq_{11}	eq_{12}	…	eq_{1n}
L_2	eq_{21}	eq_{22}	…	eq_{2n}
…	…	…	…	…
L_k	eq_{k1}	eq_{k2}	…	eq_{kn}

TU_j = the translation unit containing T_word_j;
EQ_i = a vector, containing the TEs of T_wordin language L_i: (eq_{i1} eq_{i2} … eq_{in})
More often than not the translation equivalents found for different occurrences of the target word are identical and thus identical words could appear in the EQ_i vector. If T_word_j is not translated in the language L_i, then eq_{ij} is represented by the null string. Every non-null element eq_{ij} of the EQ matrix is subsequently replaced with the set of all ILI identifiers that correspond to the senses of the word eq_{ij} as described in the wordnet of the i-language. If this set is named IS_{ij}, we obtain the matrix EQ_ILI which is the same as EQ matrix except that it has an ILI set for every cell (Table 2).

Table 2. The matrix containing the senses for all translation equivalents (EQ_ILI matrix)

	Occ #1	Occ #2	…	Occ #n
L_1	IS_{11} ={ILI_p| ILI_p identifies a synset of eq_{11}}	IS_{12} ={ILI_p| ILI_p identifies a synset of eq_{12}}	…	IS_{1n} = {ILI_p| ILI_p identifies a synset of eq_{1n}}
L_2	IS_{21} ={ILI_p| ILI_p identifies a synset of eq_{21}}	IS_{22} {ILI_p| ILI_p identifies a synset of eq_{22}}	…	IS_{2n} = {ILI_p| ILI_p identifies a synset of eq_{2n}}
…	…	…	…	…
L_k	IS_{k1} ={ILI_p| ILI_p identifies a synset of eq_{k1}}	IS_{k2} {ILI_p| ILI_p identifies a synset of eq_{k2}}	…	IS_{kn} = {ILI_p| ILI_p identifies a synset of eq_{kn}}

If some cells in EQ contain empty strings, then the corresponding cells in EQ_ILI will obviously contain empty sets. Similarly, we have for the T_word the list T_ILI = (ILI_{T1} ILI_{T2}… ILI_{Tq}).

The next step is to define our target data structure. Let us consider a new matrix (see Table 3), called VSA (Validation and Sense Assignment).

with VSA_{ij} = T_ILI ∩ IS_{ij}, if IS_{ij} is non-empty and ⊥ (undefined) otherwise.

The i^{th} line of the VSA matrix provides valuable corpus-based information for the evaluation of the interlingual linking of the the i-wordnet and T-wordnet.

Ideally, computing for each column j the set SA_j (sense assignment) as the intersection ILI_{1j}∩ ILI_{2j}…∩ILI_{kj}one should get at a single ILI identifier: SA_j=($ILI_{T\alpha}$), that is the j^{th} occurrence of the target word was used in all source languages with the same meaning,

Table 3. The VSA matrix

	Occ #1	Occ #2	...	Occ #n
L_1	VSA_{11}	VSA_{12}	...	VSA_{1n}
L_2	VSA_{21}	VSA_{22}	...	VSA_{22}
...	...	...	...	...
L_k	VSA_{k1}	VSA_{k2}	...	VSA_{kn}

represented interlingually by $ILI_{T\alpha}$. If this happened for any T_word, then the WSD problem (at least with the parallel corpora) would not exist. But this does not happen, and there are various reasons for it: the wordnets are partial and (even the PWN) are not perfect, the human translators are not perfect, there are lexical gaps between different languages, automatic extraction of translation equivalents is far from being perfect, etc.

Yet, for cross-lingual validation of interlinked wordnets the analysis of VSAs may offer wordnet developers extremely useful hints on senses and/or synsets missing in their wordnets, wrong ILI mappings of synsets, wrong human translation in the parallel corpus and mistakes in word alignment. Once the wordnets have been validated and corrected accordingly, the WSD (in parallel corpora) should be very simple. There are two ways of exploiting VSAs for validation:

Horizontal validation (HV): the development team of i-wordnet (native speakers of the language L_i with very good command of the target language) will validate their own i-wordnet with respect to the T-wordnet, that is from all VSA matrixes (one for each target word) they would pay attention only to the i-th line (the VSA(L_i) vector).

Vertical validation (VV): for each VSA all SAs will be computed. Empty SAs could be an indication of ILI mapping errors still surviving in one or more wordnets (or could be explained by lexical gaps, wrong translations etc.) and as such, the suspicious wordnet(s) might be re-validated in a focused way. The case of an SA containing more than a single ILI identifi er could be explained by the possibility of having in all i-languages words with similar ambiguity.

We exemplify the two types of validation by considering English as the target language and Romanian and Czech as source languages. At the time of this writing the Romanian wordnet contains 11698 synsets (encoding 23571 literals), all linked to ILI records. The Czech wordnet is twice as large (25240 synsets and 37451 literals).

HV: The case study language is Romanian. For the validation purposes we selected a pool of 733 English common nouns appearing in Orwell's *Nineteen Eighty-Four* (out of 3167), because all their senses were implemented in the Romanian wordnet. There were 4319 occurrences of these words in the English part of our corpus and we built, as described in the previous section, 733 VSA vectors.

Almost half of the 4319 VSA_{ij} in the 733 vectors were empty. According to the procedure discussed in the previous section, when a VSA_{ij} contains an empty set, it means that none of the senses of the word eq_{ij} could be mapped (via ILI) to any of the senses of the target word. Although the analysis is not complete yet, we identifi ed the following main explanations:

1. T_word and eq_{ij} are not related and the error is attributable to the human translator who used a wrong translation for T_word; we spoted only one such error (*darts/damă*) but systematically used four times.

2. T_word and eq_{ij} are not related and they were wrongly extracted as a translation pair by the word alignment program. By inspecting the TU_j it was easy to recognize this case and correct it; although these errors were not related to Wordnet development, and less than 15% of the analysed empty VSA_{ij} cells could be attributed to word-alignment errors, identifying them was beneficial for further development of the word aligner.
3. the right sense is defined for eq_j but it has a wrong ILI identifier (it is wrongly mapped on ILI). By inspecting TU_j and sense glosses for eq_{ij}, the i-wordnet developer may easily identify the wrong mapping and correct it appropriately. This case is very relevant for the wordnet development and we estimate around 20% of the empty VSA_{ij} cells being explained by wrong mappings.
4. the synset linked to the relevant ILI record does not include the literal eq_{ij}, meaning that not all senses of eq_{ij} are defined in the i-wordnet and it happened that one of the missing senses was used in the TU_j. This situation is easy to recognize by a native speaker and the obvious solution is to add the eq_{ij} literal (indexed with the new sense number) to the proper synset. We estimate that this case (incomplete synsets) is responsible for almost 25% of all empty VSAs cells.
5. although none of the senses of T_word and eq_{ij} points to the same ILI identifier, one could identify a sense of T_word linked to ILI_α and a sense of eq_{ij} linked to ILI_β so that ILI_α and ILI_β are closely related. Closely relatedness was considered based on a maximum of two link traversals. This is what we call a *near-miss* interlingual linking. This case was the most frequent (we estimate it to more than 35%). The near-misses might be explained either by the translator's use of a more general or more specific Romanian word for the English word (e.g. because of lexical gaps or stylistic reasons) as in case of *prettiness/frumuseţe*, *bureaucrat/funcţionar*, *dish/farfurie*, *throat/gât*, etc. or by a misguided ILI mapping in the Romanian wordnet (still close enough) such as: *emotion/emoţie*, *hero/erou*, *event/eveniment* and several other real cognates. While translation licenses are inherent, coping with them is very important for the WSD task. The relatedness measure is an effective approach to decide which senses the T_word and eq_{ij} might have. The near-misses due to wordnet builders must be corrected. Most near-misses due to mapping errors show quite a regular pattern: when mapping a Romanian synset, the lexicographer had always as options at least two ILI records characterised by very similar glosses. As expected, looking up the PWN synsets corresponding to these ILI records, more often than not they were located in the same proximity (one hyponym/hypernym or meronym/holonym relation). Without additional information and based on subjective reasoning, lexicographers' introspection was wrong in several cases.

VV: The vertical validation is exemplified for English-Romanian-Czech**.** In order to see the potential of vertical validation procedure, we conducted a very small experiment on Romanian and Czech building the VSA for the T_world *country.*The 20 occurrences of the word *country* were translated in Czech by *země* (13 times), *venkov* (twice), *stát* (twice), *vlast* (twice), and once it was not translated. In Romanian, the occurrences of *country* were translated by the words *ţară* (12 times), *tărâm* (5 times), *stat* (twice) and once it was translated by a pronoun. The distinct triples of non-null mutual translations were the following:

1. <country ¸tară země> occurring eight times;
2. <country stat stát> occurring twice;

3. <country ˛tară vlast > occurring twice;
4. <country ˛tară venkov> occurring twice;
5. <country tărâm země> occurring five times.

Computing SAs for all triples above we obtained complete disambiguation for the first two of them (ten occurrences), all corresponding to the ILI record 171-07034213-n. The disambiguated translations of these 10 occurrences of *country* were:

1') <country:1 ˛tară:1 země:3>;
2') <country:1 stat:1.1a stát:3>.
The remaining triples generated empty SAs. However, they were disambiguated as near-misses as follows:
3') <country:1 ˛tară:1 vlast:1> – vlast:1 is a hyponym of země:3 and <country:1 ˛tară:1 země:3> is uniquely interpretable as 171-07034213-n. The contexts of these occurrences were:"...they betrayed their country..." and "...you betray your country...". This example show a near miss due to a lexical gap: neither English nor Romanian uses a single word for the concept of *own country*, unlike Czech.
4') <country:4 ˛tară:5 venkov:1> – both *country*:4 and *ţară*:5 are linked to the ILI record 171-07121548-n which is closely related to the one corresponding to ILI record 172-07121859-n standing for *venkov*:1. This latter ILI record is lexicalized in English by *countryside*, the first sense of which is a hyponym of *country*:4(rural area).
5') Finally, the third group of reciprocal translations was the most interesting. All the five occurrences were in the context of "...Golden Country..." (the fantasy land Winston Smith, the main character in "*Nineteen Eighty-Four*", was dreaming of). Between English and Romanian the near-miss was disambiguated as (country:5 tărâm:1) corresponding to the ILI record 171-06996512-n. Between English and Czech, the VSA_{ij}(country, země) = (171-07034213-n 171-06771212-n) and as such the near-miss was partially disambiguated as ((country:1 země:3)(country:3 země:6)). Since the distances between country:1 and country:5 or between country:3 and country:5 were beyond our considered threshold, the global near-miss could not be disambiguated. The conclusion we reached was that in the Czech wordnet there should be another sense for *země* (in the same synset with oblast:1, území:2 and prostor:2) in order to license translations as in the example below:
In his waking thoughts he called it the Golden Country/V duchu ji nazýval Zlatá země

4 Conclusions

This preliminary experiment shows that using translation equivalents extracted from a test-bed parallel corpus may precisely pinpoint various problems in the wordnets structuring and interlingual linking. A thorough quantitative and qualitative evaluation will follow the syntactic validations of the BalkaNet wordnets.

Recently the wordnets of the Balkanet project have been remapped on an ILI that corresponds to PWN2.0.

The methodology we discussed in this paper has been implemented in a Java program called *WSDtool*. In the present stage of the project we use it as a multilingual wordnet checker and specialized editor for error correction. Once the wordnets are validated, *WSDtool* can be

used to consistently sense-tag the entire multilingual parallel corpus (hence the name). For the most part, the sense tagging can be accomplished fully automatically; in those cases where it cannot, the human annotator is offered a small set of options from which to choose, thus reducing the likelihood of error. In the Appendix there is a commented snapshot from a horizontal validation session (English-Romanian) with WSDTool.

References

1. Fellbaum, Ch. (Ed.) (1998) WordNet: An Electronic Lexical Database, MIT Press.
2. Stamou, S., Ofhzer K., Pala K., Christoudoulakis D., Cristea D., Tufis D., Koeva S., Totkov G., Dutoit D., Grigoriadou M. (2002): BalkaNet A Multilingual Semantic Network for the Balkan Languages, in Proceedings of the 1st *International Wordnet Conference*, Mysore.
3. Tufiş, D., Cristea, D. (2002): Methodological issues in building the Romanian Wordnet and consistency checks in Balkanet, In Proceedings of *LREC2002 Workshop on Wordnet Structures and Standardisation*, Las Palmas, Spain, May, 35–41.
4. Tufiş, D., Cristea, D.: Probleme metodologice în crearea Wordnet-ului românesc şi teste de consistenţă pentru BalkaNet, în Tufiş, D., F. Gh. Filip (eds.) *Limba Român˘a în Societatea Informa¸tional˘a – Societatea Cunoa¸sterii*, Editura Expert, Academia Român˘a (2002) 139–166.
5. Vossen, P. (Ed.) (1999): *EuroWordNet: a multilingual database with lexical semantic networks for European Languages*, Kluwer Academic Publishers, Dordrecht.
6. Erjavec, T., Lawson A., Romary, L. (eds.) (1998): East Meet West: A Compendium of Multilingual Resources. TELRI-MULTEXT EAST CD-ROM, 1998, ISBN: 3-922641-46-6.
7. Tufiş D., Barbu A. M., Ion R. (2003): A word-alignment system with limited language resources, Proceedings of the *NAACL 2003 Workshop on Building and Using Parallel Texts*; Romanian-English Shared Task, Edmonton, Canada, 36–39 (also available at: `http://www.cs.unt.edu/~rada/wpt/index.html#proceedings/`).
8. Budanitsky, A., Hirst, G. (2001): Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In Proceedings of the *Workshop on WordNet and Other Lexical Resources*, Second meeting of the North American Chapter of the Association for Computational Linguistics, Pittsburgh, June.

Appendix

The snapshot illustrates a horizontal validation (English-Romanian), the selected target word being "*shop*" and its translation equivalents in Romanian being displayed on the right part of the main screen. The first occurrence of "shop" appears in the Ozz.69 translation unit and clicking in the VSA cell corresponding to this occurrence on the **Check** and **Go** buttons several windows are opened:

1. the top most window shows the translation unit Ozz.69 with the translation equivalents highlighted (shops ↔ magazinele).
2. the partial networks in the Princeton Wordnet and Romanian Wordnet with the corresponding synsets as barycenters (right top and bottom left corners of the main window). Next to the barycenters are the entries in the two wordnets: [shop(1), store(1)] ↔ [magazin(1), prăvălie(1)].

The VSA cell exemplified contains one single ILI-record number (ENG171-03661978-n), signifying full disambiguation of the translation pair <shop, magazin>. The single common ILI-record number is pointed by the senses *shop(1)* and *magazin(1)*.

The VSA cell below the one exemplified contains the same ILI-record and everything discussed above holds true.

However, the VSA cell corresponding to the third occurrence of *"shop"* (visible at the bottom left corner of the main window) is empty. This occurrence of the target word was not translated in Romanian aligned sentence.

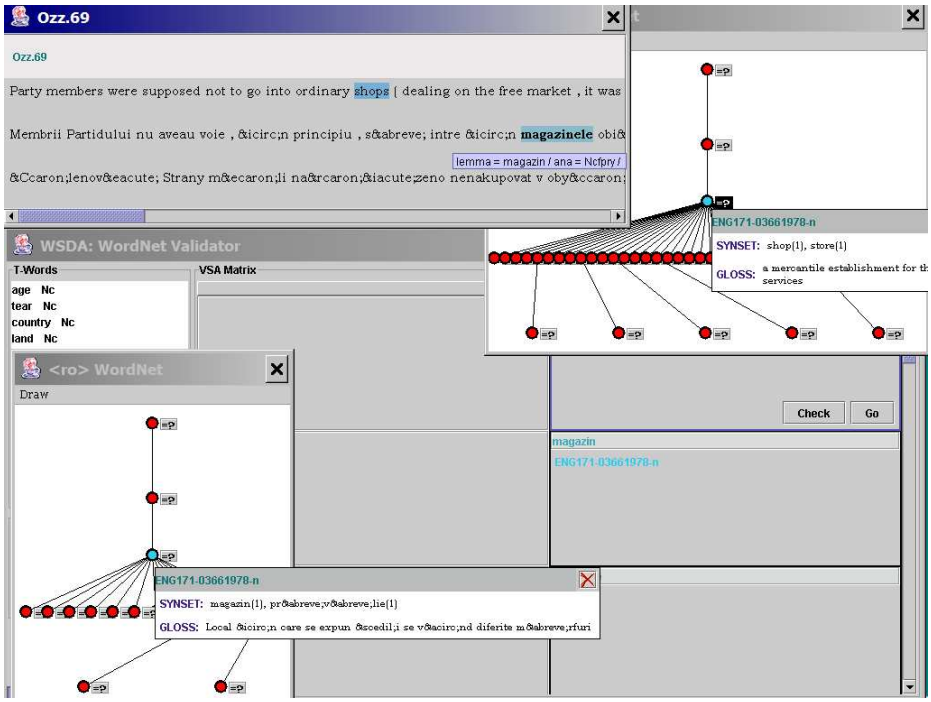


Fig. 2. A snapshot from a WSDTool HV session:
T-word is "shop", L_1 is Romanian, eq_{11} is "magazin" and VSA_{11} is {ENG171-03661978-n}

Roles: One Dead Armadillo on WordNet's Speedway to Ontology

Martin Trautwein and Pierre Grenon

Institute for Formal Ontology and Medical Information Science, University of Leipzig
Email: martin.trautwein@ifomis.uni-leipzig.de,
pierre.grenon@ifomis.uni-leipzig.de

Abstract. We assume that the ontological structure of the common-sense world, and thus of human knowledge about this world, is organized in networks rather than in hierarchies. Thus, using the taxonomies that semantic relations generate in WordNet as the only source for the reconstruction of ontological information must fail at some point. Comparing the ontological structures underlying roles to WordNet representations, we demonstrate that the power of lexical semantics to abstract over contexts distorts the taxonomic order of a conceivable ontology. Approaches trying to adjust the semantics of WordNet relations, in order to reach a higher ontological adequacy, unintentionally produce artifacts deriving from differences between the frequency of contexts, and from metonymy-like reference to ontological relations.

1 Introduction

Although WordNet was designed as a semantic dictionary, many applications have put emphasis on the fact that the semantic content of the conceptual entries of WordNet depict common-sense world knowledge and thus reflect common-sense ontology. [4] notices however that "WordNet is only really serviceable as an ontology if some of its lexical links are interpreted according to a formal semantics that tells us something about (our conceptualization of) 'the world' and not (just) about the language." In this light, the authors cited propose several improvements of the hierarchizing semantic relations in WordNet [3,4] in accordance with constraint-based Formal Ontology. In contrast, we want to put more emphasis on the question of which kinds of semantic links can be readily interpreted semantically as ontological relations and which can not. We show that in fact, some of WordNets semantic relations can locally be regarded as taxonomic, whereas the network as a whole can not or should not be converted one-to-one into a taxonomic ontological framework.

The semantic descriptions of WordNet concepts provides links to two distinct forms of knowledge: common-sense knowledge and semantic knowledge. Thus, a concept involves two aspects of information: (i) information about concrete world contexts and their spatiotemporal structure, and (ii) information about epistemology and grammaticalization, merging (our knowledge about) a set of concrete world contexts by means of abstraction. This second kind of information provides the criteria for associating an entity to a certain concept or lexical entry (classifying criteria).

Only the first aspect provides genuine ontological information. When referring and asserting (i.e. constructing concrete contexts of language usage), however, we merge both

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 341–346.

the knowledge issued from concrete contexts and the abstracted classifying criteria at hand. Moreover, in order to identify referents or to establish truth, we have to match concrete contexts with the relevant classifying criteria carried by our word semantics. Accordingly, both modules of meaning have to interact and this in a manner prone to reveal their contribution to semantic meaning. WordNet's concepts are based on the synonymy of certain word meanings in a certain set of linguistic (discourse) contexts, but in the end the lexical data is detached from context. As a consequence, most parts of the ontological assumptions and intuitions standing behind WordNet are only implicitly represented by the set-up of synsets and by the semantic relations holding between the corresponding relations. These assumptions and intuitions, however, may be regained through the analysis of the classifying criteria encoded in the semantic aspects of a word meaning. In what follows, we apply our hypothesis to the category of ROLE. We compare the WordNet strategy of representing roles to the rich system of ontological structures of reality that roles crystallize.

2 Types vs. Roles: the WordNet Strategy from the Ontological and Linguistic Standpoints

Recent publications (cf. [4,1] discuss the type-role distinction and WordNet's representation of the two corresponding categories. One of the main arguments for a clear-cut distinction between the two categories is that they differ in the way their instances inherit their properties. This distinction is also crucial for applications which use WordNet as a source for common-sense reasoning. WordNet, however, does not distinguish between roles and types, but organizes both in the same taxonomy. In accordance with the ontological approach to properties of [3,2] proposes that roles should not subsume types in the hyponyme taxonomy. A more radical ontological account, however, could even demand that roles and types should not be represented at all within the same taxonomy.

Although an ontological characterization of properties in the spirit of [3] is indispensable for a proper treatment of classifi cations and of the qualitative information they provide, such a theory does not anticipate the way in which individual natural languages grammaticalize types and roles, nor how these grammaticalized forms are applied to language usage. With regard to the synonymy criterion that constitutes the conceptualization of WordNet, it is remarkable that the lexicalizations of types and roles generally do not differ in the way they are used in linguistic contexts since in most contexts, a role expression can replace a type expression and vice versa without changing the referential or truth-conditional value. Ranking immediately behind proper names, both type and role expressions (such as *man*, *teacher*, *speaker*, or *speak*, *declare*, *verbalize*) have a strong identifying potential and thus suffi ciently restrict a class of referents, e.g. in a nominal description (cf. (1)).

Claire had a dispute with James / her friend / a theologian. (1)

Verbs do not differ in this point: the sortal verb *swallow* can be replaced using verbs such as *eat* or *ingest* which express roles of the process referred to (cf. (2)).

Claire swallowed / ate / ingested a fly. (2)

Beyond this, co-referential descriptions often use hypo/hypernymes and role-expressions as sortals in order to bridge the gap between co-referential terms (cf. sample sentences (3)

and (4); co-referring terms are underlined).

(3) *Hannah observed a hedgehog. She picked the animal up.* (via hypernymy)

(4) *She turned quickly*.*Her vigilant reaction saved the armadillo's life.* (via a role)

3 Lexicalization of Roles and Context Frequency

The usage of nouns and verbs in nominal and verbal descriptions demonstrates that, from the perspective of linguistics, the need for differentiating between the treatments of type-, subtype-, supertype-, and role-denoting lexical items is not conspicuous. Nonetheless, since distinguishing types and roles is all the more desirable from the ontologist view, some authors are looking for ways out. [1] observes that some cases of troponymy are heterogeneous, i.e. they do not fit properly in the tropo-taxonomy, e.g. ***swim*** ^{v1} $<$ ***move*** ^{v1} (troponyme) vs. ***swim*** ^{v1} $</$ ***exercise*** ^{v4} (non-troponymy).[1] In view of the expressiveness of WordNet used by NLP applications, however, [1] pleads for including links as those between ***swim*** ^{v1} and ***exercise*** ^{v4} in the lexical DB, and proposes to introduce additional, autonomous semantic relations into WordNet in order to capture the specific relationship between entities and the roles they carry. These relations, *para-hyponymy* for nouns and *para-troponymy* for verbs, omit the necessity condition which, in contrast, holds for the regular hyponymy and troponymy relations. The linguistic tests [1, p. 27f] are shown in (5) to (8).

(5)	***X**'s and other **Y**'s* & ¬(*It's an **X**, but it's not a **Y***)	**X** is a hyponyme of **Y**
(6)	***X**'s and other **Y**'s* & *It's an **X**, but it's not a **Y***	**X** is a para-hyponyme of **Y**
(7)	***X**'ing and other manners of **Y**'ing* & ¬(*It's an **X** event, but it's not a **Y** event*)	**X** is a troponyme of **Y**
(8)	***X**'ing and other manners of **Y**'ing* & *It's an **X** event, but it's not a **Y** event*	**X** is a para-troponyme of **Y**

[1] assumes that for some concepts the 'role' aspect is more important than the 'supertype' aspect, e.g. ***jog***^{v1} tends to be interpreted as a para-troponyme of ***exercise*** ^{v4} rather than a troponyme of ***move*** ^{v1}. This point signifies that the weightiness of the semantic relations is gradual and vague. The links in Fellbaum's examples seem plausible since the sample word meanings occur frequently in contexts. But what about less frequent, less common, or less expectable contexts? Attending a course of survival training, you will probably soon discover that flies can perfectly adopt the role of food. A bottle can serve as a musical instrument. Singing (and, in particular, bad singing) can not only produce sound but may also amuse people. Grinning, finally, may be more than just a sign of amusement but may offend somebody (e.g. a bad singer). So do we also have to include links for these cases?

[1] Words given in bold italics and marked with an superscript symbolize word meanings taken from WordNet v1.7.1. The superscript indicates the entry number of the word sense of the noun (n) or the verb (v) database. Words indexed in such a way stand for the entire synset to which the corresponding word meanings belongs.

(9) $$\begin{aligned} &\textbf{\textit{fly}}^{\text{n1}} <_{\text{para-hyponyme}} \textbf{\textit{food}}^{\text{n2}} \\ &\textbf{\textit{sing}}^{\text{v1,v2,v3}} <_{\text{para-troponyme}} \textbf{\textit{amuse}}^{\text{v2}} \end{aligned}$$

At least, they pass the linguistic test for para-hyponymes and para-troponymes.

(10) *flies and other forms of food & It's a fly, but it's not food.*

(11) *singing and other manners of amusing somebody & It's singing, but it's not amusing people.*

4 The Ontological Nature of Roles

The above examples indicate a general problem: some roles have (almost) no specific range of types of entities to which, in an adequate context, they may apply. Think of roles such as TOOL or INSTRUMENT which may be carried by almost every (natural or artificial) entity. The options available to the WordNet designers or users for escaping these problems are not really convenient. They (i) could drop all such links and loose the information. They (ii) could add more and more para-links to the DB. But where is the limit? Finally, they (iii) could link the more unspecific roles, not to possible candidates of a set of hyponymes of a certain concept, but to the hypernymic concept itself: ***living_thing***$^{\text{n}} >_{meronymy} \ldots$ ***nutrient***$^{\text{n}} \ldots <_{para-hyponymy}$ ***food***$^{\text{n2}}$. Thereby the semantic relations would sketch approximately the encyclopedic information which is enclosed in the gloss of ***food***$^{\text{n2}}$: "any solid substance [...] that is used as a source of nourishment". At least, this strategy would regain the information lost in the abstraction process, when the ontological richness of the possible contexts was reduced to underspecified classifying criteria. But it would also call for an ontological analysis of the world structure underlying every single concept. In order to avoid arbitrary or ad-hoc local decisions, such an ontological analysis has to satisfy certain constraints, e.g. it has to be systematic, consistent, has to optimize disambiguation, and so forth.

We will not give such an exhaustive analysis here, but restrict to some considerations which shed light on the origins of the problems mentioned above. We adopt the characterization of roles by [3] who distinguish between material and formal roles. We agree with the authors that roles generally are dependent entities, in the sense that they always depend on the existence of a further entity. A further interesting point with respect to NL encoding of roles is a different kind of dependency, however. Role expressions and their contextual interpretations always depend on a certain 'domain of obtainment', which is some kind of reference point, a certain respect, a particular perspective, or value of comparison. Examples for domains of obtainment are:

- a space of subjective mental or emotional states, e.g. ***joy***$^{\text{n1}}$, ***amuse***$^{\text{v1,v2}}$;
- a quality space or scale: ***relief***$^{\text{n7,n8,n9}}$, ***fail***$^{\text{v2, v5}}$;
- comparison to subjective expectations, e.g. ***delay***^{n1}, ***surprise***$^{\text{v1,v2}}$.

Corresponding to the distinction between common-sense world knowledge and linguistic knowledge, NL semantics always includes two aspects of meaning: on the one hand, it

denotes structural properties of possible referents; on the other hand, it restricts the set of possible contexts in which the item can be used (i.e. grammatical contexts, contexts of utterance, and world contexts in which referring expressions are linked to referents and the truth of propositions is evaluated). Therefore it is not enough to focus on the first, the denotational, aspect of word meaning in order to extract world knowledge. We also have to understand how the second, the abstracting, aspect interacts with the first aspect. Hence, an interesting factor for a characterization of roles is how these contextual dependencies and the domain of obtainment of a role-expression are realized or predetermined by lexical entries.

Material Roles in the lexicalization of WordNet are, for instance, ***student***^{n1} or ***announce***^{v2}. Lexical items encoding material roles often specify the types of entities which may carry the role, and they also co-lexicalize the domain of obtainment (e.g. cultural status). This means that these items do not so much depend on the perspective of a given linguistic context, but rather predetermine the range of possible contexts which might enclose them.

Among **formal roles**, we distinguish between **thematic roles** and **schematic roles**. The ontological structure of thematic roles is the least complex. They correspond to a formal relation holding between two entities of the same or of different top-level categories. Examples are ***causal_agent***^{n}, ***product***^{n3} or ***perform***^{v1}. Lexicalizations of thematic roles do not specify an obtainment domain. Therefore this domain has to be specified by the context. The product of an orchestral performance can be seen as a sound (i.e. a particular), but also as the joy or as a headache of (parts of) the audience.

Examples of WordNet's lexicalizations of **schematic roles** are ***speaker***^{n1}, ***food***^{n2}, ***exercise***^{v4}. Schematic roles are based on schemas, i.e. complex chains or networks of categories and relations. Schematic-role expressions may also provide more concrete information about the domains to which those concrete entities and relations belong that are given by the linguistic or world context and which fit into the schema.

So FOOD, for instance, suggests that the speaker thinks of groceries. But as we have seen in the fly case, also entities of an 'exotic' type may fall under a schema node, provided that it satisfies the ontological requirements. An illustration for the schema FOOD is given in Figure 1.

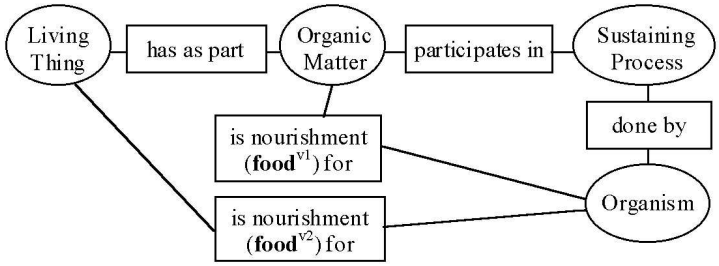


Fig. 1. A possible schema for the schematic role FOOD. Circles symbolize types, boxes stand for relations

Food may be both stuff that an organism ingests and processes (***food***^{n2}) or the narrower notion of what actually is nourishing in ***food***^{n2}, namely ***food***^{n1}. It is not surprising that some

food^{n2} will have ***food***^{n1} as parts (we abstract from the fact that food can be relative to a consumer). According to the schema in Figure 1, an instance of a living thing such as a fly is not itself ***food***^{n1}. Rather, it is a source of ***food***^{n1} in that it contains parts which may play the role of a nutrient. Source of nourishment is then ambiguous and the different senses ought not to be conflated. Claire swallowing a fly brings about a role played by the fly, not a type. The fly plays the role of nourishment insofar as it contains nutritive parts. It is food insofar as it instantiates the schema of Figure 1. The nutritive claim applying to the fly is thus in essence metonymical. It is the most direct link, ***food***^{v2}, and not the more specifi c and rigorous link, ***food***^{v1}, which applies between a living thing (e.g. the fly) and an organism (e.g. Claire).

5 Conclusions

We have used the example of roles to motivate our claim that ontologizing WordNet means unveiling the implicit ontological structures which support lexicalization rather than merely turn the network into a taxonomy. The ontological variety of kinds of roles allows to differentiate between aspects of common-sense knowledge which purely linguistically motivated features race over. Driving WordNet towards ontological adequacy will in effect transform the lexical database into a knowledge base. The challenge is to preserve the richness of semantic information while operating this transformation. This would mean mobilizing other representational tools rather than merely altering the existing semantic relations.

References

1. Fellbaum, C.: Parallel Hierarchies in the Verb Lexicon. In: Proceedings of 'The Ontologies and Lexical Knowledge Bases' Workshop (OntoLex02), 27th May 2002, Las Palmas, Spain. Online source `http://www.bultreebank.org/OntoLex02Proceedings.html` (2002).
2. Gangemi, A., Guarino, N., Oltramari A.: Conceptual Analysis of Lexical Taxonomies – the Case of WordNet Top-Level. In C. Welty, B. Smith (eds.), Proceedings of FOIS 2001, ACM press (2001) 285–296.
3. Guarino, N. and Welty, C. A Formal Ontology of Properties. In R. Dieng and O. Corby (eds.), Knowledge Engineering and Knowledge Management: Methods, Models and Tools. 12th International Conference, EKAW2000. Springer Verlag (2000) 97–112.
4. Oltramari A., Gangemi A., Guarino N., Masolo C.: Restructuring WordNet's Top-Level: The OntoClean approach. Proceedings of LREC2002 (OntoLex workshop). Las Palmas, Spain (2002).

Two Kinds of Hypernymy Faults in WordNet: the Cases of Ring and Isolator*

Yang Liu, Jiangsheng Yu, Zhengshan Wen, and Shiwen Yu

Institute of Computatinal Linguistics, Peking University
100871 Bejing, China
Email: liuyang@pku.edu.cn, yujs@pku.edu.cn, wenzs@pku.edu.cn, yusw@pku.edu.cn

Abstract. Hypernymy is the key relation that serves to form the ontology of the noun and verb concepts in WordNet and provides a common way of making induction along the hypernymy tree for the NLP researchers. However, we find 2 kinds of abnormal hypernymy in WordNet 2.0, the cases of ring and isolator for short, which can largely harass the reasoning and eventually lead to errors.

1 Introduction

As the mostly used MRD for semantic analysis nowadays, WordNet features the following items. First, the founders at Princeton University originally defined the rather abstract concept, *Concept*, by a less abstract concept, *SynSet*, which makes a *Concept* formally representable by itself. Second, they further described many kinds of relation between all these *SynSets*, which makes a *Concept* actually significative in such a semantic network.

By means of this particular organization of WordNet, the NLP researchers can, somehow, evaluate the sense of a word or phrase in its context and the *Concept* eventually emerges. The reasoning of ontology, say induction and deduction, thus gets involved.

The credibility of the reasoning lies in the description of the *Concepts* in WordNet. What really counts is that whether or not all the *SynSets* and their relations are well formed (Liu, 2002).

The relations WordNet now applied to the noun and verb *concepts* are synonymy, antonymy, hypernymy, holonymy, entailment, cause and etc., among which synonymy and hypernymy are the most important. Synonymy and hypernymy help to form the *SynSets* and their hierarchies respectively. The hypernymy tree, as the hierarchy of *Concepts*, provides a common way of making induction for the NLP researchers.

According to the specification of WordNet, the noun and verb *concepts* fall into 40 semantic categories with the noun *concepts* ranging from 04 to 28 and the verb *concepts* ranging from 29 to 43. Each category actually denotes a hypernymy tree by the hypernymy relation and its name and content list below (Fellbaum, 1999).

2 Why the Cases of Ring and Isolator Are Abnormal

In principle, hypernymy indicates the uniqueness of induction by its definition and the hypernym of a *Concept* should always be in the same category of the *Concept* proper. This is

* This research is supported by National Project 973, No.G1998030507-4 National Project 863, No. 2002AA117010-08 and Beijing Natural Science Foundation, No. 4032013.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 347–351.

Table 1. Semantic categories of the noun concepts in WordNet

Category	Name	Contents
04	Act	Nouns denoting acts or actions
05	Animal	Nouns denoting animals
06	Artifact	Nouns denoting man-made objects
07	Attribute	Nouns denoting attributes of people and objects
08	Body	Nouns denoting body parts
09	Cognition	Nouns denoting cognitive processes and contents
10	Communication	Nouns denoting communicative processes and contents
11	Event	Nouns denoting natural events
12	Feeling	Nouns denoting feelings and emotions
13	Food	Nouns denoting foods and drinks
14	Group	Nouns denoting groupings of people or objects
15	Location	Nouns denoting spatial position
16	Motive	Nouns denoting goals
17	Object	Nouns denoting natural objects "not man-made"
18	Person	Nouns denoting people
19	Phenomenon	Nouns denoting natural phenomena
20	Plant	Nouns denoting plants
21	Possession	Nouns denoting possession and transfer of possession
22	Process	Nouns denoting natural processes
23	Quantity	Nouns denoting quantities and units of measure
24	Relation	Nouns denoting relations between people or things or ideas
25	Shape	Nouns denoting two and three dimensional shapes
26	State	Nouns denoting stable states of affairs
27	Substance	Nouns denoting substances
28	Time	Nouns denoting time and temporal relations

quite true of the general linguistics theory. We, however, live in a world of reality other than theory. There do exist case that it is hard to reach the uniqueness of induction for a certain *Concept* and we can only adopt such a belief that this *Concept* might have more than one hypernym, one in its own category (the main category) and the others in other categories (the less important categories). This is an exception to the definition.

In other words, if we use H_{in} to measure the hypernyms of a certain *Concept* C_x in its own category and H_{out} to measure its hypernyms in other categories, the cases we can adopt should satisfy the condition of $0<H_{in}\leq 1$ and the value of H_{out} does not matter too much.

Then what happens to the cases not satisfying this condition? What is the meaning of these cases and whether or not this will happen in WordNet 2.0, the latest version of WordNet family by now?

The denial of $0<H_{in}\leq 1$ might be either (1) $H_{in}\geq 2$, case 1 for short, or (2) $H_{in}=0$, case 2 for short. As the root of the hypernymy tree also satisfies the condition of case 2, we strengthen the condition of case 2 by adding $H_{out}\geq 1$ to it and then get (3) $H_{in}=0$ and $H_{out}\geq 1$, case 3 for short.

(1) For case 1, $H_{in}\geq 2$ means that the current *Concept* C_x has at least 2 fathers in its own category. According to the specification of WordNet we've mentioned above, each

Table 2. Semantic categories of the verb concepts in WordNet

Category	Name	Contents
29	Body	Verbs of grooming, dressing and bodily care
30	Change	Verbs of change of size, temperature, intensity, etc.
31	Cognition	Verbs of thinking, judging, analyzing, doubting, etc.
32	Communication	Verbs of telling, asking, ordering, singing, etc.
33	Competition	Verbs of fighting, athletic activities, etc.
34	Consumption	Verbs of eating and drinking
35	Contact	Verbs of touching, hitting, tying, digging, etc.
36	Creation	Verbs of sewing, baking, painting, performing, etc.
37	Emotion	Verbs of feeling
38	Motion	Verbs of walking, flying, swimming, etc.
39	Perception	Verbs of seeing, hearing, feeling, etc.
40	Possession	Verbs of buying, selling, owning, and transfer
41	Social	Verbs of political and social activities and events
42	Stative	Verbs of being, having, spatial relations
43	Weather	Verbs of raining, snowing, thawing, thundering, etc.

category already denotes a hypernymy tree by the hypernymy relation. This condition will unavoidably lead to the case of ring in WordNet. Along these upward arcs of hypernymy of *Concept* C_x, there naturally exists C_x's most nearby ancestor, say *Concept* C_z, which has at least 2 chirldren, say *Concept* C_{y1} and C_{y2}; at the same time, *Concept* C_{y1} and $C_{y2}$are all C_x's ancestors. As WordNet is an inheritance system (Fellbaum, 1999), we can now infer that C_x shares C_{y1} and C_{y2}'s all properties, among which a pair of properties must be opposite for C_{y1} and C_{y2} have the same father C_z and hereby is distinguishable. This is paradoxical by the general linguistic theory.

(2) For case 1, H_{in}=0 means that the current *Concept* C_x has no father at all and it can be the root of the hypernymy tree. This condition doesn't lead to any faults.

(3) For case 3, H_{in}=0 and $H_{out} \geq 1$ means that the current *Concept* C_x has nothing, by the hypernymy relation, to do with any available *Concept* C_z as its father in its own category. Also, C_x has at least 1 father in other categories and actually belongs to those categories. This is nonsense and leads to the case of isolator.

In the final analysis, both the cases of ring and isolator are abnormal.

3 The Searching Algorithm and the Obtained Results

In order to explore the actual cases of ring and isolator in WordNet 2.0, we devised the searching algorithm for the noun *Concepts* demonstrated as follows. It can also apply to the verb *Concepts* after minor modification of the value of the boundary information about the semantic categories.

```
Case_Ring_Total=0
Case_Isolator_Total=0
Case_Ring_by_Category(4..28)=0
```

```
Case_Isolator_by_Category(4..28)=0
Boundary(4..28)=Begin_Offset_of_Category
Boundary(29)=Biggest_Offset_of_Dat_File
Dat_File.Recordset.MoveFirst
Do UntilDat_File.Recordset.EOF
  Number_of_IN_Hypernyms=0
  Number_of_OUT_Hypernyms=0
  Dat_File_Line_String=Data.Recordset.Fields("Dat_File_Line")
  Category=Val(Mid(Dat_File_Line_String,10,2))
  Position=InStr(Dat_File_Line_String,"@")
  Do While Position>0
    Hypernym_String=Mid(Temp,Pos+2,8)
    If Hypernym_String is between Boundary(Category)
    and Boundary(Category+1) Then
      Number_of_IN_Hypernyms=Number_of_IN_Hypernyms+1
    Else
      Number_of_OUT_Hypernyms=Number_of_OUT_Hypernyms+1
    End If
    Position=InStr(Pos+18,Dat_File_Line_String,"@")
  Loop
  If Number_of_IN_Hypernyms>=2 Then
    Record the current Dat_File_Line_String as an example of Case Ring
    Case_Ring_by_Category(Category)=Case_Ring_by_Category(Category)+1
    Case_Ring_Total=Case_Ring_Total+1
  End If
  If Number_of_IN_Hypernyms=0 and Number_of_OUT_Hypernyms>=1 Then
    Record the current Dat_File_Line_String as an example of Case Isolator
    Case_Isolator_by_Category(Category)=Case_Isolator_by_Category(Category)+1
    Case_Isolator_Total=Case_Isolator_Total+1
  End If
  Dat_File.Recordset.MoveNext
Loop
```

By this algorithm, we found 1,839 occurrences out of a total of 79,689 noun *Concepts* and 17 occurrences out of a total of 13,508 verb *Concepts* for the case of ring in WordNet 2.0. The percentages are 2.31% and 0.13% respectively. Table 3 and 4 show the detailed portion for each category.

Table 3. Cases of ring in the noun Concepts

[C04] 73	[C05] 27	[C06] 258	[C07] 12	[C08] 23
[C09] 29	[C10] 67	[C11] 5	[C12] 11	[C13] 24
[C14] 34	[C15] 205	[C16] 0	[C17] 11	[C18] 682
[C19] 7	[C20] 29	[C21] 10	[C22] 8	[C23] 13
[C24] 2	[C25] 4	[C26] 102	[C27] 193	[C28] 8

For the case of ring, there are 2,654 occurrences out of a total of 79,689 noun *Concepts* and 1,551 occurrences out of a total of 13,508 verb *Concepts* in WordNet 2.0. The percentages are 3.33% and 11.48% respectively. Table 5 and 6 show the details.

Table 4. Cases of ring in the verb Concepts

[C29] 0	[C30] 5	[C31] 0	[C32] 0	[C33] 0
[C34] 1	[C35] 4	[C36] 2	[C37] 0	[C38] 1
[C39] 0	[C40] 1	[C41] 2	[C42] 1	[C43] 0

Table 5. Cases of isolator in the noun Concepts

[C04] 65	[C05] 415	[C06] 199	[C07] 30	[C08] 93
[C09] 54	[C10] 73	[C11] 15	[C12] 42	[C13] 34
[C14] 37	[C15] 351	[C16] 6	[C17] 114	[C18] 394
[C19] 33	[C20] 286	[C21] 56	[C22] 10	[C23] 15
[C24] 72	[C25] 21	[C26] 99	[C27] 112	[C28] 28

Table 6. Cases of isolator in the noun Concepts

[C29] 104	[C30] 211	[C31] 87	[C32] 136	[C33] 69
[C34] 32	[C35] 283	[C36] 43	[C37] 36	[C38] 106
[C39] 45	[C40] 76	[C41] 197	[C42] 112	[C43] 14

4 Conclusion

To sum up, the cases of ring and isolator, as 2 kinds of hypernymy faults we've found in WordNet, can largely harass the reasoning along the hypernymy tree for the NLP researchers and eventually lead to errors. In the future, some amendments should be made to solve these issues during the evolution of WordNet.

References

1. Fellbaum, C. WordNet: an Electronic Lexical Database. Cambridge, Mass.: MIT Press (1999).
2. Liu, Y., Yu, J. S. and Yu, S. W. A Tree-Structure Solution for the Development of Chinese WordNet. GWC 2002, India (2002).
3. Liu, Y., Yu, S. W. and Yu, J. S. Building a Bilingual WordNet-Like Lexicon: the New Approach and Algorithms. COLING 2002, Taipei, China (2002).
4. Pianta, P., Pala, K. VisDic – a New Tool for WordNet Editing. GWC 2002, India (2002).
5. Vossen, P. EuroWordNet: a Multilingual Database with Lexical Semantic Networks. Dordrecht: Kluwer (1998).
6. Yu, J. S. Evolution of WordNet-Like Lexicon. GWC 2002, India (2002).

Statistical Overview of WordNet from 1.6 to 2.0

Jiangsheng Yu*, Zhenshan Wen, Yang Liu, and Zhihui Jin

Institute of Computational Linguistics, Peking University

Abstract. We defined several discrete random variables and made their statistical comparisons between different versions of WordNet, by which the macroscopical evolution of WordNet from 1.6 to 2.0 is explored. And at the same time, the examples of extreme data will be enumerated during the experimental analysis.

Keywords *WordNet, distribution, Kolmogorov-Smirnov test*

1 Introduction

For a complex machine-readable dictionary like WordNet [3], it is difficult to compare versions by all the modifications in details [12]. Yet, sometimes we indeed feel a stable trend with more updatings. In the following sections, we will define several discrete random variables and explore their statistical properties in WordNets. For convenience, only the noun and verb parts are considered.

Table 1. Amount of SynSets and words in WordNet

Amount	#NounSynSet	#VerbSynSet	#Noun	#Verb
WN1.6	66,025	12,127	94,474	10,319
WN1.7	75,804	13,214	109,195	11,088
WN2.0	79,689	13,508	114,648	11,306

The first random variable (rv), say F, is the amount of instant hypernyms that a SynSet has, whose distribution indicates the uniqueness of induction along the hypernymy tree. The second rv M describes the polysemia of English words. The third rv W measures the size of SynSet, i.e., how many words a SynSet contains. The fourth rv S depicts the amount of hyponyms a SynSet has, by which we are able to learn about the reification of concepts. Lastly, we will show the distribution of category, associated with which the distribution of category depth is studied. The examples of extreme data are enumerated during the experimental analysis and some further work will be mentioned in the conclusion.

A nonparametric method named *two-sample Kolmogorov-Smirnov goodness-of-fit test* is used in the version comparison.

* This research is supported by Beijing Natural Science Foundation, No. 4032013 and National Project 973, No. G1998030507-4. All the data we used in the paper are available at `http://icl.pku.edu.cn/yujs`.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 352–357.

2 Uniqueness of Induction

In WordNet, concept is represented by a SynSet formally. Among the SynSets various relations are defined, where the hypernymy one is the most important. It is very convenient to make induction along the hypernymy tree, which provides us an easy way of reasoning based on the semantic distances. By the fact that a SynSet may have several father-nodes in the net despite of the categories, we surveyed the random variable F, the amount of instant hypernyms each SynSet has, and summarized the data in Table 2.

Table 2. Observations of F in noun and verb SynSets

	#NounSynSet in		
F	WN1.6	WN1.7	WN2.0
0	9	9	9
1	65,144	73,997	77,594
2	852	1,751	2,016
3	18	40	54
4	2	6	12
5	0	1	3
6	0	0	1

	#VerbSynSet in		
F	WN1.6	WN1.7	WN2.0
0	617	626	554
1	11484	12557	12923
2	26	31	31

In WordNets, the noun concept that has the most hypernyms is {*Ambrose, Saint Ambrose, St. Ambrose*}, and then $\{atropine\}$.

The two-sample Kolmogorov-Smirnov goodness-of-fit test (the usual nonparametric approach to testing whether two samples are from the same population when the underlying distributions are unknown, abbreviated by *KS-test*, see [2,6]) denies that the cumulative distribution function (cdf) of F in the noun part of WordNet invariably keeps along the version updatings except from WN1.7 to 2.0 ($ks = 0.0036$ and p-value$= 0.9957$). Apropos of verb concepts, the percentage of roots is much bigger than that of noun concepts. The fact of few instances of multiple hypernyms predicates that the verb concepts are well congregated. For example, the sense 4 of *warm up* is verb concept with two hypernyms. By the KS-test, the distribution of verb hypernym varies much in every version updating. From WN1.6 to 1.7, many verb SynSets with single hypernym were added. And in the latest updating, quite a few roots have been merged. The mean of noun and verb hypernyms is 1.027 and 0.9613, respectively.

3 Polysemia

The cardinality of the meanings of each word in WordNet is a random variable, say M, that can imply the polysemia of English words [10]. The noun with the most meanings in WordNets is *head*, then *line*, and the most meaningful verb is *break*, then *make*.

The KS-test predicates that the polysemia of nouns changes little only from WN1.7 to 2.0 ($ks = 0.007$ and p-value$= 1$), and same thing happens to the verbs ($ks = 0.005$, $p - \text{value} = 0.9989$). Additionally, the mean of senses can be found in Table 6. A further work includes the estimation of sense distribution of the frequent words in practice.

Table 3. Polysemia of nouns and verbs in WordNet

	#Noun in		
M	WN1.6	WN1.7	WN2.0
1	81,910	94,714	99,365
2	8,345	9,416	9,912
3	2,225	2,710	2,859
4	873	1,027	1,113
5	451	535	565
⋮	⋮	⋮	⋮
32	0	1	1

	#Verb in		
M	WN1.6	WN1.7	WN2.0
1	5,752	5,948	6,110
2	2,199	2,499	2,508
3	979	1,085	1,094
4	502	580	604
5	318	357	360
⋮	⋮	⋮	⋮
63	1	1	0

4 Size of SynSet

The size of a SynSet, written by W, is the amount of words it contains, which provides us a cue of word substitution and corpus extension. The largest SynSet in WordNets is {*buttocks, nates,* $\cdots$, *ass*}, and then {*dohickey, dojigger,* $\cdots$, *thingummy*}. The Sense 4 of *love* is the largest verb SynSet, then the senses of *botch* and *bawl out*.

Table 4. Observations of W in noun and verb SynSets

	#NounSynSet in		
W	WN1.6	WN1.7	WN2.0
1	33,926	38,576	40,753
2	21,214	24,158	25,160
3	6,640	8,126	8,502
4	2,551	2,984	3,159
5	973	1,099	1,178
⋮	⋮	⋮	⋮
28	0	1	1

	#VerbSynSet in		
W	WN1.6	WN1.7	WN2.0
1	7,032	7,630	7,855
2	2,782	3,047	3,106
3	1,181	1,271	1,264
4	539	600	608
5	270	318	314
⋮	⋮	⋮	⋮
24	0	0	1

The KS-test detects the diverse distributions of SynSet size in WordNets, except the verb parts of WN1.7 and 2.0 ($ks = 0.0051$ and p-value$= 0.9938$). This conclusion does not contradict with that in Section 2, since SynSet size has nothing to do with the hypernymy relation. From the statistical facts of F and W, we are able to comprehend their distinct functions in lexicographic analysis. In addition, the mean size of SynSets in distinct WordNets is calculated in Table 6.

5 Reification of Concepts

The hyponyms (or troponyms) are usually used as the extension of retrieval word. For instance, the hyponyms of *disaster* $\in$ {*calamity*, *catastrophe*, *disaster*, $\cdots$} include

Table 5. Observations of S in noun and verb SynSets

	#NounSynSet in				#VerbSynSet in		
S	WN1.6	WN1.7	WN2.0	S	WN1.6	WN1.7	WN2.0
0	51,446	59,693	62,870	0	9,069	9,986	10,234
1	5,214	5,800	6,069	1	1,355	1,426	1,444
2	3,003	3,297	3,410	2	568	595	593
3	1,808	1,930	1,994	3	338	328	338
4	1,080	1,178	1,229	4	212	234	235
5	701	782	833	5	124	138	148
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
619	0	0	1	393	0	0	1

{*plague*}, {*famine*}, etc. The amount of hyponyms (or troponyms) a SynSet has is a random variable of our interest, denoted by S in this paper.

The noun concept in WordNet that has the most hyponyms is {*city, metropolis, urban center*}, then {*bird genus*}, {*writer, author*}, {*mammal genus*}. Sense 2 and 1 of *change* has the most troponyms, and then {*be*}. The KS-test verifi es that the distribution of S in verb SynSets is unaltered from WN1.7 to 2.0 ($ks = 0.0034$ and p-value$= 1$). The SynSets with no hyponyms are leaves of the hypernymy trees, whose complement is the set of inner concept nodes. For the leaves are useless for the extension of retrieval word, we examined the inner nodes and found the same result as the forenamed ($ks = 0.0092$ and p-value$= 0.9987$). For the cdf of S, the similarity between WN1.7 and 2.0 is larger than that between WN1.6 and 1.7. The data in the parentheses are the means of inner hyponyms, as a comparison of those without restrictions: see Table 6.

Table 6. Mean of senses, mean size of SynSets, and mean of (inner) hyponyms

WordNet version	Noun senses	Verb senses	Noun SynSets	Verb SynSets	Noun hyponyms	Verb hyponyms
1.6	1.231	2.138	1.762	1.820	1.013 (4.589)	0.9513 (3.772)
1.7	1.234	2.180	1.777	1.829	1.024 (4.820)	0.9550 (3.909)
2.0	1.236	2.179	1.778	1.824	1.027 (4.867)	0.9613 (3.966)

6 Distribution of Category

The amount of noun SynSets that each category contains is a random variable of interest, whose distribution represents an ontology of semantics. Although the KS-test concludes that the distribution of category varies much during the version updatings, but the shape of distribution keeps well that means the ontology of WordNet develops consistently. The numeralization of ontology and its application makes the evaluation possible.

The deepest path of induction in each category is called the *category depth*. It is not the case that the more SynSets a category has the deeper it is, e.g., category 6 and 30. Intuitively,

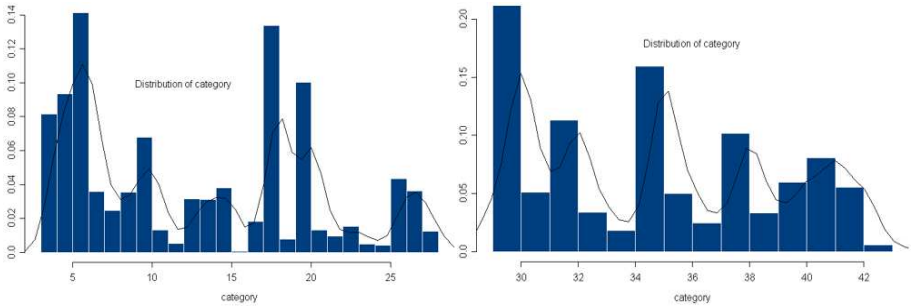


Fig. 1. Distributions of noun and verb categories

the depth of verb category varies less than that of noun category. The noun (verb) category depth reaches the maximum at category 5 (category 41), where 1, 2, 3 denotes WN1.6, 1.7, 2.0 respectively.

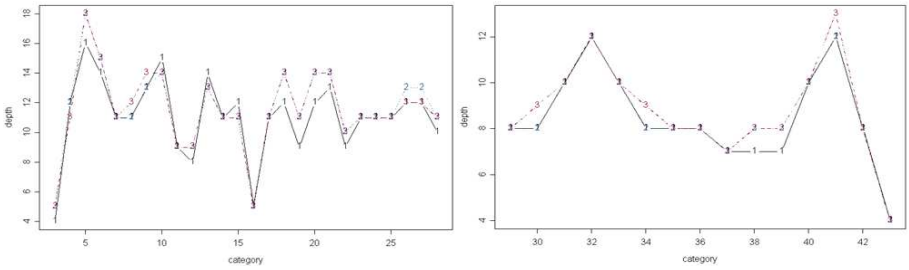


Fig. 2. Scatter plots of noun and verb category depth

Comparing the scatter plots with the histograms of category, there is no obvious relationship between the depth and the distribution. A heuristic explanation of those counterexamples is that the knowledge representation in WordNet by hypernymy tree is notable in width sometimes.

Conclusion

As a linguistic comparison, the statistical survey of Chinese Concept Dictionary (CCD, see [8,13]), the Chinese WordNet, is under consideration. Also, the similar research of EuroWordNet [11] is still worthwhile.

To improve WordNet and its widespread applications (e.g., WSD in [1], text clustering ini [5], semantic indexing in [4,9]), there is still a lot of work to do. For instance, the more advanced coding of offset, the regular patterns of frequent words and concepts, the reasonable defi nition of semantic distance between concepts in WordNet, co-training between WordNet and its application (e.g., information retrieval, text categorization, attitude identifi cation), etc.

Acknowledgement

We appreciate the persistent enthusiasm of all participants in the seminar of Machine Learning at Peking University. Special thanks are due to Prof. Shiwen Yu who is concerned about Chinese WordNet all the time.

References

1. S. Banerjee and T. Pedersen (2002), *An adapted Lesk algorithm for word sense disambiguation using WordNet*. In Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City.
2. P. J. Bickel and K. A. Doksum (2001), *Mathematical Statistics – Basic Ideas and Selected Topics* (Second Edition). Prentice-Hall, Inc.
3. C. Fellbaum (ed) (1999), *WordNet: An Electronic Lexical Database*. The MIT Press.
4. C. Fellbaum, et al (2001), *Manual and Automatic Semantic Annotation with WordNet*. In Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources: Applications, Customizations.
5. A. Hotho, S. Staab and G. Stumme (2003). *Wordnet improves text document clustering*. Submitted for publication.
6. E. L. Lehmann (1975), *Nonparametrics: Statistical Methods based on Ranks*. Hoden-Day, San Francisco.
7. Y. Liu, J. S. Yu and S. W. Yu (2002), *A Tree-structure Solution for the Development of ChineseNet*. The First Global WordNet Conference, Mysore, India, pp. 51–56.
8. Y. Liu, S. W. Yu and J. S. Yu (2002), *Building a Bilingual WordNet: New Approaches and Algorithms*. COLING 2002, Taiwan, pp. 1243–1247.
9. R. Mihalcea and D. I. Moldovan (2000), *Semantic indexing using WordNet senses*. In Proceedings of ACL Workshop on IR & NLP, Honk Kong.
10. W. Peters (2000), *Lexicalized Systematic Polysemy in WordNet*. Proc Second International Conference on Language Resources and Evaluation (LREC 2000), Athens, Greece, pp. 1391–1396.
11. P. Vossen (1999), *Euro WordNet General Document*. University of Amsterdam. Available online at `http://www.hum.uva.nl/-ewn`.
12. J. S. Yu (2002), *Evolution of WordNet-like Lexicon*. The First Global WordNet Conference, Mysore, India, pp. 134–142.
13. J. S. Yu, Y. Liu and S. W. Yu (2003), *The Specifi cation of Chinese Concept Dictionary*. Journal of Chinese Language and Computing, Vol. 13 (2), pp. 176–193.

Author Index

Agirre, E. 4, 15, 23
Alfonseca, E. 15
Alonge, A. 10
Andrikopoulos, V. 265
Atserias, J. 23
Azarova, I. 251

Bae H.-S. 91
Balkova, V. 31
Barbu, E. 332
Barbu, V. 332
Barnden, J. 7
Bentivogli, L. 39, 47
Bertagna, F. 54, 305
Bhattacharyya, P. 83, 226, 291
Bilgin, O. 60
Black, W. J. 67
Bocco, A. 39

Carroll, J. 23
Carthy, J. 112
Castillo, M. 75
Çetinoğlu, Ö. 60
Chakrabarti, D. 83
Chakrabarti, S. 291
Choi, K.-S. 91, 320
Christodoulakis, D. 265
Clough, P. 97
Cucchiarelli, A. 279

Deepa, A. 291
Devitt, A. 106
Dobrov, B. 163
Doran, W. 112
Dunnion, J. 112

Eilts, C. 157
El-Kahlout, İ. D. 118
El-Kateb, S. 67

Fankhauser, P. 326
Farreres, J. 259
Fellbaum, Ch. 3, 187
Ferretti, E. 299

Galiotou, E. 130
Gibert, K. 259
Girish, P. M. 311
Gomez, F. 124
Gonzalo, J. 5
Grenon, P. 341
Grigoriadou, M. 130

Hanks, P. 11
Horák, A. 136
Hümmer, Ch. 142

Ion, R. 332

Jiménez, D. 299
Jin, Z. 352

Kornilakis, H. 130
Koutsoubos, I. D. 265
Krstev, C. 181
Kunze, C. 150

de Lacalle, O. L. 15
Lemnitzer, L. 150
Liu, Y. 347, 352
Lloréns, J. 270
Lönneker, B. 10, 157
Loukachevitch, N. 163

Magnini, B. 23, 169
Marinelli, R. 193
Marzal, M. Á 270
Miháltz, M. 175
Morato, J. 270
Moreiro, J. 270

Navigli, R. 279
Negri, M. 169
Neri, F. 279

Obradović, I. 181
Oflazer, K. 60, 118
Orav, H. 285

Papakitsos, E. 130
Pavlović-Lažetić, G. 181
Pease, A. 187

Peters, W. 8
Pianta, E. 39, 47
Prithviraj, B. P. 291
Prószéky, G. 175

Ramakrishnan, G. 291
Real, F. 75
Rigau, G. 23, 75
Rodríguez, H. 259
Rosso, P. 299
Roventini, A. 193
Rösner, D. 242

Sagri, M. T. 305
Sharada, B. A. 311
Shin, S.-I. 320
Sinopalnikova, A. 199, 251
Smrž, P. 136, 206
Soler, C. 213
Stevenson, M. 97
Stokes, N. 112
Sukhonogov, A. 31

Teich, E. 326
Tiscornia, D. 305
Trautwein, M. 341
Tufiş, D. 332

Veale, T. 220
Velardi, P. 279
Verma, N. 226
Vidal, V. 299
Vider, K. 285
Villarejo, L. 23
Vitas, D. 181
Vogel, C. 106
Vossen, P. 23

Wen, Z. 347, 352
Wong, S. H. S. 234

Xiao, C. 242

Yablonsky, S. 31
Yu, J. 347, 352
Yu, S. 347

Colophon

The GWC 2004 Proceedings have been produced from authors' electronic manuscripts. Following guidelines, authors prepared their papers using LaTeX markup, or in Microsoft Word and sent them electronically to the editors. Files for the Methaphor Panel were collected by Birte Lönneker.

Most of the contributions in Word were converted to LaTeX to allow an easy generation of the table of contents, author index and to make the layout of the Proceedings uniform. Contributions were edited into uniform markup of custom written TeX macros and processed by one of the Proceedings editors in Brno.

Pavel Šmerk and Aleš Horák helped with entering thousands of spelling and typographical corrections into the text corpora of LaTeX files, and with the conversion from Word to TeX. The Proceedings cover was designed by Helena Lukášová.

The Proceedings were typeset in Times Roman and Math Times fonts using ε-TeX typesetting system and LaTeX macro package in a single TeX run. The accompanying CD contains electronic versions of all papers in hypertext form as PDF files. Generating of the hypertext version of Proceedings in PDF was done from the same source files. The reader will find other bonuses as a surprise on the CD.

The main editing, typesetting and proofreading stages were undertaken at the Natural Language Laboratory of the Faculty of Informatics, Masaryk University in Brno.

The Proceedings editors sincerely thank everybody who has been involved in the production. Without their hard work and diligence the Proceedings would not have been ready in time for the GWC 2004 Conference.

Petr Sojka

Book orders should be addressed to:

Pavel Mareček c/o FI MU
Botanická 68a
CZ-602 00 Brno
Phone: ++420 549 498 735
Email: marecek@kup.to

GWC 2004

Proceedings of the Second International WordNet Conference

P. Sojka, K. Pala, P. Smrž, Ch. Fellbaum, P. Vossen (Eds.)

Published by Masaryk University, Brno, 2003

Cover design: Helena Lukášová

First Edition, 2003
Number of copies 150

Printing: Konvoj, s.r.o., Berkova 22, CZ-612 00 Brno, Czech Republic
Email: konvoj@konvoj.cz WWW: http://www.konvoj.cz

55-984B-2003 02/58 3/INF

ISBN 80-210-3302-9