The Acquisition and Representation of Word Meaning

Alessandro Lenci, Simonetta Montemagni, Vito Pirrelli

ESSLLI'01

August, 2001
Helsinki, Finland

The Acquisition and Representation of Word Meaning. An Overview

Alessandro Lenci **Simonetta Montemagni** **Vito Pirrelli**

Istituto di Linguistica Computazionale (CNR)
Area della Ricerca
via Moruzzi 1
56100 PISA
Italy

{lenci,simo,vito}@ilc.pi.cnr.it

1 Introduction

Modelling the way word meanings dynamically function and combine in communicative contexts, evolve through learning and are categorised through high-level semantic classes presents one of the most difficult challenges for cognitive science, and is a large stumbling block on the way to developing advanced artificial systems for full text understanding. Although the cognitive and engineering views of the form and acquisition of semantic knowledge need not be related, data from neuroscience and psychology are indeed relevant when evaluating different ways of representing information in artificial systems, and different models for semantic acquisition.

The problem of the form of semantic knowledge typically presents itself as a representation issue: i.e. what is the aptest way of representing the meaning of words and of the complex expressions they enter into? Providing satisfactory answers to these questions is an essential requirement for explaining the effective use of semantic knowledge in concrete cognitive abilities. This is extremely important also in an engineering and computational perspective, as a key to a deeper understanding of the constructive principles underpinning the design of intelligent artifacts like robots and other artificial intelligent agents. Similarly, the issue of the acquisition of semantic knowledge has a crucial role in modelling human cognition, since cognitive agents constantly update and revise their knowledge, acquire new words, assign new meanings to already known words, etc. Actually, word meanings exhibit a protean nature, a multifaceted behaviour closer to a kaleidoscope of senses continuously changing their relations and nature depending on the vantage point from which they are observed. On the more application-oriented side, endowing systems with the capability of extending and adapting their semantic knowledge is an essential precondition to robustness, scalability, and effectiveness in tackling real-word tasks. As we will see in more detail in what follows, the two aspects of meaning form and acquisition are profoundly inter-related, and progress on both fronts can only be achieved, in our view of things, through a full appreciation of this deep interdependency.

2 The battleground

Two classical theoretical stances have traditionally approached these questions from diametrically opposed perspectives. The central tenet of the *symbolic* or *computational* paradigm is that *representing* meaning, and *processing* meaning representations, essentially consist of symbol manipulation according to explicit logical rules. Symbol manipulation is supposed to be independent of the context/environment in the sense that once a "logical theorem prover" (a Turing Machine) has been given its input representation in symbolic form, the processing can be performed *locally* (e.g. by replacing one symbol with another) and *independently* of what

happens around it. At the opposite end of the scale, Artificial Neural Networks (ANNs) have become popular in the circles of associative and cognitive psychology as models of associationist theories. Connectionist representations can be seen as a high-dimensional space of activities of neurons and connections between them, sometimes called *state space*, by analogy to state spaces in physics. From this perspective, the behavior of an ANN is mathematically defined by a set of *attractor points*, that is stable states where the ANN settles in after having been exposed to (and ultimately perturbed by) some input representations. A key assumption of this approach is that a connectionist meaning representation dynamically unfolds through continuous interaction with other meaning representations and the environment. A quick overview of the strengths and weaknesses of the two approaches is useful to illustrate the main objectives of our workshop.

The symbolic paradigm provides an economic, context-free way of stating invariant concept descriptions, generally in the form of a list of relevant defining properties, and characterises the relationship between two or more concepts/properties in terms of axioms and propositional consequences descending from them. From this perspective, induction has very little to do with concept formation, but basically consists in distilling the rules which are satisfied by all positive instances of a certain concept, and are not satisfied by any negative instance of the same concept. Symbolic representations are extrinsic and explicit, in that they translate any conceptual relation, irrespective of its own intrinsic perceptual or psychological nature, into a logical form, and are thus instrumental for traditional parsers and classical AI-oriented systems. As a result, symbolic representations are homogeneous and respond quite nicely to the language engineering demands for reusable semantic resources in machine readable form.

Last but not least, the symbolic approach to language and cognition tackles the crucial issue of structure in semantics in the most direct way. Word meanings notoriously have a complex and articulated internal architecture, which determines the way they mutually compose in text and propagate into the interpretation of larger linguistic expressions (Jackendoff 1990, Pustejovsky 1995). Symbolic representations are extremely effective in directly organising semantic knowledge into strongly compositional and potentially recursive semantic structures, whose internal constitution is intended to reflect the organisation patterns of knowledge.

On the less positive side, however, symbolic representations are timeless theoretical primitives so that the study of their evolutionary emergence is not really on the computational research agenda. To put it bluntly, there are simply no conceptual categories available for specifying the temporal course of concept formation, or describing the situation before symbols came into being. The symbolic paradigm just leaves time out of the picture, replacing it only with abstract sequences of symbolic states, corresponding to sequences of inferential steps. By the same token, while the structural aspect of meaning is explicitly represented and manipulated, very little is said about the dynamic forces acting on such structures, and determining their formation and evolution. As to the central question of how semantic structures are acquired, a strong form of nativism seems to merely turn the problem of the ontogenesis of meaning representations (i.e. how they can be learned) into a basically unsolved phylogenetic issue (i.e. how they have been developed genetically).

A successful artificial intelligent system, however, must be able to learn new properties from its interactions with the world and not only from new combinations of the given predicates. Even after an agent has learned a new concept, the meaning of the concept is bound to develop and deviate from its initial representation as a result of new experiences. Finally, meaning extension, i.e. the possibility of applying concepts developed within a certain cognitive domain to other related domains through metaphoric concept creation, remains an oddity for the symbolic approach, a fastidious accident requiring ad-hoc axiom stipulation.

Actually, the need for turning symbolic explicit representations of semantic knowledge into more dynamic structures is receiving increasing attention as a major challenge for the whole symbolic approach, as witnessed by the growing amount of work in the field of ontology learning (Staab et al. 2000). *Ontologies* are a powerful tool for knowledge representation, and are widely used to describe and categorise the

semantic content of natural language expressions. An ontology is intended to provide the explicit description of an architecture of concepts, organised according to different types of relations, thereby forming the core of a full-fledged language for knowledge representation. The power (and limit) of ontologies lies in their ability to provide a *snapshot* of a given domain of knowledge. The basic challenge is then turning this snapshot into a dynamic and evolving structure, which might really be able to tackle the complex processing of word meaning acquisition, change and extension. Issues such as the automatic enrichment and customisation of ontologies are high on knowledge engineers' agendas, and interesting interactions with the study of the cognitive dynamics of concept formation and change can be envisaged.

The sub-symbolic representations provided by ANNs are inherently both time- and context-sensitive. A significant feature of these representations is that the salience given to various aspects of a concept may vary depending on the context. The meaning of a concept is not static and homogeneous, but is inherently contextual and thus changes with the context in which it is used. This seems to be in keeping with cognitive and psychological findings on similarity judgements and their contextual grounding. If a child is eating an apple, taste becomes more salient than if the child is using the apple as a ball. By the same token, metaphoric sense extension, far from being a marginal accident of language use, lies at the very core of concept formation and language creativity. In most cases, it makes comparatively little sense to distinguish between a literal and a metaphorical use of a word. The notion of “leg” is probably learned with reference to human legs, but is then naturally extended to dog legs, chair legs, furniture legs etc. This is particularly apparent if we look at the dynamic effects induced by so-called “contrast classes” (Gärdenfors, 2000) on common-or-garden adjectives such as “tall”, “small” or “red”. “Red wine” has little in common with “red hair”, and the role played by the adjective in the two phrases seems to be that of marking a topologically similar contrast in a paradigm-like grid of perceptual discrete oppositions, rather than that of referring to the same stimulus. Incidentally, this seems to suggest the need for more complex and indirect forms of compositionality, whereby it is possible to derive the meaning of complex expressions as a non monotonic and highly non linear (i.e. non isomorphic) function of the meanings of their constituent words.

Another crucial factor involved in meaning representations has to do with learning, and in particular with non-monotonic effects in concept formation leading to the withdrawal of earlier conclusions or expectations (as in the case of non-flying birds like emus). Finally, when an ANN is exposed to partial input representations, its attractor points can be seen as suggesting what further properties of the input representation are to be expected. Accordingly, vector completion can be seen as fulfilling the role of induction.

A fundamental problem with ANNs, however, is that even if a network has learned to categorise the input in the right way, we may not be able to describe what the emerging network represents. In this respect, this problem is similar to the outcome of multidimensional scaling analyses. The interpretation of principal components as the most relevant conceptual dimensions, for example, is often difficult because orthogonal directions of highest variance have little connections with the best projections for categorisation. That is, there are no cognitive or psychological constraints on the principal components.

As to the key notion of structure in semantics, although ANNs may be right in refusing to see cognition as reducible to a discrete manipulation of symbolic structures, yet the role and presence of structures in cognition still represents a considerable challenge for this approach.

3 A way out

In the recent past, an evaluation of the pros and cons of the symbolic and connectionist perspectives on meaning representation has often been turned into a *border dispute*, as though they were describing separable domains of investigation, rather than two complementary levels of analysis and representation of basically the *same range of phenomena*. The most extreme form of the symbolic paradigm places the boundary so as to include any form of

sophisticated behaviour in the computational domain. Likewise, according to the sub-symbolic paradigm the domain of the computational paradigm is virtually empty, as anything can ultimately be explained in terms of patterns of activity values between nodes and connections.

In defining the focus of the present workshop we intended to emphasise that both forms of reductionism are basically ill-founded. The symbolic and the sub-symbolic approaches represent two profoundly inter-dependent levels of analysis, targeting the same bunch of issues at different degrees of granularity. We believe that the notion of *emergence* in non-equilibrium complex systems is considerably instrumental in tackling the issue of interdependence. As Petitot (1995: 231) puts it, "we must first understand how discrete structures can emerge from continuous substrates in the cognitive realm as in the physical realm". Orderly structures in language can emerge "for free" (Kauffman 1995) from computer models simulating the way non linguistic information is "negotiated" through linguistic signs by a group of non-coordinated agents (Batali 1998, Kirby 2000, Batali, to appear). In these models, all individuals are assumed to be endowed with general cognitive capacities remaining constant through time. The cultural evolution of complex signalling systems is a by-product of the assumed cognitive endowment, with virtually no recourse to specifically linguistic abilities.

The central hypothesis here is that language acquisition and transmission are governed by a complex system responsible for the mapping of form and meaning. We can thus hope to describe higher levels of meaning representations as emerging from systems working with subconceptual representations. What seems to be needed to mediate between the two levels is a medium-scale representation framework, capable of reducing the high dimensionality of small-scale sub-symbolic representations to more explanatory low-dimensional spaces. A crucial requirement of this representational middle-ground is that it should preserve the topological and dynamic properties of meaning representations while being able to generalise over them through higher-level categorization processes. This is likely to require something more than a simple extension of off-the-shelf machine learning technology.

For example, until recently, ANNs have mainly been used as mere pattern associators, that is engineering devices for mapping static input representations into static output representations (as in Rosenberg and Sejnowski's NETtalk, 1987). Within this narrow perspective, little insights were gained into the actual dynamic of concept formation and categorization. But, as emphasised by Elman's work on recurrent networks (1990, 1991), when the behaviour of an ANN is regular enough, the representational space staked out by its layer of hidden units is usually of a low dimension and is bound to shed light on the dynamic of the system's categorization process and on the cognitive correlates of its classificatory dimensions.

Kohonen's self-organising maps (SOMs, Kohonen 1995) are another good example of what we seem to be in need of. They tackle the prime issue of lack of structure in distributed information in the most direct way, by providing a powerful mathematical way of transforming and organizing the input into a mode that can possibly be handled at the symbolic level. The methodology makes room for cascaded application of SOMs, thus attaining gradually more abstract levels of representation by using the output of a SOM as the input of another SOM. This basically involves finding a more economic, i.e. low-dimensional representation: going from the subsymbolic to the symbolic always involves a reduction of the number of dimensions representing input stimuli.

Yet another interesting example of search for middle-ground representations is given by the recent upsurge of interest in contextual approaches to word meaning for NLP applications (Pereira *et al*, 1993, Lin 1997, 1998, Allegrini *et al.* 2000). The central claim here is that "substitutability without loss of plausibility is an important factor underlying judgements of semantic similarity" (Miller and Charles, 1991). This interest has both practical and theoretical reasons. For our present concerns, suffice it to point out that the approach has the potential of shedding light on issues of context-sensitive semantic similarity, while getting around the bottle-neck problem of sparse data. From this perspective, meaning similarity

is not dependent on the way concepts are defined in the first place (as lists of both necessary and sufficient defining properties), but acts as a determinant of concept formation, by "discretizing" the contextual space defined by the vectorial representations of word uses.

From a more general perspective any low-dimension space, be it a paradigm-like grid of discrete oppositions, a hierarchical taxonomy, or a Euclidean vector space, can be investigated as a possible solution to the problem of dimensionality reduction. We believe that this line of research has the potential of providing a unified framework addressing the basic theoretical questions in cognitive and psychological inquiry as well as the more technological demands of artificial system engineering. From an engineering standpoint, in building systems capable of performing cognitive tasks, it is often convenient to combine representation levels: different subsystems can be thus designed that work with representations defined at different levels; each system communicates the results of its processes to other subsystems. Ultimately, the number and nature of these levels can be defined as emerging from low level distributed and heterogeneous representations, combining linguistic and non linguistic knowledge (Steels, 1998).

The number of submissions for this workshop, the lively interest it raised, and the diversity of approaches documented by the accepted papers persuaded us that this conciliatory perspective is gaining momentum in the scientific community at large.

References

Allegrini, P., Montemagni, S., Pirrelli, V. (2000), Learning Word Clusters from Data Types. In Proceedings of Coling 2000, Saarbruecken, July 2000.

Batali J. (1998), Computational simulations of the emergence of grammar, in J.R. Hurford, M. Studdert-Kennedy and C. Knight (eds.) *Approaches to the Evolution of Language: Social and cognitive bases*, pp. 405-426, Cambridge University Press.

Batali J. (to appear), The negotiation and acquisition of recursive grammars as a result of competition among exemplars, in T. Briscoe (ed.) *Linguistic Evolution through Language Acquisition: Formal and Computational Models*, Cambridge University Press.

Elman, J.L. (1990), Finding structure in time, *Cognitive Science*, *14*, 179-211.

Elman, J.L. (1991), Distributed representations, simple recurrent networks, and grammatical structure, *Machine Learning*, *7*, 195-225.

Gardenfors, P. (2000), *Conceptual Spaces: The Geometry of Thought*, Cambridge, MIT Press.

Jackendoff, R. (1990), *Semantic Structures*, Cambridge, MIT Press.

Kauffman, S. (1995), At Home in the Universe. The Search for the Laws of Self-organization and complexity, Oxford University Press.

Kirby S. (2000), Syntax without natural selection: how compositionality emerges from vocabulary in a population of learners, in C. Knight, M. Studdert-Kennedy and J.R. Hurford (eds.) *Social Functions and the Origins of Linguistic Form*, pp. 303-323, Cambridge University Press.

Kohonen, T. (1995), *Self-Organizing Maps*, Berlin: Springer-Verlag.

Lin D. (1997), Using Syntactic Dependency as Local Context to resolve Word Sense Ambiguity. In Proceedings of ACL/EACL'97, Madrid, Spain, pp.64-71.

Lin D. (1998), Automatic Retrieval and Clustering of Similar Words. In Proceedings of COLINGACL'98, Montreal, Canada, August 1998.

Miller, G.A., Charles, W.G. (1991), Contextual Correlates of Semantic Similarity. In *Language and Cognitive Processes*, 6 (1), pp. 1-28.

Pereira, F., Tishby, N., Lee, L. (1993), Distributional Clustering Of English Words. In Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics, pp. 183-190.

Petitot, J. (1995), "Morphodynamics and Attractor Syntax: Constituency in Visual Perception and Cognitive Grammar", in port, R. F. and T. van Gelder (eds.), Mind as Motion. Explorations in the Dynamics of Cognition, Cambridge, MIT Press: 227-281.

Pustejovsky, J. (1995), *The Generative Lexicon*, Cambridge, MIT Press.

Rosenberg, C.R., Sejnowsky, T.J. (1987), Parallel networks that learn to pronounce English text, *Complex Systems*, *1*, 145-168.

Staab, S., Maedche, A., Nedellec, C. and P. Wiemer-Hastings eds. (2000), *Proceedings of the First*

ECAI Workshop on Ontology Learning OL'2000, Berlin, Germany, August 25.

Steels L. (1998), Synthesizing the origins of language and meaning using coevolution, self-organization and level formation, in J.R. Hurford, M. Studdert-Kennedy and C. Knight (eds.) Approaches to the Evolution of Language: Social and cognitive bases, pp. 384-404, Cambridge University Press.

Unsupervised Extraction of High-dimensional Representations from Corpora using Simple Recurrent Networks

Fermín Moscoso del Prado Martín
IWTS – Univ. of Nijmegen &
Max Planck Inst. for Psycholinguistics
P.O. Box 310
6500 AH Nijmegen, The Netherlands
fermos@mpi.nl

R. Harald Baayen
IWTS – Univ. of Nijmegen &
Max Planck Inst. for Psycholinguistics
P.O. Box 310
6500 AH Nijmegen, The Netherlands
baayen@mpi.nl

Abstract

Existing techniques for vector-space semantic representations have provided useful tools for the automatic building of semantic type systems. However, these models tend to pay little attention to the position of each element in the sequence of words. This leads to the loss of valuable information. We present an unsupervised technique that extracts rich representations encoding morphological, syntactic and semantic information from large corpora. We present results based both on a small artificial corpus, and larger realistic samples of text.

1 Introduction

This study is motivated by the morphological family size effect, the finding that in language comprehension the speed and accuracy with which a word is recognized is co-determined by the number of words in the mental lexicon in which that word occurs as a constituent (Schreuder and Baayen, 1997). Interestingly, which members of the morphological family contribute to the facilitating effect of the family size depends on the immediate morphological and syntactic context (De Jong et al., 2000), suggesting that the semantic percept of a word is co-determined by both context and by the morphologically related words in the lexicon.

Is the morphological family size effect driven by strictly morphological relations in the mental lexicon, or are general semantic and syntactic relations involved? To answer this question, we need semantic representations that are independent of the form properties of words and dynamically adjustable by context. To be useful for our purposes, a representation should be scalable both in size and in the effort required to build it.

The aim of this work is to obtain high quality representations of meaning from real untagged corpora. We extend an original idea from (Elman, 1990; Elman, 1993) of using the representations developed during the acquisition of a next word prediction task. We compare different implementations of this idea in terms of the quality of the predictions and the representations that they render, and the cost of extending them to work effectively with a large real corpus.

There have been many attempts to capture the semantic information that is present in the word co-occurrence statistics of large corpora. These techniques generally collect matrices with the frequencies with which words co-occur in the same unit of a large corpus (Lund et al., 1995) to full documents (Landauer et al., 1998). After the raw frequencies have been collected, some dimensional reduction technique is applied. This results in a collection of high-dimensional vectors and a distance structure between them.

These systems treat texts as if they were plain sets of words, disregarding the importance of the sequential order in which words appear. Some methods, e.g. (Lund et al., 1995), take word positions into account by weighting the contribution of a co-occurrence frequency with the distance between the two words inside the window. How-

ever, as these weights are then added, this will only modify the final frequency value, that is, the co-occurrence of two words that co-occur seldom at a close distance would be the same as that of words that co-occur often at a far distance.

2 Simulations on the Artificial Corpus

In this paper, we begin by describing a small artificial corpus, then we test both the predictive and representational abilities of two alternate models on that corpus, and then we will apply the technique to larger real corpora of Dutch and English text.

2.1 Data

We obtained our initial corpus from (Elman, 1993). This corpus consisted of $10,000$ sentences of different lengths generated from an artificial grammar that mirrors a number of properties of natural language. The vocabulary consisted of 23 types of different grammatical categories.

Both the verbs and the common nouns have a number of features that we want our techniques to capture. As the grammar generating the corpus did not implement any additional semantic or distributional information, we will use the features Number and Grammatical Category to evaluate the quality of the representations generated by the models. Words were represented in a completely localistic fashion, without any information about their phonological forms.

We divided the original in two parts: a 1000 (10%) sentences testing set, and the remaining $9,000$ (90%) sentences as training corpus.

All models received localist representations of the preceding words and are trained to predict the next word in the sequence.

We will now test the performance on this corpus of two candidate techniques: a case-based reasoning model, and a simple recurrent network. We use traditional n-gram language models to benchmark them both.

2.2 Case-Based Reasoning Models

The case-based reasoning models were built using the memory-based learning technique implemented in TiMBL (Daelemans et al., 2000). Memory-based learning techniques rely on explicitely storing all previous instances of a problem together with their outcomes, without doing any further abstraction. When a new example is presented, the system searches its instance base and looks for the most similar examples that it has already encountered. The target output is predicted to be the outcome of the most similar examples in the instance base. In general, there are two basic parameters that have to be considered in this technique: the number of nearest neighbors used for the decision (k), and the similarity metric used to determine the nearest neighbor set.

In our experiments we used a sliding window technique: The models have to predict the next word on the basis of the w preceding words. We built 9 models with window sizes ranging from 1 to 9 words. The words were represented by plain text labels. Examples were compared using the Modified Value Difference Metric (MVDM; Stanfill and Waltz 1986, Cost and Salzberg 1993) combined with Information Gain Ratio Feature Weighting (IGR; Quinlan 1993), the features being the different positions that figure in the prediction.

Our artificial corpus is composed of sentences that were randomly generated using a simple grammar. We filled all slots of the text window that exceeded the sentence domain with a $[NULL]$ token. Another token was added to represent the end of the sentence. All the models described here use the 11 nearest neighbors.

2.3 Simple Recurrent Network Model

Simple Recurrent Neural Networks (SRN; Jordan 1986, Elman 1990) are a class of multilayer backpropagation artificial neural networks which include a single recurrent loop in the "hidden" units. Following (Elman, 1993), an additional context layer is added to a traditional three-layered neural network. The units of this layer are an exact copy of the activation of the "hidden" layer in the previous time-tick. Their outputs are connected again to the "hidden" layer as if they were normal input units. Networks of this type implicitly code temporal information.

In his original experiments, (Elman, 1993) suggested that for a simple recurrent neural network to succeed in the task of predicting the next word in a sequence, its training regime should be manipulated to limit the amount of memory at the

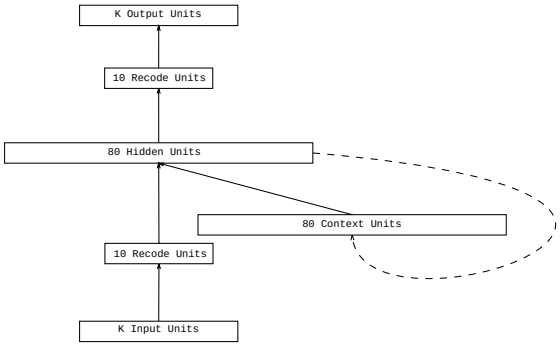


Figure 1: General architecture of the simple recurrent networks that have been used in this work. K is the size of the input/output representation, that is 24 for the localistic case and 15 for the ECOC cases.

initial stages of training. However, (Rohde and Plaut, 1999) shows that this is not necessary, and that a SRN can succeed in learning the task without any initial memory limitation.

Figure 1 shows the basic architecture of the neural networks used in this study. We used 80 hidden units and 80 context units. For the input and output layer, we used two different approaches.

1. The first approach made use of (Elman, 1990; Elman, 1993; Rohde and Plaut, 1999). A localistic representation with one bit per word was used. One additional bit was added to represent the end of sentence, totaling up to 24 bits for the input and output layers of the network.

2. The second approach used Error Correcting Output Codes (ECOC; Dietterich and Bakiri 1995) to encode both the input and the output of the network. ECOC codes, although formally distributed, are equivalent to traditional localistic representations in that the dissimilarities between any two codes follow a probabilistic distribution which does not include any information about word similarity, and present the advantage of facilitating learning (Dietterich and Bakiri, 1995). We have used a pre-designed ECOC representation consisting of 15 bits to encode 24 items.

Both networks were trained for 50 epochs using a modification of the momentum descent algorithm proposed in (Rohde and Plaut, 1999). We used a learning rate of 0.04, a momentum of 0.9 and weights were randomly initialized in the $[-1, 1]$ range. The simulations were run using the *Light Efficient Network Simulator* (Rohde, 1999). We found that after 10 epochs of training, the performance in word prediction stalled. However, we noticed important improvements in the representation quality so we decided to continue training up to 50 epochs. Above this level, performance did not show any significative improvement, neither in prediction nor in representation.

2.4 Prediction Performance

The models were evaluated on their word prediction performance using per-word cross-entropy as a measure of their quality. To have a referent against which to benchmark our models, we built traditional n-gram language models using the *CMU-Cambridge Statistical Modeling Toolkit v2*, (Clarkson and Rosenfeld, 1997). We built models with n values ranging from 2 to 5. Note that a bigram model is approximately equivalent to our case-based reasoning model with a window size of two. All models were trained on the same training corpus. For testing, the test set was divided in ten subsets, each containing 100 of the original 1000 sentences. Cross-entropy was calculated for each of these sub-corpora and then averaged.

The TiMBL system was tested by calculating the per-word cross-entropy over the probability distributions of the words predicted by the system. The cross-entropy was calculated across the 10 testing corpora and then averaged. To evaluate the performance of the neural network using the 24-bit localistic representation, we took the activation of the bits representing each word (after normalization) to be a measure of the probability with which that word was expected to appear.

The calculation of the entropy for the neural network using the 15-bit ECOC was not as straightforward. Due to the distributed aspect of the output, it is difficult to interpret it in terms of a probability distribution. Each bit in the output represents the summed probabilities of several words that share that active bit. To solve this

problem, we estimated the cross-entropy.

We started by assuming that the activation of each bit of the representation is independent from the activation of all other bits. Although this is an obvious simplification, it provides us with a lower bound to the actual probabilities. This first estimate was then corrected by taking into account the influence of the number of bits used and the number of words being encoded, so that our final estimation is that shown in Equation 1 . In the equation B stands for the number of bits in the representation, N stands for the number of types coded, n is the number of tokens in the corpus, C is the previous context and b_i is the activation value of the i-th bit of the output. According to a paired two-tailed t-test, for window sizes greater than two, our estimation was not statistically distinguishable ($t = 1.4099$, $df = 79$, $p = 0.1625$) from that obtained by the traditional method on the case-based reasoning system.

$$\widehat{H}(L,m) = -\frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{B}\log_B(P(b_i|C)) - \frac{1}{\log_2 N} \quad (1)$$

Figure 2 compares the average cross-entropy of the n-gram models, the case-based reasoning models and the neural network. For greater window sizes, the case-based systems clearly outperform the n-gram models. The cross-entropies in this graph were calculated in the traditional way. The performance of both neural networks on word prediction is somewhere in between the prediction power of a bigram model and that of a trigram model.

2.5 Representation Performance

Following training, we extracted representations from the models. In the case of the TiMBL systems, the representations we used were based on the MVDM matrices. We averaged these matrices across features using the IGR weights, and then applied a Principal Component Analysis to the columns of the resulting matrix. Four principal components accounted for approximately 92% of the variance. We used the loadings of the words on these principal components as their vector representations.

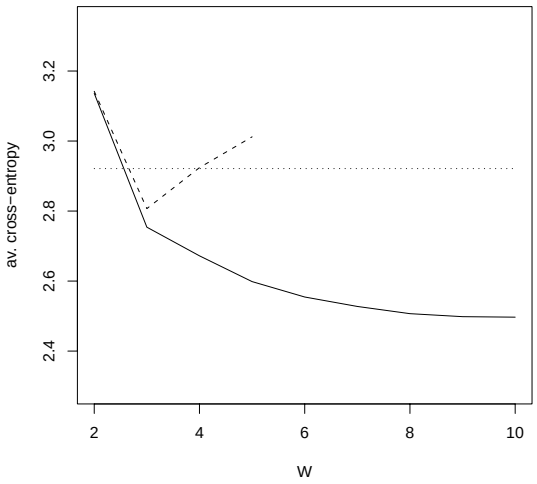


Figure 2: Comparison of the average cross-entropies produced by case-base reasoning systems, n-gram models and a ECOC SRN. The solid line represents the performance of the TiMBL systems, the dashed line represents the n-gram models and the dotted line represents a SRN.

In the case of the neural networks, we followed (Elman, 1990; Elman, 1993) and used the average activation of the hidden units of the network when a word was presented as the initial representation. We also applied Principal Component Analysis to these initial representations. We now found 4 principal components to account for 89% of the variation.

Hierarchical clustering was performed on each of these four representations. The results are shown in Figure 3. The general grammatical classification is almost perfect in all cases. However, the biggest TiMBL system, with a window size of 10, achieves a better classification in terms of number. It is also remarkable that TiMBL systems seem to outperform the networks with respect to number classification of common nouns. In general, we can say that the classification achieved by the cased-based system with a window of 10 is perfect in most aspects, except may be proper nouns and transitive verbs. On the other hand, the network models also achieve excellent performance on the classification of gram-

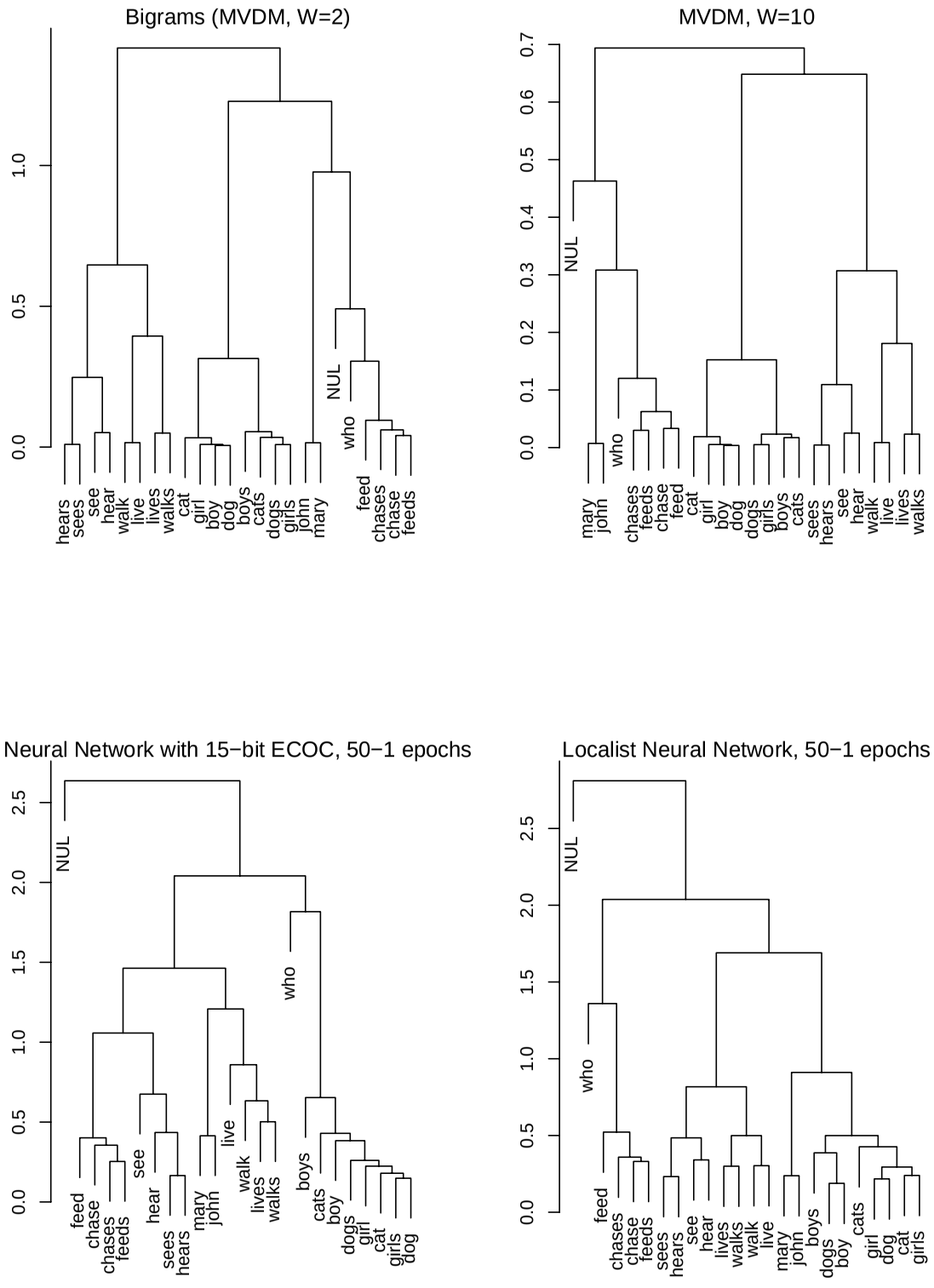


Figure 3: Clustering results from case-based reasoning systems with window sizes 2 and 10, and from the neural networks with and without the use of ECOC codes.

matical categories, especially with respect to the upper branches of the tree which correctly group nouns and verbs. But the performance of the networks is less good for the more detailed Number classification.

3 Using More Realistic Corpora

The results of the previous section seemed to indicate that a Simple Recurrent Network could be adapted to represent the meanings of large sets of words, taking them directly from their occurrence patterns in large corpora. We extended our task to the representation of a larger corpora. This time, we used two more realistic examples:

- A collection of texts consisting of $748,486$ tokens divided in $13,021$ different types. The corpus itself was a collection of Jane Austen's novels extracted from *The Project Gutenberg*. Specifically, our corpus consisted of the novels: *Emma*, *Mansfield Park*, *Northanger Abbey*, *Pride and Prejudice*, and *Sense and Sensibility*.

- As a larger set, we used a collection of Dutch newspaper articles, containing around $4,500,000$ tokens, divided into approximately $160,000$ types.

Only a very naive preprocessing was performed on the original ASCII files: they were all changed to lower-case, digits were substituted for a $[DIGIT]$ token, spaces between common punctuation marks were inserted, an end-of-paragraph token was added and all other non-alphabetical characters except common punctuation marks were removed. The texts were then divided into one-paragraph long 'examples' that were to be used in the training of the networks.

The use of a case-based reasoning system for such large corpora, would require enormous computational resources, and limit the size of our possible windows. As we showed in the previous sections, neural networks overcome these problems while achieving comparable performance.

3.1 Network Design and Training

The networks employed were adaptations of that shown in Figure 1 . The input and output layers consisted of 200 units each for the Austen Corpus, and 300 in the case of the Dutch corpus, the hidden and context layers were kept the same size as for the small artificial corpus, and the number of recoding units was extended to 20.

Although in our initial experiments with the small corpus we found that the use of ECOC codes drastically speeded the training of the networks, our experience with the large corpus showed that a semi-localistic initial representation converged faster in this case. We trained several models, differing in the nature of their input representations and in the pattern of reset applied to the context units. Best performance was achieved by the semi-localistic, paragraph driven model that we describe here.

A localistic representation with one-bit-per-word codings would require at least $13,021$ to encode the vocabulary of the Austen Corpus. Instead we used a 200 bit coding, having two active bits per word. In this way we are able to represent a maximum of $\binom{200}{2} = 19,900$ different types and we still keep them very close to being orthogonal. To build this representation, each type in our vocabulary was assigned two random active bits. This random selection of bits assures that the similarity introduced by having two, instead of one, bits per word behaves as if it were random noise, which actually improves the learning in neural networks. It also guarantees that, in the worst case, any given word can have up to 396 others that are not completely orthogonal to it (that is, they share 1 common bit). This is equivalent to having a maximum of approximately 2% random noise. In the Dutch corpus, we used 3 out of 300 active bits, which allows us to represent a maximum of $\binom{300}{3} = 4,455,100$ types.

The network was trained on word prediction for 20 epochs by modified momentum descent (Rohde and Plaut, 1999) on the Austen corpus. In the Dutch case, due to the larger size of the corpus, only 3 epochs were sufficient to obtain our representations. We used a learning rate of 0.1, a momentum of 0.9 and a batch size of 1 paragraph. The order of presentation of the examples was randomized. Context units were reset to 0.5 at the end of every paragraph. In both cases, the networks were trained for an extra epoch using a learning rate of 0.01 in order to fine-tune the rep-

resentations.

3.2 Evaluation of the Representations

For each corpus, we selected all the types that appeared with a frequency equal or greater than 10. The activation of the networks' hidden and recode units were recorded during test presentations of the full corpora after training. The vectors obtained for each token were averaged across the types. In this way we obtained, for each corpus, two different vectorial representations of the most frequent words in it, one corresponding to the hidden units and the other corresponding to the recode units.

We compared the results from both representations. The information contained turned out to be slightly superior in the 'recode' units. For this reason, we chose to use these as our final representations.

To test the quality of the representations obtained, we chose a list of words belonging to different grammatical categories, and searched for its nearest neighbors in the representation, that is, which other words were more similarly represented by the networks. Similarity was measured using *City-block* distance metric.

$$d(\overline{w}, \overline{v}) = \sum_i |w_i - v_i| \qquad (2)$$

Equation 2 shows the formula used to calculate the distance with $\overline{w}$ and $\overline{v}$ being the vectors of the two words whose distance is estimated. We also experimented using Euclidean and Cosine metrics, but results were found to be best when using City-block Distance.

Tables 1 and 2 show some examples of the results obtained. The networks tend to group nouns and adjectives together. It does not seem to be capturing much meaning in these categories. We believe that this indicates that the network is capturing mostly syntax and, in those cases where semantics can be inferred easily from syntactic patterns, some meaning as well. Interestingly, for nouns or adjectives that show strong argument structures or very defined roles in the sentence, as in family terms and proper nouns the performance is also very good in terms of semantics. For example, it is impressive that all the nearest neighbors of "Lucy" in the Austen corpus are feminine proper names. This is probably very influenced by the fact that, especially in a Jane Austen novel, a female name is likely to appear together with very specific kinds of verbs. A similar effect was found for place names in the Dutch newspapers, which again might have a lot to do with usual newspaper language. Nevertheless, although heavily guided by syntax, the network is also capturing some co-occurrence semantics. This can be seen in the nearest neighbors of "vrouw" (*woman*) and "wet"(*law*) in Dutch, which respectively include "kinderen"(*children*), "vrouwen"(*women*), and "rechter"(*judge*), "ethische"(*ethical*), "buitenland"(*abroad*) and "belangstelling"(*"interest"*).

Another interesting observation concerns the Dutch adjectives "Duits" and "Duitse" which are both forms of the adjective *German*, but in different inflectional contexts. As it can be seen, although the network hasn't achieved much success in clustering semantically related adjectives, it tends to group forms with "-e" and unsuffixed forms together.

The performance of the model is excellent for those words that contain a lot of grammatical information. The nearest neighbors of the Dutch verb "moeten" (*must*) were other modal verbs of similar meaning ("zouden"–*should*, "kunnen"–*can*, "mogen"–*may*, "zullen"–*will*) and always in the same inflectional variant, that is singular forms in the case of "moet" and plural/infinitive forms in the case of "moeten". Similarly, neighbors of the preposition "in", were all prepositions, and the pronoun "ze" (*she/they*) selects other pronouns, preferably those that are also 3rd person singular nominative. In this case it is remarkable that the nearest neighbor is "zij" which is an non-reduced version of the same pronoun.

It was also important to check whether the network was using long distance relationships to capture its representations, or whether it was just taking into account the very local contexts. To do this, a modification in the training regime was introduced: every 6 time ticks the context units were reset to 0.5. In this way we restricted the networks' memory to only the 6 previous time ticks. We found that networks trained in this way failed to capture as much information as those that had an unrestricted memory extending to a full paragraph. This is an indication that, in this task,

Word	**Nearest Neighbors**					**Average Distance**
	1st	2nd	3rd	4th	5th	
must	may	should	shall	will	might	0.606
was	were	are	*dispersed*	*grandmama*	is	0.676
said	shows	replied	continued	persuaded	*novels*	1.676
say	sold*	imagine	know	*directing*	succeed*	0.592
in	beyond	by	of	from	within	0.453
she	he	who	it	this	they	1.618
his	your	their	her	*inconveniences*	our	0.656
small	public*	perfect*	tender	*unkindness*	*comforts*	0.243
lucy	marianne	elinor	julia	edward	harriet	0.337
mother	father	harris	goddard	crawford	weston	0.574
house	case*	night*	system*	servant	ball	0.207

Table 1: Nearest neighbors for several English words as obtained by the system from the Austen corpus. Words that belong to other grammatical categories appear in italics. Words that, although correctly related in syntactic terms, do not bear any special semantic relation with the target word are marked by asterisks.

Word	**Nearest Neighbors**					**Average Distance**
	1st	2nd	3rd	4th	5th	
moet	zou	kan	zouden	zal	*wat*	0.164
moeten	zouden	kunnen	mogen	zullen	zou	0.155
zei	zegt	vindt	doe*	meent	*terwijl*	0.134
zeggen	vinden	bestaat*	zien	gezegd	weten	0.174
in	tussen	voor	naar	tijdens	bij	0.188
ze	zij	je	*daar*	*nu*	we	0.120
duitse	alle*	belgische	politieke	oude*	hogere*	0.131
duits	*dodelijke*	durft	waarschijnlijk*	goed	vermoedelijk*	0.171
vrouw	vrouwen	kinderen	*cda*	namen*	woorden*	0.099
wet	rechter	ethische	buitenland	belangstelling	mens*	0.133
groep	partij	dag*	kans*	verleden*	*officiele*	0.077

Table 2: Nearest neighbors for several Dutch words as obtained by the system from the Dutch corpus. Words that belong to other grammatical categories appear in italics. Words that, although correctly related in syntactic terms, do not bear any special semantic relation with the target word are marked by asterisks.

small window approaches cannot capture all the information required.

Results for Dutch are much better than those obtained for English. This is probably due to the much bigger and varied corpus used for Dutch.

4 Discussion

Neural networks and case-based system provide promising tools for word prediction comparable in power to standard n-gram language models. Interestingly, Figure 2 shows that n-gram models achieve their best prediction performance with $n = 3$, but that the TiMBL systems continue to improve theirs as the window-size is increased, without ever getting worse. This suggests that TiMBL systems provide a promising alternative to traditional interpolated n-gram models, in that much better performance can be achieved at a cost that is linearly proportional to that of a bigram model.

Case-based reasoning systems, when given a sufficiently large window size, outperform neural networks with respect to the more fine-grained aspects of the representations they build, although they have a slightly worse performance with respect to the general structure of the classification trees obtained, at least for the present data set. Neural network models of this kind, on the other hand, are easier to scale up to real-life corpora than case-based reasoning system, because they do not require to store the full example sets, while still achieving a fairly good performance. Neural Networks provide natural, compact and manageable representations of around 20 elements, without the need to deal with huge matrices that then need to be reduced via Principal Component Analysis or Singular Value Decomposition.

There have been many previous works on the use of neural networks to capture different aspects of syntax and semantics (Cotrell, 1989; Elman, 1993; Jain, 1989; Miikkulainen, 1993; StJohn and McClelland, 1990; Wiener et al., 1995; Schütze, 1995; Henderson and Lane, 1998; Morris et al., in press). However, most of this works impose a great amount of structure in on the networks, reflecting different assumptions on the structure of a language. In our approach, we decided to follow (Elman, 1993) in making as few assumptions as possible.

The technique we propose in this paper provides a useful tool for extracting morpho-syntactic features from large corpora, requiring very little design and processing time. This is of great convenience for the semi-automatic development of lexical resources, where new text resources for new domains or even languages need to be added. Although the representations obtained by our systems focus on syntax, we have shown that features that are traditionally considered syntactic can have semantic implications. This was specially evident in the case of verb semantics. We believe that a combination of the technique presented here, with some other traditional co-occurrence based technique such as HAL or LSA, can produce cheap, high quality high-dimensional representations that encode for syntactic and semantic information in a distributed way. This representations can also be used to enhance traditional knowledge-based semantic type systems in a semi-automatic way.

Another advantage of these kind of models is that they provide a natural way to represent context specific properties of certain tokens, both in syntactic and semantic domains. An automatic context-dependent "representator", can be of use for disambiguation or information retrieval systems.

Although at the moment we do not make any specific claims with respect to the psycholinguistic implications of this model, we believe that it shows that the syntactic and semantic information are interrelated, and moreover can be acquired from registering co-occurrence patterns from the raw corpora.

The technique remains to be tested on larger corpora, from which we expect to find a significant improvement on the quality of the representations obtained.

Acknowledgements

We wish to thank Walter Daelemans, Jeff Elman, Thomas Dietterich, David Plaut, Rob Schreuder and Nivja de Jong for helpful comments, discussion and criticism of the ideas contained in this paper. The authors were supported by the Dutch National Research Council, NWO (PIONIER grant to the second author).

References

Philip Clarkson and Ronald Rosenfeld. 1997. Statistical language modeling using the CMU-Cambridge toolkit. In *ESCA Eurospeech 1997*. ESCA.

S. Cost and S Salzberg. 1993. A weighted nearest neighbor algorithm for learning with symbolic features. *Machine Learning*, 10(57–58).

Garrison W. Cotrell. 1989. A connectionist approach to word sense dissambiguation. In *Research Notes in Artificial Intelligence*. Morgan Kaufmann, San Mateo.

W. Daelemans, J. Zavrel, K. van der Sloot, and A. van den Bosch. 2000. TiMBL: Tilburg Memory Based Learner Reference Guide. version 3.0. Technical Report ILK 00-01, Computational Linguistics Tilburg University, March.

N. H. De Jong, R. Schreuder, and R. H. Baayen. 2000. The morphological family size effect and morphology. *Language and Cognitive Processes*, 15:329–365.

T.G. Dietterich and G. Bakiri. 1995. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–286.

J. L. Elman. 1990. Finding structure in time. *Cognitive Science*, 14:179–211.

J. L. Elman. 1993. Learning and development in neural networks: The importance of starting small. *Cognition*, 48:71–99.

James Henderson and Peter Lane. 1998. A connectionist architecture for learning to parse. In *ACL 36 / COLING 17*, pages 531–537.

Ajay N. Jain. 1989. A connectionist architecture for sequential symbolic domains. Technical report CMU-CS-89-187, Carnegie Mellon University.

M.I. Jordan. 1986. Serial order: A parallel distributed approach. Institute for Cognitive Science Report 8604, University of California, San Diego.

T. K. Landauer, P. W. Foltz, and D. Laham. 1998. Introduction to latent semantic analysis. *Discourse Processes*, 25:259–284.

K. Lund, C. Burgess, and R.A. Atchley. 1995. Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, pages 660–665, Hillsdale, NJ. Erlbaum.

Riisto Miikkulainen. 1993. *Subsymbolic Natural Language Processing: An Integrated Model of Scripts, Lexicon and Memory*. MIT Press, Cambridge, MA.

W.C. Morris, G.W. Cottrell, and J.L. Elman. in press. A connectionist simulation of the empirical acquisition of grammatical relations. In Stefan Wermter and Run Sun, editors, *Hybrid Neural Symbolic Integration*. Springer Verlag.

J.R. Quinlan. 1993. *Programs for Machine Learning*. Morgan Kauffmann, San Mateo, CA.

D. L. T. Rohde and D. C. Plaut. 1999. Language acquisition in the absence of explicit negative evidence: how important is starting small? *Cognition*, 72(1):67–109.

Douglas L.T. Rohde. 1999. Lens: The light, efficient network simulator. Technical Report CMU-CS-99-164, Carnegie Mellon University, Pittsburg, PA.

R. Schreuder and R. H. Baayen. 1997. How complex simplex words can be. *Journal of Memory and Language*, 37:118–139.

Hinrich Schütze. 1995. Distributional part-of-speech tagging. In *EACL 7*, pages 251–258.

C. Stanfill and D. Waltz. 1986. Towards memory-based reasoning. *Communications of the ACM*, 29(12):1213–1228, December.

Mark F. StJohn and James L. McClelland. 1990. Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence*, 46:217–457.

Erich Wiener, Jan Pedersen, and Andreas Weigend. 1995. A neural network approach to topic spotting. In *Proceedings of SDAIR 95*, pages 317–332, Las Vegas, NV.

Using Syntactic Contexts for Measuring Word Similarity

Caroline Gasperin[⋆] **Pablo Gamallo**[†] **Alexandre Agustini**[†] **Gabriel Lopes**[†] **Vera de Lima**[⋆]

[⋆]Faculdade de Informática
PPGCC, PUCRS
Av. Ipiranga, 6681
90619-900 Porto Alegre, RS, Brasil
{caroline,vera}@inf.pucrs.br

[†]Faculdade de Ciências e Tecnologia
Universidade Nova de Lisboa
Departamento de Informática, Quinta da Torre
P-2825-114 Monte da Caparica, Portugal
{gamallo,aagustini,gpl}@di.fct.unl.pt

Abstract

This paper explores different strategies for extracting similarity relations between words from parsed text corpora. The strategies we have analysed do not require supervised training nor semantic information available from general lexical resources. They differ in the amount and the quality of the syntactic contexts used to compare words. The paper presents in details the notion of syntactic context and how syntactic information could be used to extract semantic regularities of word sequences. Finally, experimental tests with a Brazilian Portuguese corpus demonstrate that similarity measures based on fine-grained and elaborate syntactic contexts perform better than those based on poorly defined contexts.

1 Introduction

The strategies for extracting semantic information from corpora can be roughly divided into two categories, knowledge-rich and knowledge-poor methods, according to the amount of knowledge they presuppose (8). Knowledge-rich approaches require some sort of previously encoded semantic information (9; 6; 3): domain-dependent knowledge structures, semantic tagged training copora, and/or semantic resources such as handcrafted thesauri: Roget's thesaurus, WordNet, and so on. Therefore, knowledge-rich approaches may inherit the main shortcomings and limitations of man-made lexical resources: limited vocabulary size, since they can include unnecessary general words, or do not include necessary domain-specific ones; unclear classification criteria, since their word classification is sometimes too coarse and does not provide sufficient distinction between words, or is sometimes unnecessarily detailed; and, obviously, considerable time and effort required by building thesauri by hand. By contrast, knowledge-poor approaches use no presupposed semantic knowledge for automatically extracting semantic information. These techniques can be characterised as follows: no domain- specific information is available, no semantic tagging is used, and no static sources as dictionaries or thesauri are required. They attempt to extract the frequency of co-occurrences of words within various contexts to compute semantic similarity among words. More precisely, the similarity measure takes into account the contexts that words share or do not share, as well as the importance of these contexts for each word. Words which share a great number of contexts are considered as similar.

Since contexts can be defined in two different ways, two specific knowledge-poor strategies can also be distinguished: windows-based and syntactic-based techniques. Windows-based techniques consider an arbitrary number of words around a given word as forming its window, i.e., its context. The linguistic information about part-of-speech categories and syntactic groupings is not taken into account to characterise word contexts (1; 10). The syntactic-based strategy, on the contrary, requires specific linguistic information to specify the word context. First, it requires a part-of-speech tagger for assigning a morphosyntactic category to each word of the corpus. Then, the tagged corpus is segmented into a sequence of basic phrasal groupings (or chunks). Finally, simple attachment heuristics are used to specify the relations between and within the phrasal groupings. Once the syntactic analysis of the corpus is reached, each word in the corpus is associated to a set of syntactic contexts. Then, a statistical method compares the frequency of the shared contexts to judge word similarity (7; 4; 2). In both strategies, window-based and syntactic-based techniques, words will be compared to each other in terms of their contextual distribution; yet, we consider that syntactic analysis opens up a much wider range of more precise contexts than does simple windows strategy. As syntactic contexts represent linguistic dependencies involving specific semantic relationships, they should be considered as fine-grained clues for identifying semantically related words.

Since syntactic contexts can be defined in different ways, syntactic-based approaches can also be significantly different. Different pieces of linguistic informa-

tion can be taken into account to characterise syntactic contexts[1]. For instance, the information used by Lin (4) to define the notion of syntactic context is not the same than that used by Grefenstette (7). Nevertheless, the choice of a particular type of syntactic context for measuring word similarity has not been properly justified by those researchers.

This way, the main objective of this paper is to analyse the appropriateness or the inadequacy of different types of syntactic contexts for computing word similarity. More precisely, various syntactic-based strategies will be compared on the basis of different definitions of the notion of syntactic context.

For this purpose, we apply these strategies on a Brazilian Portuguese corpus from NILC (Interinstitutional Center of Computational Linguistics - USP/São Carlos/Brazil), with news documents. Experiments concerning different syntactic relations repported below illustrate and show significant differences among results.

The article is organised as follows. In the next section, various types of syntactic contexts will be analised. Special attention will be paid for the notion of syntactic context used by Grefenstette, as well as for the specific notion that we have defined. Then, in section 3.1, we will use the same statistical similarity measure to compare the appropriateness of the syntactic contexts defined in the previous section. The best results are obtained when the syntactic-based strategy relies on our notion of syntactic context. Samples of the results we have obtained are presented in the Appendix.

2 Types of Syntactic Contexts

In this section, we analyse the notion of syntactic context used by Grefenstette to compute similar words (7). Then, we extract further syntactic information from the partially parsed text in order to make syntactic contexts more elaborate. As a consequence, we obtain fine-grained contexts which contain more specific information than the one provided by Grefenstette's approach.

2.1 The Notion of Attribute by Grefenstette

Grefenstette calls "attributes" the syntactic contexts of a word. Attributes are extracted from binary syntactic dependencies between two words within a noun phrase or between the noun head and the verb head of two related phrases. A binary syntactic dependency could be noted:

$$< R, w1, w2 >$$

where R denotes the syntactic relation itself (e.g., ADJ, NN, NNPREP, SUBJ, DOBJ, and IOBJ), and w1 and w2 represent two syntactically related words. Table 1 shows some syntactic dependencies between the noun "cause" and other related words.

Then, for each word found in the text, the system selects the words that are syntactically related to it. The syntactically related words are considered the attributes of the given word, i.e., its syntactic contexts. For instance, a noun can be syntactically related to an adjective by means of the ADJ relation, to another noun by means of the NN and NNPREP relations, or to a verb by means of SUBJ, DOBJ, and IOBJ relations. These related words are taken to be the known attributes of the noun.

In order to select the attributes of "cause", the system takes as input all the binary dependencies between "cause" and other words. Then, it extracts all the specific words syntactically related to "cause", since they represent its particular attributes. For example, from the 4 dependencies illustrated in 1 between "cause" and another word, it is possible to extract 4 attributes of "cause" (see table 2).

In the Grefenstette's notation, the attributes extracted from noun modifiers (namely NN, ADJ, and NNPREP modifiers) do not keep the name of the particular syntactic relation. So, <*jaundice*>, <*possible*>, and <*death*> are attributes of "cause" even though the syntactic relations NNPREP, ADJ and NN are not explicitly represented. When extracting verbal complements, though, the specific syntactic relation is still available: <*DOBJ*, *determine*> is a verbal attribute constituted by both the word related to "cause" (i.e. the verb "determine") and the specific syntactic relation DOBJ.

2.2 Underspecified Attributes

The notion of attribute defined in the previous section does not inherit all the available syntactic information from binary dependencies. Consider one of the Portuguese expressions found in our corpus: "autorização à empresa" (*permission to the company*). From this expression, <*empresa*> (*company*) is extracted as the attribute of "autorização" (*permission*). Yet, relevant information implicitly contained in the dependency relation has been lost:

- information about the specific preposition: the attribute <*empresa*> does not convey information about the particular preposition "a" relating the two words;
- information about the opposite attribute: the attribute <*autorização*> modifying the word "empresa" is not considered.

Information about prepositions should be taken into account since they convey important syntactic and se-

[1]The choice of a particular measure of similarity may be another parameter to compare various syntactic-based approaches.

mantic information. Let's consider two prepositional expressions: "autorização à empresa" (*permission to the company*) and "autorização da empresa" (*authorization by the company*). According to the Grefenstette's notion of attribute, we should extract the same attribute, namely <*empresa*>, from both expressions. Nevertheless, preposition "a" (*to*) introduces a quite different syntactic dependency than the one introduced by preposition "de" (*by*). Whereas the preposition "a" requires <*empresa*> to be the receiver within the action of giving authorazation, the preposition "de" requires <*empresa*> to be the agent of this action. Therefore, for the purpose of extracting semantic regularities, prepositions should be considered as internal facets of syntactic contexts.

>From the Grefenstette's viewpoint, only one attribute, <*empresa*>, could be extracted from the NN-PREP expression "autorização à empresa". Whereas the modifier word (i.e., the noun after the preposition) is considered as a potential attribute, the modified word (i.e., the noun before the preposition) cannot become an attribute. The tests introduced in section 3.1 will show that the Grefenstette's notion of attribute is too restrictive for the purpose of measuring word similarity. Indeed, the less specific the attribute is, the less precise the word classification will be. In this respect, we should use as specific attributes as possible in order to improve word clustering.

2.3 More Accurate Attributes for Word Similarity Measurement

In order to take into account the implicit information contained in the dependency relationships, we will introduce a more general and flexible definition of attribute. The results of the computational tests presented in the next section will provide us with empirical evidence about the appropriateness of such a definition.

Attributes are extracted from binary syntactic dependencies. A syntactic dependency may be represented as the following binary predication:

$$r(w1^{\downarrow}, w2^{\uparrow})$$

this binary predication is constituted by the following entities:

- the binary predicate r, which can be associated to specific prepositions, subject relations, direct object relations, etc. ;
- the roles of the predicate, "$\downarrow$" and "$\uparrow$", which represent the *modified* and *modifier* roles, respectively;
- the two words holding the binary relation: w1 and w2.

In this binary syntactic dependency, the word indexed by "$\downarrow$" plays the role of *modified*, whereas the word indexed by "$\uparrow$" plays the role of *modifier*. Therefore, w1 is modified by w2 as well as w2 modifies w1. This way, two complementary attributes may be extracted from that syntactic dependency:

$$< \downarrow r, w1 >< \uparrow r, w2 >$$

where $< \downarrow r, w1 >$ is the attribute of w2 and $< \uparrow r, w2 >$ is the attribute of w1. An attribute is defined as the pair constituted by both a specific syntactic function and the word associated to this function. In particular, $\downarrow r$ represents the syntactic function of *modified*, and $\uparrow r$ the *modifier* function. Consider Table 3. The left column contains expressions constituted by two words syntactically related by a particular type of syntactic dependency. The right column contains the attributes extracted from these expressions. For instance, from the expression "autorização à empresa", it was extracted both the attribute <$\uparrow a$, *empresa*>, where "empresa" plays the role of modifier word, and the attribute <$\downarrow a$, *autorização*>, where "autorização" is the modified word. Let´s note that even though "autorização à empresa" is not truly described as a syntactic constituent by the standard syntagmatic grammar, it should be considered as a very informative syntactic context. Furthermore, information about the specific preposition connecting the words is also available. Our notion of attribute is closely related to what Lin calls "feature" (4).

These elaborate attributes provide us with fine-grained syntactic contexts. In the following section, we will compare these informative syntactic contexts to the coarse-grained contexts used by Grefenstette. This will lead us to assume that the elaborate information conveyed by our notion of attribute is able to contribute more accurately to design a suitable strategy for clustering similar words.

3 Comparing Syntactic-Based Strategies

Various semantic extraction techniques were applied to the Brazilian Portuguese corpus from NILC (Interinstitutional Center of Computational Linguistics - USP/São Carlos/Brazil), which is constituted by more than 1,400,000 word occurrences. The corpus is analysed by the parser presented in (5). Similarity was computed by measuring the syntactic information shared by 12,359 different nouns on the basis of 32,293 different attributes.

3.1 The Weighted Jaccard Similarity Measure

To compare the syntactic contexts of two words, we used as similarity measure a weighted version of the

binary Jaccard measure (7).[2] The binary Jaccard measure, noted BJ, calculates the similarity value between two words, m and n, by comparing the attributes they share and do not share:

$$BJ(w_m, w_n) = \frac{|\{w_m\ atts \cap w_n\ atts\}|}{|\{w_m\ atts \cup w_n\ atts\}|}$$

The weighted Jaccard measure considers a global and a local weight for each attribute. The global weight gw takes into account how many different words are associated with a given attribute. It is computed by the following formula:

$$gw(att_j) = 1 - \sum_i \frac{|p_{ij} \log(p_{ij})|}{nrels}$$

where

$$p_{ij} = \frac{freq\ of\ att_j\ with\ w_i}{total\ of\ atts\ for\ w_i}$$

and $nrels$ is the total number of relations extracted from the corpus. The local weight lw is based on the frequency of the attribute with a given word, and it is calculated by:

$$lw(w_i, att_j) = \log(freq\ of\ att_j\ with\ w_i)$$

The whole weight W of an attribute is the multiplication of both the global and the local weights. So, the weighted Jaccard similarity WJ between two words m and n is computed by:

$$WJ(w_m, w_n) = \frac{\sum_j \min(W(w_m, att_j), W(w_n, att_j))}{\sum_j \max(W(w_m, att_j), W(w_n, att_j))}$$

By computing the similarity measure of all word pairs in the corpus, we extracted the list of the most similar words to each word in the corpus. This process was repeated considering different types of syntactic contexts. On the one hand, we tested the relevance of the use of the prepositional information for the attributes' definition. For this purpose, we compared the results obtained from two strategies: "$+prep$–strategy" and "$-prep$–strategy". The former uses attributes containing information about the specific prepositions, while the latter does not use that information. On the other hand, we tested the adequacy of the "↓–attributes" extracted from prepositional dependencies between two noun phrases. For this purpose, we also compared two different methods: "↑↓–strategy" and "↑–strategy". The former contains both types of attributes, while the later uses only ↑–attributes.

[2]We implemented various statistical measures: coefficient of Jaccard, a different version of the weighted Jaccard, and the particular coefficient of Lin. They did not improve, though, the results obtained from the weighted Jaccard measure described in this section.

3.2 $+prep$–strategy *versus* $-prep$–strategy

We tested first the contribution of the specific prepositions to measure word similarity. The manual evaluation is based on a list of 50 randomly chosen words. The results obtained from both strategies, $+prep$–strategy and $-prep$–strategy, showed that there are no significative differences for the words sharing a great number of attributes (namely, more than 100 attributes). That is, the results are not significatively different for words frequently appearing in the corpus. Nevertheless, when the words sharing less than 100 attributes (in fact, the most abundant in the corpus) was compared, we observed that the lists obtained from the $+prep$–strategy are semanticcally more homogeneous than the lists obtained from the $-prep$–strategy. The consideration of the prepositions in the attributes become these ones less common, incresing their weights. On the other side, when we don't consider prepositions in the attributes, the number of attributes shared by words increases, because different attributes (if considering prepositions) can be seen as the same attribute. But, when the number of shared attributes is very high, these modifications became unrelevant. Table 4 shows some of the lists yielded by these strategies for less frequently appearing words.[3]

These results deserve special comments. Let's take the lists obtained from the word "verdade" (*thuth*). In the $+prep$–strategy, the attribute <↑em, *campo*> (<↑in, *field*>) is shared by "verdade" and "impressão" (*impression*). As its global weight is higher, this attribute make the two words more similar. On the contrary, in the $-prep$–strategy, the no–prepositional attribute <↑, *campo*> has a lower weight, which makes the attribute less significant when computing the similarity between "verdade" and "impressão".

Let's consider another example: the lists obtained from the word "punição" (*punishment*). In the $+prep$–strategy, the attribute <↓de, *efeito*> (<↓of, *effect*>) is shared by the words "punição" and "pena" (*penalty*). Given that its global weight is very high, it contributes to make these words semantically close. On the contrary, in the $-prep$–strategy, the no–prepositional attribute <↓, *efeito*> has a lower weight and, consequently, it cannot be considered as a significant clue when comparing the similarity between the words, so "pena" don't appear in list of most similar words of "punição".

Therefore, it can be assumed that the information about specific prepositions is relevant to characterise and identify the significant syntactic contexts used for the measurement of word similarity.

[3]We do not use a systematic evaluation methodology based on machine readable dictionaries or electronic thesaurus, because this sort of lexical resources for Portuguese are not available yet.

3.3 ↑↓–strategy *versus* ↑–strategy

We also tested the contribution of the ↓–attributes (extracted from noun phrases) to yield lists of similar words. The lists obtained from ↑↓–strategy are significantly more accurate than those obtained from the ↑–strategy, even for the frequently appearing words. Table 5 illustrates some of the lists extracted from both strategies.

On the basis of the results illustrated above, it can be assumed that the use of ↓–attributes to yield lists of similar words is extremely significant. Indeed, this type of attributes somehow provides information concerning the semantic word class. Consider the ↓–attributes <$\downarrow de$, *equipe*> (<$\downarrow of$, *team*>), <$\downarrow de$, *jogador*> (<$\downarrow of$, *player*>) and <$\downarrow de$, *partida*> (<$\downarrow of$, *match*>), shared by the words "tênis" (*tennis*) and "vôlei" (*volleyball*). As those attributes require nouns denoting the same class, namely *sports*, they can be conceived as syntactic patterns imposing the same selectional restrictions to nouns. Consequently, the nouns appearing with those specific ↓–attributes should belong to the class of sports.

In the Appendix, we compare the lists extracted by using the fine-grained techniques (i.e., both ↑↓–strategy and $+prep$–strategy) to the lists extracted by using the coarse-grained methods: ↑–strategy and $-prep$–strategy. The words constituting the lists obtained from the more informative strategies are semantically more homogeneous than those obtained from the less informative ones.

4 Final Remarks

According to the results of the tests described above, the strategies based on rich syntactic contexts are more accurate for the measurement of similar words. Experimental tests demonstrated that similarity measures relied on the fine-grained syntactic contexts we have defined in this paper perform better than those based on poorly defined contexts. This is not surprinsing since the specific syntactic data that we used to refine syntactic attributes allows us to identify more informative syntactic-semantic dependencies between words. Special attention was paid to the very informative syntactic pattern "N-PREP", even if it was not considered as a syntactic constituent in standard syntagmatic grammar.

Nevertheless, all the syntactic-based approaches (fine-grained and coarse-grained syntactic strategies) are confronted with two sorts of linguistic phenomena: both polysemic words and odd attachments of syntactic dependencies between words.

The lists of words recognised as being similar to a particular word can be semantically heterogeneous because of the lexical polysemia of the compared word. For example, the word "segundo" (*second*) appears to be similar to words describing time such as "minuto" (*minute*), "instante" (*instant*), as well as to words describing sequence such as "quarto" (*fourth*). It shares with the former group attributes requiring a quantity of time (e.g., <$\uparrow de$, *atraso*> (<$\uparrow of$, *delay*>), <$\uparrow de$, *desconto*> (<$\uparrow of$, *discount*>)), and shares with the later group attributes that could require ordinal modifiers: <$\downarrow adj$, *mundial*> (<$\downarrow adj$, *world competition*>). Various clustering proposals have been made so as to group the similar words together along sense axes (2; 7). However, the implementation of a efficient clustering method used for the groupping of semantically homogenous words remains to be future work.

Finally, attachment errors inherited from the parser should be taken into account. The number of these errors increases when the language analysed is not as syntactically rigid as English. The rate of incorrect attachments is probably related to the non predictable constraints on the Portuguese syntactic order. To palliate the noisy information inherited from the parser limitations for Portuguese, we are impelled to find more robust word similarity strategies than those used for English texts. So, the fine-grained strategies defined in this paper could be perceived as an attempt to partially palliate the poor results obtained by parsing Portuguese text corpora.

References

Carolyn J. Crouch and Bokyung Yang. 1992. Experiments in automatic statistical thesaurus construction. *5th Annual International Conference on Research and Development in Information Retrieval*. Copenhagen, pp. 77-88.

David Faure and Claire N'edellec. 1998. ASIUM: Learning subcategorization frames and restrictions of selection. *10th European Conference on Machine Learning, ECML98*. Workshop on Text Mining.

David Yarowsky. 1992. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. *Proceedings of 14th International Conference on Computational Linguistics, COLING-92*. Nantes, pp. 454-460.

Dekang Lin. 1998. Automatic Retrieval and Clustering of Similar Words. *Proceedings of the COLING-ACL'98*. Montreal.

Eckhard Bick. 2000. *Portuguese Syntax*. VISL Project. p. 114.

Francesc Ribas Framis. 1995. On Learning More Appropriate Selectional Restrictions. *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*. Dublin.

Gregory Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, USA, p. 305.

Gregory Grefenstette 1995. Evaluation Techniques for Automatic Semantic Extraction: Comparing Sy ntatic and Window Based Approaches. *Corpus processing for Lexical Aquisition*. Ed. Branimir Boguraev and James Pustejovsky, The MIT Press, pp. 205-216.

Philip Resnik. 1999. Semantic Similarity in Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research*. Vol. 11, pp. 95-130.

Young Park, Young Han and Key-Sun Choi. 1995. Automatic thesaurus construction using bayesian networks. *International Conference on Information and Knowledge Management*. Baltimore, pp. 212-217.

Appendix

Expressions	Binary Dependencies
`possible causes`	*<ADJ, cause, possible>*
`the cause of neonatal jaundice`	*<NNPREP, cause, jaundice>*
`no cause could be determined`	*<DOBJ, determine, cause>*
`death cause`	*<NN, cause, death>*

Tabela 1: Exemples of syntactic dependencies

Binary Dependencies	Attributes of `cause`
<ADJ, cause, possible>	*<possible>*
<NNPREP, cause, jaundice>	*<jaundice>*
<DOBJ, determine, cause>	*<DOBJ, determine>*
<NN, cause, death>	*<death>*

Tabela 2: Attributes of `cause`

Binary Expressions	Attributes
`autorização à empresa` (*permission to the company*)	<$\uparrow a$, *empresa*>, <$\downarrow a$, *autorização*>
`nomeação do presidente` (*appointment of the president*)	<$\uparrow de$, *presidente*>, <$\downarrow de$, *nomeação*>
`nomeou o presidente` (*appointed the president*)	<$\uparrow dobj$, *presidente*>, <$\downarrow dobj$ nomear>
`discutiu sobre a nomeação` (*disscussed about the appointment*)	<$\uparrow sobre$, *nomeação*>, <$\downarrow sobre$, discutir>

Tabela 3: Elaborate attributes

Word	Cluster of similar words	
	$+prep$–strategy	$-prep$–strategy
verdade (*truth*)	desafio, impressão, notícia, dado, responsabilidade (*challenge, impression, news, datum, responsability*)	desafio, culpa, Bélgica, impressão, notícia (*challenge, guilt, Belgium, impression, news*)
velocidade (*speed*)	ritmo, vantagem, nível, força, pressão (*rhythm, advantage, level, power, pression*)	vantagem, ritmo, nível, força, diferença (*advantage, rhythm, level, power, difference*)
valor (*value*)	preço, salário, proposta, índice, ritmo (*price, salary, proposal, index, rhythm*)	salário, proposta, preço, índice, coisa (*salary, proposal, price, index, thing*)
turno (*turn*)	set, rodada, fase, Copa do Mundo, returno (*set, round, phase, World Cup, return*)	Copa do Mundo, rodada, fase, set, Campeonato Brasileiro (*World Cup, round, phase, set, Brazilian Championship*)
tragédia (*tragedy*)	zebra, polêmica, milagre, terremoto, ferimento (*strange event, polemics, miracle, earthquake, injury*)	zebra, kart, desperdício, superfície, ferimento (*strange event, kart, wastefulness, surface, injury*)
tática (*tactics*)	sistema, fundamento, esquema, pressão, rendimento (*system, foundations, scheme, pression, income*)	fundamento, trajeto, desentendimento, sistema, circunstância (*foundations, way, misunderstanding, system, circunstance*)
talento (*talent*)	potencial, facilidade, craque, criatividade, revelação (*potential, facility, excellent player, creativity, revelation*)	reforço, potencial, movimentação, liberdade, facilidade (*reinforcement, potential, movement, freedom, facility*)
região (*region*)	centro, interior, litoral, oeste, cidade (*center, countryside, seaside, west, city*)	centro, cidade, mestre, estado, oeste (*center, city, master, state, west*)
punição (*punishment*)	suspensão, destaque, pena, atenção, briga (*suspension, prominence, penalty, attention, strife*)	suspensão, destaque, obrigação, mudança, movimentação (*suspension, prominence, obligation, change, movement*)
plano (*plan*)	esquema, programa, trabalho, sistema, objetivo (*scheme, program, work, system, objective*)	esquema, amistoso, trabalho, dinheiro, programa (*scheme, amicable game, work, money, program*)

Tabela 4: Similarity lists of less frequently appearing words (< 100 attributes) produced by contexts with and without prepositional information

Word	Cluster of similar words	
	↑↓–strategy	↑–strategy
zaga	defesa, zagueiro, goleiro, ataque, meio-campo	diretoria, meio-campo, fracasso, defesa, área
(*defence position*)	(*defence, defence player, goal-keeper, attack, middle-field*)	(*management, middle-field, failure, defence, area*)
violência	briga, pressão, confusão, festa, segurança	briga, conselho, substituição, zagueiro, cobrança
(*violence*)	(*strife, pression, confusion, party, safety*)	(*strife, counsel, substitution, defence player, exaction*)
vencedor	campeão, desafio, Japão, equipe, Grécia	Japão, Cuba, campeão, desafio, Grécia
(*winner*)	(*champion, challenge, Japan, team, Greece*)	(*Japan, Cuba, champion, challenge, Greece*)
TV	televisão, imprensa, TVA, revista, jornal	imprensa, TVA, revista, jornal, mapa
(*TV*)	(*television, press, TVA, magazine, newspaper*)	(*press, TVA, magazine, newspaper, map*)
tênis	basquete, vôlei, surfe, liga, boxe	basquete, surfe, recordista, promessa, mapa
(*tennis*)	(*basketball, voleyball, surf, league, boxe*)	(*basketball, surf, record holder, promise, map*)
surpresa	novidade, preocupação, destaque, revelação, atração	preocupação, destaque, novidade, responsabilidade, revelação
(*surprise*)	(*novelty, preoccupation, prominence, revelation, attraction*)	(*preoccupation, prominence, novelty, responsability, revelation*)
surfe	tênis, amador, WCT, vôlei, bicampeão	cinema, recordista, amador, tênis, bicampeonato
(*surf*)	(*tennis, amateur, WCT, volleyball, champion*)	(*cinema, record holder, amateur, tennis, championship*)
sessão	reunião, fila, evento, divisão, etapa	divisão, fila, turno, semana, evento
(*session*)	(*meeting, queue, event, division, degree*)	(*division, queue, turn, week, event*)
regulamento	regra, fórmula, lei, formação, tabela	fórmula, artilheiro, regra, formação, lei
(*regulation*)	(*rule, formula, law, formation, table*)	(*formula, artilleryman, rule, formation, law*)
questão	motivo, coisa, erro, destaque, diferença	segurança, destaque, motivo, erro, trabalho
(*question*)	(*reason, thing, error, prominence, difference*)	(*safety, prominence, reason, error, work*)
proposta	alternativa, intenção, oferta, convite, notícia	salário, alternativa, intenção, contrato, valor
(*proposal*)	(*alternative, intention, offer, invitation, news*)	(*salary, alternative, intention, contract, value*)

Tabela 5: Similarity lists produced by contexts with and without ↓–attributes

Word	Cluster of similar words	
	$+prep$ –strategy and ↑↓–strategy	$-prep$–strategy and ↑–strategy
zaga	defesa, zagueiro, goleiro, ataque, meio-campo	diretoria, meio-campo, fracasso, defesa, zagueiro
(*defence position*)	(*defence, defence player, goal-keeper, attack, middle-field*)	(*management, middle-field, failure, defence, defence player*)
velocidade	ritmo, vantagem, nível, força, pressão	nível, vantagem, diferença, força, ritmo
(*speed*)	(*rhythm, advantage, level, power, pression*)	(*level, advantage, difference, power, rhythm*)
vantagem	diferença, espaço, oportunidade, média, chance	espaço, velocidade, ritmo, chance, média
(*advantage*)	(*difference, space, oportunity, average, chance*)	(*space, speed, rhythm, chance, average*)
uniforme	camiseta, camisa, ataque, calção, verão	ânimo, camiseta, despedida, rivalidade, camisa
(*uniform*)	(*T-shirt, shirt, attack, ???, summer*)	(*courage, T-shirt, leave-taking, rivalry, shirt*)
TV	televisão, imprensa, TVA, revista, jornal	imprensa, TVA, revista, jornal, Globosat
(*TV*)	(*television, press, TVA, magazine, newspaper*)	(*press, TVA, magazine, newspaper, Globosat*)
tranquilidade	estrutura, virtude, confiança, facilidade, equilíbrio	potencial, estrutura, sorte, personalidade, atenção
(*calm*)	(*structure, virtue, confidence, facility, balance*)	(*potential, structure, lucky, personality, attention*)
tiro	chute, forte, finalização, pontapé, passe	pancada, aula, fama, forte, passe
(*shot*)	(*shot, powerful, finalization, kick, pass*)	(*stroke, class, fame, powerful, pass*)
término	abertura, adiamento, fim, rescisão, começo	realização, rescisão, abertura, returno, fim
(*finish*)	(*opening, postponement, end, rescission, beginning*)	(*realization, rescission, opening, return, end*)
talento	potencial, ídolo, facilidade, craque, fama	ídolo, movimentação, reforço, proteção, liberdade
(*talent*)	(*potential, idol, facility, excellent player, fame*)	(*idol, movement, reinforcement, protection, freedom*)
sessão	reunião, fila, evento, divisão, etapa	semestre, divisão, fila, turno, encontro
(*session*)	(*meeting, queue, event, division, degree*)	(*semester, division, queue, turn, appointment*)
regulamento	regra, fórmula, lei, formação, tabela	fórmula, artilheiro, regra, formação, lei
(*regulation*)	(*rule, formula, law, formation, table*)	(*formula, artilleryman, rule, formation, law*)
recuperação	preparação, preparo, rendimento, tratamento, reação	rendimento, preparo, desgaste, tratamento, estrutura
(*recuperation*)	(*preparation, preparing, yield, treatment, reaction*)	(*yield, preparing, wear, treatment, structure*)
questão	motivo, coisa, erro, destaque, diferença	destaque, motivo, coisa, trabalho, segurança
(*question*)	(*reason, thing, error, prominence, difference*)	(*prominence, reason, thing, work, safety*)
proposta	alternativa, intenção, oferta, convite, notícia	contrato, salário, novidade, intenção, valor
(*proposal*)	(*alternative, intention, offer, invitation, news*)	(*contract, salary, novelty, intention, value*)
programa	plano, corrida, evento, material, formação	corrida, livro, plano, material, esquema
(*program*)	(*plan, race, event, material, formation*)	(*race, book, plan, material, scheme*)
ponto	gol, vitória, título, chance, resultado	vitória, título, chance, gol, lugar
(*point*)	(*goal, victory, title, chance, result*)	(*victory, title, chance, goal, place*)
pai	mãe, família, irmão, filho, criança	Senna, mãe, criança, família, pessoa
(*father*)	(*mother, family, brother, son, child*)	(*Senna, mother, child, family, person*)
luta	confronto, clássico, briga, corrida, competição	corrida, clássico, confronto, lance, conquista
(*fight*)	(*confrontation, classic, strife, race, competition*)	(*race, classic, confrontation, throw, conquest*)
estratégia	mentalidade, virtude, ponto fraco, tática, visão	virtude, destino, mentalidade, endereço, sistema
(*strategy*)	(*mentality, virtue, weak point, tatics, vision*)	(*virtue, destine, mentality, address, system*)
esporte	futebol, coisa, trabalho, vôlei, mudança	vôlei, passatempo, criança, futebol, copa
(*sport*)	(*soccer, thing, work, volleyball, change*)	(*volleyball, pastime, child, soccer, cup*)
clube	equipe, técnico, seleção, futebol, piloto	piloto, atacante, equipe, técnico, futebol
(*club*)	(*team, trainer, selection, soccer, pilot*)	(*pilot, attack player, team, trainer, soccer*)
chance	oportunidade, possibilidade, condição, ponto, gol	partida, ponto, gol, vitória, lugar
(*chance*)	(*oportunity, possibility, condition, point, goal*)	(*match, point, goal, victory, place*)

Tabela 6: Similarity lists produced by two sorts of contexts: contexts with $+$prep and ↑↓–attributes and contexts with $-$prep and ↑–attributes (like Grefenstette's strategy).

Vector-Based Semantic Analysis: Representing Word Meanings Based on Random Labels

Magnus Sahlgren
RWCP Theoretical Foundation SICS Laboratory
Swedish Institute of Computer Science (SICS)
Box 1263, SE-164 29 Kista, Sweden
mange@sics.se

Abstract

Vector-based semantic analysis is the practice of representing word meanings as semantic vectors, calculated from the co-occurrence statistics of words in large text data. This paper discusses the theoretical presumptions behind this practice, and a representational scheme based on the *Distributional Hypothesis* is identified as the rationale for vector-based semantic analysis. A new method for calculating semantic word vectors is then described. The method uses *random labeling* of words in narrow context windows to calculate semantic *context vectors* for each word type in the text data. The method is evaluated with a standardized synonym test, and it is shown that incorporating linguistic information in the context vectors can enhance the results.

1 Introduction: The Meaning of Meaning

The use of vector-based models of information for the purpose of automatic acquisition and representation of semantic knowledge is an area of research that has gained considerable attention over the last decade. A number of different techniques have been suggested that demonstrate the viability of representing word meanings as semantic vectors, computed from the co-occurrence statistics of words in large text data (e.g. Deerwester et al, 1990; Schütze, 1992; Lund and Burgess, 1996).

Unfortunately, the philosophical rationale for this practice has remained tacit, which is remarkable since the various techniques purport to uncover and represent word meanings. What, one might ask, are those meanings that the words of our language apparently have? Are they perhaps some form of mental concepts that exist in the minds of language-users, or are they merely the objects named by the words, and if so: how do we represent something like that in a computer system? It seems that we need to know what it is we want to represent before we can start thinking about how to represent it in a computer system. Succinctly, it seems as if the recourse to semantics demands an explanation of the meaning of "meaning."

Ludwig Wittgenstein suggests in *Philosophical Investigations* (1953) that we should view meaning as something founded in linguistic praxis, and that the meaning of a word is determined by the rules of its use within, as he puts it, a specific *language-game*. This suggestion led to the famous dictum "meaning is use," which is sometimes referred to as a Wittgensteinian theory of meaning. The idea is that to understand the meaning of a word, one has to consider its use in the context of ordinary and concrete language behavior. To know the meaning of a word is simply to be able to use the word in the correct way in a specific language-game or linguistic praxis. This line of reasoning thus allows us to define semantic knowledge as

that which we *make use of* when successfully carrying out linguistic tasks. According to this way of thinking, meaning is the vehicle by which language travels.

Thus, it is language itself, and not the concept of meaning, that is primary to Wittgensteinian semantics. The question about the meaning of "meaning" must be answered "from within" a theory of language, since words do not (and in a stronger sense; *cannot*) have meaning outside language. That is, it does not make any sense to ask what the meaning of a word is in isolation from its use in language, since it is only by virtue of this use that the word has meaning. The lack of rigid designations regarding the concept of meaning facilitates our understanding of language as a dynamic phenomenon. What we need in order to understand the nature of meaning is therefore not so much a rigid definition of the concept of meaning, but rather a profound understanding of the inherent structures of natural language. In short, what we need is a structuralistic account of language.

1.1 The Distributional Hypothesis

Viewed from a structuralistic perspective, natural language may be characterized as a sequence of semantically arbitrary symbols (i.e. words). The symbols are semantically arbitrary since it is not their physical properties that determine their meaning. Rather, it is the relations between the symbols of the sequence that define them. This is to say that the meanings of the symbols in the sequence do not derive from any inherent semantic properties of the symbols.

This can be seen if we consider a polysemous word, for example "fly," that has a different meaning in the context of *A*, for example in texts referring to aerial activities, than in the context of *B*, for example in texts about insects. The reason that the word has different meanings in these different contexts is not because it has a different inherent semantic property in the context of *A* than in the context of *B*, for that would mean that a word could change its inherent semantic properties at any time, making linguistic communication, understanding, and hence language itself virtually impossible. The reason for the different meanings of the word is rather that the context of *A* is different from that of *B*.

This characterization of natural language as a linear sequence of semantically arbitrary elements allows us to formulate a theory about the foundation of meaning known as "the Distributional Hypothesis." The theory originates from the work of Zellig Harris, who, in his book *Mathematical Structures of Language* (1968), states that: "…the meaning of entities, and the meaning of grammatical relations among them, is related to the restriction on combinations of these entities relative to other entities."

These combinatorial restrictions can be viewed as semantic constraints that govern the distribution of entities in language. That is, if two separate entities occur in combination with the same set of other entities *C*, the distribution of the two separate entities are governed by the same semantic constraints, and these semantic constraints are manifested in the distributional pattern that consists in the co-occurrence with *C*. It is by virtue of this distributional similarity that the entities have similar meaning. This is to say that similar distributedness implies similar values of semantic information.

The merit of this hypothesis in relation to the Wittgensteinian account of meaning is that the distributional patterns of words could be thought of as manifesting language use. That is, the use of a word is manifested by its distribution in language, which in its turn is defined by the contexts in which the word occurs. This means that the context could be utilized as a measure of distribution, by which the use (and thus also the meaning) of a word could be determined. Thus, if the *meaning* of a word is determined by its *use* in natural language, and its use is manifested by its *distribution*, the distributional patterns as defined by the *contexts* of a word can be seen as viable tools for determining the meaning of that word.

The idea is that words are semantically similar to the extent that they share contexts. If two words w_1 and w_2, say "beer" and "wine," frequently occur in the same context *C*, say after "drink," the hypothesis states that w_1 and w_2 are semantically related, or, stated

more strongly, that they are semantically similar. The semantic similarity (or relatedness) of "beer" and "wine" is thus due to the similarity of usage of these words. This means that the categorization of both words as, for example, referring to alcoholic beverages is only possible because we use them in such a way; that is, for example after the word "drink" and in the vicinity of the word "drunk." The categorization is not a *cause* of usage, but a *consequence*.

In an attempt to formalize these ideas, we could say that the meaning M of a word w is determined by its distribution D in text T. D over T can be defined as the union of the contexts C in which w occurs. Thus, if $D(w,T)$ (the distribution of w in T) determines $M(w)$ (the meaning of w), and $D(w,T)$ equals:

$$\sum \{C \mid w \in C\}$$

(the contexts in which w occurs), this can be seen as a representation of $M(w)$. This representational scheme thus justifies the claim that word meanings *can* be automatically uncovered from text data and represented in computer systems. It also justifies the claim that this can be done without us having to make any ontological commitments about what these meanings are. Rather, it is because we do not demand any rigid definition of the concept of meaning that this representational scheme becomes feasible.

2 Vector-Based Semantic Analysis

This representational scheme is the communal rationale for vector-based semantic analysis. The assumption that "words with similar meanings will occur with similar neighbors if enough text material is available" (Schütze and Pedersen, 1997) is central for all the different approaches, but they implement this idea in somewhat different ways. In Latent Semantic Analysis (LSA, Landauer and Dumais, 1997), the text data is represented as a words-by-documents co-occurrence matrix where each cell indicates the frequency of a given word in a given text sample of approximately 150 words. The frequencies are normalized, and the normalized matrix is transformed with Singular Value Decomposition (SVD) into a smaller matrix with reduced dimensionality. The purpose of using SVD to reduce the dimensions of the normalized frequency matrix is that this operation appears to accomplish inductive effects that capture latent semantic structures in the text data. Words (or, more to the point, *concepts*) are thus represented in the reduced matrix by semantic vectors of n dimensionality (300 proving to be optimal in Landauer and Dumais' (1997) experiments).

In Hyperspace Analogue to Language (HAL, Lund and Burgess, 1996), the text data is represented as a words-by-words co-occurrence matrix where each cell indicates the co-occurrence counts for a single word pair (a word pair being an asymmetrical relation so that "xy" and "yx" represent different entries in the matrix). The co-occurrence counts are calculated within a 10-word sliding window. Each word is thus represented in the matrix by both a row and a column, and these row/column pairs may be concatenated to produce a co-occurrence vector for each word. Assuming an $n \times n$ co-occurrence matrix, where n is the size of the vocabulary, words are thus represented as semantic vectors of $2n$ dimensionality.

A similar approach is taken by Schütze and Pedersen (1997), who also represent the text data as a words-by-words co-occurrence matrix where each cell records the number of times a word pair has co-occurred in a window spanning 40 words. However, such a matrix with $n^2/2$ distinct entries, where n is the size of the vocabulary, will become computationally intractable for large vocabularies, so they first approximate the matrix using class-based generalization in two steps, and then transform it with SVD so that words get represented in the final reduced matrix as dense semantic vectors of n dimensionality (n = 20 in Schütze and Pedersen, 1997). The motivation for using SVD to reduce the dimensionality of the approximated matrix is, as in LSA, that it improves generalization and makes the representations more compact.

2.1 Random Indexing

We have studied the use of distributed representations for accumulating a words-by-contexts co-occurrence matrix from which

high-dimensional semantic word vectors can be extracted. The technique is called Random Indexing, and it has the same underlying mathematics as, for example, Random Mapping (Kaski, 1998) and Random Projections (Papadimitriou et al, 1998). In Kanerva et al (2000), 1,800-dimensional semantic word vectors were computed using 1,800-dimensional sparse random *index vectors* representing documents of approximately 150 words each. The index vectors were accumulated into a words-by-contexts matrix by adding a document's index vector to the row for a given word every time the word appeared in that document.

This method is comparable to LSA, except that the resulting matrix is significantly smaller than the words-by-documents matrix of LSA, since the dimensionality of the index vectors are smaller than the number of documents. By comparison, assuming a vocabulary of 60,000 words divided into 30,000 text samples, LSA would represent the data in a 60,000 × 30,000 words-by-documents matrix, whereas the matrix in Random Indexing would be 60,000 × 1,800, when 1,800-dimensional index vectors are used. This seems to accomplish the same inductive effects as those attained by applying SVD to the matrix, but without the heavy computational load that SVD requires.

In the present experiment, the high-dimensional random vectors of Random Indexing have been used to index *words* and to calculate semantic word vectors by means of *narrow* context windows consisting of only a few adjacent words on each side of the focus word. As an example, suppose that the number of adjacent words in the context window is two. This implies a window size of five space-separated linguistic units, i.e. the focus word and the two words preceding and succeeding it—what we may call a "2 + 2 sized" context window. Thus, the context for the word *is* in the sentence *This parrot is no more* would be "This parrot" and "no more," as denoted by:

[(This parrot) is (no more)]

Computing semantic word vectors using random indexing of words in narrow context windows is done by first assigning an *n*-dimensional sparse random vector called a *random label* to each word type in the text data. These random labels have a small number k of randomly distributed –1s and +1s, with the rest set to 0. The present experiment has utilized 1,800-dimensional labels with $k \approx 8.7$ (±2.9). Thus, a label might have, for example, four –1s and six +1s.

Next, every time a given word—the focus word f_n— occurs in the text data, the labels for the words in its context window are added to its context vector. For example, assuming a 2 + 2 sized context window as represented by:

$$[(w_{n-2}\ w_{n-1})\ f_n\ (w_{n+1}\ w_{n+2})]$$

the context vector of f_n would be updated with:

$$L(w_{n-2}) + L(w_{n-1}) + L(w_{n+1}) + L(w_{n+2})$$

where $L(x)$ is the label of x. This summation has also been weighted to reflect the distance of the words to the focus word. The weights were distributed so that the words immediately preceding and succeeding the focus word would get more significance in the computation of the context vectors. For the four different window sizes used in these experiments, the window slots were given weights as follows:

1 + 1: [(1) 0 (1)]
2 + 2: [(0.5, 1) 0 (1, 0.5)]
3 + 3: [(0.25, 0.5, 1) 0 (1, 0.5, 0.25)]
4 + 4: [(0.1, 0.1, 0.1, 1) 0 (1, 0.1, 0.1, 0.1)]

where the 0 in the middle represents the focus word.

This method is comparable to HAL, except that we use *distributed* representations that are more "brainlike," efficient and scalable. By comparison, assuming a vocabulary of 60,000 words, the HAL vectors would be (2 × 60,000) 120,000-dimensional, whereas our vectors are 1,800-dimensional, regardless of the size of the vocabulary. Also, we use somewhat smaller context windows to capture the meaning of words. For example, in Burgess and Lund (1998), a context window spanning 10 words were used, whereas we found in our experiments that the performance of the method degrades significantly when the window size exceeds an upper limit of 4 + 4 words.

3 Adding Linguistic Information

The only textual property that is being utilized in the creation of the context vectors is the distributional patterns of linguistic entities. This comprises, however, only a small fraction of the structural complexity of large texts, and it seems reasonable to believe that there are other inherent structural relations in natural language that might be significant for uncovering semantic information. The distributional hypothesis does tell a story about the foundation of meaning, but it might not tell the *whole* story. If the overall goal of the research is to understand how meaning resides in language, it seems unmotivated not to take these more complex linguistic features into account. Therefore, we evaluated the method using different degrees of linguistic preprocessing of the training data, such as morphological analysis and part-of-speech tagging, to see if the utilization of more sophisticated linguistic information in some way concretizes the semantic information captured in the context vectors.

The text data used as learning material in these experiments was a ten-million-word balanced corpus of unmarked English with a vocabulary of 94,000 words. The first stage of preprocessing was to reduce the number of unique word types in the text data by simply determining a frequency range so that words outside the accepted frequency range get 0 weight when they appear in the context window. The frequency range was set to 3–14,000, leaving a vocabulary of 51,000 words. Next, a rather crude method for morphological analysis was implemented by simply truncating the words. The idea was to approximate word stems by chopping off words at a certain predefined number of letters. In the present experiment, truncation lengths of 6, 8, 10 and 12 were investigated. A more reliable and linguistically justified way of extracting the word stems is to use a parser. In the present experiments, the Conexor Functional Dependency Grammar of English (FDG parser) was used to analyze the text and extract the base form of each word.

The Conexor FDG parser was also used to supply the analyzed text (the text consisting of proper word stems) with part-of-speech information in an attempt to remedy the perpetual problem of ambiguity in human languages. This is the problem that the same "word" can have several meanings, and a considerable amount of these orthographically identical but semantically dissimilar words are words that belong to different parts of speech. For example, the difference between *roll* and *roll* depends on whether it is intended to be a verb or a noun. Providing part-of-speech information by simply adding the part-of-speech tag to the beginning of each word would enable the method to discriminate between these words. For example, the verb *roll* would be represented by *vroll*, whereas the noun *roll* would be represented by *nroll*.

4 Evaluation

To repeat the fundamental presumption of this investigation, the *raison d'être* of the high-dimensional context vectors is that they, by effectively being the sum of a word's local contexts, represent the relative meaning of words. This means that they can be used to calculate semantic similarities between words, which can be verified by subjecting the method to a standardized vocabulary test, TOEFL (Test Of English as a Foreign Language). This is a standardized test employed by, for example, American universities to survey foreign applicants' knowledge of the English language. In the synonym finding part of the test, the test taker is asked to find the synonyms to certain given words. For each given word, a multiple-choice answering suggestion of four alternatives is provided, where one alternative is the intended synonym, and is supposed to be indicated by the test person. In the present experiment, 80 test items of this type have been used.

Technically, the test is executed by calculating the distance (in our case, this means the cosine of the angles between the vectors) between the target word and the four alternatives, and the highest-ranking word (i.e. the one that is closest correlated with the target word) is chosen as synonym.

Linguistic Analysis	Context Window				Average (±0.73)
	1 + 1	2 + 2	3 + 3	4 + 4	
None	64.5	67.0	65.3	65.5	65.6
Tr. 6	55.0	57.5	57.3	55.3	56.3
Tr. 8	61.5	64.3	62.0	63.3	62.8
Tr. 10	66.0	68.5	66.3	66.3	66.8
Tr. 12	64.8	65.3	63.8	64.8	64.6
WS	63.5	70.8	72.0	66.0	68.1
PoS+WS	66.0	64.5	65.0	65.5	65.3
Average (±0.56)	63.0	65.4	64.5	63.8	

Table 1: Average results (±1.5) given in percent of correct answers to the TOEFL-test, where Tr. means truncation length, WS means word stems and PoS+WS means part-of-speech tagged word stems.

5 Results

The numbers in the cells of Table 1 are the average results of five runs. The standard deviation for these results is ±1.5. All results are given in percent of correct answers to the TOEFL-test. By comparison, tests with LSA on the same text data, using the LSIBIN program from Telecordia Technologies, produced top scores at 600 factors of 58.75% using the unnormalized matrix, and 65% using a normalized one. The average result reported by Landauer and Dumais (1997) with LSA (using normalization and different text data) is 64.4%, while foreign (non-English speaking) applicants to U.S. colleges average 64.5%.

These results indicate that, by using high-dimensional random distributed representations to label words in narrow context windows, it is possible to reach a result on a standardized synonym test that is equivalent with the performance of other techniques such as, for example, LSA. Without using linguistic information, the system averages 65.6% over the four different window sizes used in these experiments. However, already when utilizing a rather naïve kind of morphological analysis in the form of carefully applied truncation (using a truncation length of 10 characters), the system's average result increases to 66.8% correct answers, although it seems imperative not to truncate too early, since this gravely affects the results. Shortening the truncation length to eight characters decreases the result to 62.8%, and shortening it to six renders a meager 56.3%. Extending the truncation length to twelve characters also decreases the result, with a 64.6% average.

The best result was produced when using proper stemming of the words, with an average of 68.1% correct answers to the synonym part of TOEFL. This indicates that, since the inclusion of morphology in the form of proper word-stem analysis or carefully applied truncation yields the best overall results, taking advantage of other inherent structural relations in text, in addition to the distributional patterns of linguistic entities, really might be significant for uncovering semantic information from text data. However, adding part-of-speech information did not further improve the performance reached when using optimal truncation or proper word-stem analysis. The average result when adding part-of-speech information to the morphologically analyzed text drops to 65.3%. This might be an effect of the increase in the number of unique words in the vocabulary that is the consequence of supplying part-of-speech information for each word.

Turning now to the different window sizes, Table 1 shows that a minimal context window with just one word on each side of the focus word yields the worst average result. This might not be surprising, since it seems reasonable *a priori* to assume that a minimal context window will not provide enough contextual information for making the comparison of distributional similarity reliable. This assumption is not categorically supported by the results however, since for the part-of-speech tagged word stems, a 1 + 1 sized context window actually produces the

best average result. The optimal size of the context window seems to be in the range of two to three words on each side of the focus word, with a 2 + 2 sized context window producing the best average result of 65.4%, but with a 3 + 3 sized context window producing the best individual result of 72% using the morphologically analyzed text. A 4 + 4 sized context window renders only a slightly better average result (63.8%) than the minimal context window, and our experiments indicated a tangible deterioration of the results when the range of the context window exceeded four words on each side of the focus word.

6 Discussion

The reason for using narrow context windows as opposed to large context regions spanning hundreds of words—i.e. whole documents as employed by, for example, LSA—is the assumption that the semantically most significant context is the immediate vicinity of a word. That is, one would expect the words closest to the focus word to be of greater importance than the other words in the text for determining the meaning of the focus word. The intuition is that a local context is more reliable for measuring semantic similarity between words.

This intuition is expressed by, for example, Dekang Lin (1997), who states that "Two different words are likely to have similar meanings if they occur in identical local contexts." Schütze and Pedersen (1997) argue that local co-occurrence statistics is both qualitatively and quantitatively more informative than document-based co-occurrence statistics, since the number of co-occurrence events will be higher when using a sliding window to define the co-occurrence region than when using documents, especially if the documents are long. Burgess and Lund (1998) also report on the merit of using narrow context windows.

If the assumption is correct that local co-occurrence statistics gives a more reliable measure of distributional similarity between words than do document-based co-occurrences, one should be able to discern an increase in performance when using narrow context windows, as opposed to documents, for calculating the semantic word vectors. The results from these experiments seem to be in favor of this assumption. Compared to the performance of techniques using large context regions exceeding a number of hundred words, such as LSA, narrow context windows fare well in an isolated linguistic task pertaining to lexical semantic knowledge.

This raises the question whether there might be a difference in what sort of semantic information that can be extracted by considering different amounts of context. A more sizeable context might give better clues to what a particular word *is about* than to what it *means*. The idea is that two words that *are about* similar things will occur in similar context *regions* (e.g. documents), while two words that have similar *meanings* will occur with similar context *neighbors* (i.e. words). This means that larger contexts might be more suited to use for tasks pertaining to what we may call *topical information*, such as information retrieval, than in tasks specifically directed towards lexical semantic competences. The applicability of LSA to information retrieval is well documented (e.g. Dumais et al, 1988; Deerwester et al, 1990).

The possibility of a discrepancy between the kinds of semantic information carried by different context sizes suggests that, although a synonym test is a fairly reliable method for measuring one kind of semantic knowledge, other conceivable methods for measuring semantic knowledge might be worth considering. Other evaluation procedures have been reported in the literature, such as comparing vector similarities with reaction times from lexical priming studies (Lund and Burgess, 1996) or using LSA for evaluating the quality of content of student essays on given topics (Landauer et al, 1997).

7 Cognitive Implications

The results from these experiments demonstrate that the technique is capable of achieving comparatively good results on a standardized synonym test. Now, the test is designed to measure word knowledge, which means that any subject capable of performing the test with scores above the level of

guessing (which statistically yields 25% correct answers) possesses a certain amount of semantic knowledge about word meanings. Therefore, the test could also be seen as a rudimentary intelligence test. Landauer et al (1998) point out that "word–word meaning similarities are a good test of knowledge—indeed, vocabulary tests are the best single measure of human intelligence."

The results achieved in these investigations are approximately parallel to the results accomplished by foreign applicants to American universities. The question is, then, if the results, viewed from this perspective, justify the conclusion that the system has acquired and applied semantic knowledge? Do the high-dimensional random distributed representations constitute a viable model of such knowledge? In short: what are the cognitive implications of the accomplishments of the method?

The keyword in this discussion is *functionality*. The performance of the method could be described as *functionally* equivalent to the linguistic behavior of a human language user in carrying out the specific predefined linguistic task of picking out synonyms to a target word. This means that, since the system-internal representations in the form of high-dimensional context vectors have been proven, by the successful execution of the TOEFL-test, to be functional for purposes pertaining to linguistic competences, we may describe the system as having acquired and applied what we may call the *computational equivalent* to the semantic knowledge humans possess when discriminating between word meanings.

This means that the relevant question is *not* whether the system-internal representations of word meanings actually *mean* anything (i.e. if they somehow correspond to how word meanings are represented in the human mind—assuming this question is meaningful), but rather whether they can be *utilized* for the purpose of modeling observable linguistic behavior. The semantic information that the context vectors carry does not reside in the vector representations by themselves, but rather in the relations between the vectors. The representation is relative rather than absolute, since it is only in relation to each other that the context vectors *mean* anything. The important point is therefore that the system-internal representations can be utilized to implement specific linguistic behavioral patterns that manifest semantic knowledge—that is, so to speak, a fragment of the *functionality* of a human language user. To repeat the point: it is by virtue of letting the meaning of "meaning" remain indeterminate that we may consider the implementation of a functional pattern as an epistemic or cognitive achievement.

8 Concluding Remarks

The present investigations have shown that random labeling of words in narrow context windows is a viable technique for computing high-dimensional semantic word vectors. The investigations have also shown that the high-dimensional context vectors can be used to perform a standardized synonym test with favorable results. These tests showed that the performance of vector-based semantic analysis can be enhanced by taking advantage of other inherent structural relations in text (i.e. morphology), in addition to the distributional patterns of linguistic entities. This inevitably raises the question whether the semantic representations would benefit from the implementation of other structural features of language such as, for example, dependency relations.

Dekang Lin (1998) demonstrates the viability of using syntactic dependencies as the basis for a comparison of distributional similarity by representing words as the frequency counts of all their *dependency triplets*. A dependency triplet consists of two words and the grammatical relationship between them in a sentence. Similarity between words is then defined as "the amount of information contained in the commonality between the objects divided by the amount of information in the description of the objects" (Lin, 1998). The technique is evaluated through a comparison with manually compiled thesauri (WordNet and Roget's) with favorable results.

This suggests that the inclusion of dependency information in the context vectors might be beneficial for the purpose of

semantic representation. Therefore, the next stage of our investigations will be to incorporate more syntactical information (for instance, dependency relations) in the context vectors to see whether grammatical information affects the semantic representation. It is not a controversial assumption that grammatical relations and syntactical structures have a significant impact on semantics. Concisely, syntax and semantics are not two essentially different things—rather; they are intimately intertwined with each other. Thus, if we want to understand how meaning resides in language, we cannot afford to neglect syntax. What we should aim for is a uniform representation of syntax and semantics, and one possible alternative for such a holistic representation is the high-dimensional vectors of Random Indexing.

Acknowledgements

I am indebted to Dr. Pentti Kanerva for his benign guidance, and to Dr. Jussi Karlgren and the SPC-lab at SICS for consultation and many helpful discussions. Thanks also to David Swanberg for help in running the experiments and for the fruitful dialogue. This research is funded by Japan's Ministry of International Trade and Industry (MITI) through Real World Computing Partnership (RWCP). The training corpus and 80 TOEFL-test items used in these experiments were provided by courtesy of Professor Thomas Landauer, University of Colorado.

References

Burgess, C. and Lund, K. 1998. The Dynamics of Meaning in Memory. In Dietrich, E. and Markman, A. B. (Eds.). 2000. *Cognitive Dynamics: Conceptual Change in Humans and Machines*. Mahwah, NJ: Lawrence Erlbaum Associates.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman, R. 1990. Indexing by Latent Semantic Analysis, *Journal of the Society for Information Science*, 41(6): 391-407.

Dumais, S. T., Furnas, G. W., Landauer, T. K. and Deerwester, S. (1988) Using Latent Semantic Analysis to Improve Information Retrieval. *Proceedings of CHI'88: Conference on Human Factors in Computing*, (pp. 281–285). New York: ACM.

Harris, Zellig. 1968. *Mathematical Structures of Language*. Interscience publishers, New York.

Kanerva, P., Kristofersson, J. and Holst, A. 2000. Random Indexing of Text Samples for Latent Semantic Analysis. In Gleitman, L.R. and Josh, A.K. (Eds.). *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (p. 1036). Mahwah, New Jersey: Erlbaum.

Kaski, Samuel. 1998. Dimensionality Reduction by Random Mapping: Fast Similarity Computation for Clustering. *Proceedings of the IJCNN'98, International Joint Conference on Neural Networks*, vol. 1 (pp. 413–418). Piscataway, NJ: IEEE Service Center.

Landauer, T. K. and Dumais, S. T. 1997. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. *Psychological Review*, 104(2):211–240.

Landauer, T. K., Laham, D., Rehder, B. and Schreiner, M. E. 1997. How Well can Passage Meaning be Derived without using Word Order? A Comparison of Latent Semantic Analysis and Humans. In Shafto, M. G. and Langley, P. (Eds.). *Proceedings of the 19th annual meeting of the Cognitive Science Society* (pp. 412–417). Mawhwah, NJ: Erlbaum.

Landauer, T. K., Laham, D. and Foltz, P. W. 1998. Learning Human-like Knowledge by Singular Value Decomposition: A Progress Report. In Jordan, M. I., Kearns, M. J. and Solla, S. A. (Eds.). *Advances in Neural Information Processing Systems 10* (pp. 45–51). Cambridge: MIT Press.

Lin, Dekang. 1997. Using Syntactic Dependency as Local Context to Resolve Word Sense Ambiguity. In *Proceedings of ACL-97*, Madrid, Spain, July, 1997.

Lin, Dekang. 1998. Automatic Retrieval and Clustering of Similar Words. *COLING-ACL98*, Montreal, Canada, August, 1998.

Lund, K. and Burgess, C. .1996. Producing High-Dimensional Semantic Spaces from Lexical Co-occurrence. *Behavior Research Methods, Instruments and Computers*, 28(2):203–208.

Papadimitriou, C. H., Raghavan, P., Tamaki, H., and Vempala, S. 1998. Latent Semantic Indexing: A Probabilistic Analysis. In *Proceedings of the 17th ACM Symposium on the*

Principles of Database Systems (pp. 159–168). ACM Press.

Schütze, H. and Pedersen, J. O. 1997. A Cooccurrence-based Thesaurus and Two Applications to Information Retrieval. *Information Processing and Management*, 33(3):307-318.

Schütze, Hinrich. 1992. Dimensions of Meaning. In *Proceedings of Supercomputing '92*, Minneapolis, USA, pp. 787–796.

Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. Translated by Anscombe, G.E.M. Oxford, Blackwell. In Swedish translation, 1992. *Filosofiska Undersökningar*. Translated by Wedberg, A. Stockholm, Thales.

Automating Acquisition in Ontological Semantics

Sergei Nirenburg
Computing Research Laboratory
New Mexico State University
Las Cruces, NM 88003, USA
sergei@crl.nmsu.edu

Victor Raskin
Department of English
Purdue University
West Lafayette, IN 47907, USA
vraskin@purdue.edu

1. Place of Acquisition in Ontological Semantics

We define theory as a set of statements that determine the format of descriptions of phenomena in the purview of the theory. A theory is effective if it comes with an explicit methodology for acquiring these descriptions. A theory associated with an application is interested in descriptions that support the work of an application. Figure 1 specifies how that schema applies to ontological semantics (the general notions are listed as headers in the four boxes; their interpretation for ontological semantics is given in the rest of the text in the boxes).

The **theory** of ontological semantics includes the format and the semantics of the text meaning representations (TMRs), the ontology, the Fact DB, the lexicons and the onomasticons as well as the generic processing architecture for analysis of meaning and its manipulation, including generation of text off of it. The **description** part in ontological semantics includes all the knowledge sources, both static and dynamic (generic procedures for extraction, representation and manipulation of meaning), implemented to provide full coverage for a language (or languages) and the world. In practice, ontological semantic description is always partial, covering only a subset of subject domains and sublanguages, and constantly under development, through the process of acquisition and as a side effect of the operation of any applications based on ontological semantics.

The **methodology** of ontological semantics consists of acquisition of the static knowledge sources and of the procedures for producing and manipulating TMRs. We addressed the latter elsewhere (see, for instance, Nirenburg and Raskin 2001, Ch. 8). Here, we focus on the former. In our presentation, we will not focus on the methodology of specific applications of ontological semantics beyond stating that TMRs may be extended in a well-defined way to support a specific application and that such an extension may require a commensurate extension and/or modification of the static resources used by the application. We will start with a general discussion of the attainable levels of automation for acquiring static knowledge sources in ontological semantics. We will then address the specific techniques of acquisition only for ontology (for the acquisition of the other static knowledge sources, see Nirenburg and Raskin 2001, Ch. 9).

2. Automating Knowledge Acquisition in Ontological Semantics

Knowledge-based applications involving natural language processing have traditionally carried the stigma of being too expensive to develop, difficult to scale up and to reuse as well as incapable of processing a broad range of inputs.[1] The opinion about the high price of development was due to the perceived necessity to acquire all knowledge manually, using highly-trained and, therefore, expensive, human acquirers. The difficulty in scaling up was believed to reflect the deficiencies in description breadth, or coverage of material, in the acquisition task for any realistic application. The all-too-real failure of knowledge-based processors on a broad range of inputs was attributed to the lack of depth (or, using our terminology, coarseness of the grain size) in the specification of world and language knowledge used by the meaning manipulation procedures.

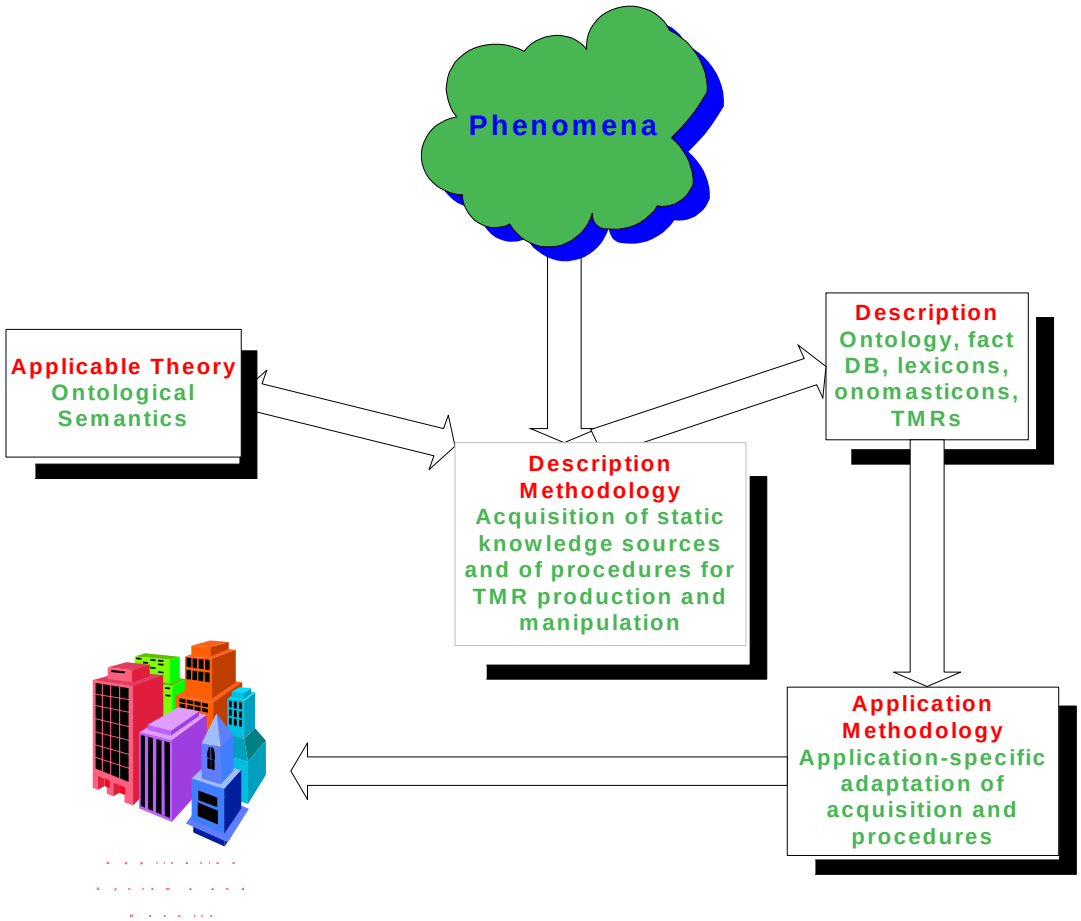


Figure 1. Interrelationships between theory, methodology, descriptions and applications in ontological semantics.

In the consecutive implementations of ontological semantics, the above problems have been progressively addressed. While we cannot claim to have completely eliminated the need for controlling the acquisition process by people, we are satisfied that ontological semantics uses about as much automation in the acquisition process as is practical within the state of the art in statistical methods of text processing and human-computer interaction. In addition to that, the acquisition methodology takes advantage of all and any possibilities for minimizing human acquisition effort and maximizing the automatic propagation of semantic information recorded earlier over newly acquired material, as applicable. The use of inheritance in ontology; of information extraction engines in acquiring facts for the Fact DB; as well as lexical rules and class-oriented syntactic dependency templates in the lexicon, are among the examples of such facilities. We have

1. "Today's state-of-the-art rule-based methods for natural language understanding provide good performance in limited applications for specific languages. However, the *manual* development of an understanding component using specific rules is costly as each application and language requires its own adaptation or, in the worst case, a completely new implementation. In order to address this cost issue, statistical modeling techniques are used in this work to replace the commonly-used hand-generated rules to convert the speech recognizer output into a semantic representation. The statistical models are derived from the automatic analyses of large corpora of utterances with their corresponding semantic representations. To port the semantic analyzer to different applications it is thus sufficient to train the component on the application- and language-specific data sets as compared to translating and adapting the rule-based grammar by hand" (Minker *et al*. 2000, xiv).

had numerous opportunities to port the resources of ontological semantics across applications, and found this task feasible and cost-effective, even within small projects. In the rest of this section, we briefly review the methodology of knowledge acquisition that has emerged over the years in ontological semantics.

Before a massive knowledge acquisition effort by teams of acquirers can start, there must be a preparatory step that includes, centrally, the specification of the formats and of the semantics of the knowledge sources, that is, the development of a theory. Once the theory is initially formulated (it is fully expected that the theory will be undergoing further development between implementations), the development of a toolkit for acquisition can start. The toolkit includes acquisition interfaces, statistical corpus processing tools, a set of text corpora, a set of machine-readable dictionaries (MRDs), a suite of pedagogical tools (knowledge source descriptions, an acquisition tutorial, a help facility) and a database management system to maintain the data acquired. In many ontology-related projects, the work on the knowledge specification format, on portability and on the acquisition interfaces becomes the focus of an entire enterprise (see, for instance, Ginsberg 1991, Genesereth and Fikes 1992, Gruber 1993, Farquhar *et al.* 1997 for a view from one particular research tradition). In such format-oriented efforts, it is not unusual to see descriptive coverage sufficient only for bootstrapping purposes. Ontological semantics fully recognizes the importance of fixed and rigorous formalisms as well as good human computer interaction practices. However, in the scheme of priorities, the content always remains the prime directive of an ontological semantic enterprise.

The preparatory step is in practice interleaved with the bootstrapping step of knowledge acquisition. Both steps test the expressive power of the formats and tools and seed the ontology and the lexicon in preparation for the massive acquisition step.

The bootstrapping of the ontology consists of:

- developing the specifications of the concepts at top levels of the ontological hierarchy, that is, the most general concepts;
- acquiring a rather detailed set of properties, the primitives in the representation system (for example, case roles, properties of physical objects, of events, etc.), because these will be used in the specifications of all the other ontological concepts;
- acquiring representative examples of ontological concepts that provide models (templates) for specification of additional concepts; and
- acquiring examples of ontological concepts that demonstrate how to use all the expressive means in ontology specification, including the use of different facets, of sets, the ways of specifying complex events, etc., also to be used as a model by the acquirers, though not at the level of an entire concept.

The bootstrapping of the lexicon for the recent implementations of ontological semantics involved creating entries exemplifying:

- all the known types of syntax-to-semantics mapping (linking);
- using every legal kind of ontological filler—from a concept to a literal to a numerical or abstract range;
- using multiple ontological concepts and non-propositional material, such as modalities or aspectual values, in the specification of a lexical entry;
- using such expressive means as sets, refsems and other special representation devices.

The main purpose of this work is to allow the acquirer during the massive acquisition step to use the example entries as templates instead of deciding on the representation scheme for a meaning from first principles. As usual, practical acquisition leads to the necessity of revising and

extending the set of such templates. This means that bootstrapping must be incremental, that is, one cannot expect for it to finish before the massive acquisition step. The preparatory step and bootstrapping are the responsibility of ontological semanticists who are also responsible for training acquirer teams and validating the results of massive knowledge acquisition. The complete set of types of work that ontological semanticists must do to facilitate a move from pure theory to an actual description includes:

- theory specification,
- acquisition tool design,
- resource collection,
- bootstrapping,
- management of acquisition teams:
  - training,
  - work process organization,
  - quality control.

At the step of massive knowledge acquisition, the acquirers use the results of the bootstrapping stage to add ontological concepts and lexicon entries to the knowledge base. It is important to understand that, in the acquisition environment of ontological semantics, acquirers do not manually record all the information that ends up in a static knowledge source unit—an ontological concept, a lexical entry or a fact. Following strict regulations, they attempt to minimally modify existing concepts and entries to produce new ones. Very typically, in the acquisition of an ontological concept, only a small subset of properties and property values are changed in a new definition compared to the definition of an ancestor or a sibling of a concept that is used as a starting template. Similarly, when acquiring a lexical entry, the most difficult part of the work is determining what concept(s) to use as the basis for the specification of the meaning of a lexical unit; the moment such a decision is made, the nature of the work becomes essentially the same as in ontological acquisition—determining which of the property values of the ontological concept to modify to fit the meaning. With respect to facts, the prescribed procedure is to use an information extraction system to fill ontologically inspired templates that become candidate entries in the fact database, so that the task of the acquirer is essentially just to check the consistency and validity of the resulting facts. At the end of the day, only a fraction of the information in the knowledge unit that is acquired at the massive acquisition step is recorded manually by the acquirer, thus imparting a rather high level of automation to the overall acquisition process.

The lists of candidate ontological concepts and lexicon entries to be acquired are included in the toolkit and are manipulated in prescribed ways. Acquirers take items off these lists for acquisition but as a result of at least some acquisition efforts, new candidates are also added to these lists. For example, when a leaf is added to an ontological hierarchy, it often becomes clear that a number of its conceptual siblings are worth acquiring. When a word of a particular class is given a lexicon entry, it is enticing to immediately add the definitions of all the other members of this class. The above mechanism of augmenting candidate lists can be called deductive, paradigmatic or domain-driven. The alternative mechanism would be inductive, syntagmatic and corpus(data)-driven and will involve adding words and phrases newly attested in a corpus to the list of lexicon acquisition candidates. Because the description of the meaning of some of such new words or phrases will require new concepts, the list of candidates for ontology acquisition can also be augmented inductively.

The results of the acquisition must be validated for breadth and depth of coverage as well as for accuracy. Breadth of coverage relates to the number of lexical entries, depth of coverage relates to the grain size of the description of each individual entry. The appropriate breadth of coverage is judged by the rate at which an ontological semantic application obtains inputs that are not attested in the lexicon. The depth of coverage is determined by the disambiguation needs and capabilities of an application that determine the minimum number of senses that a lexeme should have. In other words, the specification of meaning should not contain elements that cannot be used by application programs. Accuracy of lexical and ontological specification can be checked effectively only by using the acquired static knowledge sources in a practical application and analyzing the failures in such applications. Many

of these failures will have to be eliminated by tightening or relaxing constraints on the specification of the static knowledge sources.

3. Acquisition of Ontology

Acquisition of ontology involves the following basic tasks:

- determining whether a meaning is worth introducing a new concept;
- finding a place for the concept in the ontology, that is determining which of the existing concepts in the ontology would best serve as the parent or sibling of the newly acquired concept;
- specifying properties for the new concept, making sure that it is different from its parents, children and siblings not only on ONTOLOGY-SLOT properties (see Nirenburg and Raskin 2001, Section 7.1) but rather in a more contentful way, through other ontological properties.

The main considerations in deciding on whether a new concept is warranted are:

- the desired grain size of description; for instance, if in a question answering system we do not expect questions concerning a particular property or set of properties (or, which amounts to the same, are content with the system failing on such questions), then the corresponding property becomes too fine-grained for inclusion in the ontology; for example, in the CAMBIO/CREST implementation of ontological semantics for the application of question answering, in the domain of sports, no information was included about the regulation sizes and weights of the balls used in various games—baseball, basketball, etc., for the reason that we did not expect such questions to be asked of the system;
- the perception of whether a meaning is generic and language-independent (and, therefore, should be listed in the ontology) or a language-specific "fluctuation" of some basic meaning (and should, therefore be described in the lexicon for the language in question);
- the perception of whether a meaning is that of a concept (a type, a class of entities, a meaning, a *significatum* 'signified,' a "variable") or a fact (an instance, a token, an individual, a reference, a *denotatum*, a "constant"); for example, US-PRESIDENT is a concept, while *John Kennedy* is the name (stored in the onomasticon) of an instance of US-PRESIDENT, namely, US-PRESIDENT-35; CORPORATION is a concept; *Ford Motor Company* is the name of an instance of corporation; FORD-FOCUS, however, is a concept, a child of CAR-MAKE and CAR-MODEL; my cousin Phyllis's Ford Focus is an instance of the concept FORD-FOCUS; incidentally, if she calls her car Preston, this will probably not be general or useful enough knowledge to warrant being included in the onomasticon of an ontological semantic application;
- the perception of when the analysis and other meaning processing procedures would fail if particular concepts are not present in the ontology, e.g., the judgment that a particular disambiguation instance cannot be handled using dynamic selectional restrictions (see Nirenburg and Raskin 2001, Section 8.3.1).

With respect to language specificity, consider the example of the German *Schimmel*, 'white horse." There seems to be no reason to introduce an ontological concept for *white horse*, as this meaning is easily described in the lexicon by including in the SEM-STRUC field of the corresponding entry an instance of HORSE, with the property of COLOR constrained to WHITE. Also if this concept is introduced, the considerations of symmetry would lead to suggesting as many siblings for this concept as there are colors in the

system applicable to horses.

To generalize further, it is a useful rule of thumb in ontology acquisition not to add an ontological concept if it differs from its parent only in the fillers of some of its attributes because this is precisely the typical action involved in specifying a lexical meaning in the lexicon on the basis of a concept. It is a vote for introducing a new ontological concept if, in the corpus-driven mode of knowledge acquisition, no way can be found of relating a candidate lexeme or candidate sense of an attested lexeme to an existing concept or concepts by constraining some or all of its/their property values.

In other words, it is best to introduce new ontological concepts in such a way that they differ from their parents in the inventory of properties, not only in value sets on the properties that they share. Barring that, if the difference between a concept and its parent is in the values of relations other than the children of ONTOLOGY-SLOT (e.g., IS-A or INSTANCES) then a new concept may also be warranted. Barring that, in turn, if there are differences between a concept and its ancestor on more than one attribute, a new concept should be favorably considered. Finally, if the constraint on an attribute in the parent is an entire set of legal fillers or if a relation has as its filler a generic constraint 'OR EVENT OBJECT,' and the child introduces stricter constraints, one may consider a new ontological concept. Experience in acquisition for ontological semantics shows that applying these rules can be learned relatively reliably, and compliance with them is easy to check.

The task of finding the most appropriate place to 'hook' a concept in the ontology is also complicated. Let us assume that we have already determined, using the above criteria, that TEACH deserves to be a new ontological concept. The next task is to find one or more appropriate parents or siblings for the concept. Acquirers use a mixture of clues for placing this concept in the ontological hierarchy. Experienced acquirers, well familiar with many branches of the ontological hierarchy, may think of an appropriate place or two right off the top of their heads, based on clues inherent in concept names. In some cases, this actually does save time. The reliance on name strings is, however, dangerous, because the names are elements of the ontological metalanguage and have a semantics of their own that is different from the lexical meaning of the English words that they may resemble. Therefore, when this clue is used, the acquirer must carefully read the definition of the concept and scan its properties and values to determine its actual meaning. The more reliable, though slower, procedure involves playing a version of the game of twenty questions—comparing the intended meaning of the candidate concept with concepts at the top of the ontological hierarchy and then descending this hierarchy to find the most appropriate match.

At the very top level of the ontological hierarchy of the CAMBIO/CREST implementation of the ontology (Figure 2), the choice is relatively easy: TEACH is an EVENT. There are three

Figure 2. The top level of the ontology in all the implementations of ontological semantics.

types of events (Figure 3). Let us check whether

Figure 3. The top level of the event hierarchy.

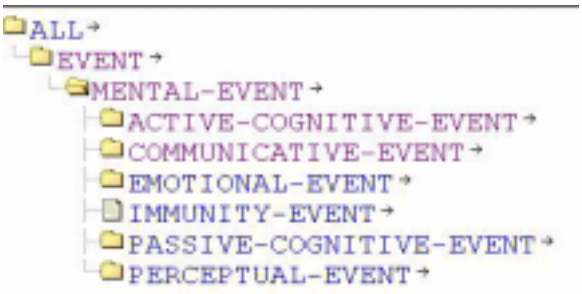


Figure 4. Some mental events.

TEACH fits into the mental event branch (Figure 4). Out of all the subclasses of mental event,

COMMUNICATIVE-EVENT (Figure 5) seems to be the most suitable. COMMUNICATIVE-EVENT has another parent, SOCIAL-EVENT (Figure 6). A quick check shows that no other children of SOCIAL-EVENT are appropriate to serve as parents of TEACH.

We need to check now whether the third child of event, PHYSICAL-EVENT or any of its descendants can also serve as a parent of TEACH. On inspection of the concept names of children of PHYSICAL-EVENT (see Figure 7), we may wish to check

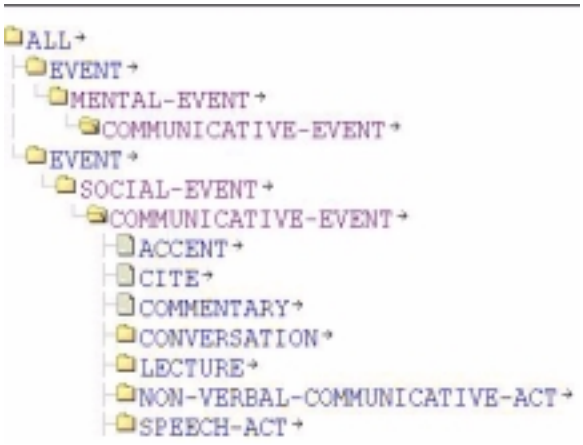


Figure 5. Multiple inheritance of communicative-event and some types of communicative events.

Figure 6. Some types of social events.

whether LIVING-EVENT has children that could be siblings of TEACH because the semantics of the concept name, *living event*, may suggest that it is appropriate. Inspection (see Figure 8) quickly demonstrates, however, that the name is, in fact, misleading in this case, as the subclasses of LIVING-EVENT do not seem to be appropriate as siblings or parents of TEACH (REAR-OFFSPRING also turns out to be a false lead). At this point, the decision can be safely made: to add TEACH as a

child of COMMUNICATIVE-EVENT.

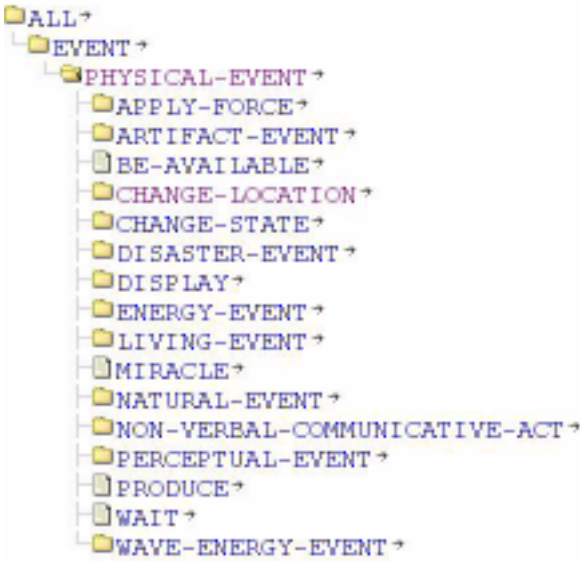


Figure 7. Some types of physical events.

The next task is to describe its meaning, that is, to check the fillers of the properties it inherits from COMMUNICATIVE-EVENT (1).

(1)

communicative-event	
agent	sem animal
theme	sem OR event object
instrument	default OR communication-device natural-language
destination	sem OR animal social-event
effect	sem OR event object
precondition	sem OR event object

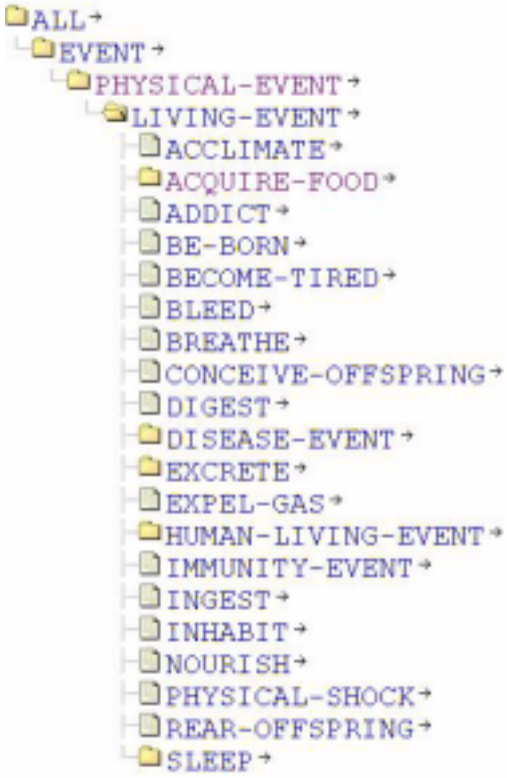


Figure 8. Some types of living events.

TEACH does, indeed, inherit all the above properties. The actual constraints (fillers) for them are shown here partially as (2). Besides the properties in (3), TEACH has an additional property, HAS-PARTS, which establishes it as a complex event (descriptions of the components of TEACH see also in Nirenburg and Raskin 2001, Section 7.1.5). The syntax of HAS-PARTS includes repeat-until loops and conditionals and thus makes the entry for a complex event essentially an algorithm that describes typical successful complex events as a time-ordered sequences of component events. References to components include the statement of the complex event of which they are a component as well as an index. For example, TEACH-KNOW-A means that this is a KNOW event which is not free-standing (there is, in fact, a know event in the ontology) but rather one of the three specially constrained "ontological instances" of KNOW that can be found in the specification of the complex event TEACH.

(2)

teach

is-a	value	communicative-event
agent	sem	human
	default	teacher
theme	sem	knowledge
destination	sem	human
	default	student
precondition	default	(teach-know-a teach-know-b)
effect	default	teach-know-c
has-parts	value	(teach-describe **repeat** (teach-request-info teach-answer) **until** teach-know-c)

Finding the appropriate fillers, if any, for the various facets of a property is a separate acquisition task. For example, if there is a candidate filler that is strongly implied when no explicit reference to it is present in the input text, it should be listed in the DEFAULT facet of the property. Thus, for the AGENT property of TEACH, the default facet will be filled with TEACHER, because in a sentence like *Math was not taught well in his high school*, the implied AGENT of TEACH is clearly a subset of instances of the concept TEACHER.

Of course, one example of this kind does not prove the point, but when combined with the acquirer's knowledge of the world, it supports a useful rule of thumb. The acquirer also knows that any (adult) human can at times perform the social role of teacher, e.g., parents teaching their teenage children to drive. Therefore, one should expect many inputs in which the constraint on the agent of TEACH is more relaxed than the one in the DEFAULT facet. This most commonly occurring constraint is recorded in the SEM facet of the property. If an input like the sentence *Gorillas teach their offspring essential survival skills* can be expected in an application system, the constraint on the AGENT should be further relaxed to ANIMATE on the RELAXABLE-TO facet (cf. Nirenburg and Raskin 2001, Section 8.2.3). However, any attempt to relax the constraint on this property further, for example, in order to accommodate the sentence *Misfortune taught him a good lesson* should be denied, because the property AGENT in ontological semantics is constrained to HUMAN or FORCE, and rather than coercing misfortune into FORCE, the meaning of this sentence should be represented, roughly, as that of the sentence *He learned a good lesson as a result of a misfortune*, thus reducing the different sense of *teach* in this sentence to a metonymic shift on the appropriate sense of *learn*.

The above procedure of finding the best place to connect a concept into the ontology is not as straightforward as may be deduced from the example. The procedure is predicated on the assumption that the constraints in the ontology become monotonically and progressively stricter as one descends the hierarchy. This was, indeed, the situation with TEACH on every one of the properties inherited. It is legal, however, for constraints in a child to be, in fact, looser than those in an ancestor. In fact, an ancestor may have inheritance on a property completely blocked using the special filler NOTHING, but a child could revert to a contentful filler. This state of affairs makes it dangerous to stop the search for the most appropriate place to include a new concept the moment some constraints become narrower than those expected in this concept. However, in practice, the monotonicity property holds in a much greater majority of cases.

Ontology acquisition may involve not only manipulation of property fillers. Sometimes (preferably, as seldom as possible, though), it is necessary to add a new property to the system. This might be necessary when a concept cannot be described using the extant inventory of properties; this typically, though not exclusively, happens when describing new subject domains. If indeed new properties must be introduced, it is highly desirable that they contain as many concepts as possible in the domain property of their definition. For example, when extending the Mikrokosmos implementation of ontological semantics to accommodate the subject domain of sports in the CAMBIO/CREST implementation, it became necessary to introduce the literal attribute COMPETITION-STAGE, whose domain property was filled with SPORTS-RESULT (a central concept for the domain) and whose range was filled with the useful constants CLASSIFICATION, FINAL, PRELIMINARY, QUALIFICATION, QUARTER-FINAL, RANKING, REPECHAGE, ROUND-

OF-16, ROUND-OF-32, ROUND-OF-64 and SEMI-FINAL. The nature of the application dictates this grain size—we do not need to know any information about the above constants than just their names and the corresponding words or phrases in the languages processed by the system.

References

Farquhar, A., R. Fikes, and J. Rice 1997. Tools for Assembling Modular Ontologies in Ontolingua, KSL-97-03, Knowledge Systems Laboratory, Department of Computer Science, Stanford University, Stanford, CA. Also in: **Proceedings of AAAI-97**, 436-441.

Genesereth, M. R., and R. E. Fikes 1992. Knowledge Interchange Format, Version 3.0 Reference Manual. Technical Report Logic-92-1, Computer Science Department, Stanford University, Stanford, CA.

Ginsberg, M. 1991. Knowledge Interchange Format: the Kif of Death. *AI Magazine*, 5-63.163.

Gruber, T. R. 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition* 5, 199-220.

Minker, W., A. Waibel, and J. Mariani 1999. **Stochastically-Based Semantic Analysis**. Boston-Dordrecht-London: Kluwer.

Nirenburg, S., and V. Raskin 2001. **Ontological Semantics**. Forthcoming. Pre-publication draft at http://crl.nmcu.edu/ Staff.pages/Technical/sergei/book/index-book.html.

A Rapidly Acquired Domain Model Derived from Markup Structure

Udo Kruschwitz
Department of Computer Science
University of Essex
Wivenhoe Park
Colchester
CO4 3SQ
United Kingdom
udo@essex.ac.uk

Abstract

Powerful and sophisticated state-of-the-art Web search engines have emerged in the last few years. But finding information on the Web can still cause a lot of headache. Part of the reason is the massive growth of available online data. But even in small domains a simple query can easily result in hundreds of matching documents being returned. This paper describes how document markup can be used to build a domain model quickly and fully automatically. This model can be incorporated in a Web search system using simple dialogues to help a user find the right subset of documents. For domains like intranets and local Web sites this approach has the advantage that it is computationally not too expensive and that it allows one to build a data driven set of hierarchies capturing the peculiarities of data stored at the site.

1 Overview

Finding Web pages relevant to a particular information need has become so much easier with the emergence of sophisticated Web search engines. Nothing seems simpler than using a Web search engine to locate up-to-date scientific papers, people, institutions, or a plumber with an emergency service in Colchester. Not even the document format seems to matter anymore. Not so long ago only some specialized search engines like ResearchIndex[1] (Lawrence et al., 1999) would have coped with formats like Postscript or PDF, now it seems common practice to retrieve PDF files alongside pure text documents. However, a common phenomenon is that users find it difficult to express their actual information need as a query. Queries submitted via search interfaces are hardly ever longer than two or three words (Silverstein et al., 1998). It seems difficult to select the most relevant of all documents from what was estimated back in 1999 to be a pool of 800 million pages on the publicly indexable web (Lawrence and Giles, 1999). How can a two-word query describe precisely what the user is after? This is true not just for searches on the entire Web but also in limited domains such as intranets and company Web sites.

What can be done? One solution is to present the user some options of how to modify the query to get to a very specific set of documents. But in order to work well, it would require some knowledge about the domain. Unlike in a number of approaches for general Web search that are based on the evaluation of the results returned by a search engine we concentrate on smaller domains and try to uncover the structure of the entire data collection independent of the user queries. The result is a world model that is built in a data driven way. Such domain knowledge can be utilized in a dialogue system helping the user navigate to the set of most relevant documents. It could also be applied as a browsing tool to get a better picture of what data can be found in the collection. This is

[1] http://www.researchindex.com

similar to using a classified directory or a manually maintained Web directory.

To pick an actual example, we will look at the user query "union", which is a frequent query according to the log files of our sample domain, the University of Essex. The user might not be aware of the fact that the query is highly ambiguous. In fact, there are Web pages in that domain presenting information about the *trade union*, *students union* and a *christian union*. Not just that, but there is a number of pages devoted to *discriminated unions*. That could well be the pages the user expects as an answer if that user is a student currently writing an assignment in the *Distributed Computing* module. A domain model that reflects these possible interpretations could be a useful aid to inform the user of what types of documents there are (or are not) in the space of possible answers. At the same time it would guide the user in the process of narrowing down the search by explicitly offering query terms.

In (Kruschwitz, 2001b) we presented a way of finding conceptual information in hypertext documents based on markup structure. Here we describe how these concepts can be automatically organized as a set of hierarchies that can easily be applied in search applications. In a separate paper we investigate how collaborative filtering may be used to improve this model.

The remainder of this paper is structured as follows. First we will look at some related work (section 2). Section 3 explains in more detail how markup is used to build the domain model and how the construction can be performed automatically. We pick some examples from our sample domain to give the reader a better picture of the type of models we get (section 4). We briefly discuss in section 5 how the model can be incorporated in a Web search application. Finally, there is some discussion of current issues (section 6), and some conclusions are given in the last section.

2 Related Work

Most of the work presented here deals with extracting conceptual information from documents based on a number of different techniques, such as clustering, classification etc. Structural information is being used in a number of approaches. However, the reviewed work does either not go as far as to use this knowledge to build a domain model, or the techniques are significantly different to what we present.

Research on the construction of concept hierarchies has so far focussed mainly on word co-occurrences or linguistic information. The motivation for the work presented in (Sanderson and Croft, 1999) is to automatically organize documents in a conceptual structure based on keywords found in the documents. Unlike in manually constructed subject hierarchies, this structure is customized to the set of documents. The presented solution is a hierarchical organization of concepts automatically derived from a set of retrieved documents. General concepts are placed in the higher levels of the hierarchies, more specific ones further down. The construction is based on pairs of terms occurring in the same document, independent of the context they were found in. A concept x is considered more general than a concept y, if the documents which y occurs in are a subset of documents containing x.

Closely related is the *Paraphrase Search Assistant* (Anick and Tipirneni, 1999). The focus is again on how to help the user find the set of most relevant documents among a large set of documents retrieved for a query. The user is presented a list of relevant terms, each of which expands to a list of compounds (containing the selected term) which can then be used to constrain the query. The assumption is the "tendency for key domain concepts within result sets to participate in families of semantically related lexical compounds". A remarkable aspect in this context is that query logs have shown that users actually make use of the presented terms for query refinement.

An interesting evaluation of different types of Web search has shown that some guidance to help the user in reformulating a query can significantly improve the relevance of the retrieved documents compared to standard Web search (Bruza et al., 2000). Part of the study was the comparison of Web search using *Google* and a search via the *Hyperindex Browser*, an interactive tool which passes the user query to a search engine and displays a list of linguistic phrases found in the top matching documents returned by the search engine. Those phrases can be selected by the user to constrain the query.

A different strand of work is layout analysis. One motivation is to convert printed collections of documents into their corresponding tagged electronic version. Autotag (Taghva et al., 1994) is a system that aims at processing many types of documents based mainly on heuristic rules. It performs a physical followed by a logical analysis. The intention is "to capture every conceivable object that may add to a document's presentation". However, it does not capture the semantic content of the documents.

Finding the logical structure of a document without knowing the actual documents or styles used in them is the focus of (Summers, 1998). This is a two-step process, *segmentation* and *classification*. The segmentation process is primarily based on layout, contour and font shape information. The classification step is based on the comparison of the text segments to predefined structure prototypes, mainly by means of geometric cues. The only linguistic knowledge used in the described prototypes is the identification of typically appearing or non-appearing symbols. If some information about the *style* of the documents is known, then this can be incorporated in the processing steps as well.

Another relevant field of related work is document clustering. Like information retrieval approaches, clustering relies very much on large amounts of data. IBM's *TaxGen* text mining project aims at the automatic generation of a taxonomy for a large corpus of unstructured news wire documents (Müller et al., 1999). A hierarchical clustering algorithm first builds the bottom clusters and then works its way upward forming higher-level clusters by grouping together related clusters. The clustering is based purely on linguistic elements in the documents, e.g. co-occurrences of words (*lexical affinities*) or names of people, organizations, locations, domain terms and other significant words and phrase from the text (*linguistic features*).

Keyphind is an example of conceptually indexing a document collection (Gutwin et al., 1999). Machine learning techniques are applied in order to extract *keyphrases* from documents in the context of browsing digital libraries. This comes close to our idea of imposing a structure on the collection by extracting "important" phrases from each document, but in *Keyphind* the documents are much longer and furthermore a manually tagged training corpus is needed to build the classifier. *Extractor* is a similar system for extracting keyphrases using supervised learning (Turney, 1999; Turney, 2000).

Clustering is also being used for *concept-based* relevance feedback for Web information retrieval (Chang and Hsu, 1999). Following a user query the retrieved documents are organized into conceptual groups. Unlike in our approach this structure is not extracted for the indexed domain but for the search results. Another example is *Grouper*, an interface to a meta-search engine which dynamically groups the search results into clusters labeled by phrases extracted from the snippets that the search engines returned with the retrieved documents (Zamir and Etzioni, 1999). These applications are quite different to traditional *offline* clustering approaches where the whole document collection is clustered, rather than a number of retrieved documents. The motivation for Zamir *et al.* is that post-retrieval document clustering has been shown to produce superior results. Moreover, clustering a few thousand documents is much quicker than doing this for millions of documents (the number of documents you would find in *general* Web search applications). However, for our problem post-retrieval clustering would miss the point since we want to build a domain model for the entire document collection.

A clustering approach for Web pages that uses words as well as incoming and outgoing links is presented in (Modha and Spangler, 2000). Starting from a set of results returned by a search engine, the presented approach clusters the resulting documents. The key is to consider not just those documents similar which share a word but also those that have the same outgoing or incoming links. All these features are represented in a single vector. More specifically, six information nuggets are encoded in the vector: summary, breakthrough, review, keywords, citation, and reference. It is reported, that compared to the HITS algorithm (Kleinberg, 1998) *reference* and *breakthrough* are conceptually analogous to *authorities*, and *citation* and *review* are analogous to *hubs*.

Document classification is closely related to clustering, the difference to clustering is that predefined categories exist, normally manually constructed, and after a training phase new documents can be classified on-the-fly like in (Chen and Dumais, 2000). Northernlight[2] is an example of a search engine where the results are grouped in dynamically created categories, so called *Custom Search Folders* which help a user to narrow down the search in case of too many matches. The search covers the whole Internet and thus has not just access to much more data than what we assume, but user queries are much more varied and thus domain independent than searches in intranets or on single Web sites.

Automatic classification of Web pages is the focus of (Attardi et al., 1999). They introduce *categorization by context* which exploits information surrounding the links pointing to a document in order to classify that document. It is seen in contrast to *categorization by content* which relies on textual information found in a document. The basic assumptions are: (1) a Web page which refers to a document must contain enough hints about its content to induce someone to read it, and (2) such hints are sufficient to classify the document referred to. Hints are mainly anchors, but also structural information found in HTML documents like headings, titles etc. However, building the profiles of the classifications are left as open issues, and the two options presented are to either build them by hand or use some learning techniques.

Ontologies and customized versions of existing language resources like *WordNet* (Miller, 1990) are being successfully employed to search product catalogues and other document collections held in relational databases (Guarino et al., 1999; Flank, 1998). Part of that research is the actual construction of ontologies (Craven et al., 1998). The cost to create the resources can be enormous and it is difficult to apply these solutions to other domains where the document structure is not known in advance.

Internal structure is being incorporated more and more in standard Web search engines. A prominent example is *Google*[3] whose development was inspired by the idea of improving the quality of search as opposed to efficiency (Brin and Page, 1998). It makes use of both link structure and anchor text. Each page gets a ranking value depending on how many other pages reference to it. Moreover the system associates Web pages not just with the text in it but also with the text found in anchors linking to this page from elsewhere. Furthermore, *Google* keeps track of some visual presentation details like font size to give for example heavier weight to words which appear in a larger font than the rest of the document. But apart from the sophisticated ranking function there is no way to help a user with too many matches or ambiguous queries.

[2] `http://www.northernlight.com`
[3] `http://www.google.com`

3 A Domain Model Based on Concepts

One of the main motivations for automatically acquiring a domain model is the inadequacy of a domain independent knowledge source like *WordNet* in search applications for limited domains. This is true for two reasons. First of all, a number of relations that may be inherent in the documents of the domain will not be present in the knowledge source, i.e. *world knowledge* rather than linguistic information (cf. the *union* example). On the other hand, a domain independent model will have a large number of links entirely irrelevant for a specific domain. To stick to the *union* example, *WordNet* senses two (*the United States during the Civil War*), three (*the act of pairing a male and female for reproductive purposes*) and five (*state of being husband and wife*) are not relevant in our domain at all, nor are some of the other senses. In addition to these problems, the domain may change over time which makes it desirable to construct a model on demand exploiting the actual data.

Extracting *concept terms* based on the structure detected in documents and relations between documents is discussed in (Kruschwitz, 2001b). The main motivations are summarized here. The basic idea is to select keywords for documents which are found in more than one of the following markup contexts: *meta* information, document *headings*, document *title* or *emphasised* parts of the document (e.g. bold or italics font). Any other text is ignored altogether. This allows us to separate important terms from less important ones. It is largely language independent because

it exploits markup more than anything else. Furthermore it is computationally very cheap. The extracted terms are called concept terms (or just concepts). And it is this set of concepts that we use to automatically create a domain model. The reasons are threefold. First of all, it was already pointed out that we consider concept terms more important than ordinary keywords. Secondly, we can define a relation between concept terms. Basically, we consider two concepts to be *related*, if there is a document for which both terms were found to be concept terms. Thirdly, the number of concept terms is just a small fraction of the number of all keywords selected for a document (roughly 2% in the sample domain). If this was not the case, it would hardly be feasible to investigate *related* concepts.

In the following we explain this in a bit more detail. The data sources we are interested in (e.g. Web pages) are *semistructured*, that means the documents do have some structure. This structure can come in a variety of forms, but most of it can be abstracted as some sort of *markup context*. Informally speaking, this is some explicit convention of highlighting or marking up a part of a document in a particular way distinguishing it from other parts. Certain HTML tags are just one example. Based on markup contexts we define concepts.

Definition 1 *An index term c is a concept term (or simply concept) of type n for document d, written as* $concept_n(c, d)$*, if c occurs in at least n different markup contexts in d.*

In addition to that we write $concept(c, d)$ if $concept_n(c, d)$ and $n \geq 2$.

Definition 2 *Two concepts* c_1 *and* c_2 *are related concepts of type n for document* d*, written as* $rel\_concepts_n(c_1, c_2, d)$*, if* c_1 *and* c_2 *are concepts of type n for document d, i.e.:*

$$rel\_concepts_n(c_1, c_2, d) \Leftrightarrow (concept_n(c_1, d) \wedge concept_n(c_2, d)) \quad (1)$$

We also define:

$$rel\_concepts_n(c_1, c_2) \Leftrightarrow (\exists d\, doc(d) \wedge rel\_concepts_n(c_1, c_2, d)) \quad (2)$$

$$rel\_concepts(c_1, c_2, d) \Leftrightarrow (rel\_concepts_n(c_1, c_2, d) \wedge n \geq 2) \quad (3)$$

$$rel\_concepts(c_1, c_2) \Leftrightarrow (\exists d\, doc(d) \wedge rel\_concepts(c_1, c_2, d)) \quad (4)$$

Moreover, we define a notion of importance of a concept.

Definition 3 *If* c_1 *and* c_2 *are two concepts for a document* d*, so that* $concept(c_1, d)$ *and* $concept(c_2, d)$*, then* c_1 *is more important than* c_2 *in document* d*, if there are m and n, such that* $m > n$ *and* $concept_m(c_1, d)$ *and* $concept_n(c_2, d)$ *and there is no* $i \geq m$ *so that* $concept_i(c_2, d)$*.*

Concept c_1 *is more important than concept* c_2*, if* c_1 *is more important than* c_2 *for all documents* d *or* c_2 *is not a concept for* d *at all.*

The definitions and examples should make it clear, that we are not interested in uncovering the type of relation that exists between two concepts. It would be too ambitious to do this in an automated fashion, in particular with limited domains where data sparsity can be a serious problem. However, what we do get is a number of relations that can guide a user in a search application. Deciding on whether a relation is useful or not is now the task of the user. That relation could be one of *hypernymy* or *synonymy*, but more likely one not linguistically motivated, rather *world knowledge* in the actual word sense, as it is the case for the two terms *union* and *christian* or the pair *union* and *bar*.

The world model we construct is a set of simple hierarchies, each with a concept as a root node. The construction is straightforward. Each hierarchy is built by going through hypothetical dialogues with the user. Starting with a concept as the original query, the task is to constrain this query in order to get more specific matches. All related concepts (in the sense of equations 2 and 4) are considered as possible terms to constrain the query. Now we create a branch between the root and one of its related concepts if there are documents in the collection matching the new query consisting of both terms. For the purpose of

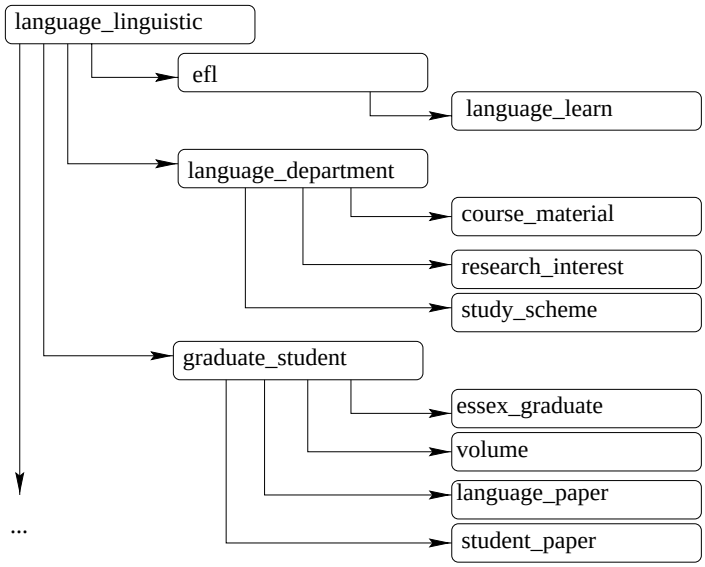


Figure 1: Concept tree for the compound *language_linguistic*

matching queries against documents we treat concepts as ordinary keywords (otherwise the resulting hierarchies would be too simplistic). Moreover, each query is processed as a Boolean query for which exactly those documents match, which contain all query terms.

The same steps are performed for each of the new nodes recursively, i.e. for each concept c for which a document d exists with $concept(c, d)$ do:

1. Create a root node labeled $node(c)$
2. For each $node(c')$ in the tree and concept c'' with $rel\_concepts(c', c'')$ create a node labeled $node(c'')$ and a directed edge from $node(c')$ to $node(c'')$ if the query consisting of all index terms collected by traversing through the graph from the root to $node(c'')$ matches a non-empty document set.
3. Perform step 2 until there are no unexplored nodes left.

Hence, once we have constructed a tree for a particular concept, then a path from the root node to any other node in this tree can be seen as adding constraints to a query. In other words, any node represents a cluster of documents in the collection, all those documents that match each of the concepts associated with the nodes passed by traversing the graph from the root to the current node.

Practically, one constructs a world model for each type of concepts, i.e. all concepts of type 2, 3, etc. The higher the number the more important the concepts and relations between concepts. Our example application distinguishes two models, one based on type-2-concepts and the other one based on type-3-concepts. The second one is obviously smaller than the first one, since by definition any type-3-concept is also a type-2-concept. The online search system which offers the user concept terms to constrain the current query checks first the *most important* concepts, i.e. type 3 in our example. If not successful, *less important* concepts are investigated.

4 Examples

When looking at the examples it must be kept in mind that no manual modification was applied to the resulting concept structure.

In order to keep the size of the index tables manageable, we only consider single words or two-word compounds. This could easily be extended to compounds consisting of more than two words. Moreover, we apply base form reduction in the examples (e.g. *linguistics* is reduced to *linguistic*). All the examples are constructed using our sample domain only considering concepts of type 3, i.e. terms that were found in at least three markup contexts in a document. When we talk

about a concept to be *more relevant* than another one, we mean more documents contain this term as a concept. The same applies to sets of concepts.

A frequent query submitted to the search engine of our sample domain is "language and linguistics". Figure 1 displays part of the tree created with *language_linguistic* as the root node. Not all concepts related to *language_linguistic* are displayed, only the three most relevant ones. The most relevant related concept is *efl* which stands for "English as a Foreign Language". There is only one more concept related to both of the terms *language_linguistic* and *efl*, which is *language_learn* (i.e. "language learning"). That is the end of this path. The construction process of the world model determined that there was no related concept that could be added to the three others which would describe a non-empty cluster of documents.

The somehow peculiar relations between concepts *graduate_student*, *volume*, *language_paper* etc. is actually something that was extracted from a collection of documents all entitled "Graduate Students Papers in Language and Linguistics, Volume ...". This looks like noise in the domain model. Ways around this problems are discussed in (Fasli and Kruschwitz, 2001). Essentially, we adopt ideas from collaborative filtering to improve the model by recording the users' selections and choices in each step of the dialogue in a Web search application.

Figure 2 depicts the top ten related concepts for *language* ("languages" is a query slightly less frequently submitted to the search engine than "language and linguistics"). Unlike in the first example we do not display the entire tree, only the first level. Naturally, the tree is much bigger than the one in the earlier example. Note, that the related concepts are not necessarily related linguistically. Not surprisingly so. The third most important match is *sulawesi*. It turns out that this term does not refer to an Indonesian island, but to some framework in the context of wearable devices that deals with natural language understanding. This is another indicator that despite the fact that an obviously existing relation between the two terms *language* and *sulawesi* was discovered in the automatic knowledge extraction process, it might not be of any relevance in a search application. But this fact can only be established by observing real users' behaviour, something that goes beyond the scope of this paper.

5 Applying the Model

Sophisticated search engines like *Google* have proven to work well on huge amounts of data. But even for smaller domains this technology will nicely handle a large proportion of the user queries. In our domain a user who wants information about "Computer Science" will most likely be happy with the homepage of the Department of Computer Science, even though there are hundreds of pages matching the query. To pick another example we could come back to the "union" query. The most relevant matches (information about the students union in our domain) will probably be ranked highest in a list of matches if a good standard search engine is being used. In both examples the user will not be interested in entering a dialogue that constrains the query. It would be best to just display the results of the standard search engine. As a consequence, our approach does not aim at abandoning established search technology but at dealing with the remaining percentage of queries that cannot be handled in the straightforward fashion just outlined. We pass the user query to the standard search engine, and in parallel we evaluate the query against the world model. Then we merge both results. As soon as the standard search is finished, the top ranked results can be displayed. Proposals for query modification are then presented alongside the matches. In the "union" example that would mean that most user requests can be satisfied by a single search engine call. But if the expected documents are not among the most highly ranked ones, the user can engage in a dialogue with the system by selecting an appropriate term from the list of proposed concepts to constrain the query (e.g. select *trade_union*). The dialogue is explained in some more detail in (Kruschwitz, 2001a).

But the model can as well be used if the search is too specific not retrieving any matching documents. For example a user asking for "course material for language learning" might not find anything, either because there are indeed no matches

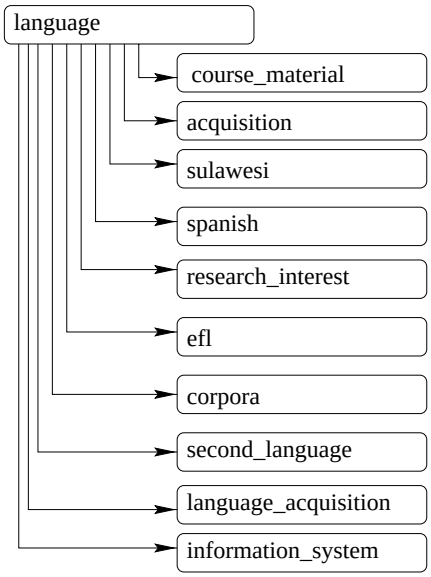


Figure 2: Concept tree for the term *language*

or because the domain was re-indexed and the pages have disappeared. Figure 1 does however give a nice picture of what structure was detected in the sources, and an obvious dialogue step would be to allow the user a relaxation of the query by searching for *language_linguistic* instead. Note however, that we found data sparsity to be a much less significant problem than queries with a large number of matching documents, despite the fairly small size of the sample domain.

6 Current Research

The approach has been shown to work well in the sample domain.

One of the problems still to be solved is that very long documents are treated just like short ones. That introduces a number of *related* concepts that should better be ignored. It is not quite clear yet whether breaking longer documents into smaller sections would help. That will at least reduce the number of extracted concepts.

In the examples we restricted the number of related concepts. That seems like a good idea. Our threshold is ten, i.e. the tree that is built for each concept will have at most ten branches to the next level down. The most relevant related concepts are selected. That results in a usable domain model. A different approach is to assign some weight to each branch and construct a much larger model.

So far we ignored one important observation altogether. It is best displayed with the help of an example. Processing our sample domain reveals that there is a concept *isabelle* which lists among others the related concepts *proof*, *theorem_prover* and *theorem*. A second set of related concepts in the same domain is *art_history*, *gallery* and *undergraduate_study*. Both of these sets are good illustrations of the intuitive meaning of what a related concept is. A search for either *isabelle* or *art_history* on the Web site results in a large number of matches. However, a query expansion using the related terms for *isabelle* result in exactly the same set of retrieved documents, while the concepts related to *art_history* are truly constraining the set of retrieved Web pages. We call the second set of concepts *dissimilar related concepts*, i.e.:

Definition 4 *Two concepts c_1 and c_2 are dissimilar related concepts of type n if there is a concept c_3, such that $rel\_concepts_n(c_1, c_3)$ and $rel\_concepts_n(c_2, c_3)$ hold, but not $rel\_concepts_n(c_1, c_2)$.*

Moreover, we observe that the "related concepts" relations in equations 1 and 3 describe equivalence relations over the domain of index terms and fixed values for type and document. The proof is trivial, since it is obvious that the

relations are reflexive and symmetric as well as transitive for a given type n and a document d, and therefore they form equivalence relations. However, this is not true for the relations defined in equations 2 and 4. As a result of this we modify the world model in the following way: each node is not a single concept but a set of concepts. Analogous to *synsets* in *WordNet* this means that each time a new node in a tree is created we will have to make sure that the new node contains a concept c together with all those concepts related to c which are not *dissimilar*. In the example, that will cause *isabelle*, *theorem*, *proof*, and *theorem_prover* to be listed in a single node.

Evaluation is currently the most important task. As mentioned earlier, the examples have not been manually modified. However, this still gives only anecdotal evidence about the quality and usability of the extracted structures. Evaluating the model as such seems not the very best idea. It must be seen as part of a search system to prove its usability. That is a complex task since it involves a number of dimensions, in particular if a standard search system is to be compared with a dialogue system. We evaluate the presented techniques as part of our dialogue driven Web search system. As a reference query corpus we use all queries submitted to the locally installed search engine between January and March this year. What can be said already is that most of the frequently submitted queries trigger the system to initiate a dialogue, because the queries match a large number of documents. We intend to publish evaluation results by the end of the year.

Finally, we want to give some figures for the current size of our sample domain:

	Number
Indexed Pages	27,333
Distinct type-3-concepts	4,383
Distinct type-2-concepts	17,876
Distinct keywords	823,116
Pairs of related type-3-concepts	45,254
Pairs of related type-2-concepts	1,282,610

The number of distinct concepts represents the number of root nodes in the model(s). However, there are some concepts with no related concepts at all (type 2 about 1,5% and type 3 about 9%). The hierarchy for those terms consists of a root node only.

The large number of distinct keyword indexes is partially caused by the fact that we do not try to find *real* compounds only, but we overgenerate (e.g. considering every sequence of nouns to be a compound) and then delete all those that fall below a certain threshold. This threshold is one, which means we only ignore those compound terms that are found only once in the entire collection.

7 Conclusions

We presented a method for building a domain model that exploits the internal document structure as it can be found in hypertext documents. The key assumption is that important keywords in a document do not necessarily turn up very frequently but that they can be found in a number of different markup contexts. Based on this we distinguish conceptual information from ordinary keywords. The conceptual information is used to automatically build a domain model which reflects the relations between concept terms. This can easily be deployed in a search application, for example to offer a user choices for constraining a query that would otherwise retrieve too many matches. It has been shown to work effectively in a search assistant that uses simple dialogues.

Acknowledgement

Thanks to Nick Webb for the title of the paper.

References

P. G. Anick and S. Tipirneni. 1999. The paraphrase search assistant: terminological feedback for iterative information seeking. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 153–159, Berkeley, CA.

G. Attardi, A. Gullí, and F. Sebastiani. 1999. Automatic Web page categorization by link and context analysis. In C. Hutchison and G. Lanzarone, editors, *Proceedings of THAI-99, European Symposium on Telematics, Hypermedia and Artificial Intelligence*, pages 105–119, Varese, Italy.

S. Brin and L. Page. 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Proceedings of the Seventh International World Wide Web Conference (WWW7)*, Brisbane.

P. Bruza, R. McArthur, and S. Dennis. 2000. Interactive Internet search: keyword, directory and query reformulation mechanisms compared. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 280–287, Athens, Greece.

C. H. Chang and C. C. Hsu. 1999. Enabling Concept-Based Relevance Feedback for Information Retrieval on the WWW. *IEEE Transactions on Knowledge and Data Engineering*, July/August.

H. Chen and S. T. Dumais. 2000. Bringing order to the web: Automatically categorizing search results. In *Proceedings of CHI'00, Human Factors in Computing Systems*, Den Haag.

M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery. 1998. Learning to Extract Symbolic Knowledge from the World Wide Web. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98) and the Tenth Conference on Innovative Applications of Artificial Intelligence (IAAI-98)*, pages 509–516, Madison, Wisconsin.

M. Fasli and U. Kruschwitz. 2001. Applying Collaborative Filtering Ideas to Improve a Dialogue System for Web Search. In *Proceedings of the International Conference on Intelligent Agents, Web Technology and Internet Commerce - IAWTIC'2001*, Las Vegas. To appear.

S. Flank. 1998. A layered approach to NLP-based Information Retrieval. In *Proceedings of the 36^{th} ACL and the 17^{th} COLING Conferences*, pages 397–403, Montreal.

N. Guarino, C. Masolo, and G. Vetere. 1999. OntoSeek: Content-Based Access to the Web. *IEEE Intelligent Systems*, May/June:70–80.

C. Gutwin, G. Paynter, I. Witten, C. Nevill-Manning, and E. Frank. 1999. Improving browsing in digital libraries with keyphrase indexes. *Decision Support Systems*, 27:81–104.

J. M. Kleinberg. 1998. Authoritative Sources in a Hyperlinked Environment. In *Proceedings of the 9^{th} ACM-SIAM Symposium on Discrete Algorithms*. American Association for Artificial Intelligence.

U. Kruschwitz. 2001a. Dialogue for Web Search Utilizing Automatically Acquired Domain Knowledge. In V. Matousek, P. Mautner, R. Moucek, and K. Tauser, editors, *Text, Speech, and Dialogue. Fourth International Conference (TSD2001)*, Lecture Notes in Artificial Intelligence. Springer Verlag. To appear.

U. Kruschwitz. 2001b. Exploiting Structure for Intelligent Web Search. In *Proceedings of the 34^{th} Hawaii International Conference on System Sciences (HICSS)*, Maui, Hawaii. IEEE.

S. Lawrence and C. Lee Giles. 1999. Accessibility of information on the web. *Nature*, 400(July 8):107–109.

S. Lawrence, C. L. Giles, and K. Bollacker. 1999. Digital libraries and autonomous citation indexing. *IEEE Computer*, 32(6):67–71.

G. A. Miller. 1990. Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3(4). (Special Issue).

D. S. Modha and S. Spangler. 2000. Clustering hypertext with applications to web searching. In *Proceedings of ACM Hypertext Conference*, San Antonio, TX.

A. Müller, J. Dörre, P. Gerstl, and R. Seiffert. 1999. The TaxGen Framework: Automating the Generation of a Taxonomy for a Large Document Collection. In *Proceedings of the 32^{nd} Hawaii International Conference on System Sciences (HICSS)*, Maui, Hawaii. IEEE.

M. Sanderson and B. Croft. 1999. Deriving concept hierarchies from text. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 206–213, Berkeley, CA.

C. Silverstein, M. Henzinger, and H. Marais. 1998. Analysis of a Very Large AltaVista Query Log. Digital SRC Technical Note 1998-014.

K. Summers. 1998. *Automatic Discovery of Logical Document Structure*. Ph.D. thesis, Cornell University.

K. Taghva, A. Condit, and J. Borsack. 1994. Autotag: A Tool for Creating Structured Document Collections from Printed Material. Technical Report TR 94-11, Information Science Research Institute, University of Nevada.

P. D. Turney. 1999. Extraction of Keyphrases from Text. Technical Report ERB-1057, National Research Council of Canada, Institute for Information Technology.

P. D. Turney. 2000. Learning Algorithms for Keyphrase Extraction. *Information Retrieval*, 2(4):303–336.

O. Zamir and O. Etzioni. 1999. Grouper: A Dynamic Clustering Interface to Web Search Results. In *Proceedings of the Eighth International World Wide Web Conference (WWW8)*, Toronto.

Word Sense Classification Based on Information Susceptibility

Oi Yee Kwong
Language Information Sciences Research Centre
City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong
rlolivia@cityu.edu.hk

Abstract

In this paper, we propose a classification of word senses according to their information susceptibility. Senses are clustered based on their responses to various types of disambiguation information available from existing lexical resources. A pilot experiment with 60 noun senses supports that senses do fall into fairly distinct groups, such that they are best characterised by different types of lexical information. These results coincide with human response in word association tests. We will also suggest how such sense types could be applied to word sense disambiguation in reverse, to adjust the way different types of disambiguation evidence are combined.

1 Introduction

Most natural language processing tasks depend on one or more lexical resources for the required lexical information. The lexicalisation of word meanings in existing lexical resources and thus the lexical knowledge available from them are often restricted to relations between words, or sometimes word senses. These resources, however, say nothing on the subtle properties intrinsic to individual senses, such as how they are best distinguished from one another. In this paper, we study the plausibility of classifying senses by their intrinsic nature. We start the investigation from the perspective of word sense disambiguation (WSD), and will see how such a classification of senses might improve WSD performance in return.

While it is generally accepted that WSD needs multiple types of knowledge (e.g. McRoy, 1992; Ng and Lee, 1996; Stevenson and Wilks, 1999), it is nevertheless not well understood how they function individually and in combination. They are often uniformly combined as if they would produce the same effect on every sense of every word. This may not be the case, as Resnik and Yarowsky (1999) called WSD a "lexically sensitive" task; and Ide and Veronis (1998) pointed out that

> "... to date there has been little systematic study of the contribution of different information types for different types of target words. It is likely that this is a next necessary step in WSD work." (Ide & Veronis, 1998:p.21)

Although the optimal combination of knowledge in certain studies has somehow addressed this issue, we believe that a more transparent picture is needed. It would be useful to know the relative strength of different kinds of evidence for different senses. This type of knowledge is especially useful in the presence of conflicting evidence.

The assumption that different information types contribute variably to different target words is essentially a corollary of the belief that different types of lexical information vary in their effectiveness to characterise a sense and distinguish it from other senses of the same word. In other words, senses have varied *"information susceptibility"* (or just *"susceptibility"* for short). However, this property

of individual senses is unfortunately absent from existing lexical resources, and is thus neglected by most WSD algorithms to date.

In this study, we aim to investigate the plausibility of separating words, or more precisely word senses, into fairly distinct groups according to their information susceptibility. We call such groups *"sense types"*. To achieve this aim, we therefore need to find out, in the first place, if senses do fall into different clusters based on their responses to particular types of disambiguation information. We study this from a sample of noun senses to see how sense types might be acquired by grouping the test samples into sense clusters according to their responses to different types of disambiguating information, and to infer properties underlying such clusters. In addition, we would see what proportion of these clusters, if any, is capable of being disambiguated by exploiting our available lexical resources. We will also suggest how WSD might benefit from this kind of knowledge.

Hence, we need to test different types of disambiguation information and find patterns from their effectiveness on disambiguating the test samples. In this way, the hypothesis that different groups of senses are best characterised by different types of information can be put to test. Grouping the test samples with respect to their responses to the various types of information can reveal the differential information susceptibility of the senses.

In Section 2, we first briefly review Quillian's (1968) model of semantic memory. Then in Section 3, we discuss how the various types of lexical information in that model could be obtained from existing lexical resources. An experiment is described and results reported in Section 4. The psychological reality of the results in relation to human response in word association tests (e.g. Jenkins, 1970; Moss and Older, 1996) and its implications will be discussed in Section 5. Finally in Section 6, we suggest how such results can be applied in reverse to help WSD.

2 Quillian's Semantic Memory

The resolution of word sense ambiguities has primarily depended on some measure of the semantic closeness between words in context with respect to some semantic relation. This practice was largely inspired by studies on human semantic processing. Quillian (1968) proposed a computational model of human memory for storing the "meanings" of words. Apparently, our memory stores not isolated but connected information. So a model of semantic memory should have, in Quillian's words, "the ability to use information input in one frame of reference to answer questions in another". Thus his model has a network structure with a mass of nodes interconnected by associative links of different kinds.

Quillian's semantic network model allows two word concepts to be compared and contrasted via the links between them. His simulation program was able to say which meaning of each word was involved in each semantic similarity or contrast found, which is essentially what is expected of WSD. The association between concepts was found by a method generally known as "spreading activation with marker passing", which also underlies many later WSD programs.[1]

For example, Hayes (1977) represented word senses as nodes in a semantic network, and superimposed some frame-like structures (called "depictions") on them. Depictions were connected by binders, and there were certain representational constructs to capture special cases of the subset-superset and part-whole relations. These structures functioned with corresponding association rules, so that only the relevant depictions and thus the appropriate senses for the words were activated.

Hirst (1987) also used marker passing with a semantic network of frames, but controlled the process with additional constraints on the

[1] Subsequent findings on semantic priming (e.g. Collins and Loftus, 1975), are in accord with Quillian's network representation of semantic memory. Although various modes of lexical access have been proposed and none is inarguably supported by empirical evidence (see (Hirst, 1987) for a thorough review), the spreading activation model is generally accepted as a plausible mechanism for human semantic processing.

number of propagations and on the nodes from which activation is to be spread. Moreover, he introduced the use of Polaroid words when no connection path was readily found for an ambiguous word.

The associative links in semantic networks such as those described above may indicate different kinds of semantic relation. We can roughly identify three kinds of association: *broad* semantic relations (essentially word associations or thematic relatedness), *narrow* semantic relations (e.g. IS-A, PART-OF, etc.), and *positional* semantic relations (e.g. selectional restriction, idiomatic collocations, etc.). Nevertheless, hand-coded semantic networks like those above are hard to form and severely limited in scale. Meanwhile, Quillian himself also remarked that the contents in his semantic network could actually be obtained by encoding dictionary definitions accordingly. This suggests that we can in effect simulate a hand-coded semantic network by assembling different types of information extracted from various places. Kwong (2001) discussed an *Integrated Lexical Resource* (ILR) which was a combination of WordNet 1.6 and Roget's Thesaurus, and how the ILR might enrich the lexical information for WSD. WordNet and Roget's Thesaurus are exemplars for representing narrow semantic relations and broad semantic relations respectively. WordNet groups words into sets of synonyms ("synsets"), with an optional textual gloss. These synsets form the nodes of a taxonomic hierarchy. In Roget's Thesaurus, there are about 1,000 numbered semantic heads, under which words are first assorted by part of speech and then grouped into paragraphs according to broad conceptual categories. Heads are in turn grouped into sections and then into classes, resulting in a hierarchical organisation. In the current study, we will obtain a range of lexical information from the ILR for WSD, as discussed below.

3 A Semantic Relation Spectrum

In this section, we will describe the semantic relations used in this study and how they are obtained computationally from the ILR. We organise the semantic relations in terms of their specificity, from the narrowest type to the broadest type. This ranking is essentially intuitive, with the number of component lexical resources involved taken into account.

3.1 Narrow Semantic Relation (*wn*)

This is the relation captured in WordNet 1.6 and we measured the similarity between two senses x and y by taxonomic distance with the simple metric from Rada et al. (1989):

$$S(x,y) = \frac{1}{1 + Dist(x,y)}$$

where $Dist(x,y)$ is the minimum number of edges separating the nodes representing senses x and y.

3.2 Broad Semantic Relation 1 (*rog*)

The similarity of two WordNet senses x and y with respect to the Roget classification (now a component of the ILR) is measured by the following empirically determined function:

$$S(x,y) = \begin{cases} |M(x) \cap M(y)| & \text{if } M(x) \cap M(y) \neq \emptyset \\ & \text{and } (M(x) \cap M(y)) \cap TopC \neq \emptyset \\ 0.25 & \text{if } M(x) \cap M(y) = \emptyset \\ & \text{and } h_1 \in M(x), h_2 \in M(y), \\ & h_3 \in TopC, S(h_1) = S(h_2) = S(h_3) \\ 0 & \text{otherwise} \end{cases}$$

where

$M(s)$ = the set of Roget heads that sense s is mapped to,

$S(h)$ = the Roget section (one level above heads) subsuming head h, and

$TopC$ = the set of Roget heads covering most words in the text. Heads at the top two positions with eight or more words in the text are included (determined empirically), and ties are preserved.

3.3 Broad Semantic Relation 2 (*def*)

This is the word overlap between pairs of sense definitions in the tradition of Lesk (1986). Sense definitions are based on the WordNet glosses, with stop words and duplicated words removed.[2] The similarity

[2]Lacking reliable mechanisms, no stemming was done. However, perfect stemming should be beneficial for definition matching.

between senses is measured by the number (≥ 2) of words common to their definitions, normalised with the Dice function:

$$S(x,y) = \begin{cases} \frac{2|X \cap Y|}{|X|+|Y|} & \text{if } |X \cap Y| > 1 \\ 0 & \text{otherwise} \end{cases}$$

where

X is the definition of sense x, and
Y is the definition of sense y.

3.4 Broad Semantic Relation 3 (*com*)

This relation directly measures the relevance of a sense to the text as a whole. Each sense (in the text) is assigned a score based on the length of overlap between its WordNet definition and the set of words covered by the Roget heads in *TopC* (as defined above).

In addition, to find out how sufficient the lexical information in the ILR is for WSD, we need to test a range of information broader than what is available from the ILR. We thus introduce two other types of information obtained from corpus data, as follows.

3.5 Syntactic Relation (*f-2g*)

We used a flexible type of part-of-speech (POS) bigram for syntactic relation. Normally, one would estimate $p(t_{i-2}t_{i-1}|s_i)$ for a left-context bigram, where s_i is the sense at position i, and t_{i-2} and t_{i-1} are the POS tags of the words at positions $i-2$ and $i-1$ respectively. In this study, however, we look for all possible bigrams within a range of three consecutive positions on the left of the target word, in order to address the problem of data sparseness. So we estimate all $p(t_j t_k|s_i)$ where $i-3 \leq j < k < i$. Similarly for the right context, we estimate all $p(t_j t_k|s_i)$ where $i < j < k \leq i+3$. Subsequently in the experiment, we combine evidence from both the left and right context. With a small corpus and thus unreliable estimates of prior probabilities, we assumed uniform prior probabilities for all s_i.

3.6 Broad Semantic Relation 4 (*col*)

This refers to simple collocation. We followed Yarowsky's (1992) approach for estimating $p(word|s_i)$ where *word* is any noun (in root form) occurring within a window of ± 50 words from s_i. Again, given a small corpus, we assumed uniform prior probabilities for all s_i.

Parameters for *f-2g* and *col* were estimated from the SEMCOR corpus which contains about 90K sense-tagged nouns. However, since data sparseness was particularly pertinent, the parameter estimation was slightly modified from the standard approaches in attempt to improve the situation. First, synonymous senses of the targets were used to increase the training data where necessary. Second, only the candidate senses which had five or more occurrences in the training data were considered, and senses without sufficient training samples were discarded.[3]

The various types of lexical information described above are lined up in Figure 1, forming a spectrum of semantic information of varied specificity. The adequacy of the ILR can hence be evaluated according to the "formula" in Figure 2.

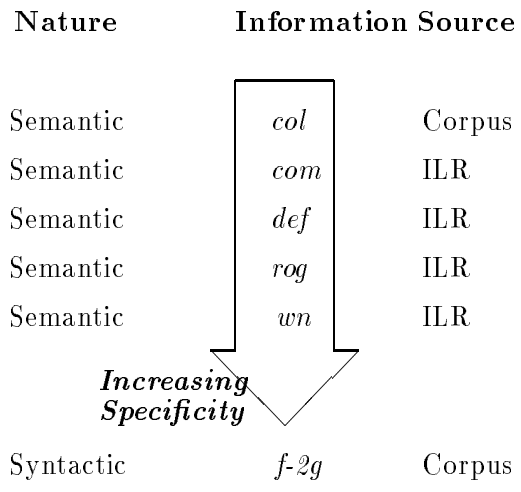


Figure 1: A Spectrum of Lexical Information

[3] Although such modification of the training data may invoke certain hidden bias in the disambiguation process, we assume it is random enough across all samples so as not to invalidate the study.

$$\left\{ \begin{array}{c} What\ can\ be \\ achieved\ with \\ ILR \end{array} \right\} = \left\{ \begin{array}{c} All \\ Samples \end{array} \right\} - \left\{ \begin{array}{c} No\ Response \\ OR \\ All\ Wrong \end{array} \right\} - \left\{ \begin{array}{c} Correct \\ ONLY\ with \\ col\ OR\ f\text{-}2g \end{array} \right\}$$

Figure 2: Formula for Estimating the Sufficiency of the ILR

4 An Experiment

4.1 Materials and Method

For this experiment, 14 short texts, each containing about 200 to 300 words with blanks to be filled in, were chosen from cloze test papers. All the blanks for polysemous nouns[4] were used for the test. The 14 texts thus resulted in a sample of 60 target nouns (with their expected senses known) for study and analysis. The average number of candidate senses for the samples was 5.4. Here is an example blank (the word in bracket is the known preferred answer and we additionally mark its expected sense from WordNet 1.6):

> ... The (rhythm/1) of the music also helps people to remember ...

The various types of disambiguating information mentioned in Section 3 were tested in combination with different context sizes to disambiguate the targets. The combinations[5] are summarised below:

Information	Context Size		
	$\pm 3 words$	$\pm 50 words$	Full
col		√	
com			√
def		√	
rog		√	
wn		√	
f-2g	√		

Apart from *com* and *f-2g*, all other types of information were applied with a window-based approach. Each sense of the target word was compared to each sense of each noun in context independently with respect to the particular type of information, and the results from these independent comparisons were combined. For each target, the sense with the best score was selected.

4.2 Results and Discussion

Table 1 shows the number of samples (out of 60) correctly disambiguated at each test condition.[6]

Information	**Context Size**		
	±3	±50	Full
col	–	23	–
com	–	–	20
def	–	15	
rog	–	20	
wn	–	17	
f-2g	13	–	–

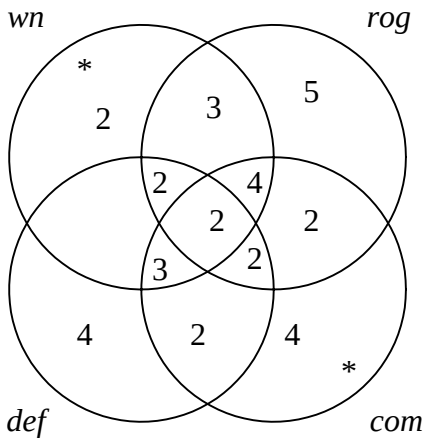
Table 1: Disambiguation Performance of Individual Information Types

As mentioned in the beginning of this paper, we would be more interested to see how the different types of information work individually and in combination. Figure 3 shows a further analysis of the above results in terms of the overlapping between information types from the ILR.

[4] The expected answers were checked to be present in WordNet 1.6.

[5] Theoretically for a controlled experiment, each type of information should undergo the same range of context sizes for comparison, although this is not always possible in practice. In this experiment, we simply assumed we are using some optimal context size for the different types of information, as learned from past studies.

[6] The numbers for *f-2g* and *col* are in fact out of a total of 34 samples, which had enough training data. However, our objective was to see how much can be achieved with the ILR, and all the information obtained from the ILR (*com*, *def*, *rog*, and *wn*) had been tested on all 60 samples. So even though *f-2g* and *col* were not applied to all 60 samples, we may only under-estimate the effectiveness of these two types of information, but not the usefulness of the ILR.



* There was also one overlapping case between *wn* and *com*.

Figure 3: Overlapping of Information Types

From Figure 3, it is obvious that while each individual information type can exclusively disambiguate some test samples, each type also overlaps with one or more of the other types at the same time. Nevertheless, there are only two out of all 60 test samples which are correctly disambiguated by all the four types of information. In other words, more often some types will produce no results or even contradicting results to others. Hence such conflicts must be resolved or at least minimised so that we can make the most of the edge held by individual information types and use them in combination for better WSD.

Assuming such an ideal situation, we may then estimate "what can be achieved with the ILR" from the results according to our "formula" in Figure 2:

Total no. of samples	=	60	
All wrong / No response	=	13	(21.67%)
Correct ONLY with *f-2g*	=	3	(5.00%)
Correct ONLY with *col*	=	4	(6.67%)
Correct ONLY with *f-2g* & *col*	=	4	(6.67%)

Thus we can work out that:

Correct with:			
ANY method	=	47	(78.33%)
SEMANTIC information	=	44	(73.33%)
SEMANTIC information in ILR	=	36	(60.00%)
1st sense from WordNet 1.6	=	32	(53.33%)
Random Choice Expected	=	–	(18.5%)[7]

We can thus see that if different types of information could be properly combined, the WSD performance would be much better than with any single type alone. However, it remains for us to find out what the proper way is.

Table 2 shows a manual clustering of the test samples according to their responses with respect to the various types of disambiguation information, with the sense number (according to WordNet 1.6) indicated after "/". We will discuss the results further in the light of this clustering in the next section.

5 From Clusters to Sense Types

Most of the test samples (about 80%, and possibly more) were successfully disambiguated in one way or another, as compared to the two baselines (53.33% with first sense and 18.5% with random choice). Also, most noun senses in our samples can be disambiguated by one or more types of semantic relation with other noun senses (i.e. Types B, C, D, and E in Table 2). We found that 17 out of 60 samples could be disambiguated with the narrowest paradigmatic association (i.e. Type B). This does not necessarily mean that those cases are not susceptible to broader semantic relations, but for them the simple IS-A relation suffices. In fact, most of our samples which were correctly disambiguated with *wn* were also correct with one or more other information types. However, we find the distinction (between broad and narrow semantic relations) important because there remains the possibility that some (untested) senses may only work with *wn*, and it is useful to gather all plausible evidence so that senses can be more confidently selected. Moreover, as suggested by Gentner and Brem (1999), thematic association and similarity comparison are distinct processes though the former often intrudes upon the latter.

The results allow us to uncover more about the susceptibility of target words (or more precisely their senses) to various types of information, and thus the effectiveness of such information for disambiguating the words.

[7] Estimated from the average number of senses 5.4.

Type	Information Susceptibility	Information Source	Examples
A	syntactic information only	non-ILR	instant/1, way/2
B	narrow semantic information	ILR	trick/5, sun/1, mirror/1
C	broad semantic information	ILR	space/5, injury/1
D	semantic information only	non-ILR	loss/2, attitude/1
E	syntactic/semantic information	non-ILR	attention/1, mystery/1
F	unknown	–	safety/1, exchange/3

Table 2: Clusters by Information Susceptibility

The clusters in Table 2 support our hypothesis that different senses are best characterised by different types of information. However, given the fuzziness of these clusters, at present we can give no more than an intuitive account of the pattern. As we go up the specificity spectrum of semantic relations, the senses susceptible to those relations seem to involve more and more abstract thinking. For example:

- Type B: What is sun/1? One can readily point to that source of heat and light in the sky (assuming a sunny day). What is mirror/1? One understands it as the instrument which reflects the image of oneself when standing in front of it.
- Type C: What is space/5? One cannot possibly point to a concrete object, but would rather think of the region with planets, satellites, orbits, rockets, and so on. Similarly, what is injury/1? It can be a bleeding wound, a broken leg, a dislocated knee, among others.
- Types D and E: What is loss/2? Loss as such is difficult to describe. A loss can be anything but not everything is a loss. What is attention/1? If it is "the concentration of one's mind", then what is "concentration" and what is "mind"?

Hence we are moving towards examples where an abstract idea can only be expressed via further abstract ideas. The relation we found between the clusters and the semantic relation spectrum also coincides with human response in word association tests. For instance, in the Birkbeck word association norms (Moss and Older, 1996), "loss" triggers associations like "death" and "grief", none of which are related through a narrow IS-A relation. On the contrary, "sun" is associated with "moon" and "trick" with "magic", among others. These associations can easily be related by the WordNet hierarchy. Such association data therefore support the claim that different strategies are employed by people to conceptualise different groups of senses, hence the idea of classifying senses by information susceptibility.

We have also seen that some senses are only identified by syntactic patterns, but it may not at all be obvious that some other senses might only be identified using higher level information such as world knowledge and inference. Some examples in Type F (unknown) may be of this kind. However, as there is an intrinsic distinction between lexical resources and encyclopaedia, we will not delve into the handling of those word senses here.

6 Future Work and Conclusion

The clusters of senses like those obtained in the present study could potentially be applied in reverse to WSD systems for tuning the relative importance of various types of disambiguation information. A scheme for doing this is proposed here.

First, the sense clusters need to be enriched with more samples. When manual clustering becomes tedious with increasing samples, the clustering could be assisted by some statistical procedure like factor analysis to extract certain prominent or core clusters.[8]

[8] The factor analysis function of SPSS extracted six factors from our data, and the first four corresponded very closely to the results of our manual clustering.

Second, once we have obtained a group of larger sized clusters of reasonable generality, other senses could be incorporated into these clusters in a more straightforward way, e.g. by using some similarity-based methods.

Third, the sense clusters could then form the sense types to tune the weighting of different types of disambiguation information in a WSD system, although how they are to be weighed would be another empirical issue.

In conclusion, our study supports that senses have varied information susceptibility, such that they are best characterised and thus best disambiguated by different types of lexical information. The results also throw light on the usefulness of the lexical knowledge contained in the ILR. Although we need more samples to help meaningfully "name" the sense clusters, classifying word senses according to information susceptibility could potentially help fine-tune WSD systems.

Acknowledgments

This work was done at the Computer Laboratory, University of Cambridge. The author would like to thank Prof. Karen Sparck Jones, Dr. Ted Briscoe and Prof. Yorick Wilks for advice at various stages. The work was financially supported by the Croucher Foundation.

References

A.M. Collins and E.F. Loftus. 1975. A spreading-activation theory of semantic processing. *Psychological Review*, 82(6):407–428.

D. Gentner and S.K. Brem. 1999. Is Snow Really Like a Shovel? Distinguishing Similarity from Thematic Relatedness. In *Proceedings of the Twenty First Annual Conference of the Cognitive Science Society*, pages 179–184, Simon Fraser University, Vancouver, BC.

P.J. Hayes. 1977. On semantic nets, frames and associations. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence (IJCAI-77)*, pages 99–107, Cambridge, MA.

G. Hirst. 1987. *Semantic Interpretation and the Resolution of Ambiguity.* Cambridge University Press.

N. Ide and J. Veronis. 1998. Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1):1–40.

J.J. Jenkins. 1970. The 1952 Minnesota word association norms. In L. Postman and G. Keppel, editors, *Norms of Word Association.* New York: Academic Press, Inc.

O.Y. Kwong. 2001. Word sense disambiguation with an integrated lexical resource. In *Proceedings of the NAACL Workshop on WordNet and Other Resources: Applications, Extensions and Customizations*, pages 11–16, Pittsburgh.

M.E. Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 1986 SIGDOC Conference*, pages 24–26, Toronto, Canada.

S.W. McRoy. 1992. Using multiple knowledge sources for word sense disambiguation. *Computational Linguistics*, 18(1):1–30.

H. Moss and L. Older. 1996. *Birkbeck Word Association Norms.* Psychology Press, Hove, U.K.

H.T. Ng and H.B. Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL '96)*, pages 40–47, Santa Cruz, CA.

M.R. Quillian. 1968. Semantic memory. In M. Minsky, editor, *Semantic Information Processing.* MIT Press, Cambridge, MA.

R. Rada, H. Mili, E. Bicknell, and M. Blettner. 1989. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1):17–30.

P. Resnik and D. Yarowsky. 1999. Distinguishing systems and distinguishing senses: new evaluation methods for Word Sense Disambiguation. *Natural Language Engineering*, 5(2):113–133.

M. Stevenson and Y. Wilks. 1999. Combining weak knowledge sources for sense disambiguation. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI-99)*, Stockholm, Sweden.

D. Yarowsky. 1992. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, pages 454–460, Nantes, France.

Analysis of functional similarities of Finnish verbs using the self-organizing map

Krista Lagus
Helsinki University of Technology
Neural Networks Research Centre
P.O.Box 5400 (Konemiehentie 2)
FIN-02015 HUT, Finland
Krista.Lagus@hut.fi

Anu Airola
University of Helsinki
Department of General Linguistics
P.O.Box 9 (Siltavuorenpenger 20 A)
FIN-00014 University of Helsinki, Finland
Anu.Airola@helsinki.fi

Abstract

Obtaining semantic or functional word categories from data in an unsupervised manner is a problem motivated both from the linguistic point of view and from that of construing language models for various language processing tasks. In this work, we use the Self-Organizing Map algorithm to visualize and cluster common Finnish verbs based on their immediate morphological contexts. Based on a data set of over 500,000 utterances of 25 verbs, we studied (1) the base forms and (2) the most common word forms of the same verbs (4764 forms). The results show that even the simple feature selection used in this experiment was found to be suitable for rough automatic categorization of verbs on the basis of data extracted from unrestricted texts. In particular, the automatically obtained organization resembles the semantic classification of verbs designed by a linguist.

1 Introduction

The data now available in the form of massive computer readable corpora has the potential to present a challenge to the way we think about the concept of linguistic theory. Especially, in theoretical corpus linguistics, one sense of the term 'data-driven' is to find new ways of deriving the theory from the data and thus bring new insights into linguistic theory itself (cf. (Sinclair, 1991; Stubbs, 1996)). However, one needs proper tools for identifying dependencies within the data and for representing them in an understandable form.

A much debated question is to what degree and which parts of our language ability are innate, and what and how much is learned from exposure to natural language in communicative settings. From the point of view of building systems capable of communication, a central question is: What is a minimal, sufficiently strong and covering set of learning principles and prior information that enables a system over time to learn versatile communicative skills? Obviously, this question defines a broad agenda for long-term research.

In this paper we discuss a sub-topic of *lexical acquisition*, i.e. the acquisition of information of semantic similarities of words (see e.g. (Boguraev and Pustejovsky, 1996)). The meaning-bearing elements in language expressions are constructions consisting of the following elements: the words themselves, morphological endings and markers, the order of words, and the intonation (see e.g. (Goldberg, 1995; Tomasello, 1998; Croft, 1999)). In actual expressions there is redundancy in the sense that several elements may contribute to communicating similar content. For example, in the sentence *I went to the store* both *went* and *to* appear to signal the presence of movement. Such redundancy allows one to deduce similarities between elements based on the contexts of their usage; in other words, to form 'contextual representations' of words based on hearing or seeing the words used in a variety of linguistic contexts (Miller and Charles, 1991; Leacock

et al., 1996).

The crucial question then is what kind of information is relevant when a system forms these contextual representations. More specifically, the following question is addressed here: How much information regarding semantic or functional similarities of verbs can, in principle, be uncovered by assuming very little, i.e., by selecting the verbs to be studied based on frequency, and by looking only at their immediate contexts in a very simple and straightforward manner. It is also assumed that (1) the system is already capable of distinguishing verbs from other syntactic word classes, and (2) able to recognize some morphological features.[1]

1.1 Statistical language modeling

Statistical language models (SLM) are models that are estimated based on large corpora and that allow one to make predictions regarding future linguistic data (e.g., future words) or interpretations of such data (e.g., which word was just heard) given some properties of previous data (e.g., previous words). Such models are applied, for example, in speech recognition. Surprising success has been achieved with very simple models, such as N-gram models (cf. (Manning and Schütze, 1999)) that record frequencies of all encountered combinations of n words and attempt to predict the next word based on the previous $n-1$ ones. However, these typically require vast amounts of learning data, and even then, some quite ordinary word combinations have not appeared in the data even once (Charniak, 1993). The central question then becomes one of generalizing existing information to new, similar but not identical cases. One way of doing this is by *clustering* the data, i.e., by grouping similar items together, and generalizing between words based on the clusters. This is sometimes called model smoothing.

A typical approach in statistical language modeling has been to apply a simple, equivalent strategy for all words, regardless of whether the words describe actions or objects. However, due to the apparent structure inherent in natural languages, a more fruitful direction might be to construct models with more structure, including different submodels for different kinds of words.

1.2 Related work

It has been previously shown that in languages with strict word order, such as English, the distribution of words in the immediate context of a target word contains a considerable amount of information regarding the syntactic and semantic category of the target word (see (Finch and Chater, 1992; Zavrel and Veenstra, 1995); for an overview of early work, cf. (Charniak, 1993)). Also the Self-Organizing Map (SOM) algorithm (cf. Sec. 2) has been applied for clustering English word forms based on the word forms in their immediate contexts in (Ritter and Kohonen, 1989; Honkela et al., 1995; Honkela, 1997), with promising results: both syntactic and semantic organization was observed. Furthermore, the word clusters obtained in such a manner have been used for encoding the meaning of documents e.g., in (Kaski et al., 1998). However, for Finnish the same approach does not seem feasible for two reasons. First, the kind of functional information that in English is coded by independent function words such as "to" and "at" is in Finnish coded by inflectional morphology (see Sec. 3 for details). Secondly, the function of word order in Finnish is to organize the flow of information at sentence and textual levels.

In models that are based on calculating combinations of word forms, the doubling of the number of target words simultaneously doubles the number of features. Moreover, with Finnish the effect is much stronger than a doubling—a single base form of a verb or a noun may have thousands of inflected forms (Karlsson, 1983). This causes immediate problems for reliable estimation of such models.

From the point of view of SLM the present article is a pilot study on how one might overcome some of the aforementioned problems. Our aim was to obtain a semantic or functional clustering for Finnish verbs. In particular, we apply the SOM algorithm for simultaneous clustering and visualization of the data. The visualization provides a means to understand and to qualitatively evaluate the obtained clustering and the feature

[1] Morphological features are often regarded as "surface features" since they are to some degree apparent in the actual words.

selection.

1.3 Structure of the article.

In Sec. 2 we briefly describe the SOM algorithm. The selected verbs and features are introduced in Sec. 3. The experiments and their results are described in Sec. 4, with conclusions and discussion in Sec. 5.

2 Self-organizing map algorithm

The Self-Organizing Map (SOM) (Kohonen, 1982; Kohonen, 1995; Kohonen et al., 1996) is an unsupervised neural network (Haykin, 1999) algorithm that is able to arrange complex and high-dimensional data so that similar inputs are, in general, found near each other on the map. On a visual display of the organized map properties of the data set can be illustrated in a meaningful manner.

The SOM algorithm automatically places a set of reference vectors—also called model vectors—into the input data space so that the data set is approximated by the model vectors. The model vectors are constrained to a (usually two-dimensional) regular grid that in effect forms an "elastic network": after application of the SOM algorithm, the map follows the data in a nonlinear fashion. The SOM algorithm obtains simultaneously *a clustering* of the data onto the model vectors and a *nonlinear projection* of the input data from the high-dimensional input space onto the two-dimensional, ordered lattice formed by the model vectors. Furthermore, the mapping defined by the model vectors can later be used to project new, unseen data onto the clusters or the map display.

3 Data and features

3.1 Selected verbs and the corpora

For this study, we selected 25 frequent Finnish verbs, shown in Table 1. All the occurrences of these verbs, a total of about 500,000 samples, were extracted from two sub-corpora of written standard Finnish (CSC, 2001), namely 'Newspapers' (13.6 million words) and 'Books' (4.1 million words). The morphological analysis of the data was conducted by the Conexor FDG, a functional dependency parser for Finnish (Tapanainen and Järvinen, 1997).[2] No effort was made to detect possible parsing errors. Furthermore, the verbs were not disambiguated semantically.

3.2 Contextual morphological features

To describe a verb we collected features from a window of a single word to the right of the target word, regardless of the syntactic function of the word in question. The following morphosyntactic properties were included: 1) a set of case endings (Table 2), 2) adpositions, i.e., postpositions and prepositions (see example 2), 3) adverbs, and 4) two nominal forms of verbs (i.e., the 1st and the 3rd infinitive, see example 3). Thus, the only information the system 'knows' is whether the next word immediately to the right of the target word bears some of the selected features.

One criterion for selecting the feature set was to prefer overtly marked features that can be identified by a system able to carry out some segmentation. These include case endings and the endings of the nominal forms of verbs. Moreover, by including 'adverb' and 'adposition' as features we assumed that the system is already capable of distinguishing parts of speech.

In Finnish, the primary means for coding various functional and syntactic dependencies between the verb and the nominal constituents in a clause is the case-marking system. The case endings are added to stems, as shown in (1), where the subject is marked by the nominative, and the end point of the movement indicated by the verb *ajaa* 'to drive' is marked by the illative.

(1)

Lapse-t	ajoivat	kaupunki-in.
child	drive	city
N-PL-NOM		N-SG-ILL

'The children drove to the city.'

In addition to the so-called local cases, location or movement can be coded by using an adposition (2), too. Adpositions include postpositions (PSP) and prepositions (PRE). They take a complement either in genitive or in partitive.

(2)

Lapse-t	ajoivat	kaupunki-a	kohti.
child	drive	city	toward
		N-SG-PTV	PSP

'The children were driving towards the city.'

Adpositions, adverbs, and local cases are grammatical categories that are at least partly func-

[2] ©Conexor oy and Anu Airola, <http://www.conexor.fi>

Verb	freq.	Translation
1. PSYCHOLOGICAL STATES AND PROCESSES, MODAL VERBS		
alkaa	17,987	begin, start, commence
haluta	14,637	want; wish; like (how do you like your tea); care for (would you care for a cup of tea?)
nähdä	14,746	see, spot
pitää	38,299	I) hold, keep, retain II) must, have to, be compelled to, be obliged to, shall; be supposed to
saada	57,815	get; receive; may, can, be allowed to, be permitted to; have to
tietää	12,678	know, be aware of, be conscious of
voida	46,474	can, be able, be capable, may
2. MOTION VERBS		
ajaa	6,372	drive; ride; run
käydä	14,447	go; walk
laskea 1	5,150	I) tr. lower, let down, haul down II) itr 1. fall, decline, go down, drop, decrease, diminish
laskea 2		count; calculate
lähteä	12,921	go, leave
mennä	18,486	go
nousta	10,889	rise; go up; ascend; climb; arise
päästä	14,809	a) get (into/out of); arrive, reach b) be let (into a room); be allowed (to enter); be admitted
tulla	43,972	come; arrive; become, get, turn, grow into
tuoda	6,840	bring
3. SPEECH ACT VERBS		
kertoa	21,474	tell; narrate, relate, recount; inform
puhua	10,389	speak, talk; tell; say; discuss
sanoa	43,015	say
vastata	7,554	answer, respond, reply, give an answer, make a reply, counter
4. ACTION VERBS		
käyttää	12,035	use (a knife)
tehdä	27,195	do; perform; commit; make
5. VERBS OF POSSESSION		
antaa	14,269	give; let; allow
kuulua	12,425	belong; be heard, be audible
ottaa	18,501	take

Table 1: The verbs divided into five categories according to Pajunen's ontological-semantic verb classification (Pajunen, 1999). The number after the verb shows the frequency of occurrences of the verb in the studied corpora. Translations are from (Hurme et al., 1984). Note that many of the verbs are polysemous and different senses should, in fact, be placed into different categories.

tionally overlapping. However, as examples (1) and (2) show, they construe the situation differently. The two clauses show an aspectual opposition; the illative in (1) indicates that the action is completed, while the postposition *kohti* 'towards' in (2) expresses duration without specifying completion.

The feature set used also includes two nominal (non-finite) verb forms, namely the 1st and the 3rd infinitive. The infinitives function in a sentence as nouns (3).

(3)
Halua-n	lähte-ä	syömä-än.
want	go	eat
V-PRES-SG1	V-INF1	V-INF3-ILL

'I want to go to eat.'

In Finnish, the lower-level constituents, e.g., NPs (noun phrases), are predominantly head-final, the order being modifier-before-head. Adjectives agree in number and case with the head-noun when they occur as attributes (4). Due to this redundancy, the function of a dependent NP can be inferred before hearing or seeing the head of the phrase.

(4)
Emmi	ajaa	punaise-lla	auto-lla.
Emmi	drive	red	car
N-PROP	V-SG3	A-SG-ADE	N-SG-ADE

'Emmi drives in a red car.'

In the clause-level, the basic, or default, word order is SVO. According to (Hakulinen et al., 1980), subjects precede finite verbs in 61% of all sentences in standard written prose. However, as Vilkuna (1989) points out, clause-level word order in Finnish shows great freedom; for example, in a simple sentence consisting of a subject, an object, a verb and one or two adverbials, all permutations are at least grammatically possible.

4 Experiments

4.1 Experiment 1: A map of 25 Finnish verbs

The purpose of this experiment was to obtain an overview of the verbs selected for study. In particular, we were interested in whether the contextual window of only one word and the set of morphological features would contain sufficient information for obtaining automatically a functional ordering or clustering for the verb base forms. As a basis for comparison, we used Pajunen's verb classification (see Table 1).

4.1.1 Data and preprocessing.

Let V be the set of 25 verbs, $v_i \in V$, $i = 1 \ldots 25$; and M the set of 18 morphological tags, $m_j \in M$, $j = 1..18$. In the data vector $\mathbf{x}_i$ for verb v_i a separate input dimension is reserved for representing each morphological tag. The value of $x_{i,j}$ is the frequency of context $v_i m_j$ in the corpus normalized by the frequency of verb v_i. In effect, $\mathbf{x}_i$ contains estimates of the conditional probabilities $Pr(m_j(t+1)|v_i(t))$ for each j (t is an index over the sequence of words). In other words, the vector $\mathbf{x}_i$ describes how typical each of the morphological tags is in the word following the verb v_i.

4.1.2 The organization and clustering of the data.

After forming the data vectors, various data analysis methods can be applied. We organized the verb representations automatically with the SOM algorithm, using the SOM ToolBox (Vesanto et al., 2000) for the data analysis tasks and for the visualizations. Default values were used for SOM learning parameters. K-means[3] was utilized for clustering the map vectors after creation of the map.

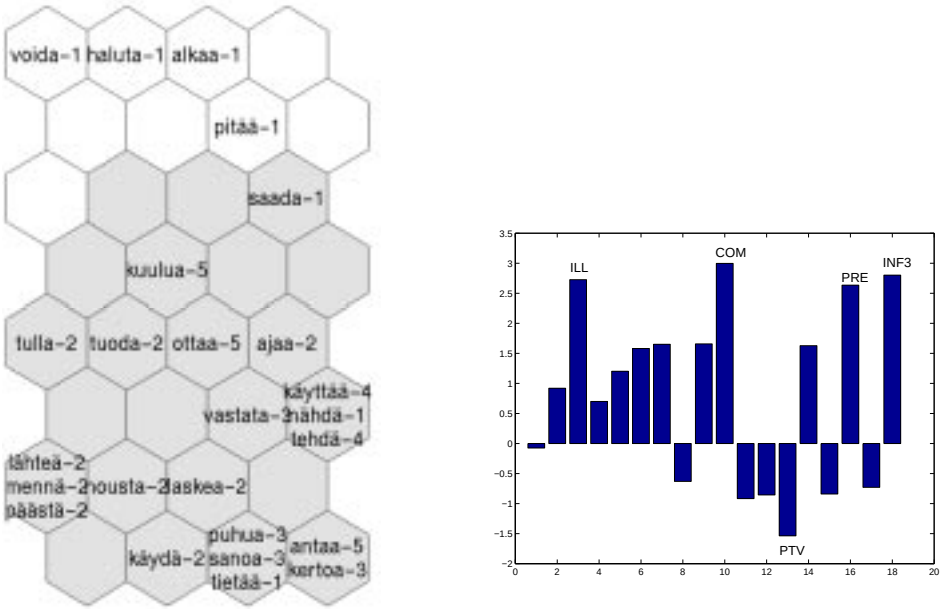
4.1.3 Results.

For each verb, the best-matching map unit (bmu) was found by locating the model vector closest to the verb vector. The verb was then written onto the bmu on the map lattice in Fig. 1. The data vectors of verbs related to two selected map units are visualized in Fig. 2. The values of individual features have been plotted on the map in Fig. 3.

When comparing the obtained functional organization to Pajunen's ontological-semantic verb classification, there seems to be a reasonable amount of agreement: verbs with the same category number are rather often found near each other on the map display.

However, even the perceived differences can be meaningfully justified. For example, the common aspect of *puhua-sanoa-tietää* ('to speak,' 'to say,' 'to know') is that they are cognitive verbs related to communicating or possessing in-

[3]K-means algorithm is obtained as a special case of the SOM by disregarding the neighborhood in the organization.



(a) (b)

Figure 1: Analysis of the 25 verbs. (a) The SOM lattice where each hexagon marks a map unit. The map units belonging to cluster 2 are shaded (bottom of the map). The best-matching unit, bmu, has been automatically searched for each of the 25 verb vectors, and each verb has been written on its bmu. For example, the bmu for the verbs *lähteä, mennä,* and *päästä* was the unit near the lower left corner of the map. The number after the verb refers to the verb class of Table 1. (b) The difference between the map unit of verbs *lähteä, mennä, päästä* and the rest of the map for each input feature. The values are normalized with standard deviation of the feature. If the value for a feature (say COM) is above zero, that feature is more frequent in this map unit than in general, and below zero means that it is here less frequent than in general. The features that most distinguish this unit from other units are ILL, COM, PRE, INF3, and the lack of PTV.

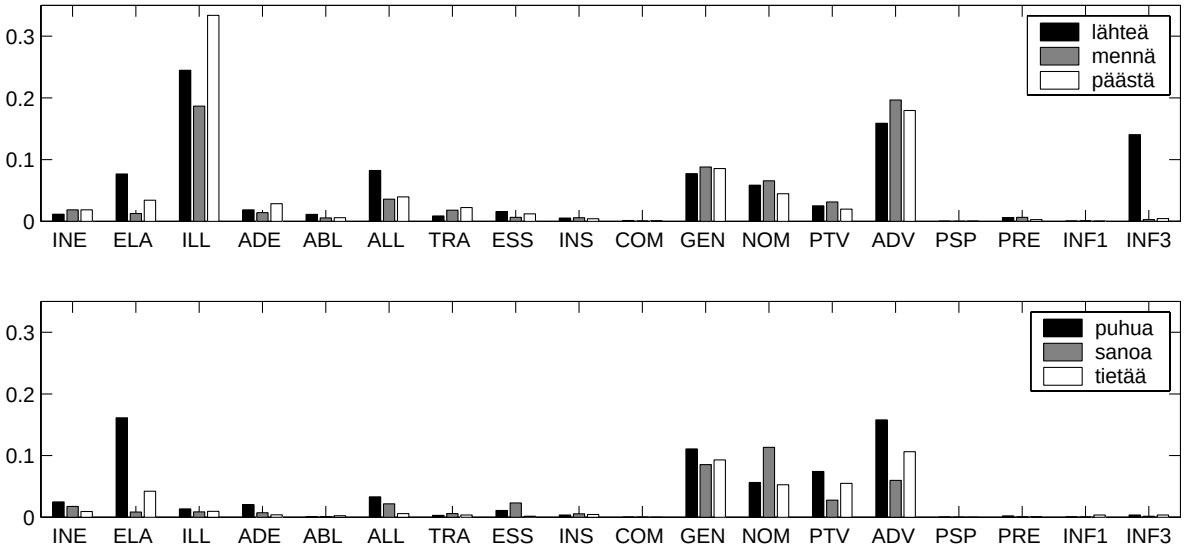


Figure 2: Data vectors for verbs in two different map units: the value for each morphological form is visualized as the height of the bar.

Case		Endings	Meaning	Example
		GRAMMATICAL CASES		
nominative	NOM	- (pl. -t)	basic form	auto 'car'
partitive	PTV	-a/-ä; -ta/-tä;	indefinite	maito+a '(some) milk'
		-tta/-ttä	quantity	vet+tä '(some) water'
genitive	GEN	-n; -den, -tten	possession	auto+n 'of the car'
		LOCAL CASES		
inessive	INE	-ssa/-ssä	inside	auto+ssa 'in the car'
elative	ELA	-sta/-stä	out of	auto+sta 'out of the car'
illative	ILL	-Vn,	into	auto+on 'into the car'
		-hVn,		maa+han 'into the country'
		-seen, -siin		Lontoo+seen 'to London'
adessive	ADE	-lla/llä	on; instrument	pöydä+llä 'on the table'
ablative	ABL	-lta/ltä	off	pöydä+ltä 'off the table'
allative	ALL	-lle	onto	pöydä+lle 'onto the table'
		OTHER CASES		
essive	ESS	-na/-nä	state	opettaja+na 'as a teacher'
translative	TRA	-ksi	change of state	opettaja+ksi '(become) a teacher'
comitative	COM	-ine-	accompanying	vaimo+ine+ni 'with my wife'
instructive	INS	-n	(idiomatic)	jala/n 'on foot'

Table 2: The cases used as input features (cf. (Karlsson, 1987): 22–23).

formation (CONTENT). In the neighboring unit, *antaa-kertoa* ('to give,' 'to tell') describe concrete or abstract change of possession (RECIPIENT). Moreover, *käyttää-nähdä-tehdä* ('to use,' 'to see,' 'to do') describe actions towards an object (PATIENT).

Pajunen's classification exhibits a certain ontological view of the world. As the comparison with our experiment suggests, an alternative perspective emerges when one concentrates on functional features and on what is typical in actual language use.

4.2 Experiment 2: A map of 4764 verb forms

The purpose of the second experiment was to study more closely the 21 verbs in the second cluster (the lower part of the map obtained in the first experiment, see Fig. 1). In particular, of interest were the various inflected verb forms, since it is to be expected that different forms of the same verb may have quite different behavior profiles.

The data vectors were formed similarly as in experiment 1, the only difference being that a vector was now collected per each verb form. To avoid unreliable estimates verb forms with fewer than 20 occurrences were discarded. As a result,

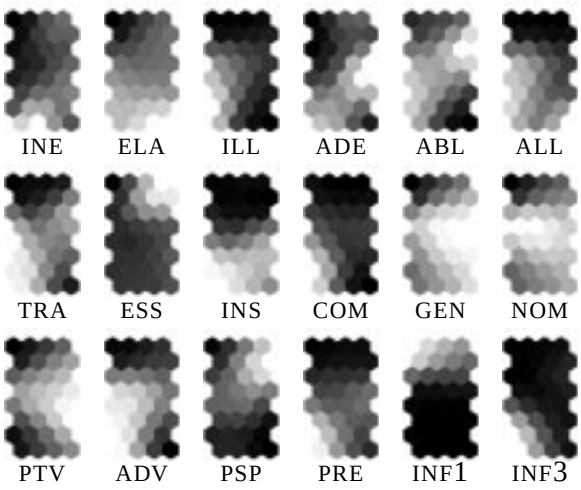


Figure 3: The values of individual features have been visualized on the map—bright colors signal large values. For example, the most prominent area of the feature INF3 is the lower left corner of the map. By comparing the visualizations of two features, dependencies between them can be identified. For example, features COM, PRE, and INF3 seem strongly correlated.

4764 data vectors were obtained. Next, a map of 11×14 units was automatically created with the SOM ToolBox. Best-matching units were then located for each data vector.

4.2.1 Results

To obtain an overall view of the map organization of the verb form map, we plotted all the occurrences of verb forms of a single base form onto a single map display. Several examples are depicted in Fig. 4. Fig. 5 shows more closely the behavior of different verb forms, such as the contrast between finite and non-finite forms of verbs.

The spread of a verb shows how coherently the forms of that verb behave in terms of the applied feature selection. By comparing spreads of two different verbs, say *mennä* and *lähteä*, one quickly obtains some overview of the differences and similarities between the verbs—in this case, that the verbs are rather similar. However, a difference is implied since the verbs fall on adjoining areas instead of overlapping.

By looking at the values of the morphological features on the map (not shown here) it can be seen that the feature INF3 (3rd infinitive) is especially strong on the area overlapping with the verb *lähteä* but not on the adjoining area of *mennä*. In fact, the illative form of the 3rd infinitive can be used after all motion verbs, indicating an action which is about to begin, as in (5):

(5) Lähden / Menen nukku-ma-an.
V-INF3-ILL
'I'll go to sleep.'

However, as the experiment shows, the 3rd infinitive illative is particularly typical of the verb *lähteä*. In contrast, the verbs *mennä* and *päästä* are used more often to indicate a 'pure' change of location.

5 Conclusions and Discussion

On the outset it may have seemed surprising that we decided to use only the word immediately following the verb for collecting information. After all, (1) many verbs take several arguments, and (2) since the word order in Finnish is free—or better, discourse-conditioned (Vilkuna, 1989)—the arguments may in some cases appear quite far apart from the verb, either before or after the verb.

Nevertheless, in Experiment 1, comparison between the map ordering and the pre-existing verb categorization shows that the selected morphosyntactic features of the next word indeed are statistically quite indicative of the functional-semantic properties of the verb, *if the information is accumulated over a large number of individual samples*. In such an accumulation, the typical cases will dominate and the less typical cases, many of which may be random (noise), are less frequent and become averaged out over time.

One might think that studying larger contexts would give a rather different view of the verbs. In fact, we repeated Experiment 1 so that both the preceding and succeeding words surrounding the verb were considered (using separate feature dimensions), with very similar results. It is possible that due to the unconstrained word order, wider contexts, if utilized in a similar manner, provide only a marginal amount of additional information. Nevertheless, especially when clustering rare words a wider context might become useful and worthwhile. Also additional features might be worth considering.

With respect to computational language modeling, the obtained clustering seems promising. In the future, larger numbers of verbs, including rare verbs, should be examined, and the approach should be extended to other kinds of words. Furthermore, the obtained clustering should be quantitatively evaluated in a specific task.

The visualizations obtained with the SOM can be extremely valuable, e.g., for a linguist interested in *actual language use*. This kind of method allows tackling larger problem areas than before. As seen from the simple examples presented here, the possibility to obtain an overview of the behavior of a group of words with a glance at the data may draw attention to some interesting differences that are otherwise hard to detect, and perhaps give rise to new perspectives or hypotheses regarding the linguistic phenomena.

References

Branimir Boguraev and James Pustejovsky, editors. 1996. *Corpus Processing for Lexical Acquisition*. The MIT Press, Cambridge, Massachusetts.

Eugene Charniak. 1993. *Statistical Language Learning*. MIT Press.

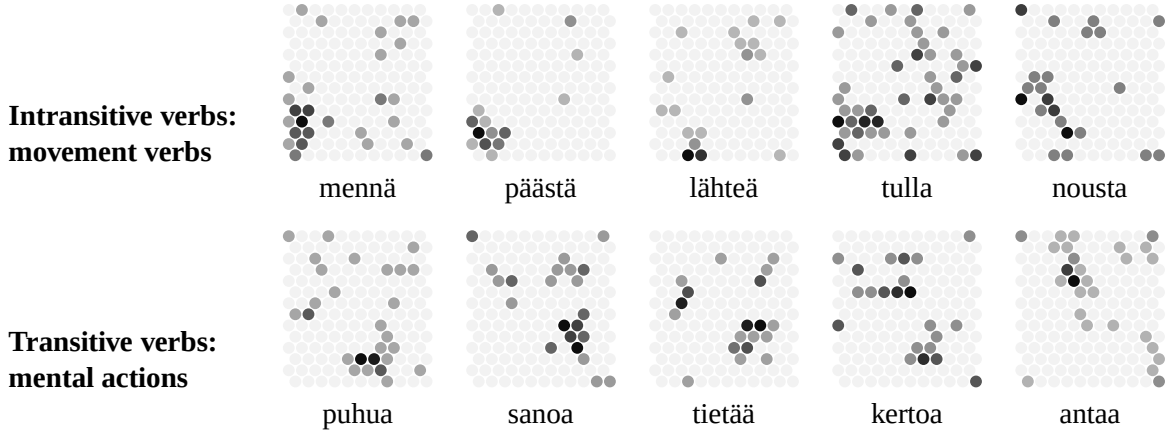


Figure 4: Verb distributions on the verb form map. The first plot depicts the occurrences of verb forms with the base form *mennä*—dark colour corresponds to a large number of occurrences. Each plot can be viewed as a "higher-order fingerprint" of the verb—similar fingerprints correspond to similar distributions of word forms in terms of their morphological contexts. The plots on the first row depict verbs that were found on the lower left side of the map in the previous experiment (see Fig. 1), whereas the second row consists of the verbs in the lower right corner of the same map.

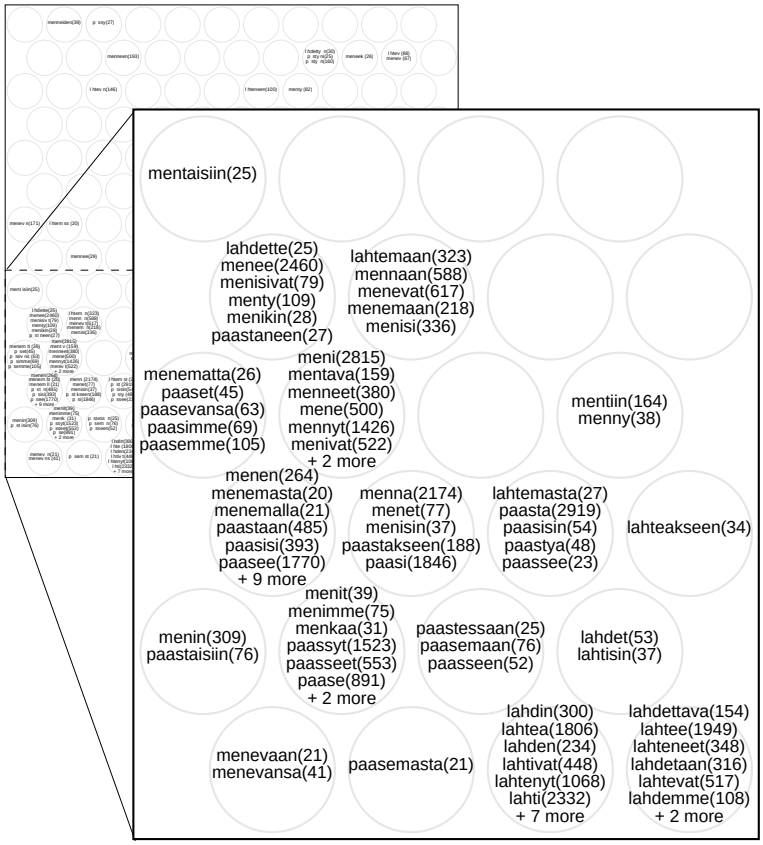


Figure 5: A close-up on the lower left corner of the map. The individual verb forms whose base form was *lähteä* 'to leave,' *mennä* 'to go,' or *päästä* 'to get into/out of' have been plotted on the map to their best-matching units (the scandinavian characters *ä* and *ö* are shown here as *a* and *o*). The numbers of occurrences of each form are shown in parenthesis. The verb forms in this area are mostly finite forms of the verbs. They seem to be in contrast to the non-finite forms, which appear as another, weaker cluster in the upper right corner of the map (not visible).

William Croft. 1999. Typology and cognitive linguistics. In Theo Janssen and Gisela Redeker, editors, *Cognitive Linguistics: Foundations, Scope, and Methodology*, pages 61–94. Mouton de Gruyter, Berlin.

2001. CSC—tieteellinen laskenta oy. Finnish Language Text Bank: Corpora 'Books' and 'Newspapers'. http://www.csc.fi/kielipankki/.

S. Finch and N. Chater. 1992. Unsupervised methods for finding linguistic categories. In I. Aleksander and J. Taylor, editors, *Artificial Neural Networks, 2*, pages II–1365–1368. North-Holland.

Adele Goldberg. 1995. *Constructions. A Construction Grammar Approach to Argument Structure*. Cognitive Theory of Language and Culture. The University of Chicago Press, Chicago.

Auli Hakulinen, Fred Karlsson, and Maria Vilkuna. 1980. *Suomen tekstilauseiden piirteitä: kvantitatiivinen tutkimus*. Publications of the Department of General Linguistics, Yliopistopaino, Number 6, Helsinki.

Simon Haykin. 1999. *Neural Networks—A Comprehensive Foundation*. Prentice Hall, 2nd edition.

Timo Honkela, Ville Pulkki, and Teuvo Kohonen. 1995. Contextual relations of words in Grimm tales analyzed by self-organizing map. In F. Fogelman-Soulié and P. Gallinari, editors, *Proceedings of ICANN-95, International Conference on Artificial Neural Networks*, volume II, pages 3–7, Paris. EC2 et Cie.

Timo Honkela. 1997. *Self-Organizing Maps in Natural Language Processing*. Ph.D. thesis, Helsinki University of Technology, Espoo, Finland.

Raija Hurme, Riitta-Leena Malin, and Olli Syväoja. 1984. *Uusi suomi-englanti suursanakirja. Finnish-English General Dictionary*. WSOY, Porvoo.

Fred Karlsson. 1983. *Suomen kielen äänne- ja muotorakenne*. WSOY, Porvoo. See also http://www.ling.helsinki.fi/˜fkarlsso/bookpaprep.html.

Fred Karlsson. 1987. *Finnish Grammar*. WSOY, Juva, the second edition edition.

Samuel Kaski, Timo Honkela, Krista Lagus, and Teuvo Kohonen. 1998. WEBSOM—self-organizing maps of document collections. *Neurocomputing*, 21:101–117.

Teuvo Kohonen, Jussi Hynninen, Jari Kangas, and Jorma Laaksonen. 1996. SOM_PAK: The Self-Organizing Map program package. Report A31, Helsinki University of Technology, Laboratory of Computer and Information Science, January.

Teuvo Kohonen. 1982. Self-organizing formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69.

Teuvo Kohonen. 1995. *Self-Organizing Maps*. Springer, Berlin.

Caludia Leacock, Geoffrey Towell, and Ellen M. Voorhees. 1996. Towards building contextual representations of word senses using statistical models. In Branimir, Boguraev, and James Pustejovsky, editors, *Corpus Processing for Lexical Acquisition*, pages 97–113. The MIT Press, Cambridge, Massachusetts.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

George A. Miller and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Anneli Pajunen. 1999. *Suomen verbirektiosta. Verbin argumenttirakenteen jäsenten valinta*. Yleisen kielitieteen julkaisuja 1. Åbo Akademis tryckeri, Åbo.

H. Ritter and T. Kohonen. 1989. Self-organizing semantic maps. *Biological Cybernetics*, 61:241–254.

John Sinclair. 1991. *Corpus, Concordance, Collocation*. Oxford University Press, Oxford.

Michael Stubbs. 1996. *Text and Corpus Analysis: Computer Assisted Studies of Language and Culture*. Blackwell, Oxford.

Pasi Tapanainen and Timo Järvinen. 1997. A non-projective dependency parser. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, Washington, D.C., April. Association for Computational Linguistics.

Michael Tomasello, editor. 1998. *The New Psychology of Language*. Lawrence Erlbaum Associates, Mahwah, NJ.

Juha Vesanto, Johan Himberg, Esa Alhoniemi, and Juha Parhankangas. 2000. Som toolbox for matlab 5. Report A57, Helsinki University of Technology, Neural Networks Research Centre, Espoo, Finland, April.

Maria Vilkuna. 1989. *Free Word Order in Finnish. Its Syntax and discourse functions*. Suomalaisen kirjallisuuden seura, Helsinki.

Jakub Zavrel and Jorn Veenstra. 1995. The language environment and syntactic word-class aquisition. In *Proceedings of the Groningen Assemly on Language Aquisition (GALA95)*, pages 365–374.

Conceptual enrichment of ontologies by means of a generic and configurable approach

Andreas Faatz (*, *) Stefan Hörmann*, Cornelia Seeberg*, Ralf Steinmetz(*, **)**

{afaatz, hoermann, seeberg, steinmetz}@kom.tu-darmstadt.de

*

Industrial process- and system-communications (KOM)

Technical University of Darmstadt
25 • 64283 Darmstadt

**

GMD IPSI
Integrated Publication and Information Systems Institute
Dolivostr. 15 • 64293 Darmstadt

***

intelligent views GmbH

Julius-Reiber-Str. 17 • 64283 Darmstadt

Abstract

In this paper we focus on systematic enrichment of ontologies by candidate concepts. To achieve this goal we compare semantic distance measures between concepts in an ontology with similarity and dissimilarity information introduced by statistical analysis of text corpora. Moreover we present methods, how the corpus extraction and the processing of the statistical information is also enhanced by a given ontology. Summing up these ontological methods we state the enrichment problem as an optimization problem. Its solution will yield two results: the best candidate concepts for the enrichment and the position of these candidates relative to the existing concepts of the ontology.

1. Introduction

Investigating the contemporary techniques of textual knowledge acquisition and formalisation, we identify two tendencies, an old one and a rather new one: on the one hand people preserve and interchange their knowledge creating natural language texts - a fundamental cultural technique of the civilized human being. On the other hand formal knowledge representation tries to find mappings of the real world or parts of the real world to a machine-readable and simplified model of the real world.

If we wish to share such a model of reality, it has to fulfil preliminary design principles [Gr]. The design principles have to hold while editing the knowledge representation. They prescribe formal rules, which are the basis of understanding, what the encoding of knowledge in the formal knowledge representation actually means.

The way we express our knowledge every day is completely different from formal knowledge engineering: we write it down in natural language. We observe a worldwide growth of the number of text documents, messages, textual archives and hypertextual informations, which are not formalised. All these 'uncontrolled' archives of knowledge grow significantly faster than the formal knowledge representations in any domain - and use natural language.

The question arises, if there exists an alternative to the strict formalisation of knowledge - an alternative, which benefits from the high quantity of textual information. In this paper we examine this question concerning the following situation: imagine there already exists a domain ontology, that is the conceptualisation of a knowledge domain [Gr]. Additionally there are large collections of text documents available, among them documents, which could form a basis for enriching a given ontology The question of opening formal knowledge representations for knowledge expressed in natural language transforms to two rather concrete ones:

- how can we extract concepts from the large collection of documents, which *fit semantically* to the knowledge, i.e. to the concepts, represented by the ontology?

- how do we *group the concepts* identified in such a manner to the concepts in the ontology?

In this paper we envisage an approach, which tries answer these questions of ontological enrichment: we extract concepts from very large collections of texts and insert them as candidate concepts of an ontology.

2. Overview

In this paper we exploit the comparison of semantic distance measures between concepts in an ontology and similarity introduced by statistical information about the usage and especially the collocated usage of words in text corpora. From the comparison we develop candidate concepts, which can for instance be presented to a knowledge engineer as a possible enrichment of the ontology.

The paper is organised as follows: in section 3.1 we characterise the formal knowledge we want to enrich. We introduce the notion of ontologies and define heuristics for a computation of conceptual distance in 3.2. Afterwards we characterise the less formally structured resources of natural language we want to use for an enrichment of the ontology (4.1 and 4.2). We are in need of an identification and an isolation of concepts from the textual resources. Founding on statistical information in section 4.3 we explain a generic and configurable approach, which originally was applied to automatic ontology generation [BiNeCa]. Since our goal is ontology enrichment we undertake a further development of the approach which is twofold: we refer to the generic features (the so called rule-base) of the approach inserting attributes, which can only be defined with an ontology at hand.

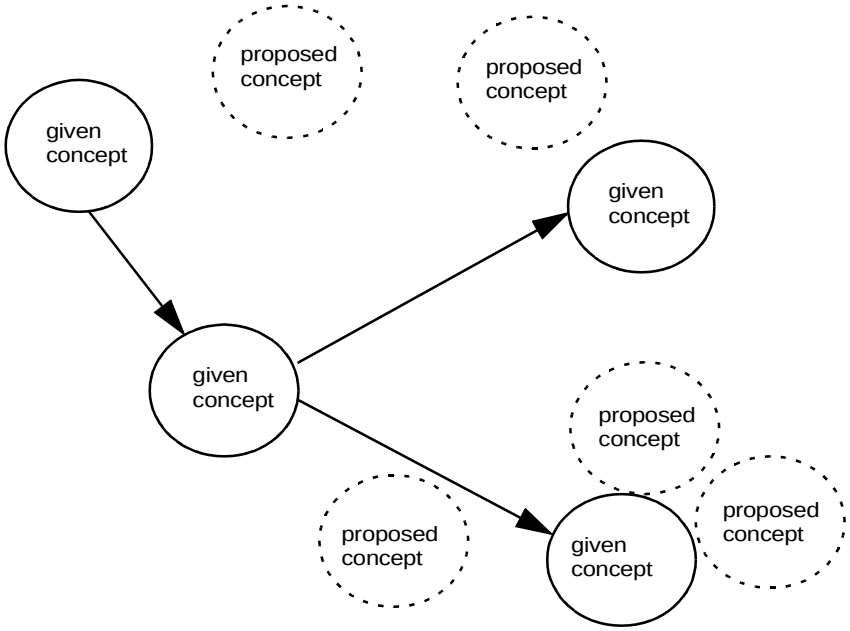


Figure 1

Secondly referring to the configurable features of the approach in section 4.4 we derive an optimization problem.

The solution of the optimization problem is able to fill a given ontology with proposals of new concepts, as shown in 4.5. The solution leads to the least possible numerical contradiction between 'similarities' in our textual and 'similarities' in our ontological world. Our vision is depicted in figure 1, where nodes indicating given concepts and arcs indicating 'is_a' relations among the given concepts attract further nodes, which means further concepts.

In section 5 we present our conclusions and future work to be done.

3. Ontologies

3.1 Definition and characterisation

Throughout this paper Ω will always denote an ontology.
In the literature about knowledge representation there exist many different definitions of ontologies. One of the most common definitions is due to Gruber [Gr], who defines an ontology to be "*a conceptualisation of a knowledge domain*". Guarino [Gu] put efforts into an at least uniform characterisation of the components an ontology necessarily holds. Together with [Gu] we state, that an ontology has the following parts: *concepts*, which represent the things existing in our knowledge domain, *relations* connecting the concepts semantically and *axioms* as formal laws for the ontology. Throughout this paper we assume that a concept has a name existing in a dictionary. This assumption is not natural for arbitrary ontologies, because for example in description logics [Ho] one could define concepts like 'all countries with the same number of inhabitants as Malta'. Clearly there does not exist a name from a dictionary for this concept. Moreover we do not want to consider ontologies containing constructed names like for example the medical ontology GALEN does [OG]. GALEN distinguishes 'skin-as-organ' and 'skin-as-tissue', contrary to that we would refer to 'skin'. Throughout this paper we assume the identity of a concept and its name.
The second ingredient of an ontology is a set of relations. A relation *r* in our case is binary and establishes statements about concepts *x, y* : we write *r(x, y)*, if it is true, that a relation *r* holds between *x* and *y*. For example, if *r* is 'is a' it may connect the concepts 'salmonella' and 'bacterium' to a statement, which we interpret as 'a salmonella is a bacterium'.
The third part of an ontology, the axioms, are responsible for the supervision of concept creation and deletion and for relation creation and deletion [StM]. For example 'is a' almost ever is a transitive relation, i.e. if for concepts x, y, z the relations *r(x, y)* and *r(y, z)* hold, then also *r(x, z)* holds. As our approach will propose extensions of a given state of an ontology, we do not examine axioms in detail.
Note that our characterisation of an ontology by concepts, relations and axioms implies that we can visualise such an ontology as a directed graph $\Gamma(\Omega)$, with nodes corresponding to concepts and edges corresponding to relations.
Our ontological enrichment or extension identifies new concepts and groups them semantically to concepts from a given set of concepts from an existing ontology Ω: we translate the concepts of Ω and the relations among them to a semantic distance measure in the ontology. Technically this is based on relational paths in $\Gamma(\Omega)$ and on topological properties of $\Gamma(\Omega)$.

3.2 Distance measures

The distance between two concepts *x, y* indicates, how strong the semantic similarity between *x* and *y* is. Semantic similarity and the distance measure are correlated: the bigger the distance between *x* and *y*, the bigger the dissimilarity between *x* and *y*. Unlike the approaches of [Tv], which is based on hierarchical structures, and [Leng], which is based formal concept lattices, we do not define our similarity measures by the attributes or the extension of concepts. Nevertheless the visualisation by [Leng], which is based on the construction of paths corresponding to an attribute-driven similarity definition of a distance, has commonalities with our premise: semantic distances between concepts in an ontology Ω are defined by path lengths in the graph $\Gamma(\Omega)$ we introduced in the previous section.
To enhance a reasonable distance measure we add a restriction to our ontology Ω, namely some our relations have to be hierarchical, for example 'is a'- or 'subconcept of'-relations. The parts of $\Gamma(\Omega)$ remaining form a multihierarchy as shown in figure 2, where the plain lines visu-

alise hierarchical and the interrupted lines express non-hierarchical relations. .

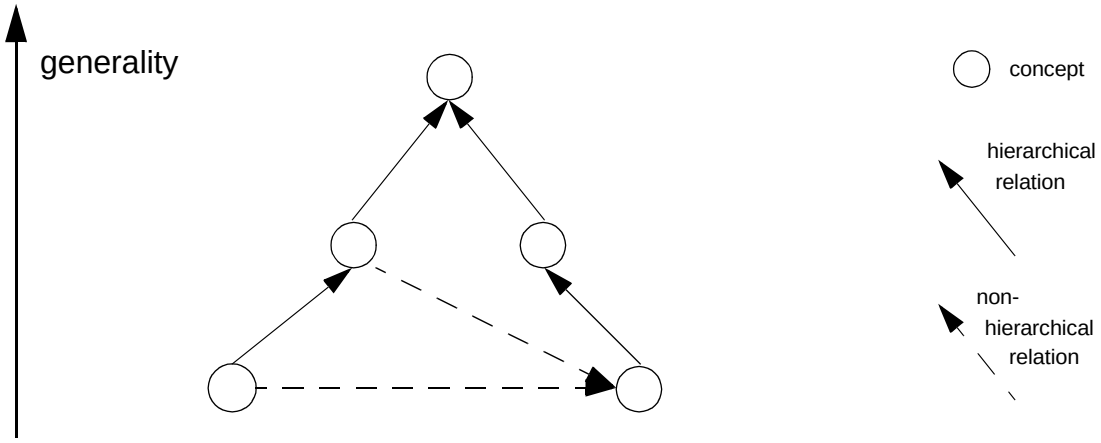


Figure 2

In figure 2 we also find an axis ('generality'), which expresses the fact, that the visualisation places the more general or abstract concepts at the top and the more special concepts to the bottom. We denote by $\Gamma'(\Omega)$ those parts of $\Gamma(\Omega)$, which are formed by the concepts (as nodes) and the hierarchical relations (as edges). Thus $\Gamma'(\Omega)$ would result from figure 2 omitting the interrupted lines. We will only use the hierarchical relations in the definition of semantic distance and claim, that the measure should fulfil four principles:

(i) longer paths in $\Gamma'(\Omega)$ are correlated with a higher semantic distance.
(ii) up-posting (generalisation) leads to a lower semantic distance than stepping down the hierarchy. Finding superconcepts is more fault tolerant than referring to senseless specialisations.
(iii) siblings increase the distance: this means the more subconcepts one concept *z* has, the higher is the distance between the subconcepts and also between *z* and its subconcepts. We posed this principle, because too many siblings often indicate missing abstraction levels in a hierarchy.
(iv) for isomorphic paths the distance between abstract or general concepts is higher than between special concepts at the bottom of the hierarchy, i.e. we judge 'the same' path on an abstract level to be vague in comparison to its pendant on a concrete level.

How do we achieve (i)-(iv) mathematically? An example distance design is given by the formula

$$(1) \qquad D(x,y) = \left[\frac{\alpha}{\gamma}\left(\frac{1}{\kappa}+\frac{1}{2\lambda}\right)\right]^{-2}$$

where the distance *D(x,y)* between two concepts *x, y* is computed with α denoting the average abstraction level of the concepts *x, y*, whereas γ denotes the average number of siblings of *x* and *y*. By κ we denote the steps upwards in the hierarchy while moving from *x* to *y* via the shortest path in $\Gamma'(\Omega)$, by λ the steps downwards in the hierarchy. Let us define exceptions for formula (1). If $x = y$ then set *D(x,y)=0*, if $\kappa = 0$ then set

$$(1a) \qquad D(x,y) = \left[\frac{\alpha}{2\gamma\lambda}\right]^{-2}$$

and analogously, if $\lambda = 0$ set

$$(1b) \qquad D(x,y) = \left[\frac{\alpha}{\gamma\kappa}\right]^{-2}$$

Formula (1), (1a) and (1b) together fulfil (i)-(iv), because

(i) increasing path lengths κ and λ increase *D(x,y)*

(ii) $(2\lambda)^{-1}$ decreases faster than κ^{-1}

(iii) as γ increases also *D(x,y)* increases

(iv) we work with an abstraction factor α, which we define as 1 divided by the average number of steps along the shortest path in $\Gamma'(\Omega)$ from *x* and *y* to one of the top level concepts, which means a concept without any hierarchical relations posting up from it. The abstraction factor α is high on a high level of abstraction and low on a low level of abstraction.

Note that *D(x,y)* and *D(y,x)* do not have to be equal. The exponent -2 stretches the distances, such that high distances become relatively even higher.

For technical reasons, namely turning semantic distance into semantic similarity to compute the optimization formula, α should be modified by a multiplication with a positive real in such a way, that $0 \le D(x,y) \le 1$ for every pair of concepts *x* and *y*. A consequent modification of α would yield 1 for the maximal occurring distance *D(x,y)*.

With *n* concepts given, we obtain $n \times n$ distances, which will serve as the input for the comparison mechanism optimising the configuration of the statistic knowledge acquisition process, which we will describe in the next section.

4. Generic and configurable acquisition and enrichment

4.1 Overview of the approach

For the remainder of this paper we consider an ontology Ω to be given. In this section we refer to a generic approach to determine similarities between concepts. It is based on computing statistics of word usage in natural language. Large text corpora are the basis for such statistics. [BiNeCa] first described the design and the capabilities of such an approach and applied it to conceptual clustering without a given ontology. Consequently the major differences between [BiNeCa] and our approach are

- we focus on creating algorithms for ontological enrichment and not on automatic ontology generation.
- we control our textual input by the concepts from Ω. This is motivated by experiments outlined in [BiNeCa]: specialised and pruned corpora seem to be the ideal input to unbias the word usage statistics.
- we only will use the initial step of [BiNeCa] and define conceptual similarities. We will modify it because we want to determine conceptual similarities within and by the help of Ω.

We undertake the following steps: we list the concepts from the ontology Ω and query the text corpus. This extracts from the corpus all natural sentences with at least one concept from the ontology Ω (see 4.2). We fill a representation of the collocation phenomena in these extracted sentences (see 4.3). From this sample representation we define similarity measures between any pair of concepts: note that a concept may belong to the *known* concepts from Ω and the additional concepts from the extracted sentences, which were formerly *unknown*. This finally yields similarities between each pair of concepts. These similarities will be optimised (see 4.4 and 4.5) in such a way, that at least the *known* concepts from the ontology Ω have similarities, which do not contradict the distances from section 3.2.

4.2 Corpus split and extraction

Again an ontology Ω and a large text corpus are given. By $B_\Omega = \{b_1, , , , b_n\}$ we denote the set of concepts from Ω. In this first step we derive a corpus C_Ω, which contains all sentences from the given corpus with at least one $b_i \in B_\Omega$. The preparation of a corpus C_Ω is motivated by the work of [BiNeCa] on the one hand, who found evidence for the fact that specialised corpora with a restricted vocabulary tend to be a good basis for ontology learning. A partial restriction of the vocabulary and the domain in our case will be due to the set of concepts B_Ω. The work of [ChGr] shows, that a corpus, which is as general as the one we use (a newspaper corpus in both cases) may be split to several artificial corpora, that were also a good basis for ontology learning. [ChGr] called their method 'corpus split' because they grouped sentences from the corpus to thematic domains: a sentence was added to the sample of the thematic domain, if it contained significantly many words from one and only one thematic domain. The thematic domains were characterised by typical words occurring in domain (such as 'judge', 'crime', 'law' for the domain if justice) together with a value of importance: for the domain of justice judge(0.8), law(0.5), crime(0.6) for example would indicate, that the occurrence of 'judge' in a sentence is highly significant for grouping a sentence containing 'judge' to the domain of jus-

tice, i.e. more significant than the occurrence of 'law' in a sentence. The domain significance information was determined by a collocation network of the classical news categories in newspapers (politics, foreign affairs, arts, media, justice etc.).

Our approach to extract sentences from a corpus works from the point of view Ω - and especially B_Ω - instead of a collocation network. Another difference to [ChGr] is, that we do not categorise the extracted sentences to get specialised text corpora. We rely on the powers of the distance measures for Ω defined in 3.2 to implicitly group the concepts from C_Ω. This will be done in the further steps.

Although the ontology enrichment approach we will show in the remainder of this paper is a technique, which may be applied to other languages, our methodology will be applied to German. For German the IDS ('Institut für deutsche Sprache', i.e. German Language Institute) located at Mannheim/Germany offers online access to very large newspaper corpora [IDS]. The general online corpus at the IDS contains several volumes of more than ten newspapers respectively. Moreover we find the online query system COSMAS I [IDS] supporting extraction queries like the one described at the beginning of 4.2. and in addition to this systematic stemming. Consequently for the German language we find an ideal environment for the establishment of the corpus C_Ω

4.3 The representation matrix

The next basic step is filling a representation matrix $M(C_\Omega, \rho)$ for a finite rule set ρ and the corpus C_Ω defined in the previous section. We firstly explain, what a matrix entry means and then we explain how we install a rule set ρ. Let us state, that C_Ω by definition contains some of the concepts B_Ω and further nouns, which we aggregate to a set $B(C_\Omega)$. B_Ω are *known* concepts from the ontology Ω, whereas $B(C_\Omega)$ are formerly *unknown* concepts.

We define $B_\Omega \cap B(C_\Omega)$ to be empty, $B(C_\Omega)$ consists of the concepts (i.e. nouns) *additionally* found in C_Ω and is the set of the potential candidates for the enrichment.

Each row in the representation matrix includes the information concerning the properties of exactly one concept from the unified set of concepts $B_\Omega \cup B(C_\Omega)$. More precisely reading the *i*-th row of $M(C_\Omega, \rho)$ verifies, if the rules in ρ are fulfilled for the *i*-th concept of $B_\Omega \cup B(C_\Omega)$. That means the *j*-th entry in the *i*-th row indicates, if the *j*-th rule is fulfilled or how often it is fulfilled in C_Ω. The rules ρ used in [BiNeCa] were of a syntactic type. The deployment of a syntactic parser found concepts, which appeared in at least one sentence from the corpus in a subject-predicate-object relationship with another concept. We give an example of a syntactical rule:

> Take the sentence '*The bakerman is baking bread*' as an example sentence occurring in the corpus. The rule ' '*bread* ' occurs as an object of '*bakerman*' ' would be fulfilled, as the sentence exists, and in the row for '*bread* ' there would be a positive entry at the *j*-th position, if *j* is the number of the rule 'concept *x* occurs as an object in sentence with *bakerman* as a subject'.

As we do not have syntactic parsers for German at hand we define a rule set ρ, which reflects another usage information. The genericity of the approach comes into play here: technically every rule set ρ reflecting usage of concepts in a corpus may design the columns of our matrix $M(C_\Omega, \rho)$.

The query system, which [IDS] offers, is able to return the words, which occurred at a maximal distance δ_W in a sentence. δ_W just counts, how many words one has to pass in a sentence from one word to another. At this point we remark, that the notion of δ_W should not be mixed up with the semantic distances from section 3.2. In our example sentence '*The bakerman is baking bread*' we obtain the values for δ_W listed in table 1

Table 1:

word	word	δ_W
bakerman	*bread*	3
bakerman	*baking*	2
baking	*bread*	1
bread	*bread*	0

For the rule set ρ we chose a fixed δ_W and constitute the *j*-th rule dependent from the ontology Ω and its set of concepts $B_\Omega = \{b_1, , , , b_n\}$. We define the *j*-th rule ρ_j to be fulfilled for a concept $x \in B_\Omega \cup B(C_\Omega)$, if *x* occurs at a prede-

fined maximal distance δ_W to the concept $b_j \in B_\Omega$ in a sentence from C_Ω.

Once we organised this particular rule set ρ we declare the entries m_{ij} of the matrix $M(C_\Omega, \rho)$ as follows:

(2) $m_{ij} = 0$, if the *i*-th concept in $B_\Omega \cup B(C_\Omega)$ does not fulfil the *j*-th rule ρ_j in ρ

(3) $m_{ij} = 1$, if the *i*-th concept in $B_\Omega \cup B(C_\Omega)$ fulfils the *j*-th rule ρ_j in ρ

For a more refined evaluation of the fact *how often* a rule ρ_j was fulfilled we could modify m_{ij} from (3) and count the number of occurrences of rule ρ_j for the *i*-th concept.

Actually all the m_{ij} from for technical reasons should be normalised to obtain $m_{ij} \in [0, 1]$. We achieve this by viewing all m_{ij} as a statistic sample, applying statistic standardisation and cutting off all values less than zero.

For the development of our approach we only focus on three characteristics of the m_{ij}:

(i) if a rule is not fulfilled for a concept, (2) applies and the entry equals zero.

(ii) if a rule is fulfilled for a concept, $m_{ij} \in (0, 1]$ must hold.

(iii) if we do not rely on the naive strategy (3), we should define m_{ij} positively correlated with the frequency of this rule appearing in the corpus C_Ω.

We now give an example of a representation matrix constructed according to (2) and (3). Note that the size of the example is not realistic. Indeed we have to deal with large corpora, so we just demonstrate the context of section 4. Let the set of known concepts be $B_\Omega = \{bread, butter, food\}$. Let C_Ω consist of two sentences:

The bakerman is baking bread.
She had bread and butter for breakfast.

The set of additionally retrieved concepts from the corpus C_Ω is $B(C_\Omega) = \{breakfast, bakerman\}$. With a rule set ρ founded on the word distance $\delta_W = 4$ in a sentence and the naive entry strategy (3) we obtain a matrix $M(C_\Omega"", \rho)$ like the inner cells we visualise in table 2:

Table 2:

	ρ_{bread}	ρ_{butter}	ρ_{food}	$\rho_{breakfast}$	$\rho_{bakerman}$
bread	1	1	0	1	1
butter	1	1	0	1	0
food	0	0	0	0	0
breakfast	1	1	0	1	0
bakerman	1	0	0	0	1

The matrix is not at all typical, regularly we expect sparse matrices in our approach. The next section will derive similarity measures from matrices $M(C_\Omega, \rho)$ by comparison of the respective rows.

4.4 Configurations, similarities and optimal configurations

We follow [BiNeCa] and assign a weight to each rule in our set of rules ρ. While [BiNeCa] assigned these weights with a focus on avoiding overgeneral clustering, a problem obscuring many approaches in automatic thesaurus and ontology construction, we will use the weights as a

configuration, which bring into line the two semantic informations we compute: the distances implicitly given by $\Gamma'(\Omega)$ (see 3.2) and the similarities we define now.

Let us assume a given representation matrix $M(C_\Omega, \rho)$. A set $k = \{k_1, \ldots, k_n\}$ of positive (or zero) reals with $|k| = |\rho|$ will be called configuration of the rule set ρ. The configuration k decides about the similarities we derive from $M(C_\Omega, \rho)$ by the following definition. We define conceptual similarity *S(x,y)* between two concepts *x*, the *i*-th concept from $B_\Omega \cup B(C_\Omega)$ and *y*, the *j*-th concept from $B_\Omega \cup B(C_\Omega)$ to be

$$(4) \qquad S(x,y) = \sum_{l=1}^{|\rho|} \frac{k_l}{|\rho|} min[(m_{il}, m_{jl})]$$

Thus a fixed configuration is needed to compute *S(x,y)*. The question arises, how we should chose such a fixed configuration *k*. Remember section 3.2: we already own - for the concepts B_Ω from the ontology Ω - a similarity, we just have to interpret semantic distance *D(x,y)* as the contrary of semantic similarity. Because we required for our distance measure $D(x,y) \in [0,1]$ the transformation of such distance measures to similarity measures, $S_\Omega(x,y)$ can be achieved easily:

$$(5) \qquad S_\Omega(x,y) = 1 - D(x,y)$$

Taking the distances the ontology Ω as an input, which approximately should be supported by the *S(x,y)* the question of finding an optimal configuration *k* reduces to the question:

> what configuration might minimise the average (squared) error expressed by the (squared) differences $(S(x,y) - S_\Omega(x,y))^2$?

Finally we present a formulation of this question in terms of a quadratic optimization formula. Searching for an optimal *k* means searching for a minimum of the expression

(6)

$$\min \sum_{i=1}^{N} \sum_{i=1}^{N} \left(1 - D(x_i, x_j) - \left(\sum_{l=1}^{|\rho|} \frac{k_l}{|\rho|} min[(m_{il}, m_{jl})] \right) \right)$$

with respect to $k = \{k_1, \ldots, k_n\}$ and $k_l \geq 0$ for all $k_l \in k$. Let us clarify the notation of formula (6): we let *i* and *j* run from 1 to $|B_\Omega| = N$. This means, that in formula (6) x_i denotes the *i*-th concept from B_Ω. Correspondingly in (6) the m_{il} are the matrix entries in $M(C_\Omega, \rho)$ in the row of x_i, whereas the m_{jl} are the matrix entries of $M(C_\Omega, \rho)$ in the row of x_j.

As the time consumed for solving the quadratic optimization problem (6) depends on the number of rules, we mention at this point, that in case of large rule sets ρ these ρ can be reduced in two ways:

- if a rule is not at all or only once fulfilled for the concepts B_Ω, it does not influence *S(x,y)* for $x, y \in B_\Omega$ andshould be skipped.
- if two rules are strongly positively correlated, for example in case of synonyms constituting the rules, one could merge these rules to one rule.

Furthermore, we suggest a careful approach towards the number of concepts from Ω, that is on $|B_\Omega|$. All steps presented also work with any subset of B_Ω.

We now get to our conclusion and explain, which concepts are candidates for the ontological enrichment.

4.5 Ontological enrichment

Once we optimised formula (6) with respect to $k = \{k_1, \ldots, k_n\}$ we obtain the configuration in need to compute all the similarity measures *S(x,y)* between all the concepts from $B_\Omega \cup B(C_\Omega)$. We propose to apply an enrichment step starting with the optimal similarity measures *S(x,y)*.

Only take into concern the *S(x,y)* with $x \in B_\Omega$ and $y \in B(C_\Omega)$. If such a distance between a formerly known concept and a formerly unknown concept exceeds a threshold, for example the average of the similarities $S_\Omega(x,y)$, it is a candidate concept, which should be communicated to

a knowledge engineer creating or maintaining the ontology Ω.

Additionally the *S(x,y)* with $x \in B_\Omega$ and $y \in B(C_\Omega)$ carry even more information, namely an optimal placement of the candidate concepts. The candidate concepts and the concepts from Ω can be presented together, a fact that simplifies the knowledge engineer's understanding of how the candidate concepts evolved, i.e. in which semantic area of Ω they might belong. We sketch how this can take place in a visualisation. Remember figure 2, which presented $\Gamma'(\Omega)$, the hierarchical parts of Ω as a graph. Figure 3 is an example of a visualisation of the enrichment process.

.

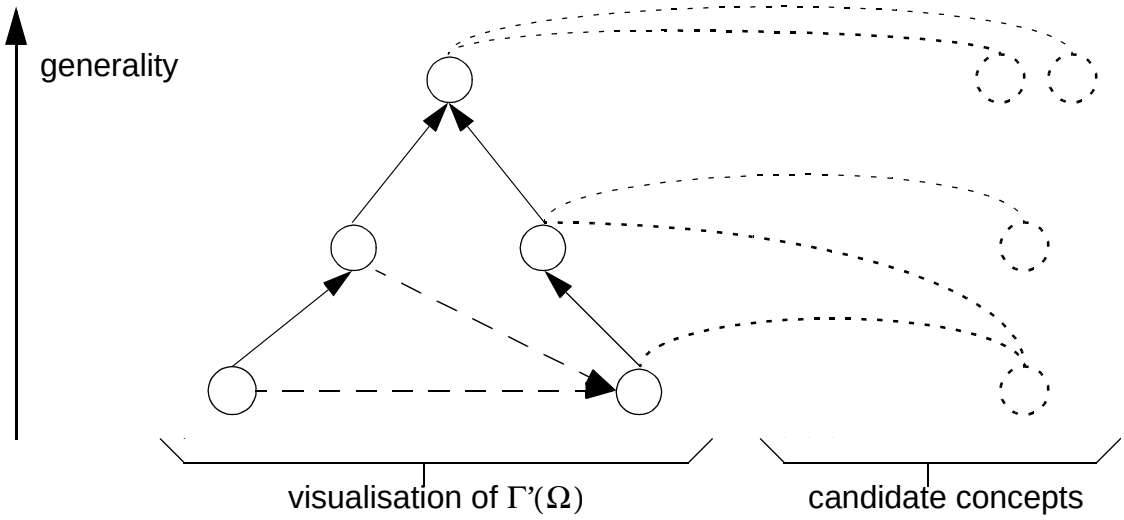


Figure 3

Figure 3 shows the visualisation of the known concepts on the left. On the right the candidate concepts were inserted. These candidate concepts have a similarity exceeding the threshold *T* we mentioned at the beginning of this subsection. Moreover, the dotted lines between candidate concepts and known concepts show, *which one* of the known concepts and *which one* of the candidate concepts have a similarity exceeding the threshold *T*. To sum it up we can visualise our ontological enrichment by drawing $\Gamma'(\Omega)$, adding all $y \in B(C_\Omega)$, for which a $x \in B_\Omega$ with S(x,y) > T exists. Finally we draw a line for each pair x, y with S(x,y) > T. Thus a candidate concept can refer to one or many existing concepts, and two or more candidate concepts can refer to one existing concept.

5. Conclusions and future work

In this paper we presented an approach of ontological enrichment. The approach leads to grouping new concepts identified in a large text corpus to concepts known in an ontology Ω. The way to achieve this goal is influenced by the given ontology Ω three times:

(i) we derived a text corpus C_Ω, which is particularly characteristic for Ω.

(ii) we created rules to determine, to which extend the usage of a concept in C_Ω was similar to the usage of another concept in C_Ω. The rule set ρ consisted of rules implied by the ontology Ω.

This step was possible because the approach has generic features principally allowing the use of many different rule sets.

(iii) during an optimization process we grouped candidate concepts to concepts from Ω. The optimization process was due to attaching a weight to each rule. The configuration of these weights was transformed into an optimization problem with semantic distances also derived from Ω.

By (i) to (iii) we extended an approach originally stemming from semi-automatic ontology engineering and conceptual clustering [BiNeCa].

The work presented in this paper should be continued with experiments. An evaluation of the experimental results from our point of view should investigate the following questions:

- do different distance measures fulfilling the requirements 3.2 (i)-(iv) yield very different enrichment results?
- what is a sensible choice for the threshold *T*? To answer this question, we must investigate, how many concepts are proposed as candidate concepts with different thresholds *T*.
- does the quality of the candidate concepts satisfy the needs of an ontology engineer?

In our opinion the latter question is crucial. Roughly speaking, there are two ways of finding answers to this evaluation question. We could take parts of existing large ontologies and see, if the parts left out evolve through our enrichment process. On the other hand, we could ask a knowledge engineer to build a (small) ontology and comment the candidate concepts found afterwards. This evaluation should include answering the questions, which concepts did the engineer expected to be added, if they did appear in the enrichment results, if the enrichment results contained unsuspected high-quality concepts and on the contrary bad candidates.

Also future work should experiment either with more complex concept extraction from the texts (like attributed nouns) or extending the candidates due to high similarity to another candidate concept.

From a more abstract point of view, future work also includes applications of the approach to ontology merging and mapping, i.e. to the question, what happens if we take more than one ontology Ω as basic input. Finally, the question arises if the approach presented here benefits automatic ontology generation, i.e. if there are also optimization problems to be stated and solved in this research area.

References

[BiNeCa] Bisson, G. and Nedellec, C. and L. Cañamero: "Designing clustering methods for ontology building - The Mo'K workbench", *Proceedings of the Ontology Learning ECAI-2000 Workshop*, August 2000

[ChGr] G. de Chalendar and B. Grau: "SVETLAN' or how to Classify Words using their Context", *proceedings of the 12th International Conference on Knowledge Engineering and Knowledge Management, EKAW 2000*, Juan-les-Pins, France, October 2000

[Gr] T.R. Gruber: "Towards principles for the design of ontologies used for knowledge sharing", International Journal of Human-Computer Studies, **43**, 1995

[Gu] N. Guarino: "Understanding, Building and Using Ontologies. A Commentary to "Using Explicit Ontologies in KBS Development"", *International Journal of Human and Computer Studies*, **46**, 1997

[Ho] I. Horrocks: "Using an expressive description logic: FaCT or fiction?" , *Principles of Knowledge Representation and Reasoning: Proceedings of the Sixth International Conference (KR'98)*, San Francisco, California, June 1998.

[IDS] Institut für Deutsche Sprache: *www.ids-mannheim.de*

[Leng] K. Lengnink: "Formalisierung vom Ähnlichkeit aus Sicht der formalen Begriffsanalyse", *PhD Thesis, Technical University of Darmstadt, Shaker Verlag*, 1996

[OG] Open GALEN: *www.opengalen.org*

[StM] Steffen Staab, Alexander Mädche: Axioms are objects, too - Ontology Engineering Beyond the Modeling of Concepts and Relations *Proceedings of the ECAI 2000 Workshop on Ontologies and Problem-Solving Methods* Berlin, 2000

[Tv] A. Tversky: "Features of similarity", *Psychological Review* **84**, 1977

An operator-based account of semantic processing

C. Anton Rytting
Department of Linguistics
The Ohio State University
rytting@ling.ohio-state.edu

Deryle Lonsdale
Department of Linguistics
Brigham Young University
lonz@byu.edu

Abstract

This paper explores issues of psychological plausibility in modeling natural language understanding within Soar, a symbolic cognitive model. It focuses on constructing syntactic and semantic representations in simulated real time, with particular emphasis on word sense disambiguation (WSD). We discuss (i) what level of WSD should be modeled and (ii) how to use resources such as WordNet to inform these models. A preliminary model of coarse-grained WSD is included to show how syntactic, semantic, and other knowledge sources interact in Soar. Finally, we explore issues of interleaving, learning, and integrating other WSD approaches with Soar's native model of learning.

1 NL-Soar, WordNet and WSD

This paper argues that the principled use of WordNet within an operator-based framework yields interesting results in modeling semantic processing in general, and particularly where issues of word-sense disambiguation are concerned. It outlines the system's semantic processing module including the use of corpus data in determining semantic constituent compatibility constraints. Current and projected research issues are explored, including further integration with other semantic processing systems and the testing of empirical claims regarding the degree of syntax/semantics interaction. Finally, the role of learning in modeling systems is discussed.

1.1 NL-Soar

Natural-Language Soar (NL-Soar) is a natural-language extension of Soar, a symbolic cognitive modeling system (Newell, 1990). Operators (in the traditional AI sense) provide the basic processing pulse of the system, encapsulating a set of associated low-level actions that effect state changes in an agent's world model. The purpose of NL-Soar is to perform large-scale, robust language comprehension and production in a psychologically plausible way.

One goal of NL-Soar is to model language comprehension as a real-time process. NL-Soar (like humans) receives each input sentence word by word, and each word must be processed (that is, integrated into the larger syntactic and semantic representations) before it decays from a temporary perceptual buffer. NL-Soar must therefore begin syntactic and semantic processing incrementally, without knowing the whole sentence beforehand. Observable effects follow; for example, its syntactic module has been used to predict the point of comprehension failure on garden-path sentences (Lewis, 1993).

This paper reports on current development of the semantic module. In particular, it examines how the operator-based approach

provides a flexible framework for modeling and testing various assumptions and theories about human semantic processing.

1.2 WordNet

NL-Soar has recently been fitted with the WordNet lexical knowledge base (Miller, 1993). This has greatly enhanced NL-Soar's capabilities, but adds a number of complications of both practical and theoretical import (Lonsdale and Rytting, 2001).

Originally designed and organized according to psycholinguistic principles, WordNet has a number of advantages over traditional lexicons in modeling the human lexicon. Most of its 100,000 or so different concepts (or synonym sets) are arranged into a semantic hierarchy. Near the top of this hierarchy are 45 lexical files (originally used for managing the lexicographers' task) based on collocation patterns: 26 noun classes, 15 verb classes, and 4 classes for other parts of speech. Examples of these classes are `v-change`, `v-communication`, `n-artifact`, and `n-food`.

Previous versions of NL-Soar used a well-known set of eight primitive semantic categories (Jackendoff, 1990), but had no principled method for assigning new vocabulary into these eight classes. Now WordNet's lexical files serve as semantic categories in NL-Soar's semantic module, adding completeness and tractability to the approach.

Since the relation between words and concepts is many-to-many, there must be a way to determine, given a word in its context, which concept is the one being conveyed—in other words, what sense of the word is being used. This task is known as word sense disambiguation (WSD).

1.3 Word Sense Disambiguation

WSD is still an active area of research within computational linguistics (Kilgarriff and Palmer, 2000). However, it is not clear to what extent it is really a problem within psycholinguistics, and how much it is a carry-over from traditional lexicography. Some hold that word-senses do not really exist as discrete entities, and hence WSD is a manufactured problem (Melby, 1995; Kilgarriff, 1997; Hanks, 2000).

Psycholinguistic research also provides some cautionary evidence that WSD (and even part-of-speech disambiguation) may not always be a necessary part of human language comprehension (Wilks, 2000). For example, it has been shown that priming across word senses, an experimental effect commonly used to support the existence of word senses in the mind, is not always reliable. The use of words in a less-frequent context may still prime more frequent senses of the word, but more frequent contexts will not always prime less frequent senses (Williams, 1992). This suggests that less-frequent word senses may be stored as extensions of more central definitions, rather than as separate entities.

Another argument against word senses as an inherent aspect of lexical knowledge is the lack of agreement between naïve native speakers. On the one hand, word-sense tagger reliability with trained experts reaches as high as 95% even for fine-grained senses (Kilgarriff and Palmer, 2000). However, the agreement between naïve speakers is significantly lower, less than 80% (Fellbaum et al., 1998).

These studies suggest that fine-grained WSD may not be an appropriate problem for a psycholinguistic model. Nevertheless, some account of the phenomenon is necessary in order to construct a workable semantic model given a large, polysemous vocabulary. If fine-grained WSD is not feasible or plausible, an approximate model of coarse-grained WSD still seems advisable (Wilks, 1997).

Moreover, some research supports limited, course-grained polysemy in the mental lexicon. The results of one sentence-based semantic classification experiment argue for the existence of a few stable word-senses (typically around three) for many nouns, rather than an exhaustive fine-grained inventory (Jorgensen, 1990).

This suggests that it would be better to model disambiguation between a few broadly defined senses of each word than to use dictionary definitions; or alternatively, to dis-

ambiguate instances of high-level ontological concepts. This latter approach might be termed semantic class disambiguation (SCD), and is exemplified by the use of Roget's categories (Yarowsky, 1992), and the use (Krymolowski and Roth, 1998) of CoreLex, a set of 126 semantic classes derived from WordNet (Buitelaar, 1998).

2 Semantic representation in NL-Soar

The present version of NL-Soar follows this second strategy, using WordNet's 45 general divisions as its semantic classes. This first approximation at a semantic model enables exploiting WordNet's vocabulary coverage without committing to the plausibility of its fine-grained concepts. As the WSD modeling granularity issue evolves, the general techniques tested here will be refined.

A semantic link in NL-Soar consists of three parts: an assigning entity, a receiving entity, and a semantic role or link type (either external, internal, internal2, or property). For example, NL-Soar's semantic model of the sentence *The dog barked.* could be schematically represented thus:

```
bark.v-body (assigner) --External--> dog.n-animal (receiver)
```

NL-Soar's semantic module currently uses three criteria to link concepts: syntactic and morphological constraints, frequency rankings, and semantic pairing constraints.

First, NL-Soar processes the syntax of an utterance and then constrains the possible semantic readings to those consistent with the syntactic model. Proposed links reflecting relevant criteria are preferred over those which do not. Syntax thus plays a role in SCD (Lonsdale and Rytting, 2001); section 5 below discusses further implications.

Research indicates that word-sense frequency may provide a plausible default ranking strategy (Fellbaum et al., 1998). When given a frequency-ordered list, subjects often pick the first sense which reasonably fits the context. This suggests that in absence of other factors (such as discourse context) word sense frequency is an important criterion for assigning meaning to a word.

For a given word, NL-Soar ranks the frequency of each semantic class according to the ranking of most frequent WordNet sense in that class. For example, the word *tooth* has 5 WordNet senses: 1 and 3 fall under the semantic class `n-body`; senses 2 and 5 fall under `n-artifact`; and sense 4 under `n-act`. Hence, NL-Soar's semantic module assigns *tooth* to these four semantic classes, with `n-body` treated as most frequent, followed by `n-artifact` and `n-act`.

The base frequency of semantic classes is not sufficient to determine the most plausible semantic links, however. In the absence of a robust model of real-world knowledge, NL-Soar uses corpus-derived "semantic collocation constraints" determined by frequencies of pairings of certain semantic classes. The strategy is similar to that of (Resnik, 1997; Ciaramita and Johnson, 2000); however, the frequencies are (currently) calculated over semantic classes rather than over specific nouns and verbs.

Canonical semantic collocations are found by searching SEMCOR, a WordNet-sense-tagged version of the Brown Corpus (Francis and Kučera, 1982), for examples of each verb class. For each verb class, a sample of approximately 50 verb instances is gathered. For each of these instances, the semantic classes for the nouns that fill the verb's external and internal roles are recorded.

For example, the following semantic pairings were found to be the most common for the semantic classes `v-body`:

```
External v-body: n-artifact, n-body, n-person
Internal v-body: n-artifact, n-attribute, n-body
```

3 Semantic Processing in NL-Soar

As a given word undergoes semantic processing, all of its possible semantic classes and parts of speech are considered, in order of frequency, as determined above. The first one to pass the selectional pairing constraints is accepted as the best choice and fitted into the semantic model.

However, NL-Soar does not perform SCD on individual words directly. Since the semantic constraints license concept pairings be-

tween noun classes and verb classes, NL-Soar must disambiguate noun-verb pairs as a unit, not separately.

For example, consider the sentence *The tooth hurts.* The appropriate (and most frequent) semantic pairing between *tooth* and *hurt* is:

```
(1) hurt.v-body --External-->  tooth.n-body.
```

Operators also propose other possible links:

```
(1') hurt.v-body --Internal-->  tooth.n-body.
(2) hurt.v-body --External-->  tooth.n-artifact
(3) hurt.v-change --External-->  tooth.n-body
(4) hurt.v-change --External-->  tooth.n-artifact (etc.)
```

but these are not considered first, since they either do not match the syntax (in case of 1') or are not most frequent. Since `(v-body, External, n-body)` is a relatively common semantic pairing, it is licensed by the constraints, and the pairing (1) is linked with the ongoing semantic model.

Although the various potential semantic links are initially considered in parallel, semantic constraints are tested serially in NL-Soar. Word-sense frequency determines which links are tested first. But if both words in the link are semantically ambiguous, it is unclear which word's semantic-class frequency ranking should carry more weight. A number of criteria could be used: e.g., part of speech, syntactic role (assigner vs. receiver), and degree of polysemy. This study assumes temporal ordering as the key factor. Since the earlier of the two words being linked is (in the absence of extra-sentential context) less constrained by previously heard context than the later word, it seems that its baseline word-sense frequency information should apply more pertinently by default. Hence, the most frequent sense of the first ambiguous word should most strongly constrain the acceptible word senses of later words.

This claim is not essential to the model but rather another working hypothesis in need of empirical verification. There seems to be preliminary neurolinguistic support: an event-related potential (ERP) study shows that in normal, congruous sentences, N400 anomaly effects for low-frequency words are greatest sentence-initially, and decrease thereafter (Van Petten, 1995). See section 6 for further implications.

4 Encoding and testing constraints

The current NL-Soar approach has been implemented and is being tested. The first preliminary test consisted of ten sentences: five from the same source as the samples used to create the semantic collocation constraints, and five from a different portion of SEMCOR. A number of syntactic simplifications were imposed on the test sentences, in order to focus on lexical and semantic issues. All simplified sentences consisted of noun phrases and finite verbs.

A preliminary trial of these ten simplified sentences was run on NL-Soar as it currently stands. After syntactic coverage difficulties are factored out, the semantic module provided three sentences with an exact match with the SEMCOR tagging, two with semantically plausible mismatches, and two with semantically implausible mismatches. These results must be interpreted cautiously, since the semantic constraints were not complete for all verb classes. However, NL-Soar provides an acceptable parse on half of the sentences, a performance better than chance when the number of variables involved are considered.

Brief mention of a few examples is instructive in appreciating the issues that arise from using WordNet and in particular SEMCOR as a data source. Each sentence was run using, for each word, all possible senses from the WordNet inventory (e.g. 15 WordNet classes for verbs). Semantic constraints were tested for external roles; internal roles were not addressed. Each sentence of interest was taken from SEMCOR, and most were simplified somewhat to abstract away from nonessential material; however, the same semantic classes were preserved for each semantic role. For example, a SEMCOR sentence such as "The antagonists were instructed to behave in such ways as to upset the manager and get him to operate in a manner for which he had been previously criticized." was simplified to "The antagonists behave annoyingly." For each sentence, the SEMCOR tags were compared with those assigned by NL-

Soar. Additional thematic role information not provided in SEMCOR (i.e. E(xternal) for agent, or deep subject, and I(nternal) for patient/theme, or deep object) but calculated by NL-Soar is also indicated in the examples that follow. An asterisk indicates that the role is left unfilled (e.g. an optionally empty Internal role). Mismatches are indicated with italics. Linkages were categorized into one of four classes: (1) match with SEMCOR, (2) acceptable mismatch (i.e., a plausible reading), (3) unacceptable mismatch, (4) fail.

Consider, for example, the (simplified) sentence *The exercise exhausts the child.*. NL-Soar's semantic linkage exactly reflects that of SEMCOR's:

```
exercise.n-act-E-exhaust.v-body-I-child.n-person
```

The class `exhaust.v-body` is the most frequent class for *exhaust* (and also the most frequent sense, out of 5 senses). Constraints that follow from semantic processing and semantic cooccurrence licensing allow both External and Internal matching[1].

Another sentence, *Clergy wear grey.* poses another interesting problem. It is an unsurprising result of training on a small data source—namely, that the class `n-group` never showed up in an External role for `v-body`. Yet it shows up in a total of three of the test sentences. This type of problem is easily overcome; the addition of one more precondition to an existing production to relax this constraint improves results tremendously. Furthermore, the use of some sort of underspecification scheme would correct this problem automatically, given that `n-group` is a natural extension of `n-person`.

One example of an unacceptable mismatch may be due to the quality of SEMCOR tagging instances. The sentence *Antagonists behave badly.* has the following SEMCOR and NL-Soar linkages:

```
SEMCOR:  antagonist.n-group-E- behave.v-body-I-*
NL-Soar: antagonist.n-body-E- behave.v-body-I-*
```

Here NL-Soar has taken the `v-body` class for *behave*, and consequently prefers an `n-body` linkage for the subject *antagonist*; even though the `n-group` class is more frequent, collocation constraints do not prefer linking `n-group` with `v-body`. On the other hand, *behave* has another class, `v-social`; if SEMCOR would have had this arguably more appropriate sense tagged instead, NL-Soar would have been able to perform the proper linkage with `n-group`.

WordNet's exhaustiveness often causes homonymy-related problems; words such as *at*, *a*, *an*, *are*, etc. have nominal readings that are not readily identifiable by non-lexicographers. Most of these lexical, open-class readings (e.g. *are* as a unit of measure) complicate syntactic and semantic processing, and consequences were seen while testing SEMCOR-derived sentences. The solution is straightforward in NL-Soar: a production was introduced that automatically prefers the functional reading for obscure homographs like these.

Finally, consider the sentence *The epidemic infected the citizens.*, an instance where the frequency as a measure of the primary or "central" word-sense fails. Note the SEMCOR and NL-Soar linkages:

```
SEMCOR: epidemic.n-event-E-infect.v-body-I-citizen.n-person
NL-Soar epidem.n-event-E-infec.v-cognition-I-citiz.n-person
```

Intuitively, the "central" meaning of infect is the medical (or `v-body`) reading: "communicate a disease to". However, a metaphorical `v-cognition` reading "fill with revolutionary ideas" is listed as the most frequent one in WordNet. Oddly enough, the original context for the test sentence (*"It is like a mysterious epidemic, which ... spreads ... through the whole town until all have been infected by it."*) was also embedded in a simile. This meshes well with the prediction that one of the chief problems facing a (fairly literal-minded) artificial agent will be detecting and processing metaphor and figurative language.

[1] An interesting future extension would be to assign the External and Internal with STIMULUS and EXPERIENCER theta-roles, or otherwise relating the External `n-act` with the Internal `n-person`, since surely their co-occurrence is not independent. One weakness of the current sampling mechanism is to treat External and Internal roles independently, and hence lose this potentially useful information. On the other hand, trying to build in too much information could lead to a sparse data problem, unless more automatic methods of sampling are developed.

As the situation stands now, "training" and "performance" are articifially separated as hand-sampling of collocation patterns replaces (or rather, holds a place for) the processing mechanisms native to Soar. A better way to look at the problem of Soar processing is the cominbation of intelligent use of heuristics and remembering the results from applying these heuristics. The Soar framework is specifically designed to support a panoply of strategies and processing modalities, and to manage the decision process involving which to use when. For exmaple, one (or even several) of the previously mentioned WSD techniques could be accessed as strategy when NL-Soar is unsure of the most appropriate word in context. NL-Soar could first search for "give-away" bigrams (as in (Pedersen, 2001)). If no such bigrams are in the previously heard context, then other techniques such as those mentioned above could be used. If insufficient data is available for a particular word, then a strategy such as (Clark and Weir, 2001) could be used to find a more suitable level on the hierarchy on which to repeat the reasoning process. There need be no assumption of a single WSD technique appropriate for all occasions.

5 Modeling: what is the target?

The approach sketched above is, of course, not the only one possible. Instead, it was adopted as a baseline strategy to serve while the basic semantic processing architecture set in place. In fact, several aspects of psycholinguistic groundwork still need to be done in the areas of WSD and on-line semantic processing; comprehensive modeling must either await such results, or else propose models that might inspire (or challenge) experimentalists to verify or refute claims put forward by speculative modeling efforts.

As computational and architectural issues remain an area of focus, an operator-based approach can support whatever solution might be followed in addressing these issues. For example, a determination of the most appropriate set of off-the-shelf semantic classes (or indeed entire ontologies) would be useful; possibilities include the Roget and CoreLex categories mentioned above. Several off-the-shelf WSD algorithms also remain to be explored (Yarowsky, 1992; Resnik, 1997; Ciaramita and Johnson, 2000; Clark and Weir, 2001).

The relationship between syntactic and semantic processing has been a subject of considerable debate within psycholinguistics. Two extreme positions exist; the first assumes complete precedence of syntax over semantics. Doubt has been cast on on the plausiblity of this extreme; consider this popular example: *After finding the book on the atom, Kim decided that the library really wasn't as bad as people had been claiming.* (Hirst, 1987)

It seems clear that *book* and *atom* must be matched with real-world concepts (along with the most plausible meaning of *on*) before the complete sentence is heard. While no-one likely takes the separation to this extreme, it would appear that some models might adopt it for reasons of simplicity.

The other extreme is that of complete two-way interaction between syntax and semantics. A considerable degree of interaction is claimed by the constraint satisfaction approach (MacDonald et al., 1994; Trueswell et al., 1994; Burgess and Lund, 1997) and assumed in some connectionist approaches. However, this claim is still controversial, and other works provide evidence against it (Frazier and Fodor, 1978; Ferreira and Clifton, 1986; Clifton et al., 1991). While much psycholinguistic research remains to be made, the parallel development of flexible and versatile frameworks will help current and future modeling efforts.

6 Interleaving and Interactivity

NL-Soar's use of operators, which can be interleaved, allows it to model an intermediate position between these two extremes. While NL-Soar currently assumes that syntax guides semantics, this need not be a temporal ordering. Rather, through interleaving of the syntactic and semantic processes, the semantics module may begin processing the syntactic module's early results while the syntactic module processes later parts of the sentence.

Misanalyses in either syntax or semantics can be corrected (to a point) through "snip" operators, but beyond that point are predicted to result in garden pathing. NL-Soar's framework is thus able to account for a great deal of interaction between modules while preserving the separation commonly assumed; however, it could be modified to accommodate full interaction between modules, should further evidence come out in favor of that hypothesis.

This leads into a further claim, mentioned in section 3, that frequency effects play a larger role in the disambiguation of early words than later words in an utterance (outside of prior disambiguating context). This claim seems consistent with either the strong interactive (constraint satisfaction) stance or with NL-Soar's current interleaving framework. It predicts the existence of "semantic garden paths" depending on early rejection of a low-frequency meaning of the first word, followed by late disambiguating material. Again, psycholinguistic exploration of this area would be valuable to establish a target for semantic modeling efforts.

It is expected that, once other issues regarding other aspects of WSD are settled, NL-Soar can be a useful tool in investigating this claim, in conjunction with corresponding psycholinguistic experiments.

7 Transfer and learning

One important property of operators is that their effects can be cast in terms of hierarchical subgoal satisfaction. When the knowledge about how to attain a goal (such as computing a semantic linkage) is lacking in a holistic sense, NL-Soar can subgoal and test various possibilities in an attempt to achieve the top-level goal. This may involve extensive levels of decomposition, each of which may in turn involve lower-level operators, constraint checking, and deliberation. Though this process may be costly in terms of processing time and memory requirements the first time it is carried out, the system also provides a significant capability for such situations—machine learning. Results (whether successful or not) can be "chunked up" automatically during subgoaling, and the results will then be available for subsequent processing. This enables the so-called recognitional mode of operation in certain contexts.

Learned operators for semantic processing thus are often available, and they leverage experience gained in accepting or rejecting certain pairings of semantic substructure as described in section 2. WordNet semantic classes, with their associated frequency information, thus are incorporated into the learning process.

Interesting issues which cannot be addressed here involve the granularity of the semantic operators: to what degree should the semantic class categories which are used as linking constraints be generalized? To what extent can the hand-derived corpora be abstracted away from, in order to increase the coverage? Language acquisition research in such areas as categorization and semantic overgeneralization could be relevant in exploring these issues.

An intriguing learning-based direction involving the semantic processing mentioned in this paper will be possible when more than one WSD method is implemented, as discussed earler. These strategies could even inform one another through memorized chunks and meta-learning strategies. For example, if one WSD mechanism is successful at identifying a unique word sense at a high degree of confidence, then it could train other strategies, perhaps by remembering salient bigrams for future expansion of Pedersen's bigram heuristic. The system may even be able to learn when some heuristics (e.g., one-sense-per-discourse (Gale et al., 1992) or one-sense-per-collocation (Yarowsky, 1993)) are valid, when they should be avoided, and when one strategy is preferable to another.

Another advantage NL-Soar derives from the operator-based architecture is the possibility of transfer. Tasks which leverage similar low-level processes can share operators for these processes; hence, cross-modal application of operators may account for facilitation and inhibition across tasks, and allow the generalization of semantic or syntactic facts be-

yond the specific context of a learning corpus. Bootstrapping of semantic knowledge through supervised learning is thus one avenue of future work.

8 Conclusions

This paper has surveyed the semantic processing component of a system that is based on an operator framework. Operators provide the flexibility to model language use according to various theoretical and empirical approaches. The system currently performs semantic model construction based on fairly commonly held assumptions, and based on cooccurrence constraints observed in hand-extracted frames from annotated corpora. This process can ultimately be automated via a wide range of possible approaches. Interesting related topics such as learning, interactivity, operator granularity, task-related transfer, and priming are currently being explored.

References

Paul Buitelaar. 1998. *CORELEX: Systematic Polysemy and Underspecification*. Ph.D. thesis, Brandeis University, Computer Science Department, February.

Curt Burgess and Kevin Lund. 1997. Modelling parsing constraints with high-dimensional context space. *Language and Cognitive Processes*, 12(2/3).

Massimiliano Ciaramita and Mark Johnson. 2000. Explaining away ambiguity: Learning verb selectional preferences with bayesian networks. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 187–193.

Stephen Clark and David Weir. 2001. Class-based probability estimation using a semantic hierarchy. In *WordNet and other lexical resources: applications, extensions, and customizations*. Association for Computational Linguistics. NAACL-2001.

Charles Clifton, Jr., Shari Speer, and Steven P. Abney. 1991. Parsing arguments: Phrase structure and argument structure as determinants of initial parsing decisions. *Journal of Memory and Language*, 30:251–271.

Christiane Fellbaum, Joachim Grabowski, and Shari Landes. 1998. Performance and confidence in a semantic annotation task. In Christiane Fellbaum, editor, *WordNet: An electronic lexical database*, pages 217–238. MIT Press, Cambridge, MA.

Fernanda Ferreira and Charles Jr. Clifton. 1986. The independence of syntactic processing. *Journal of Memory and Language*, 25:348–368.

Winthrop Nelson Francis and Henry Kučera. 1982. *Frequency analysis of English usage: Lexicon and grammar*. Houghton Mifflin, Boston.

Lyn Frazier and Janet Dean Fodor. 1978. The sausage machine: A new two-stage parsing model. *Cognition*, 6:291–325.

William Gale, Kenneth Church, and David Yarowsky. 1992. One sense per discourse. In *Proceedings of the DARPA Speech and Natural Language Workshop*, pages 233–237, New York. Harriman.

Patrick Hanks. 2000. Do word meanings exist? *Computers and the Humanities*, 34(1-2):205–215.

Graeme Hirst. 1987. *Semantic Interpretation and the Resolution of Ambiguity*. CUP, Cambridge.

Ray Jackendoff. 1990. *Semantic Structures*. Current Studies in Linguistics Series. MIT Press.

Julia C. Jorgensen. 1990. The psychological reality of word senses. *Journal of Psycholinguistic Research*, 19(3):167–190.

Adam Kilgarriff and Martha Palmer. 2000. Introduction to the special issue on SENSEVAL. *Computers and the Humanities*, 34(1-2):1–13.

Adam Kilgarriff. 1997. "I don't believe in word senses". *Computers and the Humantities*, 31:91–113.

Yuval Krymolowski and Dan Roth. 1998. Incorporating knowledge in natural language learning: A case study. In Sanda Harabgiu, editor, *Usage of WordNet in natural language processing systems: Proceedings of the workshop*, pages 121–127. COLING-ACL-98.

Richard Lewis. 1993. An architecturally-based theory of human sentence comprehension. Technical Report CMU-CS-93-226, CMU.

Deryle Lonsdale and C. Anton Rytting. 2001. Integrating WordNet in NL-Soar. In *WordNet and other lexical resources: applications, extensions, and customizations*, pages 162–164. Association for Computational Linguistics. NAACL-2001.

Maryellen C. MacDonald, Neal J. Pearlmutter, and Mark S. Seidenberg. 1994. The lexical nature of syntactic ambiguity resolution. *Psychological Review*, 101:676–703.

Alan Melby. 1995. *The Possibility of Language*. John Benjamins, Amsterdam.

George Miller. 1993. WordNet: A lexical database for English. In *Communications of the ACM*, volume 38.

Alan Newell. 1990. *Unified Theories of Cognition*. Harvard University Press.

Ted Pedersen. 2001. A decision tree of bigrams is an accurate predictor of word sense. In *2nd Meeting of the North American Chapter of the Association for Computational Linguistics: Proceedings of the Conference*, pages 79–86. Association for Computational Linguistics.

Philip Resnik. 1997. Selectional preference and sense disambiguation. *Proceedings of the ANLP-97 Workshop: Tagging Text with Lexical Semantics: Why, What, and How?*

John C. Trueswell, Michael K. Tanenhaus, and Susan M. Garnsey. 1994. Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language*, 33:285–318.

Cyma Van Petten. 1995. Words and sentences: Event-related brain potential measures. *Psychophysiology*, 32:511–525.

Yorick Wilks. 1997. Senses and texts. *Computers and the Humanities*, 31:77–90.

Yorick Wilks. 2000. Is word sense disambiguation just one more NLP task? *Computers and the Humanities*, 34:235–243.

John N. Williams. 1992. Processing polysemous words in context: Evidence for interrelated meanings. *Journal of Psycholinguistic Research*, 21:193–218.

David Yarowsky. 1992. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of COLING-92*, pages 454–460, Nantes.

David Yarowsky. 1993. One sense per collocation. In *Proceedings of the ARPA Human Language Technology Workshop*, pages 266–271. Morgan Kaufman.