

A Task-based Approach for Ontology Evaluation

Robert Porzel and Rainer Malaka ¹

Abstract. The need for the establishment of evaluation methods that can measure respective improvements or degradations of ontological models - e.g. yielded by a precursory ontology population stage - is undisputed. We propose an evaluation scheme that allows to employ a number of different ontologies and to measure their *performance* on specific tasks. In this paper we present the resulting task-based approach for quantitative ontology evaluation, which also allows for a bootstrapping approach to ontology population. Benchmark tasks commonly feature a so-called *gold-standard* defining perfect performance. By selecting ontology-based approaches for the respective tasks, the ontology-dependent part of the performance can be measured. Following this scheme, we present the results of an experiment for testing and incrementally augmenting ontologies using a well-defined benchmark problem based on a evaluation gold-standard.

1 Introduction

The employment of ontologies for system-wide representations, inferencing operations or for defining interface specifications has gained paramount importance in the development of multi-domain multi-modal dialogue systems [11, 26, 10]. Still many well-known problems remain. Two critical issues concern:

- the knowledge-acquisition bottleneck, where ontology evolution, learning and population come into play,
- the lack of formal means for evaluating a given ontology or ontology improvement in the light of the specific tasks at hand.

We will follow the general distinction between qualitative and quantitative ontology evaluation [2] and propose a quantitative scheme for a task-based ontology evaluation and population approach. The underlying question in such a quantitative evaluation is how effective a given ontology is in the light of a well-defined task.

Effective in this sense is straight-forward. It means that - if an ontology is to be employed for a given task - it can be used to perform better or worse in a measurable way. In order to measure such effectiveness the operations - whose outcome depends on the ontological model - need to be constant throughout an evaluation. An evaluation suite should therefore, be selected such that the measurable output concerning the suite's task depends as much as possible on the ontology used.

In this paper we discuss the feasibility to test and incrementally augment ontologies given a well-defined benchmark problem based

on a so-called evaluation *gold-standard*. It is important to point out that the type of ontology evaluation proposed herein can - at least at the moment - be carried out only in respect to a given task at hand, which the specific ontology has to solve. A task-independent automatic evaluation, in our minds, still remains an elusive goal for which a general solution does not exist² [8]. Such a maximal evaluation (or test) of an ontology can judge an ontology at least on three basic levels:

- the scope (or fit) of the vocabulary (we will refer to these ontology classes as *concepts*);
- the well-ness (fit) of the taxonomy, i.e. the generalization or *isa* hierarchy; and
- the adequacy of the non-taxonomic relations, i.e. the fit of the semantic relations.

Also aggregate evaluations combining the various levels are also possible. More importantly for us, is that there are meaningful translations of the commonly used *error rates* of insertions, deletions and substitutions. These error rates are used in automatic speech recognition [15], as well as concept tagging [12, 13]. After a brief overview of the state of the art in Section 2 we will introduce the fundamental schema and the ingredients for task-based ontology evaluations in Section 3, followed by an example evaluation with respect to the level of semantic relations (and its ontology population fall-out) in Section 4ff.

2 State of the Art

The need to develop a clear set of evaluation methodologies is widely acknowledged. The qualitative type of evaluations [7, 6] basically rely on user or expert judgments, whereby it is left open whether ontology engineers, system users or domain experts ought to be the judges. Additionally, to judge ontologies in terms of the principles on which their design has been based also bases on criteria defined by the "external semantics" which, again, has to be evaluated by human experts. There are even more general problems that arise from such principle-based approaches [20].

As our concern in this work is on quantitative evaluation for measuring the performance of an ontology for a given task, we will not discuss the valuable work on measuring similarities between ontologies or evaluating a given ontology against a pre-defined ontological *gold-model* [14, 18].

As we will discuss in Section 7, the ontology population that - in a sense - falls out of our evaluation is comparable to automatic means of ontology learning or ontology population [28] only with respect to

¹ European Media Laboratory GmbH, Schloss-Wolfsbrunnengasse 33, 69118 Heidelberg, Germany, {firstname.lastname@eml-d.villa-bosch.de}

² It might even be impossible, as it is widely acknowledged that ontology engineering, employment and evaluation has many task-dependent features and constraints.

the basic levels. Our approach enables us to evaluate improvements or the adequacy of ontological models (or changes brought about by some level-specific learning approach) for each level respectively and is independent from the automatic or manual means by which the ontology was crafted.

The work presented herein is, however, in principle related and in perfect agreement to that of Brewster et al [2], who state that:

The establishment of a clear set of simple application suites which would allow a number of different ontologies to be slotted in, in order to evaluate the ontologies would be an important research step. (ibid:2)

Brewster et al (2004) provide a data-driven approach that enables an ontology evaluation against given textual corpora. Employing this promising approach they arrive at a measure for *ontological fit*. This constitutes a measure of vocabulary overlap between the concepts contained in a given ontology and the terms extracted (by means of a latent semantic-based clustering algorithm) and expanded by means of a two step hyponym WordNet look up. The necessary alignment or mapping between concepts and terms is done by manual annotation³. For measuring the taxonomic fit the authors employ the *tennis metric* proposed by Stevenson [29].

An ontology, however, provides more than a vocabulary of entities and their generalization hierarchy. A substantial amount of its expressive and inferential capabilities (at least for natural language processing applications) lies in the non-taxonomic relations that hold between the concepts. For evaluating this aspect of an ontological model no solution has been proposed so far. In this work, we will examine the feasibility to fill this gap by conducting performance- or task-based evaluations that yields a measure of how well the vocabulary, taxonomy and the non-taxonomic relations are modeled for a given task at hand. For this, we will sketch out the necessary framework and its ingredients in Section 3 and describe their specific instantiations in our experiments in Section 4. The corresponding experiments and results are given in Section 5 and Section 6 followed by a discussion and concluding remarks in Section 7.

3 Task-based Evaluations

Given that, one can examine ontologies with respect to their three basic levels: vocabulary (1), taxonomy (2) and (non-taxonomic) semantic relations (3) - as these levels are also subject to different respective learning approaches - we propose that: the common notion of error rates, such as found in word- or concept-error rates [15], known from previous work, suffices for each level of evaluation. In a task-based evaluation the results should show the following shortcomings:

- *insertion errors* indicating superfluous concepts, isa- and semantic relations,
- *deletion errors* indicating missing concepts, isa- and semantic relations, and
- *substitution errors* indicating *off-target* or ambiguous concepts, isa- and semantic relations.

Given appropriate tasks and maximally independent algorithms operating on the ontology in solving these tasks and given the task evaluation gold standards we can calculate the error rates corresponding to specific ontological shortcomings as shown in the overview

³ Unfortunately they authors do not provide a measure for inter-annotator agreement on this task, which, as our experiences show is also not at all trivial.

of the translation of error rates to the three basic ontological levels displayed in Table 1.

level	insertion	deletion	substitution
1	irreverent concepts	omitted concepts	ambiguous concepts
2	isa too coarse	isa too fine	isa too polygamous
3	irreverent relations	missing relations	indirect relations

Table 1. Overview of the errors on the respective levels

With this we can provide performance measures that can:

- evaluate one or more ontologies in terms of their *performance* on a given task (ideally to measure only the ontology-specific aspect of the performance),
- quantify the respective gains and losses of the insertion, deletion and substitution errors,
- populate/improve the ontology as derived from the individual error type specific results, and
- re-evaluate the respective performance increases resulting from the improvements.

By applying this evaluation scheme we can test and measure the respective improvements that are brought about by learning approaches that target the same levels and issues in the ontology learning and population field. We can therefore roughly categorize these approaches as shown in Table 2.

level	(domain of) learning /population/evaluation	sample learning	sample evaluation
1	concepts vocabulary	[22, 28]	[2, 6]
2	hierarchy/granularity	[32]	[28]
3	semantic relations	[5]	in this work

Table 2. Overview of the learning and evaluation levels and sample approaches

Next, we describe the minimal elements and their specific constraints that are necessary for a task-based evaluation of an ontology and its entire range of relations. An overview of a such a generic task-based evaluation suite is given in Figure 1.

A Task: The task, certainly, needs to be sufficiently complex to constitute a suitable benchmark for examining a given ontology. Especially if the target of the evaluation is to include non-taxonomic relations as well, then we need to find tasks where the performance outcome hinges substantially on the way these relations are modeled within the ontology.

One (or more) Ontologies: This almost goes without saying, at least one ontology is needed for the type of evaluation proposed herein. However, note that one is sufficient, i.e. as an ontology is evaluated in terms of its *performance* on a given task, this can be done as a single ontology evaluation as well as an evaluation of how one ontology fares on the specific task as compared to another. It is, therefore, in principle the same paradigm as it is used in the TREC, MUC or SENSEVAL evaluations.

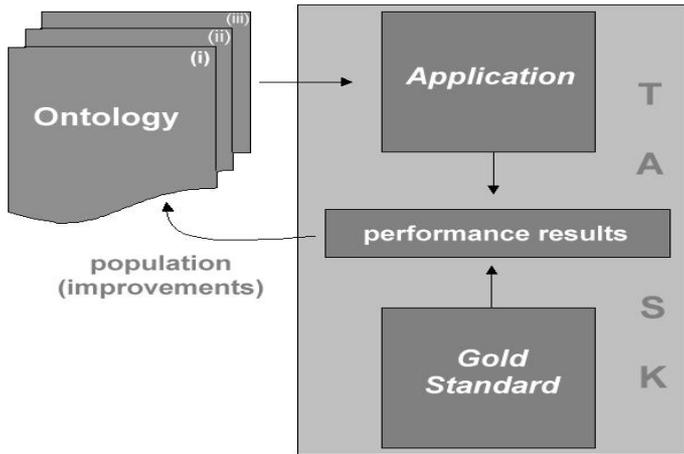


Figure 1. Test Suite Setup: Our setup requires a single task, application and gold standard and one or more ontologies

An Application: As an application we see the specific algorithm that uses the ontology to perform the task at hand. To foreshadow, in part, our conclusion, the untangling of algorithmic and ontology-related factors constitute the most difficult issue in our approach and it is vital that the algorithmic side is kept constant within an evaluation suite.

A Gold Standard: In order to evaluate the performance of any algorithms that produce so-called *keys*, whether they be part-of-speech tags, word senses or tag ontological relations, a given set of *answers* is needed. We call this perfectly annotated corpus of answers a *gold standard*.

4 Task-based Evaluations: An Example

Herein we describe how we have designed our ontology evaluation application-suite, its motivation, and background.

A Sample Task: For this evaluation suite we choose the task of tagging the *ontological relations* that hold between ontologically marked-up entities. This mark-up can be gained, e.g., from a concept tagging system (as described in [17]) and constitutes a form of sense disambiguation system, whereby the specific *senses* correspond to items of the ontology’s vocabulary.

The task can be thought of as an extension of the work by Gildea and Jurafsky [5], wherein the tagset is defined by entities corresponding to FrameNet *frame elements*. Therein, for example, given the occurrence of a *CommercialTransaction* frame the task lies in the appropriate labeling of the corresponding roles, such as *buyer*, *seller* or *goods*.

Additionally the task discussed herein features similarities to the *scenario template task* of the Message Understanding Conferences [19]. In this case predefined templates are given (e.g. *is-bought-by*(COMPANY_A, COMPANY_B) which have to be instantiated correctly, i.e. in a phrase such as:

“Stocks sky-rocketed after Big Blue acquired Softies . . .”

the specific roles, i.e. *Big Blue* as COMPANY_B and *Softies* as COMPANY_A have to be put in their adequate places within the overall template. The task of concept tagging can be considered to be solved successfully if all lexical items that have ambiguous word-to-concept mappings, such as the word *kommt*, in a German utterance like “*wie kommt man dann in Heidelberg weiter*”,⁴ are tagged with their contextually adequate senses given in our case by the ontological class inventory, such as *MotionDirected* rather than *WatchPerceptualProcess*, which would be the appropriate concept for *kommt* in an utterance such as “*Was kommt im Fernsehen*”⁵.

In this experimental set-up we take the new and less explored task of ontology relation tagging in the sense of [24]. This means to label all previously disambiguated and concept-tagged words with non-taxonomic relations, such as shown in Figure 2 .

domain			
conceptSet	score	cwr	realScore
[Broadcast] [Channel] [RecordTapeDevice]	0.87	0.5	0.87
[Broadcast] [has-channel] [Channel]			
[Channel] [has-broadcast] [Broadcast]			
[RecordTapeDevice] [has-broadcast] [Broadcast]			
[RecordTapeDevice] [has-broadcast] [Broadcast] [has-channel] [Channel]			
[Broadcast] [TwoPointRelation] [Channel] [RecordTapeDevice]	0.43	0.67	0.43

Figure 2. Tagging Ontological Relations: Given a set of concepts (Broadcast, Channel, and RecordTapeDevice) the system has to tag concept pairs with their appropriate relations (has-channel, has-broadcast).

A Sample Ontology: The independent ontology used in the experiments described herein was initially designed as a general purpose model for knowledge-based NLP. It includes a top-level developed following specific ontological principles (see the procedure outlined by [27]) and originally covered the tourism domain encoding knowledge about sights, historical persons and buildings. Then, the existing ontology was adopted in the SMARTKOM project [31, 30] and modified to cover a number of new domains, e.g., new media and program guides, pedestrian and car navigation and more [11]. The top-level ontology was re-used with some slight extensions. Further developments were motivated by the need of a *process hierarchy*.

This hierarchy models processes which are domain-independent in the sense that they can be relevant for many domains, e.g., *InformationSearchProcess*. The modeling of Process as

⁴ A loose translation of the German utterance would be “How can one then continue on in Heidelberg”.

⁵ Which is translatable as “What comes on TV”.

a kind of event that is continuous and homogeneous in nature, follows the frame semantic analysis used in the FRAMENET project [1].

The slot structure also reflects the general intention to keep abstract and concrete elements apart. A set of most general properties has been defined with regard to the role an object can play in a process: *has-agent*, *has-theme*, *has-experiencer*, *has-instrument* (or *has-means*), *has-location*, *has-source*, *has-target*, *has-path*. These general roles applied to concrete processes may also have subslots: thus an agent in a process of buying (TransactionProcess) is a *buyer*, the one in the process of cognition is a *cognizer*. This way, slots can also build hierarchical trees. The property *has-theme* in the process of information search is a required *has-piece-of-information*, in presentation process it is a *has-presentable-object*, i.e., the item that is to be presented, etc.

A Sample Application: The ONTOSCORE software runs as a module in the SMARTKOM multi-modal and multi-domain spoken dialogue system [30]. The system features the combination of speech and gesture as its input and output modalities. The domains of the system include cinema and TV program information, home electronic device control as well as mobile services for tourists, e.g. tour planning and sights information.

ONTOSCORE operates on n-best lists of speech recognition hypotheses (SRH) produced by the language interpretation module out of the ASR word graphs. It has been evaluated successfully on a number of tasks, e.g. computing a numerical ranking of alternative SRHs and thus providing an important aid to the spoken language understanding component. More precisely, this task of ONTOSCORE in the system is to identify the best SRH suitable for further processing and evaluate it in terms of its contextual coherence against the domain and discourse knowledge. Additionally, the system has been evaluated for the tasks of classifying SRHs as coherent *versus* incoherent [9], correct *versus* incorrect [23], as well as concept tagging [17].

ONTOSCORE performs a number of processing steps. At first each SRH is converted into a *concept representation* (CR). For that purpose we augmented the system's lexicon with specific concept mappings. That is, for each entry in the lexicon either zero, one or many corresponding concepts were added. A simple vector of concepts - corresponding to the words in the SRH for which entries in the lexicon exist - constitutes each resulting CR. All other words with empty concept mappings, e.g. articles and aspectual markers, are ignored in the conversion. Due to lexical ambiguity, i.e. the one to many word - concept mappings, this processing step yields a set $I = \{CR_1, CR_2, \dots, CR_n\}$ of possible interpretations for each SRH.

Next, ONTOSCORE converts the domain model, i.e. an ontology, into a directed graph with concepts as nodes and relations as edges. In order to find the shortest path between two concepts, ONTOSCORE employs the *single source shortest path* algorithm of Dijkstra [3]. Thus, the minimal paths connecting a given concept c_i with every other concept in CR (excluding c_i itself) are selected, resulting in an $n \times n$ matrix of the respective paths.

To score the minimal paths connecting all concepts with each other in a given CR, [9] adopted a method proposed by [4] to score the semantic coherence of alternative sentence interpretations against graphs based on the Longman Dictionary of Contemporary English (LDOCE). As defined by [4], $R = \{r_1, r_2, \dots, r_n\}$ is the set of direct relations (both *isa* and semantic relations) that can connect two nodes (concepts); and $W = \{w_1, w_2, \dots, w_n\}$ is the set of corresponding weights, where the weight of each *isa* relation is set to 0 and that of each other relation to 1.

The algorithm selects from the set of all paths between two concepts the one with the smallest weight, i.e. the *cheapest*. The distances between all concept pairs in CR are summed up to a total score. The set of concepts with the lowest aggregate score represents the combination with the highest semantic relatedness. The ensuing distance between two concepts, e.g. $D(c_i, c_j)$ is, then, defined as the minimum score derived between c_i and c_j . So far, a number of additional normalization steps, contextual extensions and relation-specific weighted scores have been proposed and evaluated [9, 23, 17]

For this evaluation the ONTOSCORE currently employed two knowledge sources, an ontology (about 800 concepts and 200 relations) and a lexicon (ca. 3.600 words) with word to concept mappings, covering the respective domains of the system.

A Sample Gold Standard: For this annotation we employed a concept tagged data set consisting of speech recognition hypotheses that had already been identified as being the best ones out of a set of 1.375 SRHs. For these utterance representations the ontological relations that hold between the concepts that are part of the ontology's process hierarchy and the concepts that are part of the ontology's physical object hierarchy had to be identified. As this is quite a difficult task and requires substantial knowledge of both the relation inventory and its semantics, we trained two annotators for this task to see if their inter-annotator agreement was sufficient to conclude that this is a task that human annotators can reliably undertake. The resulting inter-annotator agreement on this task amounted to 79.54%, which is significantly above chance and shows that - while indeed this task is not as easy for humans as for example identifying the best SRH out of an n-best list, where the agreements amounted to 94 % - it is still doable with a satisfying degree of reliability. The gold standard was produced by means of the annotators agreeing on mutually satisfactory solutions for the cases of disagreement.

5 The Relation Tagging Experiment

For evaluating the performance of the ONTOSCORE system we employed the semantic relation error types proposed above and listed in Table 1. We defined an accurate match, if the correct non-taxonomic relation was chosen by the system for the corresponding concepts contained therein. As inaccurate we counted the semantic relation error rates proposed herein:

- deletions, i.e. missing relations in places where - according to the annotators - a relation ought to have been identified;
- insertions, i.e. postulating any relation to hold where none ought to have been; and
- substitutions, i.e. postulating a specific relation to hold where some other ought to have been.

An example of a substitution in this task is given the SRH shown in Example 1.

(1) wie komme ich von hier zum Schloss⁶

Again in this case the concept disambiguation was accurate, so that the two ambiguous entities, i.e. *kommen* and *Schloss*, were correctly mapped onto a *MotionDirected* process and a *Sight* object - the concept *Person* resulted from an unambiguous word to concept mapping from the form *ich* (*I*). A tagged example of relations for this case is given in Figure 3.

⁶ Translatable as "how can I come from here to castle".

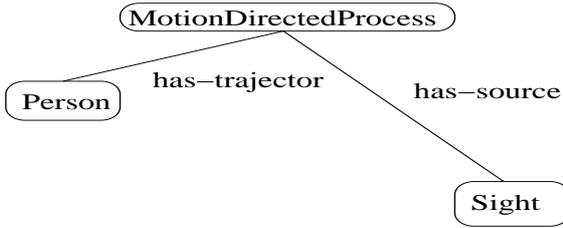


Figure 3. Substitution Type A: The gold-standard relation *has-target* was substituted with the relation *has-source*

Such substitution errors, i.e. in the case of Figure 3 (Type A) the key (*has-source*) does not fit to the gold standard answer (*has-goal*), of course, is due to missing syntactic and word information. As a special case of substitution we also counted those cases as shown in Figure 4 (Type B).

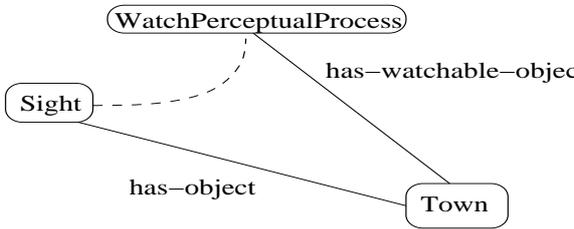


Figure 4. Substitution Type B: The gold-standard relation *has-watchable-object* was linked indirectly via the concept *town* with the relation *has-object*

Our gold standard showed cases as inaccurate where a *relation chain* was selected by the algorithm, while in principle such chains, e.g. *metonymic chains* are possible and in some domains not infrequent, in the relatively simple and short dialogues that constitute our data. Therefore cases such as the connection between *WatchPerceptualProcess* and *Sight* shown in Example 2 and Figure 4 was counted as a substitution, because a simpler one was indicated in the gold standard.

(2) ich will das Schloss anschauen⁷

As a deletion such cases were counted wherein the gold standard contained a specific relation such as *WatchPerceptualProcess has-watchable-object Sight*, was not tagged at all by the system, as shown in Figure 5.

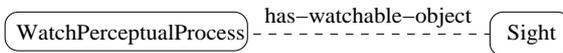


Figure 5. Deletion: The gold standard relation *has-watchable-object* was not tagged by the system

As an insertion we counted the opposite case, i.e. where any relations, e.g. between *Agent* and *Sight* in Example (2) were tagged by the system, as shown in Figure 6.

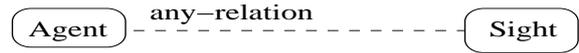


Figure 6. Insertion: Any relation was tagged where gold standard indicated none

6 Results

As compared to the human gold standard we obtained the accuracies, substitutions, deletions and insertions as shown in Table 3.

overall accuracy	76.31%
substitutions	15.32%
deletions	7.11%
insertions	1.26%
human annotation	79.54%

Table 3. Results Overview

Population Fall-out: While not every error can be mapped directly for populating the ontological model, for example in cases of Substitution Type A the appropriate conceptual instrument has been part of the model, e.g. the relation *has-target* typed on a superclass of *Sight* (i.e. *Sight* also features the role *has-target* next to *has-source*⁸ via inheritance, but the choosing of the erroneous relation was caused by the application/algorithm.

All substitutions of type B as well as Deletions can be used for populating the ontological model with new or better instruments. Note that we use *instruments* to denote that not only concepts are/can be added but also semantic relations that are missing or modeled inefficiently. Alternative applications for such a task-based evaluation are constituted by sense/concept tagging and discovering set-ups (e.g. see the corresponding concept tagging experiments [17] or learning experiments [21]). While these can lead to concept population or concept generalizations [32], the task of finding semantic/ontological relations has shown to be quite elusive. The population fall-out of running the test-suite in this case derives its content from the gold standard that was merged from the doubly annotated data. This manual input is still required and without it the *relation population* problem is still a relevant challenge for comprehensive ontology learning approaches. The result of this work makes it feasible to measure our progress along the path to better performances and better ontological models.

7 Analysis and Concluding Remarks

In this experiment concerning the tagging of the ontological relations a baseline computation as in [25] has (so far) been thwarted by the difficulties in calculating the set of markable-specific tagsets out of the ontological model and attribute-specific values found in the data. Therefore, these results do not provide a comparable measure for evaluating the performance of the ONTOSCORE system on this non-trivial task. They do, however, in our mind, clearly indicate several shortcomings in the ontology used:

⁸ As well as many more including those corresponding to the SOURCE-PATH-GOAL image schema [16].

⁷ Translatable as “I want to see the castle”.

- The 7.11% deletions indicate clear cases where a pertinent (at least for this task) relation was not modeled in the ontology,
- About 50% of the substitution errors showed inefficiencies in the model (the rest where actually a result of the algorithm's shortcomings), and
- the - however - small percentage of insertions can be regarded as superfluously modeled relations.

It now would be relatively easy to go back to the model and undertake the corresponding changes and run the evaluation again (repeating this process until the accuracy is 100 %). While this might be a sensible undertaking from an engineering perspective, i.e. as a bootstrapping approach accompanying a system's deployment, we see little scientific value in this path. A more challenging and interesting outcome of such task-based approach to ontology evaluation lies in questions concerning way to make such a scheme more general and scalable.

That is, we hope to have shown that, while it is hard for qualitative evaluations to look at a model and say whether it is *good* or *scruffy*, it is quite feasible for humans to construct ontologically annotated gold standards given a specific set of human utterances or textual corpora. Additionally, given one (or more) domain specific ontologies and given a corresponding gold standard we can perform a task-based evaluation series that yields quantifiable results about the respective qualities of each of the individual levels of the ontological model.

Acknowledgments

This work is part of the SmartWeb project - partially funded by the German Federal Ministry of Research and Technology (BMBF) under Grant number 01IMD01E and by the Klaus Tschira Foundation.

REFERENCES

- [1] Collin F. Baker, Charles J. Fillmore, and John B. Lowe, 'The Berkeley FrameNet Project', in *Proceedings of COLING-ACL*, Montreal, Canada, (1998).
- [2] Christopher Brewster, Hartith Alani, Srinandan Dasmahapatra, and Yorick Wilks, 'Data driven ontology evaluation', in *Proceedings of LREC 2004*, Lisbon, Portugal, (2004).
- [3] Thomas H. Cormen, Charles E. Leiserson, and Ronald R. Rivest, *Introduction to Algorithms*, MIT press, Cambridge, MA, 1990.
- [4] George Demetriou and Eric Atwell, 'A semantic network for large vocabulary speech recognition', in *Proceedings of AISB workshop on Computational Linguistics for Speech and Handwriting Recognition*, eds., Lindsay Evett and Tony Rose, University of Leeds, (1994).
- [5] Daniel Gildea and Daniel Jurafsky, 'Automatic labeling of semantic roles', *Computational Linguistics*, **28**(3), 245–288, (2002).
- [6] Asuncion Gomez-Perez, 'Evaluation of taxonomic knowledge in ontologies and knowledge bases', in *Proceedings of the 12th Banff Knowledge Acquisition for Knowledge-based Systems Workshop, Banff, Canada*, (1999).
- [7] Nicola Guarino, 'Some ontological principles for designing upper level lexical resources', in *Proceedings of the 1st International Conference on Language Resources and Evaluation*, Granada, Spain, 28-30 May, 1998, (1998).
- [8] Nicola Guarino and Christopher Welty, 'Evaluating ontological decisions with ontoclean'.
- [9] Iryna Gurevych, Rainer Malaka, Robert Porzel, and Hans-Peter Zorn, 'Semantic coherence scoring using an ontology', in *Proceedings of the HLT/NAACL 2003*, Edmonton, CN, (2003).
- [10] Iryna Gurevych, Robert Porzel, and Stefan Merten, 'Automatic creation of interface specifications from ontologies', in *Proceedings of the HLT/NAACL SEALTS Workshop*, Edmonton, Canada, (2003).
- [11] Iryna Gurevych, Robert Porzel, and Stefan Merten, 'Less is more: Using a single knowledge representation in dialogue systems', in *Proceedings of the HLT/NAACL Text Meaning Workshop*, Edmonton, Canada, (2003).
- [12] Ryuichiro Higashinaka, Noboru Miyazaki, Mikio Nakano, and Kiyooki Aikawa, 'A method for evaluating incremental utterance understanding in spoken dialogue systems', in *Proceedings of the International Conference on Speech and Language Processing 2002*, pp. 829–833, Denver, USA, (2002).
- [13] Ryuichiro Higashinaka, Noboru Miyazaki, Mikio Nakano, and Kiyooki Aikawa, 'Evaluating discourse understanding in spoken dialogue systems', in *Proceedings of Eurospeech*, pp. 1941–1944, Geneva, Switzerland, (2003).
- [14] Eduard Hovy, *Semantics of Relations*, chapter Comparing sets of semantic relations, Kluwer Academic Publishers, Dordrecht, NL, 2001.
- [15] Daniel Jurafsky and James Martin, *Natural Language Processing*, Springer, 1991.
- [16] George Lakoff, *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*, University of Chicago Press, 1987.
- [17] Berenike Loos and Robert Porzel, 'The resolution of lexical ambiguities in spoken dialogue systems', in *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, Boston, USA, 30-31 April 2004, (2004).
- [18] Alexander Maedche and Steffen Staab, 'Measuring the similarity between ontologies', in *Proceedings of the 13th Conference on Knowledge Engineering and Knowledge Management*, Springer, LNAI 2473, Berlin., (2002).
- [19] Elaine Marsh and Dennis Perzanowski, 'MUC-7 evaluation of IE technology: Overview of results', in *Proceedings of the 7th Message Understanding Conference*. Morgan Kaufman Publishers, (1999).
- [20] Yorick Milks, 'Ontotherapy: or how to stop worrying about what there is.', in *Proceedings of the Workshop on Ontologies and Lexical Knowledge Bases*, Las Palmas, Canary Islands, (2002).
- [21] Patrick Pantel and Dekang Lin, 'Automatically discovering word senses', in *HLT-NAACL 2003: Demo Session*, eds., Bob Frederking and Bob Younger, Edmonton, Alberta, Canada, (2003). ACL.
- [22] Fernando Pereira, Tishby Naftali, and Lee Lillian, 'Distributional clustering of english words', in *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio, 22–26 June 1993, (1993).
- [23] Robert Porzel, Iryna Gurevych, and Christof Müller, 'Ontology-based contextual coherence scoring', in *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*, Sapporo, Japan, July 2003, (2003).
- [24] Robert Porzel, Berenike Loos, and Vanessa Micelli, 'Making relative sense: From word-graphs to semantic frames', in *Proceedings of the HLT/NAACL Workshop on Scalable Natural Language Understanding*, Boston, USA, (2004).
- [25] Robert Porzel and Rainer Malaka, 'Towards measuring scalability in natural language understanding tasks', in *Proceedings of the HLT/NAACL Workshop on Scalable Natural Language Understanding*, Boston, USA, (2004).
- [26] Robert Porzel, Norbert Pflieger, Stefan Merten, Markus Löckelt, Ralf Engel, Iryna Gurevych, and Jan Alexandersson, 'More on less: Further applications of ontologies in multi-modal dialogue systems', in *Proceedings of the 3rd IJCAI 2003 Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, Acapulco, Mexico, (2003).
- [27] Stuart J. Russell and Peter Norvig, *Artificial Intelligence. A Modern Approach*, Prentice Hall, Englewood Cliffs, N.J., 1995.
- [28] Mark Stevenson, *Combining disambiguation techniques to enrich an ontology.*, In Proceedings of the 15th ECAI workshop on Machine Learning and Natural Language Processing for Ontology Engineering, 2002.
- [29] Mark Stevenson, *Word Sense Disambiguation: The Case for Combining Knowledge Sources*, CSLI, 2003.
- [30] Wolfgang Wahlster, 'SmartKom: Symmetric multimodality in an adaptive reusable dialog shell', in *Proceedings of the Human Computer Interaction Status Conference*, Berlin, Germany, (2003).
- [31] Wolfgang Wahlster, Norbert Reithinger, and Anselm Blocher, 'Smartkom: Multimodal communication with a life-like character', in *Proceedings of the 7th European Conference on Speech Communication and Technology*, (2001).
- [32] Dominic Widdows, 'Unsupervised methods for developing taxonomies by combining syntactic and statistical information', in *HLT-NAACL 2003: Main Proceedings*, pp. 276–283, Edmonton, Canada, (May 27 - June 1 2003). ACL.