Making Relative Sense: From Word-graphs to Semantic Frames

Robert Porzel **Berenike Loos** **Vanessa Micelli**

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

Abstract

Scaling up from controlled single domain spoken dialogue systems towards conversational, multi-domain and multimodal dialogue systems poses new challenges for the reliable processing of less restricted user utterances. In this paper we explore the feasibility to employ a general purpose ontology for various tasks involved in processing the user's utterances.

1 Introduction

We differentiate between controlled single-domain and more conversational multi-domain spoken dialogue systems (Allen et al., 2001). The transition from the former to the later can be regarded as a scaling process, since virtually every processing technique applicable for restricted single domain user utterances has to be adopted to new challenges, i.e., varying context-dependencies (Porzel et al., 2004) increasing levels of ambiguity (Gurevych et al., 2003a; Loos and Porzel, 2004) and less predictable input (Loeckelt et al., 2002). Additionally, for conversational multi-domain spoken dialogue systems tasks have to be tackled that were by and large unnecessary in restricted single-domain systems. In this exploration, we will focus on a subset of these tasks, namely:

- *hypotheses verification* (HV) - i.e. finding the best hypothesis out of a set of possible speech recognition hypotheses (SRH);
- *sense disambiguation* (SD) - i.e. determining the best mapping of the lexically ambiguous linguistic forms contained therein to their sense-specific *semantic representations*;
- *relation tagging* (RT) - i.e. determining adequate semantic relations between the relevant sense-tagged entities.

Many of these tasks have been addressed in other fields, for example, hypothesis verification in the field of machine translation (Tran et al., 1996), sense disambiguation in speech synthesis (Yarowsky, 1995), and relation tagging in information retrieval (Marsh and Perzanowski, 1999). These challenges also apply for spoken dialogue systems and arise when they are scaled up towards multi-domain and more conversational settings.

In this paper we will address the utility of using ontologically modeled knowledge to assist in solving these tasks in spoken dialogue systems. Following an overview of the state of the art in Section 2 and the ontology-based coherence scoring system in Section 3, we describe its employment in the task of hypotheses verification in Section 4. In Section 5 we describe the system's employment for the task of sense disambiguation and in Section 6 we present first results of a study examining the performance of the system for the task of relation tagging. An analysis of the evaluation results and concluding remarks are given in Section 7.

2 Related Work

2.1 Hypotheses Verification

While a simple one-best hypothesis interface between automatic speech recognition (ASR) and natural language understanding (NLU) suffices for restricted dialogue systems, more complex systems either operate on n-best lists as ASR output or convert ASR word graphs (Oerder and Ney, 1993) into n-best lists. Usually, this task is performed by combining the respective acoustic and language model scores calculated by the speech recognition system as described by Schwartz and Chow (1990).

Facing multiple representations of a single utterance consequently poses the question, which one of the different hypotheses corresponds most likely to the user's utterance. Several ways of solving this problem have been proposed and implemented in various systems. As mentioned above, the scores provided by the ASR system itself are used most frequently. Still, in recent works also scores provided by the NLU system have been em-

ployed, e.g. parsing scores (Engel, 2002) or discourse based scores (Pfleger et al., 2002). However, these methods are prone to assign very high scores to SRHs which are semantically incoherent and low scores to semantically coherent ones, if faced with imperfect and unpredicted input (Porzel et al., 2003a).

2.2 Sense Disambiguation

Employing the task categorization scheme proposed by Stevenson (2003), the task of creating adequate semantic representations of the individual entities occurring in the SRHs can be regarded as a form of *semantic disambiguation*. Since, in our case, a fixed inventory of *senses* is given by the lexicon and only the ambiguous lexical forms have to be disambiguated, our task falls into the corresponding subcategory of *sense disambiguation*. Following Ide and Veronis (1998) we distinguish between data- and knowledge-driven word sense disambiguation. Given the basic distinction between written text and spoken utterances, the only sense disambiguation results performed on speech data stemming from human interactions with dialogue systems have been reported by Loos and Porzel (2004), who compared both data- and knowledge-driven sense disambiguation on the same set of actual speech data.

Historically, after work on WSD had overcome so-called *early doubts* (Ide and Veronis, 1998) in the 1960's, it was applied to various NLP tasks, such as machine translation, information retrieval, content and grammatical analysis and text processing. Yarowsky (1995) used both supervised and unsupervised WSD for correct phonetizitation of words in speech synthesis. However, there is no recorded work on processing speech recognition hypotheses resulting from speech utterances as it is done in our research. In general, following Ide and Veronis (1998) the various WSD approaches of the past can be divided into two types, i.e., data- and knowledge-based approaches.

Data-based Methods Data-based approaches extract their information directly from texts and are divided into supervised and unsupervised methods (Yarowsky, 1995; Stevenson, 2003).

Supervised methods work with a given (and therefore limited) set of potential classes in the learning process. For example, Yarowsky (1992) used a thesaurus to generate 1042 statistical models of the most general categories. Weiss (1973) already showed that disambiguation rules can successfully be learned from hand-tagged corpora. However limited by the small size of his training and test corpus, an accuracy of 90% was achieved. Even better results on a larger corpus were obtained by Kelly and Stone 1975 who included collocational, syntactic and part of speech information to yield an accuracy of 93% on a larger corpus. As always, supervised methods require a manually annotated learning corpus.

Unsupervised methods do not determine the set of classes before the learning process, but through analysis of the given data by identifying clusters of similar cases. One example is the algorithm for clustering by committee described by Pantel and Lin (2003), which automatically discovers word senses from text. Generally, unsupervised methods require large amounts of data. In the case of spoken dialogue and speech recognition output sufficient amounts of data will hopefully become available once multi-domain spoken dialogue systems are deployed in real world applications.

Knowledge-based Methods Knowledge-based approaches work with lexica and/or ontologies. The kind of knowledge varies widely and machine-readable lexica are employed. The knowledge-based approach employed herein (Gurevych et al., 2003a) operates on an ontology partially derived from FrameNet data (Baker et al., 1998) and is described by Gurevych et al. (2003b).

In a comparable approach Sussna (1993) worked with the lexical reference system WordNet and used a similar metric for the calculation of semantic distance of a number of input lexemes. Depending on the type of semantic relation (hyperonymy, synonymy etc.) different weights are given and his metric takes account of the number of arcs of the same type leaving a node and the depth of a given edge in the overall tree. The disambiguation results on textual data reported by Sussna (1993) turned out to be significantly better than chance. In contrast to many other work on WSD with WordNet he took into account not only the *isa* hierarchy, but other relational links as well. The method is, therefore, similar to the one used in this evaluation, with the difference that this one uses a semantic-web conform ontology instead of WordNet and it is applied to speech recognition hypotheses. The fact, that our WSD work is done on SRHs makes it difficult to compare the results with methods evaluated on textual data such as in the SENSEVAL studies (Edmonds, 2002).

2.3 Labeling Semantic Roles and Relations

The task of representing the *semantic relations* that hold between the sense tagged entities can be thought of as an extension of the work presented by Gildea and Jurafsky (2002), where the tagset is defined by entities corresponding to FrameNet *frame elements* (Baker et al., 1998). 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 (Marsh and Perzanowski,

1999). In this case predefined templates are given (e.g. `is-bought-by(COMPANY_A,COMPANY_B)` which have to instantiated correctly, i.e. in a phrase such as *"Stocks sky-rocketed after Big Blue acquired Softsoft ..."* the specific roles, i.e. *Big Blue* as `COMPANY_B` and *Softsoft* as `COMPANY_A` have to be put in their adequate places within the overall template.

Now that speech data from the more conversational multi-domain dialogue systems have become available, we present the corresponding annotation experiments and evaluation results of a knowledge-driven hypothesis verification, sense disambiguation and relation tagging system, whose knowledge store and algorithm are presented below.

3 Ontology-based Scoring and Tagging

The Ontology Used: The ontology used in the experiments described herein was initially designed as a general purpose component for knowledge-based NLP. It includes a top-level ontology developed following the procedure outlined by Russell and Norvig (1995) and originally covered the tourism domain encoding knowledge about sights, historical persons and buildings. Then, the existing ontology was adopted in the SMARTKOM project (Wahlster et al., 2001) and modified to cover a number of new domains, e.g., new media and program guides, pedestrian and car navigation and more (Gurevych et al., 2003b). 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 a kind of event that is continuous and homogeneous in nature, follows the frame semantic analysis used in the FRAMENET project (Baker et al., 1998).

The role 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: *agent*, *theme*, *experiencer*, *instrument* (or *means*), *location*, *source*, *target*, *path*. These general roles applied to concrete processes may also have subroles: thus an agent in a process of buying (*TransactionProcess*) is a *buyer*, the one in the process of cognition is a *cognizer*. This way, roles can also build hierarchical trees. The property *theme* in the process of information search is a required *piece-of-information*, in *PresentationProcess* it is a *presentable-object*, i.e., the entity that is to be presented.

The OntoScore System: The ONTOSCORE software runs as a module in the SMARTKOM multi-modal and multi-domain spoken dialogue system (Wahlster, 2003). 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 SRHs produced by the language interpretation module out of the ASR word graphs. It computes a numerical ranking of alternative SRHs and thus provides an important aid to the spoken language understanding component. More precisely, the 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.

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 where 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 $I = \{CR_1, CR_2, \ldots, CR_n\}$ 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 (Cormen et al., 1990). 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, Gurevych et al. (2003a) adopted a method proposed by Demetriou and Atwell (1994) to score the semantic coherence of alternative sentence interpretations against graphs based on the Longman Dictionary of Contemporary English (LDOCE). As defined by Demetriou and Atwell (1994), $R = \{r_1, r_2, \ldots, r_n\}$ is the set of direct relations (both *isa* and semantic relations) that can connect two nodes (concepts); and $W = \{w_1, w_2, \ldots, w_n\}$ 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 (Gurevych et al., 2003a; Porzel et al., 2003a; Loos and Porzel, 2004)

The ONTOSCORE module currently employs 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 Motivating Example: Given the utterance shown in its transcribed form in example (1), we get as input the set of recognition hypotheses shown in examples (1a) - (1e) extracted from the word graph produced by the ASR system.

1 *wie komme ich in Heidelberg weiter.*
how can I in Heidelberg continue.

1a *Rennen Lied Comedy Show Heidelberg*
Race song comedy show Heidelberg
weiter.
continue.

1b *denn wie Comedy Heidelberg weiter.*
then how comedy Heidelberg continue.

1c *denn wie kommen Show weiter.*
then how come show continue.

1d *denn wie Comedy weiter.*
then how comedy continue.

1e *denn wie komme ich in Heidelberg*
then how can I in Heidelberg
weiter.
continue.

For our evaluations we defined three tasks and their domains as follows:

- The task of hypotheses verification to be solved successfully if the SRHs 1a to 1e are ranked in such a way that hypothesis 1e achieves the best score.
- The task of sense disambiguation to be solved successfully if all ambiguous lexical items, such as the verb *kommen* in 1e, are tagged with their contextually adequate senses given in our case by the ontological class inventory, such as `MotionDirectedTransliterated` rather than `WatchPerceptualProcess`.
- The task of semantic role labeling to be solved successfully if all concepts are labeled with their appropriate frame semantic roles, such as shown below.

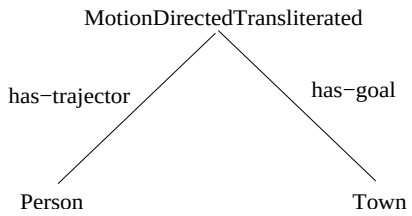

Figure 1: Tagging Relations

It is important to point out that there are at least two essential differences between spontaneous speech semantic tagging and the textual correlates, i.e.,

- a smaller size of processable context as well as
- imperfections, hesitations, disfluencies and speech recognition errors.

For our evaluations we employ the ONTOSCORE system to select the best hypotheses, best sense and best relation and compare its *answers* to *keys* contained in corresponding gold-standards produced by specific annotation experiments.

4 Hypotheses Disambiguation

4.1 Data and Annotation

The corresponding data collection is described in detail by Gurevych and Porzel (2004). In the first experiment 552 utterances were annotated within the discourse context, i.e. the SRHs were presented in their original dialogue order. In this experiment, the annotators saw the SRHs together with the transcribed user utterances. The task of the annotators was to determine the best SRH from the n-best list of SRHs corresponding to a single user utterance. The decision had to be made on the basis of several criteria. The most important criteria was how well the SRH captures the intentional content of the user's utterance. If none of the SRHs captured the user's intent adequately, the decision had to be made by looking at the actual word error rate. In this experiment the inter-annotator agreement was 90.69%, i.e. 1,247 markables out of 1,375. In a second experiment annotators had to label each SRHs as being semantically coherent or incoherent, reaching an agreement of 79.91% (1,096 out of 1,375). Each corpus was then transformed into an evaluation *gold standard* by means of the annotators agreeing on a single solution for the cases of disagreement.

4.2 Evaluation Results

The evaluation of ONTOSCORE was carried out on a set of 95 dialogues. The resulting dataset contained 552 utterances resulting in 1,375 SRHs, corresponding to an average of 2.49 SRHs per user utterance. The corpus had been annotated by human subjects according to specific annotation schemata which are described above.

Identifying the Best SRH The task of ONTOSCORE in our multimodal dialogue system is to determine the best SRH from the n-best list of SRHs corresponding to a given user utterance. The baseline for this evaluation was computed by adding the individual ratios of utterance/SRHs - corresponding to the likelihood of guessing the best one in each individual case - and dividing it by the number of utterances - yielding the overall likelihood of guessing the best one as 63.91%. The accuracy of ONTOSCORE on this task amounts to 86.76%. This means that in 86.76% of all cases the best SRH defined by the human *gold standard* is among the best scored by the ONTOSCORE module.

Classifying the SRHs as Semantically Coherent versus Incoherent For this evaluation we used the same corpus, where each SRH was labeled as being either semantically coherent *versus* incoherent with respect to the previous discourse context. We defined a baseline based on the majority class, i.e. coherent, in the corpus, 63.05%. In order to obtain a binary classification into semantically coherent and incoherent SRHs, a cutoff threshold must be set. Employing a cutoff threshold of 0.44, we find that the contextually enhanced ONTOSCORE system correctly classifies 70.98% of SRHs in the corpus.

From these results we can conclude that the task of an absolute classification of coherent *versus* incoherent is substantially more difficult than that of determining the best SRH, both for human annotators and for ONTOSCORE. Both human and the system's reliability is lower in the coherent *versus* incoherent classification task, which allows to classify zero, one or multiple SRHs from one utterance as coherent or incoherent. In both tasks, however, ONTOSCORE's performance mirrors and approaches human performance.

5 Sense Disambiguation

5.1 Data and Annotation

The second data set was produced by means of Wizard-of-Oz experiments (Francony et al., 1992). In this type of setting a full-blown multimodal dialogue system is simulated by a team of human hidden operators. A test person communicates with the supposed system and the dialogues are recorded and filmed digitally. Here over 224 subjects produced 448 dialogues (Schiel et al., 2002), employing the same domains and tasks as in the first data collection. In this annotation task annotators were given the recognition hypotheses together with a corresponding list of ambiguous lexemes automatically retrieved form the system's lexicon and their possibles senses, from which they had to pick one or select not-decidable for cases where not coherent meaning was detectable.

Firstly, we examined if humans are able to annotate the data reliably. Again, this was the case, as shown by the resulting inter annotator agreement of 78.89%. Secondly, a gold-standard is needed to evaluate the system's performance. For that purpose, the annotators reached an agreement on annotated items of the test data which had differed in the first place. The ensuing gold-standard altogether was annotated with 2225 markables of ambiguous tokens, stemming from 70 ambiguous words occurring in the test corpus.

5.2 Evaluation Results

For calculating the majority class baselines, all markables in the gold-standards were counted. Corresponding to the frequency of each concept of each ambiguous lexeme the percentage of correctly chosen concepts by means of selecting the most frequent meaning without the help of a system was calculated by means of the formula given by Porzel and Malaka (2004). This resulted in a baseline of 52.48% for the test data set.

For this evaluation, ONTOSCORE transformed the SRH from our corpus into concept representations as described in Section 2. To perform the WSD task, ONTOSCORE calculates a coherence score for each of these concept sets. The concepts in the highest ranked set are considered to be the ones representing the correct word meaning in this context. In this experiment we used OntoScore in two variations: Using the first variation, the relations between two concepts are weighted 0 for taxonomic relations and 1 for all others. The second mode allows each non taxonomic relation being assigned an individual weight depending on its position in the relation hierarchy. That means the relations have been weighted according to their level of generalization. More specific relations should indicate a higher degree of semantic coherence and are therefore weighted cheaper, which means that they - more likely - assign the correct meaning. Compared to the gold-standard, the original method of Gurevych et al. (2003a) reached a precision of 63.76% as compared to 64.75% for the new method described herein.

6 Relation Tagging

6.1 Data and Annotation

For this annotation we employed a subset of the second data set, i.e. we looked only at the hypotheses identified as being the best one (see above). For these utterance

representations the semantic relations that hold between the predicate (in our case concepts that are part of the ontology's Process hierarchy) and the entities (in our case concepts that are part of the ontology's Physical Object hierarchy) had to be identified. The inter-annotator agreement on this task amounted to 79.54%.

6.2 Evaluation Results

For evaluating the performance of the ONTOSCORE system we defined an accurate match, if the correct semantic relation (role) was chosen by the system for the corresponding concepts contained therein[1]. As inaccurate we counted in analogy to the word error rates in speech recognition:

- deletions, i.e. missing relations in places were one 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 2.

2 *wie komme ich von hier zum Schloss.*
how come I from here to castle.

In this case the sense disambiguation was accurate, so that the two ambiguous entities, i.e. *kommen* and *Schloss*, were correctly mapped onto a `MotionDirectedTransliterated` (MDT) process and `Sight` object - the concept `Person` resulted from an unambiguous word-to-concept mapping from the form *I*. The error in this case was the substitution of [has-goal] with the relation [has-source], as shown below:

```
[MDT] [has-agent] [Agent]
[MDT] [has-source] [Sight]
```

As a special case of substitution we also counted those 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 still relatively simple and short dialogues that constitute our data[2]. Therefore cases such as the connection between `WatchPerceptualProcess` (WPP) and `Sight` shown in Example 3 were counted as substitutions, because simpler ones should have been found or modeled[3].

[1]Regardless of whether they were the correct senses or not as defined in the sense disambiguation task.

[2]This, in turn, also shed a light on the paucity of the capabilities that current state-of-the-art systems exhibit.

[3]We are quite aware that such an evaluation is as much a test of the knowledge store as well as of the processing algorithms. We will discuss this in Section 7.

3 *ich will das Schloss anschauen*
I want the castle see

```
[WPP] [has-watchable_object]  [Map]
[has-object] [Sight]
```

As a deletion such cases were counted where the annotators (more specifically the ensuing gold standard) contained a specific relation such as [WPP] [has-watchable-object] [Sight], was not tagged at all 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 compared to the human gold standard we obtained an accuracy of 76.31% and an inaccuracy of substitutions of 15.32%, deletions of 7.11% and insertions of 1.26%.

7 Analysis and Concluding Remarks

In the cases of hypothesis and semantic disambiguation the knowledge-driven system scores significantly above the baseline (22,85% and 11.28% respectively) as shown in Table 1.

Task	Baseline	Agreement	Accuracy
HV	63.91%	90.69%	86.76%
SD	52.48%	78.89%	63.76%
RT	n.a.	79.54%	76.31%

Table 1: Results Overview

In the case of tagging the semantic relations a baseline computation 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. However, the performance may even be seen especially encouraging in comparison to the case of sense disambiguation. However, comparisons might well be misleading, as the evaluation criteria defined different views on the data. Most notably this is the case in examining the concept sets of the best SRHs as given potentially existing disambiguated representations. While this can certainly be the case, i.e. utterances for which these concept sets constitute the correct set can easily be imagined, the underlying potential utterances, however, did not occur in the data set examined in the case of the sense disambiguation evaluations.

A more general consideration stems from the fact that both the knowledge store used and coherence scoring method have been shown to perform quite robustly for a variety of tasks. Some of these tasks - which are not mentioned herein - are executed by different processing components that employ the same underlying knowledge model but apply different operations such as *overlay* and have been reported elsewhere (Alexandersson

and Becker, 2001; Gurevych et al., 2003b; Porzel et al., 2003b). In this light such evaluations could be used to single out an evaluation method for finding gaps and inconsistencies in the ontological model. While such a bootstrapping approach to ontology building could assist in avoiding scaling-related decreases in the performance of knowledge-based approaches, our concern in this evaluation also was to be able to set up additional examinations of the specific nature of the inaccuracies, by looking at the interdependencies between relation tagging and sense disambiguation.

There remain several specific questions to be answered on a more methodological level as well. These concern ways of measuring the task-specific perplexities or comparable baseline metrics to evaluate the specific contribution of the system described herein (or others) for the task of *making sense* of ASR output. Additionally, methods need to be found in order to arrive at aggregate measures for computing the difficulty of the combined task of sense disambiguation and relation tagging and for evaluating the corresponding system performance. In future we will seek to remedy this situation in order to arrive at two general measurements:

- a way of assessing the increases in the natural language understanding difficulty that result from scaling NLU systems towards more conversational and multi-domain settings;
- a way of evaluating the performance of how individual processing components can cope with the scaling effects on the aggregate challenge to find suitable representations of spontaneous natural language utterances.

In the light of scalability it is also important to point out that scaling such knowledge-based approaches comes with the associated cost in knowledge engineering, which is still by and large a manual process. Therefore, we see approaches that attempt to remove (or at least widen) the knowledge acquisition bottleneck to constitute valuable complements to our approach, which might be especially relevant for designing a bootstrapping approach that involves automatic learning and evaluation cycles to create scalable knowledge sources and approaches to natural language understanding.

Acknowledgments

This work has been partially funded by the German Federal Ministry of Research and Technology (BMBF) as part of the SmartKom project under Grant 01 IL 905C/0 and by the Klaus Tschira Foundation.

References

Jan Alexandersson and Tilman Becker. 2001. Overlay as the basic operation for discourse processing. In *Proceedings of the IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*. Springer, Berlin.

James F. Allen, Donna K. Byron, Myroslava Dzikovska, George Ferguson, Lucian Galescu, and Amanda Stent. 2001. Towards conversational human-computer interaction. *AI Magazine*.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of COLING-ACL*, Montreal, Canada.

Thomas H. Cormen, Charles E. Leiserson, and Ronald R. Rivest. 1990. *Introduction to Algorithms*. MIT press, Cambridge, MA.

George Demetriou and Eric Atwell. 1994. A semantic network for large vocabulary speech recognition. In Lindsay Evett and Tony Rose, editors, *Proceedings of AISB workshop on Computational Linguistics for Speech and Handwriting Recognition*, University of Leeds.

Philip Edmonds. 2002. SENSEVAL: The evaluation of word sense disambiguation systems. *ELRA Newsletter*, 7/3.

Ralf Engel. 2002. SPIN: Language understanding for spoken dialogue systems using a production system approach. In *Proceedings of the International Conference on Speech and Language Processing 2002*, Denver, USA.

J.-M. Francony, E. Kuijpers, and Y. Polity. 1992. Towards a methodology for wizard of oz experiments. In *Third Conference on Applied Natural Language Processing*, Trento, Italy, March.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Iryna Gurevych and Robert Porzel. 2004. Empirical studies for intuitive interaction. In Wolfgang Wahlster, editor, *SmartKom: Foundations in Multimodal Interaction*. Springer, Berlin.

Iryna Gurevych, Rainer Malaka, Robert Porzel, and Hans-Peter Zorn. 2003a. Semantic coherence scoring using an ontology. In *Proceedings of the HLT/NAACL 2003*, Edmonton, CN.

Iryna Gurevych, Robert Porzel, and Stefan Merten. 2003b. Less is more: Using a single knowledge representation in dialogue systems. In *Proceedings of the HLT/NAACL Text Meaning Workshop*, Edmonton, Canada.

Nancy Ide and J. Veronis. 1998. Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics*, 24/1.

Markus Loeckelt, Tilman Becker, Norbert Pfleger, and Jan Alexandersson. 2002. Making sense of partial. In *Proceedings of the 6th workshop on the semantics and pragmatics of dialogue*, Edinburgh, Scotland.

Berenike Loos and Robert Porzel. 2004. The resolution of lexical ambiguities in spoken dialgue systems. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, Boston, USA, 30-31 April 2004. To appaer.

Elaine Marsh and Dennis Perzanowski. 1999. MUC-7 evaluation of IE technology: Overview of results. In *Proceedings of the 7th Message Understanding Conference*. Morgan Kaufman Publishers.

Martin Oerder and Hermann Ney. 1993. Word graphs: An efficient interface between continuous-speech recognition and language understanding. In *ICASSP Volume 2*.

Patrick Pantel and Dekang Lin. 2003. Automatically discovering word senses. In Bob Frederking and Bob Younger, editors, *HLT-NAACL 2003: Demo Session*, Edmonton, Alberta, Canada. Association for Computational Linguistics.

Norbert Pfleger, Jan Alexandersson, and Tilman Becker. 2002. Scoring functions for overlay and their application in discourse processing. In *Proceedings of KONVENS 2002*, Saarbrücken, Germany.

Robert Porzel and Rainer Malaka. 2004. Towards measuring scalability for natural language understanding tasks. In *Proceedings of the 2th International Workshop on Scalable Natural Language Understanding*, Boston, USA, 6 May 2004. To appaer.

Robert Porzel, Iryna Gurevych, and Christof Müller. 2003a. Ontology-based contextual coherence scoring. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*, Saporo, Japan, July 2003.

Robert Porzel, Norbert Pfleger, Stefan Merten, Markus Löckelt, Ralf Engel, Iryna Gurevych, and Jan Alexandersson. 2003b. 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.

Robert Porzel, Iryna Gurevych, and Rainer Malaka. 2004. In context: Integration domain- and situation-specific knowledge. In Wolfgang Wahlster, editor, *SmartKom: Foundations in Multimodal Interaction*. Springer, Berlin.

Stuart J. Russell and Peter Norvig. 1995. *Artificial Intelligence. A Modern Approach*. Prentice Hall, Englewood Cliffs, N.J.

Florian Schiel, Silke Steininger, and Ulrich Türk. 2002. The smartkom multimodal corpus at bas. In *Proceedings of the 3rd LREC*, Las Palmas Spain.

R. Schwartz and Y. Chow. 1990. The n-best algorithm: an efficient and exact procedure for finding the n most likely sentence hypotheses. In *Proceedings of ICASSP'90, Albuquerque, USA*.

Mark Stevenson. 2003. *Word Sense Disambiguation: The Case for Combining Knowldge Sources*. CSLI.

Michael Sussna. 1993. Word sense disambiguation for free text indexing using a massive semantic network. In *Proceedings of the Second International Conference on Information and Knowledge Management*.

B-H. Tran, F. Seide, V. Steinbiss, R. Schwartz, and Y. Chow. 1996. A word graph based n-best search in continuous speech recognition. In *Proceedings of ICSLP'96*.

Wolfgang Wahlster, Norbert Reithinger, and Anselm Blocher. 2001. Smartkom: Multimodal communication with a life-like character. In *Proceedings of the 7th European Conference on Speech Communication and Technology*.

Wolfgang Wahlster. 2002. SmartKom: Fusion and fission of speech, gerstures and facial expressions. In *Proceedings of the Firsat International Workshop on Man-Machine Symbiotic Systems*, Kyoto, Japan.

Wolfgang Wahlster. 2003. SmartKom: Symmetric multimodality in an adaptive an reusable dialog shell. In *Proceedings of the Human Computer Interaction Status Conference*, Berlin, Germany.

Stephen Weiss. 1973. Learning to disambiguate. *Information Storage and Retrieval*, 9.

David Yarowsky. 1992. Word-sense disambiguation using statistical models of roget's categories trained on large corpora. In *Proceedings of the 15th International Conference on Computational Linguistics*, Nantes, France, 23-28 August 1992, volume 1.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivalling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, Cambridge, Mass., 26–30 June 1995.