Resolution of Lexical Ambiguities in Spoken Dialogue Systems

Berenike Loos   **Robert Porzel**

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

Abstract

The development of conversational multi-domain spoken dialogue systems poses new challenges for the reliable processing of less restricted user utterances. Unlike in controlled and restricted dialogue systems a simple one-to-one mapping from words to meanings is no longer feasible here. In this paper two different approaches to the resolution of lexical ambiguities are applied to a multi-domain corpus of speech recognition output produced from spontaneous utterances in a spoken dialogue system. The resulting evaluations show that all approaches yield significant gains over the majority class baseline performance of .68, i.e. f-measures of .79 for the knowledge-driven approach and .86 for the supervised learning approach.

1 Introduction

Following Ide and Veronis (1998) we can distinguish between data- and knowledge-driven word sense disambiguation (WSD). Given the basic distinction between written text and spoken utterances, we follow Allen et al. (2001) and differentiate further between controlled and conversational spoken dialogue systems. Neither data- nor knowledge-driven word sense disambiguation has been performed on speech data stemming from human interactions with dialogue systems, since multi-domain conversational spoken dialogue systems for human computer interaction (HCI) have not existed in the past. Now that speech data from multi-domain systems have become available, corresponding experiments and evaluations have become feasible.

In this paper we present the results of first word sense disambiguation annotation experiments on data from spoken interactions with multi-domain dialogue systems. Additionally, we describe the results of a corresponding evaluation of a data- and a knowledge-driven word sense disambiguation system on that data. For knowledge-driven disambiguation we examined whether the ontology-based method for computing semantic coherence introduced by Gurevych et al. (2003a) can be employed to disambiguate between alternative interpretations, i.e. concept representations, of a given speech recognition hypothesis (SRH) at hand. We will show the results of its evaluation in the semantic interpretation task of WSD. For example, in speech recognition hypotheses containing forms of the German verb *kommen*, i.e. (to) come, a decision had to be made whether its meaning corresponds to the motion sense or to the showing sense, i.e. becoming mapped onto either a `MotionDirectedTransliteratedProcess` or a `WatchPerceptualProcess` in the terminology of our spoken language understanding system. For a data-driven approach we employed a highly supervised learning algorithm introduced by Brants (2000) and trained it on a corpus of annotated data. A second set of semantically annotated speech recognition hypotheses was employed as a gold-standard for evaluating both the ontology-based and supervised learning method. Both data sets were annotated by separate human annotators.

All annotated data stems from log files of an automatic speech recognition system that was implemented in the SMARTKOM system (Wahlster et al., 2001; Wahlster, 2003). It is important to point out that there are at least two essential differences between spontaneous speech WSD and textual WSD, i.e.,

- a smaller size of processable context as well as

- imperfections, hesitations, disfluencies and speech recognition errors.

Existing spoken language understanding systems, that are not shallow and thusly produce deep syntactic and semantic representations for multiple domains, e.g. the production system approach described by Engel (2002) or unification-based approaches described by Crysmann et al. (2002), have shown to be more suitable for well-formed input but less robust in case of imperfect input. For conversational and reliable dialogue

systems that achieve satisfactory scores in evaluation frameworks such as proposed by Walker et al. (2000) or Beringer et al. (2002) for multi-modal dialogue systems, we need robust knowledge- or data-driven methods for disambiguating the sometimes less than ideal output of the large vocabulary spontaneous speech recognizers. In the long run, we would also like to avoid expensive pre-processing work, which is necessary for both ontology-driven and supervised learning methods, i.e. labor intensive ontology engineering and data annotation respectively.

2 State of the Art

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.

2.1 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. Despite 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.

2.2 Knowledge-based Methods

Knowledge-based approaches work with lexica and/or ontologies. The kind of knowledge varies widely and machine-readable as well as computer 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 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 past SENSEVAL studies (Edmonds, 2002).

The ontology-based system has been successfully used for a set of tasks such as finding the best speech recognition hypotheses from sets of competing SRHs, labeling SRHs as correct or incorrect representations of the users intention and for scoring their degree of contextual coherence (Gurevych et al., 2003a; Porzel and Gurevych, 2003; Porzel et al., 2003). In general, the system offers an additional way of employing ontologies, i.e. to use the knowledge modeled therein as the basis for evaluating the semantic coherence of sets of concepts. It can be employed independent of the specific ontology language used, as the underlying algorithm operates only on the nodes and named edges of the directed graph represented by the ontology. The specific knowledge base, e.g. written in OIL-RDFS, DAML+OIL or OWL,[1] is converted into a graph, consisting of the class hierarchy, with each class corresponding to a concept representing either an entity or a process and their slots, i.e. the named edges of the graph corresponding to the class properties, constraints and restrictions.

---

[1] OIL-RDFS, DAML+OIL and OWL are frequently used knowledge modeling languages originating in W3C and Semantic Web projects. For more details, see www.w3c.org/RDF, www.w3c.org/OWL and www.daml.org.

3 Data and Annotation Experiment

In this section we describe the data collection and annotation experiments performed in order to obtain independent data sets for training and evaluation.

3.1 Data Collection

The first data set was used for training the supervised model is described in Gurevych et al. (2002b) and was collected using the so-called *Hidden Operator Test* (Rapp and Strube, 2002). This procedure represents a simplification of classical end-to-end experiments and Wizard-of-Oz experiments (Francony et al., 1992) - as it is conductible without the technically very complex use of a real or a seemingly real conversational system. The subjects are prompted to ask for specific information and the system response is pre-manufactured. We had 29 subjects prompted to say certain inputs in 8 dialogues. 1479 turns were recorded. In our experimental setup each user-turn in the dialogue corresponded to a single illocution, e.g. route request or sights information request as described by Gurevych et al. (2002a).

The second data set was used for testing the data- and ontology-based systems and thusly will be called the *test corpus*. It 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.

3.2 Data Pre-Processing

After manual segmentation of the data into single utterances. The resulting audio files were then manually transcribed. The segmented audio files were handed to the speech recognition engine integrated in the SMARTKOM dialogue system (Wahlster, 2003). Employing the semantic parsing system described by Engel (2002) the corresponding speech recognition word lattices (Oerder and Ney, 1993) were first transformed into n-best lists of so-called *hypotheses sequences*. These were mapped onto *conceptual representations*, which contain the multiple semantic interpretations of the individual hypotheses sequences that arise due to lexical ambiguities.

For obtaining the training data, we used only the best, correct and perfectly disambiguated speech recognition hypotheses as described by Porzel et al. (2003) from the first data set of 552 utterances. For obtaining the test data we took a random sample of 3100 utterances from the second data set. This seeming discrepancy between training and test data is due to the fact that only a part of the test data set actually contains ambiguous lexical items and many of the utterances quite similar to each other.

For example, given the utterance shown in its transcribed form in example (1), we then obtained the sequence of recognition hypotheses shown in examples (1a) - (1e).

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.

3.3 Annotation

We employed VISTAE[2] (Müller, 2002) for annotating the data and for creating the corresponding gold-standards for the training and test corpora. The annotation of the data was done by two persons specially trained for the annotation tasks, with different purposes:

- First of all, if humans are able to annotate the data reliably, it is generally more feasible that machines are able to do that as well. This was the case as shown by the resulting inter annotator agreement of 78.89%.

- Secondly, a gold-standard is needed to evaluate the systems' performances. For that purpose, the annotators reached an agreement on annotated items of the test data which had differed in the first place. The resulting gold-standard represents the highest degree of correctly disambiguated data and is used for comparison with the tagged data produced by the disambiguation systems.

- Thirdly, for the supervised learning another correctly disambiguated data set is needed for training the statistical model.

---

[2]The acronym stands for **Vis**ualization **T**ool for **A**nnotation and **E**valuation.

The class-based kappa statistic of (Cohen, 1960; Carletta, 1996) cannot be applied here, as the classes vary depending on the number of ambiguities per entry in the lexicon. Also an additional class, i.e., `not-decidable` was allowed for cases as in SRH (1c), where it is impossible to assign sensible meanings. The test data set altogether was annotated with 2219 markables of ambiguous tokens, stemming from 70 ambiguous words occurring in the test corpus.

3.4 Calculating the Baselines

For calculating the majority class baseline, which in our case corresponds to the performance of a unigram tagger, we applied the method described in (Porzel and Malaka, 2004). Therefore, all markables in the gold-standard were counted and, 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 was calculated. This resulted in a baseline of 52.48% for the test data set.

4 Word Sense Disambiguation Systems

Both word sense disambiguation systems described herein were tested and developed with the SMARTKOM research framework. As one of the most advanced current systems, the SMARTKOM (Wahlster, 2003) comprises a large set of input and output modalities together with an efficient fusion and fission pipeline. SMARTKOM features speech input with prosodic analysis, gesture input via infrared camera, recognition of facial expressions and their emotional states. On the output side, the system features a gesturing and speaking life-like character together with displayed generated text and multimedia graphical output. It currently comprises nearly 50 modules running on a parallel virtual machine-based integration software called *Multiplatform*[3] described in Herzog et al. (2003).

4.1 The Knowledge-driven System

The ontology employed for the evaluation has about 800 concepts and 200 relations (apart from the *isa*-relations defining the general taxonomy) and is described by Gurevych et al. (2003b). It includes a generic top-level ontology whose purpose is to provide a basic structure of the world, i.e. abstract classes to divide the universe in distinct parts as resulting from the ontological analysis.[4] The modeling of *Processes* and *Physical Objects* as a kind of event that is continuous and homogeneous in nature, follows the frame semantic analysis used for generating the FRAMENET data (Baker et al., 1998).

---

[3]The abbreviation stands for "MUltiple Language / Target Integration PLATform FOR Modules".

[4]The top-level was developed following the procedure outlined in Russell and Norvig (1995).

The hierarchy of *Processes* is connected to the hierarchy of *Physical Objects* via slot-constraint definitions herein referred to as *relations*.

The system performs a number of processing steps. A first preprocessing step is to convert each SRH into a *concept representation* (CR). For that purpose the system's lexicon is used, which contains either zero, one or many corresponding concepts for each entry. 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, 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.

For example, the words occurring in a SRH such as (2) have the corresponding entries in the lexicon that are shown below.

2 *Ich bin auf dem Philosphenweg*
   I   am   on   the   Philosopher's Walk

```
<entry>
   <string> Ich </string>
   <concept> Person </concept>
</entry>
<entry>
   <string> bin </string>
   <concept> StaticSpatialProcess </concept>
   <concept> SelfIdentificationProcess </concept>
   <concept> NONE </concept>
</entry>
<entry>
   <string> auf </string>
   <concept> TwoPointRelation </concept>
   <concept> NONE </concept>
</entry>
<entry>
   <string> Philosophenweg </string>
   <concept> Location </concept>
</entry>
```

Since we have multiple concept entries for individual words, i.e. lexical ambiguities, we get a resulting set I of concept representations.

CR1	{Person, StaticSpatialProcess, Location}
CR2	{Person, StaticSpatialProcess, TwoPointRelation, Location}
CR3	{Person, SelfIdentificationProcess, Location}
CR4	{Person, SelfIdentificationProcess, TwoPointRelation, Location}
CR5	{Person, TwoPointRelation, Location}
CR6	{Person, Location}

The concept representations consist of a different number of concepts, because the concept `none` is not represented in the CRs. The concept `none` is assigned to lexemes which have one (or more than one) meaning outside the SmartKom domains or constitute functional grammatical markers.

The system then 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, the ONTOSCORE system 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, 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) was used in the original system.[5]

The new addition made for this evaluation was to assign different weights to the individual relations found by the algorithm, depending on their level of granularity within the relation hierarchy. For example, a broad level relation such as *has-theme* which is found in the class statement of `Process` is weighted with negative 1 as it has only one super-relation, i.e. *has-role*, whereas a more specific relation such as *has-actor* is weighted with negative 4 because it has four super-relations, i.e. *has-artist*, *has-associated-person(s)*, *has-attribute* and *has-role*.

As before, 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.[6] The set of concepts with the lowest aggregate score represents the combination with the highest semantic relatedness.

4.2 The Data-driven System

In this section we describe the implementation of the statistical learning techniques employed for the task of performing WSD on our corpus of spoken dialogue data.

For our experiments we took the general purpose statistical tagger (Brants, 2000), which is generally used for part-of-speech tagging. It employs a VITERBI algorithm for second order Markov models (Rabiner, 1989), linear interpolation for smoothing and deleted interpolation for

determining the weights. According to Edmonds (2002), WSD is in many ways similar to part-of-speech tagging as it involves labeling every word in a text with a tag from a pre-specified set of tag possibilities for each word by using features of the context and other information. This, together with the fact that we do not find cross-paradigmatic ambiguities in our data, led to the idea to use a part-of-speech tagger as a concept tagger.

In our case the tagset consisted of part-of-speech specific concepts of the SmartKom Ontology. The data we used for preparing the model consisted of a combination of three gold-standard annotations, namely the best SRHs, the correct SRHs and the correctly disambiguated SRHs as described in Section 3.3. These were listed lexeme by lexeme with their corresponding concepts in a file in the format expected by TnT. TnT used the file to produce a new model, consisting of a trigram model and a lexicon with lexemes and the concepts which corresponded to them as shown in Figure 1.
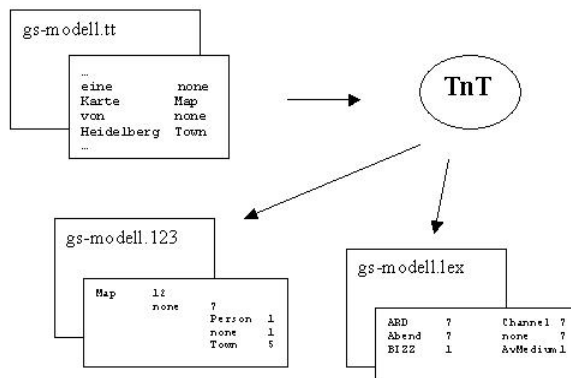

Figure 1: Training the TnT Model

As one can see in Table 1, in our corpus the concept `Greeting` occurred 38 times and was followed 20 times by `Person`, which itself was followed 13 times by `EmotionExperiencerSubjectProcess`. This is equivalent to an utterance beginning with "Hello, I want ...".

The lexicon (see Table 2) shows how often a certain lexeme was tagged with which concept. For example, the German TV channel *ARD* was tagged in all occurrences with the concept `Channel`. The German preposition *am* (at) occurred 17 times and in 12 cases it was tagged as a `TwoPointRelation`, in one case as `TemporalTwoPointRelation` and in 4 cases with `none`. In cases in which the tagger cannot decide between different concepts, because of missing context, it chooses the concept, which occurred most frequently in the model according to the lexicon.

---

[5]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.

[6]Note that more specific relations subtract more then less specific ones from the aggregate score.

1st	2nd	Tokens
Person		20
	EmotionExperiencerSubjectProcess	13
	none	3
	StaticSpatialProcess	4
none		15
	Person	3
	none	3
	WatchPerceptualProcess	3
	InformationSearchProcess	2
	MotionDirectedTransliterated	2
	TvProgram	1
	PatientMotionProcess	1
	InformationSearchProcess	3
	Person	2
	none	1
Total	Greeting Process	38

Table 1: Part of the trigram for `GreetingProcess`

Word	Concept	Tokens
ARD		7
	Channel	7
am		17
	TemporalTwoPointRelation	1
	TwoPointRelation	12
	none	4
kommen		5
	MotionDirectedTransliterated	4
	WatchPerceptualProcess	1
in		12
	TwoPointRelation	8
	none	4

Table 2: Part of the lexicon file: model.lex

5 Evaluation

The percentage of correctly disambiguated lexemes from both systems is calculated by the following formula: $R = (g + n)/a * 100$. Where R is he result in percent, g the number of lexemes that match with the gold-standard, n the number of not-decidable ones and a the number of total lexemes. As opposed to the human annotators, both systems always select a specific reading and never assign the value `not-decidable`. For this evaluation, therefore, we treat any concept occurring in a `not-decidable` slot as correct.[7]

---

[7]Such SRHs usually score below the consistency thresholds described by Gurevych et al. (2003a) and are not passed on.

5.1 Evaluation Knowledge

For this evaluation, ONTOSCORE transformed the SRH from our corpus into concept representations as described above. To perform the WSD task, ONTOSCORE calculates a coherence score for each of these concept sets in I. The concepts in the highest ranked set are considered to be the ones representing the correct word meaning in this context. OntoScore has 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 relation being assigned an individual weight as described in Section 4.1. For this purpose, 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% (f-measure = .78)[8] as compared to 64.75% (f-measure = .79) for the new method described herein (baseline 52.48%).

5.2 Evaluation Supervised

For the purpose of evaluating a supervised learning approach on our data we used the efficient and general statistical TnT tagger, the short form for Trigrams'n'Tags (Brants, 2000). With this tagger it is possible to train a new statistical model with any tagset. In our case the tagset consisted of part-of-speech specific concepts of the SmartKom ontology. The data we used for preparing the model consisted of a gold-standard annotation of the training data set. Compared to the gold-standard made for the test corpus the method achieved a precision of 75.07% (baseline 52.48%).

5.3 Evaluation Comparison

For a direct comparison we computed f-measures for the human reliability, the majority class baseline method as well as for the knowledge-based and data-driven methods in Table 3.

Method	F-measure	$Gain_{precision}$
Baseline	.68	0 %
Knowledge (original)	.78	11.28%
Knowledge (relation)	.79	12.27%
Supervised	.86	22.59%
Annotator agreement	.88	26.41%

Table 3: F-measures and gains on the test data

---

[8]We calculate the standard f-measure (Van Rijsbergen, 1979) with $\alpha = 0.5$by regarding the accuracy as precision and recall as 100%.

6 Discussion

In this paper we presented two methods for disambiguating speech recognition hypotheses. Both methods showed significant gains over the majority class baseline. The results also show that the statistical method outperforms the ontology-based method. This is congruent to findings from textual WSD methods, where the results from data-based approaches frequently yield better scores. However, labeling and training times for these methods are high and costly and they take up a significant amount of memory space. Furthermore, if new domains - featuring lexical ambiguites hitherto unseen by the statistical model - are integrated into the system, new models must consequently be trained in order to keep performance up to par. In such cases, new annotated data has to be made available.

The results of the knowledge-based approach show that ontologies can be employed for such tasks even if they have not been constructed specifically for WSD. Since ontology engineering is at least as costly as annotation and training of statistical models, alternative means for ontology construction and learning need to be pursed. Nonetheless, projects related the semantic web efforts (Heflin and Hendler, 2000) continue to increase their coverage and will become dynamically combinable so that new domains can be integrated in less time without the need of manually processed data.

Our future work will involve the testing of an unsupervised method as well as the improvement of the presented approaches. This will include a compression of the data-based model and experiments concerning the scalability of the knowledge-based approach.

Acknowledgments

References

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.

Nicole Beringer, Ute Kartal, Katerina Louka, Florian Schiel, and Uli Türk. 2002. PROMISE: A Procedure for Multimodal Interactive System Evaluation. In *Proceedings of the Workshop 'Multimodal Resources and Multimodal Systems Evaluation*, Las Palmas, Spain.

Thorsten Brants. 2000. TnT – A statistical Part-of-Speech tagger. In *Proceedings of the 6th Conference on Applied Natural Language Processing,* Seattle, Wash.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

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

Berthold Crysmann, Anette Frank, Kiefer Bernd, Stefan Mueller, Guenter Neumann, Jakub Piskorski, Ulrich Schaefer, Melanie Siegel, Hans Uszkoreit, Feiyu Xu, Markus Becker, and Hans-Ulrich Krieger. 2002. An integrated archictecture for shallow and deep processing. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*.

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.

Iryna Gurevych, Robert Porzel, and Michael Strube. 2002a. Annotating the semantic consistency of speech recognition hypotheses. In *Proceedings of the Third SIGdial Workshop on Discourse and Dialogue*, Philadelphia, USA, July.

Iryna Gurevych, Michael Strube, and Robert Porzel. 2002b. Automatic classification of speech recognition

hypothesis. In *Proceedings of the 3nd SIGdial Workshop on Discourse and Dialogue,* Philadelphia, USA, July 2002, pages 90–95.

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.

Jeff Heflin and James A. Hendler. 2000. Dynamic ontologies on the web. In *Proceedings of AAAI/IAAI*, pages 443–449, Austin, Texas.

Gerd Herzog, Heinz Kirchmann, Stefan Merten, Alassane Ndiaye, Peter Poller, and Tilman Becker. 2003. MULTIPLATFORM: An integration platfrom for multimodal dialogue systems. In *Proceedings of the HLT/NAACL SEALTS 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.

Christof Müller. 2002. *Kontextabh¨ngige Bewertung der Koh¨renz von Spracherkennungshypothesen*. Master Thesis at the Institut für Informationstechnologie der Fachhochschule Mannheim.

Martin Oerder and Hermann Ney. 1993. Word graphs: An efficient interface between continuousspeech 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.

Robert Porzel and Iryna Gurevych. 2003. Contextual coherence in natural language processing. In P. Blackburn, C. Ghidini, R. Turner, and F. Giunchiglia, editors, *Modeling and Using Context*. LNAI 2680, Sprigner, Berlin.

Robert Porzel and Rainer Malaka. 2004. Towards measuring scalability in natural laguage understanding tasks. In *Proceedings of the HLT/NAACL Workshop on Scalable Natural Language Understanding*, Boston, USA. To appear.

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

L.R. Rabiner. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77.

Stefan Rapp and Michael Strube. 2002. An iterative data collection approach for multimodal dialogue systems. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, Las Palmas, Spain.

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.

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*.

C. J. Van Rijsbergen. 1979. *Information Retrieval, 2nd edition*. Dept. of Computer Science, University of Glasgow.

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

Marilyn A. Walker, Candace A. Kamm, and Diane J. Litman. 2000. Towards developing general model of usability with PARADISE. *Natural Language Engeneering*, 6.

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.