Assigning Domains to Speech Recognition Hypotheses

Klaus Rüggenmann and Iryna Gurevych

EML Research gGmbH
Schloss-Wolfsbrunnenweg 33
69118 Heidelberg, Germany
{rueggenmann,gurevych}@eml-r.villa-bosch.de

Abstract

We present the results of experiments aimed at assigning domains to speech recognition hypotheses (SRH). The methods rely on high-level linguistic representations of SRHs as sets of ontological concepts. We experimented with two domain models and evaluated their performance against a statistical, word-based model. Our hand-annotated and tf*idf-based models yielded a precision of 88,39% and 82,59% respectively, compared to 93,14% for the word-based baseline model. These results are explained in terms of our experimental setup.

1 Motivation

High-level linguistic knowledge has been shown to have the potential of improving the state of the art in automatic speech recognition (ASR). Such knowledge can be integrated in the ASR component (Gao, 2003; Gao et al., 2003; Stolcke et al., 2000; Sarikaya et al., 2003; Taylor et al., 2000). Alternatively, it may be included in the processing pipeline at a later stage, namely at the interface between the automatic speech recognizer and the spoken language understanding component (Gurevych et al., 2003a; Gurevych and Porzel, 2003).

In any of these cases, it is necessary to provide a systematic account of domain and world knowledge. These types of knowledge have largely been ignored so far in ASR research. The reason for this state of affairs lies in the fact that the manual construction of appropriate knowledge sources for broad domains is extremely costly. Also, easy domain portability is an important requirement for any ASR system. The emergence of wide coverage linguistic knowledge bases for multiple languages, such as WordNet (Fellbaum, 1998), FrameNet (Baker et al., 1998; Baker et al., 2003), PropBank (Palmer et al., 2003; Xue et al., 2004) is likely to change this situation.

Domain recognition, which is the central topic of this paper, can be thought of as high-level semantic tagging of utterances. We expect significant improvements in the performance of the ASR component of the system if information about the current domain of discourse is available. An obvious intuition behind this expectation is that knowing the current domain of discourse narrows down the search space of the speech recognizer. It also allows to rule out incoherent speech recognition hypotheses as well as those which do not fit in a given domain.

Apart from that, there are additional important reasons for the inclusion of information about the current domain in any spoken language processing (SLP) system. Current SLP systems deal not only with a single, but with multiple domains, e.g., Levin et al. (2000), Itou et al. (2001), Wahlster et al. (2001). In fact, the development of multi-domain systems is one of the new research directions in SLP, which makes the issue of automatically assigning domains to utterances especially important. This type of knowledge can be effectively utilized at different stages of the spoken language and multi-domain input processing in the following ways:

- optimizing the performance of the speech recognizer;
- improving the performance of the dialogue

manager, e.g., if a domain change occurred in the discourse;

- dynamic loading of resources, e.g. speech recognizer lexicons or dialogue plans, especially in mobile environments.

Here, we present the results of research directed at automatic assigning of domains to speech recognition hypotheses. In Section 2, we briefly introduce the knowledge sources in our experiments, such as the ontology, the lexicon and domain models. The data and annotation experiments will be presented in Section 3, followed by the detailed description of the domain classification algorithms in Section 4. Section 5 will give the evaluation results for the linguistically motivated conceptual as well as purely statistical models. Conclusions and some future research directions can be found in Section 6.

2 High-Level Knowledge Sources

2.1 Ontology and lexicon

Current SLP systems often employ multi-domain ontologies representing the relevant world and discourse knowledge. The knowledge encoded in such an ontology can be applied to a variety of natural language processing tasks, e.g. Mahesh and Nirenburg (1995), Flycht-Eriksson (2003).

Our ontology models the domains *Electronic Program Guide*, *Interaction Management*, *Cinema Information*, *Personal Assistance*, *Route Planning*, *Sights*, *Home Appliances Control* and *Off Talk*. The hierarchically structured ontology consists of ca. 720 concepts and 230 properties specifying relations between concepts. For example every instance of the concept `Process` features the relations `hasBeginTime`, `hasEndTime` and `hasState`. A detailed description of the ontology employed in our experiments is given in Gurevych et al. (2003b).

Ontological concepts are high-level units. They allow to reduce the amount of information needed to represent relations existing between individual lexemes and to effectively incorporate this knowledge into automatic language processing. E.g., there may exist a large number of movies in a cinema reservation system. All of them will be represented by the concept `Movie`, thus allowing to map a variety of lexical items (instances) to a single unit (concept) describing their meaning and the relations to other concepts in a generic way.

We did not use the structure of the ontology in an explicit way in the reported experiments. The knowledge was used implicitly to come up with a set of ontological concepts needed to represent the user's utterance.

The high-level domain knowledge represented in the ontology is linked with the language-specific knowledge through a lexicon. The lexicon contains ca. 3600 entries of lexical items and their senses (0 or more), encoded as concepts in the ontology. E.g., the word *am* is mapped to the ontological concepts `StaticSpatialProcess` as in the utterance *I am in New York*, `SelfIdentificationProcess` as in the utterance *I am Peter Smith*, and `NONE`, if the lexeme has a grammatical function only, e.g., *I am going to read a book*.

2.2 Domain models

For scoring high-level linguistic representations of utterances we use a domain model. A domain model is a two-dimensional matrix DM with the dimensions $(\#d \times \#c)$, where $\#d$ and $\#c$ denote the overall number of domain categories and ontological concepts, respectively. This can be formalized as: $DM = (S_{dc})_{d=1,\ldots,\#d,c=1,\ldots,\#c}$, where the matrix elements S_{dc} are domain specificity scores of individual concepts.

We experimented with two different domain models. The first model DM_{anno} was obtained through direct annotation of concepts with respect to domains as reported in Section 3.2. The second domain model DM_{tf*idf} resulted from statistical analysis of *Dataset 1* (described in Section 3.1). In this case, we computed the *term frequency - inverse document frequency* (tf*idf) score (Salton and Buckley, 1988) of each concept for individual domains. In the case of human annotations, we deal with binary values, whereas tf*idf scores range over the interval [0,1].

3 Data and Annotation Experiments

We performed a number of annotation experiments. The purpose of these experiments was to:

- investigate the reliability of the annotations;

- create a domain model based on human annotations;
- produce a training dataset for statistical classifiers;
- set a *Gold Standard* as a test dataset for the evaluation.

All annotation experiments were conducted on data collected in hidden-operator tests following the paradigm described in Rapp and Strube (2002). Subjects were asked to verbalize a predefined intention in each of their turns, the system's reaction was simulated by a human operator. We collected utterances from 29 subjects in 8 dialogues with the system each. All user turns were recorded in separate audio files. These audio files were processed by two versions of our dialogue system with different speech recognition modules. Data describing our corpora is given in Table 1. The first and the second system's runs are referred to as *Dataset 1* and *Dataset 2* respectively.

	Dataset 1	*Dataset 2*
Number of dialogues	232	95
Number of utterances	1479	552
Number of SRHs	2.239	1.375
Number of coherent SRHs	1511	867
Number of incoherent SRHs	728	508

Table 1: Descriptive corpus statistics.

The corpora obtained from these experiments were further transformed into a set of annotation files, which can be read into GUI-based annotation tools, e.g., MMAX (Müller and Strube, 2003). This tool can be adopted for annotating different levels of information, e.g., semantic coherence and domains of utterances, the best speech recognition hypothesis in the N-best list, as well as domains of individual concepts. The two annotators were trained with the help of an annotation manual. A reconciled version of both annotations resulted in the *Gold Standard*. In the following, we present the results of our annotation experiments.

3.1 Coherence, domains of SRHs in *Dataset 1*

The first experiment was aimed at annotating the *speech recognition hypotheses* (SRH) from *Dataset 1* w.r.t. their domains. This process was two-staged. In the first stage, the annotators labeled randomly mixed SRHs, i.e. SRHs without discourse context, for their semantic coherence as *coherent* or *incoherent*. In the second stage, coherent SRHs were labeled for their domains, resulting in a corpus of 1511 hypotheses labeled for at least one domain category. The numbers for ambiguous domain attributions can be found in Table 2. The class distribution is given in Table 3.

Number of domains	*Annotator 1*	*Annotator 2*
1	90.06%	87.11%
2	6.94%	11.27%
3	3.01%	1.28%
4	0%	0.35%

Table 2: Multiple domain assignments in *Dataset 1*.

	Annotator 1	*Annotator* 2
Electr. Program Guide	14.43%	14.86%
Interaction Management	15.56%	15.17%
Cinema Information	5.32%	8.7%
Personal Assistance	0.31%	0.3%
Route Planning	37.05%	36%
Sights	12.49%	12.74%
Home Appliances Control	14.12%	11.22%
Off Talk	0.72%	1.01%

Table 3: Class distribution for domain assignments.

	P(A)	*P(E)*	*Kappa*
Electr. Program Guide	0.9743	0.7246	0.9066
Interaction Management	0.9836	0.7107	0.9434
Cinema Information	0.9661	0.8506	0.7229
Personal Assistance	0.9953	0.9930	0.3310
Route Planning	0.9777	0.5119	0.9544
Sights	0.9731	0.7629	0.8865
Home Appliances Control	0.9626	0.7504	0.8501
Off Talk	0.9871	0.9780	0.4145

Table 4: Kappa coefficient for separate domains.

Table 4 presents the Kappa coefficient values computed for individual categories. *P(A)* is the percentage of agreement between annotators. *P(E)* is the percentage we expect them to agree by chance. The annotations are generally considered to be reliable if $K > 0.8$. This is true for all classes except those which occur very rarely on our data.

3.2 Domains of ontological concepts

In the second experiment, ontological concepts were annotated with zero or more domain categories.[1] We

[1]Top-level concepts like *Event* are typically not domain-specifi c. Therefore, they will not be assigned any domains.

extracted 231 concepts from the lexicon, which is a subset of ontological concepts relevant for our corpus of SRHs. The annotators were given the textual descriptions of all concepts. These definitions are supplied with the ontology. We computed two kinds of inter-annotator agreement. In the first case, we calculated the percentage of concepts, for which the annotators agreed on all domain categories, resulting in ca. 47.62% (CONCabs, see Figure 1). In the second case, the agreement on individual domain decisions (1848 overall) was computed, ca. 86.85% (CONCindiv, see Figure 1).

3.3 Best conceptual representation and domains of SRHs in *Dataset 2*

As will be evident from Section 4.1, each SRH can be mapped to a set of possible interpretations, which are called *conceptual representations* (CR). In this experiment, the best conceptual representation and the domains of coherent SRHs from *Dataset 2* were annotated. As our system operates on the basis of CR, it is necessary to disambiguate them in a preprocessing step.

867 SRHs used in this experiment are mapped to 2853 CR, i.e. on average each SRH is mapped to 3.29 CR. The annotators' agreement on the task of determining the best CR reached ca. 88.93%.

For the task of domain annotation, again, we computed the absolute agreement, when the annotators agreed on all domains for a given SRH. This resulted in ca. 92.5% (SRHabs, see Figure 1). The agreement on individual domain decisions (6936 overall) yielded ca. 98.92% (SRHindiv, see Figure 1). As the Figure 1 suggests, annotating utterances with domains is an easier task for humans than annotating ontological concepts with the same information. One possible reason for this is that even for an isolated SRH of an utterance there is at least some local context available, which clarifies its high-level meaning to some extent. An isolated concept has no defining context whatsoever.

4 Domain Classification

In this section, we present the algorithms employed for assigning domains to speech recognition hypotheses. The system called DOMSCORE performs several processing steps, each of which will be described separately in the respective subsections.

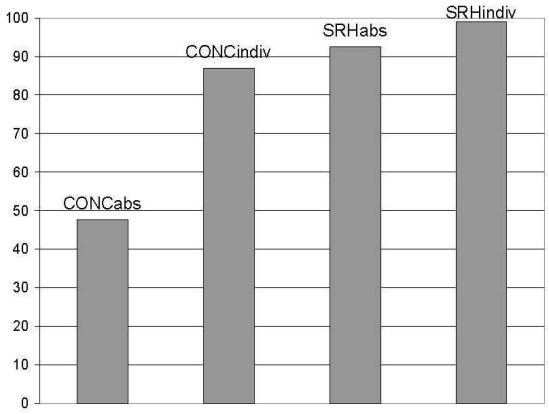

Figure 1: Agreement in % on domain annotations for concepts and SRHs. Absolute agreement (CONCabs, SRHabs) means that annotators agreed on all domains. Individual agreement (CONCindiv, SRHindiv) refers to identical individual domain decisions.

4.1 From SRHs to conceptual representations

SRH is a set of words $W = \{w_1, ..., w_n\}$. DOMSCORE operates on high-level representations of SRHs as *conceptual representations* (CR). CR is a set of ontological concepts $CR = \{c_1, ..., c_n\}$. Conceptual representations are obtained from W through the process called *word-to-concept mapping*. In this process, all possible ontological senses corresponding to individual words in the lexicon are permutated resulting in a set I of possible interpretations $I = \{CR_1, ..., CR_n\}$ for each speech recognition hypothesis.

For example, in our data a user formulated the query concerning the TV program, as:[2]

(1) *Und was für Spielfilme kommen*
And which movies come
heute abend
tonight

This utterance resulted in the following SRHs:

[2] All examples are displayed with the German original and a glossed translation.

SRH_1 *Was für* *Spielfilme* *kommen* *heute abend*
Which movies come tonight

SRH_2 *Was für* *kommen* *heute abend*
Which come tonight

The two hypotheses have two conceptual representations each. This is due to the lexical ambiguity of the word *come* as either `MotionProcess` or `WatchProcess` in German. *Movie* in SRH_1 is mapped to `Broadcast`. As a consequence, the permutation yields $CR_{1a,1b}$ for SRH_1 and $CR_{2a,2b}$ for SRH_2:

CR_{1a}: {`Broadcast, MotionProcess`}
CR_{1b}: {`Broadcast, WatchProcess`}
CR_{2a}: {`MotionProcess`}
CR_{2b}: {`WatchProcess`}

In Tables 5 and 6, the domain specificity scores S_{dc} for all concepts of Example 1 are given.

	Broadcast	*Motion*	*Watch*
Electr. Program Guide	1	0	1
Interaction Management	0	0	0
Cinema Information	0	0	1
Personal Assistance	0	0	0
Route Planning	0	1	1
Sights	0	0	1
Home Appliances Control	1	0	0
Off Talk	0	0	0

Table 5: Matrix DM_{anno} derived from human annotations.

	Broadcast	*Motion*	*Watch*
Electr. Program Guide	1	0.496	0.744
Interaction Management	0	0	0
Cinema Information	0.283	0.178	0.043
Personal Assistance	0	0	0
Route Planning	0	0.689	0.044
Sights	0	0.020	0.079
Home Appliances Control	0.494	0.027	0.147
Off Talk	0	0.238	0.374

Table 6: Matrix DM_{tf*idf} derived from the annotated corpus.

4.2 Domain classification of CR

The domain specificity score S of the conceptual representation CR for the domain d is, then, defined as the average score of all concepts in CR for this domain. For a given domain model DM, this formally means:

$$S_{CR(d)} = \frac{1}{n} \sum_{i=1}^{n} S_{d,i}$$

where n is the number of concepts in the respective CR. As each CR is scored for all domains d, the output of DOMSCORE is a set of domain scores:

$$S_{CR} = \{S_{d_1}, ..., S_{\#d}\}$$

where $\#d$ is the number of domain categories.

Tables 7 and 8 display the results of the domain scoring algorithm for the conceptual representations of Example 1.

	SRH_1		SRH_2	
	CR_{1a}	CR_{1b}	CR_{2a}	CR_{2b}
Electr. Program Guide	0.5	1.0	0	1.0
Interaction Management	0	0	0	0
Cinema Information	0	0.5	0	1.0
Personal Assistance	0	0	0	0
Route Planning	0.5	0.5	1.0	1.0
Sights	0	0.5	0	1.0
Home Appliances Control	0.5	0.5	0	0
Off Talk	0	0	0	0

Table 7: Domain scores on the basis of DM_{anno}.

	SRH_1		SRH_2	
	CR_{1a}	CR_{1b}	CR_{2a}	CR_{2b}
Electr. Program Guide	0.748	0.872	0.496	0.744
Interaction Management	0	0	0	0
Cinema Information	0.231	0.163	0.178	0.043
Personal Asssitance	0	0	0	0
Route Planning	0.344	0.022	0.689	0.044
Sights	0.01	0.04	0.02	0.079
Home Appliances Control	0.26	0.32	0.027	0.147
Off Talk	0.119	0.187	0.238	0.374

Table 8: Domain scores on the basis of DM_{tf*idf}.

In the *Gold Standard* evaluation data, SRH_1 was annotated as the best SRH and attributed the domain *Electronic Program Guide*, CR_{1b} was selected as its best conceptual representation. As can be seen in the above tables, this CR_{1b} gets the highest domain score for *Electronic Program Guide* on the basis of both DM_{anno} and DM_{tf*idf}. Consequently, both domain models attribute this domain to SRH_1.

SRH_2 was not labeled with any domains in the *Gold Standard*, as this hypothesis is an incoherent

one and hence cannot be considered to belong to any domain at all. According to DM_{anno}, its representation CR_{2a} gets a single score 1 for the domain *Route Planning* and CR_{2b} gets multiple equal scores. DOMSCORE interprets a single score as a more reliable indicator for a specific domain than multiple equal scores and assigns the domain *Route Planning* to SRH_2. On the basis of DM_{tf*idf} the highest overall score for $CR_{2a,2b}$ is the one for domain *Electronic Program Guide*. Therefore, the model will assign this domain to SRH_2.

4.3 *Word2Concept* ratio

In previous experiments (Gurevych et al., 2003a), we found that when operating on sets of concepts as representations of speech recognition hypotheses, the ratio of the number of ontological concepts n in a given CR and the total number of words w in the respective SRH must be accounted for. This relation is defined by the ratio $R = n/w$.

The idea is to prevent an incoherent SRH containing many function words with zero concept mappings, represented by a single concept in the extreme, from being classified as coherent. Experimental results indicate that the optimal threshold R should be set to 0.33. This means that if there are more than three words corresponding to a single concept on average, the SRH is likely to be incoherent and should be excluded from processing.

DOMSCORE implements this as a post-processing technique. For both conceptual representations of SRH_1 the ratio is $R = 1/3$, whereas for those of SRH_2, we find $R = 1/5$. This value is under the threshold, which means that SRH_2 is considered incoherent and its domain scores are dropped. Finally, this results in both models assigning the single domain *Electronic Program Guide* as the best one to the utterance in Example 1.

5 Evaluation

5.1 Evaluation metrics

The evaluation of the algorithms and domain models presented herein poses a methodological problem. As stated in Section 3.3, the annotators were allowed to assign 1 or more domains to an SRH, so the number of domain categories varies in the *Gold Standard* data. The output of DOMSCORE, however, is a set with confidence values for all domains ranging from 0 to 1. To the best of our knowledge, there exists no evaluation method that allows the straightforward evaluation of these confidence sets against the varying number of binary domain decisions.

As a consequence, we restricted the evaluation to the subset of 758 SRHs unambiguously annotated for a single domain in *Dataset 2*. For each SRH we compared the recognized domain of its best CR with the annotated domain. This recognized domain is the one that was scored the highest confidence by DOMSCORE. In this way we measured the precision on recognizing the best domain of an SRH. The best conceptual representation of an SRH had been previously disambiguated by humans as reported in Section 3.3. Alternatively, this kind of disambiguation can be performed automatically, e.g., with the help of the system presented in Gurevych et al. (2003a). The system scores semantic coherence of SRHs, where the best CR is the one with the highest semantic coherence.

5.2 Results

We included two baselines in this evaluation. As assigning domains to speech recognition hypotheses is a classification task, the *majority class* frequency can serve as a first baseline. For a second baseline, we trained a statistical classifier employing the *k-nearest neighbour* method using *Dataset 1*. This dataset had also been employed to create the tf*idf model. The statistical classifier treated each SRH as a *bag of words* or *bag of concepts* labeled with domain categories.

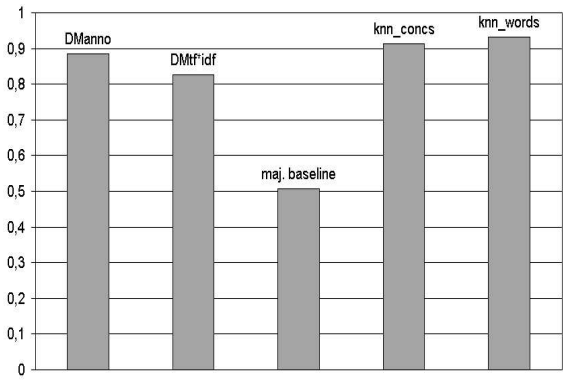

Figure 2: Precision on domain assignments.

The results of DOMSCORE employing the hand-

annotated and tf*idf domain models as well as the baseline systems' performances are displayed in Figure 2. The diagram shows that all systems clearly outperform the *majority class* baseline. The hand-annotated domain model (precision 88.39%) outperforms the tf*idf domain model (precision 82.59%). The model created by humans turns out to be of higher quality than the automatically computed one. However, the *k-nearest neighbour* baseline with words as features performs better (precision 93.14%) than the other methods employing ontological concepts as representations.

5.3 Discussion

We believe that this finding can be explained in terms of our experimental setup which favours the statistical model. Table 9 gives the absolute frequency for all domain categories in the evaluation data. As the data implies, three of the possible categories are missing in the data.

	Number of instances
Electr. Program Guide	74
Interaction Management	85
Cinema Information	0
Personal Assistance	0
Route Planning	385
Sights	150
Home Appliances Control	64
Off Talk	0

Table 9: Class distribution in the evaluation dataset.

The main reason for our results, however, lies in the controlled experimental setup of the data collection. Subjects had to verbalize pre-defined intentions in 8 scenarios, e.g. record a specific program on TV or ask for information regarding a given historical sight. Naturally, this leads to restricted man-machine interactions using controlled vocabulary. As a result, there is rather limited lexical variation in the data. This is unfortunate for illustrating the strengths of high-level ontological representations.

In our opinion, the power of ontological representations is just their ability to reduce multiple lexical surface realizations of the same concept to a single unit, thus representing the meaning of multiple words in a compact way. This effect could not be exploited in a due way given the test corpora in these experiments. We expect a better performance of concept-based methods as compared to word-based ones in broader domains.

An additional important point to consider is the portability of the domain recognition approach. Statistical models, e.g., tf*idf and *k-nearest neighbour* rely on substantial amounts of annotated data when moving to new domains. Such data is difficult to obtain and requires expensive human efforts for annotation. When the manually created domain model is employed for the domain classification task, the extension of knowledge sources to a new domain boils down to extending the list of concepts with some additional ones and annotating them for domains. These new concepts are part of the extension of the system's general ontology, which is not created specifically for domain classification, but employed for many purposes in the system.

6 Conclusions

In this paper, we presented a system which determines domains of speech recognition hypotheses. Our approach incorporates high-level semantic knowledge encoded in a domain model of ontological concepts. We believe that this type of semantic information has the potential to improve the performance of the automatic speech recognizer, as well as other components of spoken language processing systems.

Basically, information about the current domain of discourse is a type of contextual knowledge. One of the future challenges will be to find ways of including this high-level semantic knowledge into SLP systems in the most beneficial way. It remains to be studied how to integrate semantic processing into the architecture, including speech recognition and discourse processing.

An important aspect of the scalability of our methods is their dependence on concept-based domain models. A natural extension would be to replace hand-crafted ontological concepts with, e.g., WordNet concepts. The structure of WordNet can then be used to determine high-level domain concepts that can replace human domain annotations. One of the evident problems with this approach is, however, the high level of lexical ambiguity of the WordNet concepts. Apparently, the problem of ambiguity scales up together with the coverage of the

respective knowledge source.

Another remaining challenge is to define the methodology for the evaluation of methods such as proposed herein. We have to think about appropriate evaluation metrics as well as reference corpora. Following the practices in other NLP fields, such as semantic text analysis (SENSEVAL), message and document understanding conferences (MUC/DUC), it is desirable to conduct rigourous large-scale evaluations. This should facilitate the progress in studying the effects of individual methods and cross-system comparisons.

References

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

Collin F. Baker, Charles J. Fillmore, and Beau Cronin. 2003. The structure of the FrameNet database. *International Journal of Lexicography*, 16.3:281–296.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Mass.

Annika Flycht-Eriksson. 2003. Representing knowledge of dialogue, domain, task and user in dialogue systems - how and why? *Electronic Transactions on Artificial Intelligence*, 3:5–32.

Yuqing Gao, Bowen Zhou, Zijian Diao, Jeffrey Sorensen, and Michael Picheny. 2003. MARS: A statistical semantic parsing and generation-based multilingual automatic translation system. *Machine Translation*, 17(3):185 – 212.

Yuqing Gao. 2003. Coupling vs. unifying: Modeling techniques for speech-to-speech translation. In *Proceedings of Eurospeech*, pages 365 – 368, Geneva, Switzerland, 1-4 September.

Iryna Gurevych and Robert Porzel. 2003. Using knowledge-based scores for identifying best speech recognition hypotheses. In *Proceedings of ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*, pages 77 – 81, Chateau-d'Oex-Vaud, Switzerland, 28-31 August.

Iryna Gurevych, Rainer Malaka, Robert Porzel, and Hans-Peter Zorn. 2003a. Semantic coherence scoring using an ontology. In *Proceedings of the HLT-NAACL Conference*, pages 88–95, 27 May - 1 June.

Iryna Gurevych, Robert Porzel, Elena Slinko, Norbert Pfleger, Jan Alexandersson, and Stefan Merten. 2003b. Less is more: Using a single knowledge representation in dialogue systems. In *Proceedings of the HLT-NAACL'03 Workshop on Text Meaning*, pages 14–21, Edmonton, Canada, 31 May.

Katunobu Itou, Atsushi Fujii, and Tetsuya Ishikawa. 2001. Language modeling for multi-domain speech-driven text retrieval. In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*, December.

Lori Levin, Alon Lavie, Monika Woszczyna, Donna Gates, Marsal Gavalda, Detlef Koll, and Alex Waibel. 2000. The JANUS-III translation system: Speech-to-speech translation in multiple domains. *Machine Translation*, 15(1-2):3 – 25.

K. Mahesh and S. Nirenburg. 1995. A Situated Ontology for Practical NLP. In *Workshop on Basic Ontological Issues in Knowledge Sharing, International Joint Conference on Artificial Intelligence (IJCAI-95)*, Montreal, Canada, 19-20 August.

Christoph Müller and Michael Strube. 2003. Multi-level annotation in MMAX. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*, pages 198–207, Sapporo, Japan, 4-5 July.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2003. The Proposition Bank: An annotated corpus of semantic roles. *Submitted to Computational Linguistics*, December.

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*, pages 661–665, Las Palmas, Canary Island, Spain, 29-31 May.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523.

Ruhi Sarikaya, Yuqing Gao, and Michael Picheny. 2003. Word level confidence measurement using semantic features. In *Proceedings of ICASSP*, Hong Kong, April.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–373.

Paul Taylor, Simon King, Steve Isard, and Helen Wright. 2000. Intonation and dialogue context as constraints for speech recognition. *Language and Speech*, 41(3-4):493–512.

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*, pages 1547–1550.

Nianwen Xue, Fei Xia, Fu-dong Chiou, and Martha Palmer. 2004. The Penn Chinese Treebank: Phrase Structure Annotation of a Large Corpus. *Natural Language Engineering*, 10(4):1–30, June.