Ontology-based Contextual Coherence Scoring

Robert Porzel **Iryna Gurevych** **Christof E. Müller**

European Media Laboratory
Schloss-Wolfsbrunnenweg 31c
D-69118 Heidelberg
{porzel,gurevych,mueller2@eml.org}

Abstract

In this paper we present a contextual extension to ONTOSCORE, a system for scoring sets of concepts on the basis of an ontology. We apply the contextually enhanced system to the task of scoring alternative speech recognition hypotheses (SRH) in terms of their semantic coherence. We conducted several annotation experiments and showed that human annotators can reliably differentiate between semantically coherent and incoherent speech recognition hypotheses (both with and without discourse context). We also showed, that annotators can reliably identify the overall best hypothesis from a given n-best list. While the original ONTOSCORE system correctly assigns the highest score to 84.06% of the corpus, the inclusion of the conceptual context increases the number of correct classifications to yield 86.76%, given a baseline of 63.91% in both cases.

1 Introduction

Following Allen et al. (2001), we can distinguish between controlled and conversational dialogue systems. Since controlled and restricted interactions between the user and the system increase recognition and understanding accuracy, such systems are reliable enough to be deployed in various real world applications, e.g. public transportation or cinema information systems. The more conversational a dialogue system becomes, the less predictable are the users' utterances. Recognition and processing become increasingly difficult and unreliable.

Today's dialogue systems employ domain- and discourse-specific knowledge bases, so-called *ontologies*, to represent the individual discourse entities as *concepts* as well as their relations to each other. In this paper we employ an algorithm for measuring the semantic coherence of sets of concepts using such an ontology and show how its performance can be improved by means of an inclusion of the *conceptual context*. Thereby creating a method for scoring the *contextual coherence* of individual sets of concepts.

In the following, we will show how the contextual coherence measurement can be applied to estimate how well a given speech recognition hypothesis (SRH) fits with respect to the existing knowledge representation and the given conceptual context, thereby providing a mechanism that increases the robustness and reliability of dialogue systems. We can, therefore, show how the algorithm can be successfully employed by a spoken dialogue system to enhance the interface between automatic speech recognition (ASR) and natural language understanding (NLU).

In Section 2 we discuss the problem of scoring and classifying SRHs in terms of their semantic coherence followed by a description of our annotation experiments and the corresponding results in Section 3. Section 4 contains a description of the kind of knowledge representations and the algorithm employed by ONTOSCORE. In Section 5 we present the contextually enhanced system. Evaluations of the corresponding system for scoring SRHs are given in Section 6. A conclusion and additional applications are given in Section 7.

2 Semantic Coherence and Speech Recognition Hypotheses

While a simple one-best hypothesis interface between ASR and 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, given the distribution of acoustic and language model scores (Schwartz and Chow, 1990; Tran et al., 1996). For example, in our data a user expressed the wish to get from Cologne to Heidelberg and then to continue his visit in Heidelberg, as:[1]

(1) *ich möchte auf dem schnellsten Weg*
I want on the fastest way
von Köln nach Heidelberg.
from Cologne to Heidelberg.

(2) *wie komme ich in Heidelberg weiter.*
how can I in Heidelberg continue.

Looking at the SRHs from the ensuing n-best list of Example (1) we found that Example (1a) constituted the best representation of the utterance, whereas all others constituted less adequate representations thereof.

(1a) *ich möchte auf schnellsten Weg von*
I want on fastest way from
Köln nach Heidelberg.
Cologne to Heidelberg.

(1b) *ich möchte auf schnellsten Weg Köln*
I want on fastest way Cologne
nach Heidelberg.
to Heidelberg.

(1c) *ich möchte Folk Weg von Köln*
I want folk way from Cologne
nach Heidelberg.
to Heidelberg.

(1d) *ich möchte auf schnellsten Weg vor*
I want on fastest way before
Köln nach Heidelberg.
Cologne to Heidelberg.

(1e) *ich möchte vor schnellsten Weg von*
I want before fastest way from
Köln nach Heidelberg.
Cologne to Heidelberg.

Facing multiple representations of a single utterance consequently poses the question, which 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. Frequently the scores provided by the ASR system itself are used, e.g. acoustic and language model probabilities. More recently also scores provided by the NLU system have been employed, e.g. parsing scores (Engel, 2002) or discourse model scores (Pfleger et al., 2002). However, these methods often assign very high scores to SRHs which are semantically incoherent and low scores to semantically coherent ones.

In the case of Example (1) all scores, i.e. the acoustic, language model, parsing and the ONTOSCORE scores assign the highest score to Example (1a) (see Table 2 for the actual numbers). SRH 1a can consequently be chosen as the best SRH. As we will show in Section 6, the scoring of the SRHs from Example (2) differs substantially, and only the contextual coherence score manages to pick an adequate SRH. The fact that neither of the other scoring approaches systematically employs the system's knowledge of the domains at hand, can result in passing suboptimal SRHs through the system. This means that, while there was a better representation of the actual utterance in the n-best list, the NLU system is processing an inferior one, thereby causing overall dialogue metrics, in the sense of Walker et al. (2000), to decrease. We, therefore, propose an alternative way to rank SRHs on the basis of their contextual coherence, i.e. with respect to a given ontology representing the domains of the system and the given conceptual context.

3 Annotation Experiments

The experiments reported here are based on the data collected in hidden-operator tests where subjects were prompted to say certain inputs. We obtained 232 dialogues, which were divided into 1479

[1] All examples are displayed with the German original on top and a glossed translation below.

audio files with single user utterances. Each utterance corresponded to a single intention, e.g. a route- or a sight information request. Firstly, all utterances were also transcribed. Then the audio files were sent to the speech recognizer. We logged the speech recognition output, i.e. n-best lists of SRHs for all utterances. A subset of the corpus was used to log also the scores of the recognizer, parser and that of OntoScore - including context-independent and context-dependent semantic coherence scores. This trial resulted in a sub-corpus of 552 utterances corresponding to 1,375 SRHs along with the respective confidence scores.

We, then, conducted several annotation experiments with a two-fold motivation. In the first place, it was necessary to produce a hand-annotated corpus to be used as a *gold standard* for the evaluation of the contextual coherence scores. Furthermore, we wanted to test whether human subjects were able to annotate the data reliably according to our annotation schemata. We had two annotators specially trained for each of these particular annotation tasks.

In an earlier annotation experiment reported in Gurevych et al. (2002), the task of annotators was to classify a subset of the corpus of SRHs as either coherent or incoherent. Here we randomly mixed SRHs in order to avoid contextual priming.[2] In the first new experiment, a sub-corpus of 552 utterances was annotated within the discourse context, i.e. the SRHs were presented in their original dialogue order. For each SRH, a decision again had to be made whether it is semantically coherent or incoherent with respect to the best SRH representing the previous user utterance. Given a total of 1,375 markables, the annotators reached an agreement of 79.71%, i.e. 1,096 markables.

In the second new annotation experiment, the annotators saw the SRHs together with the transcribed user utterances. The task of 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 intention 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.[3] Each corpus was then tranformed into an evaluation *gold standard* by means of the annotators agreeing on a single solution for the cases of disagreement.

The aim of the work presented here, then, was to provide a knowledge-based score, that can be employed by any NLU system to select the best hypothesis from a given n-best list. The corresponding ONTOSCORE system will be described below, followed by its evaluation against the human *gold standards*.

4 The Knowledge Base and OntoScore

In this section, we provide a description of the underlying algorithm and knowledge sources employed by the original ONTOSCORE system (in press). It is important to note that the ontology employed in this and the previous evaluations existed already and was crafted as a general knowledge representation for various processing modules within the system.[4] Ontologies have traditionally been used to represent general and domain specific knowledge and are employed for various natural language understanding tasks, e.g. semantic interpretation (Allen, 1987) and in spoken dialogue systems, e.g. for discourse modeling, modality fusion and dialogue management, see also Porzel et al. (2003) for an overview. ONTOSCORE 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 independently 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 DAML+OIL

[2] As reported elsewhere the resulting Kappa statistics (Carletta, 1996) over the annotated data yields $\kappa = 0.7$, which indicates that human annotators can reliably distinguish between coherent samples and incoherent ones.

[3] A Kappa-statistic suitable for measuring the reliability of annotations is not possible in this case. The Kappa-statistic is class-based and cannot, therefore, be applied to the best SRH labeling, due to the different number of SRHs in the n-best lists. Therefore, we calculated the percentage of utterances, where the annotators agreed on the best SRH.

[4] Alternative knowledge representations, such as WORDNET, could have been employed in theory as well, however most of the *modern* domains of the system, e.g. electronic media or program guides, are not covered by WORDNET.

or OWL,[5] 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.

The ontology employed for the evaluation has about 730 concepts and 200 relations. 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.[6] 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). The hierarchy of *Processes* is connected to the hierarchy of *Physical Objects* via slot-constraint definitions. See also (Gurevych et al., 2003b) for a further description of the ontology.

ONTOSCORE performs a number of processing steps. A first preprocessing step is to convert each SRH 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.

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, we 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.

Demetriou and Atwell (1994) do not provide concrete evaluation results for the method. Also, their algorithm only allows for a relative judgment stating which of a set of interpretations given a single sentence is more semantically related.

Since our objective is to compute coherence scores of arbitrary CRs on an absolute scale, certain extensions were necessary. In this application the CRs to be scored can differ in terms of their content, the number of concepts contained therein and their mappings to the original SRH. Moreover, in order to achieve absolute values, the final score should be related to the number of concepts in an individual set and the number of words in the original SRH. Therefore, the results must be normalized in order to allow for evaluation, comparability and clearer interpretation of the semantic coherence scores.

We modified the algorithm described above to make it applicable and evaluatable with respect to the task at hand as well as other possible tasks. The basic idea is to calculate a score based on the path distances in CR. Since short distances indicate coherence and many concept pairs in a given CR may

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

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

have no connecting path, we define the distance between two concepts c_i and c_j that are not connected in the knowledge base as D_{max}. This maximum value can also serve as a maximum for long distances and can thus help to prune the search tree for long paths. This constant has to be set according to the structure of the knowledge base. For example, employing the ontology described above, the maximum distance between two concepts does not exceed ten and we chose in that case $D_{max} = 10$.

We can now define the semantic coherence score for CR as the average path length between all concept pairs in CR:

$$S(CR) = \frac{\sum_{c_i,c_j \in CR, c_i \neq c_j} D(c_i, c_j)}{|CR|^2 - |CR|}$$

Since the ontology is a directed graph, we have $|CR|^2 - |CR|$ pairs of concepts with possible directed connections, i.e., a path from concept c_i to concept c_j may be completely different to that from c_j to c_i or even be missing. As a symmetric alternative, we may want to consider a path from c_i to c_j and a path from c_j to c_i to be semantically equivalent and thus model every relation in a bidirectional way. We can then compute a symmetric score $S'(CR)$ as

$$S'(CR) = 2\frac{\sum_{c_i,c_j \in CR, i<j} min(D(c_i, c_j)D(c_j, c_i))}{|CR|^2 - |CR|}$$

ONTOSCORE implements both options. As the ontology currently employed features mostly unidirectional relations we chose the $S'(CR)$ function for the evaluation, i.e. only the best path $D(c_i, c_j)$ between a given pair of concepts, regardless of the direction, is taken into account. A detailed description of the original system can be found in (Gurevych et al., 2003a).

5 Contextual Coherence Scoring

The contextually enhanced ONTOSCORE system performs a number of additional processing steps, each of them will be described below.

5.1 Scoring Conceptual Context Representations

A necessary preprocessing step for the conceptual context scoring of SRHs is to build a conceptual context representation $CR'(SRH_{n+1})$ resulting from a pair of concept representations:

- a concept representation of the SRH to be scored, i.e. $CR(SRH_{n+1})$,
- and a concept representation of the preceding utterance's SRH, i.e. $CR(SRH_n)$.

For that purpose, the ONTOSCORE stores the best concept representation from each dialogue turn as $CR_{best}(SRH)$. By the best CR we mean the interpretation which received the highest score from the ONTOSCORE system, from the list of alternative interpretations of the utterance. For example CR_{best} for the utterance shown in Example (1) is the CR of the SRH given in (1e), i.e. {EmotionExperiencerSubjectProcess, Person, TwoPointRelation, Route, Town, Town}.

To produce a conceptual context representation for SRH_{n+1}, we build a union of each of its possible interpretations $I = \{CR_1, CR_2, \ldots, CR_n\}$ with the stored $CR_{best}(SRH_n)$ from the previous utterance. This results in a contextually augmented new set $I' = \{CR'_1, CR'_2, \ldots, CR'_n\}$ representing possible conceptual context interpretations of SRH_{n+1} as shown in Table 1.

$I(SRH_{n+1})$			$I'(SRH_{n+1})$
CR_1	$\cup$	$CR_{best}(SRH_n)$	$= CR'_1$
CR_2	$\cup$	$CR_{best}(SRH_n)$	$= CR'_2$
...		...	...
CR_n	$\cup$	$CR_{best}(SRH_n)$	$= CR'_n$

Table 1: Creating conceptual context representations

If, however, the calculated score of $CR_{best}(SRH_n)$ is below a certain threshold, meaning that even the best prior hypothesis is most likely not semantically coherent, then $CR_{best}(SRH_n) = \{\emptyset\}$. See Section 6.2 for the corresponding numbers with respect to the coherent *versus* incoherent classification. Thusly, only if $CR_{best}(SRH_n)$ is empty then solely the concept representations of SRH_{n+1} are taken into account. This is, of course, also the case at the first dialogue turn.

In order to score the alternative conceptual context representations defined by $I'(SRH_{n+1})$, the formula for $S'(CR)$ is employed. This means that we calculate a conceptual context coherence score S' for each conceptual context representation CR'. We also perform an inverse linear transformation of

the scores resulting in numbers from 0 to 1, so that higher scores indicate better contextual coherence.

5.2 ONTOSCORE at Work

Looking at an example of ONTOSCORE at work, we will examine the following discourse fragment consisting of the two sequential utterances given in Example (1) and (2). As shown in Table 2, in the case of Example (1) all scores indicate the SRH given in Example (1a) to be the best one.

SRH	recognizer	parser	OntoScore
1a	1	1	.6
1b	.74	.94	.6
1c	.63	.94	.54
1d	.78	.89	.54
1e	.74	.88	.54

Table 2: The scores for the SRHs of Example (1).

Example (2) yields the following SRHs with the corresponding context-independent CRs and context-dependent CR's:

2a *Rennen Lied Comedy Show Heidelberg*
Race song comedy show Heidelberg
weiter.
continue.
$CR\{$MusicPiece, Genre, Genre, Town$\}$
$CR'\{$MusicPiece, Genre, Genre, Town,
EmotionExperiencerSubjectProcess,Person,
TwoPointRelation, Route $\}$

2b *denn wie Comedy Heidelberg weiter.*
then how comedy Heidelberg continue.
$CR\{$Genre, Town$\}$
$CR'\{$Genre, Town,
EmotionExperiencerSubjectProcess,Person,
TwoPointRelation, Route $\}$

2c *denn wie Comedy Show weiter.*
then how comedy show continue.
$CR\{$Genre, Genre$\}$
$CR'\{$Genre, Genre,
EmotionExperiencerSubjectProcess, Person,
TwoPointRelation, Route $\}$

2d *denn wie Comedy weiter.*
then how comedy continue.
$CR\{$Genre$\}$
$CR'\{$Genre,
EmotionExperiencerSubjectProcess, Person,
TwoPointRelation, Route $\}$

2e *denn wie komme ich in Heidelberg*
then how can I in Heidelberg
weiter.
continue.
$CR\{$MotionDirectedTransliterated, Person,
Town$\}$
$CR'\{$MotionDirectedTransliterated, Person,
Town, EmotionExperiencerSubjectProcess,
TwoPointRelation, Route $\}$

Adding the conceptual context we get the results shown in Table 3 for Example (2):

SRH	recognizer	parser	OntoScore
2a	1	.25	.32
2b	.52	.2	.48
2c	.34	.2	.39
2d	.35	.12	0
2e	.52	.08	.71

Table 3: The scores for the SRHs of Example 2.

As evident from Table 3, CR'_{best} corresponds to Example 2e. This means that 2e constitutes a more contextually coherent concept structure than the alternative SRHs. This SRH was also labeled both as the best and as a coherent SRH by the annotators.

6 Evaluation

The ONTOSCORE software runs as a module in the SMARTKOM multi-modal and multi-domain spoken dialogue system (Wahlster et al., 2001). 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.

The ONTOSCORE module currently employs two knowledge sources, an ontology (about 730 concepts and 200 relations) and a lexicon (ca. 3.600 words) with word to concept mappings, covering the respective domains of the system. 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 humans subjects according to two separate annotation schemata. The results of annotation experiments are reported in Section 3.

6.1 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 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. The ONTOSCORE module without the conceptual context feature yields the accuracy of only 84.06% on the same task. This suggests that the overall results in identifying the best SRH in the speech recognizer output can by improved by taking the knowledge of conceptual context into account.

6.2 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. This indicates the improvement of 7.93% over the baseline. We also conducted the same classification experiment with ONTOSCORE without using the conceptual context feature. In this case we obtained 69.96% accuracy.

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 (see Section 3) 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.

7 Concluding Remarks

The contextually enhanced ONTOSCORE system described herein automatically performs ontology-based scoring of sets of concepts which constitute an adequate and suitable representation of a speech recognition hypothesis and the prior conceptual context. This conceptual context is an analogous conceptual representation of the previous user utterance. To date, the algorithm has been implemented in a software which is employed by a multi-domain spoken dialogue system and applied to the task of scoring n-best lists of SRH, thus producing a score expressing how well a given SRH fits within the domain model and the given discourse. In the evaluation of our system we employed an ontology that was not designed for this task, but already existed as the system's internal knowledge representation. As shown above, the inclusion of the conceptual discourse context yields an improvement of almost 3% as compared to the context-independent system.

As future work we will examine how the computation of a contextual coherence score, i.e. how well a given SRH fits within the domain model with respect to the previous discourse, can be em-

ployed to detect domain changes in complex multimodal and multi-domain spoken dialogue systems. As one would expect, a contextual coherence score as described above actually decreases when the user changed from one domain to another, which most likely also accounts for a set of the actual misclassifications. As a future enhancement we will integrate and evaluate an automatic domain change detection function, which, if activated, will cause the system to employ the context-independent scoring function. Currently, we are also investigating whether the proposed method can be applied to scoring sets of potential candidates for resolving the semantic interpretation of ambiguous, polysemous and metonymic language use (Porzel and Gurevych, 2003). Additionally, As ontology building is constly, we examine the feasibility to employ alternative knowledge sources, that are generated automatically from corpora, e.g. via self organizing maps.

Acknowledgments

There work described herein was conducted within the SmartKom project partly funded by the German ministry of Research and Technology under grant 01IL95I7 and by the Klaus Tschira Foundation.

References

James F. Allen, Georga Ferguson, and Amanda Stent. 2001. An architecture for more realistic conversational system. In *Proceedings of Intelligent User Interfaces*, pages 1–8, Santa Fe, NM.

James F. Allen. 1987. *Natural Language Understanding*. Menlo Park, Cal.: Benjamin Cummings.

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

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

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.

Ralf Engel. 2002. SPIN: Language understanding for spoken dialogue systems using a production system approach. In *Proceedings of ICSLP 2002*.

Iryna Gurevych, Robert Porzel, and Michael Strube. 2002. Annotating the semantic consistency of speech recognition hypotheses. In *Proceedings of the Third SIGdial Workshop on Discourse and Dialogue*, pages 46–49, Philadelphia, USA, July.

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

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*, Edmonton, Canada.

Martin Oerder and Hermann Ney. 1993. Word graphs: An efficient interface between continuous-speech recognition and language understanding. In *ICASSP Volume 2*, pages 119–122.

Norbert Pfleger, Jan Alexandersson, and Tilman Becker. 2002. Scoring functions for overlay and their application in discourse processing. In *KONVENS-02*, Saarbrücken, September – October.

Robert Porzel and Iryna Gurevych. 2003. Contextual coherence in natural language processing. *Modeling and Using Context*, Springer, LNCS:to appear.

Robert Porzel, Norbert Pfleger, Stefan Merten, Markus Löckelt, Iryna Gurevych, Ralf Engel, and Jan Alexandersson. 2003. More on less: Further applications of ontologies in multi-modal dialogue systems. In *Proceedings of the IJCAI'03 Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, page to appear.

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

Richard Schwartz and Ye-Lo 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*.

Bach-Hiep Tran, Frank Seide, Volker Steinbiss, Richard Schwartz, and Ye-Lo 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., pages 1547–1550.

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