Semantic Coherence Scoring Using an Ontology

Proceedings of HLT-NAACL 2003
Main Papers , pp. 9-16
Edmonton, May-June 2003

Iryna Gurevych **Rainer Malaka** **Robert Porzel** **Hans Peter Zorn**

European Media Lab GmbH
Schloss-Wolfsbrunnenweg 31c
D-69118 Heidelberg, Germany
{gurevych,malaka,porzel,zorn@eml.org}

Abstract

In this paper we present ONTOSCORE, a system for scoring sets of concepts on the basis of an ontology. We apply our system to the task of scoring alternative speech recognition hypotheses (SRH) in terms of their semantic coherence. We conducted an annotation experiment and showed that human annotators can reliably differentiate between semantically coherent and incoherent speech recognition hypotheses. An evaluation of our system against the annotated data shows that, it successfully classifies 73.2% in a German corpus of 2.284 SRHs as either coherent or incoherent (given a baseline of 54.55%).

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*, and their relations to each other. In this paper we present an algorithm for measuring the *semantic coherence* of sets of concepts against such an ontology. In the following, we will show how the semantic coherence measurement can be applied to estimate how well a given speech recognition hypothesis (SRH) fits with respect to the existing knowledge representation, thereby providing a mechanism that increases the robustness and reliability of dialogue systems.

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 experiment. Section 3 contains a description of the kind of knowledge representations employed by ONTOSCORE. We present the algorithm in Section 4, and an evaluation of the corresponding system for scoring SRHs is given in Section 5. A conclusion and additional applications are given in Section 6.

2 Semantic Coherence and Speech Recognition Hypotheses

2.1 The Problem

While a simple one-best hypothesis interface between automatic speech recognition (ASR) and natural language understanding (NLU) suffices for restricted dialogue systems, more complex systems either operate on n-best lists as ASR) output or convert ASR word graphs (Oerder and Ney, 1993) into n-best lists, 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 see a specific city map again, as:[1]

(1)	*Ich*	*würde*	*die*	*Karte*	*gerne*	*wiedersehen*
	I	would	the	map	like	to see again

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

(1a)	*Ich*	*würde*	*die*	*Karte*	*eine*	*wieder*	*sehen*
	I	would	the	map	one	again	see

(1b)	*Ich*	*würde*	*die*	*Karte*	*eine*	*Wiedersehen*
	I	would	the	map	one	Good Bye

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

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

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 or discourse scores (Litman et al., 1999; Engel, 2002; Alexandersson and Becker, 2003). However, these methods assign higher scores to SRHs which are semantically incoherent and lower scores to semantically coherent ones and disagree with other.

For instance, the acoustic and language model scores of Example (1b) are actually better than for Example (1a), which results from the fact that the frequencies and corresponding probabilities for important expressions, such as *Good Bye*, are rather high, thereby ensuring their reliable recognition. Another phenomenon found in our data consists of hypotheses such as:

(2) *Zeige mir alle Vergnügen*
Show me all pleasures

(3) *Zeige mir alle Filmen*
Show me all Films

In these cases language model scores are higher for Example (2) than Example (3), as the incorrect inflection on *alle Filmen* was less frequent in the training material than that of the correct inflection on *alle Vergnügen*.

Our data also shows - as one would intuitively expect - that the understanding-based scores generally reflect how well a given SRH is covered by the grammar employed. In many less well-formed cases these scores do not correspond to the *correctness* of the SRH. Generally we find instances where all existing scoring methods disagree with each other, diverge from the actual word error rate and ignore the semantic coherence.[2] Neither of the aforementioned approaches systematically employs the system's knowledge of the domains at hand. This increases the number of times where a suboptimal recognition hypothesis is passed 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 propose an alternative way to rank SRHs on the basis of their *semantic coherence* with respect to a given ontology representing the domains of the system.

2.2 Annotation Experiments

In a previous study (Gurevych et al., 2002), we tested if human annotators could reliably classify SRHs in terms of their semantic coherence. The task of the annotators was to determine whether a given hypothesis representsa n internally coherent utterance or not.

In order to test the reliability of such annotations, we collected a corpus of SRHs. The data collection was conducted by means of a hidden operator test. We had 29 subjects prompted to say certain inputs in 8 dialogues. 1.479 turns were recorded. Each user-turn in the dialogue corresponded to a single intention, e.g. a route request or a sight information request. The audio files were then sent to the speech recognizer and the input to the semantic coherence scoring module, i.e. n-best lists of SRHs were recorded in log-files. The final corpus consisted of 2.284 SRHs. All hypotheses were then randomly mixed to avoid contextual influences and given to separate annotators. The resulting Kappa statistics (Carletta, 1996) over the annotated data yields $\kappa = 0.7$, which seems to indicate that human annotators can reliably distinguish between coherent samples (as in Example (1a)) and incoherent ones (as in Example (1b)).

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. ONTOSCORE, the resulting system will be described below, followed by its evaluation against the human *gold standard*.

3 The Knowledge Base

In this section, we provide a description of the pre-existing knowledge source employed by ONTOSCORE, as far as it is necessary to understand the empirical data generated by the system. It is important to note that the ontology employed in this evaluation existed already and was crafted as a general knowledge representation for various processing modules within the system.[3]

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). We propose 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.

The system described herein 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 or OWL,[4] is converted into a graph, consisting of:

[2] As the numbers evident from large vocabulary speech recognition performance (Cox et al., 2000), the occurrence of less well formed and incoherent SRHs increases the more conversational a system becomes.

[3] 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.

[4] DAML+OIL and OWL are frequently used knowledge modeling languages originating in W3C and Semantic Web

- the class hierarchy, with each class corresponding to a concept representing either an entity or a process;
- the slots, i.e. the named edges of the graph corresponding to the class properties, constraints and restrictions.

The ontology employed herein 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. The top-level was developed following the procedure outlined in Russell and Norvig (1995).

In the view of the ontology employed herein, `Role` is the most general class in the ontology and represents a role that any entity or process can perform. It is divided into `Event` and `Abstract Event`. `Event` is used to describe a kind of role any entity or process may have in a "real" situation or process, e.g. a building or an information search. It is contrasted with `Abstract Event`, which is abstracted from a set of situations and processes. It reflects no reality and is used for the general categorization and description, e.g. `Number`, `Set`, `Spatial Relation`. There are two kinds of events: `Physical Object` and `Process`.

The class `Physical Object` describes any kind of objects we come in contact with - living as well as non-living - having a location in space and time in contrast to abstract objects. These objects refer to different domains, such as `Sight` and `Route` in the tourism domain, `Av Medium` and `Actor` in the TV and cinema domain, etc., and can be associated with certain relations in the processes via slot constraint definitions.

The modeling of `Process` 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). Currently, there are four groups of processes (see Figure 1):

- `General Process`, a set of the most general processes such as duplication, imitation or repetition processes;
- `Mental Process`, a set of processes such as cognitive, emotional or perceptual processes;
- `Physical Process`, a set of processes such as motion, transaction or controlling processes;
- `Social Process`, a set of processes such as communication or instruction processes.

Let us consider the definition of the `Information Search Process` in the ontology. It is modeled as a subclass of the `Cognitive Process`, which is a subclass of the `Mental Process` and inherits the following slot constraints:

- **begin time**, a time expression indicating the starting time point;
- **end time**, a time expression indicating the time point when the process is complete;
- **state**, one of the abstract process states, e.g. start, continue, interrupt, etc.;
- **cognizer**, filled with a class *Person* including its subclasses.

`Information Search Process` features one additional slot constraint, **piece-of-information**. The possible slot-fillers are a range of domain objects, e.g. `Sight`, `Performance`, or whole `set`s of those, e.g. `Tv Program`, but also processes, e.g. `Controlling Tv Device Process`. This way, an utterance such as:

(4) *I hätte gerne Informationen zum Schloss*
I would like information about castle

can also be mapped onto `Information Search Process`, which has an agent of type `User` and a piece of information of type `Sight`. `Sight` has a name of type `Castle`. Analogously, the utterance:

(5) *Wie kann ich den Fernseher steuern*
How can I the TV control

can be mapped onto `Information Search Process`, which has an agent of type `User` and has a piece of information of type `Controlling Tv Device Process`.

4 Ontology-based Scoring of SRHs

ONTOSCORE performs a number of processing steps, each of them will be described separately in the respective subsections.

4.1 Mapping of SRH to Sets of Concepts

A necessary 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 the concepts, corresponding to the words in the SRH for which concepts in the lexicon exist, constitutes the resulting CR. All other words with empty concept mappings, e.g. articles, are ignored in the conversion.

projects. For more detail, see www.w3c.org.

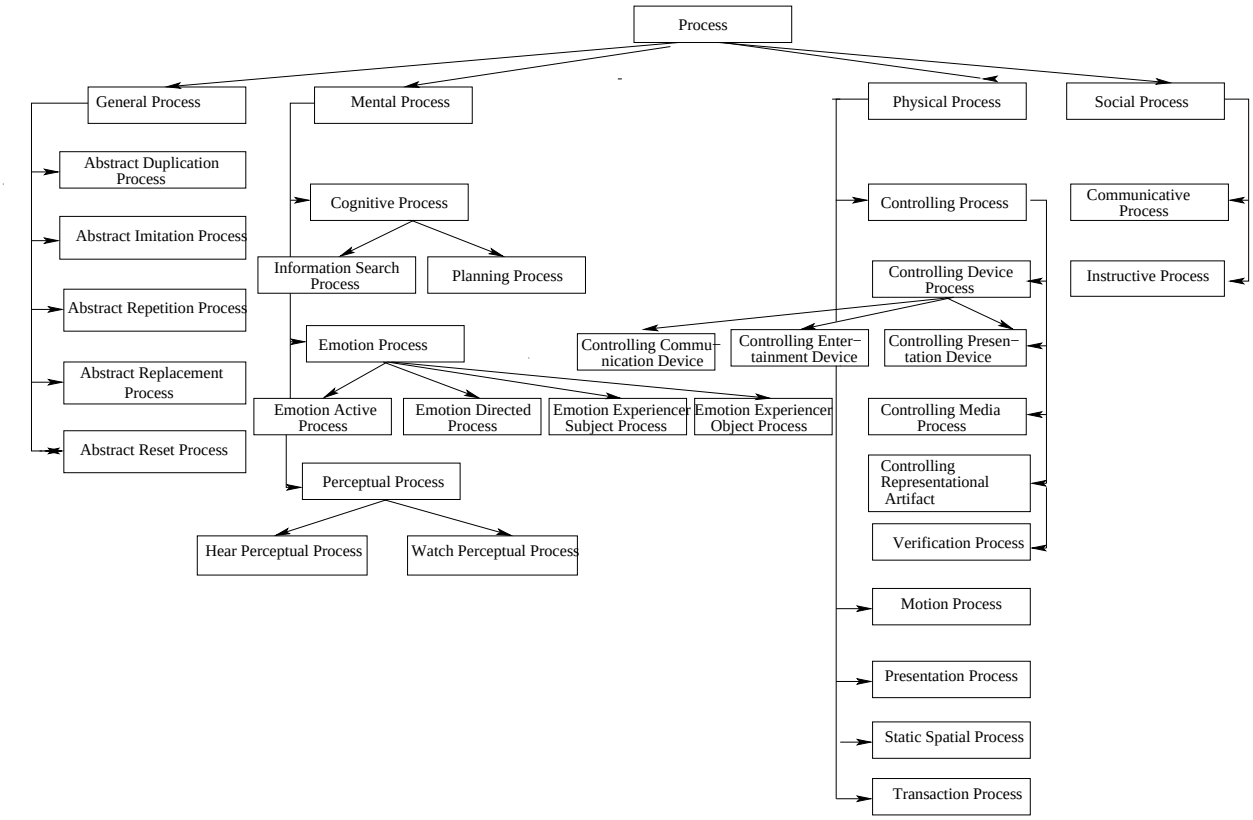

Figure 1: Upper part of the process hierarchy.

4.2 Mapping of CR to Graphs

ONTOSCORE converts the domain model, i.e. an ontology, into a directed graph with concepts as nodes and relations as edges. One additional problem that needed to be solved lies in the fact that the directed *subclass-of* relations enable path algorithms to ascend the class hierarchy upwards, but do not let them descend, therefore missing a significant set of possible paths. In order to remedy that situation the graph was enriched during its conversion by corresponding *parent-of* relations, which eliminated the directionality problems as well as avoids cycles and 0-paths. In order to find the shortest path between two concepts, ONTOSCORE employs the *single source shortest path* algorithm of Dijkstra (Cormen et al., 1990).

Given a concept representation CR $\{c_1, ..., c_n\}$, the algorithm runs once for each concept. The Dijkstra algorithm calculates minimal paths from a source node to all other nodes. Then, 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.

4.3 The Scoring Algorithm

To score the minimal paths connecting all concepts with each other in a given CR, we first 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). To construct the graph the dictionary lemmata were represented as nodes in an *isa* hierarchy and their semantic relations were represented as edges, which were extracted automatically from the 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. For each two concepts c_i, c_j the set $P = \{p_1, p_2, \ldots, p_m\}$ denotes the scores of all possible paths that link the two concepts. The score for path $k(k = 1, \ldots, m)$ can be given as:

$$p_k = \sum_{i=1}^{n} a_i w_i$$

where a_i represents the number of times the relation r_i exists in path k. The ensuing distance between two concepts c_i and c_j is, then, defined as the minimum score

derived between c_i and c_j, i.e.:

$$D(c_i, c_j) = min(p_k) \quad k = 1, 2, \ldots, m$$

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.

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 semantic coherence scores of arbitrary CRs on an absolute scale, certain extensions are 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.

4.4 Scoring Concept Representations

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 have no connecting path, we define the distance between two concepts c_i and c_j that are only connected via *isa* relations 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. In the ontology currently employed by the system some reverse relations can be found, e.g. given c_1=`Broadcast` and c_2=`Channel`, there exists a path from c_1 to c_2 via the relation **has-channel** and a different path from c_2 to c_1 via the relation **has-broadcast**. However, such reverse relations are only sporadically represented in the ontology. Consequently, it is difficult to account for their influence on $S(CR)$ in general. That is why 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.

4.5 Word/Concept Relation

Given the algorithm proposed above, a significant number of misclassifications for SRHs would result from the cases when an SRH contains a high proportion of function words (having no conceptual mappings in the resulting CR) and only a few content words. Let's consider the following example:

(6) *Wo den Informationen zu das gleiche*
Where the information to the same

The corresponding CR is constituted out of a single concept `Information Search Process`. ON TOSCORE would classify the CR as *coherent* with the highest possible score, as this is the only concept in the set. This, however, would often lead to misclassifications. We, therefore, included a post-processing technique that takes the relation between the number of ontology concepts N_c in a given CR and the total number of words N_w in the original SRH into account. This relation is defined by the ratio $V = N_c/N_w$. ONTOSCORE automatically classifies an SRH as being incoherent irrespective of its semantic coherence score, if V is less then the threshold set. The threshold may be set freely. The corresponding findings are presented in the evaluation section.

4.6 ONTOSCORE at Work

Looking at an example of ONTOSCORE at work, we will examine the utterance given in Example (1). The resulting two SRHs - SRH_1 and SRH_2 - are given in Example (1a) and (1b) respectively. The human annotators considered SRH_1 to be coherent and labeled SRH_2 as incoherent. According to the concept entries in the lexicon, the SRHs are transformed into two alternative

concept representations. As no ambiguous words are found in this example, CR_1 corresponds to SRH_1 and CR_2 corresponds to SRH_2:

CR_1: {Person; Map; Watch Perceptual Process};

CR_2: {Person; Map; Parting Process}.

They are converted into a graph. According to the algorithm shown in Section 4.3, all paths between the concepts of each graph are calculated and weighted. This yields the following non-D_{max} paths:

CR_1 $D(\texttt{WatchPerceptualProcess}, \texttt{Person}) = 1$
via the relation **has-watcher**;
$D(\texttt{WatchPerceptualProcess}, \texttt{Map}) = 1$
via the relation **has-watchable_object**.

CR_2 $D(\texttt{PartingProcess}, \texttt{Person}) = 1$
via the relation **has-agent**;

The ensuing results are:

According to S	According to S'
$S(CR_1) = 7$	$S'(CR_1) = 4$
$S(CR_2) = 8.5$	$S'(CR_2) = 7$

In both cases the results are sufficient for a relative judgment, i.e. SRH_2 constitutes a less semantically coherent structure as SRH_1. To allow for a binary classification into *semantically coherent* vs. *incoherent* samples, a cut-off threshold must be set. The results of the corresponding experiments will be presented in Section 5.2.

4.7 Word Sense Disambiguation

Due to lexical ambiguity, the process of transforming an n-best list of SRH to concept representations often results in a set of CRs that is greater than 1, i.e. a given SRH could be transformed into a set of CRs $\{CR_1, ..., CR_n\}$. Word sense disambiguation could, therefore, also independently be performed using the semantic coherence scoring described herein as an additional application of our approach. However, that has not been investigated thoroughly yet.

For example, lexicon entries for the words:

```
I - Person
am - Static Spatial Process,
     Self Identification Process, None
on - Two Point Relation, None
the - None
Philosopher's Walk - Location
```

yield a set of interpretations for an SRH such as:

(7) *Ich bin auf dem Philosophenweg*
I am on the Philosopher's Walk

CR_1 {Person, Static Spatial Process, Location}
CR_2 {Person, Static Spatial Process, Two Point Relation, Location}
CR_3 {Person, Self Identification Process, Location}
CR_4 {Person, Self Identification Process, Two Point Relation, Location}
CR_5 {Person, Two Point Relation, Location}
CR_6 {Person, Location}

and corresponding final scores:

$S(CR_1) = 6$;
$S(CR_2) = 5.23$;
$S(CR_3) = 7.45$;
$S(CR_4) = 7$;
$S(CR_5) = 5.5$;
$S(CR_6) = 10 = D_{max}$;

The examination of the resulting scores allows us to conclude that CR_2 constitutes the most semantically coherent representation of the initial SRH, CR_1 and CR_5 display a slightly lesser degree of semantic coherence, whereas CR_3, CR_4 and CR_6 are much less coherent and may, thus, be considered inadequate.

5 Evaluation

5.1 Context

The ONTOSCORE software runs as a module in SMARTKOM (Wahlster et al., 2001), a multi-modal and multi-domain spoken dialogue system. 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, 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 SRH and thus provides an important aid to the understanding component of the system in determining the best SRH. The ONTOSCORE software employs two knowledge sources, an ontology (about 730 concepts and 200 relations) and a word/concept lexicon (ca. 3.600 words), covering the respective domains of the system.

5.2 Results

The evaluation of ONTOSCORE was carried out on a dataset of 2.284 SRHs. We reformulated the problem of measuring the semantic coherence in terms of classifying the SRHs into two classes: *coherent* and *incoherent*. To our knowledge, there exists no similar software performing semantic coherence scoring to be used for com-

parison in this evaluation. Therefore, we decided to use the results from human annotation (s. Section 2.2) as the baseline.

A *gold standard* for the evaluation of ONTOSCORE was derived by the annotators agreeing on the correct solution in cases of disagreement. This way, we obtained 1.246 (54.55%) SRH classified as coherent by humans, which is also assumed to be the baseline for this evaluation.

Additionally, we performed an inverse linear transformation of the scores (which range from 1 to D_{max}), so that the output produced by ONTOSCORE is a score on the scale from 0 to 1, where higher scores indicate greater coherence. In order to obtain a binary classification of SRHs into coherent *versus* incoherent with respect to the knowledge base, we set a cutoff thresh old. The dependency graph of the threshold value and the results of the program in % is shown in Figure 1.

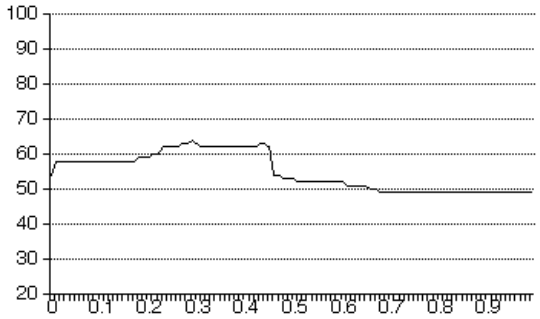

Figure 2: Finding the optimal threshold for the coherent *versus* incoherent classification

The best results are achieved with the threshold 0.29. With this threshold, ONTOSCORE correctly classifies 1.487 SRH, i.e. 65.11% in the evaluation dataset (the word/concept relation is not taken into account at this point).

Figure 3 shows the dependency graph between V, representing the threshold for the word/concept relation and the results of ONTOSCORE, given the best cutoff threshold for the classification (i.e. 0.29) derived in the previous experiments.

The best results are achieved with the $V = 0.33$. In other words, the proportion of concepts vs. words must be no less than 1 to 3. Under these settings, ONTOSCORE correctly classifies 1.672 SRH, i.e. 73.2% in the evaluation dataset. This way, the technique brings an additional improvement of 8.09% as compared to initial results.

6 Concluding Remarks

The ONTOSCORE system described herein automatically performs ontology-based scoring of sets of concepts con-

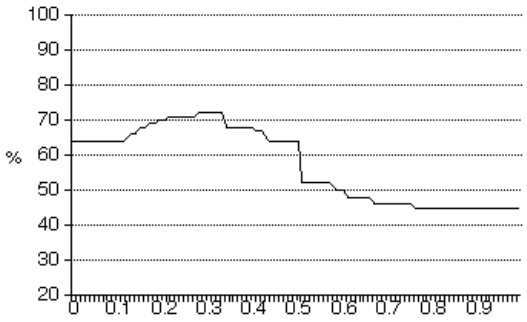

Figure 3: Finding the optimal threshold for the word/concept relation

stituting an adequate representation of speech recognition hypotheses. To date, the algorithm has been implemented in a software which is employed by a multi-domain and multi-modal 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. For this task, it provides an alternative knowledge-based score next to the ones provided by the ASR and the NLU system. 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 future work we will examine how the computation of a discourse dependent semantic coherence score, i.e. how well a given SRH fits within domain model with respect to the previous discourse, can improve the overall score. Additionally, we intend to calculate the semantic coherence score with respect to individual domains of the system, thus enabling domain recognition and domain change detection in complex multi-modal and multi-domain spoken dialogue systems. Currently, we are also beginning to investigate 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.

Acknowledgments

This work has been partially funded by the German Federal Ministry of Research and Technology (BMBF) as part of the SmartKom project under Grant 01 IL 905C/0 and by the Klaus Tschira Foundation. We would like to thank Michael Strube for his helpful comments on the previous versions of this paper.

References

Jan Alexandersson and Tilman Becker. 2003. The Formal Foundations Underlying Overlay. In *Proceedings of the Fifth International Workshop on Computational Semantics (IWCS-5)*, Tilburg, The Netherlands, February.

James F. Allen, George 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.

R.V. Cox, C.A. Kamm, L.R. Rabiner, J. Schroeter, and J.G. Wilpon. 2000. Speech and language processing for next-millenium communications services. *Proceedings of the IEEE*, 88(8):1314–1334.

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.

Diane J. Litman, Marilyn A. Walker, and Michael S. Kearns. 1999. Automatic detection of poor speech recognition at the dialogue level. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, College Park, Md., 20–26 June 1999, pages 309–316.

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.

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

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

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

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