Using Knowledge-Based Scores for Identifying Best Speech Recognition Hypothesis

Iryna Gurevych, Robert Porzel

European Media Lab GmbH
Villa Bosch, Schloss-Wolfsbrunnenweg 33
69118 Heidelberg, Germany
{iryna.gurevych, robert.porzel}@eml.villa-bosch.de

Abstract

The paper presents the evaluation of a knowledge-based scoring method applied to the problem of identifying the best speech recognition hypothesis (SRH) in a functioning multimodal dialogue system. The competing SRHs are evaluated in terms of their semantic coherence using the high-level domain knowledge encoded in the ontology. We conducted an annotation experiment and showed that humans can reliably select the best SRH in a given N-best list (agreement 95.35%). The knowledge-based method identifies correctly 88.07% of the best SRHs (given the baseline 63.91%), which is also an improvement over the automatic speech recognizer (ASR) (83.88% accuracy).

1. Introduction

One of the major challenges in making spoken dialogue system reliable enough to be deployed in more complex real world applications is recognizing the user's input correctly. In many cases both correct and incorrect representations of the user's utterances are contained in the automatic speech recognizer's N-best lists. Facing multiple representations of a single utterance poses the question which of the different hypotheses corresponds most likely to the user's utterance.

The paper is structured as follows: we present the challenge in Section 2. Section 3 introduces different approaches used to score the output of the speech recognizer. In Section 4 we describe the data and annotations underlying our experiments. Section 5 contains a description of the system employed to select the best speech recognition hypothesis, together with a working example and the integration of knowledge-based scores into a dialogue system. Finally, the results of the evaluation and some conclusions are given in Section 6.

2. Challenge

Determining the best speech recognition hypothesis in an N-best list means that the hypothesis which captures the user's intention best should be selected for further processing. The effect of selecting the non-best SRH is that an overall performance of the dialogue system decreases [1].

Below, we highlight a few examples where the inclusion of a knowledge-based score results in the improvement of the overall system's performance. In many cases, we observe that ASR and parser scores provide contradictory or incorrect information. Example 1 resulted in two alternative SRHs 1a, 1b:[1]

(1) *Erzäle mir mehr zum Schloss*
Tell me more about the castle

(1a) *Erzäle mehr zum Schloss*
Tell me about the castle

(1b) *Bitte Filme zum Schloss*
Please films about the castle

SRH	recognizer	parser	domain knowledge
1a	.36	1	.51
1b	1	.25	.42

(1)

In this case, the ASR scores are higher for the example 1b, while the parsing and knowledge-based scores favour the example 1a. In example 2, we observe the opposite situation: SRH 2a received a better score from the parser, while SRH 2b is scored higher by the ASR and the knowledge-based component.

(2) *Ich will mich ein bisschen in Heidelberg umschauen*
I want me a bit in Heidelberg look around

(2a) *Kirchen in Heidelberg umschauen*
Churches in Heidelberg look around

(2b) *ja ich ist in Heidelberg umschauen*
Yes I is in Heidelberg look around

SRH	recognizer	parser	domain knowledge
2a	.53	.25	.4
2b	.55	.08	.52

(2)

In some cases, both recognizer and parser assign a lower score to the best SRH. This results in misrecognizing the user's intention, which could be avoided by taking the system's domain knowledge into account. A typical example is:

(3) *Ich befinde mich am Philosophenweg*
I am on the Philosopher's Walk

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

(3a) *Spielfilme* *am* *Philosophenweg*
Movies on the Philosopher's Walk

(3b) *befinde* *am* *Philosophenweg*
Am on the Philosopher's Walk

SRH	recognizer	parser	domain knowledge
3a	.55	.33	.45
3b	.53	.17	.72

(3)

3. Scoring Approaches

Different methods were proposed in the literature to select the best speech recognition hypothesis in the N-best lists. Traditionally, the scores provided by the recognition system itself are used. Most typically, these are acoustic and language model features. Sometimes, also the number of words and phones in the hypothesis [2] and the number of hypotheses in an N-best list are considered. More recently, also linguistically motivated features have been used, e.g., scores provided by the parsing system [3]. Such features are based on the quality of the syntactic and/or semantic parse obtained by parsing a hypothesis into a certain representation, e.g., a semantic frame. This reflects how well the hypothesis is covered by the grammar employed in the system.

[4] employ a whole range of natural language processing and discourse features to identify understanding errors in a spoken dialogue system. They attempt to estimate the influence and contribution of individual features to the overall task by using machine learning algorithms inducing an automatic classification model. This is motivated by the previous work on identifying poor speech recognition on the dialogue level by [5]. [6] employ semantic features for scoring alternative SRHs. They measure the amount of information contained in a given utterance (semantic weight) and the difference in semantic content between two hypotheses in an N-best list (semantic distance).

In [7], an algorithm for scoring the semantic coherence of sets of concepts using an ontology is introduced (see Section 5). The algorithm is evaluated on the task of classifying SRHs in semantically coherent versus incoherent. The aim of the work presented here was to investigate the effect of using the scores relying on the high-level domain knowledge for the task of determining the best SRH in a functioning dialogue system. The performance of individual features is compared to a gold standard derived from the corresponding annotation experiments with humans.

4. Data and Annotations

The experiments are based on the data collected in hidden-operator tests. We employed an approach described in [8]. This experimental setup is slightly different from the traditional Wizard-of-Oz data collection approach. The intentions to be expressed in utterances are suggested to subjects on a per utterance basis. This means that the subjects have to stick to a pre-defined scenario for each dialogue. This way, the collected utterances are kept within the system's coverage.

In this trial, 95 dialogues consisting of 552 audio files containing single user's utterances were used. The utterances were transcribed by humans. The audio files were sent to the dialogue system. The behaviors of the recognizer, the parser and the knowledge-based scoring components were logged automatically. As a result, we obtained an N-best list for each utterance as well as confidence scores from each component. By interpreting the scores, the best SRH identified by each individual component was determined. The overall number of SRHs in the N-best lists was 1,375.

For this study an expression to be annotated, is a set of SRHs (N-best list) related to a single user's utterance. The annotators, 2 computational linguistics students, were trained for the task with the help of an annotation manual. For each utterance a single SRH had to be labeled as the best one. The guidelines for selecting the best SRH were:

- How does the respective SRH capture the intention contained in the transcribed user's utterance?
- Several SRHs capture the intention equally well, the actual word error rate had to be considered.

The Kappa statistic [9] is often applied to measure the degree of reliability of human annotations. It could not be employed in this case as the number of SRHs underlying the best SRH selection per utterance is varying. Therefore, we calculated the percentage of utterances where the annotators agreed on the correct solution, resulting in 95.35% of inter-annotator agreement. This number suggests a rather high degree of reliability for identifying the best SRH by humans. The gold standard for the evaluation was derived by the annotators agreeing on the correct solution in the cases of disagreement.

Closer examination of cases of disagreement revealed some typical problems in selecting a single SRH as the best one.

- As a result of misrecognition, all SRHs in the N-best list are semantically incoherent and wrong. Consider the following example:

  (4) *Ich* *möchte* *gern* *zur* *Peterskirche*
  I would like to the Peter's Church

  (4a) *Ich* *hätte* *gerne* *mehr* *über* *die* *Ritters*
  I would like more about the Ritters
  die *Bio* *ist*
  the Bio is

  (4b) *Ich* *hätte* *gerne* *mehr* *über* *die* *Ritters*
  I would like more about the Ritters
  die *Lust*
  the delight

- As a result of partial misrecognition, there are several SRHs, e.g., 5a and 5b, which are semantically coherent, but are wrong with respect to the initial user's utterance.

  (5) *Ich* *möchte* *ein* *Video* *sehen*
  I would like a video see

  (5a) *Ich* *möchte* *mehr* *Tiere* *sehen*
  I would like more animals see

  (5b) *Ich* *möchte* *mehr* *News* *sehen*
  I would like more news see

- The output of the speech recognizer contains several competing SRHs corresponding to the same semantic representation. If this representation is equally coherent,

ambiguity as to which of the hypotheses is to be labeled as the best one arises. As will be reported in Section 6, the same problem is observed in the data generated by the knowledge-based method. This means that multiple SRHs are scored equally, resulting in the best SRH contained in the set, but not being uniquely identifiable.

(6) *Gibt es mehr Information zur*
Is there more information about the
Heiliggeistkirche
Holy Spirit Church

(6a) *Gibt mehr Information zur*
Is more information about the
Heiliggeistkirche
Holy Spirit Church

(6b) *Gibt es Information zur*
Is there information about the
Heiliggeistkirche
Holy Spirit Church

5. Knowledge-Based Scoring Method

In this section, we provide a description of the algorithm underlying the knowledge-based scoring of SRHs and its integration into the overall dialogue system. The respective system architecture is shown in Figure 1. The parser picks an N-best list of hypotheses out of the speech recognizer's word lattice [10]. The N-best list is handed over to the knowledge-based system, which provides an additional score. The system employes two knowledge sources, an ontology of the system's domains and a lexicon.

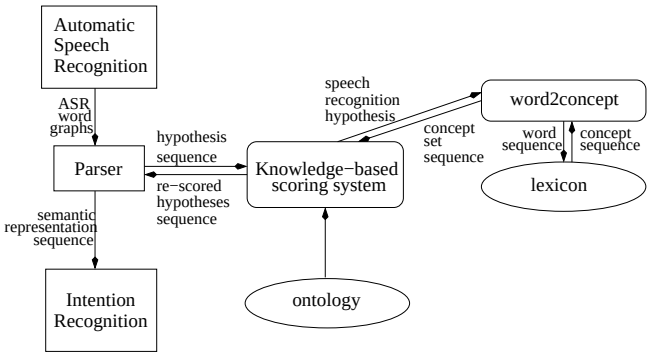

Figure 1: The subsystem responsible for the best SRH selection

The ontology employed by the knowledge-based system was built as a general knowledge representation for various processing modules within the system.[2] It consists of ca. 730 concepts representing domain and discourse entities and 200 slots representing the relations existing between them. A detailed description of the knowledge engineering approach taken to design the ontology can be found in [11]. The lexicon associated with the ontology contains 3,600 full forms. The meanings of individual lexical items are stored in terms of ontology concepts. This means that each lexicon entry is augmented with zero, one or many corresponding meanings.

The system employs ontological knowledge as the basis for evaluating semantic coherence of sets of concepts representing competing SRHs. Semantic coherence is defined as a measure reflecting how well sets of concepts fit into existing knowledge representation. The specific knowledge base, e.g. written in DAML+OIL or OWL,[3] 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 system operates on the nodes and named edges of the directed graph represented by the ontology.

In the pre-processing step, each SRH is converted into a *concept representation* (CR). This is represented by a simple vector of concepts, corresponding to the words in the SRH for which entries in the lexicon exist. All other words with empty concept mappings, e.g., articles, are ignored. 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. The *single source shortest path* algorithm of Dijkstra [12] is employed to find the minimal paths connecting a given concept c_i with every other concept in CR (excluding c_i itself). This results in an $n \times n$ matrix of the respective paths.

The semantic coherence score for CR is calculated based on the average path length between all concept pairs in CR:

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

where $D(c_i, c_j)$ is the weight of the minimum path (distance) between a pair of concepts c_i and c_j. The concept representation which has been assigned the highest score is considered to be the best.

The knowledge-based scoring system has been enhanced further to take the discourse context into account [13]. The basic idea is to consider the conceptual context of the previous utterance. Conceptual context representation for SRH_{n+1} to be scored is produced by building 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 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' is then scored according to the equation S(CR) given earlier, yielding a contextually enhanced knowledge-based score.

For example, in our data a user expressed the wish to get from Cologne to Heidelberg and then to continue his visit in Heidelberg. We will examine this discourse fragment composed of the two sequential utterances given in examples 7 and 8.

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

[2] 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.

[3] 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.

$I(SRH_{n+1})$			$I'(SRH_{n+1})$
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: Generating conceptual context representations

(8) *wie komme ich in Heidelberg weiter.*
How can I in Heidelberg continue.

Looking at the SRHs from the N-best list of example 7 we found that SRH 7a constituted the best representation of the utterance, whereas all others constituted a less adequate representation.

(7a) *ich möchte auf schnellsten Weg von Köln nach Heidelberg.*
I want on fastest way from Cologne to Heidelberg.

(7b) *ich möchte auf schnellsten Weg Köln nach Heidelberg.*
I want on fastest way Cologne to Heidelberg.

(7c) *ich möchte Folk Weg von Köln nach Heidelberg.*
I want folk way from Cologne to Heidelberg.

(7d) *ich möchte auf schnellsten Weg vor Köln nach Heidelberg.*
I want on fastest way before Cologne to Heidelberg.

(7e) *ich möchte vor schnellsten Weg von Köln nach Heidelberg.*
I want before fastest way from Cologne to Heidelberg.

As shown in Table 2, in the case of example 7 all scoring methods identify the SRH 7a as the best one.

SRH	recognizer	parser	domain knowledge
7a	1	1	.6
7b	.74	.94	.6
7c	.63	.94	.54
7d	.78	.89	.54
7e	.74	.88	.54

Table 2: The scores for the SRHs of example 7

Example 8 yields the following SRHs with the corresponding context-independent CRs and context-dependent CR's:

(8a) *Rennen Lied Comedy Show Heidelberg weiter.*
Race song comedy show Heidelberg continue.
CR\{MusicPiece, Genre, Genre, Town\}
CR'\{MusicPiece, Genre, Genre, Town, EmotionExperiencerSubjectProcess,Person, TwoPointRelation, Route \}

(8b) *denn wie Comedy Heidelberg weiter.*
then how comedy Heidelberg continue.
CR\{Genre, Town\}
CR'\{Genre, Town, EmotionExperiencerSubjectProcess,Person, TwoPointRelation, Route \}

(8c) *denn wie Comedy Show weiter.*
then how comedy show continue.
CR\{Genre, Genre\}
CR'\{Genre, Genre, EmotionExperiencerSubjectProcess, Person, TwoPointRelation, Route \}

(8d) *denn wie Comedy weiter.*
then how comedy continue.
CR\{Genre\}
CR'\{Genre, EmotionExperiencerSubjectProcess, Person, TwoPointRelation, Route \}

(8e) *denn wie komme ich in Heidelberg weiter.*
then how can I in Heidelberg 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 8:

SRH	recognizer	parser	domain knowledge
8a	1	.25	.32
8b	.52	.2	.48
8c	.34	.2	.39
8d	.35	.12	0
8e	.52	.08	.71

Table 3: The scores for the SRHs of example 8

As it is evident from Table 3, CR'_{best} corresponds to example 8e. This means that 8e constitutes a more contextually coherent concept structure than the alternative SRHs. This utterance was also labeled as the best SRH by the annotators.

6. Results

The best SRH selection of the knowledge-based system has been compared with the human gold standard. Adding the individual ratios of utterance/SRHs - corresponding to the likelihood of guessing the best SRH - and dividing it by the number of utterances in the corpus yielded the baseline of 63.91% for this evaluation. The performance of the respective individual components on this task was as follows (see Table 4): The ASR scores identified 83.88% of the best hypothesis correctly

as compared to a success rate of 87.56% of the parsing and 88.07% of the contextually enhanced knowledge-based system.

All components exceed the baseline performance. The parsing as well as the knowledge-based scores, however, yield a significant improvement over the ASR scores. This indicates their use for improving the speech recognizer's performance. We ran an additional experiment (voting) to examine how the three methods perform in combination with each other. We achieve 89.13% accuracy in identifying the best SRH if it is selected based on the majority decision. The improvement over the performance of individual components does not seem high. On the other hand, it may be considered interesting given the rather high overall figures.

Score producer	Accuracy
Baseline	63.91%
Recognizer	83.88%
Parser	87.56%
Domain knowledge	88.07%
Voting	89.13%

(4)

Our results indicate that the inclusion of knowledge-based features into spoken dialogue systems has the potential to reduce the number of misrecognitions of the user's intentions. This can improve the overall system performance by correctly identifying a higher percentage of the best SRHs in the recognizer's output.

In future work, we intend to investigate the use of knowledge-based scores in combination with other scoring methods, to automatically induce a best SRH classification model. We also intend to investigate how the knowledge-based method performs in identifying erroneous SRHs (Cf. [5]). This is motivated by our finding, as shown in the example 4, that not only identifying the best SRH is important, but also identifying whether the best SRH is correct. Also, a semantically coherent SRH is not always correct with respect to the actual user's utterance (Cf. example 5). That is why it is important to identify the best SRH as well as the correctness of this SRH to detect the speech recognition errors, i.e., problematic turns in the dialogue.

7. 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.

8. References

[1] M. A. Walker, C. A. Kamm, and D. J. Litman, "Towards developing general model of usability with PARADISE," *Natural Language Engineering*, vol. 6, 2000.

[2] S. Kamppari, "Word and phone level acoustic confidence scoring for speech understanding systems," 1999, master's thesis, MIT.

[3] R. Engel, "SPIN: Language understanding for spoken dialogue systems using a production system approach," in *Proceedings of ICSLP 2002*, 2002.

[4] M. A. Walker, J. Wright, and I. Langkilde, "Using natural language processing and discourse features to identify understanding errors in a spoken dialogue system," in *Proceedings of ICML*, Palo Alto, CA, 2000.

[5] D. J. Litman, M. A. Walker, and M. J. Kearns, "Automatic detection of poor speech recognition at the dialogue level," in *Proceedings of the 37th ACL*, 1999, pp. 309–316.

[6] C. Pao, P. Schmid, and J. Glass, "Confidence scoring for speech understanding systems," in *Proceedings of ICSLP*, Sydney, Australia, 1998.

[7] I. Gurevych, R. Malaka, R. Porzel, and H.-P. Zorn, "Semantic coherence scoring using an ontology," in *Proceedings of the HLT-NAACL Conference*, 2003.

[8] S. Rapp and M. Strube, "An iterative data collection approach for multimodal dialogue systems," in *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, 2002.

[9] J. Carletta, "Assessing agreement on classification tasks: The kappa statistic," *Computational Linguistics*, vol. 22, no. 2, pp. 249–254, 1996.

[10] M. Oerder and H. Ney, "Word graphs: An efficient interface between continuous-speech recognition and language understanding," in *ICASSP Volume 2*, 1993, pp. 119–122.

[11] I. Gurevych, R. Porzel, E. Slinko, N. Pfleger, J. Alexandersson, and S. Merten, "Less is more: Using a single knowledge representation in dialogue systems," in *Proceedings of the HLT-NAACL'03 Workshop on Text Meaning*, Edmonton, Canada, 2003.

[12] T. H. Cormen, C. E. Leiserson, and R. R. Rivest, *Introduction to Algorithms*. Cambridge, MA: MIT press, 1990.

[13] R. Porzel, I. Gurevych, and C. Müller, "Ontology-based contextual coherence scoring," in *Proceedings of the Fourth SIGdial Workshop on Discourse and Dialogue*, Sapporo, Japan, July 2003, to appear.