

Annotating Semantic Consistency of Speech Recognition Hypotheses

Iryna Gurevych, Robert Porzel and Michael Strube

European Media Laboratory GmbH

Schloß-Wolfsbrunnenweg 33

D-69118 Heidelberg, Germany

e-mail: {gurevych,porzel,strube}@eml.villa-bosch.de

Abstract

Recent work on natural language processing systems is aimed at more conversational, context-adaptive systems in multiple domains. An important requirement for such a system is the automatic detection of the domain and a domain consistency check of the given speech recognition hypotheses. We report a pilot study addressing these tasks, the underlying data collection and investigate the feasibility of annotating the data reliably by human annotators.

1 Introduction

The complete understanding of naturally occurring discourse is still an unsolved task in computational linguistics. Several large research efforts are underway to build multi-domain and multimodal information systems, e.g. the DARPA Communicator Program¹, the SmartKom research framework² (Wahlster et al., 2001), the AT&T interactive speech and multimodal user interface program³.

Dialogue systems which deal with complex dialogues require the interaction of multiple knowledge sources, e.g. domain, discourse and user model (Flycht-Eriksson, 1999). Furthermore NLP systems have to adapt to different environments and applications. This can only be achieved if the system is able to determine how well a given speech recognition hypothesis (SRH) fits within the respective domain model and what domain should be considered by the system currently in focus.

The purpose of this paper is to develop an annotation scheme for annotating a corpus of SRH with information on semantic consistency and domain specificity. We investigate

the feasibility of an automatic solution by first looking at how reliably human annotators can solve the task.

The structure of the paper is as follows: Section 2 gives an overview of the domain modeling component in the SmartKom system. In Section 3 we report on the data collection underlying our study. A description of the suggested annotation scheme is given in Section 4. Section 5 presents the results of an experiment in which the reliability of human annotations is investigated.

2 Domain Modeling in SmartKom

The SmartKom research project (a consortium of twelve academic and industrial partners) aims at developing a multi-modal and multi-domain information system. Domains include cinema information, home electronic device control, etc. A central goal is the development of new computational methods for disambiguating different modalities on semantic and pragmatic levels.

The information flow in SmartKom is organized as follows: On the input side the parser picks an N-best list of hypotheses out of the speech recognizer's word lattice (Oerder and Ney, 1993). This list is sent to the media fusion component and then handed over to the intention recognition component.

The main task of intention recognition in SmartKom is to select the best hypothesis from the N-best list produced by the parser. This is then sent to the dialogue management component for computing an appropriate action. In order to find the best hypothesis, the intention recognition module consults a number of other components involved in language, discourse and domain analysis and requests confidence scores to make an appropriate decision (s. Fig. 1).

Tasks of the domain modeling component are:

¹<http://fofoca.mitre.org>

²<http://www.smartkom.com>

³<http://www.research.att.com/news/2002/January/ISMUI.html>

- to supply a confidence score on the consistency of SRH with respect to the domain model;
- to detect the domain currently in focus.

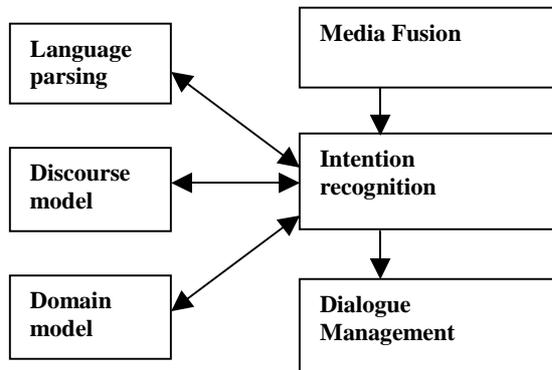


Figure 1. Information flow

These tasks are inherently related to each other: It is possible to assign SRH to certain domains only if they are consistent with the domain model. On the other hand, a consistency score can only be useful when it is given with respect to certain domains.

3 Data

We consider semantic consistency scoring and domain detection a classification task. The question is whether it is feasible to solve this task automatically. As a first step towards an answer we reformulate the problem: automatic classification of SRH is possible only if humans are able to do that reliably.

3.1 Data Collection

In order to test the reliability of such annotations we collected a corpus of SRH. The data collection was conducted by means of a hidden operator test (Rapp and Strube, 2002). In the test the SmartKom system was simulated. We had 29 subjects prompted to say certain inputs in 8 dialogues. 1479 turns were recorded. Each user-turn in the dialogue corresponded to a single intention, e.g. route request or sights information request.

3.2 Data Preprocessing

The data obtained from the hidden operator tests had to be prepared for our study to compose a corpus with N-best SRH. For this pur-

pose we sent the audio files to the speech recognizer. The input for the domain modeling component, i.e. N-best lists of SRH were recorded in log-files and then processed with a couple of Perl scripts. The final corpus consisted of ca. 2300 SRH. This corresponds to ca. 1.55 speech recognition hypotheses per user's turn.

The SRH corpus was then transformed into a set of annotation files which could be read into MMAX, the annotation tool adopted for this task (Mueller and Strube, 2001).

4 Annotation Scheme

For our study, a markable, i.e. an expression to be annotated, is a single SRH. The annotators as well as the domain modeling component in SmartKom currently do not take the dialogue context into account and do not perform context-dependent analysis. Hence, we presented the markables completely out of dialogue order and thus prevented the annotators from interpreting SRH context-dependently.

4.1 Semantic Consistency

In the first step, the annotators had to classify markables with respect to semantic consistency. Semantic consistency is defined as well-formedness of an SRH on an abstract semantic level. We differentiate three classes of semantic consistency: *consistent*, *semi-consistent*, or *inconsistent*. First, all nouns and verbs contained in the hypothesis are extracted and corresponding concepts are retrieved from a lemma-concept dictionary (lexicon) supplied for the annotators. The decision regarding consistency, semi-consistency and inconsistency has to be done on the basis of evaluating the set of concepts corresponding to the individual hypothesis.

- *Consistent* means that all concepts are semantically related to each other, e.g. "ich moechte die kuerzeste Route"⁴ is mapped to the concepts "self", "wish", "route" all of which are related to each other. Therefore the hypothesis is considered consistent.
- The label *semi-consistent* is used if at least a fragment of the hypothesis is

⁴ I'd like the shortest route.

meaningful. For example, the hypothesis "ich moechte das Video sind"⁵ is considered semi-consistent as the fragment "ich moechte das Video", i.e. a set of corresponding concepts "self", "want", "video" is semantically well-formed.

- *Inconsistent* hypotheses are those whose conceptual mappings are not semantically related within the domain model. E.g. "ich wuerde die Karte ja Wiedersehen"⁶ is conceptualized as "self", "map", "parting". This set of concepts does not semantically make sense and the hypothesis should be rejected.

4.2 Domain Detection

One of our considerations was that it is principally not always feasible to detect domains from an SRH. This is because the output of speech recognition is often corrupt, which may, in many cases, lead to false domain assignments. We argue that domain detection is dependent on the semantic consistency score. Therefore, according to our annotation scheme no domain analysis should be given to the semantically inconsistent SRH.

If the hypothesis is considered either consistent or semi-consistent, certain domains will be assigned to it. The list of SmartKom domains for this study is finite and includes the following: route planning, sights information, cinema information, electronic program guide, home electronic device control, personal assistance, interaction management, small-talk and off-talk.

In some cases multiple domains can be assigned to a single markable. The reason is that some domains are inherently so close to each other, e.g. cinema information and electronic program guide, that the distinction can only be made when the context is taken into account. As this is not the case for our study we allow for the specification of multiple domains per SRH.

⁵ I'd like the video are.

⁶ I would the map yes good-bye.

5 Reliability of Annotations

5.1 The Kappa Statistic

To measure the reliability of annotations we used the Kappa statistic (Carletta, 1996).

The value of Kappa statistic (K) for semantic consistency in our experiment was 0.58, which shows that there was not a high level of agreement between annotators⁷. In the field of content analysis, where the Kappa statistic originated, $K > 0.8$ is usually taken to indicate good reliability, $0.68 < K < 0.8$ allows to draw tentative conclusions.

The distribution of semantic consistency classes and domain assignments is given in Fig. 2.

Type	%
Consistent	51
Semi-consistent	10,3
Inconsistent	38,7

Domain	%
Route planning	33,1
Sights info	13,3
Cinema info	10,8
Electr. Program guide	15,9
Home device control	12,0
Personal assistance	1,1
Interaction Management	13,1
Other	0,7

Figure 2. Distribution of Classes

5.2 Discussion of the results

One reason for the relatively low coefficient of agreement between annotators could be a small number of annotators (two) as compared to rather fine distinction between the classes *inconsistent* vs. *semi-consistent* and *semi-consistent* vs. *consistent* respectively.

Another reason arises from the analysis of disagreements among annotators. We find many annotation errors caused by the fact that the annotators were not able to interpret the conceptualized SRH correctly. In spite of the fact that we emphasized the necessity of care-

⁷ Results on the reliability of domain assignments are not the subject of the present paper and will be published elsewhere.

ful examination for high-quality annotations, the annotators tended to take functional words like prepositions into account. According to our annotation scheme, however, they had to be ignored during the analysis.

5.3 Revisions to the annotation scheme

As already noted, one possible reason for disagreements among annotators is a rather fine distinction between the classes *inconsistent* vs. *semi-consistent* and *semi-consistent* vs. *consistent*. We had difficulties in defining strict criteria for separating *semi-consistent* as a class on its own. The percentage of its use is rather low as compared to the other two and amounts to 10.3% on average.

A possible solution to this problem might be to merge the class *semi-consistent* with either *consistent* or *inconsistent*. We conducted a corresponding experiment with the available annotations.

In the first case we merged the classes *inconsistent* and *semi-consistent*. We then ran the Kappa statistic over the data and obtained $K=0.7$. We found this to be a considerable improvement as compared to earlier $K=0.58$.

In the second case we merged the classes *consistent* and *semi-consistent*. The Kappa statistic with this data amounted to 0.59, which could not be considered an improvement.

6 Concluding Remarks

In this work we raised the question whether it is possible to reliably annotate speech recognition hypotheses with information about semantic consistency and domain specificity. The motivation for that was to find out whether it is feasible to develop and evaluate a computer program addressing the same task and implementing the algorithm reflected in the annotation scheme.

We found that humans principally had problems in looking solely at the *conceptualized* speech recognition hypotheses. This, however, should not be a problem for a machine where the word-to-concept mapping is done automatically and all so-called function words are discarded. In the future it would be interesting to have humans annotate not speech recognition hypotheses *per se*, but only their automatically generated conceptual mappings.

Another finding was that the originally proposed annotation scheme does not allow for a high level of agreement between human annotators with respect to semantic consistency. Eliminating the class *semi-consistent* led us, however, to a considerably better reliability of annotations.

We consider this study as a first attempt to show the feasibility of determining semantic consistency of the output of the speech recognizer. We plan to integrate the results into the domain modeling component and conduct further experiments on semantic consistency and domain detection.

Acknowledgements

The work presented in this paper was conducted within the SmartKom project partly funded by the German Ministry of Research and Technology under grant 01IL9517 and by the Klaus Tschira Foundation.

References

- Carletta, J. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22 (2):249-254.
- Flycht-Eriksson, A. 1999. A Survey of Knowledge Sources in Dialogue Systems. In: *Proc. of IJCAI'99 Workshop on Knowledge and Reasoning in Practical Dialogue Systems*. Stockholm, Sweden, pp. 41-48.
- Mueller, C., Strube, M. 2001. Annotating anaphoric and bridging expressions with MMAX. In: *Proc. of the 2nd SIGdial Workshop on Discourse and Dialogue*. Aalborg, Denmark, 2001, pp. 90-95.
- Oerder, M., Ney, H. 1993. Word Graphs: An Efficient Interface between Continuous Speech Recognition and Language Understanding. In: *Proc. of the International Conf. on Acoustics, Speech and Signal Processing*. IEEE Signal Processing Society.
- Rapp, S., Strube, M. 2002. An iterative data collection approach for multimodal dialogue systems. In: *Proc. of the 3rd International Conference on Language Resources and Evaluation*. Las Palmas, Canary Islands, Spain. To appear.
- Wahlster, W., Reithinger, N., Blocher, A. 2001. SmartKom: Multimodal Communication with a Life-Like Character. In: *Proc. of Eurospeech 2001*. Aalborg, Denmark, pp. 1547-1550.