Semantic Analysis in a Robust Spoken Dialog System

W. Eckert, H. Niemann

Friedrich–Alexander–Universität Erlangen-Nürnberg,
Lehrstuhl für Mustererkennung (Informatik 5),
Martensstraße 3, 91058 Erlangen, GERMANY,
E-mail: wieland.eckert@informatik.uni-erlangen.de

ABSTRACT

In this paper we describe the semantic interpretation process of utterances in a spoken dialog system for train table inquiries. *Spoken dialogs* show a large set of problems in human–machine–communication like stops, corrections, filled pauses, non grammatical sentences, ellipses, unconnected phrases etc. In our robust approach we are able to handle a substantial amount of them. While the principles of robustness are shared in several modules, in this paper we concentrate on the aspect of robust semantic analysis and domain specific interpretation of spoken utterances.

Keywords: robustness, semantic analysis, spoken dialog

1 INTRODUCTION

A very important application in the speech recognition area is the domain of information systems, accessed via telephone. We have built a system for train timetable inquiries which is connected to the public telephone network (cf. [3]). Therefore our system has to cope with untrained, naive users. In this situation typical phenomena of spontaneous speech are quite frequent and the system has to cope with them.

This imposes some constraints on the architecture and processing tasks of the whole system. Apart from plain acoustic problems like distorted signals caused by bad line connection, a critical point is the human factor: it is nearly impossible to predict the behaviour of humans. A solution towards a robust prosodically guided dialog system is already described in [7]. In this paper we describe our approach towards robust semantic analysis which includes semantic processing of spontaneous phenomena.

The structure of our system is shown in Figure 1. Operating on the recognized words, the Linguistic Processor (LP) is responsible to generate a semantic description of the user utterance. Primarily the success of the LP is caused by the use of *partial descriptions*, the so called *utterance field objects* (UFOs), whenever a complete parse is not possible. These partial parses are used for further pragmatic analysis in the dialog manager in order to find an adequate context dependent interpretation. Processing partial information is the normal mode in our system, not just a backup strategy.

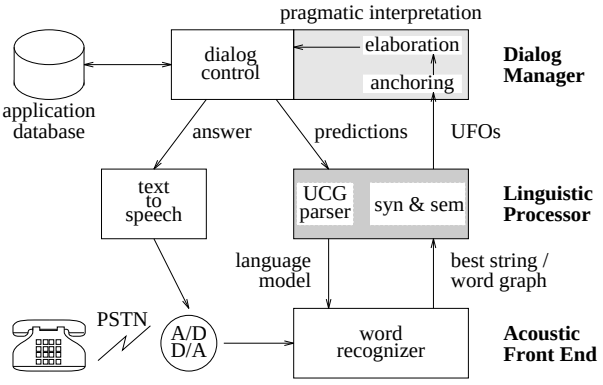

Figure 1: System structure

Inside the LP a chart parser using a UCG (unification categorial grammar) grammar (cf. [5]) is analyzing the word string returned from the acoustic front end. Syntactic and semantic structures are built in parallel by unifying complex feature structures during parsing. Thus, beginning with isolated word interpretations, the spanning edges are extended into a consistent maximal description of the utterance. Whenever no single edge spanning the whole utterance could be found, the selection of a best interpretation represented by a set of UFOs is done according to several quality measures. These measures of optimal semantic interpretation are composed of a set of simple parameters, like the saturation of necessary functor arguments, their coherence with system predictions, the spanning width and, of course, the acoustic score.

A principal demand for the system to withstand an arbitrary (naive) user is *robustness*. We argue, that for successful speech understanding systems the principle of *partial information processing* is vital. This paper describes our approach to the robust analysis of semantic information in spoken dialogs. Another approach using partial parses is described in [1]. Processing partial semantic interpretations is also discussed in [6].

2 ROBUSTNESS

A large set of phenomena of spontaneous speech have already been documented (cf. [11]). Well known problems are, e.g.

- unknown or mispronounced words,
- filled pauses (ah, uh, um . . .),
- restarts — repeating a word or phrase,
- interjections — extraneous phrases,
- ellipses,
- ungrammatical constructions.

While the first of them is "just" a problem of the recognizer, the others are to be handled in the semantic interpretation process. Identification and proper processing of those phenomena is a difficult problem causing every traditional parser to fail to provide a single closed interpretation. There are two principal approaches to overcome such inconveniences: the grammar of the parser is extended by modeling those phenomena, or the linguistic analysis itself is designed robustly.

Our approach to robust semantic analysis is the second one. Based on UCG which encodes grammatical correct expressions and their combination, the key to robust analysis is to use partial descriptions for ungrammatical (or incorrectly recognized) utterances. Two typical examples observed in our collection of phone calls are:

I want to go from – at nine from Koeln.
night train from Berlin – ah – from – from . . .

Usually, parts of the utterance are built properly, but the combination of them is out of coverage of the linguistic grammar. In these cases robust interpretation is performed by using partial results, the UFOs.

3 PARTIAL PARSING

Given a best string (or word graph) as the result of the recognizer, the LP's task is to produce a proper semantic representation. In spontaneous speech the case of incomplete parses is the regular case; dealing with complete and coherent descriptions is the exception. A chart parser is best suited for supplying partial parses, because all previous combinations of partial instances are represented as edges in the chart.

Two modes of operation are supported by our system: parsing of a single string of words and parsing of a word graph. In the *best string mode* the analysis is growing left to right. New words are entered into the chart and all possible combinations with preceding edges are evaluated. Finally, all parses of the utterance are represented in the chart, either as a full parse (i.e. an edge spanning the whole utterance) or as partial parses (i.e. several concatenated edges spanning the utterance).

The *word graph mode* is more flexible, but needs control information. Initial candidates of the word graph are entered as *seed edges*, resulting in an island driven analysis. Additionally, open seeds can be entered by semantic prediction (cf. [5]). Predictions are made for each dialog step by the dialog manager. Each edge has scores associated with it and the search is controlled by an A^* algorithm. Score components (cf. Table 1) are weighted in order to make up a compound score.

Score components consider the number of words in an edge (i.e. the spanning width), the total number of edges remaining in the resulting semantic description, different schemes of acoustic measures, the syntactic completeness of an edge (i.e. saturation of the functor) and the semantic information content of an edge. We are still investigating in different weightings of the components to constitute the final score.

Type	Measure
counting	spanning width of edge, number of edges
acoustic	word score: density, shortfall, shortfall density
syntactic	saturation of necessary arguments
semantic	information content

Table 1: Scoring measures for chart edges.

In both modes of the parser the final step is to determine the best parse, i.e. finding the optimal set of edges. Currently a combination is used which eliminates UFOs with no semantic information content as well as unsaturated necessary arguments. From the resulting set the best UFO sequence (which could be a single UFO) is extracted and sent to the pragmatic processing module.

In our current experiments, the syntactic and semantic scores are weighted extremely high, so every UFO must be syntactically and semantically well formed.

4 ROBUST SEMANTICS

A possible technique to incorporate robustness is the modification of the underlying grammar. In our system we have a very small amount of specialized grammar rules to model phenomena of spontaneous speech explicitly. An illustrative example is U3 in Figure 2, which demonstrates the very application dependent nature of this method. For other application domains there is no rational reason to represent from Koeln with a *type:go*.

However, some phenomena of spontaneous speech are solved in our approach generically. *Pauses* which are recognized as pauses are represented by a special "word" of the lexicon. This symbol is represented in the parser as a functor taking arbitrary arguments of attached edges. The resulting edge has the same properties as the original neighbour and is processed accordingly. In the final representation the embedded pause has just disappeared. *Filled pauses* (ah, uh, um . . .) are represented similarly.

Ellipses are the most natural reason for the usage of UFOs. Partial descriptions are built during the analysis and (apart from the type raising mentioned above) the resulting partial semantic interpretation is representing the user utterance.

Ellipses are a special case of general *ungrammatical constructions*. A maximally consistent subsequence of partial interpretations is constructed in a single pass. Incompatible information bits cannot be combined into a longer spanning edge, so they are left as a sequence of UFOs. This is demonstrated in U2 in Figure 2. Since the elliptic at nine o'clock is compatible neither with U1

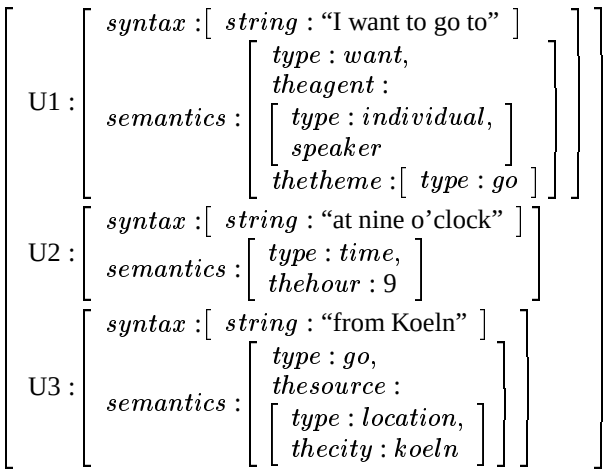

Figure 2: Analysis of *"I want to go to — at 9 o'clock from Koeln"*

nor with U3, it cannot be combined further.

Corrections are comparably simple to process. Since no grammatically correct edge could span a string like from Bremen from Berlin (sometimes with an embedded pause), the resulting interpretation will consist of a sequence of competing UFOs. In the next step of pragmatic analysis only the last instance of the corrected phrase is considered as intended by the user.

All of the above mentioned difficulties can be seen in the example shown in Figure 2. Not shown in this example is the problem of *restarts*. They are typically identified by repetitions and as a special case of corrections handled accordingly.

Unfortunately the processing of interjections is not yet solved.

Usually, parsing of *multi–phrase–utterances*, i.e. utterances consisting of more than one sentence, is seen as a major problem. In our approach, UFOs for all consistent sequences are generated and therefore it is able to process an arbitrary number of grammatical (or even ungrammatical) sentences / phrases in a single utterance.

All these phenomena are handled in an uniform way. No special care has to be taken to identify, exclude and process them in a special manner.

5 ROBUST PRAGMATICS

Finally the pragmatic interpretation process of the semantic information is using the concept of UFOs to find the task relevant parts of information pieces. Based on the processing of isolated UFOs, the two basic principles of interpretation are: the *anchoring* of UFOs into dialog step dependent contexts, and the *elaboration* of them in order to extract the task parameters. The same framework is used for representing the results of the parser and their elaboration: SIL, the *S*emantic *I*nterface *L*anguage (cf. [8]).

Figure 3 shows the result of the pragmatic analysis. Anchoring of U2 (*time*) into a more specialized slot

Figure 3: Domain dependent evaluation of *"I want to go to — at 9 o'clock from Koeln"*

specifying the departure time *sourcetime* is done by applying discourse knowledge. Additionally, predictions for subsequent user utterances are generated and cause a substantial improvement of acoustic recognizer, parser and dialog manager.

Due to the generalized usage of partial information, the pragmatic analysis component is able to interpret every UFO and to build an internal representation using a semantic network scheme. The appropriate context is generated by examining previous utterances and their intentions.

In this semantic network the domain independent semantic information represented by UFOs is reduced to domain and task specific parameters. Those parameters and their status (repeated by user, negated, modified . . .) is reported to the dialog manager. The paradigm of partial information processing is present in every part of the system.

6 RESULTS

In [3] we have reported our two system evaluation sets using microphone quality input with semi–naive subjects who have some knowledge about computers but no experience in speech recognition systems. Our current version was improved by using the robust semantic analysis described above. A third evaluation phase was launched using the telephone line. Due to the simultaneous enhancement of the acoustic recognizer the word accuracy did not degrade. We expected better results in linguistic coverage leading to a substantially higher dialog completion rate. The results are summarized in Table 2.

Measure	P1	P2	P3
word accuracy	66.5%	67.7%	65.6%
concept accuracy	n.a.	n.a.	66.9%
dialog completion	31%	38%	53%
dialog length (min)	2:59	3:03	3:40
# dialogs	255	237	57
avg. # turns	12.0	15.8	26.7
time per turn (sec)	14.9	11.6	8.1

Table 2: Results of former (P1 & P2) and current system setup (P3)

Our experiments show that the average number of turns per dialog increased substantially. This is caused by the dialog manager: when three subsequent user turns failed (i.e. contain no task relevant semantic information), the dialog manager closes down. In P1 & P2 there were many cases where LP could not find any interpretation,

whereas in P3 the system "tried harder" to understand the user utterances.

Due to the substantially increased robustness of the semantic analysis component we reached a much better dialog completion rate: the number of dialogs closed due to sequences of user utterances which seem to carry no task relevant information was reduced greatly.

A very interesting score is the *concept accuracy*, which measures the deviation of semantic concepts after acoustic recognizer and linguistic processor. These concepts are mainly task parameters (sourcecity, date) and dialog markers (yes, good bye). For evaluating the concept accuracy, we examined whether the domain relevant information given in the user utterance is present in the resulting parser output.

A traditional point of view would see the recognizer and parser both as filter operations with non–perfect performance (i.e. $\leq$ 100%). In the case of P3 the concept accuracy is higher than that of the recognizer, so the LP cures some of the recognition errors. Obviously most of the important words were recognized correctly and LP can safely ignore filler words which have no relevant semantic information.

Therefore one has to drop the idea of sequential, independent modules. Robustness means that the independent modules cooperate: one module might cure the other modules errors. In our system the acoustic front end ISADORA and the linguistic processor are well fitted, leading to a concept accuracy which is higher than the word accuracy.

7 CONCLUSION

The outlined approach overcomes the main problems of robustness in semantic analysis. Typical traditional parsers are not able to analyze only parts of the utterance (either a parse is complete or it will fail), whereas reduced parsing strategies (e.g. spotting techniques) suffer from the unconnected aggregation of isolated information pieces. Therefore typical robust approaches try to use reduced parsing as a backup strategy in case of emergency, thus consuming additional processing time for the second step.

Our approach makes no difference between full and partial parses. Partial parses are represented by UFOs accordingly and the whole system is prepared to handle multiple UFOs per utterance. We have shown that a reasonable number of phenomena of spontaneous speech are handled in an uniform way.

Experiments show that robustness against single recognition errors and spontaneous speech is one of the key issues for successful, automatic, easy–to–use, every–person dialog systems. Our approach proved to be highly robust especially against non–syntactic structures caused either by ungrammatical sentences or recognition errors.

There is still some work to be done finding a proper weighting of the score components of an edge. The processing time depends on the sorting criteria of the agenda, especially for large word graphs, whereas the quality of the results is controlled by the selection criteria for the best UFO sequence. Further investigation of the predictions' power to improve time consumption and parse results is planned for the near future. Finally, the best string will be interpreted as a special word graph, thus eliminating the string parse mode.

ACKNOWLEDGEMENTS

This research was partly supported by the CEC ESPRIT project 2218 SUNDIAL with many contributing colleagues. We wish to thank our partners at FORWISS (Erlangen) for the cooperation within the project SYSLID funded by the Daimler–Benz Research Center (Ulm).

REFERENCES

[1] P. Baggia and C. Rulent. Partial parsing as a robust parsing strategy. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, pages II/123–126, Minneapolis, 1993.

[2] A. Brietzmann, F. Class, U. Ehrlich, P. Heisterkamp, K. Mecklenburg, A. Kaltenmeier, P. Regel-Brietzmann, G. Hanrieder, and W. Hiltl. Integration of Acoustics-linguistics for a Robust Speech Dialogue System. In *Proc. Int. Conf. on Spoken Language Processing*, Yokohama, Sep 1994 (this volume).

[3] W. Eckert, T. Kuhn, H. Niemann, S. Rieck, A. Scheuer, and E. Schukat-Talamazzini. A Spoken Dialogue System for German Intercity Train Timetable Inquiries. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 1871–1874, Berlin, Germany, September 1993.

[4] W. Eckert and S. McGlashan. Managing Spoken Dialogues for Information Services. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 1653–1656, Berlin, Germany, September 1993.

[5] G. Hanrieder and P. Heisterkamp. Robust Analysis and Interpretation in Speech Dialog. In *Proceedings of the CRIM / FORWISS Workshop: Progress and Prospects of Speech Research and Technology*, Munich, September 1994 (to appear).

[6] P. Heisterkamp, S. McGlashan, and N. Youd. Dialogue Semantics for a Spoken Dialogue System. In *Proc. Int. Conf. on Spoken Language Processing*, Oct. 1992. Banff, Canada.

[7] R. Kompe, A. Kießling, T. Kuhn, M. Mast, H. Niemann, E. Nöth, K. Ott, and A. Batliner. Prosody takes over: A prosodically guided dialog system. In *Proc. European Conf. on Speech Communication and Technology*, volume 3, pages 2003–2006, Berlin, September 1993.

[8] S. McGlashan. A Proposal for SIL. Technical report, Sundial WP6 (unpublished), 1991.

[9] S. McGlashan, E. Bilange, N. Fraser, N. Gilbert, P. Heisterkamp, and N. Youd. Managing Oral Dialogues. In *6th International Workshop on Natural Language Generation*, Trento, Italy, April 1992.

[10] J. Peckham. A New Generation of Spoken Dialogue Systems: Results and Lessons from the SUNDIAL Projekt. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 33–40, Berlin, September 1993.

[11] W. Ward. Understanding spantaneous speech. In *Speech and Natural Language Workshop*, pages 137–141. Morgan Kaufmann, Philadelphia, 1989.