A ROBUST DIALOGUE SYSTEM FOR MAKING AN APPOINTMENT*

H. Brandt–Pook, G.A. Fink, B. Hildebrandt, F. Kummert, and G. Sagerer

Technische Fakultät, AG Angewandte Informatik Universität Bielefeld
Postfach 100131, 33501 Bielefeld, Germany
e-mail: hbrandt@techfak.uni-bielefeld.de

ABSTRACT

A complete dialogue system within the task domain of making an appointment is presented. It is based on a semantic network representation of linguistic knowledge and a word recognition system that communicates with the interpretation component bidirectionally. System robustness is achieved using a special metascore that evaluates the advance of the linguistic interpretation.

1. INTRODUCTION

The task of dialogue systems is to take over the part of one dialogue participant. Typical applications are database inquiry systems, e.g. [7, 6]. Within the VERBMOBIL project, however, a system for the automatic translation of dialogues restricted to the domain of appointment scheduling is being developed. In order to be able to investigate aspects of speech understanding and dialogue management more easily we simplified the scenario of two humans communicating via a translation device to one human communicating with an automated appointment scheduling system.

Our scenario can be described as follows: the user wants to visit someone e.g. a lawyer. When phoning the lawyer's office he gets connected to the automatic date-placing-system. Figure 1 shows an example for such a dialogue. Let us consider some characteristics of the dialogue. At first the system asks for a general proposal without giving instructions to the user. Thus he is not restricted in formulating his answer. The utterance is analysed and the extracted time-proposal is compared with the internal calendar. Because the day is not completely reserved, the system confirms the proposed day and asks for a detailed time. The following utterance ("at three in the afternoon") can only be understood correctly in the dialogue context, because the user gives only the new information but does not repeat the day discussed so far. This behaviour is typical for human dialogues. Thus one of the main (and most complicated) tasks of the dialogue-component of our system is the time-merging. Finally the system confirms the now precise proposal and expects a confirmation of the user before it finishes the dialogue.

*This work was funded by the German Federal Ministry of Education, Science, Research and Technology (BMBF) in the framework of the VERBMOBIL Project under Grant 01IV102G/7. The responsibility for the contents of this study lies with the authors.

We are particularly interested in robustness. In this connection robustness means a system reaction satisfying the user although the input data was corrupted and therefore the utterance could not be recognized and analysed completely. On this special topic we will focus in section 5.

System: "Sie können mit mir einen Termin vereinbaren. Bitte machen Sie einen Vorschlag!"
(You can make an appointment with me. Please make a proposal!)
User: "Ich möchte einen Termin am Dienstag oder Mittwoch."
(I would like a date on Tuesday or Wednesday)
System: "Am Dienstag sieht es gut aus. Wann genau paßt es Ihnen an diesem Tag?"
(On Tuesday it looks well. What time do you prefer at this day?)
User: "Am Nachmittag um drei."
(At three in the afternoon.)
System: "Ihr Vorschlag paßt mir sehr gut. Bitte bestätigen Sie noch einmal: Ist es richtig, daß der Termin am Dienstag um 15 Uhr stattfindet?
(Your proposal suits me very well. Please confirm once more: is it correct that we fixed the date for Tuesday at 3 p.m.?)
User: "Ja."
(Yes.)
System: "In Ordnung, dann bis Dienstag um 15 Uhr."
(Okay, see you at Tuesday at 3 o' clock p.m.!)

Figure 1: Introductory dialogue example

2. LINGUISTIC ANALYSIS

For knowledge representation the semantic network language ERNEST [5] is used. In this language three types of nodes exist. In *concepts* the a priori knowledge is stored. *Instances* establish a connection between the input data and a concept, i.e. an instance is a concept filled with individual values. Finally *modified concepts* contain partial results already obtained during the analysis. As usual the nodes are connected with links. In ERNEST three link types are

User: "Donnerstag morgen habe ich zwei Stunden Zeit."
(On Thursday in the morning I have two hours.)
System: "Ich habe Sie leider nicht verstanden. Welche Zeit hatten Sie vorgeschlagen?"
(Sorry, I didn't understand you. What time did you propose?)
User: "Es geht am Donnerstag morgen."
(It is possible on Thursday in the morning.)
System: "Am Donnerstag von sechs Uhr bis neun Uhr sieht es gut aus. Wann genau paßt es Ihnen?"
(On Thursday from 6 o'clock up to 9 o'clock it looks well. What exact time do you prefer?)
User: "Nein. Dann lieber am Samstag."
(No. In this case I prefer Saturday.)
System: "Nein. Am Sonnabend geht es nicht. Wann haben Sie sonst noch Zeit?"
(No. On Saturday it is not possible. What other time do you prefer?)

Figure 2: Example for system repetition and refusal

defined. The *part-of* link establishes a decomposition, the *specialisation* link models the relationship "general – special" whereas the *concrete* link connects nodes of different degrees of abstraction.

In our system we represent the necessary knowledge in five levels of abstraction.

The word hypothesis level establishes the traditional interface to the word recognizer. Word hypotheses of the recognizer are represented in an instance on this level.

On the syntactic level syntactic constituents and word classes are modelled as well as time constituents. E.g. the concept *T_DAY* represent utterances that contain information about a date, a week-day, or a special holiday [3].

The meaning of a single utterance is analysed on the semantic level. Deep case theory [1] forms the theoretic basis of the concepts on this level. E.g. on the one hand the verb *propose* opens deep cases for an initiator and a receiver of the proposal and on the other hand deep cases for the proposed action. Up to this level the modelling is problem independent and we use a similar implementation in other speech understanding systems.

On the pragmatic level the semantic descriptions are applied to the specific domain. Thus e.g. the knowledge stored in the semantic verb frame of the verb *to take* is now restricted. The according pragmatic intention models the pragmatic verb frame of a proposal for an appointment ("Let's take Wednesday"). All other meanings of *take* are no longer considered.

Finally the dialogue level manages the strategy of the session and realizes the dialogue memory. We want to focus on two aspects of this tasks.

The first one is concerned with the dialogue strategy. Every utterance of the user is labelled with one of the categories *proposal*, *confirmation*, or *refusal*. In figure 1 the first and second utterances are proposals whereas the third one forms a confirmation. A more complicated situation is given in figure 2. Here the user takes back his first proposal by refusing the system confirmation. At the current state of the system this utterance is classified as a (new) proposal although it contains also a refusal.

After the determination of the utterance type the adequate system reaction must be selected. We modelled four possible system reactions: final confirmation, partial confirmation, refusal, and repetition. A final confirmation (e.g. last system output in Fig. 1) will only happen if the gained information about the time of the appointment is precise enough *and* if the time is labelled as free in the calendar of the system *and* if the preceding utterance was a confirmation. Thus it is sure that finally the time stored in the dialogue memory is really the time intended by the user. If the proposed time would be adequate for an appointment but a final confirmation is not possible, the system reaction will be a partial confirmation that includes – if necessary – a request for more information (e.g. second and third system output in Fig. 1). A refusal will be given, if the uttered time is already reserved (e.g. third system output in Fig. 2). The system will repeat its last reaction if there is no other possibility, e.g. the interpretation of the utterance was completely unsuccessful (e.g. first system output in Fig. 2).

User: "Wir können uns am 7. Mai um zehn Uhr treffen."
(We could meet on May 7th at 10 o'clock.)
System: "Nein. Um zehn Uhr am Dienstag geht es nicht. Wann haben Sie am Dienstag sonst noch Zeit?"
(No. On Tuesday 10 o'clock a.m. it is not possible. What other time do you prefer on Tuesday?)
User: "Zwei Stunden später wäre möglich."
(Two hours later would be possible.)
System: "Ihr Vorschlag paßt mir sehr gut. Bitte bestätigen Sie noch einmal: Ist es richtig, daß der Termin am Dienstag um zwölf Uhr stattfindet?
(Your proposal suits me very well. Please confirm once more: is it correct that we fixed the date for Tuesday at 12 a.m.?)

Figure 3: Example for the time-merging

The second aspect we want to look at is concerned with the dialogue memory that stores the relevant information given by the user. Only on the dialogue level it is possible to merge an actual time proposal with former ones. In the second utterance of figure 3 the time-description "two hours later" can only be analysed correctly within the dialogue context. "two hours later" builds up an instance on the syntactic level. Figure 4 shows the structure *rel-time* computed on this level and the later merging with the so far actual dialogue time *dia-time*.

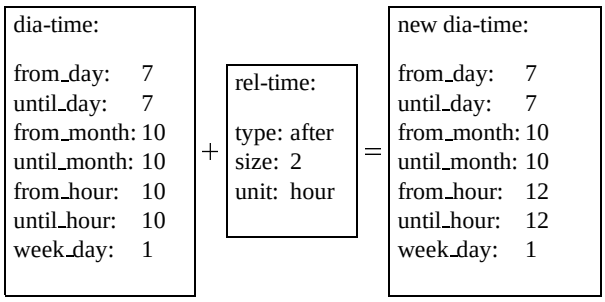

Figure 4: Time-merging

3. SPEECH RECOGNITION

The recognition module is based on the ISADORA system [8]. It provides highly flexible Markov-model-based speech recognition and the possibility to build structured acoustic descriptions from simple constituents. A rule mechanism is used to represent complex acoustic models that are capable of describing any regular language of the basic models known to the system. The building of complex HMMs from simpler models by means of these rules does not simply yield a single complex model. The hierarchy of rule constituents is preserved as a hierarchy of HMMs.

4. INTEGRATED CONTROL

We use an integrated control strategy for both recognition and understanding [2] that tries to overcome many of the disadvantages of traditional loosely coupled systems. It makes use of the possibility to process abstract constituents in our word recognizer and pass them back as complex hypotheses.

Before the analysis of an utterance for certain constituents linguistic language models are created automatically from their semantic network representation. Speech recognition and interpretation then alternate between model driven prediction steps and the verification of the predicted acoustic models by the recognizer. At the beginning of the analysis the set of all constituents allowed to start an utterance are passed to the recognizer as predictions. During the recognition process the corresponding language models are applied to generate complex constituent hypotheses. These can be mapped to structures of the linguistic interpretation automatically and thus be integrated into possibly competing analysis results easily. Depending on these partial interpretations different sets of constituents can be computed for the following predictions that are processed analogously. This incremental control strategy is continued until the end of the utterance is reached.

A severe problem of *all kinds* of language modelling are portions of an utterance that do not adhere to this model leading to erroneous results. Therefore, a special model for an unknown constituent [4] is part of *every* prediction set passed to the recognizer. This makes it possible to interpret utterances successfully though possibly only partially that are not conforming to the language model applied.

5. SYSTEM ROBUSTNESS

Each state of the analysis is represented in a scored node of a search tree. Alternative states of the incremental analysis are stored in competitive nodes. The search for the best analysis is based on the well-known A*-algorithm. To score a node we use the acoustic score of the hypotheses, the number of signal frames covered by the analysis, and linguistic and syntactic admissibility. Usually this score is good enough to control the analysis. However, an utterance "ich möchte äh äh – einen Termin am Montag" (I want er er – a date on Monday) the hesitations splits up the search tree enormously. Also other irregularities within the utterance can defer the analysis.

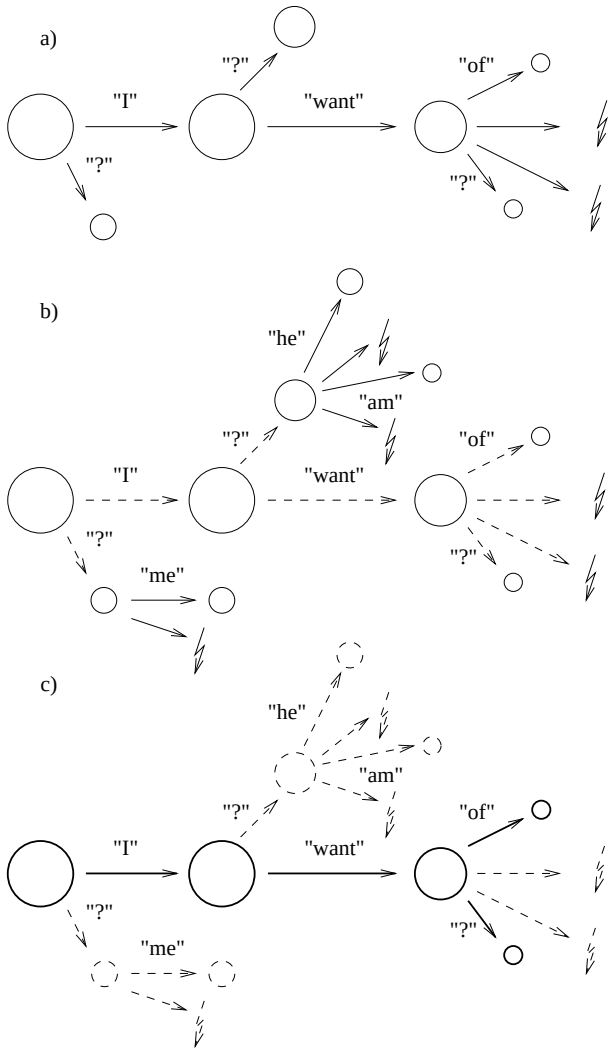

Figure 5: Influence of the meta-score on the search

Let us consider figure 5 (according to the example utterance above). In this figure the size of a circle represents the score of an search tree node – bigger circles stand for better scores. The flash symbolises inadmissibility and the question-mark stands for the hypothesis "unknown word". In figure 5a) we see the search tree at the crucial

point. The A*-algorithm causes the expansion of those nodes who refer to short but better scored parts of the utterance. That means the search concentrates on the section *before* the problematical part of the signal as shown in figure 5b).

To increase the system robustness it must be possible to overcome such situations. Therefore we added to the scores mentioned above a special meta-score that evaluates the *advance of the interpretation*, where an interpretation is defined as follows: it refers to a non empty signal-area (i.e. results that are only based on hypotheses of *unknown word* are not taken into consideration), it contains an instance of a pragmatic intention, and all predictions are terminated. We will consider an interpretation better than another one, if it contains more instances, covers more input data, and has a better score. If the interpretation stagnates (i.e. after a fixed number of tree nodes no better interpretation is found), the search tree will be cleaned up and the analysis will continue leaving the so far best interpretation and its successor nodes for further processing but omitting all the parts of the search tree containing less promising results. Figure 5c) shows the surviving nodes (solid).

6. FIRST RESULTS

We used a speaker independent word recognizer built with the ISADORA-System [8] which was trained on 5520 utterances of 48 speakers. The first tests were performed by a speaker who was familiar with the system but not involved in training the word recognizer. The analysis of 16 dialogues containing 77 utterances is shown in table 1. All in all 13 (81.3 %) of these dialogues were finished successfully, i.e. finally an satisfying appointment was made.

utterances	complete analysis	partial analysis	wrong analysis	no analysis
77	51	8	4	14
100 %	66.2%	10.4%	5.2%	18.2%

Table 1: First results

In this table “complete analysis” means that the information of the utterance (time proposal, refusal or confirmation) was totally ascertained. A “partial analysis” extracts only some of the uttered facts whereas a “wrong analysis” draws a wrong conclusion from the utterance. Finally the row “no analysis” counts the abortions (only one) and repetitions (13).

7. CONCLUSION

We presented a dialogue system for making an appointment. Linguistic knowledge is represented in different levels of abstraction within a homogeneous semantic network system. The word recognizer uses linguistic language models automatically extracted from the knowledge base. System robustness is achieved using a score that evaluates the quality of the linguistic interpretation. The system processes utterances incrementally and is speaker independent.

Our future work at this system will focus on the time modelling, because up to now not all possible formulations concerned with the time are covered by the model. Especially a model for longer time periods needs to be to developed. Furthermore, utterances containing a negation (e.g. ”um zehn Uhr kann ich nicht, aber um zwölf” (it is not possible at 10 a.m. but at 12 a.m.)) are not yet processed satisfactorily.

8. REFERENCES

1. C. Fillmore. A case for case. In E. Bach and R. T. Harms, editors, *Universals in Linguistic Theory*, pages 1–88. Holt, Rinehart and Winston, New York, 1968.
2. G. A. Fink, F. Kummert, and G. Sagerer. A Close High-Level Interaction Scheme for Recognition and Interpretation of Speech. In *ICSLP-94: Proc. of the Int. Conf. on Speech and Language Processing*, volume 4, pages 2183–2186, Yokohama, Japan, 1994.
3. B. Hildebrandt, G. A. Fink, F. Kummert, and G. Sagerer. Modeling of time constituents for speech understanding. In *Proc. European Conf. on Speech Communication and Technology*, pages 2247–2250, Berlin, 1993.
4. A. Jusek, G. A. Fink, F. Kummert, H. Rautenstrauch, and G. Sagerer. Detection of unknown words and its evaluation. In *Proc. European Conf. on Speech Communication and Technology*, pages 2107–2110, Madrid, 1995.
5. F. Kummert, H. Niemann, R. Prechtel, and G. Sagerer. Control and Explanation in a Signal Understanding Environment. *Signal Processing, special issue on 'Intelligent Systems for Signal and Image Understanding'*, 32:111–145, 1993.
6. M. Mast, F. Kummert, U. Ehrlich, G. A. Fink, T. Kuhn, H. Niemann, and G. Sagerer. A Speech Understanding and Dialog System with a Homoge neous Linguistic Knowledge Base. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 16:179–193, 1994.
7. R. Pieraccini, E. Tzoukermann, E. Gorelov, Z. an d Levin, C. Lee, and J. Gauvain. Progress Report on the CHRONUS System: ATIS Benchmark Results. In *Speech and Natural Language Workshop*, pages 67–70. Morgan Kaufmann, 1992.
8. E. G. Schukat-Talamazzini. *Automatische Spracherkennung*. Vieweg, Wiesbaden, 1995.