A GERMAN DIALOGUE SYSTEM FOR SCHEDULING DATES AND MEETINGS BY NATURALLY SPOKEN CONTINUOUS SPEECH

Daniel Willett, Arno Römer, Jörg Rottland, Gerhard Rigoll

Department of Computer Science
Faculty of Electrical Engineering
Gerhard-Mercator-University Duisburg, Germany
e-mail: willett@fb9-ti.uni-duisburg.de

ABSTRACT

In this paper, we present the basic design principles and architecture of a dialogue system for scheduling appointments. This mixed-initiative dialogue system integrates an automatic speaker-independent speech recognition engine for continuously spoken German, a speech synthesizer and a scheduler database application to build up a scheduler that is purely driven by natural continuous speech and thus, does not need any visual display device. With these properties it is a prototype for a speech driven palm-size computer application and could be integrated in miniature computers that come along with no display device at all.

1. INTRODUCTION

Dialogue systems enable the user to fulfill some well defined interaction with the machine by natural conversational speech in a spoken dialogue, in which the computer takes the part of one of the dialogue participants. The techniques and principles for the development of robust dialogue systems have attracted a lot of attention in the recent years. In the following, we will describe our experiences with these design principles and outline some of the most important features of a dialogue system developed by the authors for scheduling dates and meeting in naturally spoken German speech. A user-oriented approach [7] was chosen for the design of this mixed-initiative [1, 6] dialogue system.

At ICSLP 1996, Brandt-Pook et al. presented a German dialogue system for making appointments [3]. In their system, they focused on the interaction between the linguistic interpretation unit and the speech recognizer. They saw the major application in automatic appointment arrangements over the telephone. Hence, they restricted the system to dialogues for arranging appointments and did not allow queries and deletions.

2. DESIGN GOALS

The major design goals of our dialogue system were to preserve as much naturalism from human-to-human dialogues as possible, and at the same time to have a high degree of usability. Naturalism in this sense means, that the system constrains the user's utterances as little as possible, and makes him feel like talking to a human being as much as possible. Usability in this case mainly means a small number of misunderstandings and a rapid correction of those.

Figure 1 displays an example dialogue to give an impression of how the processed dialogues look like. Lots of similar dialogues were collected in a Wizard-Of-Oz (WOZ) scenario [4] before and during the development of the system, so that the system is able to handle very different kinds of approaches towards it. It has to be considered that different users have different opinions about what is the most "natural" approach to address a dialogue system.

System: Hallo.	Hello.
User: Ich habe morgen was vor.	I've something to do tomorrow.
System: Was soll ich für morgen eintragen?	What shell I enter for tomorrow?
User: Ich treffe mich mit Peter.	I'm going to meet Peter.
System: Wann beginnt das Treffen mit Peter?	When does the meeting with Peter start?
User: Abends, so um halb sechs wahrscheinlich.	In the evening, at around five thirty.
System: Soll ich ein Treffen mit Peter am 4.7.1998 um 17 Uhr dreißig eintragen?	Shell I enter a meeting with Peter for the 4.7.1998 at 17.30?
User: Ja, bitte.	Yes, do so.
System: Eintragung erfolgt.	Done.
User: Was liegt denn heute so an?	What's up today?
System: Soll ich die Termine vom 3.7.1998 anzeigen?	Shell I list the dates for the 3.7.1998?
User: Ja.	Yes.
System: Am 3.7.1998 ist ein Termin um 11 Uhr bis 13 Uhr und ein Termin um 16 Uhr beim Zahnarzt	On the 3.7.98 there is a date at 11 o'clock until 13 o'clock and a date at 16 o'clock at the dentinst.
User: Danke.	Thanks.

Figure 1. Example dialogue to be handled by the system

Contrary to other dialogue systems that are mainly used for the retrieval of information from databases (e.g. time-tables) or allow only a very limited write-access to the database (e.g. ticket-booking), the developed dialogue system has to handle complex data input like "I'm gonna meet Dave on Sunday at ten" or "Hey computer, what's up on Saturday?" (translated). In order to avoid misunderstandings, that are particularly dangerous when adding new data to the database, the dialogue system comes along with a very sensitive and secure behavior. The system repeats any data that it extracted from the user's utterance (in synthesized speech) in order to let the user check and possibly correct it. However, another somewhat opposing goal of the dialogue design was to avoid unnecessary repetitions that reduce the processing speed and tend to bore the user.

Furthermore, an important issue is that the user should never feel left alone in the dialogue not knowing the dialogue status and his options. Therefore, the system has to generate questions that guide the (unexperienced) user through the dialogue. However, the (more experienced) user should be free in addressing the system with only as few restrictions as possible. In [6], Larsen introduced the term Mixed-Initiative for the kind of dialogue systems that we intended to set up.

3. SYSTEM ARCHITECTURE

The system's architecture is illustrated in Figure 2. Beside the essential components of such systems, the speech recognition and synthesis engine, the database and the keyword-based dialogue manager, the system contains a dialogue memory in order to be able to access data from previous dialogues and in order to collect the needed data in multiple questions and answers. The speech re-

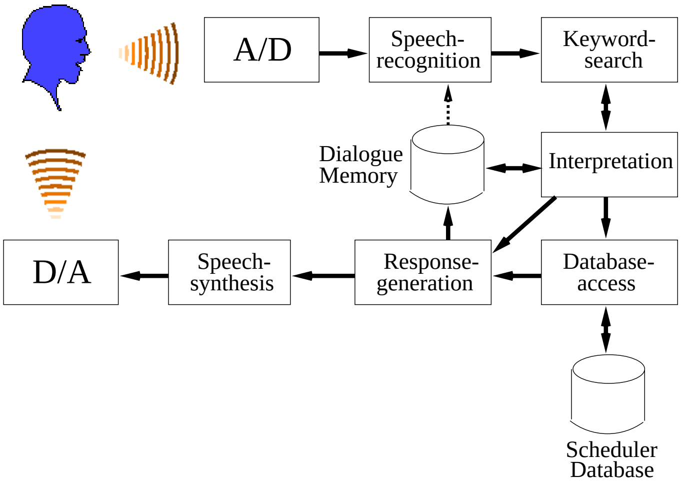


Figure 2. Basic architecture

cognition component sends a sequence of words to the keyword extraction unit. This unit is linked bidirectional to the linguistic interpretation unit. The linguistic interpretation is based upon the keywords found and upon the dialogue memory including the actual state that the current dialogue is in. The interpretation either results in a database access and an answer-generation or in a question asking for further details. The influence of the dialogue memory on the speech recognition component in cases of multiple finite recognition grammars is discussed in Section 7.

The interpretation of the user's utterance is based upon the spotting of keywords. It is a rule-based evaluation, set up by the WOZ experimental prototyping method [2]. The extraction of time and place specific information is independent of the dialogue act interpretation. These issues are discussed in the following sections.

The dialogue memory is a record of several registers. Its components are listed in Figure 3. For each of the dialogue acts, Query, Input and Deletion, several sets of key components are defined that have to be specified before this kind of dialogue act can be confirmed and a command can be directed to the database. In Figure 3 for example, the dialogue act has been determined as Input. the system keeps in its memory that the user wants to input a business dinner with Peter and Paul in the evening. The user has said something like "In the evening, I'm gonna meet Peter and Paul for a business dinner.". A required field however for this type of appointment is an exact date. As this has not yet been specified, the status remains more-specifications-needed and the system asks for the date and does not allow a confirmation of this input. Detailed information on the state space and the interpretation pipeline can be found in [8].

Dialogue act:	Input		
State:	need more specifs.		
Appointment type:	business dinner		
Participant:	Peter, Paul		
Time: from			
to			
approx:	in the evening		
Date: from			
to			

Figure 3. The elements of the dialogue memory

4. DIALOGUE ACTS

The interpretation unit, as displayed in Figure 2 fulfills the linguistic analysis. As a first task, the three major types of dialogue acts have to be distinguished. These are defined as Input, Query and Deletion, and symbolize the user's basic intension, the reason why he addresses the system. Figure 4 displays the dialogue acts as a state automata. From the baseline state it has to be decided what type of dialogue act the user's intension belongs to. Observations on the WOZ data showed that the dialogue act can be determined safely from the user's first utterance (as long as there is no recognition error). Once, the user's intension has been determined the purpose of the state model is to guide the further question and answer procedure until the user has given enough information about what he wants to know, enter or delete or until the user cancels the dialogue. An important feature is the possibility to cancel the running dialogue at any point by natural speech. (For matters of simplicity, these connections are omitted in Figure 4.) Examples of such canceling commands were gained in the WOZ scenario as well. This way the user is always able to reset the dialogue, whenever it got totally wrong because of recognition errors and other types of misunderstanding.

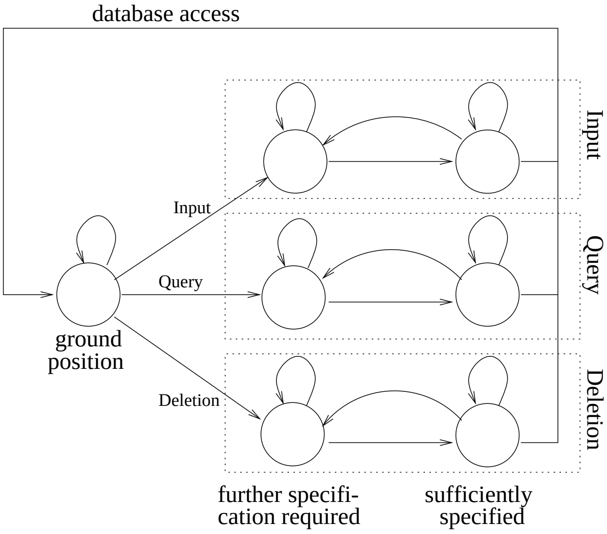
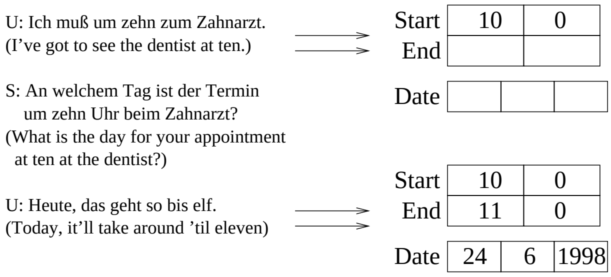


Figure 4. The baseline state-graph of the linguistic interpretation unit

5. EXTRACTING INFORMATION ABOUT TIME AND PLACE

Some work on the extraction of time and date information from natural (German) speech has been published in [5]. As proposed in [5], the dialogue system has a record of several fields that stores the starting- and ending-time and date. This record is part of the dialogue memory. Whenever key-phrases such as "um acht Uhr" (at eight o'clock) or "bis zum zwanzigsten" (until the twentieth) are observed, these fields are filled. Figure 5 illustrates this procedure. A portion of the word sequences that define time constituents is illustrated as a word graph in Figure 6. A likewise graph in defined for the specification of date information. Especially


U: Ich muß um zehn zum Zahnarzt.
(I’ve got to see the dentist at ten.)

S: An welchem Tag ist der Termin
um zehn Uhr beim Zahnarzt?
(What is the day for your appointment
at ten at the dentist?)

U: Heute, das geht so bis elf.
(Today, it’ll take around ’til eleven)

Figure 5. Extracting time constituents during the dialogue

relative expressions like ”morgen” (tomorrow) or ”nächsten Mittwoch” (next Wednesday) deserve a careful treatment. Some refer to a date (i.e. the begin-date) given in previous utterances, some refer to the real date that the dialogue is spoken at.

In addition to the time and date specification, each database entry allows the specification of an appointment type and a list of people that the specific appointment is associated with. At this point, a major disadvantage of the common approaches to speech recognition comes into effect. Due to the limited vocabulary, only a limited number of appointment types and a limited number of Christian and family names can be supported. In order to overcome this problem, at least the expansion of the recognition vocabulary should be made possible, so that the user is able to adjust the system to his needs. This issue will be further discussed from the viewpoint of continuous speech recognition in Section 7.

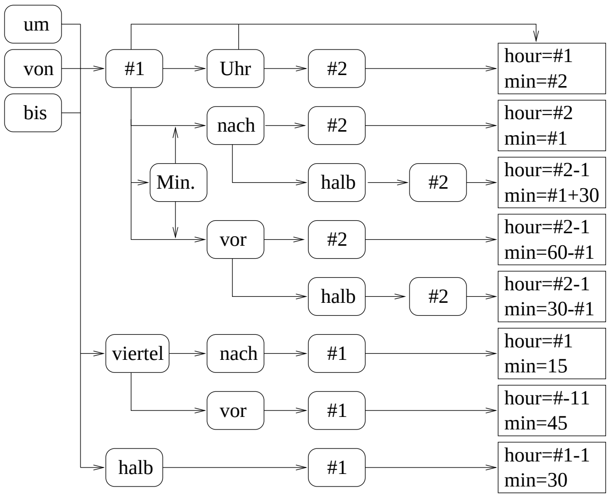


Figure 6. Understanding time constituents by graph-based key-phrase spotting

6. CORRECTING MISUNDERSTANDINGS

Misunderstandings occur in dialogue systems just as they occur in natural dialogues. These misunderstandings are due to misrecognized words and they are due to ambiguities that appear in every language. Therefore, no dialogue system will ever be able to omit all misunderstandings. However, the user has to detect misunderstandings and has to be enabled to correct them. In our dialogue system, this is accomplished in a very straight forward way, that we observed to be very effective. All the data in the dialogue memory can be overwritten by new data of the same type found in newer dialogue utterances. An example dialogue that demonstrates this behavior is given in Figure 7. After the system misunderstood the time and type of the meeting, the rejection of the confirmation with the time and type specification given again corrects these fields and can afterwards be confirmed. Finally, the whole dialogue did not have to be canceled because of the misunderstandings caused by the speech recognizer.

U: Ich muß morgen um 15 Uhr zum Zahnarzt.	U: Tomorrow at 15 o’clock I’ll see the dentist.
(recognized: Ich muß morgen um 5 Uhr zu zahlen.)	(recognized: Tomorrow at 5 o’clock to pay.)
S: Soll ich einen Termin um 5 Uhr am 8.7.98 eintragen?	S: Shall I enter a date at 5 o’clock for 8.7.98?
U: Nein, um 15 Uhr muß ich zum Zahnarzt.	U: No, at 15 o’clock I see the dentist.
S: Soll ich einen Termin um 15 Uhr am 8.7.98 beim Zahnarzt eintragen?	S: Shell I enter a date at 15’clock for 8.7.98 at the dentist?
U: Ja, na also.	U: Yes, you did it.
S: Eintragung erfolgt	S: Done.

Figure 7. Correcting misrecognitions

7. SPEECH RECOGNITION COMPONENT

The Verbmobil continuous speech corpus consists of spontaneous face-to-face dialogues from the domain of appointment arrangements. The original idea was that a speaker-independent speech recognizer based on this corpus should to be ideal for the task of recognizing speech within the appointment dialogues.

7.1. Statistical language models

In the originally applied recognizer the hidden Markov models, that model the acoustic observation likelihoods, as well as the vocabulary and the language model were both set up on the Verbmobil corpus. However, it turned out that the application of this recognition system within the dialogue system leads to several new problems. These are mainly due to the circumstance that the way people address dialogue systems widely differs from how they address other human beings. Therefore, the vocabulary and especially the language model obtained on the face-to-face dialogues fails to capture the common dialogues observed in the Wizard-Of-Oz experiments. Hence, the number of successfully finished dialogues tends to zero, when simply using the Verbmobil-based speech recognizer.

7.2. Finite grammar

Because of the unsuccessful experiences with the bigram-based continuous speech recognition, we followed a different approach in the further development. Namely the usage of a (complex) finite grammar, set up on the WOZ example dialogues, and the application of word-based confidence measures in order to reduce the number of false word hypothesis [9]. Of course, this limits the naturalness to some extend. However, as the grammar is loose enough to cover all the sample dialogues and a lot of additional variations, it is well suited for recognizing utterances of the dialogue domain.

A further improvement in recognition accuracy can be gained by setting up different finite grammars for each of the dialogue states. However, this affords even more text data, as several robust finite grammars have to be set up. For the presented system we tried to have one grammar for the baseline state, when the user has lots of freedom concerning the possible utterances, and one for the rest of the dialogue states, when the user’s intension (the dialogue act) has been determined and he only gives further specification, cancels or confirms his input.

A further advantage of using finite grammar(s) over the application of stochastical language models is the possibility of an easy extension with new words, especially names and appointment types. This allows the user to easily adjust the speech recognition component to his needs.

8. FIRST PERFORMANCE EVALUATIONS

For the evaluation of dialogue systems, several approaches have been proposed [7]. On the whole, it has to be considered that the success of a spoken dialogue system largely depends on the knowledge and education of the user and on the user's cooperativeness. Our tests were all run with cooperative users with a considerable technical knowledge.

The WOZ dialogues that were setup before and during the system design are processed by 100%. The system was specified to do so, therefore this is not too surprising and doesn't say much about the system's ability to generalize from these dialogues. In experiments run with the speech recognition component replaced by a keyboard in order to have no errors caused by the speech recognition component, about 80% of the 'dialogues' were processed correctly. Furthermore, more than half of the misunderstandings that occurred in the remaining 20% could be corrected (see Section 6).

With the bigram language model trained on the Verbmobil face-to-face dialogue corpus, most of the dialogues could not be accomplished satisfactorily. The rate of correctly finished dialogues was around 30%, but has not been estimated accurately.

With the finite grammar set up on the WOZ training dialogues the rate of successfully finished dialogues could be raised to 60%. The application of acoustical confidence measures to compensate for the strict grammar raised this number to 80%, although it lengthens the dialogues, as sometimes correctly recognized words and word sequences fall below the confidence threshold and have to be repeated. The application of two different finite grammar as proposed in Section 7 provided no measurable improvements.

9. CONCLUSION

A dialogue system that manages a scheduling application purely driven by natural speech has been presented. It has been shown how, with some constraints concerning the speech recognition component, this system is capable of completing weakly constrained dialogues successfully. The system handles all the training dialogues, set up in a WOZ scenario. The application of confidence measures for the hypothesized words of the speech recognizer helps reducing misunderstandings. Within the system's architecture, a basic design principle of dialogue systems allowing read-, write- and delete-access to a database has been presented.

10. REFERENCES

1. N. O. Bernsen, L. Dybkjaer, H. Dybkjaer: "A Dedicated Task-Oriented Dialogue Theory in Support of Spoken Language Dialogue Systems Design", Proc. ICSLP, Yokohama, 1994, pp. 875–878.
2. N. O. Bernsen, L. Dybkjaer, H. Dybkjaer: "Principles for the Design of Cooperative Spoken Human-Machine Dialogue", Proc. ICSLP, Philadelphia, 1996, pp. 729–732.
3. H. Brandt-Pook, G. A. Fink, B. Hildebrandt, F. Kummert, G. Sagerer: "A Robust Dialogue System for Making an Appointment", Proc. ICSLP, Philadelphia, 1996, pp. 693–696.
4. H. Dybkjaer, N. O. Bernsen, L. Dybkjaer: "Wizard-Of-Oz and the Trade-off between naturalness and recognizer constraints", Proc. ICSLP, Yokohama, 1994, pp. 947–950.
5. B. Hildebrandt, G. A. Fink, F. Kummert, G. Sagerer: "Modeling of Time Constituents for Speech Understanding", Proc. EUROSPEECH, Berlin, 1993, pp. 2247–2250.
6. L. B. Larsen: "A Strategy For Mixed-Initiative Dialogue Control", Proc. EUROSPEECH, Rhodes, 1997, pp. 1331–1334.
7. C. Müller, F. Runge: "Dialogue Design Principles - Key for Usability of Voice Processing", Proc. ICSLP, Yokohama, 1994, pp. 943–946.
8. A. Römer: "The Development of a Natural Language Dialogue System for Scheduling Appointments", Diploma-thesis (in German), Duisburg University, 1998.
9. D. Willett, A. Worm, C. Neukirchen, G. Rigoll: "Confidence Measures for HMM-based Speech Recognition", somewhere in these proceedings.