EVALUATION OF SPOKEN LANGUAGE UNDERSTANDING AND DIALOGUE SYSTEMS

B. Hildebrandt, H. Rautenstrauch[*]*, G. Sagerer*

Universität Bielefeld, AG Angewandte Informatik,
Postfach 100131, 33501 Bielefeld, Federal Republic of Germany
[*]SAP-AG, Neurotstr.16, 69190 Walldorf, Federal Republic of Germany

ABSTRACT

A spoken language understanding and dialogue system in the domain of appointment schedule is presented. The system is capable of understanding complex times, e.g. it correctly combines discontinuous constituents and resolves ambiguities. A distributed representation of surface structure models and an incremental semantic analysis is used to manage the complexity. An elaborate evaluation of the system based on measurements of accuracy was carried out. Our approach combines pattern recognition with linguistic aspects forming a system of measurement consisting of word accuracy, constituent accuracy, and concept accuracy.

1. INTRODUCTION

In a speech understanding system there are several levels of analysis and interpretation. Firstly, the system has to recognize single words in a torrent of speech sounds. Secondly, the system combines words to constituents, i.e. it performs a syntactic analysis. It also has to reconstruct the meaning of the utterance in question, i.e. it performs a semantic interpretation. Every step can be executed independently. However, it is more efficient, if the system follows a strategy alternating between data and model driven phrases [5]. This becomes obvious in the analysis and interpretation of time constituents, which is important for many applications of spoken language understanding systems (e.g. VERBMOBIL).

The area our research is concerned with is appointment schedules. Since diverse time constituents are distributable in variable positions within an utterance, problems of modelling can emerge. For example, in German it is possible to say "Ich kann Sie *Dienstag vor acht Uhr abends* treffen." (I can meet you *on Tuesday before 8 o'clock in the evening*.) as well as "*Am Dienstag* kann ich Sie *abends vor acht Uhr* treffen.". Both sentences have the same meaning and in spoken language even the following version would be acceptable: "*Vor acht Uhr* kann ich Sie *am Dienstag* treffen *abends*." [2, 3].

This work was funded by the Federal Ministry of Education, Science, Research and Technology (BMBF) in the framework of the Verbmobil Project under Grant 01IV102AO. The responsibility for the content of this study lies with the authors.

2. MODELLING OF TIME CONSTITUENTS

Time constituents are one kind of syntactic units, which organize semantic concepts. In a corpus of domain specific spontaneous dialogues several types of time constituents can be found; for example:

- The *time of day* can be expressed in a phrase which is built up coordinating hour and minute ('um acht Uhr und zehn Minuten', 'at eight hours and ten minutes') or relating them ('um zehn Minuten vor acht Uhr', 'at ten to eight'). A phrase can also consist of hour and another unit of time ('um dreiviertel acht', 'at a quarter to nine').
- The *section of day* which can be a phrase with an adverb as nucleus ('für abends', no equivalent in English) or with a noun as nucleus ('am frühen Abend', 'early in the evening').
- The *date* which can be expressed by an adverb ('für morgen', 'tomorrow') or by a noun ('am Dienstag', 'on Tuesday'). In addition, the *time point of speech* can be relevant for the interpretation of date ('am kommenden Dienstag', 'next Tuesday') or date can be expressed in a rather absolute way ('am fünften Mai', 'May 5th').

In spoken German there does not seem to be any rule that restricts the position or the combination of time constituents in a sentence.

The first step in order to manage this complexity and variability is the syntactic analysis on *phrase structure level*. Each time constituent in an utterance is analyzed independently and tested for its syntactic coherence. For this purpose several *modalities* are defined, which represent different variations of constituent structures. A modality itself is defined with *categories* for which *adjacencies* are stipulated. Therefore, 'für halb fünf ' (half past four) and 'für Viertel nach fünf' (at a quarter past five) are correct phrases, but not *'für halb nach fünf'.

A semantic interpretation of the time constituent has to follow. However, it is not sufficient to reconstruct the correct time point. The system also has to detect whether the intended appointment is supposed to start within an interval, i.e. it happens between two time points. A semantic representation is necessary, which is capable of reflecting such conditions.

The interpretation of intended action at a time point is represented by a so-called time table, i.e. a frame in which both rows of a slot

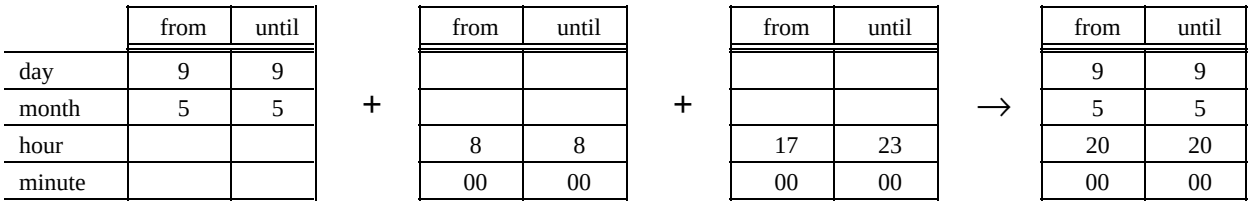

Figure 1: Merging of time constituents

are instantiated, as can be seen in Figure 1. The interpretation of intended action before the time point mentioned implies that the starting point is unknown; thus, only the *until row* is instantiated. The interpretation of the intended action after the time point mentioned is treated analogously, i.e. the *from row* is instantiated while the until row is left open. Surely, every type of time constituent needs a different algorithm to compute the correct time. Such an algorithm often depends on a modality.

Example (1):

Am neunten Mai kann ich Sie um acht Uhr abends treffen.
(On May 9th I can meet you at 8 o'clock in the evening.)

The second step consists of analysis and interpretation on *sentence structure level*. Then, after the treatment of discontinuous time constituents, the time interpretation needs to be tested for consistency and merged into one single representation; i.e. if the system has two or more items of information, let us say the date and the time of day, it can merge them. On top of that, on the sentence structure level, it is often possible to resolve ambiguities. In German single time expressions are often ambiguous; e.g. 'um acht Uhr' can refer to 8 a.m. or 8 p.m.. If sufficient information is given such an ambiguity can be resolved (see example 1). Figure 1 shows the steps of interpretation. Firstly, the date, the time of day and the section of day are interpreted independently: 'abends' (in the evening) is currently defined as 5-11 p.m., and taking 'um acht Uhr' (at 8 o'clock) literally in German means 08.00 a.m.. Secondly, date, time of day and section of day are merged, resolving the ambiguity of 'um acht Uhr' into 08.00 p.m..

As a last step, interpretations of time constituents must be merged on *dialogue level*. If the system has found an utterance with time constituents, the system tells the user whether it can offer a date at the time in question. Doing this it repeats the time it has computed in order to offer an opportunity for verification to the user. A user seldom replies just 'yes' or 'no'. Mostly they add new information about time. In such a case, the system begins to construct a time representation as shown above. Then, it tries to merge the new time table with the old one. If the user confirms the system's interpretation, the merging algorithm is similar to the one at sentence structure level. An explicit negation by the user does not mean inevitably that they negate the whole interpretation, which would compel the system to start anew. Quite often, the negation refers only to the date or to the time of day (see example 2). In spontaneous dialogues the user sometimes also corrects the systems's interpretation without explicit negation. Thus, the system detects nothing but an inconsistency between old and new time interpretation. In both, explicit and implicit cases of correction, the system has to find out, what kind of negation is meant and has to react accordingly. The strategy used to cope with this problem is to select all those unmerged interpretations made on phrase structure level which are consistent with the most recently merged interpretation. The old ones selected provide the basis for a new merging process.

Example (2):

User: Kann ich Sie morgen um acht Uhr treffen.
(Can I meet you tomorrow at eight o'clock.)
System: Ja, morgen habe ich von acht bis zwölf Uhr Zeit.
(Yes, I can meet you tomorrow from eight until twelve o'clock.)
User: Nein, um acht Uhr abends.
(No, at eight o'clock in the evening.)

Furthermore, it appears to be reasonable that the system is able to ask for uncertain or missing bits of information immediately. An uncertain item of information is for example an ambiguous time constituent like 'um zwei Uhr' (at two o'clock). A user seldom wants to make an appointment at 2 o'clock at night, although it might happen. In those cases in which clues for disambiguation are missing the system should be able to ask for exact information. The system should react in the same way, if it has no information about the day.

3. THE SPOKEN DIALOGUE SYSTEM

The linguistic analysis of the spoken dialogue system is based on semantic network representation of linguistic knowledge using the ERNEST formalism [6, 8]. ERNEST enables a uniform representation of all knowledge needed for linguistic analysis. However, the description of linguistic knowledge distinguishes between various levels of abstraction.

- The *hypothesis level* forms the traditional interface between acoustic recognition and linguistic analysis. The speech recognition system incrementally generates word graphs [9, 11] that provide word hypotheses for the linguistic module.
- The *syntactic level* consists of concepts describing the structure of syntactic constituents.
- On the *semantic level* the meaning of syntactic components is described by a framework that uses problem independent noun frames and verb frames of the *deep case theory* [1].
- The *pragmatic level* consists of concepts that represent task specific knowledge. Semantic descriptions are

restricted to their specific use in the task domain of appointment schedule.

- On *dialogue level* the potential utterances of the user and system´s replies and the relationship between them are described. A specific dialogue step of the system depends on the user´s utterance and a system time table with reserved slots.
- A special *time level* is integrated into the knowledge base to model adequately the complexity and variability of time constituents. The special time level provides the syntactic analysis as well as the semantic interpretation.

Our implementation of the system is currently restricted to time constituents needed to agree on an appointment within one week. Thus, time constituents of the type *month* or the type *year* are not integrated yet.

4. EVALUATION

There are various approaches to evaluate speech understanding and dialogue systems [4, 10]. Each of these deals with the performance of whole systems, i.e. to what degree the system is accepted by a human user. But the performance of a complex system depends on the quality of single modules and the interactions between them. Thus, a more detailed evaluation is necessary. We distinguish between qualitative evaluation and quantitative evaluation. By means of *qualitative* evaluation the system's responses in complex dialogues are tested for their plausibility. Examples of such dialogues are given in [2].

The degree of the correctness of the system's interpretation of the user's utterance is investigated by means of *quantitative* evaluation. Usually the quality of a speech recognizer is measured in terms of *word accuracy* based on the Levenshtein distance [7]. For this purpose correctly detected words as well as inserted words and deleted words are taken into consideration: for example, if the utterance 'um vierzehn Uhr' (at 2 p.m.) is partially incorrectly recognized by the system as 'um Pforzheim Uhr' the word accuracy is 66,67%. On the level of semantics nothing can be understood because the relevant word 'vierzehn' is not detected. The following example is more complex: the prepositional phrase 'am Abend' (in the evening) is detected as adverbial phrase 'abends' (in the evening). Although no word is detected correctly in cases like these a robust system can find the correct interpretation.

In order to deal appropriately with such cases our approach combines pattern recognition with linguistic aspects forming a system of measurement consisting of:

- *Word accuracy*, which is the quantitative measurement on the lexical level of processing. It is used to express the quality of a speech recognizer.
- *Constituent accuracy*, which is the quantitative measurement on the syntactic level of processing. It is used to express the quality of a parser.
- *Concept accuracy*, which is the quantitative measurement on the semantic level of processing. It is used to express the quality of an understanding system.

For the usual evaluation based on word accuracy the acoustic signal has to be transcribed. The corpus has to be prepared similarly regarding its syntactic constituents and its semantic concepts, i.e. it has to be tagged.

The relation between word accuracy, constituent accuracy, and concept accuracy becomes clear in the example shown in Figure 2. There is an optimal path regarding the utterance ´am nächsten Dienstag´ (next Tuesday), which is also the reference word sequence used to analyse the accuracy. The optimal path is the sequence ´ab nächsten dienstags´ (nodes 1-3-6-11-12), which is used as test word sequence. Though the word 'Dienstag' is detected this path is not acceptable because it does not lead to the final node of the graph. Only one word of the optimal path, namely 'nächsten', is correctly detected. But neither a syntactic constituent nor a semantic interpretation can be based on 'nächsten'. Two words are substituted: 'ab' (from) instead of 'am' (on) and 'dienstags' (on Tuesday) instead of 'Dienstag' (Tuesday). The dialogue system is able to generate the intended interpretation with 'dienstags' although the syntactic category, noun vs. adverb, is not correct. The divergent semantic of the preposition 'ab' instead of 'am' is not taken into account because the phrase *'ab nächsten dienstags' is rejected by the parser.

An evaluation based on our measurement methods brings about the following results: 33.33% word accuracy, 0% constituent accuracy, and 100% concept accuracy. In contrast assuming that the speech recognizer has detected a pause which links 'Dienstag' (node 10) and the final node of the word graph one can see that the prepositional phrase 'ab nächsten Dienstag' is parsed. In this case we get 66.67% for word accuracy and 100% constituent

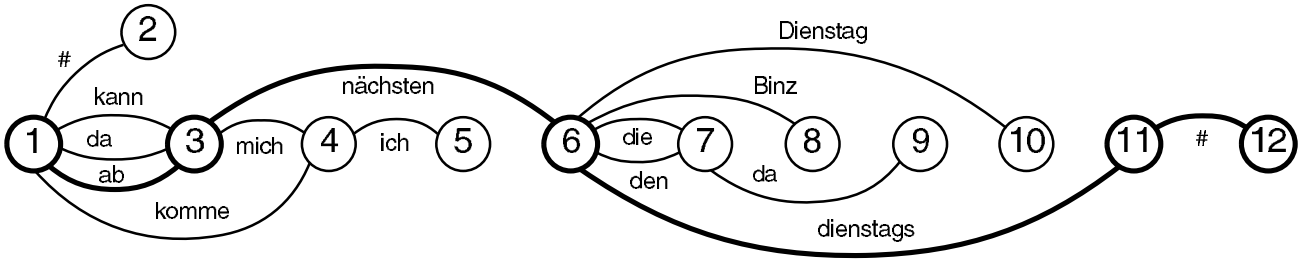

Figure 2: Incrementally generated word graph of the utterance 'am nächsten Dienstag' (next Tuesday).

Accuracy	ACC (accuracy)		SUB (substitutions)		DEL (deletions)		INS (insertions)	
	SRS	TIM	SRS	TIM	SRS	TIM	SRS	TIM
Word	81.12	58.67	9.69	11.73	5.61	27.55	3.57	2.04
Constituent	88.12	80.20	0.99	3.96	7.92	10.89	2.97	4.95
Concept	80.77	70.19	8.65	14.42	9.62	11.54	0.96	3.83

Table 1: Accuracy of the speech recognition system (SRS) and the time component (TIM)

accuracy. As mentioned above, the semantic interpretation differs ('am' vs. 'ab') and the concept accuracy is 0%. Anyhow, our approach of evaluating reflects complex relations between the different components of systems and makes a detailed investigation possible. In addition, it permits the evaluation of a single component of a system as will be shown in the next section.

5. RESULTS

The speech understanding system will not be evaluated as a whole. Only one component will be discussed here, namely the modelling of time constituents (for an elaborate evaluation see [2]). In respect to time constituents, word accuracy of the speech recognizer is 81.12%; after the linguistic interpretation it is 58.67%. There is obviously a big difference. In contrast to this, the difference concerning the constituent accuracy is very small: speech recognizer 88.12% vs. linguistic analysis 80.20%. These data reveal that the head of a linguistic phrase is mostly detected by the speech recognizer. And they reveal that time constituents can be parsed correctly. Unfortunatly, if the head of a constituent is not detected by the speech recognizer, no word depending on the head can be parsed by the time component. Thus, these words are deleted. This is the reason for the high rate of word deletions (27.55%) after the linguistic analysis. The difference between speech recognizer and time component is moderate concerning constituent deletions and concept deletions. The concept accuracy resembles the constituent accuracy: speech recognizer 80.77% vs. linguistic interpretation 70.19%. These data can be interpreted as follows: if the head of a phrase is detected correctly by the speech recognizer the linguistic analysis will find the correct interpretation.

In respect to these data and the good dialogue performance of the system it appears to be necessary to focus our future efforts on improving the speech recognizer system.

6. REFERENCES

[1] C. Fillmore. A case for case. In E. Bach and R. T. Harms, editors, *Universals in Linguistic Theory*, pages 1-88. Rinehart and Winston, New York, 1968.

[2] B. Hildebrandt. *Struktur und Bedeutung temporaler Konstituenten in einem sprachverstehenden Dialogsystem*. Infix, Sankt Augustin, 1995.

[3] B. Hildebrandt, G. A. Fink, F. Kummert, and G. Sagerer. Modelling of time constituents for speech understanding. In *Proc. European Conf. on Speech Communication and Technology*, Berlin, 1993.

[4] L. Hirschman, M. Bates, D. Dahl, W. Fisher, J.Garofolo, D. Pallet, K. Hunicke-Smith, P. Price, A. Rudnicky, and E. Tzoukermann. Multi-site Data Collection and Evaluation in Spoken Language Understanding. *Human Language Technology. Proc. of a ARPA Workshop*, 19-24, 1993.

[5] F. Kummert. *Flexible Steuerung eines sprachverstehenden Systems mit homogener Wissensbasis*. Infix, Sankt Augustin, 1992.

[6] 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.

[7] K.-F. Lee. *Automatic Speech Recognition: the Development of the SPHINX System*, Kluwer Academic Publisher, Boston, 1989.

[8] H. Niemann, G. Sagerer, S. Schröder, and F. Kummert. ERNEST: A Semantic Network System for Pattern Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:883-905, 1990.

[9] M. Oerder, and H. Ney. Word Graphs: An efficient interface between continuous speech recognition and language understanding. *Proc. of the Inter. Conf. on Acoustics, Speech and Signal Processing*, 119-122, 1993.

[10] P. Price, L. Hirschman, E. Shriberg, and E. Wade. Subject-based Evaluation Measures for Interactive Spoken Language Systems. *Speech and Natural Language. Proc. of a DARPA Workshop*, 34-39, 1992.

[11] H. Rautenstrauch, G. A. Fink, F. Kummert, and G. Sagerer. Schritthaltende Generierung von Wortgraphen. *Fortschritte der Akustik. Proc. DAGA '94*. 1261-1264, 1994.