# A HYBRID APPROACH TO SPOKEN DIALOGUE UNDERSTANDING: PROSODY, STATISTICS AND PARTIAL PARSING

*Elmar Nöth*[1] *Manuela Boros*[2] *Jürgen Haas*[1] *Volker Warnke*[1] *Florian Gallwitz*[1]

[1] *Chair for Pattern Recognition*
*University of Erlangen-Nuremberg*
*Martensstrasse 3*
*D-91058 Erlangen, Germany*
*e-mail: noeth@informatik.uni-erlangen.de*

[2] *Bavarian Research Center for*
*Knowledge Based Systems (FORWISS)*
*Am Weichselgarten 7*
*D-91058 Erlangen, Germany*
*e-mail: boros@forwiss.uni-erlangen.de*

## ABSTRACT

Linguistic processing in spoken dialogue systems has to be robust against a large number of phenomena such as recognizer errors, spontaneous speech phenomena and out-of-vocabulary (OOV) words. A commonly used solution to this problem is partial parsing, that aims at detecting only parts of sentences/utterances that are vital for the respective task of the parser. In our paper we present a framework for robust linguistic processing in our spoken dialogue system EVAR for train timetable information. The linguistic processor combines partial parsing with prosody and statistical concept prediction. Parsing is restricted to the detection and analysis of those parts of an utterance that are crucial for its understanding by the system. In order to accomplish this task most efficiently, the parser operates not only on word lattices as delivered by the recognizer, but also on prosodic information and statistical concept prediction.

## 1. INTRODUCTION

Linguistic analysis in spoken dialogue systems has to cope with two main problems. First, spontaneous speech very often is fragmented, ungrammatical or exceeds the system's boundaries (e.g. out-of-vocabulary words). Second, word recognition in spoken dialogue systems produces errors, thus rendering utterances ungrammatical on the syntactic as well as the semantic level. In order to cope with these problems, methods of robust parsing have been established. E.g. partial parsing methods restrict syntactic analysis to sub-units of utterances only, therefore reducing the above mentioned problems to these sub-units. Different methods of partial parsing have been successfully employed in spoken dialogue systems, such as the systems described in [1] and [2].

Linguistic processing in spoken dialogue systems usually operates on scored word hypotheses as delivered by the word recognizer, and, in some cases, semantic predictions of the system's dialogue manager on the contents of the actual user utterance. However, partial parsing in dialogue systems becomes even more efficient if more sophisticated sources of information, beyond acoustically scored word graphs and dialogue predictions, can be used to guide the linguistic processor. In our work we concentrated on the integration of prosodic information, extracted from the speech signal, and statistically detected semantic concepts in utterances as additional support for the parser, thus resulting in a hybrid approach to language understanding.

Partial parsing reduces the syntactic and semantic analysis of an utterance to the analysis of specific sub-units, whose definition in speech understanding is based rather on semantic than on purely syntactic criteria. In our approach, these units correspond to semantic concepts (e.g. time, date, source or target location for train timetable inquiries), that are vital for the correct interpretation of the utterance in the actual domain. The parser will identify and analyze these concepts, assigning a semantic representation to each.

For each concept and its possible surface realizations grammar fragments are defined, that may be used by the parser upon request. The parser is guided by prosodic information on phrase boundaries and phrase accents, telling it, where to start the partial analysis. Statistical concept detection provides information on which semantic concepts are included by the actual utterance, thus helping the parser to choose the appropriate grammar fragments. The use of grammar fragments has two major advantages: the danger of false alarms in parsing is drastically reduced, as well as the time consumed by the parser and the efforts for grammar development.

In the following sections we will first introduce the methods and modules used for the extraction of prosodic information (section 2) and semantic concept detection (section 3) before describing the partial parser and grammar fragments in section 4. Section 5 will show preliminary results of experiments run with the integrated system, that prove the increase in efficiency and robustness of the hybrid approach. Conclusions will be drawn in section 6, where also an outlook on further work will be given.

## 2. PROSODY IN WORD GRAPHS

The use of prosodic information for spoken dialogue systems becomes more and more important. In the VERBMOBIL project [4] prosodic information was successfully used in a speech understanding systems for the first time. We use neuronal networks (NN), which use prosodic features derived from the pitch-contour, the energy-contour and word durations as input and classify phrase boundaries, phrase accents and sentence mood [6].

Here we want to use this prosodic information to determine the salient regions in a phrase. These regions are the parts of a sentence, which hold the most important content words e.g. time expressions and locations and which most of the time are 'in focus', i.e., are the carrier of the focal accent. To get information for those regions, we use a NN trained on a part of the VERBMOBIL database with a topology of 276 nodes in the input layer (one node for each used prosodic feature), one hidden layer with 60 nodes and an output layer with 2 nodes (a word is accentuated $\mathsf{A}$ or not $\neg\mathsf{A}$). Using $Score(A \mid w)$ and $Score(\neg A \mid w)$ from the output nodes of the NN for each word $w$ we can estimate the probability $P(A \mid w)$ by using the following formula

$$P(\mathsf{A} \mid w) = \frac{Score(\mathsf{A} \mid w)}{Score(A \mid w) + Score(\neg A \mid w)}.$$

This work was partailly funded by the DFG (German Research Foundation) under contract number 810 830-0 and by the German Federal Ministry of Education, Science, Research and Technology (*BMBF*) in the framework of the VERBMOBIL Project under the Grant 01 IV 701 K5. The responsibility for the contents of this study lies with the authors.

Now we are able to estimate the probability $P(A \mid w)$ for each word of an utterance and we decide for a focused region by using a threshold. In Figure 1 an example is given for a German utterance.

The estimation of stressed regions in a given utterance offers two possible ways to use this knowledge in combination with the parser:

1. we rank the regions by their prosodic scores and offer the ranking list to the parser, which has to find the best expression for the given context
2. we get a list of possible expressions from the parser and disambiguate them using the prosodic score from the NN.

Both ways can efficiently be used to find the best expression the parser is searching for in the context the concept predictor (3) has estimated. The first way seems to be the better one if working on word hypotheses graphs, because the parser only has to search in the best scored paths and thus search effort is smaller.

### 2.1. Experiments

In this section we present results for determining stressed words for different dialogue acts (see [5]) in the Verbmobil database using the above described NN. In Verbmobil there are 42 illucotionary dialogue acts defined which are grouped into 18 dialogue act classes used for template based translation. For these classes we estimated the most frequent stressed words of a subset of the Verbmobil database using the above described method. For this approach only those words are considered, whose stress probability exceeds a threshold of 0.8 and that were seen stressed in more than 80% of their occurrences. In Table 1 the ten most often seen automatically estimated stressed words for all dialogue act classes together are shown. Table 2 shows the five most often seen detected stressed words for the most frequent dialogue act classes SUGGEST and ACCEPT. In both tables the words are ranked by their frequency of occurrence in the observed data set.

| $P(A \mid w) > 0.8$ | | |
|---|---|---|
| *Rank* | *% stressed* | *word (translation)* |
| 1 | 88.57 | Freitag (Friday) |
| 2 | 82.69 | Wiederhören (bye) |
| 3 | 84.31 | Donnerstag (Thursday) |
| 4 | 90.91 | Samstag (Saturday) |
| 5 | 95.35 | neunzehnten (19th) |
| 6 | 81.82 | August (August) |
| 7 | 96.15 | vierundzwanzig. (24th) |
| 8 | 87.50 | achten (8th) |
| 9 | 86.96 | wunderbar (marvellous) |
| 10 | 100.00 | sechsundzwanzig. (26th) |

Table 1. Automatically determined stressed words for all dialogue acts.

The results from Tables 1 and 2 show, that the detection of content words of an utterance is possible through determining the stressed words. This fact is very important for the use of this method to estimate the focused regions by only using acoustic features to decide for semantically important information.

## 3. STATISTICAL CONCEPT DETECTION

As a second additional information source for the hybrid partial parsing, we examine a statistical approach using $n$-gram language models as semantic concept predictors. The model has to decide about the occurrence of special semantic concepts in word chains. We prove its usability on a corpus collected with the above mentioned information retrieval system containing the utterances used for the grammar development. We present here two predictors one for time expressions and one for date expressions. The predictor should be able to decide whether there appears such a time/date expression in an utterance or not.

| ACCEPT | | |
|---|---|---|
| $P(A \mid w) > 0.8$ | | |
| *Rank* | *% stressed* | *word (translation)* |
| 1 | 100.00 | einverstanden (ok) |
| 2 | 100.00 | Ordnung (alright) |
| 3 | 100.00 | wunderbar (marvellous) |
| 4 | 85.71 | Freitag (Friday) |
| 5 | 85.71 | frei (free) |

| SUGGEST | | |
|---|---|---|
| $P(A \mid w) > 0.8$ | | |
| *Rank* | *% stressed* | *word (translation)* |
| 1 | 82.22 | Montag (Monday) |
| 2 | 87.80 | Freitag (Friday) |
| 3 | 83.33 | Donnerstag (Thursday) |
| 4 | 82.76 | Mittwoch (Wednesday) |
| 5 | 93.10 | Samstag (Saturday) |

Table 2. Automatically determined stressed words for dialogue acts ACCEPT and SUGGEST.

### 3.1. Language Model Predictor

The language model we use computes estimations for the occurrence of a word $w_i$ under the assumption of its predecessor words $w_{i-1}, \ldots, w_{i-n+1}$. To smooth the probabilities we combine the $n$-gram probability with smaller $n$-grams. There exist a lot of possible interpolation techniques for those probabilities – a very common one is the linear interpolation. Another interpolation method which performs quite well for the prediction task is the rational interpolation (cf. [7] for details) where the probabilistic model looks like this:

$$p(w|v) = \frac{\sum_{i \in \mathcal{I}} \lambda_i \cdot g_i(v) \cdot p_i(w|v)}{\sum_{i \in \mathcal{I}} \lambda_i \cdot g_i(v)} \qquad (1)$$

$p_i$ gives the probability for observing word $w$ based on some appropriate portion of the sentence history $v$, $g_i(v)$ is a history dependent weight function and the denominator operates as normalizing term.

If we now want to use $n$-gram language models as a semantic concept predictors we have to claim for a word chain $\mathbf{w}$ whether the concept we are looking for is expressed in $\mathbf{w}$ or not. For this purpose we build two different language models. The first one is trained with word chains expressing the semantic concept and the second one with the utterances not expressing it. During analysis we compute the two scores for the incoming word chain – when using word graphs we choose the best word chain in the graph – and we decide for the higher probability. The results presented in the following section 3.2 work on this classification.

A more detailed analysis is possible when we split the training data into three different sets, which are used to train language models. The first set comprises all word chains where the semantic concept does not appear, the second one all utterances where only the concept is expressed and no other semantically relevant information is present and the third set all word chains where the interesting semantic concept appears along with additional information. The decision rule again decides for the highest probability of the three scores. If we combine the second and the third part we have a probability estimation for the same partition as above. Results on that task can be found in [8].

### 3.2. Experiments

In the experiments we want to prove if the language models are able to be used as predictors for semantic concepts. For training and test purposes we use 20406 sentences collected with the system EVAR for train timetable informa-

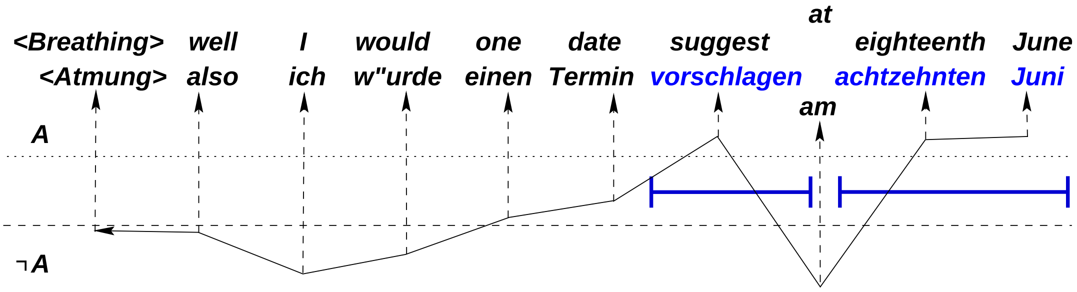

Figure 1. A German sentence from Verbmobil with probability $P(A \mid w)$ for each word $w$ and the two estimated focused regions with word to word translation.

tion. Here we concentrate on the detection of time and date expressions in the examined word chains. Therefore we mark for each sentence, based on the transliteration of the utterance, whether the semantic concept *time* (TIME) is present or not (NOTIME) and equivalently we do the labeling for *date* (DATE vs. NODATE).

The available data is split 2/3 to 1/3 for training and test purposes. The number of sentences for each class is presented in Table 3. Since a word sequence from the test set might have been used by a different speaker from the training set (i.e. if the system asks *"where do you want to leave from?"* different users answered with the same city name expression), the column 'test ≠ train' gives the number of sentences from the column 'test' that were not observed during training.

| | train | test | test ≠ train |
|---|---|---|---|
| TIME | 2366 | 1145 | 759 |
| NOTIME | 11238 | 5657 | 2025 |
| DATE | 2482 | 1232 | 763 |
| NODATE | 11122 | 5570 | 2025 |

Table 3. Number of word chains for training and testing

For our experiments we clustered the observed words automatically in categories. In order to find an optimal working point for the predictor we tested different sizes of category systems and different interpolation strategies and context lengths. The results we report here are obtained with a system consisting of 25 categories, using rational interpolation and a context of three words. The 'Semantic Concept Predictor' results are shown in Table 4 as confusion matrices. From these we see that our language model approach to the prediction task performs quite well and could therefore be used as a predictor for the semantic concept analysis. For the problem of detecting time expressions we obtain a recognition rate of 95.6%; if we measure the performance of the models as being a time expression spotter we get a recall of 98.4% and a precision of 80.0%. For date expressions we have a recognition rate of 95.7% and a recall of 96.7% and a precision of 82.4%.

| | TIME | NOTIME |
|---|---|---|
| TIME | 1127 | 18 |
| NOTIME | 282 | 5375 |

| | DATE | NODATE |
|---|---|---|
| DATE | 1191 | 41 |
| NODATE | 254 | 5316 |

Table 4. Confusion Matrices for time and date expressions

## 4. PARTIAL PARSING WITH GRAMMAR FRAGMENTS

The partial parser described here is an agenda driven chart parser, operating as an island parser (cf. [3]). The island parsing strategy allows to start the parsing process at arbitrary edges in the graph (*islands*) and expand these islands successively to the left and right until a full-spanning edge has been generated. The initial islands have to be chosen carefully so that relevant parts of the utterance will be analyzed as soon as possible. Island parsing has proven to be very robust against spontaneous speech phenomena.

Our approach restricts the linguistic analysis to the analysis of semantic concepts. Lexicon and grammar of the parser therefore only need to cover the relevant syntactic realizations for each concept, thus resulting in several grammar fragments, rather than one full grammar. Island parsing on the basis of these grammar fragments means, that each of the maximal islands, the parser will find, corresponds to one relevant part of an utterance, and that these islands will not be combined to one single edge but remain separately. We coded a grammar fragment for each of the semantic concepts in terms of a context-free phrase structure grammar.

### 4.1. Integration of Further Information

As for each semantic concept there exists a corresponding grammar module, the predictions on the occurrence of concepts in user utterances can be used to guide the parsing process. This is done by using only those grammar fragments for parsing, that correspond to semantic concepts predicted by the concept detection module. In parallel to the recognizer output, the predicted concepts are passed to the parser that will take into account only the appropriate grammar fragments for the next parse.

In order to further improve efficiency of the parsing process, the use of prosodic information is included into the parsing process. Each word hypothesis comes with a prosodic accent score, in addition to the usual acoustic score. This information can then be used for choosing the initial islands: only those hypotheses, that are marked to carry accent are chosen for initial islands. We made experiments with different thresholds of accent scores to determine the best value. In section 5 results for different thresholds are shown.

### 4.2. The Parsing Algorithm

Before starting the parsing process, the chart is initialized with the lexical entries for the hypotheses in the word graph. As for each parse not every grammar fragment is used, many hypotheses are unknown, thus leaving gaps in the chart. In parallel to the chart, two agendas are initialized that will guide the flow of the analysis. The first agenda (*seed agenda*) contains all hypotheses that will serve as initial islands. The second agenda (*non-seed agenda*) contains the remaining hypotheses and is only used as fall-back. Each hypothesis , whose accent score exceeds a given threshold, is inserted into the seed agenda,

| | NIL | lex. | 100%<br>Pred. | real<br>Pred. | Pros.<br>0.5 | Pros.<br>0.7 | Pros.<br>0.9 | 0.5+<br>100% | 0.7+<br>100% | 0.9+<br>100% | 0.5+<br>real | 0.7+<br>real | 0.9+<br>real |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Parses time | 871 | 346 | 123 | 136 | 219 | 201 | 151 | 120 | 119 | 108 | 124 | 123 | 109 |
| date | 871 | 292 | 141 | 169 | 244 | 233 | 202 | 130 | 125 | 115 | 133 | 127 | 114 |
| Seeds time | 2761 | 836 | 420 | 431 | 439 | 377 | 241 | 281 | 251 | 181 | 285 | 255 | 182 |
| date | 2761 | 605 | 416 | 416 | 447 | 414 | 340 | 309 | 284 | 240 | 308 | 282 | 237 |
| D (time/date) | 1/0 | 1/0 | 1/0 | 1/0 | 1/0 | 3/0 | 10/5 | 1/0 | 1/2 | 10/5 | 1/0 | 1/2 | 10/5 |
| I (time/date) | 2/2 | 2/2 | 3/2 | 3/1 | 2/2 | 2/2 | 2/4 | 3/2 | 3/2 | 3/3 | 3/1 | 3/2 | 4/2 |

Table 5. Number of necessary parses and possible island seeds with different levels of information sources and the number of deletions (D) and insertions (I) for *date* and *time*

the remaining ones to the non-seed agenda. Within both agendas, entries are sorted according to their acoustic score. Agenda entries may not only be the initial lexical entries (*seed entries*) but also pairs of chart edges (*non-seed entries*) that comprise pointers to two adjacent chart edges and a list of grammar rules that might combine these two edges to a new one. Until the seed agenda does not contain any more entries, the following steps will be performed:

1. Take best scored agenda entry $E$ from seed agenda.
2. If $E$ is a seed entry go to 3, else go to 4.
3. For each adjacent chart edge to $E$ look for rules that can be applied to both and generate an agenda pair for both and sort it into seed agenda; go to 1.
4. For each grammar rule in $E$: apply this rule to both edges, insert new edge (if rule can be applied) into chart, generate new agenda pairs for this new edge and insert them into seed agenda; go to 1.

This is done for each of the predicted semantic concepts using the respective grammar fragments. Only in case no valid semantic representation for a concept can be found in the chart after parsing, the process is re-started with the non-seed agenda.

## 5. EXPERIMENTS AND RESULTS

First experiments were done for the semantic concepts `time` and `date`. The respective grammar fragments comprise 177 lexical entries and 36 grammar rules for `time` and 418 lexical entries and 16 grammar rules for `date`. The numbers in Table 5 denote the following: we examine the two concepts *time* and *date* and for our test database we count for each set of integrated knowledge how many sentences we have to parse i.e. how often the parser is applied. This number is given in the column Parses. Additionally we count how many words are on the initial seed agenda (column Seeds) as all these words must be considered when applying a grammar fragment. The used information set is composed from the following parts. As one information source we have the lexicon which is always applied except in the first column (NIL) where no knowledge is used. As prediction (Pred.) we can either use the reference which gives 100% prediction rate or our LM classifiers giving a real prediction rate. For the prosodic scores we define a threshold and all words whose accentuation score is higher than that threshold – which we choose 0.5, 0.7 and 0.9 for our tests – are put on the seed agenda. In the last two columns we give the numbers of deletions (D) and insertions (I) the analysis does for the two concepts.

In Table 5 we see that the more knowledge we use for our hybrid approach the less parses we have to perform and the less island seeds have to be considered so analysis time will be shorter. The first column with the results when no knowledge is used corresponds to an analysis using a full grammar. Through applying grammar fragments the numbers of column two (lexicon) are obtained and so on. The big advantage of the hybrid approach is that we can perform the linguistic analysis faster without increasing the error rate. The number of deletions and insertions changes only significantly when we turn the prosodic threshold to 0.9 but even then it is not intolerably high. Keep in mind that these results are produced when we use only the seed entries. The errors can be reduced if we use a fall-back strategy that tells us to look also on the non-seed entries if the first analysis of seed entries did not succeed.

## 6. CONCLUSION AND FURTHER WORK

In this paper we presented a hybrid approach to speech understanding using grammar fragments along with statistical and prosodic information. We have shown that we reduce the necessary analysis and therefore the required time drastically without increasing the error rate.

The nest steps we plan to do is the implementation of the fall-back strategy on non-seed entries, we have to rerun our experiments on the recognized word chains and finally we want to use word graphs, where we expect much more gain from prosody as a lot of unsuccessful paths are eliminated very fast using the prosodic scores.

## REFERENCES

[1] D. Albesano, P. Baggia, M. Danieli, R. Gemello, E. Gerbino, C. Rullent. A Robust System for Human-Machine Dialogue in Telephony-Based Applications. In *International Journal of Speech Technology*, Vol.2, pp. 101-111, 1997.

[2] H. Aust, M. Oerder, F. Seide, V. Steinbiss. The Philips automatic train timetable information system. In *Speech Communication*, Vol.17, pp. 249-262, 1995.

[3] K. Mecklenburg, P. Heisterkamp, and G. Hanrieder. A robust parser for continuous spoken language using Prolog. In *Proceedings of the Fifth International Workshop on Natural Language Understanding and Logic Programming (NLULP 95)*, pages 127–141, Lisbon, Portugal, 1995.

[4] T. Bub and J. Schwinn. 1996. Verbmobil: The Evolution of a Complex Large Speech-to-Speech Translation System. In *Int. Conf. on Spoken Language Processing*, volume 4, pages 1026–1029, Philadelphia, 1996.

[5] S. Jekat, A. Klein, E. Maier, I. Maleck, M. Mast, and J. Quantz. 1995. Dialogue Acts in Verbmobil. Verbmobil Report 65, April 1995.

[6] Andreas Kießling. 1997. *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung.* Berichte aus der Informatik. Shaker Verlag, Aachen.

[7] E.G. Schukat-Talamazzini, F. Gallwitz, S. Harbeck, and V. Warnke. Rational Interpolation of Maximum Likelihood Predictors in Stochastic Language Modeling. In *Proc. European Conf. on Speech Communication and Technology*, pages 2731–2734, Rhodes, Greece, 1997.

[8] J. Haas, E. Nöth, H. Niemann. Semantigrams – Polygrams Detecting Meaning . In *Proc. 2nd SQEL Workshop on Multi-Lingual Information Retrieval Dialogs*, pages 65–70, Pilsen, Czech Republic, 1997.