USING PHRASE ACCENT INFORMATION FOR DIALOGUE ACT RECOGNITION IN SPONTANEOUS GERMAN SPEECH

Matthias Nutt[1], *Anton Batliner*[2], *Volker Warnke*[2] *and Elmar Nöth*[2]
nutt@forwiss.uni-erlangen.de

[1]Bavarian Research Center for Knowledge-Based Systems (FORWISS)
Knowledge Processing Research Group
Am Weichselgarten 7, D-91058 Erlangen, Germany.

[2]Friedrich-Alexander-University of Erlangen-Nuremberg
Chair for Pattern Recognition
Martensstraße 3, D-91058 Erlangen, Germany

ETRW on Dialogue and Prosody
Veldhoven, The Netherlands
September 1-3, 1999

ABSTRACT

This paper describes an approach in which phrase accent information is used for dialogue act recognition in German spontaneous speech. This application is an example of how automatically computed prosodic information can be used in automatic speech recognition. Usually the important intention conveyed by an utterance is found in the focused area, which is often accentuated. When all the words of an utterance are used for dialogue act classification, the best result is achieved only if all probabilities (e.g. of n-grams) are known. In real life applications this not the case. Because utterances can be very similar to one another, but belong to different dialogue act classes, it may be possible to distinguish the classes on the basis of characteristic words. For this reason dialogue act classification is often based on keyword detection. The selection of keywords is crucial. Better recognition relies on better chosen keywords. This paper shows how keyword selection can be improved by using two additional information sources: lexical POS information and prosody. POS and prosodic information is used to build subsets of the vocabulary to improve recognition. Experiments are conducted on a sub-sample of the VERBMOBIL corpus. The aim is to distinguish between four dialogue act sub-classes of the general class SUGGEST.

1. INTRODUCTION

The research reported in this paper is embedded in the VERBMOBIL project, that combines speech technology with machine translation [12]. The aim is to develop a prototype for the translation of spontaneous speech in face-to-face dialogues about business appointments. To this effect, a large database of spontaneous German speech has been collected. A subset has been used for the experiments discussed here.

Prosodic information about boundaries is a powerful tool to reduce the number of possible readings of utterances. This is especially helpful for parsing the utterance [4, 5]. In this paper we, report on experiments with the other type of prosodic information which is often used, namely accents.

The essential topic of an utterance is called the focus. The focus can be marked by the syntactic structure of a sentence, but is normally marked by accentuation as well. So the accentuated words contain parts of the intention. The underlying intentional meaning of an utterance can be mapped onto different 'dialogue acts' (see [3, 10, 11] for details).

In our approach, the position of the word accent in a word is trivially provided by lexicon look-up. We are interested in which words of a phrase are accentuated. The most prominent word in a phrase is called **phrase accent** in our terminology (see [1] for a discussion on this). We are not interested in the specific phonetic form of accents, (e.g. loudness, pitch accent and/or speaking rate) (see [4]).

Different user utterances can be quite similar to one another, but express different dialogue acts, as in the following two examples: *Let us meet at ONE (o' clock)*, *Let us meet at HOME*. The first belongs to SUGGEST-SUPPORT-DATE, the second to SUGGEST-SUPPORT-LOCATION. But function words (FW) are also characteristic of dialogue act classes. Consider "*It is possible at one.*" as an example of SUGGEST-SUPPORT-DATE. If the FW *not* is added, the utterance belongs to SUGGEST-EXCLUDE-DATE. Preliminary experiments implied that some words or word classes are characteristic of specific dialogue act classes. The investigations further confirm the importance of accentuated content words (CW).

Other researchers also report improvements using "short phrases, that appear frequently in dialogues and convey a significant amount of discourse information" [9], called cue phrases. They can be collected automatically by n-grams. The disadvantage of this method is that no prosodic information is taken into account. Using however prosody for dialogue act classification is experimentically proven to be useful; e.g. to distinguish questions from statements [11] (Example: *At one o'clock?* vs. *At one o'clock!*).

Which cue phrases or words are characteristic of a certain dialogue act class certainly depends on the domain. In [8], dialogue act prediction is done by

This work was funded 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.

Dialogue act class	# Train	# Test
SUGGEST-EXCLUDE-DATE	707	71
SUGGEST-SUPPORT-DATE	4281	366
SUGGEST-SUPPORT-DURATION	86	16
SUGGEST-SUPPORT-LOCATION	120	12
SUGGEST-EXCLUDE-DURATION	4	0
SUGGEST-EXCLUDE-LOCATION	0	0

Table 1. Data material taken from VERBMOBIL

a keyword spotter. The keywords for each dialogue act class have been automatically computed by a Keyword Classification Tree [6]. Consider the words 'home', 'office' or 'hotel' in the VERBMOBIL domain. These are examples of CW which belong with a high probability to class SUGGEST-SUPPORT-LOCATION. In the *Let us meet ...*-examples above, the last word is accentuated. So we suggest that accentuation information can be helpful for dialogue act classification, especially to collect keywords or cue phrases. Disambiguation of dialogue acts based on accentuated words should be possible.

The experiments presented below show how discriminative these words are. In the following we describe the speech material, its annotation and the results of the experiments conducted.

2. CORPUS COLLECTION AND ANNOTATION

The speech material is taken from the VERBMOBIL corpus [12]. It consists of all parts of utterances belonging to the dialogue act class SUGGEST, the most often prominent dialogue act class in VERBMOBIL. We use just this class, because the other classes do not provide sufficient data for the reliable estimation of probabilities.

We distinguish the sub-classes shown in Table 1. The Table also shows the number of phrases in the test and train corpus for each class. Due to a lack of data, we cannot take into account the classes SUGGEST-EXCLUDE-DURATION and SUGGEST-EXCLUDE-LOCATION, although they also belong to the general class SUGGEST.

In the first stage, a multi layer perceptron (MLP), trained on hand-labeled data, computes an **accentuation score** for each word. The MLP uses 276 prosodic features, described in [4]. Because of the design of the MLP training and an additional normalisation procedure, this score can be interpreted as a **probability**. Next, a proportional relation between the probability of being accentuated and the strength of the accentuation is assumed. This is a critical point, because the accentuation probability is not the accentuation strength. But the **accentuation strength** cannot be computed and a high accentuation probability is often an indicator for a strong accent and vice versa. So the word with the maximum accentuation probability should be the most accentuated word of a phrase, the phrase accent.

Additionally, the part-of-speech (POS) of each word is defined and coded in a special dictionary. In this paper only the two main POS classes, CW and FW, are considered. FW include articles, pronouns, interjections, modal verbs, and copula. CW are e.g. verbs, nouns, names, or letters (see [1, 2, 7] for details).

3. EXPERIMENTS AND RESULTS

In this section the most important results of a set of 12 experiments are presented. The classification task is performed by language models. We use unigrams to be able to reflect the influence of isolated words. We also benefit from another feature of unigrams. The word order does not influence the recognition. This allows reordering the word chain without negative consequences and thus simplifies data preparation for the experiments.

For each SUGGEST class a unigram is trained. The training database consists of transliterations of all utterances, including all words of the spoken word chain belonging to this class. For the classification experiments, each word chain of the transliterated utterance of the test database is reduced. Which word will be removed from the word chain depends on the different experiments. There are two orthogonal reasons why a word is dropped. The first criterion is based on the accentuation probability. As we will see in the first experiments, a word is neglected if its accentuation probability is above or below a given threshold. This criterion can be regarded as a very simple focus detector. There are other experiments which use only the first n words of a list, sorted in ascending or descending order by accentuation probability.

The second reason to remove a word is based on the POS. Because the POS for each word is available, this information can be used during the data preparation. Each experiment was conducted on the whole test data set, only on the CW and only the FW of the test data set. For the last two cases, the utterances are, of course, shortened, because the words of a specific POS will be deleted. This allows to establish the influence of each of the two word classes (CW/FW) on the classification task. The following sections describe the experiments in detail and report the recognition rates achieved. But first the naming convention for the experiments is presented.

To easily distinguish the different experiment setups, they are named using the following scheme: The first characters determine which words are used: ALL words, only CW, or only FW. The last character indicates if a threshold (Θ) or the number of words (#) is used for an experiment. A '+' or '-' between these characters means that we are using accentuated or unaccentuated words, respectively. This means in the case of threshold experiments that the threshold is a lower or an upper bound, which results in using only those words with accentuation probability above or below Θ. If the number of words is used, a '+' or '-' means that the word list is sorted in descending or ascending accentuation probability order, which results in using the n most accentuated or the n least accentuated words. Consider CW+Θ for example. In this experiment only content words (CW) with an accentuation probability above a threshold are used for the classification task.

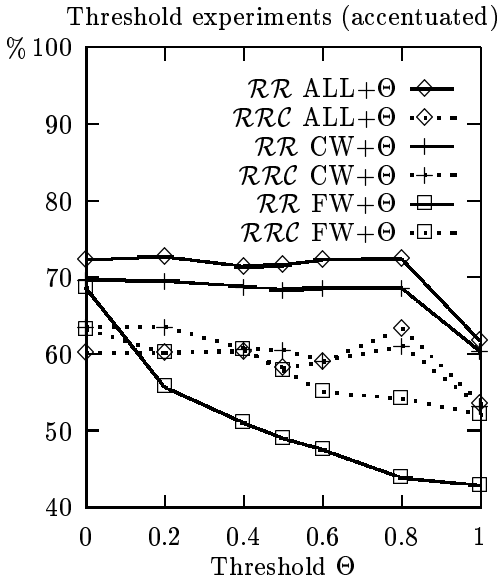

Figure 1. Recognition rates using selected accentuated words. Their selection depends on their accentuation probability and the word class (ALL, CW, FW).

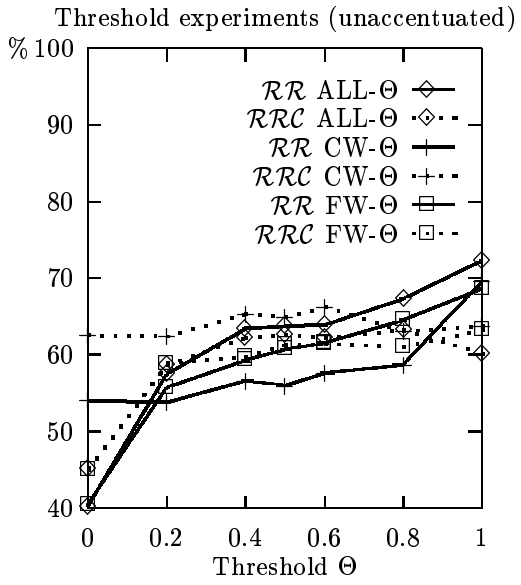

Figure 2. Recognition rates using selected unaccentuated words. Their selection depends on their accentuation probability and the word class (ALL, CW, FW).

3.1. Threshold experiments

The first experiments examine how the accentuation probability influences dialogue act recognition. For the first experiments, only words with an accentuation probability above a threshold Θ are considered. So the threshold Θ controls which words are considered as accentuated. Only the accentuated words remain in the word chain. If there is no word in the utterance with this attribute, the word with the maximum accentuation probability (the most accentuated word) is used instead. $\Theta = 1.0$ means that only the most accentuated word, the phrase accent, remains, and $\Theta = 0.0$ means that no word is removed from the word chain. Because no additional information is needed for the last experiment with $\Theta = 0.0$, it is the baseline for all others. Figure 1 shows the recognition results for this and two other experiments depending on Θ. The solid lines in figure 1 show the progression of the recognition rates ($\mathcal{RR}$) for each experiment. The dotted lines display the average of the classwise recognition rates ($\mathcal{RRC}$).

In general the recognition rates ($\mathcal{RR}$) for the first experiment ALL+Θ remain quite constant. But the best recognition is achieved for $\Theta = 0.8$. The $\mathcal{RR}$ is quite the same as in the baseline experiment, but the $\mathcal{RRC}$ is 3% better than in the baseline experiment. This shows that taking only words with high accentuation probability (strongly accentuated words) into account for the classification task increases $\mathcal{RR}$ and $\mathcal{RRC}$. Note, as well, the difference for $\mathcal{RR}$ between $\Theta = 0.8$ and $\Theta = 1.0$. Using only the most accentuated word is suboptimal. The average number of words remaining at $\Theta = 0.8$ is 3, whereas the average number of words for $\Theta = 0.0$ is 9. It can be concluded that a better result is achieved using fewer words. All recognition rates are above the chance level of 25%.

For the next series of experiments only those words belonging to the selected POS class (CW/FW) remain. Again the accentuation probabilities of these words have to be above Θ, figure 1 contains these results too. In experiment CW+Θ only CW are chosen, in experiment FW+Θ only FW. The recognition rates of experiment CW+Θ are almost parallel to the first experiment, the deviation is rather small. This shows the importance of CW for the classification task. Remarkable is the course of the $\mathcal{RR}$ for exp. FW+Θ. The $\mathcal{RR}$ is always below the others and drops already for $\Theta = 0.2$.

The next experiments are similar to the experiments above, but now only those words with an accentuation probability below Θ remain. So $\Theta = 0.0$ means now that only the least accentuated word is used. No word is deleted for $\Theta = 1.0$. The results are presented in figure 2. In experiment ALL-Θ all words are used, but when no strongly accentuated words remain ($\Theta = 0.8$) the $\mathcal{RR}$ is only 67.3 % or below (see figure 2). Although the overall classification rate $\mathcal{RR}$ decreases using only unaccentuated words, the classes with fewer elements are recognized better. This can be concluded from an increasing $\mathcal{RRC}$ in all experiments with unaccentuated words.

To summarize, it can be seen that using accentuation information improves the recognition slightly (see exp. ALL+Θ).

3.2. Accent ordering experiments

In every experiment described above, a threshold is used as a criterion to sort out words. The threshold criterion does not inform about the number of words remaining. It is possible that every word of the utterance remains or that only one word remains. To examine the effect of the number of remaining words on the recognition rate, the most n accentuated or unaccentuated words are extracted. This can be easily implemented because the words of an utterance can be sorted according to their accentuation probability.

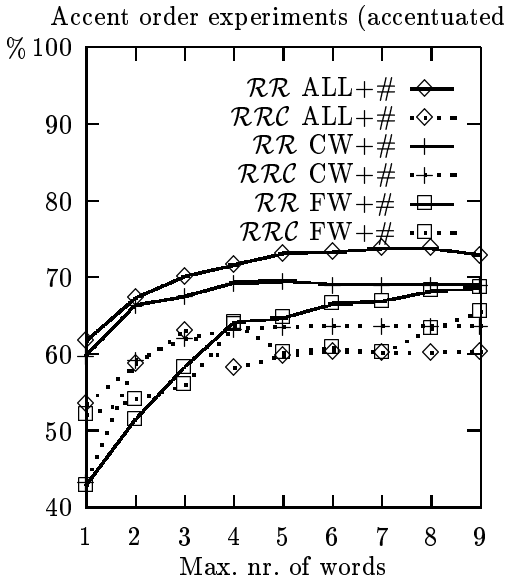

Figure 3. Recognition rates using the up to n most accentuated words. Their selection depends on their accentuation probability and the word class (ALL, CW, FW).

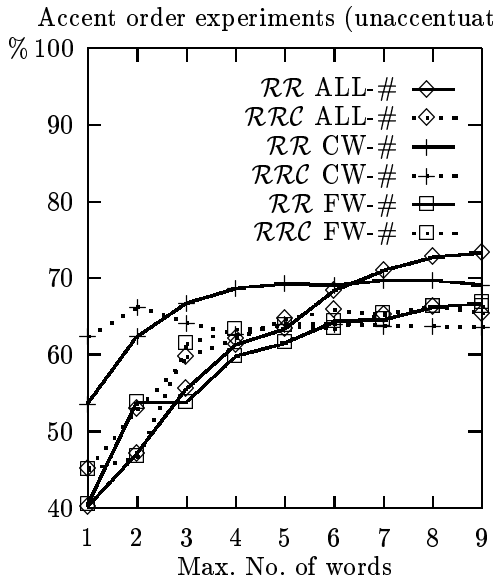

Figure 4. Recognition rates using the most up to n least accentuated words. Their selection depends on their accentuation probability and the word class (ALL, CW, FW).

Then it is easy to determine the first up to n members of that list and use them for the classification task.
The next series of experiments use only the first up to n words of the descending order sorted list (see figure 3). These words are the most probable accentuated words of the utterance. In figure 3, it can be seen that the $\mathcal{RR}$ in exp. ALL+# improves with every word until a phrase length of 7. From 5 to 7 words the $\mathcal{RR}$ improves only slightly. The best $\mathcal{RR}$ is achieved with 7 words. In experiment CW+# resp. and FW+# only CW or FW are considered. Note the poor performance in the beginning of exp. FW+#.
The same experiments with ascending sort order give the results shown in figure 4. Until only 6 words are considered in exp. ALL-#, the $\mathcal{RR}$ is not really good. But with 9 words we get the best combination of $\mathcal{RR}$ and $\mathcal{RRC}$ in all experiments, which is about 2% or 5% higher than the $\mathcal{RR}$ or $\mathcal{RRC}$ of the baseline experiment. A brief data inspection shows that there are many phrases with only a few CW and a limited number of (unaccentuated) FW. Combined with the knowledge that CW are often accentuated, it can be concluded, that in exp. ALL-# both kinds of relevant words exist: unaccentuated FW and accentuated CW, if only the phrase is long enough (here 9 words). From this point of view, it is not surprising, that this experiment produces the best recognition rates.

3.3. Keywords

We are also interested in which words are potential keywords for each dialogue act sub-class. Only the two classes SUGGEST-SUPPORT-DATE and SUGGEST-EXCLUDE-DATE provide sufficient data for this. So we extract the words with an accentuation probability above 0.8. This is done independently for CW and FW. The CW list contains mostly the names of months, numbers, weekdays and the words 'Zeit' (time) and 'Termin' (date). For FW the words 'nachmittags' (in the afternoon), 'ja' (yes), 'vielleicht' (maybe), 'oder' (or), 'könnte' (could) are on top of the list. Also the words with an accentuation probability below 0.2 are extracted. The most frequent words are: 'I', 'it', 'at', 'and', 'we', 'is' and 'not'.

4. CONCLUSION

The detection of dialogue acts is a very important task in VERBMOBIL. The recognition is usually based on determining keywords or cue phrases, without using prosodic information. This paper shows that prosodic and syntactic information can be useful to distinguish between dialogue act classes. The key is to carefully select the words which should remain in the word chain for the classification task. This paper has presented a scheme on how words can be selected for dialogue act classification. The selection is based on the POS and prosody. The POS is coded in a special dictionary. The accentuation probability is automatically computed by a MLP.
Two kinds of experiments have been conducted. Experiments which use words whose accentuation probability matches a threshold criterion and experiments which rely on the ranking of accentuation probabilities. For every experiment, the influence of the POS was also examined.
From the experimental results, it can be concluded that excluding the 'wrong' words improves the recognition rates. A selection criterion can be based on prosody (see exp. ALL+Θ). An overall result for all experiments is that using only FW is worse than using only CW. It is always better to use FW and CW together. The best result is achieved with a combination of accentuated CW and unaccentuated FW (exp. FW-#), with $\mathcal{RR} = 73.3\%$ and $\mathcal{RRC} = 65.4\%$. The recognition rates of the baseline experiment wi-

thout using prosodic or syntactic information are $\mathcal{RR} = 72.3\%$ and $\mathcal{RRC} = 60.1\%$, in comparison.

REFERENCES

[1] **A. Batliner, M. Nutt, V. Warnke, E. Nöth, J. Buckow, R. Huber and H. Niemann:** *Automatic annotation and classification of phrase accents in spontaneous speech*, to appear in EUROSPEECH, Budapest, 1999

[2] **A. Batliner, V. Warnke, E. Nöth, J. Buckow, R. Huber and M. Nutt:** *How to label accent position in spontaneous speech automatically with the help of syntactic–prosodic boundary labels*, VERBMOBIL Report 228, 1998

[3] **S. Jekat, A. Klein et al.:** *Dialogue acts in VERBMOBIL*, VERBMOBIL Report 65, 1995

[4] **A. Kiessling:** *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung*, Shaker Verlag, Aachen, Germany, 1997

[5] **R. Kompe:** *Prosody in Speech Understanding Systems*, Springer Verlag, Berlin, Germany, 1997

[6] **R. Kuhn:** *Keyword Classification Trees for Speech Understanding Systems*, Ph.D. Thesis, School of Computer Science, McGill University, Montreal, 1993

[7] **M. Nutt:** *Automatische Bestimmung von Akzenten und ihr Einsatz bei der Dialogakterkennung*, Diplomarbeit, Lehrstuhl für Mustererkennung, Univ. Erlangen–Nürnberg, Erlangen, Germany, 1998

[8] **N. Reithinger, R. Engel, M. Kipp and M. Klesen:** *Utilizing Statistical Dialogue Act Processing in VERBMOBIL*, VERBMOBIL Report 80, 1995

[9] **K. Samuel, S. Carberry and K. Vijay-Shanker:** *Computing dialogue acts from features with transformation-based learning*, Department of Computer and Information Science, University of Delaware, Newark, Delaware, USA, 1998

[10] **R. Searle:** *Speech Acts*, University Press, Cambridge, UK, 1969

[11] **E. Shriberg et al.:** *Can Prosody Aid the Automatic Classification of Dialogue Acts in Conversational Speech?*, Langugage and Speech 41(3-4): 439-487. Special Issue on Prosody and Conversation, 1998

[12] **W. Wahlster:** *VERBMOBIL-Translation of Face-to-Face Dialogs.* Technical Report, German Research Center for Artificial Intelligence (DFKI), 1993.