BMBF

Verbmobil
Verbundvorhaben

DIALOG ACT CLASSIFICATION WITH THE HELP OF PROSODY

M. Mast, R. Kompe,
S. Harbeck, A. Kießling,
H. Niemann, E. Nöth

F.–A.–Universität Erlangen–Nürnberg

Vm

Report 130
Mai 1996

Mai 1996

M. Mast, R. Kompe,
S. Harbeck, A. Kießling,
H. Niemann, E. Nöth

Lehrstuhl für Mustererkennung (Inf. 5)
Friedrich–Alexander–Universität Erlangen–Nürnberg
Martensstr. 3
D–91058 Erlangen

Tel.: (09131) 85 - 7799
e-mail: mast@informatik.uni-erlangen.de

Gehört zum Antragsabschnitt: 3.11, 3.12, 10.1

Das diesem Bericht zugrundeliegende Forschungsvorhaben wurde mit Mitteln des Bundesministers für Bildung, Wissenschaft, Forschung und Technologie unter dem Förderkennzeichen 01 IV 102 H/0 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei den Autoren.

DIALOG ACT CLASSIFICATION WITH THE HELP OF PROSODY

M. Mast[1], *R. Kompe*[1], *S. Harbeck*[1], *A. Kießling*[1], *H. Niemann*[1], *E. Nöth*[1],
E.G. Schukat-Talamazzini[2], *V. Warnke*[1]

[1]Universität Erlangen–Nürnberg, Lehrstuhl für Mustererkennung, 91058 Erlangen, Germany
http://www5.informatik.uni-erlangen.de/
[2]Universität Jena, Fakultät für Mathematik und Informatik, 07743 Jena, Germany

ABSTRACT

This paper presents automatic methods for the segmentation and classification of dialog acts (DA). In VERBMOBIL it is often sufficient to recognize the sequence of DAs occurring during a dialog between the two partners. Since a turn can consist of one or more successive DAs we conduct the classification of DAs in a two step procedure: First each turn has to be segmented into units which correspond to a DA and second the DA categories have to be identified. For the segmentation we use polygrams and multi–layer perceptrons, using prosodic features. The classification of DAs is done with semantic classification trees and polygrams.

1. INTRODUCTION

In the VERBMOBIL–Project an automatic translation system with the application of appointment scheduling dialogs is developed [11]. The scenario is the following: the dialog partners are assumed to have at least a passive knowledge of English. They usually will speak English until they do not know how to express themselves. Then they can switch to their mother tongue. In this case they have to press a button and VERBMOBIL translates the utterance. This means a translation is given on demand and only for parts of the dialog. Nevertheless the system must follow the entire dialog in order to catch the dialog context so that e.g. the resolution of anaphora or the generation of predictions is possible [1]. Thus, also when both dialog partners speak in English the system has to keep track of the dialog history, which means at least to recognize all dialog acts. In our approach this is done within two steps: First, a turn is segmented into DA units (DAU), and, second, these segments are classified into DA categories (DAC). For the segmentation of DAUs the same methods can be applied as for the segmentation of turns into sentences, cf. [5]. Note however, that a DAU can comprise more than one sentence. In rare cases there can even be a topic shift within one sentence so that it can be further segmented into different DAUs.

Few related work has been done: In [12] the segmentation of spontaneous speech utterances into sentences by classification trees has been investigated; mainly textual, only few prosodic features were used. Empirical studies showed that is important for the discourse structure [10] and for the correct interpretation of cue phrases, which have discourse governing function [3].

2. DIALOG ACTS IN VERBMOBIL

In VERBMOBIL the dialog as a whole is seen as a sequence of DAs, which means that DAs are the basic units on the dialog level. The DACs are defined according to their illocutionary force, e.g., ACCEPT, SUGGEST, REQUEST, and can be subcategorized as for their functional role or their propositional content, e.g., DATE or LOCATION depending on the application. We defined 18 DACs on the illocutionary level and 42 subcategories [4].

In Figure 1 the examples each show one turn hand–segmented into DAUs and hand–labeled with the appropriate DAC. Each DAU corresponds to one (cf. example 2) or more (cf. example 1) DA(s). Since in spontaneous speech many incomplete and incorrect syntactic structures occur, e.g., a lot of elliptical sentences or restarts, it is not easy to give a quantitative and qualitative definition of the term DA. We defined criteria for the manual segmentation of turns based on their textual representation and for the manual labeling of these segments with DACs [7]. Examples are:

- All 'material' that belongs to the verb frame of a finite verb belongs to the same DA. That way it is guaranteed that both the obligatory and the optional elements of a verb are included in the same DA, cf. example 2.
- Conventionalized expressions that do not necessarily contain a verb are seen as one unit even if they do not contain a verb. Typical phrases are: *hello, good morning, thanks.*

Prosodic information is not taken into account in order to be able to label dialogs without having to listen to them and

*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 01 IV 102 H/0. The responsibility for the contents lies with the authors. We wish to thank A. Batliner for helpful comments.

Ex. 1	uh Matt this is Brian here again	INTRODUCE_NAME
	I have to meet you sometime uh uhm this month to uh discuss the documentation for the code you have written	SUGGEST_SUPPORT_DATE, MOTIVATE_APPOINTMENT
Ex. 2	well I have a meeting all day on the thirteenth	SUGGEST_EXCLUDE_DATE
	and on the fourteenth I am leaving for my bob sledding vacation until the nineteenth	SUGGEST_EXCLUDE_DATE
	uh how 'bout the morning of the twenty second or the twenty third	SUGGEST _SUPPORT_DATE

Figure 1: Two turns segmented into DAUs and labeled with the respective DACs.

thus to reduce the labeling effort. Nevertheless, we will prove in Section 4.2. that for the automatic detection of DAUs prosodic markers are very important cues, cf. also [3]. These manually created labels are used as reference for the training and evaluation of our stochastic models, as described in the remainder of this paper.

3. METHODS USED

3.1. Multi–layer Perceptrons (MLP)

Multi–layer perceptrons were trained to recognize the DA–boundaries in a similar way as the prosodic phrase boundaries described in [5]. For each word–final syllable a vector of prosodic features c_i is computed automatically from the speech signal modelling prosodic properties over a context of six syllables taking into account duration, pause, F0–contour and energy. This is based on a time alignment of the phoneme sequence corresponding to the spoken words. The MLP has one output node for the DA boundaries (D) and one for the other word boundaries (¬D). During training the desired output for each of the feature vectors is set to one for the node corresponding to the reference label; the other one is set to zero. With this in theory the MLP estimates posterior probabilities. However, in order to balance for the a priori probabilities of the different classes, during training the MLP was presented with an equal number of feature vectors from each class. The best classification result so far (cf. below) was obtained with 117 prosodic features for each word–final syllable and an MLP with 60/30 nodes in the first/second hidden layer.

3.2. Polygrams Language Models (LM)

A certain kind of n–gram language models – so called polygrams [9] – are used for the segmentation and classification of DAs. Polygrams are a mixture of n–grams with varying size of n. They are superior to standard n–gram models because n can be chosen arbitrarily large and the probabilities of higher order n–grams are interpolated by lower order ones. The interpolation weights are optimized using the EM algorithm. However, we found that limiting n to 3 is sufficient for the experiments described in this paper. In [8] the DA classification results improved up to a maximum of $n = 5$, due to a larger amount of training data.

For the *segmentation* of turns into DAUs we trained LMs which model the probability for the occurrence of a boundary after the current word given the neighboring words, cf. [5]. For each word boundary symbol sequences $\ldots w_{i-2} w_{i-1} w_i v_i w_{i+1} w_{i+2} \ldots$ are considered, where w_i denotes the i-th word in the spoken word chain and v_i is either D or ¬D. During training v_i is replaced by the appropriate reference label and the n–gram probabilities are estimated based on these symbol sequences. In the test v_i is replaced by D and by ¬D; for both resulting symbol sequences probabilities are computed and combined with the output activations $P(v_i|c_i)$ of the MLP, which are assumed to approximate probabilities:

$$Q_{v_i} = P(v_i|c_i) P^{\xi}(\ldots w_{i-1} w_i v_i w_{i+1} w_{i+1} \ldots)$$

where c_i is the acoustic prosodic feature vector computed for word w_i and ξ denotes an optimized weight. For every v_i the score Q_{v_i} is normalized to a probability:

$$P(v_i) = \frac{Q_{v_i}}{\sum_{v_i \in \{D, \neg D\}} Q_{v_i}}$$

Polygrams are also used for the *classification* of the different DACs. For this purpose a separate LM is trained for each of the 19 DACs on the corresponding word sequences obtained from the hand–segmented turns. During classification a word sequence is scored with each of these LMs. We decide for the DAC with the highest LM probability.

3.3. Semantic Classification Trees (SCT)

SCTs are classification trees which operate on word sequences and are for example used for semantic concept classification [6]. The SCTs can be viewed as rule based systems where the rules are not hand–coded but are trained on a text corpus. The questions (decision rules) in the nodes of an SCT refer to regular expressions consisting of keywords (represented by w) and non–zero gaps (represented by "+"). The keywords and the specific rules are determined automatically during the training process from the given corpus. Only the type of rules is predefined. At the root of the tree the known regular expression is <+>, representing all non–zero symbol sequences. "<" and ">" represent the beginning and the end of the sequence, respectively. Questions are formed by replacing the known regular expression by another

		classified as	
reference	#	D	¬D
D	662	85.0	15.0
¬D	7317	6.8	93.2

Table 1: Confusion matrix for the recognition of DA boundaries with MLP+LM.

regular expression consisting of gaps and/or keywords. E.g., the first question could be: "Is the already known structure $<+>$ of the form $<+w+>$, i.e., does w appear somewhere in the structure?", where w represents a specific keyword. A subsequent question in the "yes" branch could than refer to the expression $<+w+w'+>$.

To avoid an over expansion of the tree, which means that it is totally adapted to the training corpus and therefore loses its generality, stopping rules are needed to end the expansion process. These stopping rules define when a node is declared to be a terminal node which is then labeled with the corresponding category from the inventory.

The rule which selects the best question or decides that the present node should be a leaf node is based on an impurity measure I, where the impurity is always non–negative and takes a maximum value for a node containing equal proportions of all possible categories, and is zero for a node containing only one of all possible categories. Therefore in each node the question is chosen which maximizes ΔI. The node is not split if the impurity cannot be reduced any further. Our algorithm uses the Gini criterion as measure of impurity which always lies between 0 and 1 (see [6]).

For growing the tree a strategy with two stages is used. First a much too large tree is grown using a simple stopping rule, e.g., that each terminal node contains fewer than N items (N close to 1) or that the maximum value of ΔI is 0. Based on the same impurity measure the tree is then pruned upwards starting at the leaves using an independent data set. Thereafter the tree is expanded on this second data set and pruned on the first. This process is iterated until no change in the tree occurs between two iterations (a prove for the convergence of this procedure can be found in [2]).

4. RESULTS

4.1. Data

All classification experiments are based on the same subsets of the German VERBMOBIL spontaneous speech corpus. We did not use English data because much more German data is available at present and the prosody module has not yet been adapted to English. For training 96 dialogs (2459 turns of 57 different female and 58 male speakers, approx. 5.5 hours of speech) are considered; the test set comprises 31 dialogs (453 turns) of 20 different speakers (3 female, 17 male; approx. 1 hour of speech). The training set consists of 6496 and the test set of 1107 DAs. These data sets have been chosen for the reason of compatibility with other research in VERBMOBIL and differ from the data used for our previous DA classification experiments, which were presented in [8]. In the experiments we distinguish between 18 DACs and an additional GARBAGE class for constructions, which are semantically meaningless due to interruptions or not well recorded utterances.

θ	*acc*	*corr*	*del*	*ins*
0.95	45.2	48.8	25.3	3.6
0.93	45.8	51.6	20.8	5.8
0.86	44.4	55.7	12.9	11.4
0.79	43.2	56.5	11.0	13.3
0.50	29.3	61.7	4.8	32.8

Table 2: Classification results for DACs based on automatically segmented word sequences.

4.2. Segmentation

For the segmentation of turns into DAUs MLPs and LMs were used. With MLPs an average recognition rate of 83.6% for D vs. ¬D was reached. The LM alone yielded in a recognition rate of 90.7%, the best result (92.5%) could be achieved with a combination of MLP and LM. A confusion matrix can be found in Table 1. We also tried an MLP trained on perceptual–prosodic clause boundaries for the classification of Ds; the recognition rate of the MLP alone was slightly better (84.4%), however the combination with the LM yielded only a recognition rate of 91.3%. All these results do not take into account the end of turns which by default are labeled as D and the classification of which is trivial.

4.3. Classification of Dialog Acts

For the classification of the 19 different DACs based on the hand–segmented data two methods were used: First, a semantic classification tree was trained where a recognition rate of 46.4% was obtained. Second, LM classifiers were trained as described in Section 3.2., yielding a recognition rate of 59.7%. In [8] we improved the SCT approach by using dialog–state dependent SCTs and achieved results comparable to the ones for polygrams. We currently modify the training algorithm and will repeat the experiments described in the following for polygrams with SCTs.

4.4. Segmentation and Classification of Dialog Acts

With respect to the integration in the VERBMOBIL system the DA classification has to deal with automatically segmented word sequences. For the segmentation we used the combination of MLP and LM as described in Section 4.2. The DACs are classified with the LMs. Note, that at this stage of our research we still work on the spoken word chain, thereby simulating 100% correct word recognition.

The classification is conducted as follows: First, we compute for each word boundary the probabilities $P(D)$ and $P(\neg D)$. Second we classify each boundary as D if $P(D) > \theta$ and as $\neg\mathsf{D}$ else. Third the word chains between each subsequent pair of D is extracted and classified with the LM into one out of the 19 DACs.

For the evaluation it has to be taken into account that DAUs may be deleted or inserted. Therefore, we align for each turn the recognized sequence of DAC class symbols with the reference. The alignment is performed with respect to the minimization of the Levenshtein distance. The percentage of correct ($corr$) classified DACs is given together with the percentage of deleted (del) and inserted (ins) segments in Table 2. Furthermore, the recognition accuracy (acc) measures the combined classification and segmentation performance; it is defined as $100 - subs - del - ins$, where $subs$ denotes the percentage of misclassified DACs. Note, that in this evaluation a DA is considered as classified correctly if it is mapped onto the same DA category in the reference no matter if the segment boundaries agree with the hand–segmented boundaries. In this context the most important numbers are the correctly classified DAs versus the insertions. In the table results for different thresholds θ are given. The smaller θ the smaller the number of deleted and the larger the number of inserted segments.

5. CONCLUSION

The segmentation and classification of DAs is an important upcoming issue, because in real dialogs a turn can consist of more than one DA. Especially in the context of VERBMOBIL the segmentation and classification of DAs is necessary for keeping track of the dialog history. As for the classification of DAs we compared SCTs and polygram LMs. We expected SCTs to be better suited for the classification of DAs than LMs because the SCTs can model long–distance dependencies to a greater extent than LMs. With the larger training material and dialog–step dependent SCTs we previously achieved comparable results [7]. We believe that for the SCTs the amount of training data was not sufficient.

In this paper we especially showed that DAs can be reliably classified based on automatically detected segments. As for the detection of the segment boundaries the best results were achieved with the combination of an MLP classifying acoustic–prosodic feature vectors with a polygram LM. Compared with the LM alone the combination of both classifiers reduced the error by over 19%. Thus the prosodic information proved to be very useful for this task. The results for the DA classification on the automatically segmented word sequences vary depending on the number of detected segments. The best accuracy achieved is 45.8%. The highest percentage of correctly classified DAs is 61.7%. This is even higher than the recognition rate of 59.7% when using the hand–segmented data. The reason for this might be that on the one hand most of the deleted segments are misclassified on hand–segmented data and on the other hand the high number of insertions increases the likelihood for a correct class being among the recognized ones. For the use in the VERBMOBIL system we have to find a trade–off between correctly classified DAs and inserted segments. Therefore, it is important to analyze how the inserted segments are classified and if this disturbs the task of dialog history tracking. Additional improvement can be expected by using LMs also for modelling the sequence of the different DACs within a turn. Furthermore, we want to investigate if we can reduce the number of insertions by using a rejection threshold.

Note, that our approach can as well be used for topic spotting, especially for utterances containing topic shifts.

6. REFERENCES

1. J. Alexandersson, E. Maier, and N. Reithinger. A Robust and Efficient Three–layered Dialogue Component for a Speech–to–speech Translation System. In *Proc. of the 7th Conference of the European Chapter of the ACL (EACL-95)*, pages 188–193, Dublin, 1995.
2. S. Gelfand, C. Ravishankar, and E. Delp. An Iterative Growing and Pruning Algorithm for Classification Tree Design. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 13:302–320, 1991.
3. J. Hirschberg and D. Litman. Empirical Studies on the Disambiguation of Cue Phrases. *Computational Linguistics*, 19(3):501–529, 1993.
4. S. Jekat, A. Klein, E. Maier, I. Maleck, M. Mast, and J. Quantz. Dialogue Acts in VERBMOBIL. Verbmobil Report 65, 1995.
5. R. Kompe, A. Kießling, H. Niemann, E. Nöth, E.G. Schukat-Talamazzini, A. Zottmann, and A. Batliner. Prosodic scoring of word hypotheses graphs. In *Proc. European Conf. on Speech Communication and Technology*, volume 2, pages 1333–1336, Madrid, 1995.
6. R. Kuhn and R. De Mori. The Application of Semantic Classification Trees to Natural Language Understanding . *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 17:449–460, 1995.
7. M. Mast, E. Maier, and B. Schmitz. Criteria for the Segmentation of Spoken Input into Individual Utterances. Verbmobil Report 97, 1995.
8. M. Mast, E. Nöth, H. Niemann, and E.G. Schukat-Talamazzini. Automatic Classification of Speech Acts with Semantic Classification Trees and Polygrams. In *International Joint Conference on Artificial Intelligence 95, Workshop "New Approaches to Learning for Natural Language Processing"*, pages 71–78, Montreal, 1995.
9. E.G. Schukat-Talamazzini. Stochastic Language Models. In *Electrotechnical and Computer Science Conference*, Portorož, Slovenia, 1995.
10. M. Swerts, R. Geluykens, and J. Terken. Prosodic Correlates of Discourse Units in Spontaneous Speech . In *Int. Conf. on Spoken Language Processing*, volume 1, pages 421–424, Banff, 1992.
11. W. Wahlster. Verbmobil — Translation of Face–To–Face Dialogs. In *Proc. European Conf. on Speech Communication and Technology*, volume "Opening and Plenary Sessions", pages 29–38, Berlin, 1993.
12. M.Q. Wang and J. Hirschberg. Automatic Classification of Intonational Phrase Boundaries. *Computer Speech & Language*, 6(2):175–196, 1992.