SEMANTIC BOUNDARIES IN MULTIPLE LANGUAGES

J. Haas[1], V. Warnke[1], H. Niemann[1], M. Cettolo[2], A. Corazza[2], D. Falavigna[2], G. Lazzari[2]

[1] *Universität Erlangen–Nürnberg,*
Lehrstuhl für Mustererkennung (LME),
D-91058 Erlangen, Germany
e-mail: haas@informatik.uni–erlangen.de
(http://www.mustererkennung.de/)

[2] *ITC-Irst, Centro per*
la Ricerca Scientifica e Tecnologica,
via Sommarive, 18,
I-38050 Povo, Trento, Italy
e-mail: cettolo@irst.itc.it
(http://www.itc.it/irst/)

10.21437/Eurospeech.1999-138

Abstract

This paper presents the results obtained for the task of detecting Semantic Boundaries (SBs) in spoken language using two different methods on the same data set. Hence we first introduce the two approaches developed by ITC-Irst in Trento (Italy) and the LME of the University Erlangen (Germany) and discuss the individually obtained results. The basis for the decision upon SBs in both cases are textual and prosodic features. The LME has already worked for several years on the computation and application of prosodic features in automatic speech processing within the VERBMOBIL* project. The approaches developed in that project were adapted to work on the data collected at IRST in the Italian language. Finally we compare the results we obtain with the German SB detection against the Italian result with regard to precision and recall.

1. INTRODUCTION

For robust spoken language processing it is not always necessary to analyse a user's utterance completely as one coherent segment. Often it is sufficient to split the utterance at certain points, which we call Semantic Boundaries (SBs), to get independently analyzable segments. The task we address in this contribution is the detection of SBs. Prosodic features characterizing energy, fundamental frequency F0, their contours, speaking rate and so on proved to be helpful for the detection of SBs [11, 14, 12, 10, 7]. Since the detection of semantic boundaries using only prosodic features is not reliable enough, also information about the words corresponding to the input signal is considered, either by using the best word sequence or the word hypothesis graph. In summary, two kinds of information are used for semantic segmentation: *prosodic features* and *words*.
In the following sections we discuss in detail the two approaches for the detection of SBs and the used features and methods. Furthermore the portability of the German approach to Italian data is proven and the results of the two cooperating institutes are compared.

2. LME APPROACH

The approach of the LME was developed in the VERBMOBIL project for the purpose of the segmentation and classification of dialog acts (DA). VERBMOBIL is a speech-to-speech translation project in the domain of appointment scheduling. The framework is that of two persons trying to fix a date where the partners talk in their native language and VERBMOBIL translates the utterance in the other language. For the system it is important to keep track of the dialog in order to know about the dialog state. This tracking is provided by terms of dialog acts (e.g. greeting or suggesting a date) which are to be found in the current utterance. Obviously an utterance most often covers more than only one DA. As said above in VERBMOBIL the task has two aspects. First we have to find the boundaries between two sequential dialog acts meet, i.e. we have to detect an SB, and second we have to classify the determined segment in order to match one out of 18 possible dialog act categories. Those categories are e.g. INTRODUCTION, SUGGEST_DATE, DELIBERATION [5]. In [7] we presented a two step approach for the segmentation and classification of DAs and in [13] we established an integrated approach using A*-search which proved to be better.

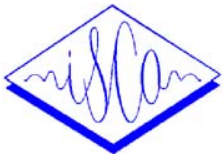
2.1. Used Methodology

The detection of semantic boundaries and the succeeding classification of the hypothesized segments into categories is done using several information sources and classification procedures, namely we use multi layer perceptrons (MLP) to recognize SBs from prosodic features and language models (LM) to hypothesize those boundaries on the current word sequence. The A*-search then uses the two measures calculated from the MLP and the LM, combines them with additional information and looks for the optimal sequence of words and SBs. As result we get the segmentation and consequently the semantic boundaries and the sequence of attached dialog act categories.

MLP Classification The MLP is trained to recognize SBs in an equivalent way as described in [6]. For each word–final syllable we compute several prosodic features automatically from the speech signal. Those features characterize 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 SB and one for ¬SB.

We assume that the MLP estimates posterior probabilities. However, in order to balance for the a priori probabilities of the different classes, the MLP is

*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.



6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

trained with an equal number of feature vectors from each class. For the classification we compute the prosodic features for each word–final syllable and use an MLP with 60/30 nodes in the first/second hidden layer.

LM Classification A certain kind of n–gram language models – so called polygrams [8] – are used for the segmentation and classification of dialog acts in VERBMOBIL. Polygrams are a set 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. There are several interpolation methods possible for the polygrams, which are described in detail in [8, 9]. For the segmentation of utterances we use LMs, which model the probability for the occurrence of an SB after the current word given the neighboring words, cf. [6]. For each 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 SB or $\neg$SB. Note that theoretically, we should model sequences $\ldots w_{i-1}v_{i-1}w_i v_i w_{i+1}v_{i+1}\ldots$; experiments showed, however, that this yields worse results. In this case the polygram obviously is not able to cover a sufficiently large word context.

A*-search Together with the above two measures we use a score computed by a dialog act dependent language model - for each dialog act we have one specialized LM – and a score from a LM modeling the sequence of dialog acts. All these scores are weighted with corresponding factors and the sum of them defines the cost function used in the search process. The remaining costs in the experiments we present here are always set to zero so that we always have a complete search.

2.2. Results on VERBMOBIL

In Table 1 we present the best results we obtain using the two step approach presented in [7]. As the LME uses the detection of SBs in combination with the classification of dialog acts, the results in [7] and [13] are given with respect to the classification of dialog acts so that we do not present them here.

C	I	D	Recall	Precision
563	498	99	85%	53%

Table 1: SB detection results on VERBMOBIL

3. IRST APPROACH

3.1. Lexical Information

Sentence texts of the corpus can be seen as sequences of strings that are either words or a dummy symbol (e.g. `SB`) , which does not correspond to a spoken word and indicates the presence of a semantic boundary.

A trigram LM can be trained on such sequences. Once a n-gram LM is estimated on a training set, there are several ways to find the most likely segmentation of a test/input sentence. One possibility is to score and sort all its possible segmentations.

If a sentence consists of m words, all the possible segmentations are 2^{m-1}, and the problem becomes intractable for large m. Heuristics can be introduced to limit the number of segmentation hypotheses to be scored. A possible approach is to put a threshold on the difference between the probabilities of the word sequence without and with the SB, that is, between $Pr(w_{i-n+1}\ w_{i-n+2}\cdots w_i)$ and $Pr(w_{i-n+1}\ w_{i-n+2}\cdots w_{i-1}\ SB)$. Then an SB is allowed between words w_{i-1} and w_i only when the difference is minor than the threshold. Another possible heuristic is to decide the maximum number of boundaries which can be present in the sentence: only the $q < m$ boundaries corresponding to the q lower differences are considered.

Once all the allowed segmentation hypothesis are scored and ordered, the best one can be taken. If another knowledge source is available, it is also possible to use it to rescore the k-best segmentations. This can be a way for integrating scores based on prosodic features.

The number k of the best hypotheses can be fixed a priori, or be decided by considering the k segmentations whose scores differ from the best one for less than a certain quantity, defined by a factor $\delta \in [0, 1]$.

3.2. Prosodic Information

Given a test/input sentence, a vector $\vec{\theta}_i$ of prosodic features can be computed at the end-time of each word w_i. A boolean label can be associated to the vector: `True` when w_i is the last word of a semantic unit but it is not the last word of the sentence; `False` when w_i and w_{i+1} belong to the same semantic unit, or when w_i is the last word of the sentence.

A Binary Classification Tree (BCT) [2] can be trained to recognize the presence of a SB on the basis of the feature vector $\vec{\theta}$. Given a segmented sentence $\vec{\mathbf{v}}$, the BCT is asked to give the probability of all end-time words. The product of these probabilities over all words gives the "prosodic plausibility" of that particular segmentation.

Computed prosodic features are related to speaking rate, energy and `F0` contours. Their description can be found in [3].

3.3. LM and Prosody Integration

The integration of lexical and prosodic information was done by rescoring the k-best segmentations, hypothesized by the LM, with their prosodic plausibilities. In particular, the one giving the best score obtained with the weighted product of its LM probability ($\mathrm{Pr}^{LM}(\vec{\mathbf{v}}_j)$) and its prosodic plausibility ($\mathtt{Pl}^{\mathtt{pros}}(\vec{\mathbf{v}}_j)$) is chosen as follows:

$$\hat{\vec{\mathbf{v}}} = \underset{j=1\ldots k}{\mathtt{argmax}} \left(\mathrm{Pr}^{LM}(\vec{\mathbf{v}}_j)\right)^{\alpha} \times \left(\mathtt{Pl}^{\mathtt{pros}}(\vec{\mathbf{v}}_j)\right)^{\beta} \qquad (1)$$

4. CORPUS DESCRIPTION

Experiments were carried out on a dialog corpus collected at ITC-Irst [1], composed of monolingual person-to-person Italian conversations for which acoustic signals, word transcriptions and linguistic

annotations are available. The two speakers were asked to fix an appointment, observing the restrictions shown on two calendar pages they were given; they did not see each other and could hear the partner only through headphones. The conversations took place in an acoustically isolated room and were naturally uttered by the speakers, without any machine mediation.

The dialogs were transcribed by annotating all extra-linguistic phenomena such as mispronunciations, restarts and human noises, with the exception of pauses.

	Training	Test	Whole Corpus
# dialog	169+12/2	20+12/2	201
# turn	2680	406	3086
# DA	5421	877	6298
# SB	2741	471	3212
size (non-noise words)	27786	4683	32469
|V| (non-noise words)	1291	627	1433

Table 2: Training and test set statistics.

The whole corpus was then divided into training and test sets (see Table 2), paying attention to avoid speaker overlap between the two sets. The test set consists of all the sentences uttered by 11 speakers, resulting in 20 complete dialogs and 12 half dialogs, for a total of 406 turns.

5. COMPARISON OF RESULTS

In this section we report about the adaptation of the LME approach to the Italian data on appointment scheduling. The results we present are generated using the two-step LME approach without the A*-search. Even though we obtained better results with the integrated approach, we decided to start with our previous approach, since it is directly comparable with the two-step IRST approach.

5.1. Preparation and Preprocessing

For the application of the LME method on the IRST data we first of all had to prepare the data to match our requirements. Most of the work done for this data preparation is due to implementation details, for sure, but nevertheless it takes quite a while to get the programs work with "foreign" data. For example we had to adapt the lexicon as the Italian data did not include syllable boundaries but they are needed for the LME prosodic features. Another point was the different formats of the alignment data for words and phonemes. After preparing the data some first preprocessing could be done, e.g. the computation of statistics concerning the duration of phonemes and computing the basic prosodic features energy and fundamental frequency.

5.2. IRST Results

Results using LM In order to make the number of semantic segmentation hypotheses manageable, only a maximum of $q = 14$ SBs (see Subsection 3.1) was allowed inside each sentence. This means that at the most $2^{14} = 16384$ different segmentations had to be scored for each test/input sentence.

In Table 3 results are reported by aligning the *1*-best output against the hand labelled test data. Performance is given in terms of correct detection (C), insertions (I) and deletions (D) of SBs, and recall and precision measures. The LM employed was a Shift-β trigram LM described in [4].

type	C	I	D	Recall	Precision
LM	285	115	186	60.5%	71.3%

Table 3: SB detection results using the LM *1*-best output.

Results using Prosody To check the relevance of the three types of prosodic features, three different BCTs were built: one for the 3 speaking rate features (`ros`), one for the 25 features related to the energy contour (`ene`), and one for the 18 features derived from the `F0` curve (`F0`). Finally, a general BCT was trained to handle all the 46 prosodic features considered (`all`).

In Table 4 results obtained on the test set, by aligning the outputs of the BCTs against the hand labelled test data, are reported.

type	#feat.	C	I	D	Recall	Precision
ros	3	141	639	330	29.9%	18.1%
ene	25	171	510	300	36.3%	25.1%
F0	18	133	622	338	28.2%	17.6%
all	46	211	520	260	44.8%	28.9%

Table 4: SB detection results using prosody.

Effects of Integration The integration of LM and prosody was then applied as explained in Subsection 3.3. The average number k of segmentation hypotheses to be rescored was 5.4, derived setting δ to 0.980. Weights α and β were empirically chosen, and set to 0.8 and 1.0 respectively. Results are reported in Table 5.

type	C	I	D	Recall	Precision
LM$\oplus$prosody	296	116	175	62.8%	71.8%

Table 5: SB detection results using LM and prosody.

5.3. LME Results

Results using LM For the SB detection task only with LM we train a trigram i.e. we have a two words context for the decision upon an SB, we use rational interpolation ([9]) and employ the method explained in Subsection 2.1. In Table 6 we report the results in the same terms as for the IRST evaluation. For an easier comparison between the IRST and LME results we tuned the parameters of the LM in a way so that the precision is equal to the IRST precision.

Results using Prosody For testing the quality of the prosodic features we trained an MLP for the detection of boundaries using only prosodic features. We computed the 276 features and fed them in the network with two hidden layers where the first one had 60 nodes and the second 30 and decided on the

type	C	I	D	Recall	Precision
LM	314	127	157	67%	71%

Table 6: SB detection results using the LM with rational interpolation and two words context.

occurrence of an SB or its non-occurrence. The results are shown in Table 7.

type	#feat.	C	I	D	Recall	Precision
all	276	281	186	190	60%	60%

Table 7: SB detection results using prosody.

Effects of Integration The results obtained when combining lexical and prosodic information are shown in Table 8. We want to emphasize that we did not use the A*-search to produce those results but only our first approach presented in [7]. Here again we tuned parameters in a way that the precision matches the IRST result.

type	C	I	D	Recall	Precision
LM⊕prosody	372	152	99	79%	72%

Table 8: SB detection results using LM and prosody.

5.4. Discussion

From the above reported results and the cooperation we draw some useful conclusions:

We proved successfully that the adaptation of the methods developed in VERBMOBIL are easily portable to new languages with only little effort and work. In this paper we showed it for the Italian language, some work for English and Japanese is also already done.

The LME results using only prosodic features are better since the LME approach uses more features – 6 times as much as IRST – and additionally the LME has already worked for several years on computing prosodic features and their application in speech processing.

By using only lexical information, the LME approach performs better than the IRST one in a significant way. In our opinion, this is mainly due to the quantity of training data, that results to be enough for a robust training of the *local* LME approach, unlike for the *global* IRST approach. The results obtained on a different and larger corpus (about twice the size), on which the IRST approach performs slightly better than the LME one, supports this idea.

Heuristics in IRST approach to SB detection based on lexical information does not introduce notable performance degradation, while they allow to take quickly a global decision. The observation results from experiments performed at IRST for comparing the algorithm presented here and an admissible one (A*).

As a result from the better performance of the LME methods used separately the integration of both knowledge sources also performs better.

6. SUMMARY

In this paper we presented two different approaches for the detection of semantic boundaries in spoken language. The methods where developed for Italian and German and in a further step we adapted the German approach so that it works for the Italian data. Results showed that the application is quite easily feasible. We compared the results obtained by the two cooperating institutes (IRST and LME) and discussed the differences.

7. References

1. B. Angelini, M. Cettolo, A. Corazza, D. Falavigna, and G. Lazzari. Multilingual Person to Person Communication at IRST. In *ICASSP*, Munich, Germany, 1997.
2. L. Breiman, J.H. Friedman, R.O. Olshen, and C.J. Stone. *Classification and Regression Trees.* Wadsworth, Pacific Grove, Cal., 1984.
3. M. Cettolo and D. Falavigna. Automatic Detection of Semantic Boundaries based on Acoustic and Lexical Knowledge. In *ICSLP*, Sidney, Australia, 1998.
4. M. Federico, M. Cettolo, F. Brugnara, and G. Antoniol. Language Modeling for Efficient Beam-Search. *Computer Speech and Language*, 9:353–379, 1995.
5. S. Jekat, A. Klein, E. Maier, I. Maleck, M. Mast, and J. Quantz. Dialogue Acts in Verbmobil. Verbmobil Report 65, 1995.
6. 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.
7. M. Mast, R. Kompe, S. Harbeck, A. Kießling, H. Niemann, E. Nöth, and V. Warnke. Dialog Act Classification with the Help of Prosody. In *Int. Conf. on Spoken Language Processing*, volume 3, pages 1728–1731, Philadelphia, 1996.
8. E.G. Schukat-Talamazzini. Stochastic Language Models. In *Electrotechnical and Computer Science Conference*, Portorož, Slovenia, 1995.
9. 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*, volume 5, pages 2731–2734, Rhodes, Greece, 1997.
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. M. Swerts and M. Ostendorf. Prosodic and Lexical Indications of Discourse Structure in Human–machine Interactions. *Speech Communication*, 22(1):25–41, 1997.
12. M.Q. Wang and J. Hirschberg. Automatic Classification of Intonational Phrase Boundaries. *Computer Speech & Language*, 6(2):175–196, 1992.
13. V. Warnke, R. Kompe, H. Niemann, and E. Nöth. Integrated Dialog Act Segmentation and Classification using Prosodic Features and Language Models. In *Proc. European Conf. on Speech Communication and Technology*, volume 1, pages 207–210, 1997.
14. C.W. Wightman and M. Ostendorf. Automatic Labeling of Prosodic Patterns. *IEEE Trans. on Speech and Audio Processing*, 2(3):469–481, 1994.