Gefördert vom

bmb+f

Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie

Verbmobil

Verbundvorhaben

A Frame and Segment based Approach for Topic Spotting

E. Nöth, S. Harbeck
H. Niemann, V. Warnke

F.–A.–Universität Erlangen–Nürnberg

Vm

Report 217
September 1997

September 1997

E. Nöth, S. Harbeck
H. Niemann, V. Warnke

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

Tel.: (09131) 85 - 7883
e-mail: `warnke@informatik.uni-erlangen.de`

Gehört zum Antragsabschnitt: 4.8

Das diesem Bericht zugrundeliegende Forschungsvorhaben wurde mit Mitteln des Bundesministers für Bildung, Wissenschaft, Forschung und Technologie unter dem Förderkennzeichen 01 IV 701 K/5 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei den Autoren.
ISSN 0949-6084

In Proc. EUROSPEECH '97, Rhodes, Vol. 1, pages 275–278

A Frame and Segment Based Approach for Topic Spotting

E. Nöth, S. Harbeck, H. Niemann, V. Warnke

Universität Erlangen–Nürnberg, Lehrstuhl für Mustererkennung,
Martensstr. 3,
91058 Erlangen, Germany
Tel.: +49 9131 / 857888
Fax.: +49 9131 / 303811
noeth@informatik.uni-erlangen.de http://www5.informatik.uni-erlangen.de

Abstract

In this paper we present a new approach for topic spotting based on subword units (phonemes and feature vectors) instead of words. Classification of topics is done by running topic dependent polygram language models over these symbol sequences and deciding for the one with the best score. We trained and tested the two methods on three different corpora. The first is a part of a media corpus which contains data from TV shows for three different topics (IDS), the second is part of the Switchboard corpus, the third is a collection of human machine dialogs about train timetable information (EVAR corpus). The results on Switchboard are compared with phoneme based approaches which were made at CRIM (Montréal) and DRA (Malvern) and are presented as ROC curves; the results on IDS and EVAR are compared with a word based approach and presented as confusion tables. We show that a surprisingly little amount of recognition accuracy is lost when going from word to subword based topic spotting.

1. Introduction

In most approaches in the field of topic spotting, words or word sequences are used for identifying a topic [8]. This is done by word recognition with large vocabularies or special word spotters [14]. To train such recognizers in both cases a huge amount of tediously labeled data has to be available.

For the training of our topic spotter with phonemes and vector quantized feature vectors we do not need the data to be labeled as exactly as for training a word spotter or a word recognition system. We only need the speech signals labeled with their topic rather than a word–to–word transliteration. Using either a vector quantizer or phoneme segmentizer (see section 2.), we segment the speech signal into a symbol sequence. With these sequences we train stochastic language models (LMs) for each of the topics to be identified. In the test phase we run all LMs in parallel and decide for the topic with the maximum a posteriori probability (see section 3.).

The advantage of our approach is evident when changing to a new domain. Doing topic spotting with a large vocabulary speech recognizer, one has to adapt the lexicon if not retrain all the acoustic models with domain dependent transliterated speech. With a word spotter, new keywords have to be identified and trained. In our approach only the language models have to be retrained. The training labels, i.e. the assignment of the topic to the training utterances, can be done very fast. Using the vector quantizer, one can even switch the language without the need for a more detailed labeling of the training data.

For our experiments we understand a topic in two different ways: in the first case (IDS [13] and Switchboard [4]), the topic of an utterance is given by the topic of the TV show (i.e. *folk music* vs. *politics*) or by instructing the speaker to talk about a subject; in the second case (EVAR, [1]) we analyze user utterances which were collected in a field test with our information retrieval dialog system. Each utterance is represented on a semantic level as containing a certain number of semantic attributes (like *relative time expression*). In this case we interpret the presence or absence of a semantic attribute as topic of the utterance. Notice that we do not consider topic changes (i.e. one utterance belongs to exactly one topic), and that the definition of topic w.r.t. IDS and Switchboard is much more broad than the definition w.r.t. EVAR[2].

2. Subword Units

For representing the speech signal as feature vectors, we use the mel-frequency-cepstrum-coefficients (MFCCs). A feature vector $\mathbf{c}$ is calculated for a 10ms part of the speech signal and contains the energy and the first 11 MFCCs. To create useful *codebook class sequences* (CCSs) eight neighboring feature vectors $\mathbf{c}$ are concatenated to a new feature vector $\hat{\mathbf{c}}$, which describes a context of 80ms with 96 coefficients. We use this time window, because the average length of a phoneme is about 80ms across many languages. With this feature vector we calculate an initial codebook $q(\hat{C})$ with 256 classes. Using the *linear discriminant*

[1]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. We would like to thank our colleagues from CRIM and DRA for the fruitful discussions and for providing us with the phoneme sequences for Switchboard.

[2]The accuracy of a topic spotter heavily depends on the closeness of the competing topics, whether we allow the speaker to digress from the given topic, and how much time to decide upon the topic.

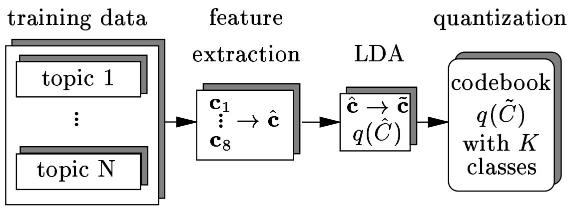

Figure 1: Partitioning of the feature space with feature reduction using LDA

analysis (LDA) on the codebook $q(\hat{C})$ we optimize the intra- and inter-class-distance of the 256 codebook classes and transform the feature vector $\hat{\mathbf{c}}$ with the 96 components in a new smaller vector $\tilde{\mathbf{c}}$ with 24 components. This feature vector is used to train a new codebook $q(\tilde{C})$ with 65 classes (see figure 1). For training the codebook we use the well known *LBG*-algorithm [7], which minimizes the expected quantization error

$$\varepsilon = \mathcal{E}[d(\mathbf{C}, q(\mathbf{C}))] \tag{1}$$

With the resulting codebook we are able to segment the utterances of the training and test data into equidistant codebook class sequences with a segment length of 80ms. Using these CCSs we train a LM for every topic of interest.

For representing the speech signal as phonemes sequences (PS) we use a standard HMM recognizer. In the case of the IDS corpus we use a monophone recognizer trained on the Verbmobil task [2]; for the Switchboard corpus we use the phoneme recognizer from CRIM [6] to allow for better comparison of the results (see Section 5.).

3. Polygram Language Model

In most cases language models are used to calculate the probability of a word sequence $\omega = \omega_1 \ldots \omega_m$ in a given language or context. We use *polygram language models* [10], which are a special kind of *stochastic* language models to estimate the probability of a *symbol sequence* where a symbol could be a phoneme or a codebook class.

Using polygrams the probability of the symbol sequence $\omega_1 \ldots \omega_m$ is calculated with the help of

$$P(\omega_1 \ldots \omega_m) = P(\omega_1) \cdot \prod_{n=2}^{m} P(\omega_n \mid \underbrace{\omega_1 \omega_2 \ldots \omega_{n-2} \omega_{n-1}}_{\text{history}}) \tag{2}$$

Because the younger history $\omega_{m-n+1} \ldots \omega_{m-1}$ of the symbol sequence $\omega_1 \ldots \omega_m$ is more important for modeling and to restrict the number of free parameters inside the LM, we only use the last $n-1$ symbols instead of the whole history.

$$P(\omega_m \mid \omega_1 \ldots \omega_{m-1}) \simeq P(\omega_m \mid \underbrace{\omega_{m-n+1} \ldots \omega_{m-1}}_{(n-1)}) \tag{3}$$

With this shorter history we can estimate $P(\omega_m \mid \omega_{m-n+1} \ldots \omega_{m-1})$ from a given training corpus using the interpolation scheme

$$\hat{P}(\omega_m \mid \omega_1 \ldots \omega_{m-1}) = \frac{\#(\omega_1 \ldots \omega_m)}{\#(\omega_1 \ldots \omega_{m-1})}, \tag{4}$$

where # is a function which counts how often a symbol sequence is seen in the training data. To handle symbol sequences that were not seen in the training data we need an interpolation formalism.

Linear interpolation

The first interpolation method we use is the *linear interpolation* [10] ($L = \textit{lexicon size}$):

$$\begin{aligned}\tilde{P}(\omega_m \mid \omega_1 \ldots \omega_{m-1}) &= p_0 \cdot \frac{1}{L} \\ &+ p_1 \cdot \hat{P}(\omega_m) \\ &+ p_2 \cdot \hat{P}(\omega_m \mid \omega_{m-1}) \\ &+ \sum_{i=3}^{n} p_i \cdot \hat{P}(\omega_m \mid \omega_{m-i+1} \ldots \omega_{m-1}).\end{aligned} \tag{5}$$

The interpolation coefficients p_i can be estimated using the *Expectation Maximization (EM)* algorithm [9] on a given validation set. Using this method an unseen symbol sequence is modeled by its subsequences weighted with the interpolation coefficients.

Rational interpolation

The *rational interpolation* method is the second interpolation we apply [10, 11]:

$$P(\omega_m \mid \omega_1 \ldots \omega_{m-1}) = \frac{\sum_{i=1}^{n} p_i \cdot (1/L)^{n-i} \cdot \#_i(\omega_1 \ldots \omega_{m-1}\omega_m)}{\sum_{i=1}^{n} p_i \cdot (1/L)^{n-i} \cdot \#_i(\omega_1 \ldots \omega_{m-1})}, \tag{6}$$

where $\#_i$ counts the i predecessors $(\omega_{m-i} \ldots \omega_{m-1})$ in a given sequence ω. In this interpolation formalism it is also possible to estimate the interpolation coefficients using the EM-algorithm on a given validation set. This interpolation gives more weight to the symbol sequences which have often been present in the training data and are in the nearest neighborhood of the observed symbol. New methods for polygram interpolation are presented in [11].

Language models as classifiers

Training topic dependent language models and running all of them in parallel we can use the language models as topic classifiers. Each one estimates the a-posteriori probability $P(T_i \mid \tilde{\omega})$ for each topic $T_i \in \mathcal{T}$ of interest. We decide for the topic with the maximum a posteriori probability.

$$\frac{P(T_i \mid \omega) = P(\omega \mid T_i) \cdot P(T_i)}{\sum_{j=1}^{N} P(\omega, T_j).} \tag{7}$$

4. Corpora

4.1. IDS Corpus

We performed experiments on a small part of the IDS Media corpus (Institut für Deutsche Sprache). This

approach	*CCS*		*PS*		*WS*	
Interpolation	R	L	R	L	R	L
language	78	70	87	70	83	82
Politics	78	85	74	81	81	96
Culture	50	87	75	81	56	75
average	71	**80**	**79**	77	76	**86**

Table 1: Recognition results in percent for CCS, PS and WS at the end of the utterance

corpus contains data from German TV shows for the three topics *speech, politics, and culture.* It is not easy to say, which TV show corresponds to which topic. So we assigned, for example, one TV show to the topic "politics", if it was announced that it is a political discussion about the "gulf war". All utterances of this TV show were then assigned to the topic "politics". 316 utterances from 11 different TV shows were divided into speaker disjunctive training (250 utterances) and test sets (66 utterances). The length of an utterance is 39 seconds in the average.

4.2. Switchboard Corpus

The Switchboard corpus is an example for a broad definition of topic. The users were given the task to call another person who is unknown to them. They were instructed to talk about a certain subject like *family life* or *gun control.* The length of an utterance is approximately 5 minutes. We used the exact same subset of Switchboard as [6], i.e. 10 topics with a total of 507 files, arranged in the ratio 9:1 training:test in 10 different ways (leave one out). The Switchboard is a somewhat artificial task since the speakers were instructed to talk about a given topic. Consequently they often "gave the topic away" within the first sentence (*so we are supposed to talk about ...*). In the case of subword units the weakness of the corpus has not as much effect as when evaluating keyword based methods.

4.3. EVAR Corpus

For the experiments on topic spotting as part of a shallow linguistic analysis we used the corpus described in [3] which is collected using the train-timetable information system EVAR. For training and test purposes we use a set of 10114 sentences (2/3 for training and 1/3 for test). The average length of an utterance is about 3.5 words. For these sentences we have a semantic annotation which we use as a reference for the detection of semantic attributes.

5. Experiments and Results

For the IDS corpus we calculated polygrams for CCS, PS and — simulating 100 percent word accuracy — the spoken word sequence (WS). The maximum context of the polygrams were trigrams. We interpolated them with either linear (L) or rational (R) interpolation. Table 1 shows the results we achieved for the three class problem. It is surprising that we only loose 6 percent recognition accuracy when going from the spoken word chain to codebook classes. However one has to keep in mind that there is an

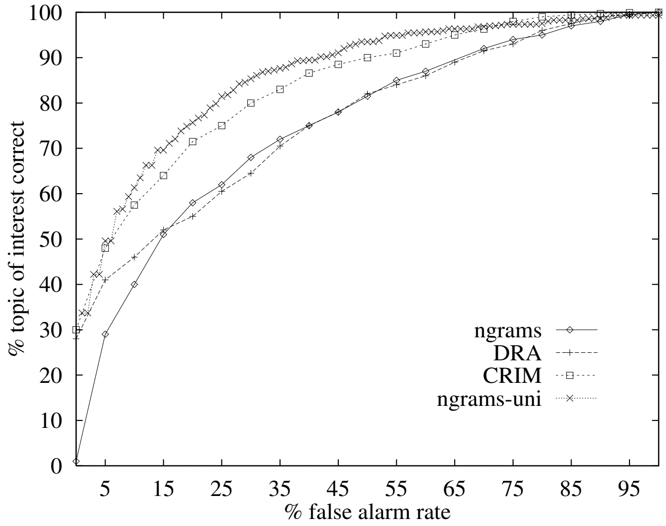

Figure 2: Topic Spotting Performance on Switchboard

order of magnitude less observations for training the language models on WSs. We observed that linear interpolation was more robust with smaller amounts of training data (WS) or more consistent data in case the amount the training was the same. This can be seen when comparing the CCS with the PS results: the amount of training material is the same but PS is a better representation of the acoustic event than the crude segmentation into equidistant 80ms segments. For a more detailed discussion of the different interpolation methods and spotting results with smaller segments than the whole utterance cf. [12, 13].

Using the data from the Switchboard corpus we only took the phoneme sequences produced by the CRIM recognizer as input to the polygram classifiers. We can thus compare our results directly to the approach from the CRIM (Montréal) who experimented with a nearest neighbor and a decision tree based approach as well as to the results from DRA (Malvern) who work with a DP based ngram approach on the same phoneme sequences (see [6] for a description of these phoneme based approaches). All approaches in [6] decide for a class based on fragments of the phonemes sequences ("key phoneme sequences"). These fragments were selected automatically with a discriminative training approach. In Figure 2 our approach is compared to the most successful CRIM approach (Euclidean Extended Pruning - Length 4 in [6]) and the most successful DRA approach (DP-ngrams Version 2 in [6]). Above 50 percent detection rate our approach gives comparable results to the DRA approach, CRIM's approach is clearly superior.

For the EVAR corpus we grouped the utterances w.r.t. containing the semantic attributes *CITY*, *TIME* or *DATE* exactly once (*CITY_ONLY*, ...), at least once and some other semantic attributes (*CITY_PLUS*, ...) or not at all (*CITY_NO*, ...). We only took the codebook sequences and compared the results with polygram language models over the spoken and the recognized word sequence from [5]. Again as in the case of the IDS corpus the loss of recognition

	spoken word sequence				recognized word sequence				codebook sequence			
	ONLY	PLUS	NO	RR	ONLY	PLUS	NO	RR	ONLY	PLUS	NO	RR
CITY_ONLY	**82**	17	0		**52**	29	19		**42**	28	30	
CITY_PLUS	10	**90**	0	**86**	11	**79**	11	**74**	15	**67**	18	**64**
NO_CITY	6	9	**85**		8	14	**78**		14	15	**71**	
DATE_ONLY	**92**	8	0		**72**	11	17		**33**	30	47	
DATE_PLUS	12	**84**	5	**85**	6	**60**	34	**79**	5	**62**	33	**69**
NO_DATE	6	9	**85**		5	13	**83**		6	22	**72**	
TIME_ONLY	**91**	9	0		**71**	11	18		**11**	31	58	
TIME_PLUS	9	**89**	2	**85**	6	**66**	28	**81**	1	**65**	34	**77**
NO_TIME	7	8	**85**		7	9	**84**		2	18	**80**	

Table 2: Recognition results in percent for CCSs in comparison with results on the spoken and recognized word sequence

accuracy is surprisingly low (see Table 2), especially when taking into account the very short length of the EVAR utterances (see Section 4.). In the case of the CITY detector the loss is higher due to a higher amount of unknown words (this can also be seen when going from spoken to recognized word sequences).

6. Conclusion

We presented two approaches for topic spotting with the use of subword units. Because we use vector quantization or a phoneme segmentizer, we only need the utterances of the training set labeled with their topics. The main difference between both approaches is the fact, that we use CCSs of equidistant length in our vector quantizer and a symbol sequence of variable length in our phoneme based approach. In both approaches the topics are modeled with the help of stochastic language models. During the recognition task for all topics the a posteriori probability is calculated in parallel and the topic with the maximum probability is chosen. We showed that depending on the definition of topic and the amount of training data our approach performed almost as good as a spotter which uses a "perfect" word recognition module. The approach proved to show good results with three corpora which were very different w.r.t. distance between the topics of the corpus and time to decide.

7. References

1. M. Boros, W. Eckert, F. Gallwitz, G. Görz, G. Hanrieder, and H. Niemann. Towards understanding spontaneous speech: Word accuracy vs. Concept accuracy. In *Int. Conf. on Spoken Language Processing*, volume 2, pages 1005–1008, Philadelphia, 1996.
2. T. Bub and J. Schwinn. 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.
3. W. Eckert, E. Nöth, H. Niemann, and E. Schukat-Talamazzini. Real Users Behave Weird — Experiences made collecting large Human–Machine–Dialog Corpora. In P. Dalsgaard, L.B. Larsen, L. Boves, and I. Thomsen, editors, *Proc. of the ESCA Tutorial and Research Workshop on Spoken Dialogue Systems*, pages 193–196, Vigsø, Denmark, 1995. ESCA.
4. J. Godfrey, Hlliman E., and J. McDaniel. SWITCHBOARD: Telephone Speech Corpus for Research and Development. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 517–520, 1992.
5. J. Haas, E. Nöth, and H. Niemann. Semantigrams – Polygrams Detecting Meaning. In *Proc. of the 2nd SQEL Workshop on Multi-Lingual Information Retrieval Dialogs*, pages 65–70, Pilsen, 1997. University of West Bohemia.
6. R. Kuhn, P. Nowell, and C. Drouin. Approaches to Phoneme–based Topic Spotting: An Experimental Comparison. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 3, pages 1819–1822, München, 1997.
7. Y. Linde, A. Buzo, and R.M. Gray. An Algorithm for Vector Quantizer Design. *IEEE Trans. on Communications*, 28(1):84–95, 1980.
8. E. Parris and M. Carrey. Topic spotting with task independent Models. In *Proc. European Conf. on Speech Communication and Technology*, pages 2133–2136, Madrid, Spain, September 1995.
9. E. G. Schukat-Talamazzini. *Automatische Spracherkennung – Grundlagen, statistische Modelle und effiziente Algorithmen.* Vieweg, Braunschweig, 1995.
10. E.G. Schukat-Talamazzini. Stochastic Language Models. In *Electrotechnical and Computer Science Conference*, Portorož, Slovenia, 1995.
11. 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.
12. V. Warnke. Topik– und Topikgrenzen–Spotting. Diplomarbeit, Lehrstuhl für Mustererkennung (Informatik 5), Universität Erlangen–Nürnberg, 1996.
13. V. Warnke, S. Harbeck, H. Niemann, and E. Nöth. Topic Spotting using Subword Units. In *9. Aachener Kolloqium "Signaltheorie", Bild– und Sprachsignale*, pages 287–291, 1997.
14. M. Wintraub. Keyword-spotting using sri's decipher large vocabulary speech recognition system. In *Proceedings International Conference on Automatic Speech and Signal Processing*, pages 463–466, 1993.