Online Topic Segmentation of Russian Broadcast News

Maxim Korenevsky, Irina Ponomareva, Kirill Levin
Speech Technology Center, Saint-Petersburg, Russia

Abstract

This paper deals with topic segmentation of continuous speech. We propose an online segmentation method that relies on the information about sentence boundaries obtained from an automatic sentence boundary detection system. We show that using information about sentence boundaries to divide continuous speech into fragments for topic classification provides an increase in classification accuracy of about 25-30%, compared to the method where only a threshold on the number of words is used to divide continuous speech into fragments. The highest average classification F-measure for 5 topics obtained in our experiments is 0.79.

1. Introduction

The use of information extraction technologies from audio data enables the examination of a much wider range of data sources than when using only text. In the modern world, many sources (e.g., interviews, conversations, news broadcasts) are available only in audio form.

One of the current issues in speech information retrieval is automatic topic segmentation of continuous speech. Solving this problem is essential both for direct applications (topic classification of Broadcast News, telephone conversations with call-centers operators ect.), and indirect applications (for example, improvement of speech recognition results after specifying the topic of the conversation or the message) [1].

Methods of topic classification of texts are well studied and have a rich history beginning with 1960s research on naive likelihood models of texts classification [2, 3, 4]. Now the commonly applied methods are Naive Bayes classifiers [5], geometrical classifiers (such as Nearest Neighbors and Rocchio methods) [6], support vector machine (SVM) methods [7, 15, 22] and more complicated likelihood models - LSI, Aspect, LDA [10]. All of them yield steadily good results in various applications.

But direct application of topic classification methods focused on work with text to recognized speech generates a number of problems. The first one is the decrease in topic classification quality, caused by the presence of errors (deletion, substitution, insertion) inevitably occurring in automatic speech recognition of continuous speech. The second is the absence of the auxiliary printing information (headings, paragraphs, capital letters and punctuation marks) that establishes sentence and phrase boundaries. The presence of such information would allow to preliminarily segment speech into blocks which are then subject to topic classification.

Preliminary automatic detection of sentence and phrase boundaries is the first step on the way to reliable topic segmentation of continuous speech. Various approaches to this problem have been proposed: for example, a simple limitation of the number of recognized words in speech blocks [8], as well as prosody based [9] or lexical-based [10, 11, 12] automatic segmentation of speech into sentences.

However, the results presented in the literature show a rather low topic segmentation reliability when classification is performed online: values of the F-measure (see section 3.3) in different works range from 0.4 to 0.65 [13, 14]. To compare, the results of offline topic classification of separate messages (i.e., when it is known in advance that the speech fragments fed into the classification system do not contain topic change events) show sufficiently high reliability: F-measure ~ 0.95 [15].

In this paper we present an online topic segmentation method that makes use of a specially developed technology for preliminary automatic segmentation of continuous speech into sentences.

2. Target Settings and Data Preparation

This paper deals with the problem of online topic segmentation of Russian Broadcast News with a predetermined set of topics. We suppose that a speech fragment can belong to various topics simultaneously, so we apply the so-called “multitag” topic classification: the output of the system is the decision about the set of topics that the speech fragment corresponds to. The following 5 subjects of news are considered as the target topic set: crime, economy, politics, public events and culture/life.

A specially collected text database of Russian Broadcast News was used for training the classification models and algorithms. On average it contained about 10 Mb of text for each of the 5 topics. 4/5 of the training data set was used for training the classification models, and 1/5 for training the weight factors of the Fusion Method (see section 3.2).

Speech recognition was carried out by the ASR system developed at Speech Technology Center using a general language model (trained on a general Broadcast News database, dictionary size – 17000 words). In our experiments WER (word error rate) is 39%.

3. Methods

3.1. Preliminary automatic segmentation of speech into sentences

There are two basic approaches to sentence boundary detection for continuous speech: detection of coherent speech fragments (the lexical approach) and the analysis of intonational (prosodic) features. However, when the text is obtained as a result of speech recognition, the presence of some level of word error can lead to loss of syntactic and semantic coherence of the text. Prosodic features, which can be examined without analysing the contents of the text, are a more reliable basis for solving this problem [10, 16, 17].

The basic idea of the method we use is online processing of speech and prediction of the most probable positions of sentences boundaries. For that purpose, speech is segmented into 10ms fragments, for each of which the following characteristics are calculated: speech/not speech, F0 value, energy value. On the basis of these characteristics, a set of features for classifying the fragments into "boundary" and "non boundary" is formed. For this classification we used SVM (support vector machines) and Decisions Trees. The first classifier was developed at Speech Technology Center and the second is an open source project developed by ALGLIB Project company [18].

EER (equal error rate) of the system of sentence boundary detection ranges from around 17% (in case of speech database that contained studio records consisting mostly of reading) to 40% (in case of speech database containing records of various quality and different channels, with both reading and spontaneous speech).

It should be noted that a serious problem is the absence of large annotated corpora which would include information about sentence breaks. For this reason the system was trained on a number of small corpora (from 1900 to 5400 sentences boundaries in various training sets) which were available for this task. We expect that increasing the size of training databases will lead to a considerable improvement of the results.

3.2. The topic segmentation method

Topic segmentation of speech is carried out in online mode. This means that the stream of recognized words together with sentence boundary probabilities is transmitted to the classifier input. The duration of recognized text fragments fed into the classification system is regulated by means of two parameters (which can be used both separately and simultaneously): a threshold on the number of recognized words and the requirement of reaching a sentence boundary.

In section 4, we present experimental results that illustrate the dependence of topic classification reliability on the method of splitting recognized text into fragments.

The text fragment (document) which is subject to topic classification is represented in the form of a multidimensional vector whose components depend on the occurrence of each term in the given document and other documents. To define the distance between documents, the standard tf-idf weighting function (proposed in the early 1970s [19] and now actively employed by other researchers [20, 21]) is used.

In our topic segmentation system we use the following groups of classification algorithms:

- Naive Bayes classifiers (based both on the multinomial and multivariate models);
- Nearest Neighbors and Rocchio classifiers;
- Support Vector Machine (SVM): the Linear classifier, and also classifiers with Polynomial and RBF (Radial Basis function) kernels.

For the generalized classification decision, the Fusion Algorithm, suggested by Niko Brummer [23, 24] is used. This algorithm performs discriminative (logistic regression) training to form a calibrated fusion of the scores of multiple classifiers. Training is performed on a supervised database of scores.

3.3. Estimating the topic segmentation accuracy

To estimate the reliability of the topic segmentation method, we compare the automatic topic segmentation of speech data to an expert manual segmentation. The resulting F-measure (the balanced Van Rijsbergen's effectiveness measure) was calculated separately for each topic, and an average was calculated for all topics:

$$F = \frac{2 \cdot precision \cdot recall}{precision + recall}.$$

Various approaches to determining precision and recall in classification are possible. If smoothing post-processing methods of the segmentation results are applied (for example, merging speech fragments that are classified as belonging to the same topic and are divided by small time intervals), it makes sense to take into account the number of correct topic boundaries. In our experiments post-processing methods were not applied, so classification precision and recall were based on the duration of correctly classified speech:

$$precision = \frac{CorrectTopicDuration_{AUTOMAT}}{TopicDuration_{AUTOMAT}},$$

$$recall = \frac{CorrectTopicDuration_{AUTOMAT}}{TopicDuration_{EXPERT}},$$

where: $TopicDuration_{EXPERT}$ is the complete duration of the speech fragments belonging to the topic in the expert manual marking; $TopicDuration_{AUTOMAT}$ is the complete duration of the speech fragments belonging to the topic in the automatic segmentation; $CorrectTopicDuration_{AUTOMAT}$ is the complete duration of the speech fragments correctly identified as belonging to the topic in the automatic segmentation (i.e. the overlap of the automatic and expert marking).

4. Experimental Results and Discussions

The experimental database consisted of recordings of Russian broadcast news (radio and TV). The test set comprised about 3 hours of speech and contained both studio recordings and telephone conversations that were broadcast live. About 50% of the test set were fragments of prepared speech or reading and the other 50% was spontaneous speech. Part of the test speech data (about 20%) had background music, which is typical for radio.

Table 1 shows the amounts of data belonging to different topics in the test set. The test set contained 128 events of topic change.

Table 1. Distribution of topics in the test set.

Topic	**Speech Duration, minutes**
Crime	8
Economics	14
Public Events	20
Politics	81
Culture / Life	26

We compared two approaches to segmenting speech into the fragments used for classification:

- A method that does not use any information about sentence boundaries. In this case, the length of speech fragments that were to be classified was limited only by specifying the number of recognized words. The results are shown in Table 2.
- A method that uses information about sentence boundaries. In this case, the length of speech fragments that were to be classified was limited both by specifying the number of recognized words and by the requirement of reaching a sentence boundary. The results are shown in Table 3.

The word number thresholds were 5 and 15 words. Numbers exceeding 15 were not used because the database contained a large number of small news messages.

Table 2. Topic segmentation results (without sentence boundary information)

Topic	**F-measure (precision, recall)**	
	5 word threshold	15 word threshold
Crime	0.26 (0.18, 0.47)	0.32 (0.25, 0.46)
Economics	0.41 (0.34, 0.54)	0.48 (0.4, 0.6)
Public Events	0.34 (0.31, 0.38)	0.34 (0.35, 0.33)
Politics	0.48 (0.8, 0.35)	0.62 (0.84, 0.49)
Culture / Life	0.41 (0.27, 0.83)	0.44 (0.31, 0.79)
All topics	0.41 (0.37, 0.46)	0.50 (0.47, 0.53)

Table 3. Topic segmentation results (with sentence boundary information)

Topic	**F-measure (precision, recall)**		
	0 word threshold + sentence boundaries	5 word threshold + sentence boundaries	15 word threshold + sentence boundaries
Crime	0.66 (0.72, 0.61)	0.37 (0.36, 0.38)	0.40 (0.44, 0.36)
Economics	0.67 (0.74, 0.61)	0.76 (0.69, 0.85)	0.79 (0.74, 0.85)
Public Events	0.74 (0.71, 0.77)	0.60 (0.63, 0.58)	0.60 (0.63, 0.58)
Politics	0.84 (0.86, 0.82)	0.89 (0.87, 0.9)	0.90 (0.87, 0.93)
Culture / Life	0.67 (0.59, 0.78)	0.70 (0.67, 0.73)	0.71 (0.68, 0.73)
All topics	0.77 (0.76, 0.77)	0.78 (0.76, 0.8)	**0.79 (0.78, 0.81)**

The results show that using sentence boundary information increases the accuracy of online topic classification of continuous speech by the average 25-30%. Maximum classification accuracy (F-measure = 0.79) is obtained when both sentence boundary information and a word number threshold of 15 words are used. However, even when only the information about sentence boundaries (without a word number threshold) is available, the results show a high level of classification accuracy (F-measure = 0.77).

For the topics “Crime” and “Public events” there is a sharp decline in classification accuracy when a word number threshold is set (Table 3). This can be explained by the fact that these topics are mostly present in news digests and not in long thematic programs, so they generally have relatively short duration.

We have also noticed that frequently, the following situation occurs: long programs (lasting several minutes) that were marked by the expert as belonging to one topic are also marked by the classifier as belonging to the correct topic; however, inside this program the classifier finds shorter fragments that are classified as belonging to a different topic. In fact, listening to these short fragments taken out of context confirms that they indeed belong to the “wrong” topic. For instance, in a program about parliamentary elections, classified by the expert as belonging to the topic “Politics”, the speaker talked for a while about the economical situation in the country, which went unnoticed by the expert but was correctly identified by the automatic classifier. However, even though such “sensitivity” of the classifier seems justifiable, such situations were regarded as errors because the main task was to maximize the similarity between the classifier’s output and the expert’s. The decision to treat such situations as “correct” instead of “false alarm error” can be taken in specific applications of the classifier.

5. Conclusions

We propose the online method of topic classification of speech that uses information about sentence boundaries obtained from an automatic sentence boundary detector. The experimental results demonstrate that using sentence boundary information improves the performance of the topic classifier. The topic classifier that uses this method was shown in our experiments to have the F-measure = 0.79.

For topic classification in offline mode (when fragments that are to be classified do not contain topic change events), F-measure in our experiments exceeds 0.9 for every topic. This is the level of accuracy that we would like to reach in the online topic classification method as well. We expect that training the sentence boundary detector on more data will lead to its improved performance and to increased accuracy of the topic classifier.

References

1. *Nikolenko S.I., Levin K.E., Khokhlov Yu.Yu.* Two-pass automatic speech recognition with use of text mining. (*Николенко С.И., Левин К.Е., Хохлов Ю.Ю.* автоматическое распознавание речи с использованием интеллектуального анализа текстов. Труды конференции «Интегрированные модели, мягкие вычисления, вероятностные системы и комплексы программ в искусственном интеллекте» (ИММВИИ-2009). В 2-х тт., Физматлит, 2009. Т. 1. С. 192-202.)

2. *Minsky M.* Steps toward Artificial Intelligence. Proceedings of the IRE 49(1): 8-30, 1961

3. *Maron M. E.* Automatic Indexing: An Experimental Inquiry. Journal of the ACM (JACM) 8(3): 404–417, 1961

4. *Harold Borko, Myrna Bernick.* Automatic Document Classification. J. ACM 10(2): 151-162, 1963

5. *Zhang H.* The Optimality of Naive Bayes. Proceedings of the FLAIRS Conference / Ed. by V. Barr, Z. Markov, 2004

6. *Joachims T.* A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. Proceedings of the 14th International Conference on Machine Learning (ICML-97). P. 143–151, 1997

7. *T. Hofmann, B. Scholkopf , A. J. Smola.* Kernel Methods in Machine Learning. The Annals of Statistics, Vol. 36, No. 3, 1171–1220, 2008

8. *C. Chemudugunta, P. Smyth, M. Steyvers*. Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model. Advances in Neural Information Processing Systems, 2007, 19

9. *Sadaoki Furui, Katsutoshi Ohtsuki and Zhi-Peng Zhang*. Japanese Broadcast News Transcription and Information Extraction. COMMUNICATIONS OF THE ACM, Vol. 43, No. 2, February 2000

10. *E.Shriberg, A.Stolcke, D.Hakkani-Tur, and G.Tur*. Prosody based automatic segmentation of speech into sentences and topics. Speech Comm., 32(1-2):127-154, 2000

11. *N. Stokes*. Spoken and written news story segmentation using lexical chains. In Proc. of the Student Workshop at HLT-NAACL2003, 49.53, 2003

12. *D. Beeferman, A. Berger, and J. Lafferty*. Statistical models for text segmentation. Machine Learning, 31:177.210, 1999

13. *A. Rosenberg, Ju. Hirschberg*. Story Segmentation of Brodcast News in English, Mandarin and Arabic. Human Language Technology Conference of the North American Chapter of the ACL, pages 125–128, New York, June 2006

14. *C.Guinaudeau, G.Gravier, P.Sebillot*. Improving ASR-based topic segmentation of TV programs with confidence measures and semantic relations. Interspeech 2010, pages 1365 – 1368, Chiba, September 2010

15. *R.Torres, Sh.Takeuchi, H.Kawanami, T.Matsui, H.Saruwatari, K.Shikano*. Comparison of Methods for Topic Classification in a Speech-Oriented Guidance System. Interspeech 2010, pages 1261 – 1264, Chiba, September 2010

16. *Y. Liu, A. Stolcke, E. Shriberg, and M. Harper*. Using conditional random fields for sentence boundary detection in speech, in Proc. of ACL-05, pp. 451–458, 2005

17. *B. Roark, Y. Liu, M. Harper, R. Stewart, M. Lease, M. Snover, I. Shafran, B. Dorr, J. Hale, A. Krasnyanskaya, and L. Yung*. Reranking for sentence boundary detection in conversational speech. In ICASSP, 2006

18. http://www.alglib.net/aboutus.php

19. *Sparck Jones K.* A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation. Vol. 28, N. 1. P. 11–21, 1972

20. *Berry M.* Survey of Text Mining: Clustering, Classification, and Retrieval. Springer, 2003

21. *Feldman R., Sanger J.* The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, 2006

22. *Joachims T.* Learning to Classify Text using Support Vector Machines: Methods, Theory, and Algorithms. Kluwer Academic Publishers, Springer, 2002

23. *Niko Brummer*, Measuring, refining and calibrating speaker and language information extracted from speech, Ph.D. dissertation, Stellenbosch University, to be submitted 2007

24. *Niko Brummer*. FoCal Multi-class: Toolkit for Evaluation, Fusion and Calibration of Multi-class Recognition Scores. Tutorial and User Manual. Spescom DataVoice, June 2007