CUNI at MediaEval 2013 Similar Segments in Social Speech Task

Petra Galuščáková and Pavel Pecina
Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Prague, Czech Republic
{galuscakova,pecina}@ufal.mff.cuni.cz

ABSTRACT

We describe our experiments for the Similar Segments in Social Speech Task at MediaEval 2013 Benchmark. We mainly focus on segmentation of the recordings into shorter passages on which we apply standard retrieval techniques. We experiment with machine-learning-based segmentation employing textual (word n-grams, tag n-grams, letter cases, lexical cohesion, etc.) and prosodic features (silence) and compare the results with those obtained by regular segmentation.

1. INTRODUCTION

The main aim of the Similar Segments in Social Speech Task is to find segments similar to the given ones (query segments) in the collection of audio-visual recordings containing English dialogues of a university student community. In addition to the human and automatic (ASR) transcripts (both transcripts are given separately for each speaker), the collection also contains prosodic features and metadata. The training data consists of segments manually assigned to similarity sets of the query segments. The details of the task and data are described in the task description [7].

2. APPROACH DESCRIPTION

In our experiments, the queries are created from the human transcripts of the query segments. The recordings are segmented into overlapping passages (identified by their starting and ending times) which are then indexed using the Terrier IR Platform [6]. The set of potential jump-in points needed in retrieval then consists of the known beginnings of the acquired segments.

For the indexing, we use the default settings, which outperformed our most successful setting from previous experiments in the Search and Hyperlinking MediaEval Benchmark [3]. We remove stopwords and apply stemming using the Porter stemmer. Ranked lists of retrieved segments are pruned by removing segments overlapping with those ranked higher.

As both transcripts are given in separated tracks for each speaker, we join these tracks into a single one. In the human transcripts, we sort sentences from both transcripts according to their beginnings to acquire single sequential transcript. Similarly, we sort the speakers' segments given in the ASR transcripts. While in the ASR transcripts the exact playback time is given for each word, in the human transcripts such information is available only on sentence level and therefore we approximate it by assuming equal duration of words in a sentence.

2.1 Query processing

The query segments are specified by their starting and ending time. The queries are constructed by including all words lying within the boundaries of the query segment in both tracks.

We tried to expand the queries by adding words appearing in the vicinity of the query segment (allowing ± 5, ± 10, ± 15, ± 20, ± 30, and ± 60 seconds) but none of these experiments improved the results.

We also attempted to generate the queries from both the human and ASR transcripts and apply them to search in both types of transcripts. The queries created from the human transcripts achieved higher scores when applied on both the human and ASR transcripts, therefore they are used in the experiments presented in this paper.

2.2 Segmentation

In this work, we mainly focus on segmentation of the recordings, which appears to be crucial for segment retrieval [2]. We experiment with regular segmentation and two methods based on (supervised) machine learning (ML).

In regular segmentation, the recordings are divided into equilong segments of 50 seconds (which is approximately equal to the average segment length in the collection). The shift between the segments (and the overlap) is also regular, set to 25 seconds, since according to our experience from the 2012 Search and Hyperlinking task, the shift of 10 to 30 seconds achieves optimal results [2].

In the first ML approach, we identify segment boundaries using classification trees [1], implemented in the *rpart* library in R. For each word in the transcripts, we assume that it belongs to a segment and detect whether it is followed by a segment boundary, or the segment continues. Class distribution in this task (*segment boundary* vs. *segment continuation*) is highly unbalanced and the corresponding weights must be set accordingly to prevent too short segments. We set the weight of *segment boundary* misclassified as *segment continuation* in the loss matrix to 21, the weight of the *segment continuation* misclassified as *segment boundary* to 11, and the complexity parameter to 0.

In the second ML approach, we apply a similar process to detect beginnings of segments which are then set to be 50

MediaEval 2013 Workshop, October 18-19, 2013, Barcelona, Spain

Segmentation beginings	Segmentation ends	Normalized SUR	Normalized Recall	F-measure
REG	REG	0.57	0.78	0.58
ML	REG	**0.65**	**0.90**	**0.67**
ML	ML	0.59	0.80	0.61

Table 1: Retrieval results on the human transcripts.

Segmentation beginings	Segmentation ends	Normalized SUR	Normalized Recall	F-measure
REG	REG	**0.87**	**1.19**	**0.90**
ML	REG	0.70	1.00	0.72
ML	ML	0.65	0.90	0.67

Table 2: Retrieval results on the ASR transcripts.

seconds long (naturally, the segments can overlap). In this case, we aim at higher recall of the decision process to find all possible segment beginnings, but still keep the number of created segments reasonable. We set the weight of the *segment boundary* misclassified as *segment continuation* in the loss matrix to 61, the weight of the *segment continuation* misclassified as *segment boundary* to 1, and the complexity parameter to 0.

For comparison, the classification models trained and tuned on the human transcripts are also applied on the ASR transcripts despite their mutual inconsistency. The transcripts differ in the length of silence (which is in human transcripts only approximated as the duration between the imprecise word beginnings), tokenization, and letter capitalization. Therefore, our future plans include to train the classification model on the ASR transcripts too.

2.3 Features

Our classification model exploits the following features: cue words and cue tags, letter cases, length of the silence before the word, division given in transcripts, and the output of the TextTiling algorithm [4].

The cue words are the words that appear frequently at the segment boundary and often do not carry special meaning. Based on the training data, we have identified words which frequently stand at the segment boundary and words which are the most informative for the segment boundary (the mutual information between these words and the segment boundary is high). We have also defined our own set of words which might occur at such boundary and created sets for unigrams, bigrams and trigrams, for words and tags (obtained by Featurama tagger [5]) for both segment beginnings and ends. Occurrence of each n-gram is captured by a separate feature. An additional feature indicates whether at least one feature from the set (n-grams for frequent words, informative words and defined words for either beginning or end) occurs.

As the TextTiling algorithm is based on calculating similarity between adjacent regions, utilizing its output, we can also employ lexical cohesion into our decision process.

3. RESULTS

We employ three automatic evaluation measures: Normalized Searcher Utility Ratio (SUR), Normalized Recall, and the F-measure (for details, see the task description [7]). The results for various types of segmentation for the human transcripts are displayed in Table 1 and for the ASR transcripts in Table 2.

In the experiment utilizing human transcripts, the ML-based segmentation outperforms the regular segmentation. However in the experiments with the ASR transcripts, the regular segmentation wins. In both cases, ML-based segmentation searching for segment beginnings outperforms ML-segmentation searching for entire segments.

In the overall results, the ASR transcripts surprisingly outperform human transcripts. This is probably caused by the approximation of word timing and duration in the human transcripts – in the ASR transcripts, we are able to determine precise segment beginning and end times but the times in the human transcripts are inaccurate.

4. CONCLUSIONS AND FUTURE WORK

The overall best result is achieved using regular segmentation on the ASR transcripts. For the human transcripts, however, the proposed ML-based segmentation outperformed the regular segmentation, which is very promising and we will attempt to project this results into experiments using the ASR transcripts. In our future work, we would also like to employ a joint model for identification of both segment beginnings and the segment ends.

5. ACKNOWLEDGMENTS

This research is supported by the Charles University Grant Agency (GA UK n. 920913) and the Czech Science Foundation (grant n. P103/12/G084).

6. REFERENCES

[1] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.

[2] M. Eskevich, G. J. Jones, R. Aly, R. J. Ordelman, S. Chen, D. Nadeem, C. Guinaudeau, G. Gravier, P. Sébillot, T. de Nies, P. Debevere, R. V. de Walle, P. Galuščáková, P. Pecina, and M. Larson. Multimedia information seeking through search and hyperlinking. In *Proc. of ICMR*, pages 287–294, Dallas, Texas, USA, 2013.

[3] P. Galuščáková and P. Pecina. CUNI at MediaEval 2012 Search and Hyperlinking Task. In *MediaEval 2012 Workshop*, volume 927, Pisa, Italy, 2012.

[4] M. A. Hearst. TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. *Computational Linguistics*, 23(1):33–64, 1997.

[5] M. Spousta. Featurama – a library that implements various sequence-labeling algorithms. http://sourceforge.net/projects/featurama/.

[6] Terrier IR Platform. An open source search engine. http://terrier.org/.

[7] N. G. Ward, S. D. Werner, D. G. Novick, E. E. Shriberg, C. Oertel, L.-P. Morency, and T. Kawahara. The Similar Segments in Social Speech Task. In *MediaEval 2013 Workshop*, Barcelona, Spain, October 18-19 2013.