 ISBN 978-80-7378-224-5

Application of Topic Segmentation in Audiovisual Information Retrieval

P. Galuščáková

Charles University in Prague, Faculty of Mathematics and Physics, Prague, Czech Republic.

Abstract. Segmentation into topically coherent segments is one of the crucial points in information retrieval (IR). Suitable segmentation may improve the results of IR system and help users to find relevant passages faster. Segmentation is especially important in audiovisual recordings, in which the navigation is difficult. We present several methods used for topic segmentation, based on textual, audio and visual information. The proposition of our approach to topic segmentation based on the fusion of audio and visual data is presented in the article.

Introduction

Information Retrieval could be defined as *"finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)"* [Manning et al. (2008)]. We are interested in the case when the documents are in audiovisual format and the queries on the collection are in textual format. This kind of retrieval could be converted into retrieval in textual documents using automatically processed audio track of the recording [Heeren et al. (2009)].

But the retrieval in a speech or audiovisual documents still differs from the retrieval in textual documents. The biggest difference is that no indentation is given in speech. The sentence boundaries are difficult to identify in speech, similarly to the segmentation into more complex units. The transcripts also contain numerous errors and the vocabulary used in speech differs from the vocabulary used in texts. On the other hand, speech contains different type of information, such as pauses, which could be useful for segmentation into coherent parts. Such topic segmentation may, in turn, influence the quality of information retrieval in audiovisual recordings.

Topic Segmentation

According to Morris and Hirst (1991) the sentences and the phrases of any sensible text refer to the same topic and text thus has a quality of unity. This cohesion could be achieved using particular means, such as back-reference, conjunction and semantic word relations.

The nature of topics is hierarchical. For instance, a whole book could be considered as one topically coherent segment. In a closer view, chapters could be recognised as book's subsegments and the paragraphs are the subsegments of the chapters. Sometimes, one sentence or n-gram form a topically coherent segment. Therefore, the exact definition of a topic varies with the utilization of the segmentation.

Eskevich and Jones (2011) showed that the quality of topic segmentation used in the recordings affects the quality of information retrieval. When comparing TextTiling and C99 algorithms, the retrieval worked significantly better if TextTiling segmentation algorithm was employed in their experiments. Eskevich et al. (2012) then applied several segmentation methods, including TextTiling segmentation and regular segmentation according to the number of words in text, and compared the results of the information retrieval in dependency on a segmentation type.

Topic segmentation of an audiovisual recording could be based on the segmentation of the transcript of the audio track of the recording. The influence of the transcripts errors on the segmentation quality is ambiguous. Malioutov et al. (2007) showed differences in the evaluation of segmentation algorithms in dependency on employing either manual or automatic transcripts. On the other hand, Hsueh and Moore (2007) showed that despite the word recognition error (word error rate is equal to 39.1%), their segmentation system was not significantly worse if it worked with automatically processed transcripts than if it worked with reference transcripts. But Hsueh and Moore (2007) used also audio and visual features, which could have reduced the number of errors.

Approaches to Topic Segmentation

Topic segmentation algorithms can be divided into three categories:

1. **Lexical-Cohesion-Based**
   These systems use only textual information. They are based on the assumption that the segment we are looking for is lexically coherent (for instance it uses coherent vocabulary). Examples of such systems are TextTiling [Hsu et al. (2004)], C99 [Choi (2000)], LCSeg [Galley et al. (2003)], MinCut [Malioutov and Barzilay (2006)], Dotplot [Kumar et al. (2011)], IClustSeg [Pérez and Pagola (2010)], TextLec [Rojas and Pagola (2007)], DivSeg [Song et al. (2011)], NM09 [Niekrasz and Moore (2009)], U00 [Utiyama and Isahara (2001)], JSeg [Bartkova (2006)], Transeg [Labadié and Prince (2008)], LCP [Kozima (1993)], TopSeg [Hsu et al. (2004)].

2. **Feature-Based**
   Particular features are mined from the text or recording when this system is being used. Then, a selected machine learning technique is applied. Examples of such systems are [Hsueh and Moore (2007)], [Franz et al. (2003)], [Kauchak and Chen (2005)], [Passonneau and Litman (1997)], [Tür et al. (2001)], [Kauchak and Chen (2005)].

3. **Model-Based**
   These systems are based either on Hidden Markov Models [Jeong and Titov (2010); Tür et al. (2001)] or on some sampling methods [Eisenstein and Barzilay (2008); Utiyama and Isahara (2001)].

Lexical-Cohesion-Based Systems

The basic property of the lexical-cohesion-based systems is a metric which is applied to calculate similarities between the segments. The systems also approach differently in the way of extracting the segments (between which the similarity is then calculated) and postprocessing calculated similarity.

C99 and TextTiling are two most frequently used lexical-cohesion-based systems. Both are based on the cosine similarity. The cosine similarity between segments x and y is calculated as follows:

$$sim(x,y) = \frac{\sum_j f_{x,j} * f_{y,j}}{\sqrt{\sum_j f_{x,j}^2 * \sum_j f_{y,j}^2}}$$

where $f_{i,j}$ denotes the frequency of word j in segment i.

In C99 algorithm the similarity between each pair of sentences is calculated and the similarity matrix is based on these values. Subsequently, ranked matrix is created according to the values of the similarity matrix. Each value in the similarity matrix is replaced by its rank – a number of neighbouring segments with a lower similarity value [Choi (2000)]. Finally, clustering is applied and the regions with maximum density are located in the ranked matrix. These regions form coherent segments.

In contrast, a fixed-length window is slid through the text in TextTiling algorithm. The cosine distance between the left and the right side of the window is calculated and this information is converted into a digital signal. Signal is eventually smoothed and the shape of this post-processed signal determines the segment breaks [Hearst (1993)].

Feature-Based Systems

The advantage of this approach comparing with the lexical-cohesion-based approach is that the information from all modalities can be used. Authors who used this approach present many possible features but not all of them have been actually used. We can divide these features according to their modality:

1. **Textual Features**

   Lexical Features such as cue words and n-grams that often stand at the beginning or at the end of the segment (e.g. now, okay, let's, um, so, good night ...) [Hsueh and Moore (2007)]. Or, only selected word classes can be used, for instance Franz et al. (2003) uses only nouns.

   Contextual Features such as dialogue act type [Hsueh and Moore (2007)] or tense [Niekrasz and Moore (2009)]. It is also possible to use the role of a speaker, if it is available (e.g. interlocutor, guest or project manager).

   Vocabulary used in a segment. Part of speech tags, some groups of words such as months, days, country names or named entities can be used as well. Hsueh and Moore (2007) claim that a interlocutor often mentions agenda items (e.g. presentation, meeting) around the segment boundary. Information about whether the segment contains a pronoun, a number or a direct speech could also be used.

Lexical Chains and information about whether the word appears in the next or previous few word/sentences are suggested to be used [Beeferman et al. (1999); Kauchak and Chen (2005)].

2. **Audio Features**

   Prosodic Features such as fundamental frequency (maximum, mean or its pattern across the segment boundary) [Tür et al. (2001)], energy at multiple points, the pitch contour [Tür et al. (2001)], which must be relative to the speaker's baseline, the rate of speech (number of words and the number of syllables spoken per second), silence [Bartkova (2006)] and the duration of pauses [Shriberg et al. (2000)], vowels [Bartkova (2006)] and especially final vowels and final rhymes [Tür et al. (2001)]. The length of silence seems to be particularly important [Hsueh and Moore (2007)]. Shriberg et al. (2000) observed that the segment is likely to start with higher pitched sounds and with a lower rate of speech and the speaker have a tendency to fall in pitch and, again, to decelerate at the end of the segment.

   Conversational Features for instance the amount of overlapping speech [Hsueh and Moore (2007)] and the frequency of speaker's change [Niekrasz and Moore (2009)].

3. **Visual Features**

   Color Similarity which is mainly based on a histogram.

   Motion Similarity usually uses pixel comparison. Hsueh and Moore (2007) showed that frontal shots and a hand movement could be a significant feature. Eisenstein and Barzilay (2008) introduced segmentation based on gestural features, especially on the eye gaze behaviour. Some authors recall edges detection as another applicable feature.

   Bag of Visual Words is a method based on the detection of several points and regions of interest and their further classification according to the created dictionary of visual features.

More features could be also combined into a single one. For instance, some authors proclaim that lexical features should be particularly combined with the audio and the visual features, such as pauses or speaker changes.

Multimodal Fusion Models

Llinas et al. (2004) defined information fusion as *"an information process that associates, correlates and combines data and information from single or multiple sensors or sources to achieve refined estimates of parameters, characteristics, events and behaviours"*. Fusion thus can be an effective method for mixing various data modality in the segmentation process. It could especially help to make segmentation more robust.

There are three basic levels of fusion: early, intermediate and late fusion. The early fusion strategy concatenates all possible features and only one decision is taken using all of these features. In contrast, the late fusion strategy takes one decision for each data source. In the intermediate fusion strategy different feature vectors are created and further processed by Hidden Markov Models.

Proposed Solution

At first, we would like to examine topical segmentation in Search subtask of Search and Hyperlinking task in MediaEval[1] Benchmark. The main aim of the Search subtask is to find exact segments relevant to the given query in the collection of audiovisual data. Therefore, we will segment the recordings and use IR system to retrieve the relevant segment. Created system should be further adapted and applied in Dialogy corpus[2]. As a result, system will enable information retrieval in this corpus.

As far as the proposed system is concerned, there are several points that must be taken into account. For instance, the most of the recordings provided by MediaEval Benchmark are in English but the most of the recordings in Dialogy corpus are in Czech. Our system should therefore be language independent. The type of the data also differs; MediaEval works with the unprofessional news from Blip.tv[3] and Dialog corpus consists of Czech Television discussion programmes. The size of training data in MediaEval and

[1] http://www.multimediaeval.org

[2] http://ujc.dialogy.cz

[3] http://blip.tv

in Dialogy corpus is expected to be small but it would be possible to use another type of data for training, such as TDT[4] or Malach[5] corpora.

We would like to use all kinds of modality in the proposed system. Therefore, machine learning method based on the features appears to be the best solution of our problem. Different modality features will be mixed together using a fusion. Our method thus will be similar to the approach described by Hsu et al. (2004). Hsu employs Maximum Entropy statistical model with following features: the anchor face, commercials detection, pitch jump, length of pauses, speech segments, speech rapidity and a variety of lexical, semantic and structural features calculated from ASR transcripts. The most of the features used in our system will be text-based. The output of cohesion-based algorithm (TextTiling) will be one of them because this system achieved good results (comparing to another cohesion-based algorithm) in [Eskevich and Jones (2011)]. Smaller amount of visual features will be used in our experiments. We will put the emphasis on using shot detection, which is available with MediaEval data and could also be relatively easy to obtain in Dialogy corpus.

We would also like to enable users to correct the segmentation in Dialogy corpus. Therefore, it should be possible to incorporate these corrections to improve the segmentation. Some kind of active learning technique may be used for this purpose.

For the retrieval in the set of created segments we will use one of available open source IR software. Middleton and Baeza-yates (2007) give an overview of such systems with their positive sides and drawbacks. Precision and recall of retrieval are especially important for us. According to the results of the comparison, we will use Terrier[6] search engine. Not only Terrier achieved good results in the comparison, but it also offers a wide range of possible settings.

Conclusion

We described various methods used for segmentation of audiovisual recordings to allow their further usage in information retrieval process. The closer description of several algorithms based on a lexical cohesion was given, as well as the list of features which may be used by the feature-based segmentation algorithms. Eventually, the projects in which we would like to employ the segmentation were described and a solution was proposed with a respect to the given attributes of these projects.

Acknowledgments. This research was supported by the Ministry of Culture of the Czech Republic (project AMalach, program NAKI, grant n. DF12P01OVV022), the Czech Science Foundation (grant n. P103/12/G084) and Ministry of Education of the Czech Republic (project LINDAT-Clarin, grant n. LM2010013) and SVV project number 265 314.

References

Bartkova, K., How far can prosodic cues help in word segmentation?, in *Proceedings of the 3rd International Conference on Speech Prosody SP2006*, 2006.

Beeferman, D., Berger, A., and Lafferty, J., Statistical models for text segmentation, in *Machine Learning*, pp. 177–210, 1999.

Choi, F. Y. Y., Advances in domain independent linear text segmentation, in *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, NAACL 2000, pp. 26–33, Association for Computational Linguistics, Stroudsburg, PA, USA, 2000.

Eisenstein, J. and Barzilay, R., Bayesian unsupervised topic segmentation, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pp. 334–343, Association for Computational Linguistics, Stroudsburg, PA, USA, 2008.

Eskevich, M. and Jones, G. J. F., Dcu at mediaeval 2011: Rich speech retrieval, in *Working Notes Proceedings of the MediaEval 2011 Workshop, Santa Croce in Fossabanda, Pisa, Italy, September 1-2, 2011*, edited by M. Larson, A. Rae, C.-H. Demarty, C. Kofler, F. Metze, R. Troncy, V. Mezaris, and G. J. F. Jones, vol. 807 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2011.

Eskevich, M., Jones, G. J. F., Wartena, C., Larson, M., Aly, R., Verschoor, T., and Ordelman, R., Comparing Retrieval Effectiveness of Alternative Content Segmentation Methods for Internet Video Search, in *CBMI 2012*, pp. 1–6, URL `http://dx.doi.org/10.1109/CBMI.2012.6269810`, 2012.

Franz, M., Ramabhadran, B., Ward, T., and Picheny, M., Automated transcription and topic segmentation of large spoken archives, in *INTERSPEECH*, ISCA, 2003.

Galley, M., McKeown, K., Fosler-Lussier, E., and Jing, H., Discourse segmentation of multi-party conversation, in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pp. 562–569, Association for Computational Linguistics, Stroudsburg, PA, USA, 2003.

[4] `http://projects.ldc.upenn.edu/TDT/`

[5] `http://malach.umiacs.umd.edu/`

[6] `http://terrier.org/`

Hearst, M. A., Texttiling: A quantitative approach to discourse, Tech. rep., University of California at Berkeley, Berkeley, CA, USA, 1993.

Heeren, W., van der Werff, L., de Jong, F., Ordelman, R., Verschoor, T., van Hessen, A., and Langelaar, M., Easy Listening: Spoken Document Retrieval in CHoral, *Interdisciplinary Science Reviews*, *34*, 236–252, 2009.

Hsu, W., Chang, S.-f., Huang, C.-w., Kennedy, L., Lin, C.-y., and Iyengar, G., Discovery and fusion of salient multi-modal features towards news story segmentation, in *IS&T/SPIE Electronic Imaging*, 2004.

Hsueh, P.-Y. and Moore, J. D., Combining multiple knowledge sources for dialogue segmentation in multimedia archives, in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 1016–1023, Association for Computational Linguistics, Prague, Czech Republic, 2007.

Jeong, M. and Titov, I., Multi-document topic segmentation, in *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pp. 1119–1128, ACM, New York, NY, USA, 2010.

Kauchak, D. and Chen, F., Feature-based segmentation of narrative documents, in *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, FeatureEng '05, pp. 32–39, Association for Computational Linguistics, Stroudsburg, PA, USA, 2005.

Kozima, H., Text segmentation based on similarity between words, in *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, ACL '93, pp. 286–288, Association for Computational Linguistics, Stroudsburg, PA, USA, 1993.

Kumar, N., Rai, P., Pulla, C., and Jawahar, C. V., Video scene segmentation with a semantic similarity, in *Proceedings of the 5th Indian International Conference on Artificial Intelligence, IICAI 2011*, edited by B. Prasad, P. Lingras, and R. Nevatia, pp. 970–981, IICAI, 2011.

Labadié, A. and Prince, V., Lexical and semantic methods in inner text topic segmentation: A comparison between c99 and transeg, in *Proceedings of the 13th international conference on Natural Language and Information Systems: Applications of Natural Language to Information Systems*, edited by E. Kapetanios, V. Sugumaran, and M. Spiliopoulou, vol. 5039 of *NLDB '08*, pp. 347–349, Springer-Verlag, Berlin, Heidelberg, 2008.

Llinas, J., Bowman, C., Rogova, G., Steinberg, A., Waltz, E., and White, F., Revisiting the jdl data fusion model ii, in *In P. Svensson and J. Schubert (Eds.), Proceedings of the Seventh International Conference on Information Fusion (FUSION 2004*, pp. 1218–1230, 2004.

Malioutov, I. and Barzilay, R., Minimum cut model for spoken lecture segmentation, in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pp. 25–32, Association for Computational Linguistics, Stroudsburg, PA, USA, 2006.

Malioutov, I., Park, A., Barzilay, R., and Glass, J., Making sense of sound: Unsupervised topic segmentation over acoustic input, in *In Proceedings, ACL*, 2007.

Manning, C., Raghavan, P., and Schütze, H., *Introduction to Information Retrieval*, Cambridge University Press, New York, NY, USA, 2008.

Middleton, C. and Baeza-yates, R., A comparison of open source search engines, Tech. rep., Universitat Pompeu Fabra, Barcelona, Spain, 2007.

Morris, J. and Hirst, G., Lexical cohesion computed by thesaural relations as an indicator of the structure of text, *Comput. Linguist.*, *17*, 21–48, 1991.

Niekrasz, J. and Moore, J., Participant subjectivity and involvement as a basis for discourse segmentation, in *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '09, pp. 54–61, Association for Computational Linguistics, Stroudsburg, PA, USA, 2009.

Passonneau, R. J. and Litman, D. J., Discourse segmentation by human and automated means, *Comput. Linguist.*, *23*, 103–139, 1997.

Pérez, R. A. and Pagola, J. E. M., Text segmentation by clustering cohesion, in *Proceedings of the 15th Iberoamerican congress conference on Progress in pattern recognition, image analysis, computer vision, and applications*, CIARP'10, pp. 261–268, Springer-Verlag, Berlin, Heidelberg, 2010.

Rojas, L. H. and Pagola, J. E. M., Textlec: a novel method of segmentation by topic using lower windows and lexical cohesion, in *Proceedings of the Congress on pattern recognition 12th Iberoamerican conference on Progress in pattern recognition, image analysis and applications*, CIARP'07, pp. 724–733, Springer-Verlag, Berlin, Heidelberg, 2007.

Shriberg, E., Stolcke, A., Hakkani-Tür, D., and Tür, G., Prosody-based automatic segmentation of speech into sentences and topics, *Speech Commun.*, *32*, 127–154, 2000.

Song, F., Darling, W. M., Duric, A., and Kroon, F. W., An iterative approach to text segmentation, in *Proceedings of the 33rd European conference on Advances in information retrieval*, ECIR'11, pp. 629–640, Springer-Verlag, Berlin, Heidelberg, 2011.

Tür, G., Stolcke, A., Hakkani-Tür, D., and Shriberg, E., Integrating prosodic and lexical cues for automatic topic segmentation, *Comput. Linguist.*, *27*, 31–57, 2001.

Utiyama, M. and Isahara, H., A statistical model for domain-independent text segmentation, in *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, pp. 499–506, Association for Computational Linguistics, Stroudsburg, PA, USA, 2001.