CUNI at MediaEval 2014 Search and Hyperlinking Task: Visual and Prosodic Features in Hyperlinking

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

Martin Kruliš, Jakub Lokoč
Charles University in Prague
Faculty of Mathematics and Physics
Department of Software Engineering
{krulis,lokoc}@ksi.mff.cuni.cz

ABSTRACT

In this report, we present our experiments performed for the Hyperlinking part of the Search and Hyperlinking Task in MediaEval Benchmark 2014. Our system successfully combines features from multiple modalities (textual, visual, and prosodic) and confirms the positive effect of our former method for segmentation based on Decision Trees.

1. INTRODUCTION

The main aim of the Hyperlinking sub-task is to find segments similar to a given (query) segment in the collection of audio-visual recordings. Created hyperlinks enable users to browse the collection and thus improve exploratory search ability and add entertainment value to the collection [2].

The data consists of 1335 hours of BBC Broadcast recordings available for training and 2686 hours available for testing. In our experiments, we exploit subtitles, automatic speech recognition transcripts by LIMSI [9], LIUM [11] and NST-Sheffield [10], visual features (shots and keyframes) [5], and prosodic features, all available for the task [4].

2. SEARCH SYSTEM

Our search system for the Hyperlinking sub-task is identical to the system used in the Search sub-task [6]. We apply the same retrieval model with the same settings and segmentation methods – the fixed-length segmentation and the segmentation employing Decision Trees (DT). The length of the segment used in the Hyperlinking was tuned on the training data and set to 50 seconds. Similarly to the Search sub-task, we also exploit metadata by appending metadata of the recordings to the text (subtitles/transcripts) of each its segment and apply post-filtering of retrieved segments which partially overlap with another higher ranked segment. In addition, we also remove all retrieved segments which partially overlap with the query segment.

3. HYPERLINKING

In the Hyperlinking sub-task, we first transformed the query segment into a textual query by including all the words of the subtitles lying within the segment boundary. Then, we extended the segment boundary by including the context surrounding the query segment. The optimal length of the surrounding context was tuned on the training data. We used a 200-seconds-long passage before and after each segment.

3.1 Visual Similarity

The visual modality was employed in the following way. First, we calculated distance between each keyframe in the collection and each query segment keyframe using the Signature Quadratic Form Distance [3, 8] and Feature Signatures [7] (the parameter of the method was tuned on the training data). Then, we calculated the $VisualSimilarity$ between each query/segment pair as the maximal similarity $(1-distance)$ between keyframes in the query and keyframes in the segment. The calculated $VisualSimilarity$ was used to modify the final score of the segment in the retrieval for a particular query segment as follows (the $Weight$ parameter was tuned on the training data and $Score(segment/query)$ is the output of the retrieval on the subtitles/transcripts):

$$FinalScore(segment/query) = Score(segment/query) + Weight * VisualSimilarity(segment/query).$$

3.2 Prosodic Similarity

The eight prosodic features provided in the data (energy, loudness, voice probability, pitch, pitch direction, direction score, voice quality, and harmonics-to-noise ratio) were used to construct 8-dimensional prosodic vectors each 10 ms of the recordings. We took overlapping sequences of 10 vectors appearing up to 1 second from the beginning of the query segment and found the most similar sequence of the vectors in each segment.

Similarity between the vector sequences was calculated as the sum of differences between the corresponding vectors of the sequence. These differences were calculated as the sum of the absolute values of the differences between the corresponding items of the prosodic vectors. To ensure that all prosodic features have equal weights, the difference of each item of the prosodic vector was normalized to have component values between 0 and 1. Due to the computational costs, we only took into account the vector sequences lying at most 1 second far from the beginning of the segment. The final score of each segment was calculated in the same way as the final score for the visual similarity. The $Weight$ for the audio similarity was tuned on the training set.

4. RESULTS

The results of the Hyperlinking sub-task are displayed in Table 1. We report the following evaluation measures: Mean

MediaEval 2014 Workshop, October 16-17, 2014, Barcelona, Spain

Transcripts	Segment.	Weights	Metadata	Overlap	MAP	P5	P10	P20	MAP-bin	MAP-tol
Subtitles	Fixed	None	No	No	0.1618	0.4786	0.4107	0.2893	0.1423	0.1216
Subtitles	Fixed	Visual	No	No	0.1660	0.4929	0.4143	0.3000	0.1483	0.1245
Subtitles	Fixed	None	Yes	No	0.4301	0.8600	0.7767	0.5483	0.2689	0.2465
Subtitles	Fixed	Visual	Yes	Yes	**4.1824**	**0.9667**	**0.9567**	**0.8967**	**0.3080**	0.0996
Subtitles	Fixed	Visual	Yes	No	0.4366	0.8667	0.7700	0.5633	0.2724	**0.2580**
Subtitles	Fixed	Prosodic	Yes	No	0.4321	0.8533	0.7767	0.5517	0.2687	0.2473
Subtitles	DT	Visual	Yes	No	0.8253	0.8867	0.8567	0.7383	0.2525	0.1991
LIMSI	Fixed	None	No	No	0.1043	0.3071	0.2571	0.1982	0.1028	0.0742
LIMSI	Fixed	Visual	No	No	0.1054	0.3500	0.3071	0.2161	0.1051	0.0775
LIMSI	Fixed	None	Yes	No	0.4166	0.8533	0.7133	0.5450	0.2659	0.2297
LIMSI	Fixed	Visual	Yes	Yes	**4.0331**	**0.9400**	**0.9233**	**0.8983**	**0.3042**	0.0950
LIMSI	Fixed	Visual	Yes	No	0.4168	0.8667	0.7333	0.5400	0.2692	**0.2414**
LIMSI	DT	Visual	Yes	No	0.5196	0.8333	0.7233	0.5817	0.2681	0.1976
LIUM	Fixed	None	No	No	0.0817	0.3000	0.2429	0.1768	0.0873	0.0604
LIUM	Fixed	Visual	No	No	0.0870	0.3286	0.2607	0.1804	0.0913	0.0632
LIUM	Fixed	None	Yes	No	0.4226	0.8333	0.7300	0.5433	0.2593	0.2547
LIUM	Fixed	Visual	Yes	Yes	**3.8916**	**0.9267**	**0.8800**	**0.8800**	**0.2880**	0.0993
LIUM	Fixed	Visual	Yes	No	0.4212	0.8400	0.7367	0.5350	0.2622	**0.2632**
LIUM	DT	Visual	Yes	No	0.5195	0.8200	0.7467	0.5733	0.2674	0.2134
NST-Sheffield	Fixed	None	No	No	0.1147	0.3071	0.2786	0.2036	0.1021	0.0792
NST-Sheffield	Fixed	Visual	No	No	0.1211	0.3643	0.3000	0.2375	0.1072	0.0838
NST-Sheffield	Fixed	None	Yes	No	0.4072	0.8067	0.7000	0.5417	0.2611	0.2237
NST-Sheffield	Fixed	Visual	Yes	Yes	**3.9822**	**0.9267**	**0.9267**	**0.8817**	**0.2949**	0.0914
NST-Sheffield	Fixed	Visual	Yes	No	0.4160	0.8267	0.7167	0.5483	0.2655	**0.2440**
NST-Sheffield	DT	Visual	Yes	No	0.6889	0.8400	0.8067	0.6500	0.2666	0.1955

Table 1: Results of the Hyperlinking sub-task for different transcripts, segmentation types, weightening types, metadata, and removal of overlapping retrieved segments. The best results for each transcript are highlighted.

Average Precision (MAP), Precision at 5 (P5), Precision at 10 (P10), Precision at 20 (P20), Binned Relevance (MAP-bin), and Tolerance to Irrelevance (MAP-tol) [1].

The highest scores of MAP, MAP-bin and the precision-based measures are not surprisingly reached in the cases when overlapping segments are preserved in the results [6]. Unlike in the Search sub-task, the segmentation employing the Decision Trees outperforms the fixed-length segmentation for most of the measures. There is a constant improvement in the case that the visual weights were used and small, but promising improvement in MAP, P20, and MAP-tol measures in the case when prosodic features were used. The concatenation with the context and metadata is also proved to be beneficial; the improvement is on all transcripts and the MAP score raised more than 5 times on the LIUM transcripts when metadata and context were used.

5. ACKNOWLEDGMENTS

This research is supported by the Czech Science Foundation, grant number P103/12/G084, Charles University Grant Agency GA UK, grant number 920913, and by SVV project number 260 104.

6. REFERENCES

[1] R. Aly, M. Eskevich, R. Ordelman, and G. J. F. Jones. Adapting Binary Information Retrieval Evaluation Metrics for Segment-based Retrieval Tasks. *CoRR*, abs/1312.1913, 2013.

[2] R. Aly, R. J. F. Ordelman, M. Eskevich, G. J. F. Jones, and S. Chen. Linking Inside a Video Collection: What and How to Measure? In *Proc. of WWW*, pages 457–460, Rio de Janeiro, Brazil, 2013.

[3] C. Beecks, M. S. Uysal, and T. Seidl. Signature Quadratic Form Distance. In *Proc. of CIVR*, pages 438–445, Xi'an, China, 2010.

[4] M. Eskevich, R. Aly, D. N. Racca, R. Ordelman, S. Chen, and G. J. F. Jones. The Search and Hyperlinking Task at MediaEval 2014. In *Proc. of MediaEval*, Barcelona, Spain, 2014.

[5] F. Eyben, F. Weninger, F. Gross, and B. Schuller. Recent Developments in openSMILE, the Munich Open-source Multimedia Feature Extractor. In *Proc. of ACMMM*, pages 835–838, Barcelona, Spain, 2013.

[6] P. Galuščáková and P. Pecina. CUNI at MediaEval 2014 Search and Hyperlinking Task: Search Task Experiments. In *Proc. of MediaEval*, Barcelona, Spain, 2014.

[7] M. Kruliš, J. Lokoč, and T. Skopal. Efficient Extraction of Feature Signatures Using Multi-GPU Architecture. In *MMM (2)*, volume 7733 of *LNCS*, pages 446–456. Springer, 2013.

[8] M. Kruliš, T. Skopal, J. Lokoč, and C. Beecks. Combining CPU and GPU Architectures for Fast Similarity Search. *Distributed and Parallel Databases*, 30(3-4):179–207, 2012.

[9] L. Lamel and J.-L. Gauvain. Speech Processing for Audio Indexing. In *Proc. of GoTAL*, pages 4–15, Gothenburg, Sweden, 2008.

[10] P. Lanchantin, P.-J. Bell, M.-J.-F. Gales, T. Hain, X. Liu, Y. Long, J. Quinnell, S. Renals, O. Saz, M.-S. Seigel, P. Swietojanski, and P.-C. Woodland. Automatic Transcription of Multi-genre Media Archives. In *Proc. of SLAM Workshop*, pages 26–31, Marseille, France, 2013.

[11] A. Rousseau, P. Deléglise, and Y. Estève. Enhancing the TED-LIUM Corpus with Selected Data for Language Modeling And More TED Talks. In *Proc. of LREC*, pages 3935–3939, Reykjavik, Iceland, 2014.