LinkedTV at MediaEval 2014 Search and Hyperlinking Task

H.A. Le[1], Q.M. Bui[1], B. Huet[1], B. Červenková[2], J. Bouchner[2], E. Apostolidis[3], F. Markatopoulou[3], A. Pournaras[3], V. Mezaris[3], D. Stein[4], S. Eickeler[4], and M. Stadtschnitzer[4]

[1]Eurecom, Sophia Antipolis, France. huet@eurecom.fr

[2]University of Economics, Prague, Czech Republic. barbora.cervenkova@vse.cz

[3]Information Technologies Institute, CERTH, Thessaloniki, Greece. bmezaris@iti.gr

[4]Fraunhofer IAIS, Sankt Augustin, Germany. daniel.stein@iais.fraunhofer.de

ABSTRACT

The paper presents the LinkedTV approaches for the Search and Hyperlinking (S&H) task at MediaEval 2014. Our submissions aim at evaluating 2 key dimensions: temporal granularity and visual properties of the video segments. The temporal granularity of target video segments is defined by grouping text sentences, or consecutive automatically detected shots, considering the temporal coherence, the visual similarity and the lexical cohesion among them. Visual properties are combined with text search results using multimodal fusion for re-ranking. Two alternative methods are proposed to identify which visual concepts are relevant to each query: using WordNet similarity or Google Image analysis. For Hyperlinking, relevant visual concepts are identified by analysing the video anchor.

1. INTRODUCTION

This paper describes the framework used by the LinkedTV team to tackle the problem of Search and Hyperlinking inside a video collection [3]. The applied techniques originate from the LinkedTV project[1], which aims at integrating TV and Web documents, by enabling users to access additional information and media resources aggregated from diverse sources, thanks to automatic media annotation. Here follows the description of our media annotation process. Shot segmentation is performed using a variation of [1], while the selected keyframes (one per shot) are analysed by visual concept detection [9] and Optical Character Recognition (OCR) [11] techniques. For each video, keywords are extracted from the subtitles, based on the algorithm presented in [12]. Finally, video shots are grouped into longer segments (scenes) based on 2 hierarchical clustering strategies. Media annotations are indexed at 2 levels (video level and scene level) using the Apache Solr platform[2]. At the video level, document descriptions are limited to text (title, subtitle, keywords, etc...), while the scene level documents are characterized by both text (subtitle/transcript, keywords, ocr, etc...) and float fields. Each float field corresponding to a unique visual concept response.

[1]http://www.linkedtv.eu/

[2]http://lucene.apache.org/solr/

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

1.1 Temporal Granularity

Three temporal granularities are evaluated. The first, termed *Text-Segment*, consists in grouping together sentences (up to 40) from the text sources. We also propose to segment videos into scenes which consist of semantically correlated adjacent shots. Two strategies are employed to create scene level temporal segments. Visually similar adjacent shots are merged together to create *Visual-scenes* [10], while *Topic-scenes* are built by jointly considering the aforementioned results of visual scene segmentation and text-based topical cohesion (exploiting text extracted from ASR transcripts or subtitles).

1.2 Visual Properties

In MediaEval S&H 2014, queries are composed of a few keywords only (visual-cues are not provided). Hence, the identification of relevant visual concepts is more complex than last year. We propose two alternatives to this problem. On one hand, WordNet similarity is employed to map visual concepts with query terms [8]. On the other hand, the query terms are used to perform a Google Image search. Visual concept detection (using 151 concepts from the TRECVID SIN task [6]) is performed on the first 100 returned images and concepts obtaining the highest average score are selected.

2. SEARCH SUB-TASK

2.1 Text-based methods

In this approach, relevant text and video segments are searched using Solr using text (TXT) only. Two strategies are compared: one where search is performed at the text segment level directly (S) and one where the first 50 videos are retrieved at the video level and then the relevant video segment is locate using the scene-level index. The scene-level index granularity is either the Visual-Scene (VS) or the Topic-Scene (TS). Scenes at both granularities are characterized by textual information only (either the subtitle (M) or one of the 3 ASR transcripts ((U) LIUM [7], (I) LIMSI [4], (S) NST/Sheffield [5])).

2.2 Multimodal Fusion method

Motivated by [8], visual concept scores are fused with text-based results from Solr to perform re-ranking. Relevant visual concepts, out of the 151 available, for individual queries are identified using either the WordNet (*WN*) or the GoogleImage (*GI*) strategy. For those multi-modal (*MM*) runs only visual scene (*VS*) segmentation is evaluated.

3. HYPERLINKING SUB-TASK

Pivotal to the hyperlinking task is the ability to automatically craft an effective query from the video anchor under consideration, to search within the annotated set of media. We submitted two alternative approaches; One using the MoreLikeThis (*MLT*) Solr extension, and the other using Solr's query engine. *MLT* is used in combination with the sentence segments (*S*), using either text (*MLT1*) or text and annotations [2] (*MLT2*). When Solr is used directly, we consider text only (*TXT*) or with visual concept scores of anchors (*MM*) to formulate queries. Keywords appearing within the query anchor's subtitles compose the textual part of the query. Visual concepts whose scores within the query anchor exceed the 0.7 threshold are identified as relevant to the video anchor and added to the Solr query. Both visual (*VS*) and topic scenes (*TS*) granularities are evaluated in this approach.

4. RESULTS

4.1 Search sub-task

Table 1 shows the performance of our search runs. Our best performing approach (*TXT_VS_M*), according to MAP, relies on manual transcript only segmented according to visual scenes. Looking at the precision scores at 5, 10 and 20, one can notice that multi-modal approaches using WordNet (*MM_VS_WN_M*) and Google images (*MM_VS_GI_M*) boost the performance of text only approaches. There is a clear performance drop whenever ASR (*I, U or S*) are employed, instead of subtitles (*M*).

Table 1: Results of the Search sub-task

Run	map	P_5	P_10	P_20
TXT_TS_I	0,4664	0,6533	0,6167	0,5317
TXT_TS_M	0,4871	0,6733	0,6333	0,545
TXT_TS_S	0,4435	0,66	0,6367	0,54
TXT_TS_U	0,4205	0,6467	0,6	0,5133
TXT_S_I	0,2784	0,6467	0,57	0,4133
TXT_S_M	0,3456	0,6333	0,5933	0,48
TXT_S_S	0,1672	0,3926	0,3815	0,3019
TXT_S_U	0,3144	0,66	0,6233	0,48
TXT_VS_I	0,4672	0,66	0,62	0,53
TXT_VS_M	**0,5172**	0,68	0,6733	**0,5933**
TXT_VS_S	0,465	0,6933	0,6367	0,5317
TXT_VS_U	0,4208	0,6267	0,6067	0,53
MM_VS_WN_M	0,5096	**0,7**	**0,6967**	0,5833
MM_VS_GI_M	0,509	0,6667	0,68	**0,5933**

4.2 Hyperlinking sub-task

Table 2 shows the performance of our hyperlinking runs. Again, the approach based on subtitle only (*TXT_VS_M*) performed best (MAP=0,25) followed by the approach using MoreLikeThis (*TXT_S_MLT1_M*). Multi-modal approaches did not produce the expected performance improvement. We believe this is due to the significant duration reduction of anchors compared with last year.

Table 2: Results of the Hyperlinking sub-task

Run	map	P_5	P_10	P_20
TXT_S_MLT2_I	0,0502	0,2333	0,1833	0,1117
TXT_S_MLT2_M	0,1201	0,3667	0,3267	0,2217
TXT_S_MLT2_S	0,0855	0,2067	0,2233	0,1717
TXT_VS_M	**0,2524**	**0,504**	**0,448**	**0,328**
TXT_S_MLT1_I	0,0798	0,3	0,2462	0,1635
TXT_S_MLT1_M	0,1511	0,4167	0,375	0,2687
TXT_S_MLT1_S	0,1118	0,3	0,2857	0,2143
TXT_S_MLT1_U	0,1068	0,2692	0,2577	0,2038
MM_VS_M	0,1201	0,3	0,2885	0,1923
MM_TS_M	0,1048	0,3538	0,2654	0,1692

5. CONCLUSION

The results of LinkedTV's approaches on the 2014 MediaEval S&H task show that it is difficult to improve over text based approaches when no visual cues are provided. Overall, our S&H algorithms performance on this year's dataset have decreased compared to 2013, showing that task definition changes have made the task harder to solve.

6. ACKNOWLEDGMENTS

This work was supported by the European Commission under contract FP7-287911 LinkedTV.

7. REFERENCES

[1] E. Apostolidis and V. Mezaris. Fast shot segmentation combining global and local visual descriptors. In *2014 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, p 6583–6587, Italy.

[2] M. Dojchinovski and T. Kliegr. Entityclassifier.eu: Real-Time Classification of Entities in Text with Wikipedia. In H. Blockeel, K. Kersting, S. Nijssen, and F. Železný, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 8190 of *Lecture Notes in Computer Science*, pages 654–658. Springer, 2013.

[3] 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 *MediaEval 2014 Workshop*, Spain.

[4] J.-L. Gauvain, L. Lamel, and G. Adda. The LIMSI broadcast news transcription system. *Speech Communication*, 37(1):89–108, 2002.

[5] T. Hain, A. El Hannani, S. N. Wrigley, and V. Wan. Automatic speech recognition for scientific purposes-webasr. In *Interspeech*, Australia, pages 504–507, 2008.

[6] P. Over *et al.* TRECVID 2012 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. In *Proceedings of TRECVID 2012*. NIST, USA, 2012.

[7] A. Rousseau, P. Deléglise, and Y. Estève. Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks. In *LREC 2014*, Iceland.

[8] B. Safadi, M. Sahuguet, and B. Huet. When textual and visual information join forces for multimedia retrieval. In *ACM ICMR 2014*, Glasgow, Scotland.

[9] P. Sidiropoulos, V. Mezaris, and I. Kompatsiaris. Enhancing Video concept detection with the use of tomographs. In *IEEE ICIP 2013*, Australia.

[10] P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, H. Meinedo, M. Bugalho, and I. Trancoso. Temporal Video Segmentation to Scenes Using High-Level Audiovisual Features. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(8):1163–1177, Aug. 2011.

[11] D. Stein, S. Eickeler, R. Bardeli, E. Apostolidis, V. Mezaris, and M. Müller. Think Before You Link – Meeting Content Constraints when Linking Television to the Web. In *Proc. NEM Summit*, Nantes, France, Oct. 2013.

[12] S. Tschopel and D. Schneider. A lightweight keyword and tag-cloud retrieval algorithm for automatic speech recognition transcripts. In *Interspeech*, Japan, 2010.