JRS at Search and Hyperlinking of Television Content Task

Werner Bailer, Harald Stiegler
JOANNEUM RESEARCH – DIGITAL
Steyrergasse 17, 8010 Graz, Austria
{firstName.lastName}@joanneum.at

ABSTRACT

This paper describes the work done by the JRS team for the linking sub-task. We submitted eight pairs of runs: four with different textual resources only, two using reranking based on visual similarity, and two using concept detection results. Each of the pairs contains of one run using the anchor segment only, and one using a longer context segment. The results show higher variance between anchors than for the 2013 task, also the differences between runs using different textual resources are more salient. The use of the context does not generally improve results, and visual reranking provides small improvements.

1. INTRODUCTION

The MediaEval 2014 Search and Hyperlinking of Television Content Task addresses the scenario of performing search in a video collection (search sub-task) and subsequent exploration of related video segments (hyperlinking sub-task). This paper describes the work done by the JRS team for the linking sub-task. Details on the task and the data set can be found in [3].

2. LINKING SUB-TASK

For the linking sub-task, we combine textual/metadata similarity and visual similarity. The textual/metadata similarity is based on matching terms and named entities, and provides a basic set of result segments. In some runs, visual similarity based on local descriptors is used for reranking. In the following, we briefly summarise the approach.

The textual/metadata based approach uses the automatic speech recognition (ASR) transcript or subtitles and the metadata about the broadcast (using title, description and short synopsis of episodes). All these textual resources are preprocessed by removing punctuation, normalizing capitalization and removing stop words and very short words (less than three characters). We then select a basic set of terms $T = T_a \cup T_m$, which are the words T_a from the anchor and T_m from the metadata, that are found in DBpedia[1]. Some runs use the results of concept detection, treating the annotated concepts T_c like terms extracted from the text of the segment. In this case, the set of terms is defined as $T = T_a \cup T_m \cup T_c$.

For the ASR transcript or subtitles, we then broaden the set of terms and select specific classes. As a first step, we add synonyms for the terms in T from WordNet[2], obtaining a set S_T. We then select a set of connected entities C_T for the terms in T from FreeBase[3]. For the subset of terms $T_g \subset T$, which FreeBase identifies as related to a geographic location, we also add the set of connected geographic entities G_{T_g} from GeoNames[4]. Thus the set of terms used for matching is $T^* = T \cup S_T \cup C_T \cup G_{T_g}$.

For matching two segments, we match the terms related to these segments with different weights:

$$\begin{aligned} w(t) &= w_o, t \in T, \\ w(t) &= w_g, t \in G_{T_g}, \\ w(t) &= w_s, t \in S_T \cup C_T, \text{with } w_s < w_g < w_o. \end{aligned} \quad (1)$$

For multiple occurrences K in a segment, the weights of each occurrence decrease, with the total weight defined as $\widehat{w}(t) = \sum_{k=1}^{K}(1/k)w(t)$. For a pair of video segments (v_1, v_2) the similarity is determined as $\sum_{t \in T^*(v_1) \cap T^*(v_2)} w(t)$, with $T^*(v_i)$ being the extended set of terms of segment v_i.

For initial text-based matching, the videos have been segmented into segments of equal lengths of 20 seconds. In the experiments, we cut the lists at a normalized similarity score of 0.1, keeping at most 500 result items.

For visual matching we use VLAT [7] on SIFT [6] descriptors extracted from difference of Gaussians (DoG) interest points. In order to avoid possible side effects of interlaced content, only one field is used if interlacing artifacts are detected. Descriptors are extracted from every fifth frame (every tenth field) and detecting several hundred key points (limiting to the best 500). We use the VLAT Wise variant with a dictionary size of 128. Visual matching is applied to the top results from textual matching (score ≥ 0.35, at most 50 result items) and for these items are ranked using only the scores from visual similarity.

3. SUBMITTED RUNS

We submitted in total eight pairs of runs, each containing one using only the exact anchor segment, and one using the anchor plus context item. This decision is based on the conclusions from the linking sub-task at MediaEval 2013, where the use of a longer segment including context significantly improved the results. As no anchor item segment is

[1] dbpedia.org

[2] wordnet.princeton.edu

[3] www.freebase.com

[4] www.geonames.org

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

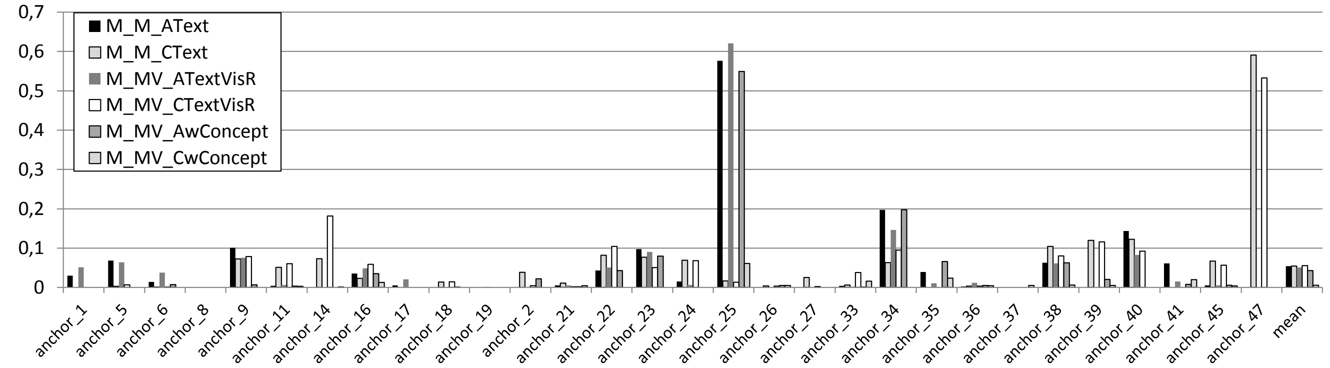

Figure 1: Results for selected runs based on subtitles, using the anchor segment (A) or also context (C), and with visual reranking or use of concept detections.

defined in the task input data in 2014, we created segments by adding three minutes before and after anchor, as this was shown in [1] to provide comparable results. All runs produce fixed segments of 20 seconds. Four of the pairs use only text resources, i.e., the metadata and one of the three types of ASR transcripts (LIMSI [4], LIUM [8], NST/Sheffield [5]) or the subtitles. For the pairs using LIMSI and subtitles, we additionally generated a version with reranking of the top results based on visual similarity. Finally, two pairs of runs used the concept detection results from University of Oxford [2], one in addition to metadata and subtitles and only the concepts without any textual metadata. We decided to only use the full set of transcripts for the text only runs, and not use them in combination with each other feature, as the results from 2013 showed rather small differences between runs using different text resources.

4. RESULTS

For the runs that use the same method as in 2013 (apart from necessary changes for parsing metadata) the overall scores are clearly lower, and the variance of the results is much higher. This indicates that the anchors used this year might be more challenging, and more diverse. Using context segments does only slightly improve results, but the impact depends very much on the anchor. For example, for anchor 25 we obtain significantly better results when only using the anchor, while for anchor 47 results are much better when using the context. There are also clearer differences between different text resources than in 2013, with manual subtitles clearly providing best results, LIMSI and NST/Sheffield have comparable results, and LIUM resulting in lower results. Using visual reranking provides in most cases a slight improvement (this is in line with the results from 2013). Visual concepts alone result in a performance that is an order of magnitude lower than using text results, combining them with text slightly lowers the performance for the anchor only runs, and significantly for runs using the context segment. An overview of the scores of some selected runs is shown in Figure 1.

5. CONCLUSION

Comparable approaches yield overall lower results than in 2013, and there is higher variance in the queries. This includes also the usefulness of context segments for the anchors. Visual information makes small contributions to the overall scores, however, visual concepts need to be treated specially, as their confidence is still low.

Acknowledgments

The authors thank Michał Lokaj for his contributions to the implementation. The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 610370, "ICoSOLE – Immersive Coverage of Spatially Outspread Live Events" (`http://www.icosole.eu/`).

6. REFERENCES

[1] W. Bailer, M. Lokaj, and H. Stiegler. Context in video search: Is close-by good enough when using linking? In *Proc. ICMR*, Glasgow, UK, 2014.

[2] K. Chatfield and A. Zisserman. VISOR: towards on-the-fly large-scale object category retrieval. In *Proc. ACCV*, pages 432–446, 2012.

[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*, Barcelona, ES, 2014.

[4] J.-L. Gauvain, L. Lamel, and G. Adda. The LIMSI Broadcast News transcription system. *Speech Communication*, 37(1-2):89–108, 2002.

[5] P. Lanchantin, P. Bell, M. Gales, T. Hain, X. Liu, Y. Long, J. Quinnell, S. Renals, O. Saz, and M. Seigel. Automatic transcription of multi-genre media archives. In *Proc. SLAM Workshop*, Marseille, FR, 2013.

[6] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[7] R. Negrel, D. Picard, and P. H. Gosselin. Web-scale image retrieval using compact tensor aggregation of visual descriptors. *IEEE MultiMedia*, 20(3):2433, 2013.

[8] A. Rousseau, P. Deléglise, and Y. Estève. Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks. In *Proc. LREC*, pages 26–31, Reykjavik, IS, 2014.