TOSCA-MP at Search and Hyperlinking of Television Content Task

Michał Lokaj, Harald Stiegler, Werner Bailer
JOANNEUM RESEARCH – DIGITAL
Steyrergasse 17, 8010 Graz, Austria
werner.bailer@joanneum.at

ABSTRACT

This paper describes the work done by the TOSCA-MP team for the linking subtask. We submitted three sets of runs: text-only with fixed segments, text-only aligned with shot boundaries, and text and visual with fixed segments. Each of these sets consists of six runs, using combinations of three different types of text resources and for each using only the anchor segment or anchor plus context as input. The results show significant improvements by taking the context of the anchor into account, and smaller improvements when additionally using visual features.

1. INTRODUCTION

The MediaEval 2013 Search and Hyperlinking of Television Content Task addresses the scenario of performing text-based known item search in a video collection (search subtask) and subsequent exploration of related video segments (hyperlinking subtask). Such a scenario is well-aligned with the goals TOSCA-MP project, which aims at developing task-adaptive analysis and search tools for professional media production. Some content needs for media production are not always sharply defined, in some cases a comprehensive coverage of a topic may be needed, or media creators aim at finding more diverse, less well-known material. All these cases fit into the pattern of performing a first search task, and using the result set for further interactive exploration of the video collection.

This paper describes the work done by the TOSCA-MP team for the linking subtask. Details on the task and the data set can be found in [1].

2. LINKING SUBTASK

2.1 Approach

For the linking subtask, 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. The textual/metadata based approach uses the ASR transcript or subtitles, the metadata about the broadcast (synopsis) and the text of the query related to the anchor as inputs. 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 $T = T_a \cup T_q \cup T_m$, which are the words from the three cleaned text resources (anchor, query, metadata) that are found in DBpedia[1]. 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:

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

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 0.35, keeping at least 75 items. On these raw results, reranking based on visual features or alignment with shot boundaries is applied.

The visual matching approach is based on the well-known SIFT descriptor [3], extracted from DoG interest points from the video. Only one field of the input image is used in order to avoid possible side effects of interlaced content. Descriptors are extracted from every fifth frame (every tenth field) and detecting several hundred key points (the number depends on the resolution of the content and the structure of the content itself). We have performed complete pairwise matching of a set of candidate link segments (result of textual matching). Both descriptor extraction and matching are implemented on the GPU using NVIDIA CUDA.

In order to align candidate segments with shot bound-

[1] dbpedia.org

[2] wordnet.princeton.edu

[3] www.freebase.com

[4] www.geonames.org

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

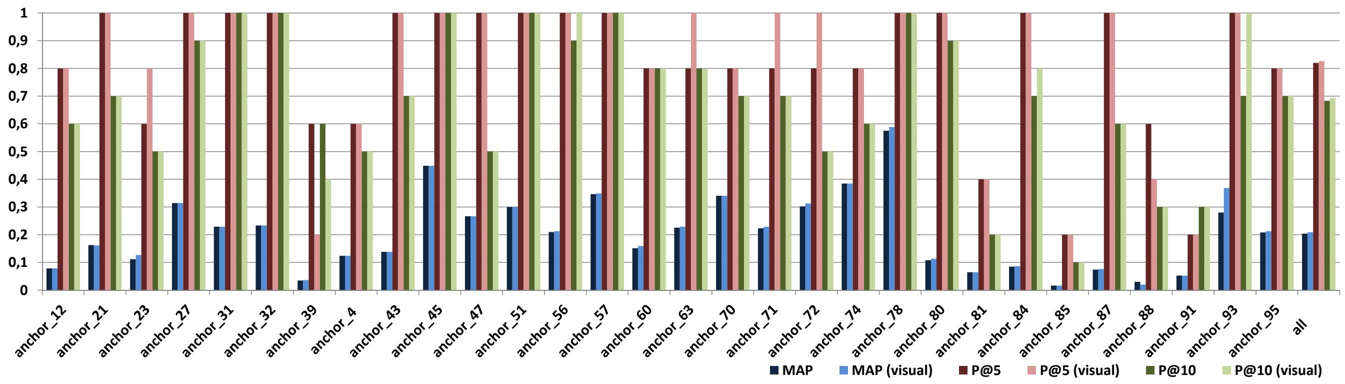

Figure 1: Comparison of a text/metadata-based run and its reranked version using visual features.

aries, all adjacent candidate segments have been matched to the respective shots. In order to avoid a bias for text matching from the shots, only scores within 30 seconds have been counted. However, the entire shot has been reported as result. According to the guidelines, the segments have been cut to at most 120 seconds, even if some result shots exceed this length.

2.2 Experiments and Results

We generated three sets of runs: text-only with fixed segments, text-only aligned with shot boundaries, and text and visual with fixed segments. Each of these sets consists of six runs, using combinations of each of the three different types of text resources (two ASR transcripts and subtitles) and for each using only the anchor segment or anchor plus context as input. For the runs using visual features, a higher weight for visual features was used (0.7 vs. 0.3 for text features), in order to make results different from text-only runs. This may not be the optimal weight combination.

The best runs reached up to 0.21 mean average precision (MAP), median average precision is 0.21 as well. The results are quite consistent across the anchors, with the median AP being very similar to the mean AP. The results for the top ranks are much better, with mean precision at rank 5 up to 0.83 and at rank 10 up to 0.7. At the top 5 ranks, the median precision is even higher, reaching 1.00 for three of the runs. Using the context of the video around the anchor had a very strong impact on the results. In terms of MAP the increase is about 0.10, i.e. MAP roughly doubles.

Reranking using visual features slightly improves the results in all cases, with most improvement at the top 5 ranks. The use of the visual features has quite different effects on the different types of anchors, causing an increase for some and a decrease for others (see Figure 1), depending on whether a query focuses more on topical/textual content or is more visual, be it because the query is more descriptive, or the scene or dominant objects are shared by all relevant segments. Using a shot-based segmentation did not in general improve the results.

Looking at the result segments, we did not expect significant differences between the runs with the same configuration but different types of transcripts. The distribution between the types of terms is also quite similar for all text resources (about 45% named entities, and another 45% from words and synomyms, the other type 1-2% each). Only LIMSI has a slightly lower fraction of matching query terms and metadata than the others. The results for the different textual resources are quite similar, with the LIUM-based [4] runs performing slightly better than those using LIMSI/Vocapia [2] transcripts or manual subtitles. There is one outlier run based on subtitles, which performs significantly worse when shot boundaries are used, but only for the non-context case. This seems to be a particular issue of the alignment of the transcript with the boundaries for the segments involved, but not a general pattern.

3. CONCLUSION

The results are quite encouraging, and show that the proposed method yields useful results for the linking subtask, especially at the top five to ten ranks. Taking the context of the anchor into account provides a significant improvement of the results. The use of visual reranking segments provides small but consistent improvements. The use of shots as result segments does not generally improve the results.

Acknowledgments

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 287532, "TOSCA-MP - Task-oriented search and content annotation for media production" (http://www.tosca-mp.eu).

4. REFERENCES

[1] Maria Eskevich, Gareth J.F. Jones, Shu Chen, Robin Aly, and Roeland Ordelman. The Search and Hyperlinking Task at MediaEval 2013. In *MediaEval 2013 Workshop*, Barcelona, Spain, October 18-19 2013.

[2] Lori Lamel and Jean-Luc Gauvain. Speech processing for audio indexing. In *Advances in Natural Language Processing (LNCS 5221)*, pages 4–15. Springer, 2008.

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

[4] A. Rousseau, F. Bougares, P. Deléglise, H. Schwenk, and Y. Estèv. LIUM's systems for the IWSLT 2011 Speech Translation Tasks. In *Proceedings of IWSLT 2011*, San Francisco, USA, 2011.