UTwente does Brave New Tasks for MediaEval 2012: Searching and Hyperlinking

Danish Nadeem, Robin Aly, Roeland Ordelman
University of Twente
Enschede, The Netherlands
(d.nadeem, r.aly, ordelman)@ewi.utwente.nl

ABSTRACT

In this paper we report our experiments and results for the brave new searching and hyperlinking tasks for the MediaEval Benchmark Initiative 2012. The searching task involves finding target video segments based on a short natural language sentence query and the hyperlinking task involves finding links from the target video segments to other related video segments in the collection using a set of anchor segments in the videos that correspond to the textual search queries. To find the starting points in the video, we only used speech transcripts and metadata as evidence source, however, other visual features (for e.g., faces, shots and keyframes) might also affect results for a query. We indexed speech transcripts and metadata, furthermore, the speech transcripts were indexed at speech segment level and at sentence level to improve the likelihood of finding jump-in-points. For linking video segments, we computed k-nearest neighbours of video segments using euclidean distance.

1. INTRODUCTION

While containing a wealth of information, the content of a video can be quite long. Therefore, anyone searching for a known-item in a particular video, may wish to be pointed at the offset time of relevant segments (jump-in-points or entry points) within that video. Furthermore, a searcher may also wish to get links to video segments relevant to the already found video segment. In this paper, we describe our approaches for searching and linking of video segments, as our contribution to the Brave New Tasks in the MediaEval Benchmark Initiative 2012, see [5] for a detailed task description. This year's search sub-task is related to the previously Rich Speech Retrieval (RSR) task at MediaEval 2011 [1]. In continuation to earlier work, we focus on setting up a baseline system that uses 1-best speech transcripts and metadata as evidence sources. However, for the novel hyperlinking task, we returned a set of nearest video segments that were similar in the concepts they contained.

The paper is outlined as follows. In Section 2 we describe our methods for searching video segments and linking video with other related video segments. In Section 3 we describe the details of our experiment and show the evaluation results for our submitted runs, and finally we conclude with discussion in Section 4 and future plans in Section 5.

MediaEval 2012 Workshop, October 4-5, 2012, Pisa, Italy

2. METHOD

In the following sub-sections, we describe our methods used for the searching and hyperlinking sub-tasks.

2.1 Searching

For searching, we use the following evidence sources for each video segment: the 1-best output of an ASR system and user-generated metadata, such as title, tags and a short description about videos. We used two different types of result units: speech segments that are time intervals which the ASR system deemed to be of one speaker, and sentences which are sub-parts of the previous segments divided by a period. We assume that considering different result units may improve the likelihood of finding a suitable jump-in-point.

We adapt the ranking function, earlier defined in [1], to combine the scores of speech segments and metadata. The final ranking function is defined as the following:

$$s_d = \lambda \frac{s_{seg}}{max(s_{seg})} + (1 - \lambda) \frac{s_{meta}}{max(s_{meta})} \quad (1)$$

where s_d is the final document score, s_{seg} is the speech segment score, λ is the influence of the speech segment score on the ranking and s_{meta} is the score of the metadata for the corresponding document. Note that, if a video or segment does not appear in a ranking we assume a score of zero. The combination method in Equation 1 results in a ranking of speech segments that we use to select the jump-in-points for a video segment.

2.2 Linking

We represent video segments by confidence scores of 508 concepts trained by a method described in [2]. For a given source segment, we determine the top-5 target segments according to their euclidean distance to the source segment. The result is a ranked list of target video segments. The list may contain all the video segments found within the same video as well as segments from other videos in the collection.

3. EXPERIMENTS

In this section, we describe the experimental setup and submitted runs for searching and linking tasks. For the searching task, short natural language queries were provided, an example from a development set query is: *How much Obama spend on his election campaign?* With respect to each search query, a ranked list of video segments were retrieved in the decreasing order of their likelihood. For the linking task, a set of anchor segments in the videos were

retrieved that corresponded to the textual queries.

3.1 Setup

Two different sets of speech transcripts from the same collection of videos were generously provided by LIMSI [7] and LIUM [8]. From the LIUM transcripts, 1-best hypotheses were used. The metadata for each video were provided by the original video uploader to blip.tv. Apart from user-generated content, metadata also contained information about *license type*, *filename*, *duration* and *uploader id*. We used LIMAS [3] to run our experiments. It uses hbase as a main index and lucene to obtain retrieval scores for text. The uniform weighting 1 scheme was set in the LIMAS configuration. We wrote python scripts to index speech segments, sentences and metadata. Also, query scripts were written to pre-process the query definitions (e.g., making it well-formed xml format) and to run them using LIMAS. For concept-based linking, we used a python script to search for nearest neighbours for a given segment using euclidean distance and returned the 5-nearest neighbour segments, based on the terms found in the query.

3.2 Submitted runs

We submitted the following runs[1]:

run1	Speech segments from LIMSI: *limsiSegments*
run2	Sentences from LIMSI: *limsiSentences*
run3	Sentences from LIUM: *liumSentences*
run4	Conept-based links: *conceptlinking*

3.3 Results

The search results are given below in terms of Mean Reciprocal Rank (MRR), mean Generalized Average Precision (mGAP) that takes into account the distance to the actual relevant jump-in point, and Mean Average Segment Precision (MASP) metric that takes into account time information in terms of both precision of the retrieved segments and the distance of the beginning of the retrieved segment to the real start of the relevant content (see [6]).

Runs	Window size	MRR	mGAP	MASP	MAP
run1	60 sec	0.156	0.122	0.085	–
	30 sec	0.155	0.088	0.085	–
	10 sec	0.093	0.033	0.050	–
run2	60 sec	0.074	0.054	0.111	–
	30 sec	0.073	0.035	0.112	–
	10 sec	0.034	0.002	0.076	–
run3	60 sec	0.213	0.161	0.124	–
	30 sec	0.204	0.131	0.129	–
	10 sec	0.136	0.081	0.122	–
run4	–	–	–	–	0.405

From the results it is clear that LIUM sentences *run3* shows the strongest performance using all measures. LIMSI segments *run1* performs better than the LIMSI sentences *run2* in the MRR and mGAP measure. In MASP measure, *run2* shows better performance.

4. DISCUSSION AND CONCLUSIONS

In our current tasks, we setup a baseline for future participations. We used speech segments, sentences and video-level metadata for searching video segments. We found that, according to the MRR, mGAP and MASP measure, using LIUM sentences showed the best performance. Comparing sentences and segment types for LIMSI showed inconsistent results according to the considered measure: for the MRR and mGAP measure segments performed more strongerly, while in terms of MASP using sentences showed better performance. We conclude that speech segments are better evidence for retrieval, however, sentences are better for the retrieval of jump-in point.

For hyperlinking, we used concept representations and computed the euclidean distances to find the 5 closest links to a given video segment.

5. FUTURE PLANS

This was our first participation to searching and hyperlinking task, in future, we plan to improve searches and linking using information about the presence of faces in videos [4]. We believe that similar faces might provide good hints for linking. Therefore, we can use face results in the fusion scheme as well.

[1] Common prefix *me12sh1_UTwente2012_*.

References

[1] R. Aly, T. Verschoor, and R. Ordelman. UTwente does rich speech retrieval at mediaeval 2011. In *MediaEval*, volume 807 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2011.

[2] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *British Machine Vision Conference*, 2011.

[3] S. Chen, K. McGuinness, R. Aly, N.E. O'Connor, and F. de Jong. The AXES-lite video search engine. In *Image Analysis for Multimedia Interactive Services (WIAMIS), 2012*, pages 1–4. IEEE, 2012.

[4] R. G. Cinbis, J. Verbeek, and C. Schmid. Unsupervised Metric Learning for Face Identification in TV Video. In *International Conference on Computer Vision*, Barcelona, Spain, November 2011.

[5] M. Eskevich, G. J. F. Jones, S. Chen, R. Aly, R. Ordelman, and M. Larson. Search and Hyperlinking Task at MediaEval 2012. In *MediaEval 2012 Workshop*, Pisa, Italy, October 4-5 2012.

[6] M. Eskevich, W. Magdy, and G. J. F. Jones. New metrics for meaningful evaluation of informally structured speech retrieval. In *Proceedings of the 34th European conference on Advances in Information Retrieval*, ECIR'12, pages 170–181. Springer Berlin Heidelberg, 2012.

[7] L. Lamel and J. Gauvain. Speech processing for audio indexing. In Bengt Nordström and Aarne Ranta, editors, *Advances in Natural Language Processing*, volume 5221 of *Lecture Notes in Computer Science*, pages 4–15. Springer Berlin Heidelberg, 2008.

[8] A. Rousseau, F. Bougares, P. Deléglise, H. Schwenk, and Y. Estève. LIUM's systems for the IWSLT 2011 speech translation tasks. In *International Workshop on Spoken Language Translation*, San Francisco (USA), 8-9 Sept 2011.