UTwente does Rich Speech Retrieval at MediaEval 2011

Robin Aly, Thijs Verschoor, Roeland Ordelman
University Twente
P.O. Box 217,
7500AE Enschede, The Netherlands
{r.aly, t.verschoor, ordelman}@ewi.utwente.nl

ABSTRACT

This paper describes the participation of the University of Twente team at the Rich Text Retrieval Task of the Media Eval Benchmark Initiative 2011. The goal of the task is to find entry points of relevant parts of videos to reduce the browsing effort of searchers. This is our first participation, therefore our main focus is to create a baseline system which can be improved in the future. We experiment with different evidence sources (ASR and meta data) together with a basic score combination function. We also experiment with different entry points relative to the segments found by the contained evidence.

1. INTRODUCTION

When searching in videos, it is especially important to return an entry point where the relevant part for a searcher begins. The reason is that videos can be multiple hours long, and unlike our ability to quickly scan text for relevant parts, scanning a video is much more time consuming. In this paper, we describe the methods we used for the participation in the Rich Text Retrieval Task of the Media Eval Benchmark Initiative 2011 [2].

This paper is structured as follows: Section 2 describes the evidences we considered to calculate the likelihood for the relevance of a segment, the combination of those evidences, and the alternative entry points. Section 3 details the experiments we undertook to evaluate our approach. Section 4 concludes this paper.

2. SEARCHING AND ENTRY POINT SELECTION

In this section, we describe the methods we used to search for segments in videos which are likely to contain relevant information, and given one such segment, how to determine the entry point into the video presented to the user.

2.1 Evidence

We used the following types of evidence to identify suitable entry points. First, we used the *meta data* which was provided on a video level. Second, we used transcripts from provided automatic speech recognition. The transcripts are divided into *speech segments* which the recognition system believed to originate from a speaker. Since the returned speech segments were relatively short, we also considered *speaker turns*, the transcripts of all consecutive speech segments of the same speaker, as an alternative evidence (we always used either speech segments or speaker turns). Finally, words in a speech segment also influence the likelihood that an entry point can be found in the remaining speech segments [5], therefore we also considered the transcript for the *whole document* as evidence.

For each of the three evidence sources, meta data, speech or speaker turn segments, and the transcript of the whole document, we create a ranking using two standard retrieval models, see Section 3. We refer to scores for the speech segment or the speaker turn as s_{base} because we determine the entry point relative to these segments. To the scores on the meta data, we refer to by s_{meta}, and to the score on the transcript for the whole video as s_{doc}. Therefore, while s_{base} can be be used to find entry points, s_{meta} and s_{doc} provide general evidence about the relevance of a video and could be combined with s_{base} to promote segments within a video with many relevant entry points.

2.2 Combination Methods

There are good reasons for the combination of the above scores. For example, an important query word might be said only in neighboring segments of the segment close to the ideal entry point. Furthermore, the language of the searcher and the speaker in the video might be different, and different sources, such as the available meta data can be useful to enrich the findings. Therefore, it is desirable to combine the findings from all evidence sources. For our first participation in this task, we choose a heuristic-based approach, on which we plan to improve in future work: because scores between evidences are not comparable, we scale the scores for each evidence to the interval $[0:1]$, see [3]. We combine the evidence scores linearly. The final ranking function is defined as the following:

$$s_d = \lambda_1 \frac{s_{base}}{max(s_{base})} + \lambda_2 \frac{s_{doc}}{max(s_{doc})} + (1-\lambda_1-\lambda_2) \frac{s_{meta}}{max(s_{meta})} \quad (1)$$

with

$$0 \leq \lambda_1, \lambda_2 \leq 1, \text{ and } \lambda_1 + \lambda_2 \leq 1$$

where s_d is the final document score, s_{base} the previously described base score, λ_1 is the influence of the base score on the ranking, s_{doc} is the score for the transcript of the corresponding document, λ_2 is the influence of the document transcript on the ranking, and s_{meta} is the score for the available meta data for the corresponding document. Note that, if a video or segment does not appear in a ranking we assume a score

MediaEval 2011 Workshop, September 1-2, 2011, Pisa, Italy

of zero. The combination method in Equation 1 results in a ranking of either speech segments or speaker turns. From this information, we then select the entry point.

2.3 Entry Point Selection

The ranking based on Equation 1 provides a ranking of intervals where suitable entry points could be. However, they do not necessarily need to be the beginning of this segment. Here, we investigate the following four entry points (EP) relative to the found speech or speaker turn segment:

1. the beginning of the retrieved speech segment (SS),
2. the beginning of the retrieved speaker turn (ST),
3. the beginning of the shot that contains the beginning of the segment (SHOT),
4. the time of the key frame of the latter shot, which is usually close to the middle of the shot (KF).

3. EXPERIMENTS

In this section, we describe the runs we performed to evaluate the evidence sources, parameter settings of the combination method in Equation 1, and methods to select an entry point based on obtained ranking.

For the experiments, we used the search engine PFTijah [1]. We used two different retrieval models (RM): the language models (LM) by the author of the engine, and the okapi retrieval model (BM25), see [4]. We considered the two different versions of the automatic speech recognition (ASR) output from 2010 and 2011, referred to by this number, which both provided a segmentation into speech segments, and inferred speaker turns therefrom. For the retrieval function in Equation 1, we performed a grid search with a step-size of 0.1 for the parameters λ_1 and λ_2. In the performance figures below, we used either only the segment evidence ($\lambda_1 = 1.0$) or the best performing combination of the grid search on the development dataset. We performed our evaluation using the three prescribed window sizes of 10, 30 and 60 seconds and a granularity factor of 10 according to [2], but only report results of window size 60 because of space requirements.

Submitted runs.

In the following we list the results of our official runs:

No	RM	ASR	EP	λ_1	λ_2	mGAP
(1)	BM25	2011	SS	1.0	0.0	0.266
(2)	BM25	2011	SS	0.2	0.3	0.221
(3)	BM25	2011	Shot	1.0	0.0	0.172
(4)	BM25	2011	Shot	0.2	0.3	0.157
(5)	BM25	2010	Shot	1.0	0.0	0.118

Run (1) was our baseline run. From the difference between (1) : (2) as well as (3) : (4) we see that combination of evindence sources decreases the performance. Because this is counter intuitive, we plan to investigate the reason for this in future work. Note that the weight of the segment is lower than the document transcript and the meta data. This suggests, that it is more important to first rank the video and only later the entry point. The version of the ASR transcripts (3) : (5) also performed rather different. Finally, the differences (1) : (3) and (2) : (4) suggest that speaker segments are better entry points than shots.

4. CONCLUSIONS

This paper described a basic approach to combine evidence to find segments in a video which might contain relevant information to a user's query. The ranking function for segments linearly combined a normalized score of the evidence found for each segment by a text retrieval model together with evidence found in the transcript of the whole video and its meta data. Relative to a found segment, we investigated different alternatives for the entry point returned to the user. Among a large set of combinations, we found that the entry point alternative is the most influential. Overall, using the BM25 retrieval model together with the beginning of shot which contained the start of the found segments as an entry point produced the strongest performance. The performance produced by our system was low compared to other systems, which we plan to investigate in the future.

References

[1] D. Hiemstra, H. Rode, T. van Os, Roel, and J. Flokstra. Pftijah: text search in an xml database system. In *Proceedings of the 2nd International Workshop on Open Source Information Retrieval (OSIR), Seattle, WA, USA*, pages 12–17. Ecole Nationale Supérieure des Mines de Saint-Etienne, 2006.

[2] M. Larson, M. Eskevich, R. Ordelman, C. Kofler, S. Schmiedeke, and G. Jones. Overview of MediaEval 2011 Rich Speech Retrieval Task and Genre Tagging Task. In *MediaEval 2011 Workshop*, Pisa, Italy, September 1-2 2011.

[3] J. H. Lee. Analyses of multiple evidence combination. In *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '97, pages 267–276, New York, NY, USA, 1997. ACM. ISBN 0-89791-836-3. doi: 10.1145/258525.258587.

[4] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 232–241, New York, NY, USA, 1994. Springer-Verlag New York, Inc. ISBN 0-387-19889-X.

[5] T. Westerveld, J. C. van Gemert, R. Cornacchia, D. Hiemstra, and A. P. de Vries. An integrated approach to text and image retrieval the lowlands team at TRECVid 2005. In *Proceedings of the 3rd TRECVid Workshop*, 2005.