DCLab at MediaEval2014 Search and Hyperlinking Task

Zsombor Paróczi
Inter-University Centre for Telecommunications and Informatics
paroczi@tmit.bme.hu

Bálint Fodor
Inter-University Centre for Telecommunications and Informatics
fodor@aut.bme.hu

Gábor Szűcs
Inter-University Centre for Telecommunications and Informatics
szucs@tmit.bme.hu

ABSTRACT

The aim of the paper was to support the answer to a query with a ranked list of video segments (search task) and to generate possible hyperlinks (in ranked order) to other video segments in the same collection that provide information about the found segments (linking task). Our solution is based on concept enrichment i.e. the set of words is extended with their synonyms or other conceptually connected words. The other contribution is the content mixing using the combination of all transcripts and manual subtitles of the videos.

1. INTRODUCTION

Our paper is about a user who searches for different segments of videos within a video collection that address certain topic of interest expressed in a query. If the user finds the segments that are relevant to his initial information need, he may wish to find additional information connected to these segments [1]. Our aims were to support the answer to a query with a ranked list of documents (search task) and to generate a ranked list of video segments in the same collection that provide information about the found segments (linking task). Both sub-tasks represent ad-hoc retrieval scenario, and were evaluated by organizer of the challenge.

We used the same collection of the BBC videos as a source for the test set collection. Collection of BBC consists of video keyframes, audio content, 3 sets of automatic speech recognition (ASR) transcripts: LIMSI/Vocapia [3, 5], LIUM [7] , NST/Sheffield [6, 4] furthermore 1 manual subtitles, metadata and prosodic features [2].

2. SYSTEM OVERVIEW

During the tasks we developed a small system for processing the data. Our solution is solely based on textual analysis, we only used the subtitles and ASR transcripts. It has 5 distinctive stages: data normalization (2.1), shot cutting (2.2), concept enrichment (2.3), content mixing (2.4), indexing and retrieval (2.5).

2.1 Normalization

The data set was given in various forms, so the first step was to normalize the data formats and to convert all data to the same scale. We used the time dimension as scale and csv as the common data format.

2.2 Shot cutting

Since in the data set each file represented a whole television program and we wanted to work on 'shot' level we created a tool, that based on the provided 'scenecut' description cuts each input data into shots. Using this method we created more than 300000 small files, each representing one shot with only one metric (like LIMSI transcript).

Our main goal was to create a concept enrichmented so called 'shot-document' file for each shot with each metric, by doing this the content can be found using synonyms in the search query. For example if the search query is "dog" and there is a shot-document which has the word 'puppy' in it, the aim is to connect them and return the needed result.

2.3 Concept enrichment

Our concept enrichment stage consists of three text transformation stages. First, each word in the shot-documents is analysed by the phpMorphy[1] morphology engine. This engine can create the normal form (stem) of each word using basic grammatical rules and a large dictionary. This engine also works with German and Russian besides English. In our work we replaced every word with its normal form. In this point the shot-document is only a bag of words.

After this step we filtered out the stop words, we used 702 different English stop words[2] for that, including search term like words e.g.: less, important. This way we narrowed down the word list of a shot-document to its core concept.

For a better match we needed to enrich this list with synonyms and conceptually connected other items. For this we used the well known ConceptNet 5[3] system, which can give us other words / phrases connected to each word in a shot-document. We experimented with a wider range solution: including 50 conceptually connected words for each word in the shot document and a smaller range solution: including only 10 connected word. In the results the (C2) notates the smaller range result. We introduced a weight for each word, the "original" words inside the shot-document's weight is 1, the weight of connected words are lower (for wide range: 0.2, for small range: 0.1). At aggregation

[1]http://sourceforge.net/projects/phpmorphy
[2]http://www.ranks.nl/stopwords, 'Long Stopword List' section
[3]http://conceptnet5.media.mit.edu/

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

all of the enriched words there can be duplicates (like 'home' is connected to 'school' and 'teacher' is connected to 'school'), we aggregate them by a simple weight sum. Using this method the weight represents the importance of a word in the conceptual graph (sum of all words in the shot-document).

2.4 Content mixing

We created multiple shot-document types (3 transcripts and manual subtitles), furthermore a combined type, so called "All transcript and subtitle". This later case was created by taking each shot-document with word weights and put together by the same sum method explained before. This way we could represent each and every possible word in our concept file, but on the other side we added a lot of conceptual noise to the originally clean document.

2.5 Indexing and retrieval

For indexing each shot-document we used Apache Solr[4] since it supports text indexing and retrieval with a lot of adjustable variables. Each shot-document is considered in Solr as a single continuous text stream, the order of the words represented the weight in the shot-document. Important note is that during the word reordering we kept concept phrases as one entity.

In the search sub-task the retrieval we only included the following steps: stop word filtering for the query, creating the norm form for each word in the query, using the query as a search input in Solr. The result was limited to 30 retrieved items.

In the linking sub-task we used the shot-document representing the needed section as a search query, but we removed the concept enriched words from it. So it was basically the core concept of the shot used as a simple text search input.

3. RESULTS AND CONCLUSIONS

The whole dataset was more than 3700 hours of video and the evaluation was on a shot level base (sometimes less than 5 seconds).

3.1 Searching sub-task

	P@5	P@10	P@20
Manual subtitles	.1778	.2000	.1407
LIMSI transcripts	.1481	.1667	.1185
LIUM transcripts	.1630	.1444	.1148
NST/Sheffield transcripts	.1769	.1308	.0981
All transcripts and subtitles	.1517	.1345	.1017
Manual subtitles (C2)	.3407	.3074	.2074
LIMSI transcripts (C2)	.3111	.2926	.2204
LIUM transcripts (C2)	.3704	.2815	.2204
NST/Sheffield transcripts (C2)	.2846	.2231	.1692
All transcripts and subtitles (C2)	.1655	.1586	.1190

Table 1: P@N result for the searching sub-task

In the search sub-task we reached a quite stable result for each subtitle / transcript. Using a manually written transcript is much better since it can include visual clues, non-spoken informations (e.g. texts) and it has lower error rate, on the other hand in the transcripts there can be 'misheard' sentences.

The biggest surprise is the failure of our context mixing method since it underperformed in almost all cases.

3.2 Linking sub-task

	P@5	P@10	P@20
Manual subtitles	.0750	.0500	.0312
LIMSI transcripts	.0444	.0333	.0167
LIUM transcripts	.0533	.0400	.0200
NST/Sheffield transcripts	.0400	.0467	.0233
All transcripts and subtitles	.0370	.0407	.0222
Manual subtitles (C2)	.1818	.1000	.0500
LIMSI transcripts (C2)	.0500	.0625	.0375
LIUM transcripts (C2)	.0526	.0316	.0184
NST/Sheffield transcripts (C2)	.0300	.0350	.0175
All transcripts and subtitles (C2)	.0143	.0250	.0196

Table 2: P@N result for the linking sub-task

In the linking sub-task the Manual subtitles gave us the best result, but it is interesting to note that for 2 of the anchors we cannot find any relevant items among all of our results, that is why the P@N results are so low. These anchors are *anchor_22* and *anchor_27*.

4. ACKNOWLEDGMENTS

The publication was supported by the TÁMOP-4.2.2.C-11/1/KONV-2012-0001 project. The project has been supported by the European Union, co-financed by the European Social Fund.

5. REFERENCES

[1] M. Eskevich, R. Aly, D. Racca, R. Ordelman, S. Chen, and G. J. F. Jones. The Search and Hyperlinking Task at MediaEval 2014. In *Proceedings of the MediaEval 2014 Multimedia Benchmark Workshop*, Barcelona, Spain, 2014.

[2] F. Eyben, F. Weninger, F. Gross, and B. Schuller. Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 835–838, Barcelona, Spain, 2013. ACM.

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

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

[5] L. Lamel. Multilingual speech processing activities in QUAERO: Application to multimedia search in unstructured data. In *Human Language Technologies–The Baltic Perspective: Proceedings of the Fifth International Conference Baltic HLT 2012*, volume 247, pages 1–8. IOS Press, 2012.

[6] P. Lanchantin, P. Bell, M. Gales, T. Hain, X. Liu, Y. Long, J. Quinnell, S. Renals, O. Saz, M. Seigel,

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

S. Swietojanski, and P. Woodland. Automatic transcription of multi-genre media archives. *First Workshop on Speech, Language and Audio in Multimedia (SLAM 2013)*, 2013.

[7] A. Rousseau, P. Deléglise, and Y. Estève. Enhancing the TED-LIUM corpus with selected data for language modeling and more TED Talks. In *The 9th edition of the Language Resources and Evaluation Conference (LREC 2014)*, Reykjavik, Iceland, 2014.