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

Abstract

1. INTRODUCTION

TODO: cite the test paper

2. SYSTEM OVERVIEW

During the task we developed a small system for processing the data we have. Our solution is solely based on textual analysis, we only used the subtitles and transcripts. It has 5 well distinctive stages: data normalization (i), shot cutting (ii), concept enrichment (iii), content mixing (iv), indexing and retrieval (v).

2.1 Normalization

The data set was given in various forms, so the first step of our system 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 have been cut each input data into shots. Using this method we created more than 300.000 small files, each representing one cut in one data dimension (like LIMSI transcript).

Our main goal was to create a concept enrichmented so called 'shot-document' file for each cut in each dimension, that way even if the user is using some synonym will find the content. For example the search was for "dog" and we have a shot-document which has the word 'puppy' in it we can make the connection and return the needed result.

2.3 Concept enrichment

Our concept enrichment stage consists of three text transformation stage. First each word in the shot-documents is analysed by the phpMorphy [1] morphology engine. This engine can create the normal form of each word using basic grammatical rules and a large dictionary. This engine also works in the German and Russian language 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 for that, including search term like words like: less, important.

This way we narrowed down the word list of a shot-document to its core concept. But for a better a wider match we needed to enrich this list with synonyms and conceptually connected other items. For this we used the well known ConceptNet 5 [2] system, which can give us other words / phrases connected to each word in our shot-document. We experimented with a wider range solution: including 50 conceptually connected word 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 gave each word a weight, the "original" words inside the shot-document's weight is one, every connected word gets a lower weight (for wide range: 0.2, for small range:0.1). When we aggregate all of the enriched words there can be duplicates (like 'home' is connected to 'school' and 'teacher' connected to 'school'), we aggregate them by a simple weight sum. This way we can grab the conceptual graph of each words by using weight as a "connectedness" and "importance" value in the shot-document.

2.4 Content mixing

We created multiple shot-document types, one of them consists of more than one type of 'base documents'. The "All transcript and subtitle" 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 down side we added a lot of conceptual noice to the originally clean document.

2.5 Indexing and retrieval

For indexing each shot-document we used Apache Solr [3] since it supports text indexing and retrieval with a lot of adjustable values. We gave each shot-document to Solr as a single continuous text stream, using the weight as the order of the words. Important note is that we kept the whole phases as one word during processing.

In the search subtask the retrieval 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 in Solr. The result was limited to 30 items.

[1] Source: http://sourceforge.net/projects/phpmorphy

[2] Source: http://conceptnet5.media.mit.edu/

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

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

In the linking subtask 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.

3. RESULTS

The whole dataset was more than 3700 hours of video and the evaluation was done on a shot level base (sometimes less than 5 seconds). Picking random shots as results should give us 0 relevancy since the topics covered in the dataset are wide enough to avoid accidental matches.

3.1 Searching subtask

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

Table 1: P@N result for the searching subtask

In the search subtask we reached a quite stable result for each subtitle / transcript, but it's clear, that using a manually written transcript is much better since it can include visual clues, non-speaken informations (like texts) and has a low error rate. In the transcripts there can be 'misheared' sentences.

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

3.2 Linking subtask

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

Table 2: P@N result for the linking subtask

In the linking subtask subtasks gave us the best result, but it's interesting to note that from the 8 anchors specified we cannot fight a single relevant items among all of our tests. That is why the average relevancy is this low. These anchors are *anchor_22* and *anchor_27*.

4. CONCLUSIONS

5. 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.

6. REFERENCES