Ghent University-iMinds at MediaEval 2013: An Unsupervised Named Entity-based Similarity Measure for Search and Hyperlinking*

Tom De Nies[1], Wesley De Neve[1,2], Erik Mannens[1], Rik Van de Walle[1]

[1]Ghent University - iMinds - Multimedia Lab
[2]Korea Advanced Institute of Science and Technology (KAIST) - IVY Lab
{tom.denies, wesley.deneve, erik.mannens, rik.vandewalle}@ugent.be

ABSTRACT

In this paper, we describe our approach to the Search and Hyperlinking task at the MediaEval 2013 benchmark. This task focuses on video retrieval and linking in the context of a large and rich dataset provided by the BBC. Our approach makes use of one of three types of audio transcripts, enriched with Named Entities. To compute similarity, we adapt the Jaccard metric to use Named Entities. This results in an unsupervised and computationally inexpensive way of searching and linking multimedia content.

1. INTRODUCTION

The Search and Hyperlinking task at MediaEval 2013 [3] provides a benchmark for retrieval and linking of video segments, constructed and evaluated with the aid of crowdsourcing and a dataset of broadcast material provided by the BBC. The dataset consists of 1260 hours of video, corresponding textual metadata, and manually and automatically generated transcripts for the speech in each video. The task consists of two parts: the search task and the linking task. For the search task, 50 known-item queries are provided, for which the corresponding video segments need to be retrieved. For the linking task, 98 anchor segments are provided, for which a ranked list of related segments needs to be created, which is then evaluated using Human Intelligence Tasks (HITs). For more details about the task, we refer to the corresponding overview paper [3].

For our approach, we only make use of the transcripts provided with the dataset. For each video, three types of transcripts are provided: human-generated subtitles, and two ASR transcripts, one provided by LIUM [7], and one provided by LIMSI [5]. We use the timing information of these transcripts to divide the videos into time-based segments. Next, each of these segments is enriched using a Named Entity Recognition (NER) service. These Named Entities (NEs) are then used to facilitate the search and linking process. In the next section, we describe each of these steps in detail.

*These research activities were funded by Ghent University, iMinds, the IWT Flanders, the FWO-Flanders, and the European Union.

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

2. PROPOSED APPROACH

Our approach consists of two main phases: **ingestion** and **run execution**.

As it is time consuming, the ingestion phase is performed before the runs are executed. It has three steps: content representation, segmentation, and enrichment. As **content representation** for the videos, we use one of the three provided transcripts, either the subtitles, LIUM transcripts, or LIMSI transcripts, selected at runtime. For the **segmentation** step, the timing information of the transcripts is used to divide the videos into segments. Due to the success of time-based segmentation in the 2012 benchmark [2], we choose to employ fixed-length temporal segments of L seconds (in our submitted runs, $L = 30$). Note that the actual length of these segments might vary, due to silences and non-speech fragments in the audio. Each of these segments is then **enriched** by extracting NEs from it using an NER service. In our case, the annotation feature of DBpedia Spotlight[1] is used to extract all types of NEs it can find. As this is a network-based service, this is the most time consuming step in our approach.

The execution phase consists of two steps: similarity calculation and result selection. In case of the search task, the **similarity** between the queries and the segments in the dataset is calculated, whereas in case of the linking task, this is done between the anchor segments and the other segments in the dataset. To calculate the similarity, we opt for a completely unsupervised similarity metric. This way, no computationally expensive training step is necessary, that would have to be repeated upon expansion of the dataset, as would be the case when using a supervised similarity metric. In our case, we use the Jaccard metric, applied to NEs. We calculate the similarity between two enriched documents (be it segments, videos, or queries) A and B as follows:

$$Sim(A,B) = \frac{|\{e : e \in E(A) \cap E(B)\}|}{|\{e : e \in E(A) \cup E(B)\}|}, \qquad (1)$$

where $E(A)$ and $E(B)$ denote the sets of extracted NEs from document A and B, respectively. Note that no NEs could be extracted from some of the queries provided for the search task, due to their short length. When this is the case, we revert to a fallback mechanism, using a slightly altered metric. The fallback mechanism consists of creating a set of keywords for both the query and segments using a naive keyword extraction algorithm. This algorithm extracts all

[1]`http://spotlight.dbpedia.org/`

distinct words in the query or segment as keywords, with all stop words[2] removed. We then employ the same calculation as in Equation 1, with the difference that E(A) and E(B) now represent the sets of keywords. For each segment of length L, the optimal result segment length is determined, by maximizing the similarity score for x consecutive segments, with $x = 1, 2, ..., W$ for a maximum window size W and maximum segment length L_{max} (in our case, $W = 4$, and $L_{max} = 2$ *minutes*). Finally, the N segments with the highest similarity score are **selected** and ranked, to be returned as results of the run (in case of our runs for the search task, $N = 500$; in case of the linking task, $N = 20$).

3. RELATED WORK

There are other approaches found in literature that make use of NEs, sometimes dubbed as "concepts", for the purpose of measuring similarity between documents. These concepts are mostly used to determine weighting schemes, such as CF-IDF [4], or used as direct input for a similarity metric, such as NESM [6]. For our Search and Hyperlinking task submission 2012 we used a NE-based weighting scheme as a component in a supervised late fusion approach [1].

4. EXPERIMENTS AND EVALUATION

We submitted five runs in total, three for the search task, and two for the linking task. Our three runs for the search task were performed using the approach described in Sect. 2, with each run using one of the three transcript forms: subtitles (S), LIUM (U), or LIMSI (I). Before the runs, the queries were enriched with NEs using the same NER service as used for the segments. Then, the similarity between the queries and all the segments in the dataset was computed as described in Sect. 2. In Table 1, we present the mean reciprocal rank (MRR), mean generalized average precision (mGAP), and mean average segment precision (MASP) of the three search runs. For more information about these evaluation metrics, we refer to the task overview paper [3].

Run type	MRR	mGAP	MASP
I	0.0322	0.0222	0.0268
U	0.0546	0.0322	0.0515
S	**0.149**	**0.0906**	**0.123**

Table 1: Results of the submitted search task runs

For the linking task, we submitted two runs, each making use of the subtitles, since these lead to the best result in the search task. The first run (A) was performed using only the segment that contained the anchor itself as input, whereas the second run (C) also made use of the context of surrounding segments, defined by the user who chose the anchor. The runs were evaluated by human users, resulting in precision at rank $x \in \{5, 10, 20\}$ (P_x), and mean average precision (MAP). The results are shown in Table 2.

Run type	MAP	P_5	P_{10}	P_{20}
A	0.0375	0.3200	0.2800	0.1667
C	**0.0459**	**0.3867**	**0.3500**	**0.2050**

Table 2: Results of the submitted linking task runs

[2] `http://users.ugent.be/~tdenies/util/stopwords.txt`

5. DISCUSSION AND FUTURE WORK

When we compare the results of both tasks with those of our (partially) supervised late fusion approach, submitted to the benchmark of 2012 [1], we observe that for both tasks, we obtain less accurate results with our unsupervised approach than with a supervised one, as could be expected. Note that these two years are not entirely comparable, since a different dataset was used. When inspecting the results of the 2013 linking task, we see that for our second run (C), 35% of the top ten (P_{10}) and 38.67% of the top five links found (P_5) are evaluated as relevant to the anchor segment, which is a promising result. This also suggests that considering more context leads to better link quality. The unsupervised approach is certainly more flexible and computationally efficient, considering that video datasets typically receive frequent additions and removals. With a supervised approach, this would result in frequent re-training and re-indexing.

In future work, we aim to experiment further with unsupervised similarity measures, adapted to work with NEs. Also, the influence of the NER service used needs to be evaluated. Other, more accurate NER services than DBpedia Spotlight exist. However, their free versions are mostly limited in number of requests, making them less suitable for datasets of this magnitude, without significant costs. Application of our approach to multilingual content also remains an important challenge we aim to address. We also aim to exploit more of the semantic features of NEs, taking advantage of the similarity between individual concepts.

6. REFERENCES

[1] T. De Nies, P. Debevere, D. Van Deursen, W. De Neve, E. Mannens, and R. Van de Walle. Ghent University-IBBT at MediaEval 2012 Search and Hyperlinking: Semantic Similarity using Named Entities. In *MediaEval 2012 Workshop*, 2012.

[2] M. Eskevich, G. J. Jones, R. Aly, R. J. Ordelman, S. Chen, D. Nadeem, C. Guinaudeau, G. Gravier, P. Sébillot, T. De Nies, et al. Multimedia information seeking through search and hyperlinking. In *3rd ACM conference on International conference on multimedia retrieval (ICMR)*, pages 287–294. ACM, 2013.

[3] M. Eskevich, G. J. F.. Jones, S. Chen, R. Aly, and R. Ordelman. The Search and Hyperlinking Task at MediaEval 2013. In *MediaEval 2013 Workshop*, Barcelona, Spain, October 18-19 2013.

[4] F. Goossen, W. IJntema, F. Frasincar, F. Hogenboom, and U. Kaymak. News personalization using the CF-IDF semantic recommender. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, page 10. ACM, 2011.

[5] L. Lamel and J.-L. Gauvain. Speech Processing for Audio Indexing. *Advances in Natural Language Processing. (LNCS 5221)*, pages 4–15, 2008.

[6] S. Montalvo, V. Fresno, and R. Martínez. NESM: a Named Entity based Proximity Measure for Multilingual News Clustering. *Procesamiento del lenguaje natural*, 48:81–88, 2012.

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