Eurecom-Polito at TRECVID 2017: Hyperlinking task

Benoit Huet
EURECOM
Sophia Antipolis, France
huet@eurecom.fr

Elena Baralis, Paolo Garza, and Mohammad Reza Kavoosifar
Department of Control and Computer Engineering
Politecnico di Torino
Torino, Italy
{name.surname}@polito.it

ABSTRACT

This paper describes the system we designed to address the Hyperlinking task at TRECVID 2017 and the achieved results. Our contribution explores the potential of a solution based on the combination of textual and visual features in order to consider the different facets of the input videos. In particular, our approaches combined automatically generated transcripts (LIMSI), visual concepts, Meta-data, the text extracted by means of a Name Entity Recognition technique and a concept mapping tool. The four submitted runs aimed at analyzing the impact of the considered features on the quality of the retrieved hyperlinks.

1. INTRODUCTION

In this paper, we describe the framework used by the Eurecom-Polito team to address the Hyperlinking task inside a video collection at TRECVID 2017 [2].

The Hyperlinking task aims at linking anchors related to a temporal segment of a video. In this task, one of the main challenges is the ambiguity regarding what criteria are to be followed to generate these links. There is uncertainty about what the user expectations are regarding these links, as well as little information about what is considered relevant to the user in the video segment.

The data used in the TRECVID 2017 competition consists of 14,838 videos, for a total of 3,288 hours, provided by blip.tv.

We have proposed a system that exploits (i) automatic speech recognition transcripts [5, 7], (ii) visual concepts, (iii) the entities extracted by means of a Name Entity recognition technique, and (iv) a concept mapping technique, which is based on WordNet [3] for identifying relevant concepts.

The paper is organized as follows. Section 2 introduces the proposed system and the exploited video features. Section 3 describes the configurations of the four submitted runs and discusses how they have been selected, while Section 4 discusses the achieved results. Finally, Section 5 draws conclusions.

2. SYSTEM OVERVIEW

For the Hyperlinking task, we developed a system based on both textual and visual features. We exploited all the data and meta-data provided by the task organizers, except visual concepts. Specifically, we decided to use the visual concepts extracted by using the Caffe framework with the BVLC GoogLeNet model [9]. We also considered some other extra features. Specifically, to identify the more relevant terms and concepts in each query we used the Stanford Named Entity Recognizer (NER) [4] software to find entities and a Concept mapping technique based on WordNet [3].

The proposed system uses (i) automatic speech recognition transcripts (LIMSI) [5, 7], (ii) visual concepts, based on the Caffe framework, (iii) meta-data of the videos (specifically, title, description and tags have been considered), and (iv) query reformulation (based on Named-entity recognition and Concept mapping).

The core of all runs is composed of three stages: Data segmentation (Section 2.1), Indexing (Section 2.2), Query formulation and Retrieval (Section 2.3).

2.1 Data segmentation

The first step that is applied on the video collection consists in splitting the videos in segments. We used a Fixed-segmentation, for which we considered 120 sec fixed segments.

All the textual data associated with the segments have been preprocessed to remove irrelevant words. Specifically, we used a punctuation removal tool and we also removed stop words. We used 665 different English stop words for that. This way we narrowed down the word list of each segment to its core concepts.

2.2 Indexing

We used Apache Solr [1] to index the textual and visual features associated with each segment. We created multiple indexes for the segments. Specifically, we created indexes based on the LIMSI transcripts and the visual concepts of the segments and also on the meta-data of the videos. The index created by Solr is known as an inverted index. An inverted index stores, for each term, the list of documents containing it. This makes term-based searches very efficient [6].

2.3 Query formulation and Retrieval

In this stage, we first transform the anchor (query) segment into a textual query by including in the text of the query all the textual information associated with the anchor (i.e., the LIMSI transcripts and the relevant visual concepts) and also the meta-data of the video containing the anchor (i.e., title and tags of the video containing the anchor).

Named-entity recognition is applied on LIMSI in order to extract the important names inside the query and give them

a higher relevance. Named Entity Recognition (NER) labels sequences of words in a text which are related to the names of things, such as person and company names, or gene and protein names.

We also exploited a concept mapping technique that is based one WordNet. It is used to find the most relevant visual concepts inside each query. For each anchor, the mapping is done by using the words appearing in meta-data of the video con taining the anchor and the list of concepts associated with the anchor.

After the query preparation phase, a tool executes it by using Apache Solr and returns the related segments ranked by relevance.

3. SUBMITTED RUNS

For the Hyperlining sub-task, we submitted four runs by using four different approaches. The considered approaches use different features and/or combine them by using different strategies. Before selecting the configurations of the four runs, we performed a set of experiments on the development anchors to evaluate the impacts of the two available transcript tools (LIUM vs LIMSI [5, 7]) and two video segmentation techniques (shot segmentation vs fixed length segmentation). On the average, on the development anchors, the LIMSI transcripts allow achieving better results than the LIUM ones and Fixed-segmentation allows retrieving more relevant segments than the shot segmentation-based approach. Hence, the four submitted runs use the LIMSI transcripts and fixed-segmentation.

The approaches associated with the four submitted runs aimed at analyzing the impacts of some of the salient components of our system. Specifically, the characteristics of the four submitted runs are the followings:

Run 1. Automatic Feature Selection (AFS): In the approach associated with this run, we used the following features: Meta-data, the LIMSI transcripts and Visual concepts. We also applied a Named-entity recognition (NER) technique to identify entities and a Concept mapping technique to identify the most relevant visual concepts. During the execution of the query, a higher importance is given to entities and the visual concepts selected by the concept mapping technique. This run exploits all the available features and all the building blocks/components of our system.

The AFS approach is based on two steps. In the first step, AFS considers one feature at a time and selects the subset of relevant segments for each feature. In the second step, the subsets of segments retrieved in the first step are merged and ranked in terms of relevance score. The output of this second step is the final result of this approach.

Run 2. Meta-data based approach: Similarly to Run 1, also this second run uses all the components of our system. Specifically, it considers all the features and also the named-entity recognition (NER) and the concept mapping techniques. However, differently from Run 1, Meta-data are used to perform an initial filter on the videos that could contain interesting segments. In the second step, the same technique used in Run 1 is applied to select the most relevant segments only from the subsets of segments of the videos selected in the first step. However, only LIMSI and visual concepts are considered in this second step of Run 2.

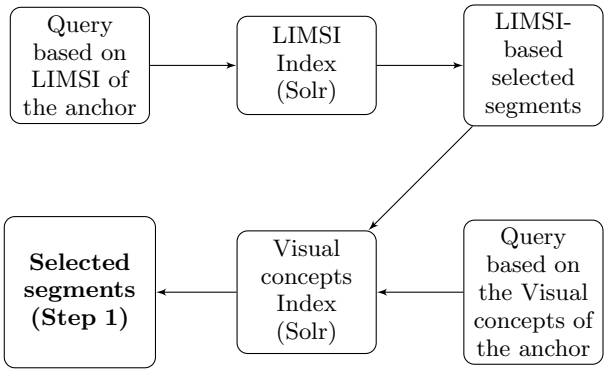

Figure 1: Pipeline approach: Step 1

Table 1: Evaluation result

Measure	Run 1	Run 2	Run 3	Run 4
P@5	0.8400	0.7040	0.7250	0.8080
P@10	0.8080	0.5560	0.6667	0.7480
MAP	0.1638	0.0815	0.0930	0.1135
MAiSP	0.2527	0.1320	0.1547	0.1851

Run 3. LIMSI-NER: In this approach, we considered only the LIMSI transcript feature and we applied the named-entity recognition (NER) technique on the queries. The aim of this approach is to analyze the differences between Monomodal and Multimodal approaches. We selected the LIMSI transcript feature because, on the development anchors, it usually performs better than the other features.

Run 4. Pipeline approach: For the fourth run, we used only two features: LIMSI and Visual concepts. Also in this run, we applied the Named-entity recognition (NER) and the Concept mapping techniques. In the Pipeline approach, for each anchor, we first select the top-k relevant segments by using a query based on LIMSI and then we refine the result by querying the subset of returned segments by means of a query based on the visual concept feature (see Figure 1). The same operation is then performed by switching the order of the two queries. Finally, the two subsets of returned segments are merged and ranked in terms of relevance score.

4. RESULTS

For the evaluation of the results, a set of metrics have been used: precision at rank 5 and 10 (P@5 and P@10), MAP, and MAiSP [8]. The results of the four runs we submitted at the Hyperlinking task are reported in Table 1.

Run 1 (Automatic Feature Selection) yields the best results in term of all the considered metrics. We recall that it exploits all the available features (LIMSI transcripts, visual concepts, and Metadata). Also the Pipeline approach (Run 4) is characterized by high values for all metrics. However, it performs worse than Run 1. Hence, pipeline the queries seem to have a negative impact on the final results.

Another difference between Run 4 and Run 1 is that in Run 4 we do not consider the Meta-data feature. Hence, in some cases, it probably allows selecting relevant segments.

Run 2 (the Meta-data based approach) achieved the lowest result. This was slightly unexpected as performance of this run on the development anchors was higher in compare to those of Run 3 and Run 4. This is most likely due to the fact that using the Meta-data for pre-filtering videos would raise the problem of selecting very few related videos for some anchors. Hence, for some anchors this approach returns few segments.

Finally, the results confirm that the exploitation of more features is usually better than using one single feature (the results of Run 3, which is based only on LIMSI, are on the average lower than those of Run 1 and Run 4).

5. CONCLUSION

The proposed system has explored the use of textual and visual features for solving the Hyperlinking task. Specifically, we have considered the LIMSI transcripts, visual concepts and Meta-data. Moreover, named-entity recognition and a concept mapping technique have also been considered.

The achieved results show that the proper combination of several features performs better than single features.

6. REFERENCES

[1] Apache Solr. http://lucene.apache.org/solr.

[2] G. Awad, A. Butt, J. Fiscus, D. Joy, A. Delgado, M. Michel, A. F. Smeaton, Y. Graham, W. Kraaij, G. Quénot, M. Eskevich, R. Ordelman, G. J. F. Jones, and B. Huet. Trecvid 2017: Evaluating ad-hoc and instance video search, events detection, video captioning and hyperlinking. In *Proceedings of TRECVID 2017*. NIST, USA, 2017.

[3] C. Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

[4] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 363–370. Association for Computational Linguistics, 2005.

[5] J.-L. Gauvain. The quaero program: Multilingual and multimedia technologies. In *International Workshop on Spoken Language Translation (IWSLT)*, 2010.

[6] J. Kumar. *Apache Solr Search Patterns*. Packt Publishing Ltd, 2015.

[7] L. Lamel. Multilingual speech processing activities in quaero: Application to multimedia search in unstructured data. In *Baltic HLT*, pages 1–8, 2012.

[8] D. N. Racca and G. J. Jones. Evaluating search and hyperlinking: An example of the design, test, refine cycle for metric development. In *MediaEval*, 2015.

[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.