UNED at MediaEval 2010: exploiting text metadata for Automatic Video Tagging

David Hernández-Aranda, Rubén Granados, Juan Cigarran,
Álvaro Rodrigo, Víctor Fresno and Ana García-Serrano
{daherar, rgranados, juanci, alvarory, vfresno, agarcia}@lsi.uned.es
NLP & IR Group, UNED, Madrid

ABSTRACT

In this paper we present the first participation of the NLP&IR group at UNED in the Tagging Task (Professional Version): prediction of semantic theme. This categorization task was carried out by an information retrieval approach, together with language models and clustering using only metadata associated with the videos. The results show that language models are useful for enriching the representation of information associated with videos.

1. INTRODUCTION

MediaEval is an expansion of the VideoCLEF track. We have participated at the Tagging Task (Professional Version), where systems receive a set of videos, associated metadata, and they have to tag them using a set of given labels. Since this has been our first participation at this task, we decided to use only the textual metadata associated to videos.

We have considered only "description" and "description_abstract" fields, which summarize the contents of the video footage, from the metadata set. Since the metadata set was in Dutch language and we did not have resources to work in this language, we translated the selected metadata subset into English using the Google Translator API.

2. SYSTEM OVERVIEW

We can distinguish two main phases in our system: a) the representation step, where the video collection is processed and organized, and; b) the categorization step, where the candidate videos to be labelled are retrieved and ranked in order to decide their tags.

2.1 Representation

We have selected clustering as the modelling approach for representing the metadata associated to videos into the document space. Thus, similar videos were grouped into clusters using a partition algorithm belonging to Cluto package [2]. We applied the *rbr* algorithm, where the desired k-way clustering solution is computed by performing a sequence of $k-1$ repeated bisections, being k the number of clusters to generate and fixed to the number of labels plus one. The main reason for this, was to have a cluster for each label and one extra cluster for not tagged documents. Our aim in performing clustering was not only to find similarities between videos, but also to generate a high level representation where the basic unit of information to be indexed and retrieved is a cluster. After this, the tag given to a cluster is given to all the documents of that cluster.

In order to dispose of a more accurate representation, we used language models for ranking the terms of a document according to their relevance with respect to the other documents in the collection. We applied the Kullback-Leibler divergence (KLD) [3]. Language models can be applied in two points of our pipeline (depending on the experiment) : a) before the clustering process, and b) after the clustering process. In the former case, KLD is applied to calculate the divergence between the language model of each document with respect to the collection, and the output is the weighted term-document matrix used to feed the clustering algorithm. In the later case, we represented each cluster as a large document obtained as the merge of all its items (i.e. documents). So, the divergence was computed between the language model of each cluster with respect to the collection of clusters.

2.2 Classification

At this step, we applied an Information Retrieval (IR) approach which uses candidate labels as queries and the generated clusters as indexed items. However, the use of standalone labels to query the system could drastically reduce the recall performance. This is why we decided to expand the query using semantic information directly extracted from WordNet. More specific, we added synonyms and hyponyms from all the synsets of the label. In order to avoid a possible lost of precision, we expanded each label into several queries. Actually, we created for each synset a query with its synonyms, and a query for each hyponym of that synset. As a result of this process, the final number of queries associated to each label can vary (e.g. 100 queries for the label *actor*).

The IR module applies a Vector Space Model (VSM) ranking function to calculate the similarities [1]. Since the indexed items are clusters, it is necessary to represent them in order to compute tf-idf weights. We generate the cluster representation by concatenating all the documents contained within each cluster and using the generated vector as input to the IR system.

The use of several queries per label produced several cluster rankings per label. Thus, it was necessary to perform some kind of fusion among them. We decided to select the most promising cluster attending to the position given in the different rankings. Our intuition was that the most relevant

MediaEval 2010 Workshop, October 24, 2010, Pisa, Italy

clusters should appear more times in the first positions of each ranking. So, we calculated a score for each cluster depending on its position across the different rankings. More in detail, given a cluster and a ranking we calculated a value as the result of the position of that cluster in the ranked list divided by the total number of clusters retrieved. Then, the local score given to the cluster in that ranking was equal to one minus the obtained value. Thus, the first cluster of the ranked list received the best local score (i.e. a value close to 1). Once all the local scores were calculated, it was possible to compute a global score for each cluster as the sum of all the local scores for each specific cluster across the different ranked lists. According to the experiments performed at the development stage, we decided to tag only those documents contained in the cluster with the best global score.

3. RUNS SUBMITTED

We submitted four different runs combining the system modules, presented in section 2, in different ways. More in detail, the configuration for each run was as follows:

Run 1: As this run was used as a baseline, it did not apply language models on the processing pipeline. It consisted on performing clustering first, which is calculated using term frequency information, and then the application of the IR engine to the cluster set using the expanded query set as explained before. Finally, the cluster selection method was applied to label the documents.

Run 2: It includes language models to improve the representation of the clusters generated. In this case, language models were applied after the clustering, which means that the clustering process did not rely on the specificity of terms into the documents, but on a term frequency based calculation. The objective of applying language models after clustering was to identify the cluster specific terminology to be used in the indexing and retrieval processes and to remove the noisy ones. In our experiments we represented each cluster using the first 50% best ranked terms.

Run 3: We applied language models before clustering, which implied to identify terminology of each document with respect to the whole collection. The idea behind this process was to guide the clustering in order to generate terminology related clusters. The retrieval and tagging processed were done as explained in run 1.

Run 4: It is a combination of runs 2 and 3. That is, language models were applied both before clustering and after it. The main goal of this run was to study the effect of modelling documents and clusters using language models and compare the results with the previous approaches.

4. ANALYSIS OF RESULTS

We present in Table 1 the results obtained by our runs according to Mean Average Precision (MAP), which was the official measure in MediaEval, as well as precision, recall and its harmonic mean (F-measure). We include these three last measures because we can see this task as a classification problem, and precision, recall and F-measure are common metrics for evaluating classifiers.

The four runs achieved a similar performance for both MAP and F-measure, which means that the different configurations did not contribute too much to changing the global results. This is due to the fact that the experiments were very similar and there were only small differences among them. It is important to remark that, as we labelled all the items within the clusters with the same relevance, we do not consider any ranking of videos. This decision has some implications on the MAP evaluation measure, which takes into account the ranking of relevant results.

Run	MAP	Precision	Recall	F-measure
#1	0.1512	0.2612	0.2475	0.2541
#2	0.1407	0.2487	0.2325	0.2403
#3	0.1694	0.2781	0.2600	0.2687
#4	0.1578	0.2654	0.2475	0.2561

Table 1: Results of the submitted runs

As it has been shown above, the only difference among runs is the inclusion of KLD for feature selection, and where it is placed. The use of language models before the clustering step (run 3) achieved the best performance, which means that terminology extraction before clustering leads to meaningful clustering sets. On the other hand, the extraction of terminology after the clustering (run 2) seems to get worse results than the baseline (run 1). This is because the removal of a terminology subset after the clustering process makes clusters less representative. Finally, the combination of both language model approaches (run 4) performs worse than considering KLD only before clustering for the reasons exposed above.

5. CONCLUSIONS AND FUTURE WORK

We have described in this paper an approach based only on the use of metadata associated to videos for the Tagging Task (Professional Version). Our system applies clustering, language models and IR in order to decide the labels for tagging videos. According to the results, language models offer an important information for representing the information associated to a video. However, we must be careful about where to apply them.

Future work is focused on considering also transcriptions, as well as performing deeper analysis as for example the recognition of Named Entities.

Acknowledgments

This work has been partially supported by the BUSCAMEDIA Project (CEN-20091026), the CREASE project (TIN-2009-14317-C03-03), the Education Council of the Regional Government of Madrid and the European Social Fund.

6. REFERENCES

[1] A. García-Serrano, X. Benavent, R. Granados, E. de Ves, and J. M. Goñi. Multimedia retrieval by means of merge of results from textual and content based retrieval subsystems. In *Multilingual Information Access Evaluation II. Multimedia Experiments*. LNCS 6242, 2010.

[2] G. Karypis. CLUTO: A Clustering Toolkit. Technical Report Technical Report 02-017, University of Minnesota, 2002.

[3] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1), 1951.