QMUL @ MediaEval 2010 Tagging Task: Semantic Query Expansion for Predicting User Tags

Krishna Chandramouli, Tomas Kliegr, Tomas Piatrik, Ebroul Izquierdo
Multimedia and Vision Research Group, School of Electronic Engineering and Computer Science,
Queen Mary, University of London, Mile End Road, E1 4NS, London, UK
{first name.last name}@elec.qmul.ac.uk

ABSTRACT
This paper describes our participation in "The Wild Wild Web Tagging Task @ MediaEval 2010", which aims to predict user tags based on features derived from video such as speech, audio, visual content or associated textual or social information. Two tasks were pursued: (i) closed-set annotations and (ii) open-set annotations. We have attempted to evaluate whether using only a limited number of features (video title, filename and description) can be compensated by semantic expansion with NLP tools and Wikipedia and Wordnet. This technique proved successful on the open-set task with approximately 20% generated tags being considered relevant by all manual annotators. On the closed-set task, the best result (MAP 0.3) was achieved on tokenized filenames combined with video descriptions, indicating that filenames are a valuable tag predictor.

Categories and Subject Descriptors
H.3 [Information storage and Retrieval]: H3.1 Content Analysis and Indexing; H3.7 Digital libraries

General Terms
Measurement, Performance, Experimentation

Keywords
Media annotation, Wikipedia, Wordnet, Semantic similarity

1. INTRODUCTION
The approach presented aims at exploiting the available complementary resources for extracting meaningful semantic information related to the video. Our goal was to evaluate the potential of the various types of textual information relating to the video for a future fusion with a visual classifier. The main strategy we investigated was to take as little information as possible about each video and expand it using complementary resources such as Wikipedia and Wordnet. The paper is organized as follows. Section 2 describes our approach to using complementary resources. Section 3 uses individual classification techniques and Section 4 summarizes the results

2. Exploiting Complementary Resources
As we considered the metadata (i.e. video title, video description, ASR) to be of value in determining the nature of tags, we processed the metadata with GATE[1] NLP framework. The framework included a tokeniser, sentence splitter, and Part-of-Speech (POS) tagger. In addition to the basic text components, we also included a Gazetteer in order to identify entity names in the text based on lists of predefined words. Also, for extraction additional semantic information we included the Java Annotation Pattern Engine (JAPE) to extract hypernyms from Wikipedia. Finally, we also included OpenCalais[2] plugin for extraction of named entities from the textual metadata.

The output annotations from the GATE NLP framework were categorized into six categories as "person", "location", "date", "organization", "opencalais" and "unknown". The entities belonging to person and unknown categories were looked up in WordNet[3] for the existence of a synset. If they exist then the entities are added to a list of potential tags. Entities for which no synset is found and/or they are categorized as "unknown", were further processed using a JAPE hypernym extraction system using Wikipedia as the corpus. The system locates Wikipedia articles that might define the unlabeled entity using a similarity measure that combines text relevance with popularity of the article [3]. From the selected article, a JAPE implementation of Hearst patterns was used to extract a hypernym. This hypernym was then looked up in Wordnet, thus a link between the entity and a Wordnet synset was established.

3. Tag Assignment
In this section, we briefly explain different strategies adopted for selecting the best possible tags for each video to be assigned.

3.1 Closed set Annotation
As the objective is to select best possible tags for each video from the list of tags, the semantic similarity measure previously described is used to derive a measure of relatedness for selecting the best possible tags.

3.1.1 Wordnet-based classification
The system computes the similarity between the synset representing the entity and each of the target tags represented as Wordnet synsets using the Lin Wordnet similarity measure as described in [2]. Both entity and target tags are mapped to Wordnet synsets using the approach described in Section 2.

3.1.2 File Name Similarity
We noticed that a valuable information might be hidden in the file name. People are using filenames as a natural solution for

MediaEval 2010 Workshop. October 24. 2010. Pisa. Italy.

[1] http://gate.ac.uk/

[2] http://www.opencalais.com/

[3] http://wordnet.princeton.edu/

organizing files from a desktop environment, where tags are generally not available. Our hypothesis is that this habit is transferred also to the web environment as well as filenames for content originating from a desktop. To evaluate this idea, we have built a simple filename-based classifier, which links assigned tags to tokens extracted from filename.

3.1.3 Wikipedia-based classification

The input are entities extracted from the text each mapped to a Wikipedia article using the approach described in Section 2. The target tags are also mapped to Wikipedia entity articles. Entity classification is done by computing cosine similarity between the TF-IDF vector (TV) created from the Wikipedia article of the entity and each of the target tags, selecting the tag with the highest similarity. We tried two approaches to compute the IDFs: 1) using train set only and using entire Wikipedia. To make the TVs denser, we also tried aggregating article TVs with TVs of articles it links to [1].

3.1.4 Term-Vector similarity

We built ID3 classifiers on the term-vector representation (both TF and IDF was attempted) from simple tag presence, description, title, ASR [4] and a tokenized filename. The classifier was trained on videos from the development set.

3.2 Open set Annotation

For the open-set annotation, while extracting the potential entities, the number of repetitions for each tag was counted and accordingly the entities with the highest number of repetitions were considered as potential tags.

4. Evaluation

In this section, we present an overview of the evaluation methodology we adopted for both closed-set and open-set annotation.

4.1 Closed set annotation

For the closed-set annotation, the evaluation was treated as a retrieval problem and using the TRECVID evaluation tool, we obtained MAP measure for different runs.

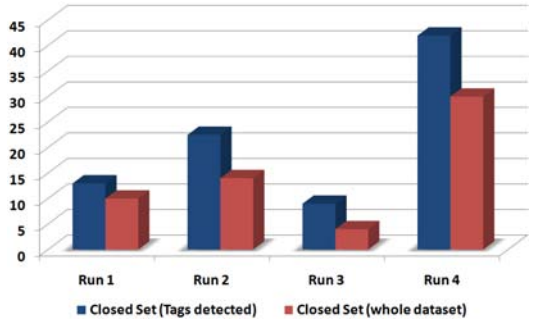

Figure 1 – MAP results from closed set annotation results

In Figure 1, although 1727 videos are present in the dataset, due to either the absence of title and/or description or the absence of named entities from these textual resources, tags were extracted only for 1671 videos. Therefore, the first set of evaluation namely "tags detected" was evaluated against the tags generated for 1671 videos and the second set of evaluation namely "whole dataset" was evaluated against ground truth (tags for 1727 videos). Out of multiple experiments, we decided to submit the following four runs:: Run1 - TV Similarity on ASR+DESC+TITLE, Run2 - TV Similarity on DESC+TITLE, Run 3 - entities from complementary resources, Run 4 – filenames + run 2 predictions for unclassified videos.

4.2 Open set annotation

In order to provide a fair evaluation on the open-set annotation, we randomly selected 40 videos and had seven annotators to manually label if the tags associated to each video are "*relevant*" or "*irrelevant*". As a measure of relevance, we considered the "inter-annotator" agreement [3] among any three or more annotators and the results are summarized in Figure 2.

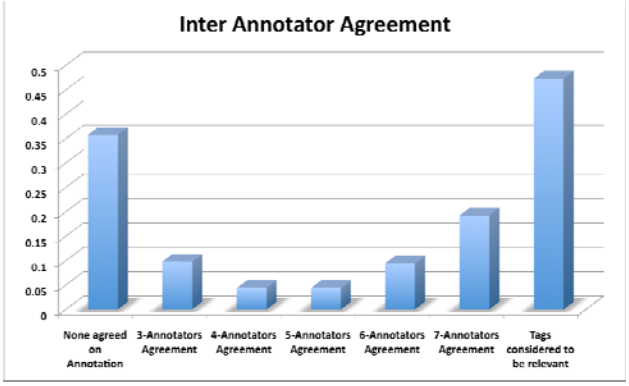

Figure 2 – Inter annotator agreement on the tags

A total of 296 tags were generated for the 40 videos considered for the evaluation and among them, 35.8% tags were considered to be irrelevant by all annotators. As represented in Figure 2, approximately 20% of the tags generated were considered to be relevant by all seven members of the annotators. Considering a tag with more than 3 inter-annotator agreement, then 47.3% of the tags generated were considered to be relevant and with 4 inter-annotator agreement, the percentage drops to 37.5%. For the total dataset of 1727 videos we obtained 6095 unique tags.

5. Conclusion and Future Work

The presented approach is based on detecting named entities from text and furthermore expanding the named entities detected through Wikipedia article search. For the closed-set task, we obtained the best results in terms of MAP (0.3) using a filename-based classifier, which indicates that a tokenized filename is a very strong tag predictor. For the open-set task approximately 20% of the tags generated were considered to be relevant by all seven annotators.

6. ACKNOWLEDGMENTS

The research was partially supported by the European Commission under contract FP7-216444 PetaMedia.

7. REFERENCES

[1] T. Kliegr, "Entity Classification by Bag of Wikipedia Articles", to appear in Doctoral Consortium, CIKM 2010.

[2] K. Chandramouli, T. Kliegr, V. Svatek, E. Izquierdo, "Towards Semantic Tagging in Collaborative Environments", 16th International Conference on Digital Signal Processing, 2009.

[3] Tomas Kliegr, Krishna Chandramouli, Jan Nemrava, Vojtech Svatek and Ebroul Izquierdo, "Combining Captions and Visual Analysis for Image Concept Classification", In MDM/KDD'08: Proceedings of the 9h International Workshop on Multimedia Data Mining. ACM, 2008.

[4] Gauvain, J.-L., Lamel, L. and Adda, G., The LIMSI broadcast news transcription system. Speech Communication 37,89-108,2002.