# Genre tagging of videos based on information retrieval and semantic similarity using WordNet *

José M. Perea-Ortega, Arturo Montejo-Ráez, Manuel C. Díaz-Galiano and M. Teresa Martín-Valdivia
SINAI Research Group, Computer Science Department
University of Jaén
23071 - Jaén, Spain
{jmperea,amontejo,mcdiaz,maite}@ujaen.es

## ABSTRACT

In this paper we propose a new approach for the genre tagging task of videos, using only their ASR transcripts and associated metadata. This new approach is based on calculating the semantic similarity between the nouns detected in the video transcripts and a bag of nouns generated from WordNet, for each category proposed to classify the videos. Specifically, we have used the Lin measure based on WordNet, which calculates the semantic distance between two *synsets*. Obviously, this approach has been only applied on the English test videos due to the use of WordNet, an English lexical resource. As base case, we have applied an information retrieval system as a classifier, using the generated bag of nouns for each category as index data and the ASR transcripts from each test video as query. Several experiments have been submitted, one of them combining both approaches (information retrieval and semantic similarity). As main conclusion we have shown that, using this combination of semantic similarity and information retrieval, we can improve the results obtained using the information retrieval approach only.

## Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing - Indexing methods

## Keywords

Genre video tagging, Video categorization, Automatic Speech Recognition

## 1. MOTIVATION AND RELATED WORK

Multimedia data are usually tagged with some relevant information in order to make the retrieval easier. In fact, the efficient use of textual data associated to other types of information such as images can improve multimedia IR systems [1, 4]. However, the provided labelling of multimedia videos may not contain sufficient context for locating data of interest in a large database. Detailed annotation is required, so that users could quickly locate clips of interest without having to go through entire databases.

The Genre Tagging task in MediaEval 2011 attempts to automatically generate genre labels to organize videos [2]. In this paper we present some experiments on automatic genre tagging of videos making use of their Automatic Speech Recognition (ASR) transcripts and metadata associated. We have worked during last years in the field of video categorization, participating in VideoCLEF [6, 7] and MediaEval 2010 [5].

## 2. DESCRIPTION OF THE TASK

In the Genre Tagging task of MediaEval 2011, participants are required to automatically assign thematic subject labels to videos using features derived from speech, metadata, audio or visual content. It is important to note that this is not a multilabel tagging task, so a given video can only be assigned to one label. The data set provided are the same as those used in MediaEval 2010 Wild Wild Web Tagging Task (ME10WWW) [2]. The development and test data sets consisted of 247 and 1,727 videos respectively. From the test videos, 1,673 videos are in English, 16 are in French, 25 are in Spanish and 13 are in Dutch. We have only worked with the English videos. The list of genre classes consisted of 25 tags, providing a "*default category*" for those videos that do not fit in any other classes.

## 3. SYSTEM OVERVIEW

Our main approach is based on using an Information Retrieval (IR) system as a classifier. On the one hand, we have generated a XML document or *bag of words* for each category proposed, making use of an external lexical resource like WordNet[1]. Specifically, we have included *synonyms*, *hyponyms* and *domain terms* related to the category. For example, for the "*educational*" category we have generated a XML document including terms such as *instruction*, *teaching*, *pedagogy*, *didactics*, *training*, etc. On the other hand, the preprocessed ASR transcripts (stemming and stop word removal) from test videos have been used as queries, without any expansion. Finally, the Terrier[2] IR system has been used to obtain a measure of relatedness (RSV, Retrieval Status Value) between each video and the generated bags of words.

*This work has been partially supported by a grant from the Fondo Europeo de Desarrollo Regional (FEDER), project TEXT-COOL 2.0 (TIN2009-13391-C04-02) from the Spanish Government, a grant from the Andalusian Government, project GeOasis (P08-TIC-41999) and Geocaching Urbano research project (RFC/IEG2010).


*MediaEval 2011 Workshop,* September 1-2, 2011, Pisa, Italy

[1] http://wordnet.princeton.edu
[2] http://terrier.org

As a second approach, we have used the formula proposed by Lin [3], which is based on WordNet, to measure the semantic similarity between the nouns detected in each test video and the bags of words generated for each category. Firstly, for calculating the semantic similarity between a video and the XML document generated for a category, we have obtained the Lin semantic similarity between each pair of nouns from both, accumulating those similarity scores that exceed a threshold set at 0.75. With the use of this threshold, we have tried to minimize the effect of the size of the ASR transcripts, since some of the videos contain more words than others. Secondly, the accumulated similarity score has been divided by the number of words detected in the video, obtaining the final semantic similarity score. Those videos with a final semantic similarity score of less than 0.25 were considered in the *default category*.

# 4. EVALUATION OF RESULTS

Several experiments were carried out under both approaches. As baseline, we have considered the use of the preprocessed ASR transcripts from test videos as query (experiment *IR_ASR*). Then, we have tried to evaluate the addition of the metadata provided, carrying out an expansion of the ASR transcripts using such metadata (experiments *IR_ASR+MD* and *IR_ASR+MD+TAGS*). Regarding the second approach, we have submitted the experiment "*SIMSEM-ASR*", in which we only calculate the semantic similarity between each video and each category (its bag of words), without using the IR approach. Finally, we have combined the IR and the semantic similarity approaches (experiment "*SIMSEM+IR-ASR*"), merging both lists of results. First, we have normalized the RSV score from the baseline. Then, for each test video, we have added their normalized RSV and semantic similarity scores. The results obtained are shown in Table 1, using the *Mean Average Precision* (MAP) measure. We also show the MAP obtained considering only the English test videos.

| **Run name** | $MAP_{official}$ | $MAP_{English}$ |
|---|---|---|
| *IR_ASR* | 0.1031 | 0.1044 |
| *IR_ASR+MD* | 0.1073 | 0.1088 |
| *IR_ASR+MD+TAGS* | 0.1115 | 0.1129 |
| *SIMSEM-ASR* | 0.0547 | 0.0559 |
| *SIMSEM+IR-ASR* | **0.1266** | **0.1288** |

**Table 1: Experiments and results obtained by SINAI in the MediaEval 2011 Genre Tagging task**

Analyzing the official results we can observe that the expansion of the ASR transcripts using the provided metadata improves the result obtained when metadata is not used (+4% and +8.15% better for the experiments *IR_ASR+MD* and *IR_ASR+MD+TAGS*, respectively), as it was expected. On the other hand, the combination of the semantic similarity and the IR approaches seems to be interesting because it improves the MAP value obtained for the baseline using the IR approach only (+22.79%). Taking into account the test *groundtruth* file provided by the MediaEval organizers, 185 videos of the 1,673 English videos (11.06%) belong to the *default* category, while our best experiment assigned only 18 videos (1.08%) to such category. This was motivated by the low threshold used to assign a video to the *default* category (0.25), which allowed to classify videos in categories that really did not correspond due to its low similarity score. Nevertheless, for some categories (*art*, *politics*, *religion* and *sports*), we obtained good results, achieving high MAP individual scores (e.g. 0.6176 for the *politics* category). This is due to such categories are more general concepts or genres than others (*business*, *comedy*, *documentary*, etc.), so it was easier to find more nouns semantically related, increasing the size of the XML document generated for such categories and, therefore, the probability of success.

# 5. CONCLUSIONS

In this paper we propose the use of the semantic similarity based on WordNet combined with the IR approach in order to solve the genre tagging task of videos. Because our research field of interest is Natural Language Processing (NLP), we have only worked with the ASR transcripts from videos and their metadata. It was shown that combining the semantic similarity score with the RSV score obtained from the IR approach, we obtained a significant improvement. Nevertheless, it seems clear that working only with the ASR transcripts generally get poor results. For future work, we will study other resources in order to increase the size of the bag of words generated for each category, adding more terms semantically related with such categories.

# 6. REFERENCES


[1] Bozzon, A., and Fraternali, P. Multimedia and multimodal information retrieval. In *SeCO Workshop* (2009), S. Ceri and M. Brambilla, Eds., vol. 5950 of *Lecture Notes in Computer Science*, Springer, pp. 135–155.

[2] Larson, M., Eskevich, M., Ordelman, R., Kofler, C., Schmeideke, S., and Jones, G. Overview of MediaEval 2011 Rich Speech Retrieval Task and Genre Tagging Task. In *MediaEval 2011 Workshop* (Pisa, Italy, September 1-2 2011).

[3] Lin, D. An information-theoretic definition of similarity. In *Proc. of the 15th Int'l. Conf. on Machine Learning* (1998), pp. 296–304.

[4] Martín-Valdivia, M. T., Díaz-Galiano, M. C., Montejo-Ráez, A., and Ureña-López, L. A. Using Information Gain to Improve Multimodal Information Retrieval Systems. *Information Processing & Management 44* (2008), 1146–1158.

[5] Perea-Ortega, J. M., Montejo-Ráez, A., Díaz-Galiano, M. C., and Martín-Valdivia, M. T. SINAI at Tagging Task Professional in MediaEval 2010. In *Working Notes Proceedings of the MediaEval 2010 Workshop, Pisa, Italy, October 24, 2010* (2010).

[6] Perea-Ortega, J. M., Montejo-Ráez, A., Díaz-Galiano, M. C., Martín-Valdivia, M. T., and Ureña-López, L. A. Using an information retrieval system for video classification. In *Evaluating Systems for Multilingual and Multimodal Information Access* (2009), vol. 5706 of *Lecture Notes in Computer Science*, Springer, pp. 927–930.

[7] Perea-Ortega, J. M., Montejo-Ráez, A., Martín-Valdivia, M. T., and Ureña-López, L. A. Using support vector machines as learning algorithm for video categorization. In *CLEF, Part II* (2010), vol. 6242 of *Lecture Notes in Computer Science*, Springer, In Press.