MIRACLE at VideoCLEF 2008:
Classification of Multilingual Speech Transcripts

Julio Villena-Román[1,3], Sara Lana-Serrano[2,3]

[1] Universidad Carlos III de Madrid
[2] Universidad Politécnica de Madrid
[3] DAEDALUS - Data, Decisions and Language, S.A.

jvillena@it.uc3m.es, slana@diatel.upm.es

Abstract

This paper describes the participation of MIRACLE research consortium at the VideoCLEF track at CLEF 2008. We took part in both the main mandatory Classification task that consists in classifying videos of television episodes using speech transcripts and metadata, and the Keyframe Extraction task, whose objective is to select keyframes that represent individual episodes from a set of supplied keyframes (one from each shot of the video source). For the first task, our system is composed of two main blocks, the first in charge of building the core system knowledge base, and then the set of operational elements that are needed to classify the speech transcripts of the topic episodes and generate the output in RSS format. For the second task, our approach is based on the assumption that the most representative fragment (shot) of each episode is the one whose distance to the whole episode is the lowest, considering a vector space model. 4 runs were submitted in all. Regarding the classification task, we ranked 3rd (out of 6 participants) in terms of precision and 2nd in terms of recall.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.2 Information Storage; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital libraries. **H.2 [Database Management]**: H.2.5 Heterogeneous Databases; **E.2 [Data Storage Representations]**.

Keywords

Video retrieval, domain-specific vocabulary, thesaurus, linguistic engineering, information retrieval, indexing, relevance feedback, multilingual speech transcripts. VideoCLEF, VID2RSS, CLEF, 2008.

1. Introduction

MIRACLE team is a research consortium formed by research groups of three different universities in Madrid (Universidad Politécnica de Madrid, Universidad Autónoma de Madrid and Universidad Carlos III de Madrid) along with DAEDALUS, a small/medium size enterprise (SME) founded in 1998 as a spin-off of two of these groups and a leading company in the field of linguistic technologies in Spain. MIRACLE has taken part in CLEF since 2003 in many different tracks and tasks, including the main bilingual, monolingual and cross lingual tasks as well as in ImageCLEF [4] [8], Question Answering, WebCLEF, GeoCLEF and VideoCLEF tracks.

This paper describes our participation in the VideoCLEF task, a new track for CLEF 2008. The goal of this track is to develop and evaluate tasks in processing video content in a multilingual environment. This track for 2008 is dedicated to Vid2RSS task which comprises a number of subtasks including topic classification performed on dual language videos. The main objective involves assigning topic class labels to videos of television episodes. Speech recognition transcripts, metadata records (containing title and description) and video keyframes (and shot boundaries) for each episode are supplied. The video data are Dutch television documentaries and contain Dutch as a dominant language, but also contain a high proportion of spoken English (i.e., interview guests often speak in English). The output format is a set of RSS-feeds, one for each topic class, created by concatenating the metadata records for the episodes assigned to a given topic class.

We have participated in the main mandatory Classification task that consists in classifying videos of television episodes using speech transcripts and metadata, and in the Keyframe Extraction task, whose objective is to select keyframes that represent individual episodes from a set of supplied keyframes (one from each shot of the video source).

2. Classification task

The objective of the mandatory Classification task is to perform the speech recognition transcript-based topic classification (i.e., classify the videos of the television documentary episodes using the speech recognition output only). Videos include dual language (Dutch + English) episodes. Output is 10 topic-based feeds, each containing the episodes that have been classified in that topic category. The defined topic classes are Archeology, Architecture, Chemistry, Dance, Film, History, Music, Paintings, Scientific research and Visual arts. While classification task I is based on episode speech transcripts only, classification task II allows to use episode metadata for classification.

Figure 1 shows the logical architecture of our system. The system is composed of two main blocks. The first block is in charge of building a corpus that can be used as the core system knowledge base. The second block includes the set of operational elements that are needed to classify the speech transcripts of the topic episodes and generate the output in RSS format. Those operational elements cover an information retrieval system and a classifier, as well as auxiliary modules for text extraction, filtering and RSS generation.

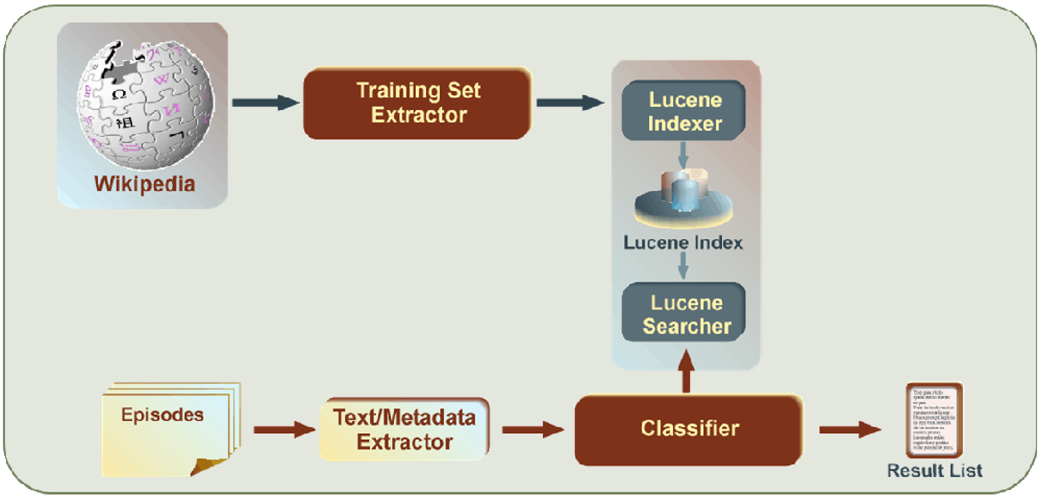

Figure 1. Overview of the classification system

The first step is to obtain the necessary training data for the classifier, as part of the task is, given the description of the subject class, to gather the necessary data to train the classifier. Our knowledge base for training the classifier was generated from Wikipedia articles. In order to do so, we first established a matching between the topic classes provided for the task and the classification topics that Wikipedia uses for articles, encoded in metadata. The next step was to obtain a list of Wikipedia articles belonging to each of the 10 topic class, for each task language, i.e. English and Dutch. Table 1 shows the number of articles for each topic class and language.

Table 1. Articles in the training set

Topic	EN	NL
Archaeology	666	1280
Architecture	1409	2002
Chemistry	2207	3130
Dance	497	3248
Film	338	4934
History	1655	7822
Music	506	8691
Paintings	612	213
Scientific research	2845	15307
Visual arts	754	1280

Next, each document is processed through the following sequence of operations:

1. **Text extraction:** Ad-hoc scripts are run to obtain the actual text content of articles, filtering out Wikipedia tags.
2. **Diacritics removal and conversion to lowercase:** all terms are normalized by removing diacritics (in the case of Dutch) and changing all letters to lowercase.
3. **Filtering:** All words recognized as stopwords are filtered out. Stopwords in the two target languages were initially obtained from [6] and afterwards extended using several other sources [3] as well as our own knowledge and resources.
4. **Stemming:** This process is applied to each one of the terms to be indexed or used for retrieval. Standard stemmers from Porter [5] have been used.

The processed corpus is indexed with Lucene retrieval engine [1] to allow a fast and efficient access to information needed for the classification. Two different indexes are built, one for each language.

The classifier is based on the k-Nearest Neighbour algorithm [7]. To find the class for a given episode, the content is first processed as explained before. Then the whole set of resulting terms is used to build a query that is given to the Lucene search engine to obtain the list of the top k most relevant (i.e., most similar) articles in the Wikipedia-based corpus. Finally, the class of the given episode is the most frequent class in the top k results. After some preliminary experiments, a value of k set to 10 was chosen for our runs.

3. Keyframe extraction task

Our system is based on the assumption that, in the context of a vector space model representation [2], the most representative fragment (shot) of each episode (represented by a vector) is the one whose distance to the whole episode (also a vector) is the lowest. The contents of both each shot and the whole episode are first processed as explained before, after extracting the text from the speech transcription. Based on the vector space model, a weighted vector is built for each episode and set of shots, representing the term frequency of the main most significant terms in the given episode. Finally the keyframe extraction module selects the keyframe belonging to the most representative shot in the episode, which is the shot whose vector has the lowest distance from (i.e., is nearest) the vector of the whole episode. The metric used here was the cosine distance [2], although Euclidean distance could also be valid. An overview of the system architecture is shown in Figure 2.

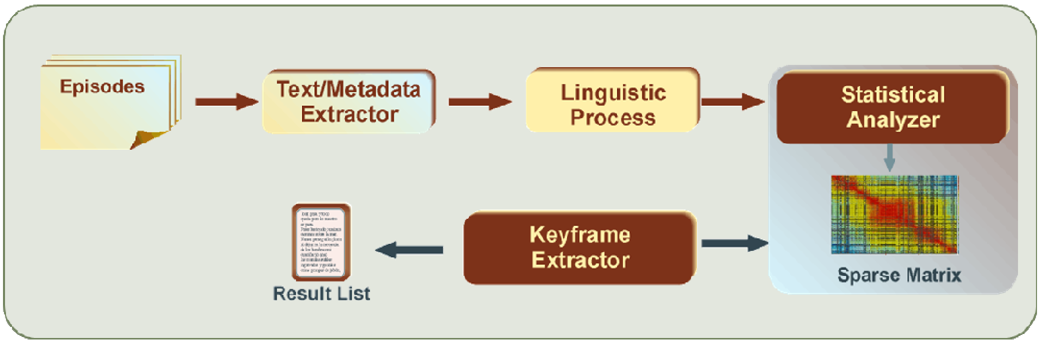

Figure 2. Overview of the keyframe extraction system

4. Experiments and results

We have submitted different runs for each proposed subtask: three for the classification task and one for the keyframe extraction. Table 2 shows the list of submitted runs.

Table 2. List of runs

Run Identifier	Language	Task
MIRACLE-CNL	NL	Classification I + Keyframe_Extraction I
MIRACLE-CNLEN	NL+EN	Classification I
MIRACLE-CNLMeta	NL	Classification II

In short, “CNL” run only uses the index for the Dutch corpus, “CNLEN” uses both indexes and gathers together results from any of them, and “CNLMeta” uses the Dutch index but also includes the episode metadata to build the query for the retrieval engine.

Table 3 shows the values of precision and recall obtained by the different runs in the classification task. Classes marked in italics are those whose number of relevant documents is equal to 0, i.e., no episode belongs to this class, according to the task organizers.

It can be observed that the best precision is achieved with the “CNL” run in which only the Dutch transcription is used. When the knowledge base and the transcription in English are involved, results are noticeably and significantly worse. This could be directly motivated by the fact that the dominant language of the episodes is Dutch. However, the best modelled class is “Music”, corresponding to one of the classes that own a higher number of Wikipedia articles in the training set. This may suggest other explanations, such as the fact that training set (the knowledge base) for English is much smaller than the one available for Dutch. Another possible explanation could be that the voice recognition system for English is not as good as for Dutch. Obviously, these issues have to be further studied.

Table 3. Classification task results

	Precision			**Recall**		
	CNL	**CNLEN**	**CNLMeta**	**CNL**	**CNLEN**	**CNLMeta**
Archaeology	0.25	0.25	0.40	0.14	0.14	0.29
Architecture	*0.00*	*0.00*	*0.00*	*1.00*	*1.00*	*1.00*
Chemistry	*0.00*	*0.00*	*0.00*	*1.00*	*1.00*	*1.00*
Dance	0.00	0.00	0.13	0.00	0.00	0.67
Film	1.00	0.25	1.00	0.00	0.33	0.00
History	0.25	0.26	0.38	0.30	0.50	0.60
Music	**0.64**	**0.65**	**0.65**	**0.95**	**1.00**	**1.00**
Paintings	1.00	0.00	1.00	0.00	0.00	0.00
Scientific Research	0.29	0.21	0.21	1.00	1.00	1.00
Visual Arts	1.00	0.20	0.14	0.00	0.40	0.20
ALL (microaveraged[1])	**0.43**	0.29	0.37	0.51	0.61	**0.65**
ALL (macroaveraged[1])	**0.44**	0.18	0.39	0.44	0.54	**0.58**

[1] By definition, macroaverage values are computed as the mean of values of all classes that are first individually calculated. In contrast, the microaverage measures first obtain the aggregate counts for all classes and then calculate the values of precision and recall.

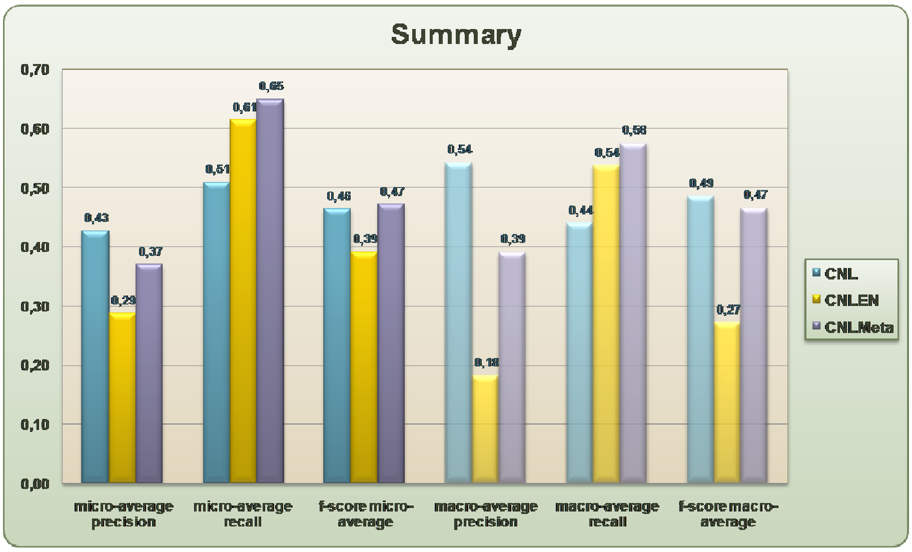

Figure 3. Summary of classification task results

Although results for the classification task II, i.e. including metadata, seem to be worse than results for classification task I, this is misleading. For all classes except Scientific research and Visual arts, precision values are higher if metadata is used, but the average value is worse specially due to the very low value for Visual arts. This issue still has to be analyzed.

Comparing to other groups, we successfully ranked 3rd out of 6 participants in terms of precision, 2nd in terms of recall and also f-score (not shown in the table).

Regarding the keyframe extraction task, MIRACLE was the only participant who submitted results. Thus, the evaluation has been manually made. Five native Dutch speakers (in the age range 20-50) were presented with the title and the description of each video episode along with two keyframes, one manually extracted and one automatically extracted provided by us. They were asked to choose which keyframe they preferred. Of the 40 videos, 1 did not have a keyframe. In two cases, the keyframe chosen manually and that chosen by the system was the same. Thus, each subject was asked about 37 different pairs of keyframes.

On average, the subjects chose the automatic over the manually selected keyframe in 15.2 cases (41.08%) and the manually over the automatic in 21.8 cases (58.92%). Table 4 shows the results of the evaluation process. The "Automatic keyframe" and "Manual keyframe" column represent the number of episodes that the evaluator has selected as more adequate between the automatically or manually extracted keyframe. The last column tries to show the number of correctly extracted keyframes, assuming that the evaluator has chosen correctly.

Table 4. Keyframe extraction task results

	Automatic keyframe	**Manual Keyframe**	**Well selected?**
Subject 1	16	21	18
Subject 2	14	23	16
Subject 3	16	21	18
Subject 4	17	20	19
Subject 5	13	24	15
Average	**41.08%**	**58.92%**	**44.10%**

These promising figures indicate that the automatically extracted keyframes may be strong competitors with the manual ones in the short- or middle-term future.

5. Conclusions and Future Work

After a preliminary analysis of results obtained in the classification task, we can conclude that there seems to be a direct relationship between the knowledge base associated to a given class and results achieved in it. This probably indicates that the architecture of the system and provided algorithms are useful, but more effort must be invested to improve the knowledge base, both in its volume (coverage) and the pre-processing activities.

Despite the subjectivity of the keyframe extraction task and lack of any reference experiment to which compare our own system, we can say that these results are promising and encourage us to keep on this line of research for future participations.

Acknowledgements

This work has been partially supported by the Spanish R+D National Plan, by means of the project BRAVO (Multilingual and Multimodal Answers Advanced Search – Information Retrieval), TIN2007-67407-C03-03 and by Madrid R+D Regional Plan, by means of the project MAVIR (Enhancing the Access and the Visibility of Networked Multilingual Information for the Community of Madrid), S-0505/TIC/000267.

References

[1] Apache Lucene project. On line http://lucene.apache.org [Visited 10/08/2008].

[2] Baeza-Yates, R., Ribeiro-Prieto B.: Modern Information Retrieval. Addison Wesley (1999).

[3] CLEF 2005 Multilingual Information Retrieval resources page. On line http://www.computing.dcu.ie/~gjones/CLEF2005/Multi-8/ [Visited 10/08/2008].

[4] Martínez-Fernández, J.L.; Villena-Román, Julio; García-Serrano, Ana M.; González-Cristóbal, José Carlos. Combining Textual and Visual Features for Image Retrieval. Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, Revised Selected Papers. Carol Peters et al (Eds.). Lecture Notes in Computer Science, Vol. 4022, 2006. ISSN: 0302-9743.

[5] Porter, Martin. Snowball stemmers and resources page. On line http://www.snowball.tartarus.org [Visited 10/08/2008].

[6] University of Neuchatel. Page of resources for CLEF (Stopwords, transliteration, stemmers …). On line http://www.unine.ch/info/clef [Visited 10/08/2008].

[7] Witten, Ian H.; Frank, Eibe. Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[8] Villena-Román, Julio; Lana-Serrano, Sara; Martínez-Fernández, José Luis; González-Cristóbal, José Carlos. MIRACLE at ImageCLEFphoto 2007: Evaluation of Merging Strategies for Multilingual and Multimedia Information Retrieval. Working Notes of the 2007 CLEF Workshop, Budapest, Hungary, September 2007.