Video shot classification using lexical context

Stéphane Ayache, Georges Quénot, Mbarek Charhad

Introduction

Associating concepts to video segments is essential for content-based video retrieval. We present here a semantic classifier working from text transcriptions coming from automatic speech recognition (ASR). The system is based on a Bayesian classifier, it is fully linked with a knowledge base which contains an ontology and named entities from several domains. The system is trained from a set of positive and negative examples for each indexed concept. It has been evaluated using the TREC VIDEO protocol and conditions for the detection of visual concepts. Three versions are compared: a baseline one, using only word as units, a second, using additionally named entities, and a last one enriched with semantic classes information.

Associating visual concepts in video by lexical analysis

Detection of visual concepts in video documents is usually achieved by categorizing key images from signal information. These approaches use low-level extraction processes for color, texture and motion features and a supervised learning phase such as KNN, SVM, or NN methods. The audio stream can also help to identify concepts. The approach studied here is based on the idea that a lexical context (word distribution) is associated to the presence of a concept in a video document. This kind of approach has been used with success for emotion detection in oral dialogues [2]. We have thus developed and experimented a classifier based on a lexical analysis of transcripted speech. Since our approach is supervised, we must train the system for each concept. This is done by the following 3 steps:

- Extract text from ASR around instances of concepts: in order to catch lexical context of concepts, we define temporal offsets around the shots containing a concept. We choose offsets for each concept by computing cross validation in the development data.
- Textual analysis: the simplest way to analyze textual information is to extract every 1-gram terms. This approach is our baseline model for textual analysis. Experiments compare models enriched using a knowledge base with this approach.
- Compute the probability p_{we} of each term w being in the class e.

Learning a semantic class by lexical analysis aim to perform a co-occurrence-like process between semantic and lexical information. In this way, the following lines are the top 6 entries of the "Madeleine Albright" model:

0.029062	State	0.022457	U.S.
0.027741	Secretary	0.013210	Iraq
0.026420	Albright	0.013210	Madeleine

Table 1: Top 6 entries from basic "Madeleine Albright" model

During the detection process, the system assigns a score value V_{se} for each shots s being in semantic class e according to a matching function, for example:

$$V_{se} = \frac{1}{L} \sum_{w} \frac{P(w|e)}{P(w)} \tag{1}$$

where $P(w)$ is the probability of w being in the general model computed on a development set), and $P(w|E)$ is the probability of having w known a model e. L is the length of the filtered text.

Toward a lexical and semantic analysis

Based on the idea that semantic information can enhance the detection process, we simply merge lexical entities and ontology leaves or nodes. With this approach, the textual extraction process aims to tag the text using our specific knowledge base by finding named-entities, class information, or applying stemming and stop-lists. We also define a set of entities referring the same concept or very closed entities, such as train and locomotive.

Ontology design

We are interested in named entities with the point of view of Information Retrieval. Named entities can improve topic classification and text desambiguisation. Thus, we designed a named entities extraction tool based on a domain specific ontology and patterns to identify persons, locations, acronyms etc. The ontology contains about 10000 instances of concepts organised in three specific classes : people (with activities), geography (with continents) and organization (full names or acronyms). This choice is justified by the kind of video document that we use as corpora (TRECVID 2003 and 2004 collections contain TV broadcast news).

Named entities enriched model

We enrich the basic class model by adding named entities, which have been extracted from text data. Thus, a model is not only defined by 1-gram terms, but also by N-grams terms, like: "Madeleine Albright" and "Secretary of State".

Semantic class label enriched model

Semantic classes label are node names (not leaves) of the ontology, such as "European-Politics", "Middle-East-Places", "Actors", "Football Players", etc. Since semantic classes are specialized enough and obviously domain dependent, we expect them to improve the accuracy of the classifier. Thus, we construct a

semantic class label model by adding node names probabilities of the extracted named entities. Also, in order to evaluate this approach, we build a model containing only node names entries.

Experiments and results

We have experimented our classifier using the TREC VIDEO corpus and protocol. Learning and tuning was done using the TRECVID 2003 collection and the evaluation was done using the TRECVID 2004 collection. Lexical context based classification was performed using the LIMSI ASR transcription [1]. In order to our approach, we computed the classification with just 1-gram terms (Baseline), enriched models with named entities, enriched models with semantic classes information only, and enriched models with named entities and classes information. We experiment on TRECVID 2004 corpus the first six high-level features. Table 2 shows a comparison of our 4 runs.

Feature	Baseline	Named Entity	Class Info	NE + CI
Ship/Boat	0.0024	0.0611	0.0013	0.0563
Madeleine Albright	0.0338	0.0702	0.0192	0.0715
Bill Clinton	0.1082	0.1144	0.0687	0.1200
Train	0.0613	0.2029	0.00	0.1530
Beach	0.0024	0.0145	0.0006	0.0139
Basket scored	0.0436	0.0548	0.0202	0.0353

Table 2: Mean Average Precision on TREC VID 2004 high level features

The named entities enriched model approach perform globally better than the baseline approach, since the semantic feature appears in a well established context. For instance, we saw on the development set that trains appears frequently in broadcast news to report a train accident. Such events are well modeled by enriched lexical analysis. However, classes label don't contribute to a good accuracy. Since our domain specific ontology is not sufficiently rich of information (there are few node names), it can't enhance accuracy of classification. The global performance is quite low but the searched concepts are quite difficult and the approach considered here uses information only from tha audio track. It could be fused with other approaches using information from the image track.

Acknowledgments

This work is supported by the Peng project.

References

[1] L.Lamel J.L.Gauvain and G.Adda. *The LIMSI broadcast news transcription system.* Speech Communication, 37(1-2):89-108, 2002.

[2] I.Vasilescu L.Devillers. *Détection des émotions à partir dindices lexicaux, dialogiques et prosodiques dans le dialogue oral.* TALN-JEP, Maroc, 2004.