The University of Amsterdam at VideoCLEF 2008

Jiyin He Xu Zhang Wouter Weerkamp Martha Larson

{j.he,x.zhang,w.weerkamp,m.a.larson}@uva.nl
ISLA, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam

Abstract

The University of Amsterdam (UAms) team carried out the Vid2RSS classification task, the primary sub-task of the VideoCLEF track at CLEF 2008. This task involves the assignment of thematic category labels to dual language (Dutch/English) television episode videos. UAms chose to focus on exploiting archival metadata and speech transcripts generated by both the Dutch and English speech recognizers. Exploratory experimentation completed prior to the start of the task on external data motivated choosing a Support Vector Machine (SVM) with a linear kernel as the classifier. As a SVM toolbox to carry out the experiments, the Least Square-SVM (LS-SVM) toolbox was selected. Wikipedia was chosen as the source of the training data because it is multilingual and contains content with broad thematic coverage. The results of the experimentation showed that archival metadata improves performance of classification, but the addition of speech recognition transcripts in one or both languages does not yield performance gains. Although the overall performance of the classifiers was less than satisfactory, adequate performance was achieved in several classes, suggesting that there is concrete potential for future work to achieve performance improvements, especially if more suitable training data could be obtained.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries

General Terms

Measurement, Performance, Experimentation

Keywords

Video classification, SVM, speech recognition

1 Introduction

In this paper we describe our participation in the VideoCLEF track of the Cross Language Evaluation Forum, CLEF.[1] The University of Amsterdam (UAms) organized the VideoCLEF 2008 track together with Dublin City University and also took part in the track as a task participant. The goal of the University of Amsterdam's involvement in the VideoCLEF track is to further our understanding of the problem of providing intelligent access to large audio-visual archives and to spoken audio content on the internet. We aim to continue refinement of our techniques for classification and retrieval of conversational speech in

[1] http://www.clef-campaign.org

a multilingual setting. Such technologies are being developed within the framework of the MultiMatch project [1]. Development of and participation in benchmark tests that help to coordinate research progress across sites are an important part of this framework.

In this year's VideoCLEF track, the task, called Vid2RSS,[2] is classification performed on a video corpus containing episodes of dual language television programs. The video corpus was provided by the audio/video archive Netherlands Institute of Sound and Vision,[3] in Dutch called *Beeld & Geluid*. The videos are predominantly documentaries. Dutch is the main language, which we call the *matrix* language, and English, which we call the *embedded* language, is spoken during interviews. The classification task is formulated as follows: given a number of thematic categories (Archeology, Architecture, Chemistry, Dance, Film, History, Music, Paintings, Scientific research and Visual arts), the participants should assign the right category label or labels to each video. For each category, an topic-based RSS-feed is then created. A topic-based feed consists of a concatenation of XML items that represent the videos that were assigned a particular thematic category label. The feed item for a video contains its title and description drawn from the archival metadata, as well as a hand-picked keyframe. The generation of the RSS-feed is trivial and the challenge of the task lies with the classification. The data set includes speech recognition transcripts and archival metadata for each video. Two sets of speech transcripts are provided, one generated by a Dutch speech recognizer and one by an English speech recognizer. Each recognizer transcribed the entire spoken content of each video: no separation of content by language was performed. The archival metadata is Dutch only. A final characteristic of the task is the lack of training data: participants should develop methods of collecting their own training data for training a classifier.

In the next section we address the dimensions the Vid2RSS task offers us. Section 3 provides details on the training data we use and our choice of classifier. These decisions are both based on initial exploratory experiments. Section 4 describes our experimental set up. Sections 5.1 and 5.2 report on the results of our approach on different dimensions, and finally we draw some preliminary conclusions and present discussion in Section 6.

2 Task dimensions

Given the task at hand, classifying dual language video based on speech recognition transcripts and metadata, without training data provided, we identify four dimensions that are worth exploring. These dimensions are discussed in this section. Since the number of runs in CLEF is limited (to five), we need to chose a highly restricted number of dimensions for our experiments. As reported below, we ultimately arrive at the decision to explore only the dimensions *metadata* and *language*.

First we take a step back and look at the four dimensions we identify: (i) classifier, (ii) training data, (iii) metadata, and (iv) language. The classifier dimension focuses on the choice of classifier and, the even more detailed decision of the parameter settings of each classifier. This dimension can be expected to influence the experiments. We did some limited exploratory tests to support ourselves in making our choice, but in the end we decided that this is not the most interesting aspect of the task and we fix this dimension. Training data refers to the different sources we can use as training data in this task, how to collect the data, and what the performance is of different sources. Although this dimension will certainly have a high impact on performance, we choose to fix this dimension. We note that in order to have an exact match between the training data and the test data for this task, it would be necessary to be able to collect training data that consists of speech recognition transcripts from Dutch television documentaries. Since such data is not available in large quantities, we are faced with the situation that the match of test and training data will be necessarily approximative. Instead of exploring the variation due to different data sources, which we expect to be wide, we chose to fix the source of the training data and to devote our Vid2RSS runs to exploring other issues. In Section 3, we discuss our choices on both the classifier and the training data dimension.

The two remaining dimensions are the main focus of our participation this year: metadata and language. Exploring the metadata dimension involves attempting to understand the impact of using the archival metadata associated with the video as a source of features for classification. The rationale behind choosing this

[2] http://ilps.science.UvA.nl/Vid2RSS

[3] http://www.beeldengeluid.nl

dimension is quite straightforward. Archival metadata is often available with television programs and, if available, it should be exploited in order to perform classification. Unlike speech recognition transcripts, metadata is very clean, nearly error-free, nicely structured text. Archival metadata is created by archivists with the goal of annotating video with high-level semantics and making video retrievable from the archive. For this reason, the terms contained in archival metadata can be expected to be meaning bearing and reflect the thematic content of the video. In contrast, the terms contained in the speech recognition transcripts can be expected to be characteristic of the interviews and discussions contained in the television programs. Although terms that reflect video topic are with out doubt present, speech transcripts are diluted with the kinds of vocabulary typical of conversational speech, namely reflecting social convention, expressing feelings and opinions and drawing connections between entities and concepts not always explicitly mentioned. We expect that using metadata in the classification process will increase performance. The second dimension, language, refers to the fact that the the videos contain two languages and that each video is accompanied by both a Dutch and a English speech transcript. As task participants, we can chose which transcript to use. As mentioned before, we distinguish between the *matrix* language (the underlying, main language of the program) and the *embedded* language(s). In our case, Dutch is the matrix language, used to introduce the foreign speakers and to glue the program together, and English is the embedded language. Using an embedded language might introduce more noise, but it could also add specific information regarding the topic that is not present in the matrix language transcript. We expect that using both matrix and embedded language in classifying videos will lead to better results than using only one (matrix) language.

Based on the choices described in this section we end up with five runs based on the different combinations given the two dimensions. Table 1 lists these five runs and their division over the dimensions.

Table 1: Submitted runs and their division over two dimensions.

Run	**Dutch transcripts**	**English transcripts**	**metadata**
uams08m			X
uams08asrd	X		
uams08masrd	X		X
uams08asrde	X	X	
uams08masrde	X	X	X

3 Fixed dimensions

Although we are not exploring the impact of the two previously mentioned dimensions, the classifier and the training data, we do need to decide on how to fix these dimensions for our experiments. In order to arrive at a basic intuition about the performance of different options, we use the development set made available to the participants for exploratory experimentation. This development set only contains ten videos and results on this set can therefore only be seen as indications, and do not allow us to draw strong conclusions. Below, we first discuss the classifier, followed by the training data.

3.1 Choice of classifier

As mentioned before, we set the choice of classifier as one of the fixed dimensions, we stick to a Support Vector Machine (SVM) classifier for all our runs. We chose SVM since it has been reported perform well in general on text classification problems [3, 2] and this performance has been demonstrated to transfer to speech recognition transcripts that contain a significant level of noise [5]. The results of exploratory experimentation suggested that the correct choice of a kernel was a linear kernel. These results were consistent with previous work [5]. Since our training data and test data are from very different resources, the generalization ability is more important than fitting the training data very well. We believe that this mismatch contributes to the fact that the linear kernel out-performed other kernel choices during the exploratory experimentation phase.

3.2 Training data

Because no training data is supplied with the Vid2RSS task, we are left with the problem of selecting appropriate data. Four considerations play a role in selecting training data: (i) data should be thematically similar to the test data, i.e., treat the same topics (ii) data should be similar in style to the test data, i.e., reflect the particular mix of planned and conversational speech characteristic of the test collection (ii) data should be dual language, containing the same proportions of Dutch and English that are present in the test data (iv) data should be available. Ideally, we would construct training data from speech transcripts of television documentary videos on the right topics, with the right mixture of Dutch and English; such a data set is not available, however. Note that we do not mention the issue of speech recognition errors in our list of considerations. Previous work has shown that classifiers that are used to classify speech transcripts (high error rates) can be trained on appropriate born text data (vanishingly low error rates) [5]. For this reason, we assumed that text data would appropriately serve our needs for classifier training. In order to chose training data, we decided to ease the first two constraints, opening up more options of sources for training data, e.g., news collections, blogs, and encyclopedic sources.

The collection that seems most suitable is online encyclopedia Wikipedia.[4] The data from this source is freely available, contains material treating the right topics, and is available in multiple languages (including Dutch and English). Additionally, Wikipedia offers high quality content and and a structure (i.e. categories) that makes it possible to automatically extract articles in particular thematic categories. These two additional characteristics make Wikipedia an interesting option for collecting training data. The main disadvantage of using Wikipedia is its dissimilarity to the test data: Wikipedia content is created to be read and is much more structured and uses a language style more formal than that of the conversational speech present in the task video. Throughout our experiments we use crawls of the Dutch and English Wikipedia from August 2007. The Dutch Wikipedia contains 590,385 pages, the English Wikipedia contains 515,884,4 pages.

Given Wikipedia as our document collection, we need a way of selecting the proper content for each category (or label). We experiment with two ways of selecting this content: (i) selecting pages that belong to Wikipedia categories similar to the Vid2RSS thematic category label, and (ii) retrieving pages using the Vid2RSS thematic category label as query. We tested both methods on the development set and decided on using the second option, the label as query. Besides its slightly better performance, this method can also be used on other sources that do not have a category structure like Wikipedia.

4 Experiments

4.1 Settings

For the implementation of the SVM classifier, we use the Least Square-SVM (LS-SVM) toolbox.[5] The LS-SVM is similar to Vapnik's SVM formulation, but instead of solving the Quadratic Programming (QP) problem, it solves a set of linear equations.

As discussed in previous section, we select the training data by performing retrieval in the Wikipedia collection using the class label as the query. For each class, we collect top 200 relevant Wikipedia pages as positive examples. Since the task is a multi-class problem, we use the "one-against-all" strategy to construct the training set, i.e., using the target class as positive example and all the rest of the classes as negative examples. Therefore, for each class, a classifier is trained.

For tuning the parameters, we do a grid search with leave-one-out cross-validation on the training data. In this case, since we are using the linear kernel, the only parameter needs to be estimated is the C which is used to control the error rate.

4.2 Feature selection

After initial experiments, we decide to perform a feature selection prune away the irrelevant features. Although in SVM, the model complexity depends on the number of data points, instead of the dimensionality

[4] http://www.wikipedia.org

[5] http://www.esat.kuleuven.ac.be/sista/lssvmlab/

measure	uams08m	uams08asrd	uams08masrd	uams08asrde	uams08masrde
micro average precision	**0.13**	0.08	0.11	0.07	0.07
micro average recall	**0.28**	0.16	0.23	0.14	0.14
micro f-score	**0.18**	0.10	0.15	0.09	0.09
macro average precision	0.11	**0.44**	**0.44**	0.09	0.32
macro average recall	0.38	0.38	**0.46**	0.35	0.35
macro f-score	0.17	0.41	**0.45**	0.14	0.33

Table 2: Classification results for various runs

of the original dataset, reducing the dimensionality would still speed up the processing time as well as improving the classification result.

In our experiments, we use the χ^2 statistics to select the relevant features for each class. The χ^2 approach is not the only method we could have chosen for feature selection, but we take this approach since it has proven useful in previous work [4]. Specifically, χ^2 feature selection functions as follows: for each class, we calculate the χ^2 value for each term and rank them in decreasing order. The χ^2 measures the degree of dependence (independence) between an observed probability distribution and an expected distribution. In the context of text classification, it measures the dependency between the observed term frequency distribution in the training set and its expected frequency distribution across classes. The high value of χ^2 of a term with respect to certain class indicates high dependency between the term and the class. For our experiments, we heuristically select 80 top features given the χ^2 values.

5 Results and observations

The results of the official runs submitted by UAms are listed in Table 2. Below, we discuss our observation for the dimensions metadata and language.

5.1 Impact of metadata

The first dimension we explore is the impact of metadata on the classification of videos. When data is evaluated using micro-averaging, the use of archival metdata clearly shows a positive impact on performance (i.e., cf. uams08m vs. uams08asrd). The classifiers have the tendency to assign their thematic category labels to too many documents, reflected in the relatively high recall compared to the precision. The picture is different, when results are evaluated using macro-averaging. Here, results are best when metadata is combined with speech transcripts. The reason for the difference between macro-averaging and micro-averaging is that macro-averaging can assign disproportionately large weights to thematic categories whose classifiers perform well, even though these classifiers are relevant to only a small number of videos. For uams08masrd, the classifiers over-assignment of class labels to videos is concentrated in a smaller number of classes. This behavior can be desirable or undesirable depending on the application. We would like to note some of our best performing single classifiers are the "music" classifier and the "history" classifier for uams08m. The "music" classifier achieves a precision of 0.57 and a recall of 0.36 and the "history" classifier achieves precision of 0.36 and recall of 0.40. We believe that this relatively high performance is due to satisfactory match between the vocabulary used by the archivists to create the archival metadata and the vocabulary used in Wikipedia to describe music. Both with respect to micro-averaging and to macro-averaging, adding metadata to the speech recognition transcripts improves, or at least does not hurt results (i.e., compare runs uams08asrd vs. uams08masrd, and uams08asrde vs. uams08masrde). In general, the results point to the conclusion that archival metadata serves to enhance classification performance.

5.2 Impact of speech transcripts from two languages

The second dimension that we explore is the impact of using both Dutch and English speech recognition transcripts for classification. Recall that the videos for the task contain two languages: Dutch as matrix

language (the underlying, main language of the show), and English as embedded language. The run using Dutch speech transcripts alone (uams08asrd) is not improved by the addition of English speech transcripts (cf. uams08asrde). Nor is the run using Dutch speech transcripts with archival metadata (uams08masrd) improved by the addition of English speech transcripts (cf. uams08masrde).

We had anticipated that there would be more useful Dutch features than English features in the data since more Dutch than English is spoken in the collection. However, we believed that it would be possible to also exploit information present in the English features. It proved, however, not to be the case that the speech transcripts of the embedded language improved classification performance.

6 Conclusions

In this pilot year of the VideoCLEF track, the University of Amsterdam participated in the Vid2RSS classification task and explored two dimensions: the use of archival metadata and the use of speech transcripts from both the matrix and the embedded language in the video. We drew our training data from the Dutch and the English editions of Wikipeida and used an SVM with a linear kernel to carry out classification.

The results show that metadata is very useful in classifying videos in topic classes: highest scores on both macro and micro level are achieved by runs using metadata. Regarding the use of one (matrix) or two (matrix and embedded) languages in the classification process, we conclude that adding an extra language does not lead to improved results. Performance of runs using only the matrix language are consistently higher than using two languages.

Overall, the results achieved on this task fell short of being satisfactory. However, individual classifiers in individual runs proved promising (e.g., "music" and "history" in the metadata only run), suggesting that further development of our methods could be successful in generalizing this performance to more topic classes. In the future, we would like to examine the classes that performed the poorest in order to ascertain the reason for their failure. We suspect that the root of the problem lies in the mismatch between the training data and the test data. However, we will also prioritize the optimization of our feature selection method and of our parameters settings. Additionally, we would also like to experiment with methods to filter the speech recognition transcripts and discard those portions where the Dutch recognizer is producing output while English is being spoken and vice versa. We believe that a more judicious selection of speech-based features from the transcripts will serve to make them useful to the classification process.

7 Acknowledgements

This research was supported by the E.U. IST programme of the 6th FP for RTD under project MultiMATCH contract IST-033104.

References

[1] G. Amato, J. Cigarran, J. Gonzalo, J Peters, and P Savino. Multimatch - multilingual/multimedia access to cultural heritage. *Lecture Notes in Computer Science*, 4675:505–508, 2007.

[2] H. Drucker, Donghui Wu, and V. N. Vapnik. Support vector machines for spam categorization. *Neural Networks, IEEE Transactions on*, 10(5):1048–1054, 1999.

[3] Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. *Machine Learning: ECML-98*, Volume 1398/1998:137–142, 1998.

[4] E. Leopold, J. Kindermann, G Paass, S. Volmer, R. Cavet, M. Larson, S. Eickeler, and T. Kastner. Integrated classification of audio, video and speech using partitions of low-level features. In *Proceedings of the Workshop on Multimedia Discovery and Mining*, 2003.

[5] Gerhard Paass, Edda Leopold, Martha Larson, Jörg Kindermann, and Stefan Eickeler. SVM classification using sequences of phonemes and syllables. In *PKDD '02: Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery*, pages 373–384, 2002.