Chemnitz at VideoCLEF 2009: Experiments and Observations on Treating Classification as IR Task

Jens Kürsten and Maximilian Eibl
Chemnitz University of Technology
Faculty of Computer Science, Dept. Computer Science and Media
09107 Chemnitz, Germany
[jens.kuersten | maximilian.eibl] at cs.tu-chemnitz.de

Abstract

This paper describes the participation of the Chemnitz University of Technology in the VideoCLEF 2009 classification task. Our motivation lies in its close relation to our research project sachsMedia[1]. In our second participation in the task we experimented with treating the task as IR problem and used the Xtrieval framework [3] to run our experiments. We proposed a automatic threshold estimation to limit the number of documents per label since too many returned documents hurt the overall correct classification rate. Although the experimental setup was enhanced this year and the data sets were changed we found that the IR approach still works quite well. Our query expansion approach performed better than the baseline experiments in terms of mean average precision. We also showed that combining the ASR transcriptions and the archival metadata improves the classification performance, unless query expansion is used in the retrieval phase.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing

General Terms

Measurement, Performance, Experimentation

Keywords

Automatic Speech Transcripts, Video Classification

1 Introduction and Motivation

This article describes a system and its configuration that was used for our participation in the *VideoCLEF classification task*. The task was to categorize dual-language video into 46 given classes based on provided ASR transcripts [5] and additional archival metadata. In a mandatory experiment only the ASR transcripts of the videos had to be used as source for classification. Furthermore each of the given video documents can have none, one or even multiple labels. Hence the task can be characterized as a real world scenario in the field of automatic classification.

Our participation in the task is motivated by the its close relation to our research project *sachsMedia*[1]. The main goals of the project are twofold. The first main objective is automatic extraction of low level

[1] Funded by the *Entrepreneurial Regions* program of the German Federal Ministry of Education and Research from April 2007 to March 2012

features from audio and video for automated annotation of poorly described material in archives. On the other hand *sachsMedia* aims to support local TV stations in Saxony to replace analog distribution technology with innovative digital distribution services. A special problem of the broadcasters is the accessibility of their archives for end users. Though we are currently developing algorithms for automatic extraction of low-level metadata the *VideoCLEF classification task* is a direct use case within our project. The remainder of the article is organized as follows. In section 2 we briefly review existing approaches and describe the system architecture and its main configuration. In sections 3 and 4 we present the results of preliminary and officially submitted experiments and interpret the results. A summary of our observations and experiences is given in section 5. The final section concludes the experiments with respect to our expectations and gives and outlook to future work.

2 System Architecture and Configuration

Since the classification task was an enhanced modification of last years *VideoCLEF* classification task [4], we give a brief review on previously used approaches. There were mainly two distinct ways to approach the classification task: (a) collecting training data from external sources like general Web content or Wikipedia to train a text classifier or (b) treat the problem as information retrieval task. Villena and Lana [8] combined both ideas by obtaining training data from Wikipedia and assigning the class labels to the indexed training data. The metadata from the video documents were used as query on the training corpus and the dominant label of the retrieved documents was assigned as class label. Newman and Jones [6] as well as Perea-Ortega et. al. [7] approached the problem merely as IR task and achieved similar strong performance. Kürsten et. al. [2] and He et. al. [1] tried to solve the problem with state of the art classifiers like k-NN and SVM. Both used Wikipedia articles to train their classifiers.

2.1 Resources

Given the impressions from last year's evaluation and the huge success of the IR approaches as well as the enhancement of the task to a larger number of class labels and more documents, we decided to treat the problem as an IR task. Hence we used the *Xtrieval framework* [3] to create an index on the provided metadata. This index was composed of three fields, one with the ASR output, one with the archival metadata and a third one containing both. To process the tokens a language specific stopword list[2] and the Dutch stemmer from the Snowball project[2] was applied. We used the class labels to query our video document index. The Lucene[4] retrieval core with the default vector-based IR model was utilized within our framework. In the retrieval phase we used an English thesaurus[5] in combination with the Google AJAX language API[6] for query expansion purposes.

2.2 System Configuration and Parameters

The following list briefly explains some of our system parameters and their values for the experimental evaluation.

- *Query Expansion (QE):* The most frequent term from the top-5 documents was used to reformulate the original query.
- *Thesaurus Term Query Expansion (TT):* Thesaurus term query expansion was used for those queries, which returned less than two documents (even after QE).
- *Multi-label Limit (DpL):* DpL denotes the maximum number of assigned documents per class label and it was used to manually set a threshold for the document cut-off in the result sets.

[2] http://snowball.tartarus.org/algorithms/dutch/stop.txt
[3] http://snowball.tartarus.org/algorithms/dutch/stemmer.html
[4] http://lucene.apache.org
[5] http://de.openoffice.org/spellcheck/about-spellcheck-detail.html#thesaurus
[6] http://code.google.com/apis/ajaxlanguage/documentation

- *Source Field (SF):* The metadata source was variated to indicate which source is most reliable and whether their combination yields to improvement of the classification or not.

Due to the problem of determining the document cut-off level a priori we calculated the following threshold for each query. The threshold T_{DpL} is based on the scores of the retrieved documents per class label. Thereby RSV_{avg} denotes the average score and RSV_{max} is the maximum score of the documents retrieved. Num_{docs} stands for the total number of document retrieved for a specific class label.

$$T_{DpL} = RSV_{avg} + 2 * \frac{RSV_{max} - RSV_{avg}}{Num_{docs}}$$

3 Experiments and Results

In this section we report results that were obtained by running various system configurations on the provided training data. In table 1 columns 2-5 refer to specific system parameters that were introduced in section 2.2. Please note that the utilization of the threshold formula is denoted with x in column DpL, which means that the number of assigned documents can be different for each class label.

Regarding the evaluation of the task we had a problem with calculating the measures. We report two values for MAP due to a peculiarity in our *Xtrieval framework*, which allows the system to return two documents with identical RSV. The trec_eval[7] tool seems to penalize this behavior by randomly reordering the result set. Thus the MAP values reported by trec_eval and our framework (labeled MAP* in the following tables) have marginal variations. Unfortunately we were neither able to correct the behavior of our system nor could we find out when or why the trec_eval tool reorders our result sets. Thus, we decided to report both MAP values for our experiments in agreement with the task organizers.

3.1 Experiments on the Training Data

For evaluation of the classification performance the total number of assigned labels (SumL), the ratio of correct assigned labels (CR), averaged recall (AR) over all class labels and mean average precision (MAP) are reported. Table 1 is divided into three sections with respect to the used metadata sources. In the five rightmost columns the best values for each section of the table are emphasized bold and the best value over all sections is marked bold and italic.

Table 1: Evaluation Results on the Training Data

ID	*SF*	*QE*	*TT*	*DpL*	*SumL*	*CR*	*AR*	*MAP**	*MAP*
cut1_l1_base	asr	no	0	1	33	**0.3333**	0.0558	0.0485	0.0485
cut2_l0_qe	asr	yes	5	∞	**1,566**	0.0390	**0.3096**	0.1072	**0.1099**
cut3_l1_qe	asr	yes	5	1	181	0.1602	0.1472	0.0993	0.1006
cut4_l1_base	meta	no	0	1	70	***0.4714***	0.1675	0.1554	0.1546
cut5_l0_qe	meta	yes	5	∞	**1,932**	0.0813	**0.7970**	0.4933	***0.4999***
cut6_l1_qe	meta	yes	5	1	188	0.3617	0.3452	0.2969	0.2985
cut7_l2_qe	meta	yes	5	2	312	0.3013	0.4772	0.3890	0.3928
cut8_l3_qe	meta	yes	5	3	368	0.3043	0.5685	0.4349	0.4395
cut9_lx_qe	meta	yes	5	x	395	0.2886	0.5787	0.4361	0.4407
cut10_l1_base	asr + meta	no	0	1	108	**0.4537**	0.2487	0.2177	0.2163
cut11_l0_qe	asr + meta	yes	5	∞	***1,999***	0.0795	***0.8071***	0.5036	**0.4975**
cut12_l1_qe	asr + meta	yes	5	1	205	0.3659	0.3807	0.3056	0.3059
cut13_l2_qe	asr + meta	yes	5	2	336	0.3036	0.5178	0.4035	0.3993
cut14_l3_qe	asr + meta	yes	5	3	414	0.2874	0.6041	0.4574	0.4523
cut15_lx_qe	asr + meta	yes	5	x	470	0.2681	0.6396	0.4741	0.4689

[7] http://trec.nist.gov/trec_eval

The following observations can be made by analyzing the experimental results. No matter which metadata source was used, the experiment without limitation of the class labels per document had the best performance in terms of AR and MAP (see ID's cut2, cut5 and cut11). The drawback of those runs is that they have very low correct classification rates (CR) of about 3% for the ASR data and about 8% when using archival metadata alone or in combination with ASR data. In contrast to that the experiments without any form of query expansion (see ID's cut1, cut4 and cut10) had the highest correct classification rates (CR) from 33% up to 47%. However, this is more a result from limiting to one document per label, which also yields to lower performance in terms of AR and MAP. Numerous experiments with either manual or automatic thresholds to limit the assigned documents per label were conducted. The results show that it is possible to improve CR substantially and almost sustain the best MAP values (compare cut5 to cut9 and cut11 to cut15). Nevertheless for those runs the AR was significantly lower.

3.2 Experiments on the Test Data

In this section we report the experimental results on the evaluation data set. Please note that we run all configurations from section 3.1 again, because we wanted to figure out if our observations on the training data are also valid on the test data set. Experiments that were submitted for official evaluation by the organizers of the task are denoted with *. Again in table 2 columns 2-5 contain parameters of our system, which are briefly explained in section 2.2. The performance of the experiments is reported with respect to overall sum of assigned label (SumL), the average ratio of correct classifications (CR) as well as average recall (AR) and mean average precision (MAP). Corresponding to section 3.1 table 2 is also divided into three sections with respect to the used metadata sources. In the five rightmost columns the best values for each section of the table are emphasized bold and the best value over all sections is marked bold and italic.

Table 2: Evaluation Results on the Test Data

ID	*SF*	*QE*	*TT*	*DpL*	*SumL*	*CR*	*AR*	*MAP**	*MAP*
cut1_l1_base*	asr	no	0	1	27	0.0741	0.0101	0.0104	0.0067
cut2_l0_qe	asr	yes	5	∞	***1,966***	0.0310	**0.3065**	0.1015	**0.1010**
cut3_l1_qe*	asr	yes	5	1	171	**0.1111**	0.0958	0.0848	0.0842
cut4_l1_base	meta	no	0	1	63	***0.6349***	0.2010	0.2004	0.2003
cut5_l0_base	meta	yes	5	∞	**1,778**	0.0889	***0.7940***	0.4478	***0.4505***
cut6_l1_base	meta	yes	5	1	194	0.3763	0.3668	0.2863	0.2867
cut7_l2_base	meta	yes	5	2	300	0.3300	0.4975	0.3693	0.3706
cut8_l3_base	meta	yes	5	3	354	0.3051	0.5427	0.3974	0.4006
cut9_lx_base	meta	yes	5	x	389	0.2853	0.5578	0.4039	0.4073
cut10_l1_base*	meta + asr	no	0	1	112	**0.5000**	0.2814	0.2541	0.2586
cut11_l0_qe	meta + asr	yes	5	∞	**1,885**	0.0838	***0.7940***	0.4404	**0.4389**
cut12_l1_qe*	meta + asr	yes	5	1	196	0.3622	0.3568	0.2552	0.2531
cut13_l2_qe	meta + asr	yes	5	2	328	0.3018	0.4975	0.3712	0.3704
cut14_l3_qe*	meta + asr	yes	5	3	393	0.2723	0.5379	0.3837	0.3813
cut15_lx_qe	meta + asr	yes	5	x	444	0.2455	0.5478	0.3869	0.3844

3.3 Observations and Interpretation

In general we see similar behavior on both the training and the test data set. For all data sources used the best correct classification rate (CR) is achieved without using any form of query expansion (see ID's cut1, cut4 and cut10). The best overall (CR) was achieved by only using archival metadata in the retrieval phase. Since the archival metadata consists of intellectual annotations this is a very straightforward finding. Another obvious observation is, that the best overall results in terms of MAP and AR were also achieved on the archival metadata. Nevertheless the gap to the best results when combining ASR output with archival metadata is very small (compare cut5 to cut11). Regarding our proposed automatic threshold calculation for limitation of the number of assigned documents per label the results are twofold. On the one hand there is a slight improvement in terms of MAP and AR compared to low manually fixed thresholds between 1 and 3 assigned documents per label. On the other hand the overall correct classification rate (CR) decreases in the same magnitude MAP and AR are increasing, which is another very straightforward finding.

The interpretation of our experimental results led us to the conclusion that using MAP for evaluating a multi-label classification task is somehow questionable. The main reason in our point of view is that MAP does not take into account the overall correct classification rate CR. Let us take a look on the two best performing experiments using archival metadata and ASR transcriptions either in table 1 or 2 (see ID's cut10 and cut15). The difference in terms of MAP is about 6% or 12%, but the gain in terms of CR is about 293% or 337% respectively. In our opinion in a real world scenario were assignment of class labels to video documents should be completely automatic it would be essential to take into account the overall ratio of correct assigned labels. Our prosposal for future evaluations is to combine measures that take into account the position of the correct assigned labels in a result set (like MAP or averaged R-Precision) with the micro or macro correct classification rate.

4 Result Analysis - Summary

The following list provides a short summary of our observations and findings from the participation in the *VideoCLEF classification task* in 2009.

- *Classification as an IR task:* According to the experiences from last year, we conclude that treating the given task as a traditional IR task with some modifications is a quite successful approach.
- *Query Expansion:* Both types of query expansion improved the results in terms of MAP and AR but had very low correct classification rates CR.
- *Metadata Sources:* Combining both ASR output and archival metadata improves MAP and AR when no query expansion is used. For those experiments where query expansion was used there is no gain in terms of MAP and AR comparing archival metadata runs to experiments which used both data sources.
- *Label Limits:* We compared an automatically calculated threshold to low manual set thresholds and found that the automatic threshold works better in terms of MAP and AR.
- *Evaluation Measure:* In our opinion using MAP as evaluation measure for a multi-label classification task is questionable. We would prefer a measure that takes into account both correct classification rate and averaged recall.

5 Conclusion and Future Work

This year we used the *Xtrieval framework* for the *VideoCLEF classification task*. In our experimental evaluation we can confirm the observations from last year, where approaches treating the task as IR problem were most successful. We proposed an automatic threshold to limit the number of assigned documents per class label to keep high correct classification rates. This seems to be the main issue that could be worked on in the future. A manual limitation of assigned documents per label is not an appropriate solution to a comparable real world problem, where possibly tens or hundred of thousand video documents should be labeled with maybe hundreds of different topic labels. Furthermore one could try to evaluate different retrieval models or try to combine the results from those models to gain a better overall performance. Finally it should be evaluated if assigning field boosts to the metadata sources could improve performance in the combined retrieval setting.

Acknowledgments

We would like to thank the *VideoCLEF* organizers and the Netherlands Institute of Sound and Vision (Beeld & Geluid) for providing the data sources for the task.

This work was accomplished in conjunction with the project *sachsMedia*, which is funded by the *Entrepreneurial Regions*[8] program of the German Federal Ministry of Education and Research.

References

[1] Jyin He, Xu Zhang, Wouter Weerkamp, and Martha Larson. The University of Amsterdam at VideoCLEF 2008. *Working Notes for the CLEF 2008 Workshop, 17-19 September, Aarhus, Denmark*, 2008.

[2] Jens Kürsten, Daniel Richter, and Maximlian Eibl. VideoCLEF 2008: ASR Classification based on Wikipedia Categories. *Working Notes for the CLEF 2008 Workshop, 17-19 September, Aarhus, Denmark*, 2008.

[3] Jens Kürsten, Thomas Wilhelm, and Maximilian Eibl. Extensible Retrieval and Evaluation Framework: Xtrieval. *LWA 2008: Lernen - Wissen - Adaption, Würzburg, October 2008, Workshop Proceedings*, 2008.

[4] Martha Larson, Eamonn Newman, and Gareth Jones. Overview of VideoCLEF 2008: Automatic Generation of Topic-based Feeds for Dual Language Audio-Visual Content. *Working Notes for the CLEF 2008 Workshop, 17-19 September, Aarhus, Denmark*, 2008.

[5] Martha Larson, Eamonn Newman, and Gareth Jones. Overview of VideoCLEF 2009: New Perspectives on Speech-based Multimedia Content Enrichment. In Francesca Borri, Alessandro Nardi, and Carol Peters, editors, *Working Notes of CLEF 2009*, September 2009.

[6] Eamonn Newman and Gareth J. F. Jones. DCU at VideoClef 2008. *Working Notes for the CLEF 2008 Workshop, 17-19 September, Aarhus, Denmark*, 2008.

[7] José M. Perea-Ortega, Arturo Montejo-Raéz, and M. Teresa Martín-Valdivia. SINAI at VideoCLEF 2008. *Working Notes for the CLEF 2008 Workshop, 17-19 September, Aarhus, Denmark*, 2008.

[8] Julio Villena-Román and Sara Lana-Serrano. MIRACLE at VideoCLEF 2008: Classification of Multilingual Speech Transcripts. *Working Notes for the CLEF 2008 Workshop, 17-19 September, Aarhus, Denmark*, 2008.

[8] The Innovation Initiative for the New German Federal States