Using a Divergence Model for MediaEval's Tagging Task (Professional Version)

Christian Wartena
Novay
Brouwerijstraat 1
7523 XC Enschede
The Netherlands
christian.wartena@novay.nl

ABSTRACT

Novay participated in MedieaEval Tagging Task (professional version). For this task videos have to be ranked according to their relevance for a number of different concepts. Our approach was based solely on the abstracts of the videos. A divergence model has been used for retrieval in which both the query and the document model are extended by a Markov chain. The results could be improved by using a small set of synonyms to represent each concept, by enlarging the basis for computing the language models of the query terms, and by also taking into account the rank the concept has for the abstract. The best result was obtained by combining all of these possibilities.

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.3.7 Digital Libraries

General Terms

Measurement,Performance,Experimentation

1. INTRODUCTION

Keywords are an important type of metadata to describe a text or a video at a high level. Thus many archival institutions use keywords to describe their content. In the task under consideration programs from Dutch public television, archived by the Dutch Institute for Sound and Vision, are used. The task of the benchmark was to find the videos to which a given term has been assigned by a human cataloguer.

Cataloguers assigning keywords at Sound and Vision often consult available contextual information such as the synopsis of a broadcast [3]. Moreover, cataloguers use a more or less restricted vocabulary for assigning keywords. Thus if a word from this vocabulary is present in the synopsis, it is very likely to become a keyword.

We use the frequency of a query term in the synopsis (i.e. the textual description: the title and the abstract) as a baseline for determining the relevance of a video for that term. The basic idea for improving on this baseline is that not only the number of occurrences of a term in a text determines its importance. If many closely related words appear in the text, the word also becomes more likely to be selected as keyword. The idea that co-occurring terms contribute to the relevance of the search term is underlying common techniques like query expansion and pseudo-relevance feedback. We basically follow the model-based feedback approach of [7], but also incorporate a form of thesaurus-based query expansion.

MediaEval 2010 Workshop, October 24, 2010, Pisa, Italy

2. DESCRIPTION OF THE APPROACH

In order to find appropriate keywords for each video we have used the titles and abstracts only. In a first preprocessing step this information was extracted from the XML-metadata and stored as plain text. Subsequently, these texts were analyzed and annotated by a standard UIMA ([1]) pipeline used at Novay. This pipeline consists of word segmentation, sentence boundary detection, part-of-speech tagging and lemmatization. All methods described below work only with the lemmas and only consider open class words, i.e. nouns, verbs, adjectives and adverbs. Most labels are given in plural form, due to archival conventions. In order to match the labels with the lemmas we have converted all labels to singular form.

As a baseline for the relevance of a word for an abstract we use the probability of a word in that document being the given term. The results of this baseline are given in the first line of the first column of the results table (Table 1). Of the 41 labels used in the task only one label (*kunstenaars*) does not occur in any of the synopses.

In the divergence model from [4] and [7] the retrieval problem is essentially equivalent to the problem of estimating the query and document language model. Following [4] we use Markov chains to obtain these models. Given a document collection D the language model for a query term q is defined as

$$\bar{p}_q(t) = \sum_{d \in D} p(t|d)p(d|q). \qquad (1)$$

where $p(t|d)$, the *term distribution* of d is the probability that a term from d is an instance of t, and where $p(d|t)$, the *source distribution* of t, is the probability that a randomly selected occurrence of t has source d. We refer to [5],[6] for details. Assuming that q is a term like other terms, we call this distribution also the co-occurrence distribution of q. The language model of a document d could simply be the distribution $p(t|d)$ of terms of the document. However, as usual we take a smoothed version. We again use the Markov

Table 1: MAP for different methods

Method	no syn.	synonyms
Frequency	0,37	0,42
Divergence	0,42	0,47
Max. Entropy	0,43	0,48
Divergence (incl. dev. set)	0,45	0,48
Max Entropy (incl. dev. set)	0,46	0,49

chain to obtain the smoothed distribution

$$\bar{p}_d(t) = \sum_{d',t'} p(t|d')p(d'|t')p(t'|d) = \sum_z p(z|d)\bar{p}_z(t). \quad (2)$$

For the comparison of the co-occurrence distribution and the document distribution we use the Jensen-Shannon divergence ([2]). Results of this relevance measure are given in the first results column, second line of Table 1.

The evaluation for the task involves a ranking of documents for a given query. If we use the divergence of document and co-occurrence distribution of a term to rank documents, as described above, we assume that these divergences can be compared among several documents. However, this is not the case. E.g. very short documents tend to have larger divergences to all terms, and therefore always will be ranked very low. Nevertheless, it might be a very likely candidate for the query term that has the smallest divergence for that document. To obtain a potentially better relevance model we linearly combine several measures. The relevance of a document d for a query term q now becomes:

$$r(d|q) = \alpha + \beta\, p(q|d) - \gamma\, \mathrm{JSD}(\bar{p}_q, \bar{p}_d) - \delta\, \mathrm{rank}(q,d)/n_l \quad (3)$$

where $\mathrm{JSD}(\overline{p_q}, \overline{p_d})$ is the Jensen-Shannon divergence between $\overline{p_w}$ and $\overline{p_d}$, $\mathrm{rank}(q,d)$ is the rank of q for d and n_l the total number of labels used. The coefficients were determined using a maximum entropy model on the test set ($\alpha = 1.0$, $\beta = 2.0$, $\gamma = 1.0$, $\delta = 0.17$). The results of this relevance measure are given in the first results column, third line of Table 1. As expected, this gives indeed a slight improvement over the run using only the divergence.

The co-occurrence distribution of a term can be seen as a proxy for its semantics. In this sense the distribution will improve if we take more documents into account for the computation of the co-occurrence probabilities. Thus in the next two runs we have used the synopses from the test and the development set to compute the co-occurrence distributions. Again we can use only the divergence or combine it with the other features. The results of these two runs are given in the last two lines, again showing a slight improvement.

Finally, the labels provided are in some cases rather formal and official terms, that do not occur very frequently in the texts. E.g. the term *buitenlandse werknemers* is much less frequently used than the common term *gastarbeider*. Similarly, in Dutch the term *acteur* is only used to denote male actors, while female actors are called *actrice*. Thus we expect further improvement if we take such synonyms and alternative terms into account. A list of synonyms was manually constructed. We do not use these synonyms for query expansion in the literal sense, but we represent a document by a bag of concepts, rather than a bag of words, considering a set of synonyms as a concept. The results of the runs using the synonym list are given in the third column of Table 1. In all cases the usage of synonyms gives better results. The baseline benefits directly from the synonyms, while in the other cases the main effect is that more documents are taken into account to compute the co-occurrence distribution.

3. DISCUSSION

Standard techniques for text based retrieval seem to yield good results for the given data set. This mainly demonstrates the value of editorial abstracts for video retrieval. Given the good results of our baseline, it seems likely that an important part of the result should be contributed to the carefully implemented preprocessing.

We have used three techniques (language modeling, query expansion and maximum entropy modeling) to improve on the baseline. We have shown that each of these techniques improves the results. The best result was obtained by combining them all, which could be done in a natural way. Finally, by considering the query language model not as a result of pseudo-feedback, but as a proxy for the semantics of the query term, we could improve this model using more data. The improved query language model again improved the results consistently.

Acknowledgments

We would like to thank Martha Larson, Wout Slakhorst and Rogier Brussee for helpful discussions. The research presented here was done within the MyMedia project funded by the European Community's Seventh Framework Program (FP7/2007-2011) under grant agreement Nº 215006.

4. REFERENCES

[1] Apache uima. http://incubator.apache.org/uima/.

[2] T. Cover and J. Thomas. *Elements of information theory*. Wiley-Interscience, 2006.

[3] L. Gazendam, C. Wartena, V. Malaise, G. Schreiber, A. de Jong, and H. Brugman. Automatic annotation suggestions for audiovisual archives: Evaluation aspects. *Interdisciplinary Science Reviews, 34*, 2(3):172–188, 2009.

[4] J. D. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. In W. B. Croft, D. J. Harper, D. H. Kraft, and J. Zobel, editors, *SIGIR*, pages 111–119. ACM, 2001.

[5] C. Wartena and R. Brussee. Topic detection by clustering keywords. In *DEXA Workshops*, pages 54–58. IEEE Computer Society, 2008.

[6] C. Wartena, R. Brussee, and W. Slakhorst. Keyword extraction using word co-occurrence. In *DEXA Workshops*, pages 54–58. IEEE Computer Society, 2010.

[7] C. Zhai and J. D. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *CIKM*, pages 403–410. ACM, 2001.