DCU at MediaEval 2010 – Tagging Task WildWildWeb

Ágnes Gyarmati
Centre for Digital Video Processing
Dublin City University
Dublin 9, Ireland
agyarmati@computing.dcu.ie

Gareth J. F. Jones
Centre for Digital Video Processing
Dublin City University
Dublin 9, Ireland
gjones@computing.dcu.ie

ABSTRACT

We describe our runs and results for the fixed label Wild Wild Web Tagging Task at MediaEval 2010. Our experiments indicate that including all words in the ASR transcripts of the document set results in better labeling accuracy than restricting the index to only words with recognition confidence above a fixed level. Additionally our results show that tagging accuracy can be improved by incorporating additional metadata describing the documents where it is available.

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; I.2 [**Artificial Intelligence**]: I.2.7 Natural Language Processing—*Speech recognition and synthesis*

General Terms

Measurement, Experimentation

Keywords

information retrieval, automatic speech recognition, text tagging

1. INTRODUCTION

The *Tagging Task "Wild Wild Web"* held as part of MediaEval 2010 required participants to automatically predict tag annotations for a set of internet videos from `blip.tv`. The videos are accompanied by automatic speech recognition (ASR) generated transcripts provided by Vecsys Research[1] and LIMSI-CNRS[2] and various other metadata (e.g. titles, descriptions, twitter messages related to the videos). The topic, genre, quality and language of the videos vary, hence the name of the task ''Wild Wild Web." Tags originally assigned manually by the uploaders are used as the basis of the evaluation of the automatic tagging carried out in this task. Since there is no tagging convention in usage at `blip.tv`, there may be inconsistencies in the assignment of these manual tags, potentially making it challenging to emulate the behaviour of manual taggers.

Two flavours of "Wild Wild Web" tagging task were offered at MediaEval 2010, with participants being allowed to submit up to 5 runs based on variants of their tagging approach: Closed-set tagging: a list of possible tags was provided with the data, and the task was to automatically predict assignment of these on the test video data. Open-set tagging: by disregarding the tag list provided, the task was to select and assign tags to videos without vocabulary constraint.

Tagging is usually considered as a categorization task, but it can be approached as an information retrieval (IR) task. Adopting the IR approach has been shown to have potential for tasks such as this, for example at VideoCLEF 2009 [2].

In addition to words in the most likely transcript from the ASR transcript, each word hypothesis included a confidence score which was available for use. One might expect words with higher confidence scores to lead to more reliable or accurate assignment of tags. In related work Zechner and Waibel demonstrated positive effects using confidence information in summarization [4], however in IR using this information has shown minimal impact on effectiveness [3].

2. SYSTEM DESCRIPTION

For our participation in MediaEval 2010 we decided to tackle the "closed set" variant of the task. We approached tagging as an IR task, where the documents to be retrieved were the videos (i.e. their transcripts, optionally combined with some metadata: titles and descriptions). Queries were formed using the closed set of possible tags.

We used the *Indri* model of the open source `Lemur` Toolkit[3] for indexing and retrieving. English texts were stemmed using `Lemur`'s own built-in stemmer, while Dutch, French and Spanish texts were stemmed using `Oleander`'s[4] implementation of `Snowball`'s[5] Porter stemmer algorithm for various languages. We used stopword lists provided by Snowball for all languages involved in this task. No preprocessing (stemming or stopping) was done if the language was intentionally considered as "unknown".

The same basic searching method was applied for each run. Our four submitted runs used the transcripts incorporating varied thresholds on the word confidence scores extracted from the transcripts and did not distinguish be-

[1]http://www.vecsysresearch.fr/
[2]http://www.limsi.fr/

[3]http://www.lemurproject.org/
[4]http://sourceforge.net/projects/porterstemmers/
[5]http://snowball.tartarus.org/

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

tween the document languages. Additional runs incorporated metadata fields in the indexed collection and indexed the documents differently according to their language.

3. RUN CONFIGURATIONS AND RESULTS

3.1 Submitted Runs

We submitted four runs for the MediaEval 2010 "Wild Wild Web" fixed vocabulary tagging task. These made no distinction between the languages of the transcripts, and they were all combined into a single search index. Since the documents' languages were considered as "unknown", no stemming and stopping was performed on either the transcripts or the query labels. The difference between the submitted runs was the text used for indexing: whether all the words of the transcripts were used (Run 1), or only words with a confidence score better than a fixed threshold of 70%, 80% and 90% for Run 2, Run 3, and Run 4 respectively. Each tag was assigned to a maximum of 20 videos, by selecting the top 20 relevant items returned by the IR system.

Table 1 shows the Mean Average Precision (MAP) values for our submitted runs. While the differences between the values are small, there is a noticeable tendency of decreasing performance as the confidence threshold is set higher.

Run	MAP
Run 1	0.155
Run 2	0.145
Run 3	0.139
Run 4	0.129

Table 1: Official results (MAP values)

as "unknown" language	ASR	+metadata
all words	0.169	0.273
conf.sc > 70%	0.158	0.269
conf.sc > 80%	0.151	0.266
conf.sc > 90%	0.141	0.264
as distinct languages	ASR	+metadata
all words	0.162	0.254
conf.sc > 80%	0.150	0.253

Table 2: Unofficial results (MAP values)

3.2 Additional Runs

We repeated the "official" runs but now with no limit on the number of returned documents (i.e. videos) to which the labels were assigned, hence the slight changes in performance compared to the official runs. We also performed retrieval on an indexed collection that included metadata (title and descriptions fields) information as well. All other details (confidence scores, lack of preprocessing) were the same, results are shown in the upper half of Table 2. Combining metadata with the transcript texts can be seen to improve the performance, as one would expect.

We also investigated indexing the transcripts separately for each language. For each collection a language specific stop word list was used and stemming applied, optionally texts were also combined with metadata. Stopping and stemming was similarly applied to query labels. Note that since these were not labelled by language, all labels were used to search each collection with the stopping and stemming matching the transcript language applied. If categorised by language, labels might be used multilingually, either with the same meaning (e.g. international words, proper names), or with a different meaning (the same word form coincidentally existing in several languages). Search was performed in each language with the same conditions as before (confidence scores, metadata). The four results for each query label were then merged into single file sorted by matching score. The results for the language-sensitive runs are shown in the lower half of Table 2. These runs are slightly outperformed by runs considering language as unknown. The reason for this is not immediately clear and will be the subject of further investigation.

4. CONCLUSIONS AND FURTHER WORK

All our official and unofficial runs seem to confirm that taking confidence scores into account does not necessarily help in IR, a slight decrease in performance is noticeable here. Recognising and dealing with language differences is also an issue that needs further investigations. In further work we also plan to explore the use of existing methods for determining different cut off points for individual tags such as [1], and also to examine other possible techniques to do this, as some tags will be associated with many items and others with only a small number of them. Also we plan to extend use of this IR based tagging method to the task to open-set tagging. This will necessitate development of a system component to automatically identify potential tags, as well as their assignment to the items to be annotated.

5. ACKNOWLEDGEMENTS

This work is funded by a grant under the Science Foundation Ireland Research Frontiers Programme 2008 Grant No: 08/RFP/CMS1677.

6. REFERENCES

[1] J. Kürsten and M. Eibl. Video classification as IR task: Experiments and observations. In C. Peters, B. Caputo, J. Gonzalo, G. J. F. Jones, J. Kalpathy-Cramer, H. Müller, and T. Tsikrika, editors, *Multilingual Information Access Evaluation II. Multimedia Experiments*, volume 6242 of *Lecture Notes in Computer Science*, pages 377–384. Springer Berlin / Heidelberg, 2010. 10.1007/978-3-642-15751-6_49.

[2] M. Larson, E. Newman, and G. J. F. Jones. Overview of videoclef 2009: New perspectives on speech-based multimedia content enrichment. In C. Peters, B. Caputo, J. Gonzalo, G. J. F. Jones, J. Kalpathy-Cramer, H. Müller, and T. Tsikrika, editors, *Multilingual Information Access Evaluation II. Multimedia Experiments*, volume 6242 of *Lecture Notes in Computer Science*, pages 354–368. Springer Berlin / Heidelberg, 2010. 10.1007/978-3-642-15751-6_46.

[3] M. Sanderson and X. M. Shou. Search of spoken documents retrieves well recognized transcripts. In *ECIR*, pages 505–516, 2007.

[4] K. Zechern and A. Waibel. Minimizing word error rate in textual summaries of spoken language. In *Proceedings of NAACL-ANLP-2000*, 2000.