The University of Amsterdam at the CLEF Cross Language Speech Retrieval Track 2007

Bouke Huurnink
ISLA, University of Amsterdam
bhuurnin@science.uva.nl

Abstract

In this paper we present the contents of the University of Amsterdam submission in the CLEF Cross Language Speech Retrieval 2007 English task. We describe the effects of using character n-grams and field combinations on both monolingual English retrieval, and cross-lingual Dutch to English retrieval.

Keywords

Speech Retrieval, Cross-Language Information Retrieval, Text Transformations, Field Combination

1 Introduction

Even in a well-funded archive, it is often infeasible to manually annotate all documents in the collection. The digitisation of multimedia collections opens the door to automatic techniques for discovering interesting documents, provided that we can leverage automatically generated annotations to their best advantage. The University of Amsterdam participated in the CLEF CL-SR 2007 English task in the hope of applying lessons learned there to the retrieval of documents from large Dutch audio-visual archives, in particular the Netherlands Institute for Sound and Vision[1] which stores the nation's public television broadcasts. These archives contain a lot of spoken material, some of which has been manually annotated by a team of archivists. A significant portion, however, has not been annotated at all. Therefore we investigated strategies both for search using only automatically generated text, as well as combining this text with manually generated annotations.

Our focus was on simple techniques that can easily be transferred to other domains. In our experiments we explored the use of character n-grams to improve the retrieval of documents using automatically generated text. We also explored the combination of manually generated with automatically generated text. In both cases we contrasted monolingual retrieval of English documents using English queries with cross-lingual retrieval of English results using Dutch queries.

The remainder of this paper is structured as follows. We first describe the setup of the retrieval system and experiments in Section 2. This is followed by the runs and results in Section 3. Finally we present our conclusions in Section 4.

2 Experimental Setup

We work with the CLEF CL-SR experimental English spoken document collection, which consists of a series of English language interviews that have been manually split into short segments. Each segment has been associated with manually and automatically assigned metadata, including

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

manual summaries, manually assigned keywords, automatic speech transcriptions, and a number of fields containing automatically assigned keywords. For experiments using only automatically assigned information we use the ASRTEXT2006B (speech transcription) and ASRKEYWORD2004A2 (automatic keyword) fields. Other automatic transcripts and keywords were available, but we chose to use only one of each, which may have had a negative impact on our results. For experiments including manual annotations we also added the MANUALKEYWORD (manual keyword) and SUMMARY (manual summary) fields.

The CLEF CL-SR benchmark provided 63 training topics with a ground truth, as well as 33 test topics. The original topic descriptions are in English, and have the traditional TREC *title - description - narrative* structure. Also available were manually created Dutch topic translations, donated by the University of Twente. We used these Dutch topics for the cross-lingual runs.

2.1 Retrieval Infrastructure

All documents were indexed and retrieved using the Indri engine from the Lemur retrieval toolkit[2]. This engine allows for fielded search in a language modeling framework. As is standard in English text retrieval, commonly occurring stop words were removed. Terms were stemmed to their morphological roots using the Porter [3] stemming algorithm. Retrieval parameters were optimised for automatic monolingual retrieval on the ASRTEXT2006B field, using the training topics to find the best combination.

As for the topics, the title and description fields of each topic were combined to make a text query. The Dutch topics were automatically translated to English using online resources, in order to be able to retrieve the English documents. As different translation systems perform better for different topics [4], we used two different online tools to translate the topics from Dutch to English. We used the SYSTRAN[3] and FreeTranslation.com[4] systems, and combined the results to form a large 'bag-of-words' cross-lingual query. Some of the differences between translations can be seen in the example given in Table 1. For instance, the word 'acts' is translated into 'deeds' by FreeTranslation.com and 'prowesses' by SYSTRAN.

Table 1: Sample Topic Translations (Topic 15602)

Original English Topic	**FreeTranslation.com**	**SYSTRAN**
Heroic survival stories. Stories of heroic acts or activities that led to the survival of one or more individuals are desired.	Heroic survivals story. Tell of heroic deeds or heroic actions that led till the (save) [survive] of an or several individuals have been wished.	Herosche overlevingsverhalen. Tales of prowesses or herosche action which led to (save) [survive] of one or more individuals have been needed.

2.2 Character n-Gram Experiments

Character n-gram tokenisation has been shown to boost retrieval in certain situations [2], such as retrieval from English newspapers [1]. We were interested to see whether this would also prove useful for the specific situation of (cross-lingual) retrieval of automatically generated text. To test this, we followed the tokenisation strategy in [2], and created overlapping, cross-word character n-grams of the text before it was indexed. An example is shown in Table 2. In designing the experiment, we used only the (weighted) ASRTEXT2006B and AUTOKEYWORD2004A2 fields.

We evaluated MAP for retrieval at different n-gram sizes on the training topics prior to submission, and found that 4-grams provided the best performance. Likewise, we evaluated different

[2]http://www.lemurproject.org/
[3]http://www.systran.co.uk/
[4]http://www.freetranslation.com/

Table 2: n-Gram Tokenisation Example

Original Text	4**-Grams**
heroic survivals story	hero eroi roic oic* ic*s c*su *sur surv urvi rviv viva ival vals als* ls*s s*st *sto stor tory

weightings for the ASRTEXT2006B and AUTOKEYWORD2004A2 fields. Here we found the best setting to be ASRTEXT2006B = 0.75 and AUTOKEYWORD2004A2= 0.25. These, then, are the settings that we used in our officially submitted runs.

2.3 Field Combination Experiments

We evaluated field combination, as we may later wish to apply this technique to retrieving the annotated portion of multimedia documents in a audio-visual archive. Combination was done using the Indri query language, giving different fields different weights. The fields that we used were MANUALKEYWORD, SUMMARY, ASRTEXT2006B, and AUTOKEYWORD2004A2.

As with the n-gram experiments, we determined the optimal combination setting on the set of 63 training topics that were provided, using MAP as our evaluation measure. We found that the best weighting for monolingual retrieval was MANUALKEYWORD = 0.375, SUMMARY = 0.375, ASRTEXT2006B = 0.125, and AUTOKEYWORD2004A2 = 0.125. For the cross-lingual task, the automatic keywords gave no contribution to retrieval performance and the best weighting was MANUALKEYWORD = 0.375, SUMMARY = 0.375 and ASRTEXT2006B = 0.25.

3 Runs and Results

Table 3 shows the results of the official runs submitted to CLEF CL-SR. Also shown are two runs that were generated post-hoc to allow fair comparison of the n-gram techniques to a baseline. The post-hoc runs were both generated using stopped and stemmed text from both the ASRTEXT2006B and AUTOKEYWORD2004A2 fields, weighted as described in Section 2.2.

Examining the n-gram runs, we found that monolingual retrieval of the automatic fields using character 4-grams decreased MAP by 9.6% . Cross-lingual retrieval, on the other hand, benefited from the use of 4-grams with an increase in MAP of 4%.

The combination runs, which included both manual and automatic information, performed much better than runs containing only automatically derived text. This is not surprising, it has been demonstrated in previous CLEF CL-SR tracks that manual annotation allows much better retrieval than automatic information alone. The weightings derived in the training phase indicate that automatically generated keywords are helpful for monolingual retrieval, but do not help for this specific case of cross-lingual retrieval. Automatically recognised speech, however, was useful for both monolingual and cross-lingual retrieval.

4 Conclusions

This paper has described the setup and performance of the University of Amsterdam's entry in the CLEF CL-SR 2007 English retrieval task. We investigated the effect of using n-grams to retrieve automatically generated text, finding that they decreased monolingual performance but improved cross-lingual performance. Furthermore, we examined the effects of combining manual and automatically generated text, and saw that both can be useful. We hope that the lessons learned here will aid us in practice, and help us enhance search through Dutch audio-visual archives.

Table 3: Results of CLEF CL-SR runs

Run ID	Type	Fields	MAP
UvA_1_base	monolingual baseline	ASRTEXT2006B	0.0430
UvA_2_en4g	monolingual 4-grams	ASRTEXT2006B, AUTOKEYWORD2004A2	0.0444
UvA_3_nl4g	cross-lingual 4-grams	ASRTEXT2006B, AUTOKEYWORD2004A2	0.0400
UvA_4_enopt	monolingual combination	MANUALKEYWORD, SUMMARY, ASRTEXT2006B, AUTOKEYWORD2004A2	0.2088
UvA_5_nlopt	cross-lingual combination	MANUALKEYWORD, SUMMARY, ASRTEXT2006B, AUTOKEYWORD2004A2	0.1408
unsubmitted run	monolingual	ASRTEXT2006B, AUTOKEYWORD2004A2	0.0491
unsubmitted run	cross-lingual	ASRTEXT2006B, AUTOKEYWORD2004A2	0.0385

Acknowledgments

This research was supported by the Netherlands Organisation for Scientific Research (NWO) MUNCH project under project number 640.002.501.

References

[1] V. Hollink, J. Kamps, C. Monz, and M. de Rijke. Monolingual document retrieval for European languages. *Inf. Retr.*, 7(1-2):33–52, 2004.

[2] P. McNamee and J. Mayfield. Character n-gram tokenization for European language text retrieval. *Inf. Retr.*, 7(1-2):73–97, 2004.

[3] M. F. Porter. An algorithm for suffix stripping. In *Readings in information retrieval*, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.

[4] Jacques Savoy. Report on CLEF-2003 multilingual tracks. In Carol Peters, Julio Gonzalo, Martin Braschler, and Michael Kluck, editors, *CLEF*, volume 3237 of *Lecture Notes in Computer Science*, pages 64–73. Springer, 2003.