University of Chicago at the CLEF 2007 Cross-language Speech Retrieval Track

Gina-Anne Levow
University of Chicago
levow@cs.uchicago.edu

Abstract

The University of Chicago participated in the CLEF 2007 CL-SR track, performing monolingual retrieval for both English and Czech and cross-language French-English retrieval. English experiments considered the impact of automatically generated keywords on retrieval. Czech experiments explored the effect of different stemming approaches on retrieval for this morphologically rich language. The best results for English employed automatically generated keywords, and the best results for Czech employed stemming strategies which significantly outperformed unstemmed techniques.

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

Multilingual retrieval, Cross-language Retrieval, Stemming

1 Introduction

The University of Chicago participated in the CLEF 2007 cross-language speech retrieval (CL-SR) track employing both English and Czech document collections. For the English documents, we submitted four official runs, two in the required English monolingual title+description condition and two in the French-English cross-language condition. For English documents, contrastive experiments focused on the contribution of automatically generated keywords to retrieval effectiveness and the utility of different query translation strategies. For Czech documents, we submitted three official runs in the required monolingual Czech title+description condition. These experiments explored the utility of a range of different stemming procedures for this highly inflected language. We found that the use of automatically generated keywords and high quality free translation software in English and aggressive stemming techniques for Czech yielded the best results for each of the conditions evaluated.

In the remainder of the paper, we describe our processing and retrieval system in detail. We introduce the baseline monolingual English system and describe the impact of keyword inclusion. We then describe the query processing for the French-English cross-language system and comparative results. We present the Czech retrieval experiments, stemming strategies and impact. Finally, we conclude with some overall observations as well as plans for future work.

Query	ASR	ASR & Keyword
Mono	0.0512	0.0571

Table 1: Monolingual retrieval on English spoken documents, with and without automatic keywords.

2 English Baseline System

We describe the query formulation, document creation, and retrieval processing for the monolingual English retrieval system.

2.1 Query Formulation

All retrieval experiments employed the title and description fields of the original topic specifications. The components were simply concatenated together, employing the weighted sum operator (#sum) from the InQuery system, with default stemming and stopword removal. No additional removal of stop structure was performed.

2.2 Document Creation

Two document formulations formed a primary contrast for monolingual and cross-language retrieval for English documents, one configuration included automatically generated keywords and the other did not. In both cases, we employed the manual document segmentation provided by the track organizers and the ASRTEXT2006B automatic speech recognition output field as the core document representation, based on its prior assessed effectiveness.[4] In the automatic keyword condition, all automatically generated keywords (AK1 and AK2) were added to the core document representation by concatenation.

2.3 Retrieval Engine

For the experiments reported below, we used the InQuery information retrieval system (version 3.1p1) [1] developed at the University of Massachusetts, Amherst, with a design motivated by inference networks, which normalizes the individual term weights when they are computed and then uses an unnormalized inner product to produce retrieval status values. The documents are then sorted in order of decreasing retrieval status value to form a list in an order that approximates a decreasing degree of relevance to the searcher's query. For English, stopwords were removed based on the default stopword list, and stemming was performed with the default *kstem* algorithm. We adopt the convention that values of $p < 0.05$ for a Wilcoxon signed ranks test on a pair of retrieval results is considered significant.

2.4 Monolingual English Results

The baseline ASR transcript only yields a mean average precision of 0.0512. When augmented with automatically generated keywords, the results rise to 0.0571. The difference does not quite reached significance (p <= 0.5068) by Wilcoxon Signed-Ranks test. The automatically generated keywords enrich the document representation with topical terms which appear to enhance retrieval and reduce the effects of the variance of the noisy output of the recognition of the automatic recognition of spontaneous speech.

3 French-English Cross-language Retrieval

For the cross-language retrieval conditions, the official runs employed a publicly available translation tool, while contrastive runs employed a dictionary-based word-for-word translation strategy,

Query	ASR	ASR & Keyword
Mono	0.0512	0.0571
FR + G	0.0322	0.0406
FR + wd	0.0256	0.0241

Table 2: Mono- and cross-language retrieval on English documents, with and without automatic keywords.

with a freely available dictionary and statistically derived stemming. We discuss the query translation procedure below; all document processing and retrieval components were identical to the monolingual English configuration.

3.1 Query Translation

For the official cross-language French-English retrieval runs, we employed the publicly available translation tool provided by Google (http://translate.google.com) to translate the queries. Queries were created by concatenating the title and description fields of the French topics, analogous to the procedure used for English query formulation.

For contrastive runs, we employed a dictionary-based word-for-word translation procedure consistent with [2]. We obtained a freely available French-English bilingual term list from http://www.freedict.com. For the word-for-word translation process, all terms are first converted to lowercase and all accent diacritics are removed for consistency with the translation resource. Next, the translation procedure applies a backoff stemming strategy [5], to support matching with highest precision between the query terms and the dictionary, but backing off to stemmed forms to enhance recall. We attempt initially to match the unstemmed forms in the query with unstemmed forms in the bilingual term list. Only if no match is found, do we perform stemming, attempting to match the stemmed query term in the unstemmed term list, the surface form of the query in the stemmed term list, and finally the stemmed query term in the stemmed term list. The stemming procedure employed stemming rules derived by a statistical stemming process as in [3]. 27% of the query terms remained untranslated, and all untranslatable terms were retained.

In both cases, the resulting English translations are stemmed and stopwords are removed, consistent with earlier document processing.

3.2 Cross-language Retrieval Results & Discussion

We find a substantial drop in mean average precision from the monolingual to the cross-language conditions, and from system-based to dictionary-based translation. Results appear in Table 2. With comparable document representations, effectiveness for the cross-language condition drops 29-37% from monolingual levels for comparable document representations. A larger drop is observed for the baseline condition without automatically generated keywords than for that with keywords. Furthermore, the drop in retrieval effectiveness is even more pronounced in the word-for-word case, and is low enough that the ASR-based retrieval is very similar to keyword-based retrieval.

These results indicate overall good effectiveness for the online translation tool for the French-English language pair and the limitations of the small dictionary-based translation strategy. However, the degradation from monolingual retrieval is quite large, and alternate strategies will be needed to overcome it. The less formal and highly variable character of the spontaneous speech materials limits the effectiveness of retrieval on ASR transcripts alone, while enrichment with automatically generated keywords appears to provide more useful representations of topical information, though differences to not reach significance. Additional strategies for enrichment and denoising of the query and document content will be necessary to overcome these challenges.

Query	No Stem	Light Stem	Aggressive Stem
Czech TD	0.0126	0.0189	0.0203

Table 3: Czech monolingual speech retrieval results with different stemming strategies.

4 Monolingual Czech Retrieval: Stemming Strategies

The Czech language poses special challenges for information retrieval. In contrast with English which has a relatively impoverished morphology, Czech employs a very rich morphology. As a result, to support effective matching between the document content and the specification of the user's information need, some means of overcoming the surface variance due to morphology is required. Here we explore the impact on retrieval effectiveness of three different stemming strategies: no stemming, light stemming, and aggressive stemming. We employ two freely available java-based Czech stemmers from the University of Neuchatel (http://members.unine.ch/jacques.savoy/clef/index.html) to perform light and aggressive stemming for Czech. "Light" stemming in these cases refers to removing affixes only for nouns and adjectives.

The basic query formulation, indexing, and retrieval processes are consistent with those described above for monolingual English with three contrasts, in stemming, stopword removal, and document formation. First, we apply one of the three stemming strategies - none, light, or aggressive - to both queries and documents to enhance matching and retrieval in this morphologically rich language. We also incorporate a freely available Czech stopword list from the same source as the stemmers, to support stopword removal. [1] Finally, we employ the playback point based document segmentation provided by the track organizers. Results thus employ the mGAP measure for Czech retrieval [4].

4.1 Results and Discussion

We find that the best results are obtained with the aggressive stemming strategy, followed by light stemming, and lastly no stemming. All results appear in Table 3. The results for aggressive stemming were the second best title+description based results for Czech in this evaluation. Clearly, the stemming approaches more effectively overcome the high degree of surface form variation of Czech terms. The unstemmed case is significantly outperformed by both the aggressive stemmer ($p <= 0.002$) and the light stemmer ($p <= 0.01$), although the two stemmers are not significantly different from each other.

However, the overall effectiveness of spontaneous speech retrieval in Czech is still quite limited. We speculate that not only must term matching be improved, for example through improved stemming and enhanced retrieval through pseudo-relevance feedback query or document expansion, but also through improvements in basic transcription accuracy and handling of spontaneous speech phenomena.

5 Conclusions and Future Work

Experiments in monolingual English and cross-language French-English speech retrieval obtained the best results by augmenting ASR transcripts with automatically generated keywords. Experiments in monolingual Czech speech retrieval demonstrated the importance of stemming to overcome surface form variation for this morphologically rich language.

However, retrieval from spontaneous speech in both languages remains a very challenging task due both to the difficulties of speech recognition and to the challenging structure of spontaneous speech. In future work we plan to explore approaches to minimize the impact of speech recognition errors and variations in lexical choice through denoising strategies such as Generalized Latent Semantic Analysis to enhance document and query similarity even without lexical overlap. Also,

[1] For compatibility with the retrieval system, we also removal all accent diacritics after stemming.

while the current work has only employed the document segmentations provided, we hope to explore novel approaches to automatic segmentation of spontaneous speech which is important for broadening access to lengthy recorded materials and also poses many interesting challenges to integrating lexical and acoustic evidence of structure in spontaneous speech.

References

[1] James P. Callan, W. Bruce Croft, and Stephen M. Harding. The INQUERY retrieval system. In *Proceedings of the Third International Conference on Database and Expert Systems Applications*, pages 78–83. Springer-Verlag, 1992.

[2] Gina-Anne Levow, Douglas W. Oard, and Philip Resnik. Dictionary-based techniques for cross-language information retrieval. *Information Processing and Management: Special Issue on Cross-language Information Retrieval*, 41(4), 2005.

[3] D. W. Oard, G.-A. Levow, and C. I. Cabezas. *CLEF Experiments at Maryland: Statistical Stemming and Backoff Translation.* Springer, 2001.

[4] D. W. Oard, Jianqiang Wang, Gareth J.F. Jones, Ryan W. White, Pavel Pecina, Dagobert Soergel, Xiaoli Huang, and Izhak Shafran. Overview of the clef-2006 cross-language speech retrieval track. In *CLEF-2006*, 2006.

[5] Philip Resnik, Douglas W. Oard, and Gina-Anne Levow. Improved cross-language retrieval using backoff translation. In *Proceedings of Human Language Technology Conference (HLT) 2001*, pages 153–155, 2001.