Waterloo Experiments for the CLEF05 SDR Track

Charles L. A. Clarke
School of Computer Science, University of Waterloo, Canada
claclark@plg.uwaterloo.ca

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval

General Terms

Measurement, Performance, Experimentation

Keywords

Spoken Data Retrieval, Okapi, Query Expansion

1 Introduction

This year is the first year that the Information Retrieval Group at the University of Waterloo participated in CLEF. For the Cross-Language Spoken Document Retrieval track we submitted five official runs — three English automatic runs (title-only, title+desc, and title+desc+narr), a Czech automatic run (title-only) and a French automatic run (title-only). All official runs used a combination of several query formulation and expansion techniques, including phonetic n-grams and pseudo-relevance feedback expansion over a topic-specific external corpus crawled from the Web. In addition, a large number of un-official runs were generated, including German and Spanish runs. This brief report provides an overview of our experiments, which are summarized in figure 1.

2 Retrieval Methods

All our runs were generated by the Wumpus retrieval system[1] using Okapi BM25 as the basic retrieval method.

The Wumpus implementation of Okapi BM25 is a variant of the formula given by Robertson et al. [3]. Given a term set Q, a document d is assigned the score:

$$\sum_{t \in Q} q_t \cdot log\,(D/D_t)\,\frac{(k_1+1)d_t}{K+d_t} \tag{1}$$

where

$$\begin{aligned} D &= \text{number of documents in the corpus} \\ D_t &= \text{number of documents containing } t \\ q_t &= \text{frequency that } t \text{ occurs in the topic} \\ d_t &= \text{frequency that } t \text{ occurs in } d \\ K &= k_1((1-b)+b \cdot l_d/l_{avg}) \end{aligned}$$

[1] www.wumpus-search.org

$$\begin{aligned} l_d &= \text{length of } d \\ l_{avg} &= \text{average document length} \end{aligned}$$

All CLEF 2005 runs used parameter settings of $k_1 = 1.2$ and $b = 0.75$.

Many of our runs incorporated pseudo-relevance feedback, following the process described in Yeung et al. [1]. For feedback purposes, we augmented the CLEF 2005 SDR corpus with a 2.5GB corpus of Web data, generated by a topic-focused crawl, seeded from 17 sites dedicated to the holocaust. Each query was first executed against this augmented corpus. Terms were extracted from the top results and added to the initial query, which was then executed against the SDR Corpus.

As an alternative to stemming, many runs were based on phoneme 4-grams. For these runs, NIST's text-to-phone tool[2] was applied to translate the words in the corpus into phoneme sequences, which were then split into 4-grams and indexed. Queries were pre-processed in a similar fashion before execution.

Several runs, including our official English-language submissions, were generated by fusing word and n-gram runs. For these runs, fusion was performed using the standard CombMNZ algorithm [2].

Our non-English runs used translated queries supplied by the University of Ottawa group. The reader should consult their CLEF 2005 paper for further information.

3 Discussion

On the training data, the fusion of feedback and phonetic n-gram runs produced a substantial performance improvement over the baseline Okapi runs. Unfortunately, this the improvement was not seen on the test data, where feedback produced only a modest improvement and the phonetic n-grams generally harmed performance.

Next year, we hope to expand our participation in CLEF, including the evaluation of additional speech-specific techniques in the context of the SDR track.

References

[1] David L. Yeung, Charles L. A. Clarke, Gordon V. Cormack, Thomas R. Lynam, and Egidio L. Terra. Task-Specific Query Expansion (MultiText Experiments for TREC 2003). In *Twelfth Text REtrieval Conference*. National Institute of Standards and Technology, 2003.

[2] E. A. Fox and J. A. Shaw. Combination of multiple searches. In *Second Text REtrieval Conference*. National Institute of Standards and Technology, 1994.

[3] S. E. Robertson, S. Walker, and M. Beaulieu. Okapi at TREC-7: Automatic ad hoc, filtering, VLC and interactive track. In *Seventh Text REtrieval Conference*. National Institute of Standards and Technology, 1998.

[2] `www.nist.gov/speech/tools/`

Lang	run	map	bpref	Fields	Description
E	uw5XET	0.090	0.113	T	stemming, no feedback
E	uw5XETD	0.099	0.128	TD	stemming, no feedback
E	uw5XETDN	0.116	0.147	TDN	stemming, no feedback
E	uw5XETfb	0.100	0.127	T	stemming, feedback
E	uw5XETDfb	0.110	0.140	TD	stemming, feedback
E	uw5XETDNfb	0.116	0.142	TDN	stemming, feedback
E	uw5XETph	0.087	0.114	T	phonetic 4-grams, no feedback
E	uw5XETDph	0.097	0.120	TD	phonetic 4-grams, no feedback
E	**uw5XETfs**	0.098	0.127	T	fusion of uw5XETfb and uw5XETph
E	**uw5XETDfs**	0.112	0.139	TD	fusion of uw5XETDfb and uw5XETDph
E	**uw5XETDNfs**	0.114	0.141	TDN	fusion of uw5XETDNfb and uw5XETph
C	uw5XCT	0.039	0.061	T	stemming, no feedback
C	uw5XCTD	0.054	0.091	TD	stemming, no feedback
C	**uw5XCTph**	0.047	0.093	T	phonetic 4-grams, no feedback
C	uw5XCTDph	0.055	0.095	TD	phonetic 4-grams, no feedback
F	uw5XFT	0.094	0.121	T	stemming, no feedback
F	uw5XFTD	0.108	0.137	TD	stemming, no feedback
F	**uw5XFTph**	0.085	0.116	T	phonetic 4-grams, no feedback
F	uw5XFTDph	0.101	0.122	TD	phonetic 4-grams, no feedback
G	uw5XGT	0.079	0.112	T	stemming, no feedback
G	uw5XGTD	0.077	0.112	TD	stemming, no feedback
G	uw5XGTph	0.064	0.105	T	phonetic 4-grams, no feedback
G	uw5XGTDph	0.072	0.108	TD	phonetic 4-grams, no feedback
S	uw5XST	0.087	0.109	T	stemming, no feedback
S	uw5XSTD	0.092	0.121	TD	stemming, no feedback
S	uw5XSTph	0.086	0.122	T	phonetic 4-grams, no feedback
S	uw5XSTDph	0.095	0.117	TD	phonetic 4-grams, no feedback
E	uw5XMT	0.224	0.224	T	MANUAL FIELDS, stemming, no feedback
E	uw5XMTD	0.235	0.243	TD	MANUAL FIELDS, stemming, no feedback
E	uw5XMTDN	0.251	0.260	TDN	MANUAL FIELDS, stemming, no feedback
E	uw5XMTfb	0.226	0.244	T	MANUAL FIELDS, stemming, feedback
E	uw5XMTDfb	0.258	0.264	TD	MANUAL FIELDS, stemming, feedback
E	uw5XMTDNfb	0.255	0.270	TDN	MANUAL FIELDS, stemming, feedback

Figure 1: Summary of runs and results. The name of submitted runs appear in **boldface**.