UB at CLEF2004: Part 1
Monolingual and Multilingual tasks

Miguel E. Ruiz Munirathnam Srikanth
State University of New York at Buffalo
{meruiz,srikanth}@buffalo.edu

Abstract

This paper presents the results of the University at Buffalo in CLEF 2004. Our efforts concentrated in two main tasks: Multilingual task using English as the initial query language, and medical image retrieval. The paper has been divided into two parts that will appear in the respective areas of the proceedings of workshop. Our Adhoc retrieval work used the TAPIR toolkit developed in house by the second author. Our approach focused on the validation and adaptation of the language model system to work in a multilingual environment and in exploring ways to merge results from multiple collections into a single list of results. The second part of the paper describes our experiments in multilingual image retrieval. Our work in image retrieval explores automatic query expansion using pseudo relevance feedback on the case descriptions to improve ranking of images retrieved by a CBIR system. Our results are quite good and show significant improvements with respect to our baseline.

1 Introduction

For CLEF 2004 we participated in the adhoc mono and multilingual retrieval as well as in the medical image retrieval. The goal of our participation in the adhoc retrieval task is to explore language modeling approaches for retrieval from non-English collections using the TAPIR (Text Analysis and Processing for Information Retrieval) toolkit which was originally developed for English and later modified to support ISO-Latin-1 encoding and Porter stemmers for European languages. Part one of this paper will present in detail our results for the monolingual retrieval task in French, Finnish and Russian, as well as the multilingual task using English queries to retrieve information in English, Finnish, French and Russian collections.

The second part of this paper presents our results for the medical image retrieval task. In this track our goal is to improve image retrieval by using retrieval feedback on the related case descriptions to re-rank the images retrieved by a CBIR system. Because our Language model system did not support retrieval feedback (which is a feature that was still under development by the time we worked on this task) we decided to use a version of the SMART retrieval system that we used in our participation in CLEF 2003.

2 Multilingual Task

The two step process of mono-lingual retrieval followed by result combination was used in our multilingual submissions. The language modeling approach to information retrieval using smoothed unigram models was experimented with for both mono-lingual and multi-lingual experiments. For multilingual retrieval, the topics were translated using Intertran translation system [1] from English to the other three languages (French, Finnish and Russian) defined in the task. The translated

[1] www.intertran.com

queries are used queried against a search index for the corresponding language. Our experiments concentrated on techniques to combine the mono-lingual retrieval results for multilingual task. Monolingual retrieval was done using statistical language modeling approaches discussed briefly in the next section.

2.1 Monolingual Retrieval using Statistical Language Models

Statistical language models have been shown to be very effective for document retrieval. Experiments in English document collections have shown significant improvements over traditional vector space and probabilistic models. A language model is a probability distribution defined on strings of an alphabet. A language model is associated with a document in the document collection to indicate or capture its unique properties. Given a query, Q, the documents are ranked based on the likelihood of their language model generating the query,$P(Q|M_d)$ [2]. The query-likelihood probability is estimated using smoothed unigram language models.

$$P(Q|M_d) = \prod_i P(q_i|M_d) \tag{1}$$

The query term probability is estimated from document and corpus counts of the query term smoothed using Dirichlet priors. In Bayesian smoothing using Dirichlet priors, the language model is assumed to be multinomial with the conjugate prior for Bayesian analysis as the Dirichlet distribution $\{\mu P_C(w_i)\}$. The Dirichlet prior smoothed term probability is given by

$$P(w|M_D) = \frac{n(w,d) + \mu p_C(w)}{\sum_v n(v,d) + \mu} \tag{2}$$

where μ is the Dirichlet prior parameter, $n(w,d)$ is the count of occurrence of term w in document d. $p_C(w)$ is the corpus probability of term w. A fixed value of $\mu = 1000$ was used in the experiments.

2.2 Results Merging

Different weighting schemes and merging methods have been experimented for multilingual retrieval. Documents are reweighted for multilingual retrieval and ranked based on the reweighted relevance value. We explored in our CLEF 2004 experiments the use of *query ambiguity* or *clarity score* for reweighting documents for multilingual retrieval. Clarity score, proposed by Cronen-Townsend and Croft [3], is defined for a query as a measure of lack of ambiguity in the given query with respect to a document collection. A query language model is generated for a given query based on the word usage in documents relevant to the given query. The simplest query model is a unigram language model based on word counts in documents deemed highly-relevant to the given query. The clarity score is computed as the relative entropy between the query model and the overall collection language model. Using Lavrenko and Croft's Method 1 [1], the query language model is given by

$$P(w|Q) = \sum_{D \in R} P(w|D)P(D|Q) \tag{3}$$

where the summation is over documents deemed highly-relevant to the given query. The top 100 documents returned using the smoothed unigram language model were used as the relevant set in our experiments. The query-likelihood probability, $P(Q|D)$ is estimated using smoothed unigram language model given by Equations 1 and 2. The clarity score is given by the Kullback-Leibler divergence between the query language model and the collection language model,

$$\text{clarity}(Q) = \sum_{w \in V} P(w|Q) log_2 \frac{P(w|Q)}{P_C(w)} \tag{4}$$

The clarity score, $cl(Q,L)$ is computed for each query-language pair. For multilingual experiments, the source language query in English is translated to other languages, the monolingual

retrieval using smoothed unigram language models is performed and the clarity scores for each query-language pair is computed. Result merging uses the clarity score to reweight the relevance status value of the documents from the monolingual results.

The clarity scores can be used "as-is" as weights assigned to different languages for a given query and relevance weight of a document for a given query can be adjusted as

$$RSV_{ASIS}(D,Q,L) = RSV_{mono}(D,Q,L) \times cl(Q,L) \tag{5}$$

However, the clarity score are not comparable across the document collections in different languages. The range of values taken by clarity score depends on the characteristics of the document collection in a particular language and the retrieval performance using such a weighting scheme is expected to match a merging method that uses a fixed multiplier values for a language across queries.

Instead of using the absolute values of the clarity scores for different query-language pairs, we experimented with different normalization methods. The clarity score can be compared and normalized across languages as they correspond to the same query.

$$RSV_{BYLANG}(D,Q,L) = RSV_{mono}(D,Q,L) \times \frac{cl(Q,L)}{\sum_l cl(Q,l)} \tag{6}$$

Normalization can also be performed across queries as the clarity scores were computed for different queries with respect to a document collection in one language.

$$RSV_{BYQUERY}(D,Q,L) = RSV_{mono}(D,Q,L) \times \frac{cl(Q,L)}{\sum_q cl(q,L)} \tag{7}$$

We also experimented with normalizing the clarity scores, first across queries and then across languages (BYQUERYLANG) and also the reverse – normalize first across languages, then across queries (BYLANGQUERY). In all the above reweighting formulas, the relevance status value of a document for a given query is normalized across the documents deemed relevant to the query. This makes the comparison of relevance status values of documents across language for a given query meaningful.

Merging retrieval results using interleaving of documents with same rank has been experimented before for multilingual retrieval. While interleaving the results a fixed order of the languages is selected and documents with the same rank are listed based on the pre-selected order of their respective languages. The language order is usually fixed at random. We experimented with different normalized and unnormalized clarity scores to check if it provides any clues for the order in which documents with same rank can be interleaved.

3 Experimental Results

The document collections for different languages were indexed separately using the TAPIR (Text Analysis and Processing for Information Retrieval) toolkit – an in-house information retrieval system that supports different retrieval models (VSM, language models) and languages. Document and collection statistics along with position information is collected and stored in the indexing system. The TAPIR toolkit was used to perform monolingual retrieval using the original and translated queries using the smoothed unigram retrieval model.

The mono-lingual retrieval results are given in Table 1.

Figure 1 plots the difference between our submitted results and the median average precision values for French, Finnish and Russian. Our submission seems to have been the only submission for monolingual retrieval in English and hence is not included in the figure. Our retrieval results performed well above average or the median values in the French corpus and below average or median values for the Finnish collection. Russian performance seems to be around the median values. These performances correspond to the simple statistical retrieval model with no special linguistic processing other than stemming and stopword removal.

	AvgP.	Recall	R-Prec.	InitPr
English	0.5167	361/375	0.4608	0.7444
French	0.4629	863/915	0.4239	0.7146
Finnish	0.4599	318/413	0.4545	0.6263
Russian	0.2978	78/123	0.2807	0.4947

Table 1: Monolingual retrieval performance using smoothed unigram language models

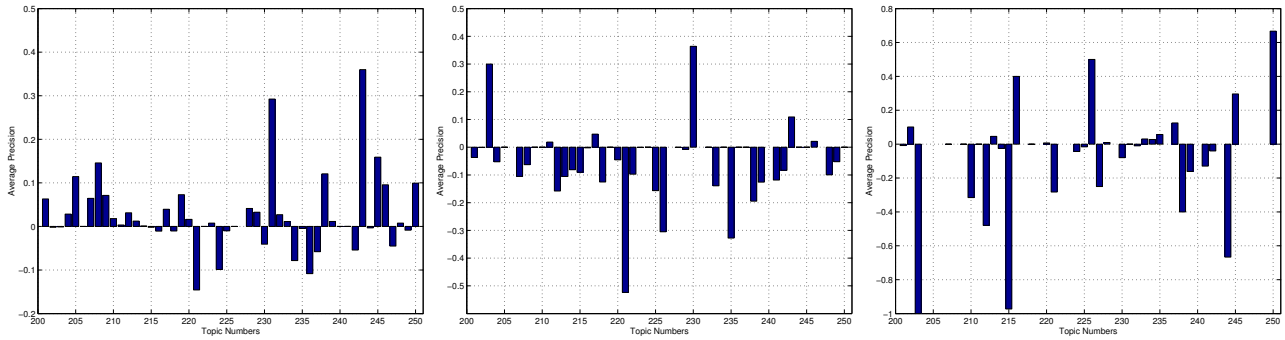

Figure 1: Difference in Average Precision comparing our official submission and median average precision values for French, Finnish and Russian respectively.

Two runs were submitted as official runs for CLEF2004 multilingual retrieval task[2]. These correspond to merging documents by weighting their relevance status value using the clarity score (ASIS) and score reweighting using normalized relevance score, where the normalization is first performed across languages and then across queries (BYLANGQUERY). Table 2 includes performance metrics for the two official submitted runs (ASIS and BYLANGQUERY) and other experimental runs using different normalization conditions for clarity scores.

	AvgP.	Recall	R-Prec.	InitPr
ASIS	0.1453	1135/1826	0.1884	0.4771
BYLANGQUERY	0.1709	1092/1826	0.2003	0.4635
BYLANG	0.1711	1092/1826	0.2003	0.4639
BYQUERY	0.1163	857/1826	0.1501	0.6259
BYQUERYLANG	**0.1769**	1094/1826	0.2043	0.4965

Table 2: Multilingual retrieval performance

Normalizing across languages seems to give a significant improvement to the average precision values. However, normaling the clarity scores both across query and language gives the best performance. It is noted that the monolingual retrieval runs that are combined to obtain these results are based on smoothed unigram language models. The merging strategy is independent of the underlying retrieval model used for mono-lingual retrieval. Improvements using better query representation and relevance feedback for monolingual retrieval is expected to reflect positively on multilingual retrieval results. The clarity scores provides some clues on weighting the monolingual results. Appropriate methods need to be devised to incorporate such clues in the re-weighting document scores. We intend to explore such methods in future.

We experimented with merging using interleaving of documents with same rank. Table 3 gives the performance of different interleaving options. The different runs correspond to the selection of

[2]We submitted a third run (UBmulti03) that was mistakenly labeled as automatic but was actually a run that combined the manual translations and will not be considered in the analysis.

the order in which documents from different monolingual retrieval with same rank are selected. In IL-ASIS a fixed order of languages is used and results are interleaved. In IL-RANDOM, the order is randomly selected for each query. Random selection looks better that IL-ASIS based on the metrics in the table. However, it corresponds to one particular run and one can expect the performance measures to take values around the IL-ASIS performance values. There is not justification for either of these selection methods. The last four entries correspond to the language order selection based on normalized clarity scores. For a given query, the clarity scores are normalized either across languages or across queries or both. The language order is decided based on the ranking of the normalized clarity scores for a given query. Normalizing the clarity scores across queries and use them to decide the language order for interleaving seems to provide best performance for CLEF2004 topics.

	AvgP.	Recall	R-Prec.	InitPr
IL-ASIS	0.1381	1156/1826	0.1785	0.4657
IL-RANDOM	0.1443	1156/1826	0.1771	0.5262
IL-BYLANG	0.1489	1156/1826	0.1787	0.5205
IL-BYQUERY	**0.1553**	1156/1826	0.1794	0.6343
IL-BYLANGQUERY	0.1489	1156/1826	0.1787	0.5205
IL-BYQUERYLANG	0.1508	1156/1826	0.1787	0.5394

Table 3: Multilingual retrieval performance - merging results using interleaving of ranked documents

While the performance of interleaving using clarity scores to select the language order does not perform as well as using the clarity scores directly as multipliers for the document weights for multilingual retrieval, it can be used as a metric for the results merging process. The improvement in average precision of interleaving using clarity scores normalized across queries is statistically significant than using a fixed language order.

4 Conclusion and future work

In an effort to build the baseline multilingual retrieval systems, we extended an IR system developed to work with English document collections to handle non-English document collections and performed monolingual retrieval on English, French, Finnish and Russian using a smoothed unigram language model. For multilingual task we experimented with clarity scores of the queries in their document collections for merging the results of monolingual retrieval. Clarity score was used as a multiplier for the document weight as well as a mechanism to determine language order in the case of merging using interleaving of documents at the same rank. Using clarity scores as multipliers to reweight the document scores improves retrieval performances. Appropriate methods to incorporate the clues provided by clarity scores towards improving retrieval needs to be investigated and this is one of the areas of our future work.

References

[1] Lavrenko, V. and Croft, W. B. Relevance-based Language Models. In *Proceedings of SIGIR'01*, pages 120–127. ACM, New York, 2001.

[2] Ponte, J. M. and Croft, W. B. A language modeling approach to information retrieval. In *Proceedings of SIGIR'98*, pages 275–281. ACM, New York, 1998.

[3] Cronen-Townsend, S. and Croft, W. B. Quantifying Query Ambiguity. In *Proceedings of HLT'02*, 2002