Multilingual Query Expansion for CLEF Adhoc-TEL

Ray R. Larson
School of Information
University of California, Berkeley, USA
ray@sims.berkeley.edu

Abstract

In this paper we will briefly describe the approaches taken by the Cheshire (Berkeley) Group for the CLEF Adhoc-TEL 2009 tasks (Mono and Bilingual retrieval). Recognizing that many potentially relevant documents in each of the TEL sub-collections are in other languages, we tried to use multiple translations of the topics for searching each subcollection, combined into a single query. Overall this strategy performed very poorly compared to the the basic monolingual approach used last year (and repeated for one run in each language this year). We haven't yet completed our analysis of the reasons for this (we suspect that results were evaluated expecting the retrieved items to also be in the same language as the topic).

Once again this year we used probabilistic text retrieval based on logistic regression and incorporating blind relevance feedback for all of the runs. All translation for bilingual tasks was performed using the LEC Power Translator PC-based MT system. Our results this year, however, were surprising poor compared to last year's results. Some testing has shown that, for some cases, unexpected hyphenations in the machine translation and untranslated words were to blame. It may also be the case that others have significantly improved their approaches for this task.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval

General Terms

Algorithms, Performance, Measurement

Keywords

Cheshire II, Logistic Regression

1 Introduction

Each the collections used in the CLEF Adhoc TEL track are considered to be "mainly" in a particular language (English for BL, French for BNF, and German for ONB), according to the language codes of the records, only about half of each collection was in that main language, with virtually all other languages represented by one or more entries in one or another of the collections. German, French, English, and Spanish records were available in all of collections. This overlap of languages presents an interesting multilingual search (and evaluation) problem, and we attempted

to address it this year by using tranlations of topics into each of the other languages and combining those translations with the original topic in some of our submissions.

This paper concentrates on the retrieval algorithms and evaluation results for Berkeley's official submissions for the Adhoc-TEL 2008 track. All of the runs were automatic without manual intervention in the queries (or translations). We submitted nine Monolingual runs (three German, three English, and three French) and 12 Bilingual runs (four for each target language German, English and French, with both expanded and unexpanded topics).

This paper first describes the retrieval algorithms used for our submissions, followed by a discussion of the processing used for the runs. We then examine the results obtained for our official runs, and finally present conclusions and future directions for Adhoc-TEL participation.

2 The Retrieval Algorithms

Note that this section is virtually identical to one that appears in our papers from previous CLEF participation and appears here for reference only[8, 7] The basic form and variables of the *Logistic Regression* (LR) algorithm used for all of our submissions was originally developed by Cooper, et al. [5]. As originally formulated, the LR model of probabilistic IR attempts to estimate the probability of relevance for each document based on a set of statistics about a document collection and a set of queries in combination with a set of weighting coefficients for those statistics. The statistics to be used and the values of the coefficients are obtained from regression analysis of a sample of a collection (or similar test collection) for some set of queries where relevance and non-relevance has been determined. More formally, given a particular query and a particular document in a collection $P(R \mid Q, D)$ is calculated and the documents or components are presented to the user ranked in order of decreasing values of that probability. To avoid invalid probability values, the usual calculation of $P(R \mid Q, D)$ uses the "log odds" of relevance given a set of S statistics, s_i, derived from the query and database, such that:

$$\log O(R \mid Q, D) = b_0 + \sum_{i=1}^{S} b_i s_i \tag{1}$$

where b_0 is the intercept term and the b_i are the coefficients obtained from the regression analysis of the sample collection and relevance judgements. The final ranking is determined by the conversion of the log odds form to probabilities:

$$P(R \mid Q, D) = \frac{e^{\log O(R|Q,D)}}{1 + e^{\log O(R|Q,D)}} \tag{2}$$

2.1 TREC2 Logistic Regression Algorithm

For Adhoc-TEL we used a version the Logistic Regression (LR) algorithm that has been used very successfully in Cross-Language IR by Berkeley researchers for a number of years[3]. The formal definition of the TREC2 Logistic Regression algorithm used is:

$$\begin{aligned}
\log O(R|C,Q) &= log\frac{p(R|C,Q)}{1-p(R|C,Q)} = log\frac{p(R|C,Q)}{p(\overline{R}|C,Q)} \\
&= c_0 + c_1 * \frac{1}{\sqrt{|Q_c|}+1} \sum_{i=1}^{|Q_c|} \frac{qtf_i}{ql+35} \\
&+ c_2 * \frac{1}{\sqrt{|Q_c|}+1} \sum_{i=1}^{|Q_c|} \log \frac{tf_i}{cl+80} \\
&- c_3 * \frac{1}{\sqrt{|Q_c|}+1} \sum_{i=1}^{|Q_c|} \log \frac{ctf_i}{N_t}
\end{aligned} \tag{3}$$

$$+ \quad c_4 * |Q_c|$$

where C denotes a document component (i.e., an indexed part of a document which may be the entire document) and Q a query, R is a relevance variable,

$p(R|C,Q)$ is the probability that document component C is relevant to query Q,

$p(\overline{R}|C,Q)$ the probability that document component C is *not relevant* to query Q, which is 1.0 - $p(R|C,Q)$

$|Q_c|$ is the number of matching terms between a document component and a query,

qtf_i is the within-query frequency of the ith matching term,

tf_i is the within-document frequency of the ith matching term,

ctf_i is the occurrence frequency in a collection of the ith matching term,

ql is query length (i.e., number of terms in a query like $|Q|$ for non-feedback situations),

cl is component length (i.e., number of terms in a component), and

N_t is collection length (i.e., number of terms in a test collection).

c_k are the k coefficients obtained though the regression analysis.

If stopwords are removed from indexing, then ql, cl, and N_t are the query length, document length, and collection length, respectively. If the query terms are re-weighted (in feedback, for example), then qtf_i is no longer the original term frequency, but the new weight, and ql is the sum of the new weight values for the query terms. Note that, unlike the document and collection lengths, query length is the "optimized" relative frequency without first taking the log over the matching terms.

The coefficients were determined by fitting the logistic regression model specified in $\log O(R|C,Q)$ to TREC training data using a statistical software package. The coefficients, c_k, used for our official runs are the same as those described by Chen[1]. These were: $c_0 = -3.51$, $c_1 = 37.4$, $c_2 = 0.330$, $c_3 = 0.1937$ and $c_4 = 0.0929$. Further details on the TREC2 version of the Logistic Regression algorithm may be found in Cooper et al. [4].

2.2 Blind Relevance Feedback

In addition to the direct retrieval of documents using the TREC2 logistic regression algorithm described above, we have implemented a form of "blind relevance feedback" as a supplement to the basic algorithm. The algorithm used for blind feedback was originally developed and described by Chen [2]. Blind relevance feedback has become established in the information retrieval community due to its consistent improvement of initial search results as seen in TREC, CLEF and other retrieval evaluations [6]. The blind feedback algorithm is based on the probabilistic term relevance weighting formula developed by Robertson and Sparck Jones [9].

Blind relevance feedback is typically performed in two stages. First, an initial search using the original topic statement is performed, after which a number of terms are selected from some number of the top-ranked documents (which are presumed to be relevant). The selected terms are then weighted and then merged with the initial query to formulate a new query. Finally the reweighted and expanded query is submitted against the same collection to produce a final ranked list of documents. Obviously there are important choices to be made regarding the number of top-ranked documents to consider, and the number of terms to extract from those documents. For ImageCLEF this year, having no prior data to guide us, we chose to use the top 10 terms from 10 top-ranked documents. The terms were chosen by extracting the document vectors for each of the 10 and computing the Robertson and Sparck Jones term relevance weight for each document. This weight is based on a contingency table where the counts of 4 different conditions for combinations

Table 1: Contingency table for term relevance weighting

	Relevant	Not Relevant	
In doc	R_t	$N_t - R_t$	N_t
Not in doc	$R - R_t$	$N - N_t - R + R_t$	$N - N_t$
	R	$N - R$	N

of (assumed) relevance and whether or not the term is, or is not in a document. Table 1 shows this contingency table.

The relevance weight is calculated using the assumption that the first 10 documents are relevant and all others are not. For each term in these documents the following weight is calculated:

$$w_t = log \frac{\frac{R_t}{R - R_t}}{\frac{N_t - R_t}{N - N_t - R + R_t}} \tag{4}$$

The 10 terms (including those that appeared in the original query) with the highest w_t are selected and added to the original query terms. For the terms not in the original query, the new "term frequency" (qtf_i in main LR equation above) is set to 0.5. Terms that were in the original query, but are not in the top 10 terms are left with their original qtf_i. For terms in the top 10 and in the original query the new qtf_i is set to 1.5 times the original qtf_i for the query. The new query is then processed using the same LR algorithm as shown in Equation 4 and the ranked results returned as the response for that topic.

3 Approaches for Adhoc-TEL

In this section we describe the specific approaches taken for our submitted runs for the Adhoc-TEL task. First we describe the indexing and term extraction methods used, and then the search features we used for the submitted runs.

3.1 Indexing and Term Extraction

The Cheshire II system uses the XML structure of the documents to extract selected portions for indexing and retrieval. Any combination of tags can be used to define the index contents.

Table 2: Cheshire II Indexes for Adhoc-TEL 2006

Name	Description	Content Tags	Used
recid	Document ID	id	no
names	Author Names	dc:creator, dc:contributor	no
title	Item Title	dc:title, dcterms:alternate	no
topic	Content Words	dc:title, dcterms:alternate dc:subject, dc:description	yes
anywhere	Entire record	record	no
date	Date of Pub.	dcterms:issued	no
lang	Language	dc:language	no
subject	Subject terms	dc:subject	no

Table 2 lists the indexes created by the Cheshire II system for the Adhoc-TEL database and the document elements from which the contents of those indexes were extracted. The "Used" column in Table 2 indicates whether or not a particular index was used in the submitted Adhoc-TEL runs. As the table shows we used only the topic index, which contains most of the content-bearing parts

Table 3: Submitted Adhoc-TEL Runs

Run Name	Description	Type	MAP
MODET2FB	Monolingual German	TD auto	0.1478 *
MODET2FBX	Monolingual German +English and French Trans.	TD auto	0.1230
MODET2FB3	Monolingual German +English and French Trans.	TD auto	0.1331
MOENT2FB	Monolingual English	TD auto	0.3267 *
MOENT2FBX	Monolingual English +German and French Trans.	TD auto	0.2224
MOENT2FB3	Monolingual English +German and French Trans.	TD auto	0.2291
MOFRT2FB	Monolingual French	TD auto	0.2070 *
MOFRT2FBX	Monolingual French +German and English Trans.	TD auto	0.1456
MOFRT2FB3	Monolingual French +German and English Trans.	TD auto	0.1684
BIENDET2FB	Bilingual English⇒German	TD auto	0.1031
BIENDET2FBX	Bilingual English⇒German + French and English	TD auto	0.1150 *
BIFRDET2FB	Bilingual French⇒German	TD auto	0.0991
BIFRDET2FBX	Bilingual French⇒German + French and English	TD auto	0.0882
BIDEENT2FB	Bilingual German⇒English	TD auto	0.2238
BIDEENT2FBX	Bilingual German⇒English + German and French	TD auto	0.1598
BIFRENT2FB	Bilingual French⇒English	TD auto	0.2478 *
BIFRENT2FBX	Bilingual French⇒English + French and German	TD auto	0.1666
BIDEFRT2FB	Bilingual German⇒French	TD auto	0.1652
BIDEFRT2FBX	Bilingual German⇒French + German and English	TD auto	0.1131
BIENFRT2FB	Bilingual English⇒French	TD auto	0.1677 *
BIENFRT2FBX	Bilingual English⇒French + English and German	TD auto	0.1365

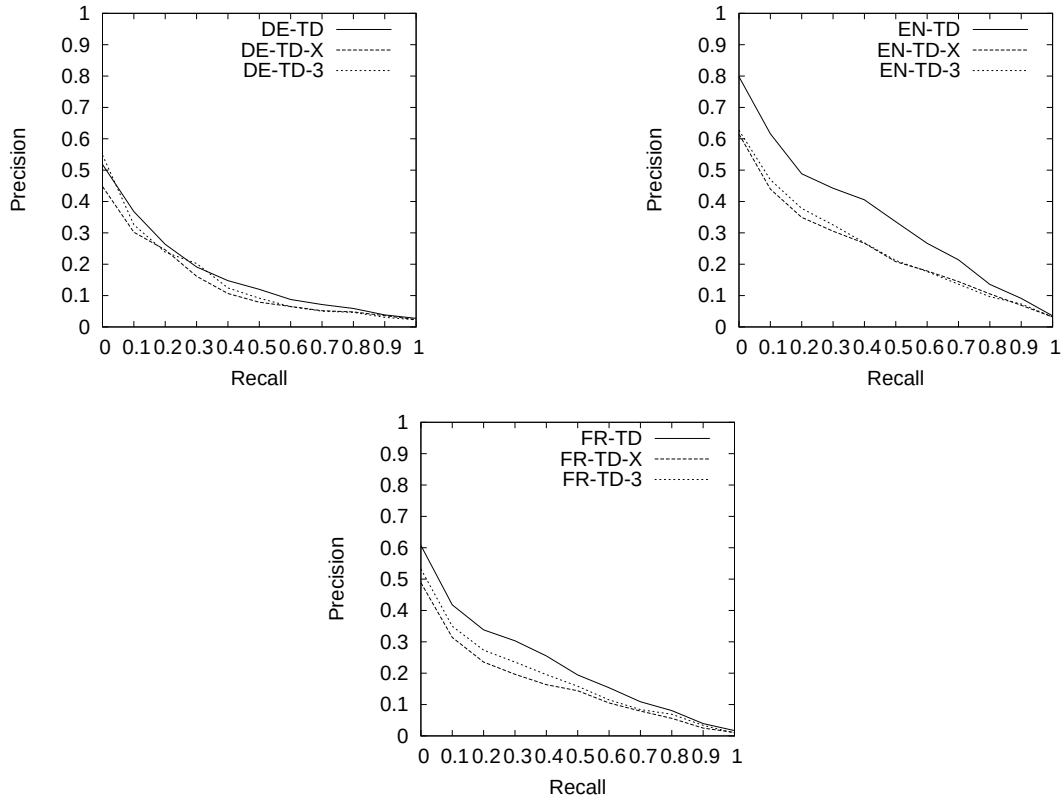
of records, for all of our submitted runs. These tables and the indexes extracted are identical to last year's for Adhoc TEL.

For all indexing we used language-specific stoplists to exclude function words and very common words from the indexing and searching. The German language runs *did not* use decompounding in the indexing and querying processes to generate simple word forms from compounds. The Snowball stemmer was used by Cheshire for language-specific stemming.

3.2 Search Processing

Searching the Adhoc-TEL collection using the Cheshire II system involved using TCL scripts to parse the topics and submit the title and description from the topics. For monolingual search tasks we used the topics in the appropriate language (English, German, and French), for bilingual tasks the topics were translated from the source language to the target language using the LEC Power Translator PC-based machine translation system.

Figure 1: Berkeley Monolingual Runs – German (top left), English (top right) and French (lower)



For query expansion in the monolingual tasks we took two approaches. The first (denoted by an "X" at the end of names in Table 3) used the topic in the specific language as a basis for machine translation to the other main languages (e.g. for English, the English topics were translated to French and German) and the translations were added to the topic. The second (denoted by "3" at the ends of the names in Table 3) used the supplied monolingual topics in the other main languages (e.g., for English, the monolingual French and German topics were added to the English).

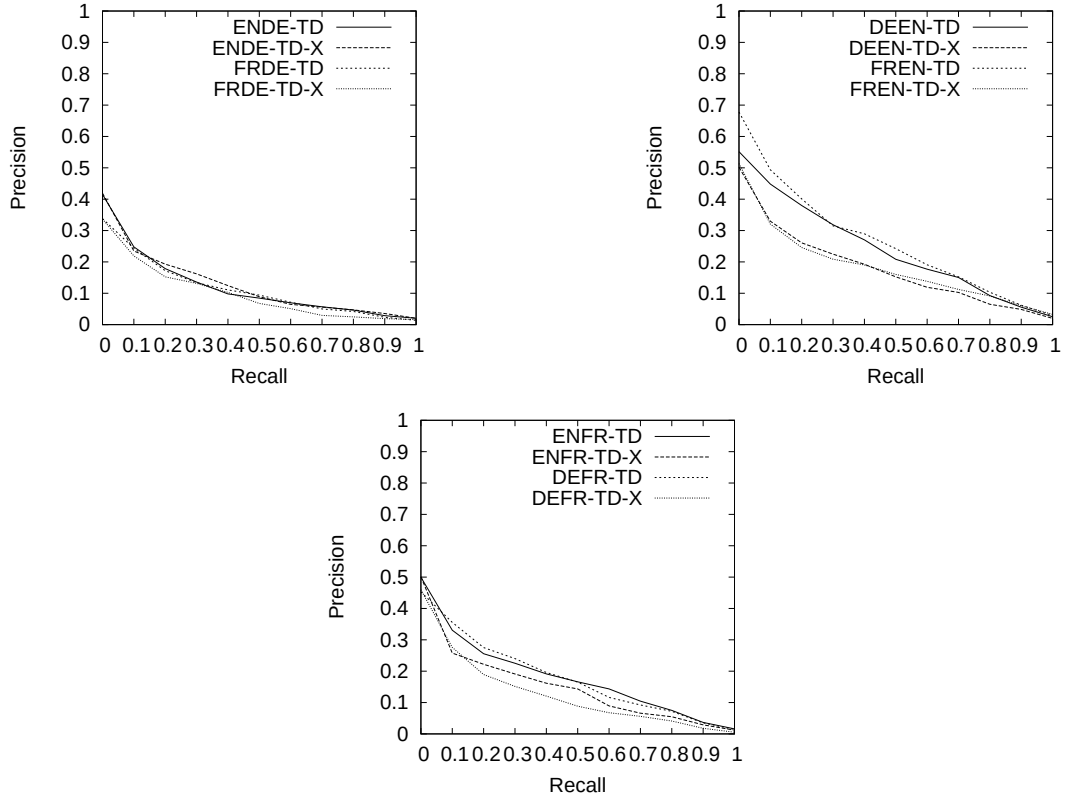
Query expansion in the bilingual tasks (denoted by "X" at the end of the names in Table 3) added the source topics from the translation and an additional translation of the topics to the other main language (e.g., for English topics translated to German, the original English was added to the translated German and an English to French translation was also added). In effect, the expanded monolingual and bilingual topics were actually multilingual topic descriptions.

The scripts for each run submitted the topic elements as they appeared in the topic or expanded topic to the system for TREC2 logistic regression searching with blind feedback. Both the "title" and "description" topic elements were combined into a single probabilistic query and searched using the "topic" index as described in Table 3.

4 Results for Submitted Runs

The summary results (as Mean Average Precision) for the submitted bilingual and monolingual runs for English German and French are shown in Table 3, the Recall-Precision curves for these runs are also shown in Figures 1 (for monolingual) and 2 (for bilingual). In Figures 1 and 2 the

Figure 2: Berkeley Bilingual Runs – To German (top left), To English (top right) and To French (lower)



names for the individual runs represent the language codes, which can easily be compared with full names and descriptions in Table 3 (since each language combination has only a single run).

Table 3 indicates runs that had the highest overall MAP for the task by asterisks next to the run name.

The results in Table 3 show, for the most part, the type of query expansion that we tried was a dismal failure. The only exception was in bilingual German where the expanded English to German topic achieved a very slight performance edge over the unexpanded topic. Overall, we see no benefit to this kind of expansion in the results.

Once again we obtained particularly poor performance in monolingual German, due in part to our lack of support for decompounding (affecting many topics this year).

5 Conclusions

Our overall results this year compared poorly with others, which was a bit of a surprise considering the how the same approach fared last year. We are starting to conduct some analyses to try to determine the causes of variation between last year and this. One very obvious change is that a new version of the MT software was used this time. Because this is a commercial product and new installations replace the old, we cannot do comparative testing directly, but we do have the translations produced last year for last year's topics, so we plan to do a comparison on that basis. One thing that we noticed with this year's topics was that translations from German often had compound terms included in the translation as hyphenated terms (e.g., "color-therapy" for

"Farbentherapie"). To see what effect this might have had in some runs we translated the hyphens in such cases to spaces and reran some experiments. The results of this re-test showed that for the German to English bilingual task we were able to obtain a MAP of 0.2613 compared to 0.2238 in our official results.

Just as this paper was about to be submitted, it occurred to us that the data, unlike the German and other data we had used in other CLEF tracks, was in UTF-8 instead of ISO-8859-1 encoding. We realized that the version of the Snowball stemmer we had used for all of our submitted runs was based on the ISO encoding and not UTF. This could explain the low scores (particularly for French and German) since the stemming process was ineffective and identically inflected stems only were matched in retrieval.

Often it is the simplest things overlooked that lead to problems.

References

[1] Aitao Chen. Multilingual information retrieval using english and chinese queries. In Carol Peters, Martin Braschler, Julio Gonzalo, and Michael Kluck, editors, *Evaluation of Cross-Language Information Retrieval Systems: Second Workshop of the Cross-Language Evaluation Forum, CLEF-2001, Darmstadt, Germany, September 2001*, pages 44–58. Springer Computer Scinece Series LNCS 2406, 2002.

[2] Aitao Chen. *Cross-Language Retrieval Experiments at CLEF 2002*, pages 28–48. Springer (LNCS #2785), 2003.

[3] Aitao Chen and Fredric C. Gey. Multilingual information retrieval using machine translation, relevance feedback and decompounding. *Information Retrieval*, 7:149–182, 2004.

[4] W. S. Cooper, A. Chen, and F. C. Gey. Full Text Retrieval based on Probabilistic Equations with Coefficients fitted by Logistic Regression. In *Text REtrieval Conference (TREC-2)*, pages 57–66, 1994.

[5] William S. Cooper, Fredric C. Gey, and Daniel P. Dabney. Probabilistic retrieval based on staged logistic regression. In *15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, Denmark, June 21-24*, pages 198–210, New York, 1992. ACM.

[6] Ray R. Larson. Probabilistic retrieval, component fusion and blind feedback for XML retrieval. In *INEX 2005*, pages 225–239. Springer (Lecture Notes in Computer Science, LNCS 3977), 2006.

[7] Ray R. Larson. Cheshire at geoclef 2007: Retesting text retrieval baselines. In *8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, LNCS 5152, pages 811–814, Budapest, Hungary, September 2008.

[8] Ray R. Larson. Experiments in classification clustering and thesaurus expansion for domain specific cross-language retrieval. In *8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, LNCS 5152, pages 188–195, Budapest, Hungary, September 2008.

[9] S. E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, pages 129–146, May–June 1976.