Pseudo-Relevance Feedback for CLEF-CHiC Adhoc

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 CHiC Adhoc tasks (Monolingual, Bilingual and Multilingual retrieval for English, French and German). We used multiple translations of the topics for searching each of the CHiC Europeana English, French and German subcollections, employing Google Translate as our translation system. In addition we combined the original topics for various multilingual runs.
Once again this year our approach was to use probabilistic text retrieval based on logistic regression and incorporating pseudo relevance feedback for all of the runs.
The results overall, when viewed using the multilingual qrels based on the entire set of languages for the CHiC collection, were not good, while the individual monolingual runs using only the collection-specific qrels, appear to have performed reasonably well. There is some question about the qrels for the the entire multilingual collection, since there appear to be no relevant documents at all from the English collection.

1 Introduction

The collections used for the Cultural Heritage in CLEF (CHiC) track are derived from the Europeana Digital Library and contain metadata describing materials ranging from text documents to photographs, museum objects, historic locations and persons.

Each the collections used in the CLEF Adhoc CHiC track are considered to be "mainly" in a particular language (at least in the English, French, and German collections), according to the language codes specified the in the language attribute of the metadata tag, however records also included descriptive metadata in virtually all other languages as part of the "enrichment:concept_label" elements. Thus, and English record would also include concept names in other languages as well. This overlap of languages presents an interesting multilingual search problem, and we explored it by using tranlations of topics into each of the other languages (we used only English, French and German for this) and combining those translations with the original topics in some of our submissions.

In this paper we review the retrieval algorithms and evaluation results for Berkeley's official submissions for the Adhoc-CHiC 2013 track. All of the runs were automatic without manual intervention in the queries (or translations). We submitted ten Multilingual runs (using various combinations of query languages and searching the English, French and German collections), six Bilingual runs (two for each target language German, English and French) and three monolingual runs for the three target languages.

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-CHiC 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-CHiC 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:

$$
\begin{aligned}
\log O(R|C,Q) = log\frac{p(R|C,Q)}{1-p(R|C,Q)} &= log\frac{p(R|C,Q)}{p(\overline{R}|C,Q)} \\
&= c_0 + c_1 * \frac{1}{\sqrt{|Q_c|}+1}\sum_{i=1}^{|Q_c|}\frac{qtf_i}{ql+35} \\
&+ c_2 * \frac{1}{\sqrt{|Q_c|}+1}\sum_{i=1}^{|Q_c|}\log\frac{tf_i}{cl+80} \\
&- c_3 * \frac{1}{\sqrt{|Q_c|}+1}\sum_{i=1}^{|Q_c|}\log\frac{ctf_i}{N_t} \\
&+ c_4 * |Q_c|
\end{aligned} \tag{3}
$$

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 Pseudo Relevance Feedback

Instead of performing direct retrieval of documents using the TREC2 logistic regression algorithm described above, we have implmented a "pseudo relevance feedback" step as part of the search processing. The algorithm used for pseudo relevance feedback was by Chen [2] for an earlier search system used in CLEF. Pseudo relevance feedback has become well-known in the information retrieval community primarily because of its apparent ability to provide consistent improvements over initial search results as seen in TREC, CLEF and other retrieval evaluations [6]. While the most commonly used algorithm for pseudo relevance feedback is the Rocchio algorithm originally developed for the SMART system, we have adopted a probabilistic approach based on the probabilistic term relevance weighting formula developed by Robertson and Sparck Jones [9].

Pseudo 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 CHiC this year, we chose to use the top 10 terms from 10 top-ranked documents, since these parameters have worked well in similar evaluations. 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 of (assumed) relevance and whether or not the term is, or is not in a document. Table 1 shows this contingency table.

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

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 CHiC Adhoc

In this section we describe the specific approaches taken for our submitted runs for the CHiC Adhoc 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-CHiC 2006

Name	Description	Content Tags	Used
URI	Europeana URI	europeana:uri dc:identifier	no
contributor	Contributing Inst.	dc:contributor	no
subject	Topics	dc:subject	no
title	Title	dc:title dc:subject, dc:description	no
topic	Entire record	ims:fields	yes

Table 2 lists the indexes created by the Cheshire II system for the three Adhoc-CHiC databases 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-CHiC runs. As the table shows we used only the topic index, which contains most of the content-bearing parts of records, for all of our submitted runs. In addition to the databases for the individual language collection, we used a Cheshire II feature that allows searching across multiple databases and merging the ranked results for searches of all three languages.

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.

Table 3. Mean Average Precision for Different QRels

RunID	EN	FR	DE	ENFRDE	ML
BerkMonoEN01	0.1842*	0.0000	0.0000	0.0645	0.0375
BerkMonoFR02	0.0000	0.2014*	0.0000	0.0648	0.0325
BerkMonoDE03	0.0000	0.0000	0.1757*	0.0857*	0.0440*
BerkBiDEEN04	0.1942*	0.0000	0.0000	0.0641	0.0371*
BerkBiFREN05	0.1744	0.0000	0.0000	0.0579	0.0333
BerkBiFRDE06	0.0000	0.0000	0.1474	0.0679	0.0330
BerkBiENFR07	0.0000	0.1608*	0.0000	0.0555	0.0284
BerkBiDEFR08	0.0000	0.1296	0.0000	0.0391	0.0217
BerkBiENDE09	0.0000	0.0000	0.1785*	0.0726*	0.0331
BerkMLEN10	0.0000	0.0274	0.0379	0.0350	0.0166
BerkMLFR11	0.0000	0.0896*	0.0185	0.0474	0.0231
BerkMLDE12	0.0000	0.0205	0.1260*	0.0734	0.0385
BerkMLDU13	0.0000	0.0244	0.0368	0.0337	0.0198
BerkMLFI14	0.0000	0.0369	0.0336	0.0406	0.0273
BerkMLIT15	0.0000	0.0222	0.0196	0.0207	0.0123
BerkMLSP16	0.0000	0.0200	0.0254	0.0260	0.0148
BerkMLALL17	0.0000	0.0485	0.0758	0.0736	0.0357
BerkMLSPENFRDEIT18	0.0000	0.0597	0.0673	0.0737	0.0353
BerkMLENFRDE19	0.0000	0.0642	0.0763	0.0803*	0.0393*

3.2 Search Processing

Searching the Adhoc-CHiC collection using the Cheshire II system involved using TCL scripts to parse the topics and submit the title element from the topics. For monolingual search tasks we used the topics in the appropriate language (English, German, and French) and searched the specific language collection.

For bilingual tasks the topics were translated from the source language to the target language using Google Translate (this involved converting the original topics to HTML (as if the set of topics was a web page) and then asking Google Translate to translate the web page, then the translated page was converted back to the topic schema, again the matching language database for the translated topics was searched.

For multilingual search tasks we used the original topics, but searched that topic (without translation) in all three (EN, FR, DE) databases and combined the results based on MINMAX normalized scores. The topics used include additional languages other than the main language of the database (e.g., Dutch, Finnish and Italian) without translation, and combinations of multiple languages for a given topic. The

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 pseudo feedback. Only the "title" topic element was used, and where appropriate multiple titles in different languages were combined into a single probabilistic query and searched using the "topic" index as described in Table 2.

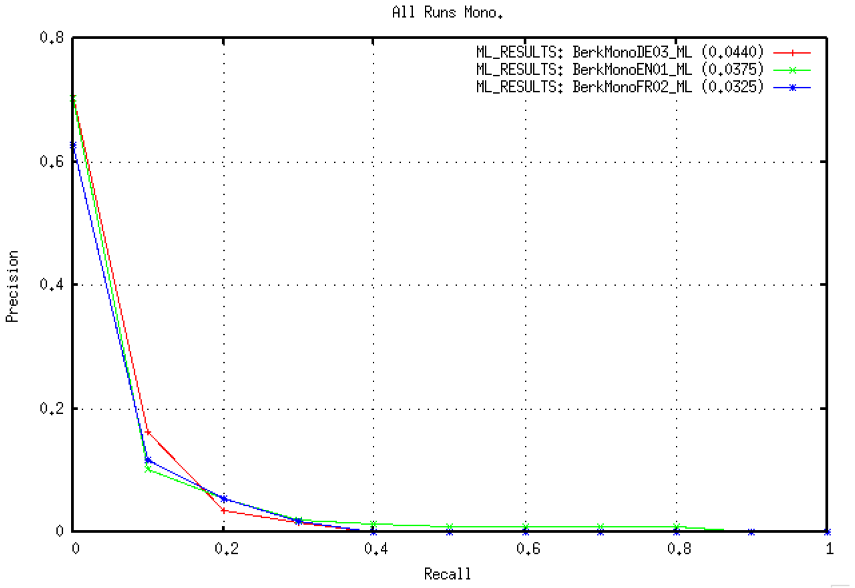

Fig. 1. Berkeley Monolingual Runs

4 Results for Submitted Runs

The summary results, as Mean Average Precision (MAP) for the submitted monolingual, bilingual and multilingual runs for English German and French collections are shown in Table 3. In this table each column represents the MAP results for a particular set of the task qrels. The EN column uses only qrels for the English subcollection, similarly for German (DE) and French (FR). The ENFRDE column combines the qrels for the three languages and while the ML column is the official MAP calculated using all language collections.

The Recall-Precision curves for these runs are also shown in Figures 1 (for monolingual) and 2 (for bilingual) and 3 (for Multilingual). In Figures 1 and 2 the 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 within a given qrel set and subtask (monolingual, translated bilingual and multilingual) by asterisks next to the scores.

The results in Table 3 show, as might be expected, the best results are usually the pure monolingual approach for a single language subcollection. We did have the interesting case again (which has happened before in other CLEF and NTCIR evaluations for us) that an bilingual version of a query in another language translated to English outperforms the "native" English query on the same collection. In this case the bilingual German to English translation did better in searching the English collection than the original English topic.

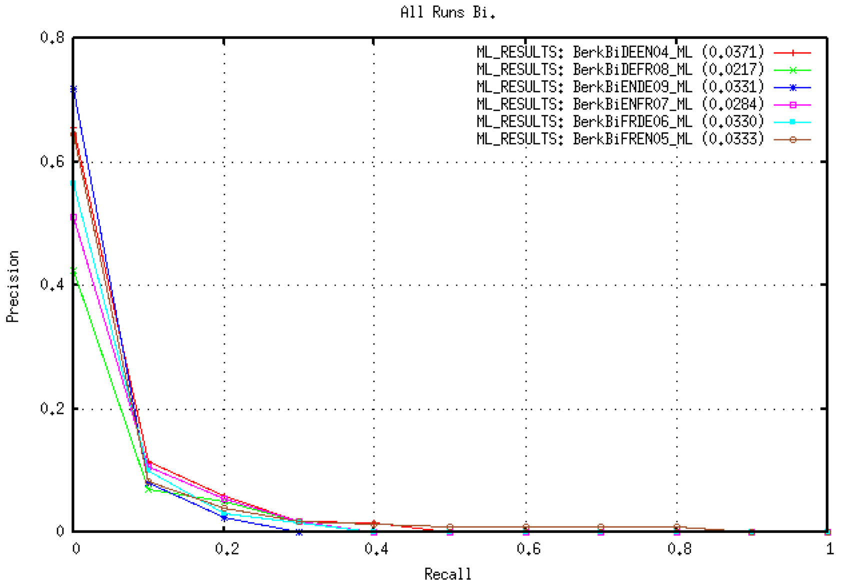

Fig. 2. Berkeley Bilingual Runs

Also not very surprising was the fact that particular languages used in searching the collections of different languages typically ended up with a MAP of zero or very near zero. What is rather strange however, is that it appears that no results were obtained from the English collections for all multilingual searches (which were supposed to be searching the English, German and French collections). Since MINMAX normalization of scores was used, it seems unlikely that this was the result of a scoring difference. It may be, however, that some system error that is causing the system to fail to combine all of the result for all three language collections. We will be checking into this later.

5 Conclusions

Our overall results this year compared poorly with the best performing systems, based on the track summary data on the DIRECT system. However, even our best runs included only the results from the English, French, and German subcollections, and no other languages, this alone apparently was enough to drastically lower the results. Interestingly, our monolingual run against only the German collection was our best performing run (when evaluated using the full multillingual qrels), which seems to indicate a predominance of German documents in the overall relevant. What was also interesting was that single language searching the three collections using languages outside of the main content of the collections including Dutch, Finnish, Italian and Spanish, provided non-zero results from each of the qrel sets, except for English. This would seem to indicated

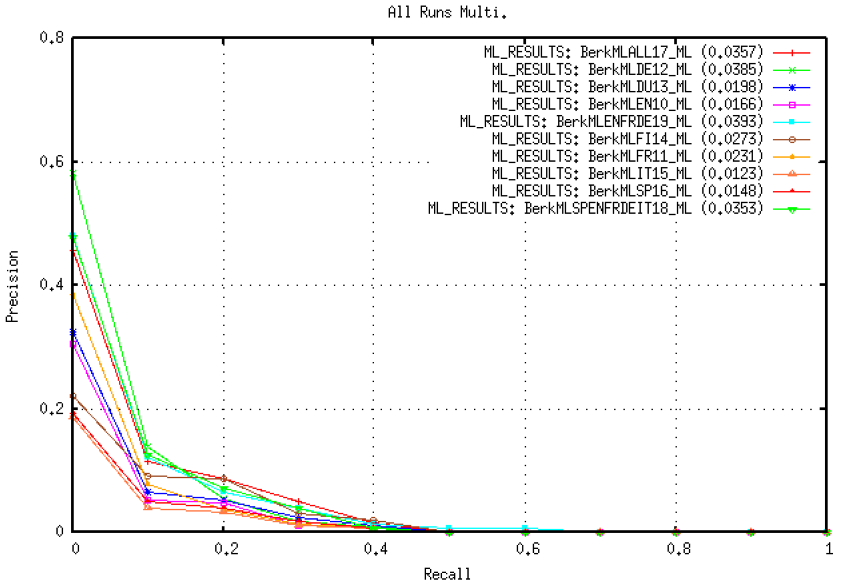

Fig. 3. Berkeley Multilingual Runs

that the multilingual topical metadata included in the Europeana collections is providing a boost for multilingual search.

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.