Logistic Regression for Metadata: Cheshire takes on 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 Berkeley Cheshire Group for the Adhoc-TEL 2008 tasks (Mono and Bilingual retrieval). Since the Adhoc-TEL task is new for this year, we took the approach of using methods that have performed fairly well in other tasks. In particular, the approach this year 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. This approach seems to be a fit good for the limited TEL records, since the overall results show Cheshire runs in the top five submitted runs for all languages and tasks except for Monolingual German.

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

The CLEF Adhoc-TEL collections are different from most of the data used for testing in the various CLEF tasks. The three sub-collections – British Library (BL), Biblioteque Nationale de France (BNF), and the Austrian National Library (ONB) – each represent about 1 million bibliographic records from The European Library union catalog (TEL). The records, we can assume, were originally in some version of MARC (Machine Readable Cataloging) before they were converted to a much more simplified bibliographic record based on the Dublin Core metadata schema. Each of the subcollections use somewhat differing encoding of the (assumed) original MARC data, not always including all of the fields that might be useful in retrieval.

Although each the collections were 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. Although this overlap of languages

presents an interesting multilingual search (and evaluation) problem, it was not addressed in our experiments this year.

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 six Monolingual runs (two German, two English, and two French) and nine Bilingual runs (each of the three main languages to both of the other main languages (German, English and French). In addition we submitted three runs from Spanish translations of the topics to the three main 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-TEL participation.

2 The Retrieval Algorithms

Note that this section is virtually identical to one that appears in our papers from previous CLEF participation[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 of records, for all of our submitted runs.

Table 3: Submitted Adhoc-TEL Runs

Run Name	Description	Type	MAP
M-DE-TD-T2FB	Monolingual German	TD auto	0.1742
M-DE-T-T2FB	Monolingual German	T auto	0.1980 *
M-EN-TD-T2FB	Monolingual English	TD auto	0.3466 *
M-EN-T-T2FB	Monolingual English	T auto	0.2773
M-FR-TD-T2FB	Monolingual French	TD auto	0.2438 *
M-FR-T-T2FB	Monolingual French	T auto	0.1931
B-ENDE-TD-T2FB	Bilingual English⇒German	TD auto	0.1556 *
B-ESDE-TD-T2FB	Bilingual Spanish⇒German	TD auto	0.1165
B-FRDE-TD-T2FB	Bilingual French⇒German	TD auto	0.1291
B-DEEN-TD-T2FB	Bilingual German⇒English	TD auto	0.1847
B-ESEN-TD-T2FB	Bilingual Spanish⇒English	TD auto	0.2694
B-FREN-TD-T2FB	Bilingual French⇒English	TD auto	0.2825 *
B-DEFR-TD-T2FB	Bilingual German⇒French	TD auto	0.1885 *
B-ENFR-TD-T2FB	Bilingual English⇒French	TD auto	0.1749
B-ESFR-TD-T2FB	Bilingual Spanish⇒French	TD auto	0.1767

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 or the title alone 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.

The scripts for each run submitted the topic elements as they appeared in the topic to the system for TREC2 logistic regression searching with blind feedback. When both the "title" and "description" topic elements were used, they were combined into a single probabilistic query. Table 3 shows which element were used in the "Type" column, T for title only and TD for title and description.

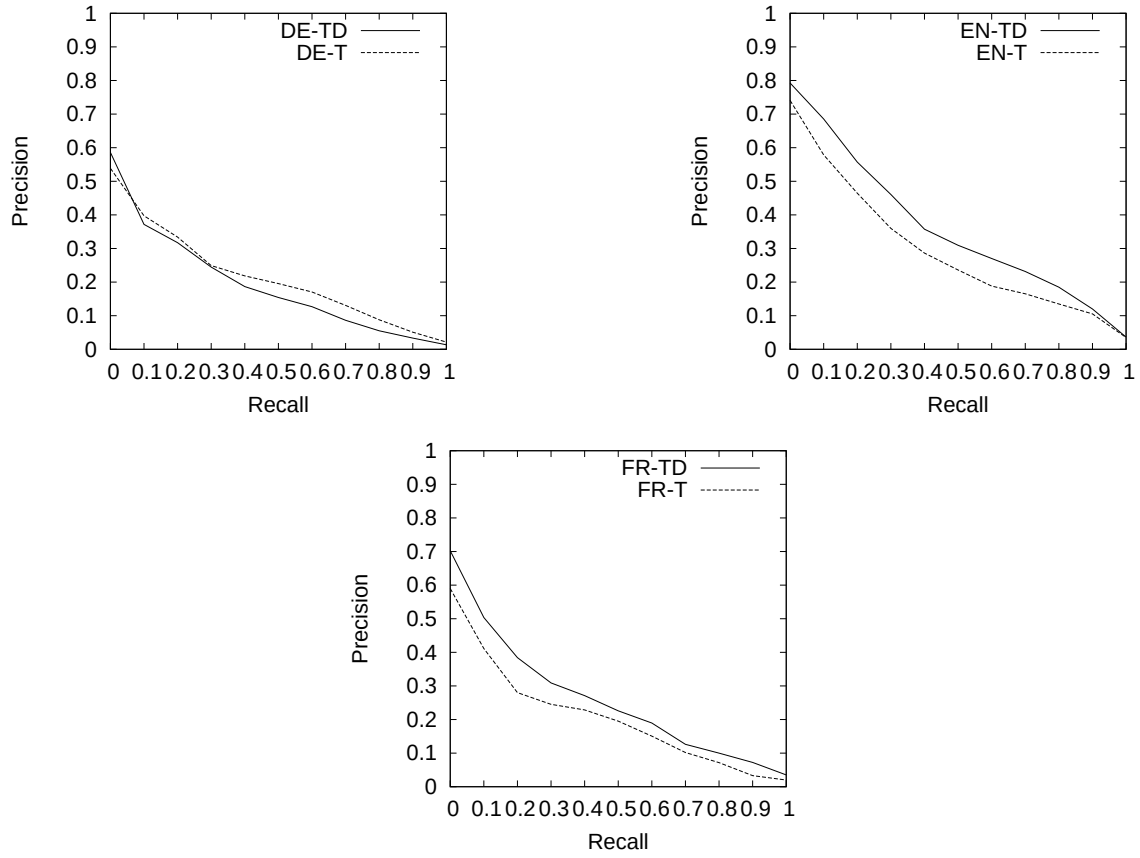
4 Results for Submitted Runs

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 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.

Obviously the "weak man" in our current implementation remains monolingual German. This may be due to decompounding issues, but the higher results for title-only monolingual seem anomalous, since for each other language, the combination of title and description performed better than title alone.

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



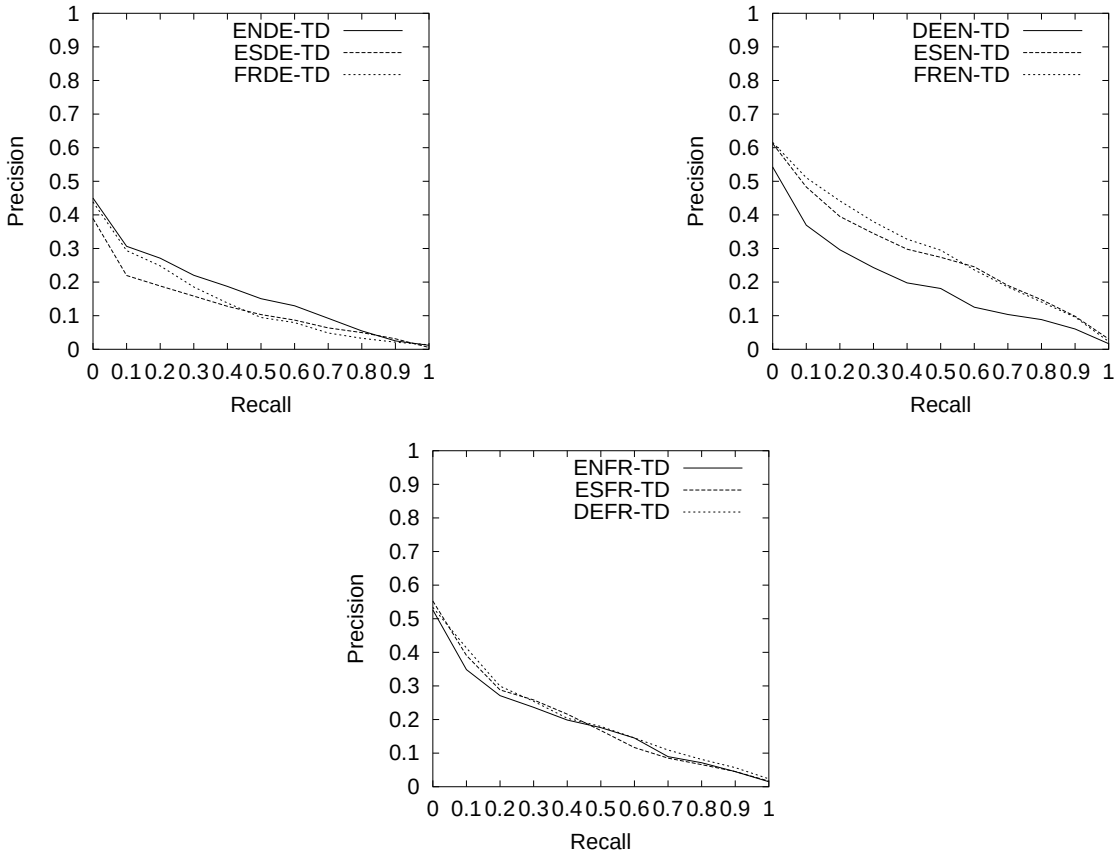
In spite of this relatively poor performance in monolingual German, we had the rather surprising results that for bilingual English to German our submitted run B-ENDE-TD-T2FB was ranked third overall among the bilingual "to German" runs submitted, and our German to French bilingual run B-DEFR-TD-T2FB was ranked first in the bilingual "to French" task well ahead of our English to French run. This would seem to indicate that the our translation system works quite well with the Adhoc-TEL topics.

5 Conclusions

In looking at the overall results for the various Adhoc-TEL tasks, it would appear that the basic logistic regression with blind relevance feedback approach, coupled with the LEC translation system is a fairly good combination. Since Adhoc-TEL is a new task, we took a fairly conservative approach using methods that have worked well in the past.

In our experiments for other tracks (GeoCLEF for example) we reintroduced fusion approached for retrieval that performed quite well and could be easily applied to this task as well. For future work we intend to test these approaches as well as some other approaches that would incorporate external supplementary topical indexing for the books (primarily) represented by Adhoc-TEL records.

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



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, page to appear 2008, 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, page to appear 2008, 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.