GeoCLEF Text Retrieval and Manual Expansion Approaches

Ray R. Larson* and Fredric C. Gey
School of Information*
University of California, Berkeley, USA
ray@sims.berkeley.edu

Abstract

In this paper we will describe the Berkeley approaches to the GeoCLEF tasks for CLEF 2006. This year we used two separate systems for different tasks. Although of the systems both use versions of the same primary retrieval algorithm they differ in the supporting text pre-processing tools used.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.7 Digital Libraries

General Terms

Algorithms, Performance, Measurement

Keywords

Cheshire II, Logistic Regression, Data Fusion

1 Introduction

This paper describes the retrieval algorithms and evaluation results for Berkeley's official submissions for the GeoCLEF track. Two separate systems were used for our runs, although both used the same basic algorithm for retrieval. Instead of the automatic expansion used in last year's GeoCLEF, this year we used manual expansion for a selected subset of queries for only 2 out of the 18 runs submitted. The remainder of the runs were automatic without manual intervention in the queries (or translations). We submitted 12 Monolingual runs (2 German, 4 English, 2 Spanish, and 4 Portuguese) and 6 Bilingual runs (2 English⇒German, 2 English⇒Spanish, and 2 English⇒Portuguese). We did not submit any Biligual X⇒English runs.

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 GeoCLEF participation.

2 The Retrieval Algorithms

Note that this section is virtually identical to one that appears in our ImageCLEF and Domain Specific papers. 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 GeoCLEF 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]. We used two different implementations of the algorithm. One was in stand-alone experimental software developed by Aitao Chen, and the other in the Cheshire II information retrieval system. Although the basic behaviour of the algorithm is the same for both systems, there are differences in the sets of pre-processing and indexing elements used in retrieval. One of the primary differences is the lack of decompounding for German documents and query terms in the Cheshire II system. 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 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 [8].

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

	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

Table 1: Contingency table for term relevance weighting

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:

$$w_t = log \frac{\frac{R_t}{R - R_t}}{\frac{N_t - R_t}{N - N_t - R + R_t}} \quad (4)$$

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 GeoCLEF

In this section we describe the specific approaches taken for our submitted runs for the GeoCLEF 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 standalone version treats all text as a single "bag of words" that is extracted and indexed. For German documents it uses a custom "decompounding" algorithm to extract component terms from German compounds.

The Cheshire II system uses the XML structure and extracts selected portions of the record for indexing and retrieval.

Name	Description	Content Tags	Used
docno	Document ID	DOCNO	no
pauthor	Author Names	BYLINE, AU	no
headline	Article Title	HEADLINE, TITLE, LEAD, LD, TI	no
topic	Content Words	HEADLINE, TITLE, TI, LEAD BYLINE, TEXT, LD, TX	yes yes
date	Date of Publication	DATE, WEEK	no
geotext	Validated place names	TEXT, LD, TX	no
geopoint	Validated coordinates for place names	TEXT, LD, TX	no
geobox	Validated bounding boxes for place names	TEXT, LD, TX	no

Table 2: Cheshire II Indexes for GeoCLEF 2006

Table 2 lists the indexes created by the Cheshire II system for the GeoCLEF 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 GeoCLEF runs. The geotext, geopoint, and geobox indexes were not created on the Cheshire2 for the Spanish and Portuguese sub-collections due to the lack of a suitable gazetteer in those languages.

Because there was no explicit tagging of location-related terms in the collections used for GeoCLEF, we applied the above approach to the "TEXT", "LD", and "TX" elements of the records of the various collections. The part of news articles normally called the "dateline" indicating the location of the news story was not separately tagged in any of the GeoCLEF collection, but often appeared as the first part of the text for the story.

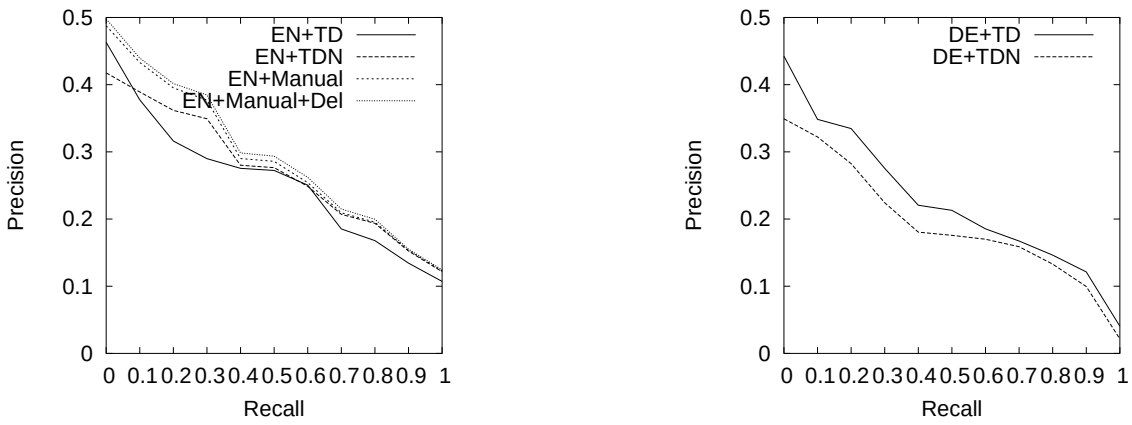

Figure 1: Berkeley Monolingual Runs – English (left) and German (right)

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 used decompounding in the indexing and querying processes to generate simple word forms from compounds.

The Snowball stemmer was used by both systems for language-specific stemming.

3.2 Search Processing

All of the runs for Monolingual English and German, and the runs for Bilingual English⇒German used the standalone retrieval programs developed by Aitao Chen. The Monolingual Spanish and Portuguese, and the Bilingual English⇒Spanish and English⇒Portuguese runs all used the Cheshire II system.

The English and German Monolingual runs used language-specific decompounding of German compound words. The Bilingual English⇒German also used decompounding.

Searching the GeoCLEF collection using the Cheshire II system involved using TCL scripts to parse the topics and submit the title and description or the title, description, and narrative from the topics. For monolingual search tasks we used the topics in the appropriate language (Spanish and Portuguese), for bilingual tasks the topics were translated from the source language to the target language using the L&H PC-based machine translation system. In all cases the various topic elements were combined into a single probabilistic query.

We tried two main approaches for searching, the first used only the topic text from the title and desc elements (TD), the second included the narrative elements as well (TDN). In all cases only the full-text "topic" index was used for Cheshire II searching.

Two of our English Monolingual runs used manual modification for topics 27, 43, and 50 by adding manually selected place names to the topics, in addition, one of these (which turned out to be our best performing English Monolingual run) also manually eliminated country names from topic 50.

Also after two initial runs for Portuguese Monolingual were submitted (BKGeoP1 aand BKGeoP2), a revised and corrected version of the topics was released, and two additional runs (BKGeoP3 and BKGeoP4) were submitted using the revised topics, retaining the original submissions for comparison.

4 Results for Submitted Runs

The summary results (as Mean Average Precision) for the submitted bilingual and monolingual runs for both English and German are shown in Table 3, the Recall-Precision curves for these runs are also shown in Figures 1 and 2 (for monolingual) and 3 and 4 (for bilingual). In Figures

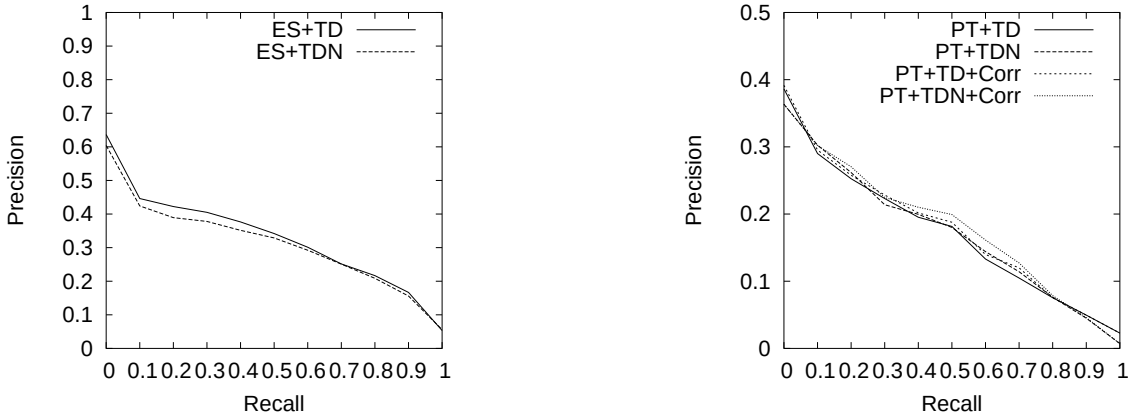

Figure 2: Berkeley Monolingual Runs – Spanish (left) and Portuguese (right)

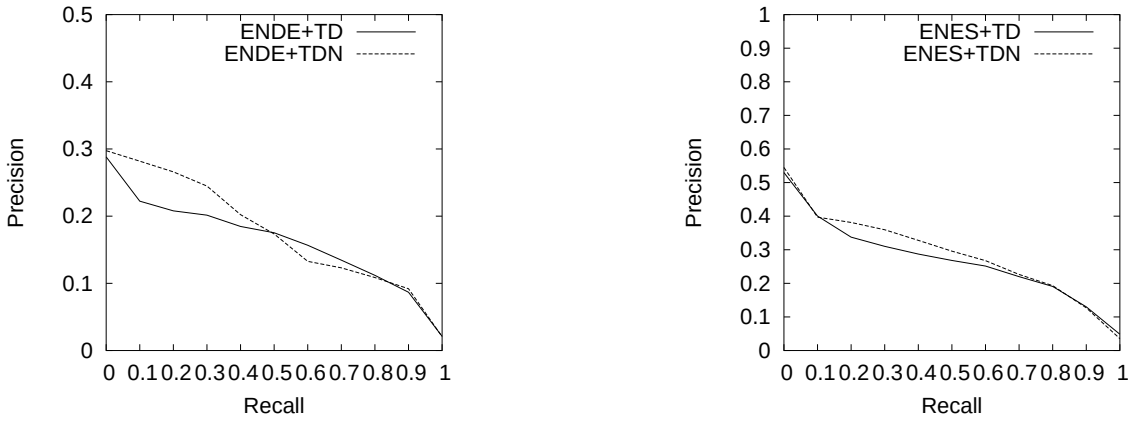

Figure 3: Berkeley Bilingual Runs – English to German (left) and English to Spanish (right)

1-4 the names for the individual runs represent the language code and type of run, which can be compared with full names and descriptions in Table 3.

Table 3 indicates runs that had the highest overall MAP for the task by asterisks next to the run name. Single asterisks indicate the the highest MAP values among our own runs, while double asterisks indicate the runs where the MAP is the maximum recorded among official submissions.

As can be seen from the table, Berkeley's cross-language submissions using titles, descriptions, and narratives from the topics were the best performing runs for the Bilingual tasks overall. Our Monolingual submissions, on the other hand did not fare as well, but still all ranked within the top quartile of results for each language except Portuguese where we fell below the mean. This result was surprising, given the good performance for Spanish. We now suspect that errors in mapping the topic encoding to the stored document encoding, or possibly problems with the Snowball stemmer for Portuguese may be responsible for this relatively poor performance.

Last year's GeoCLEF results (see [7]) also reported on runs using different systems (as Berkeley1 and Berkeley2), but both systems did all or most of the tasks. Table 4 shows a comparison of Average precision (MAP) for the best performing German and English runs for this year and for the two systems from last year. The German language performance of the system this year for both Bilingual and Monolingual tasks shows a definite improvement, while the English Monolingual performance is somewhat worse that either system last year. The "Berk2" system is essentially the same system as used this year for English and German runs.

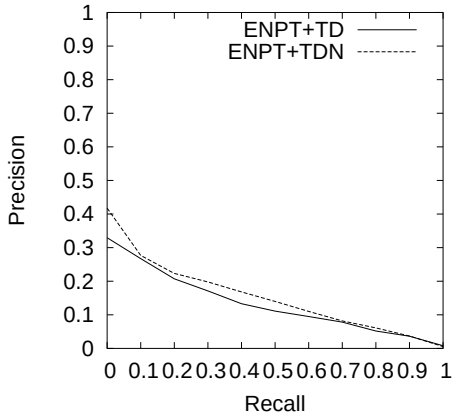

Figure 4: Berkeley Bilingual Runs – English to Portuguese

5 Conclusions

Manual expansion of selected topics shows a clear, if small, improvement in performance over fully automatic methods. In comparing to Berkeley's best performing English and German runs for last year, it would appear that either the English queries this year were much more difficult, or that there were problems in the English runs. This year, while we did not use automatic expansion of toponyms in the topic texts, this was done explicitly in some of the topic narratives which may explain the improvements in runs using the narratives. It is also apparent that this kind of explicit toponym inclusion in queries, as might be expected, leads to better performance when compared to using titles and descriptions alone in retrieval.

Although we did not do any explicit geographic processing for this year, we plan to do so in the future. The challenge for next year is to be able to obtain the kind of effectiveness improvement seen with manual query expansion, in automatic queries using geographic processing.

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.

Run Name	Description	Type	MAP
BKGeoED1	Bilingual English⇒German	TD auto	0.15612
BKGeoED2**	Bilingual English⇒German	TDN auto	0.16822
BKGeoES1	Bilingual English⇒Spanish	TD auto	0.25712
BKGeoES2**	Bilingual English⇒Spanish	TDN auto	0.27447
BKGeoEP1	Bilingual English⇒Portuguese	TD auto	0.12603
BKGeoEP2**	Bilingual English⇒Portuguese	TDN auto	0.14299
BKGeoD1*	Monolingual German	TD auto	0.21514
BKGeoD2	Monolingual German	TDN auto	0.18218
BKGeoE1	Monolingual English	TD auto	0.24991
BKGeoE2	Monolingual English	TDN auto	0.26559
BKGeoE3	Monolingual English	Manual	0.28268
BKGeoE4*	Monolingual English	Manual	0.28870
BKGeoS1*	Monolingual Spanish	TD auto	0.31822
BKGeoS2	Monolingual Spanish	TD auto	0.30032
BKGeoP1	Monolingual Portuguese	TD auto	0.16220
BKGeoP2	Monolingual Portuguese	TDN auto	0.16305
BKGeoP3	Monolingual Portuguese	TD auto	0.16925
BKGeoP4*	Monolingual Portuguese	TDN auto	0.17357

Table 3: Submitted GeoCLEF Runs

TASK	2006 NAME	MAP 2006	Berk1 MAP 2005	Berk2 MAP 2005	Pct. Diff Berk1	Pct. Diff Berk2
GC-BILI-X2DE-CLEF2006	BKGeoED2	0.16822	0.0777	0.1137	116.50	47.95
GC-MONO-DE-CLEF2006	BKGeoD1	0.21514	0.0535	0.133	302.13	61.76
GC-MONO-EN-CLEF2006	BKGeoE4	0.28870	0.2924	0.3737	-1.26	-22.74

Table 4: Comparison of Berkeley's best 2005 and 2006 runs for English and German

[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, Fredric C. Gey, and Vivien Petras. Berkeley at GeoCLEF: Logistic regression and fusion for geographic information retrieval. In *Cross-Language Evaluation Forum: CLEF 2005*, pages 963–976. Springer (Lecture Notes in Computer Science LNCS 4022), 2006.

[8] 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.