Cheshire at GeoCLEF 2007: Retesting Text Retrieval Baselines

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 Berkeley for the main GeoCLEF 2007 tasks (Mono and Bilingual retrieval). This year we used only a single system in the research, and were not able to do much in the way of interesting geographic work due to a number of factors, not the least of which was time competition from other tasks. 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. Since data on overall performance relative to other participants were not available at the time of writing, our discussion in this paper is limited to comparison between our submitted runs.

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

This paper describes the retrieval algorithms and evaluation results for Berkeley's official submissions for the GeoCLEF 2007 track. Instead of the expansion approaches used in last year's GeoCLEF, this year we used only unexpanded topics in querying the database. All of the runs were automatic without manual intervention in the queries (or translations). We submitted 3 Monolingual runs (1 German, 1 English, and 1 Portuguese) and 9 Bilingual runs (each of the three main languages to each each other language, and 3 runs from Spanish to English, German, and Portuguese).

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]. 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:

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

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 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 GeoCLEF 2006

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 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 georeferencing indexing subsystem of Cheshire II was used for the geotext, geopoint, and geobox indexes. This subsystem is intended to extract proper nouns from the text being indexed and then attempts to match them in a digital gazetteer. For GeoCLEF we used a gazetteer derived from the World Gazetteer (http://www.world-gazetteer.com) with 224698 entries in both English and German. The indexing subsystem provides three different index types: verified place names (an index of names which matched the gazetteer), point coordinates (latitude and longitude coordinates of the verified place name) and bounding box coordinates (bounding boxes for the matched places from the gazetteer). All three types were created, but due to time constraints, and lack of time to upgrade the gazetteer or fix bugs in coordinate assignments, ended up not using any of the geographic indexes in this year's submissions. Because we do not use complete NLP parsing techniques, the system is unable to distinguish between proper nouns for places from

Table 3: Submitted GeoCLEF Runs

Run Name	Description	Type	MAP
BerkMODEBASE	Monolingual German	TD auto	0.1392
BerkMOENBASE*	Monolingual English	TD auto	0.2642
BerkMOPTBASE	Monolingual Portuguese	TD auto	0.1739
BerkBIENDEBASE	Bilingual English⇒German	TD auto	0.0902
BerkBIENPTBASE	Bilingual English⇒Portuguese	TD auto	0.2012
BerkBIDEENBASE*	Bilingual German⇒English	TD auto	0.2208
BerkBIDEPTBASE	Bilingual German⇒Portuguese	TD auto	0.0915
BerkBIPTDEBASE	Bilingual Portuguese⇒German	TD auto	0.1109
BerkBIPTENBASE	Bilingual Portuguese⇒English	TD auto	0.2112
BerkBIESDEBASE	Bilingual Spanish⇒German	TD auto	0.0724
BerkBIESENBASE	Bilingual Spanish⇒English	TD auto	0.2195
BerkBIESPTBASE	Bilingual Spanish⇒Portuguese	TD auto	0.1924

those for individuals. This leads to errors in geographic assignment where, for example, articles about Irving Berlin might be tagged as refering to the city.

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 collections, but often appeared as the first part of the text for the story.

Geographic indexes were not created for the Portuguese sub-collection due to the lack of a suitable gazetteer. We plan for later work to substitute the "GeoNames" database which is much more detailed and provides a more complete geographical hierarchy in its records, along with alternate names in multiple 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. Although we tried again this year to make this work within the Cheshire system, we again lacked the time needed to implement it correctly.

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

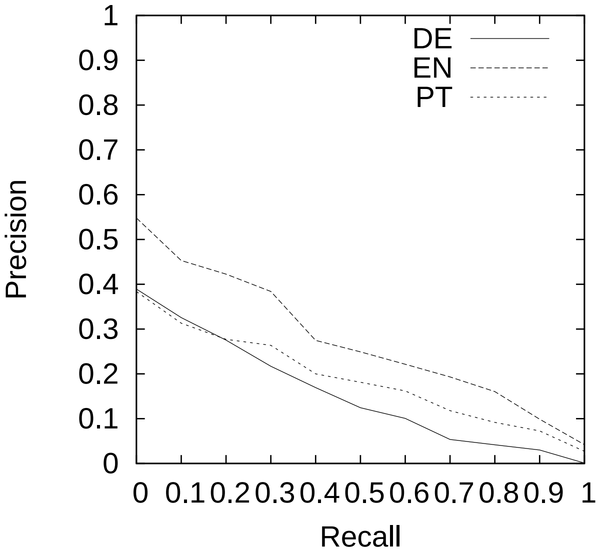
3.2 Search Processing

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 (English, German, and Portuguese), 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. In all cases the "title" and "desc" topic elements were combined into a single probabilistic query. We consider all of these runs to be the simplest "baseline" for our system, and we plan to implement more elaborate processing approaches for subsequent testing.

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 (for monolingual) and 2 (for bilingual). In Figures 1 and 2 the

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



TASK	MAP 2006	MAP 2007	Pct. Diff.
Monolingual English	0.2499	0.2642	5.7222
Monolingual German	0.2151	0.1392	-54.5259
Monolingual Portuguese	0.1622	0.1739	7.2133
Bilingual English⇒German	0.1561	0.0902	-73.0599
Bilingual English⇒Portuguese	0.12603	0.2012	59.6825

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

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.

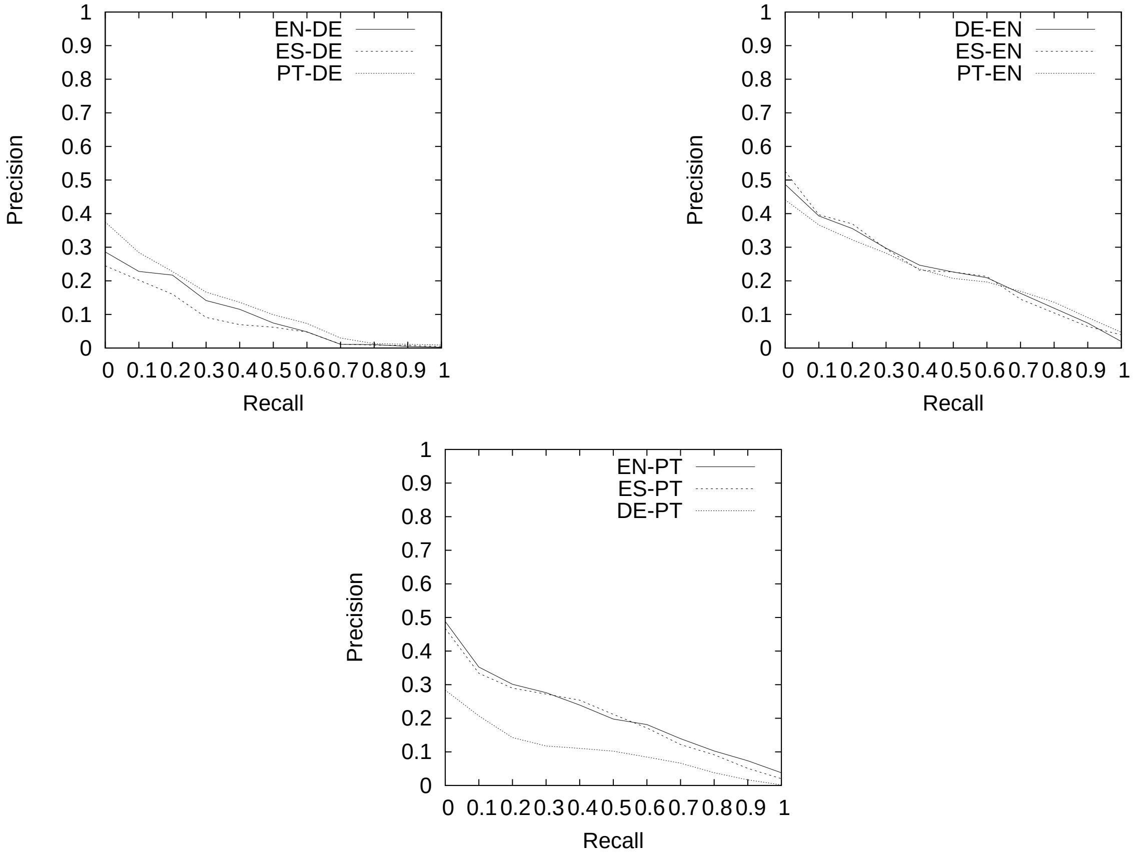
Once again we found some rather anomalous results among the official runs. For example, it is not at all clear, given the same basic approach used for all of the runs, why the bilingual runs for English⇒Portuguese (MAP 0.2012), and Spanish⇒Portuguese (MAP 0.1924) should have performed better than our Monolingual Portuguese run (MAP 0.1739).

Obviously the "weak man" in our current implementation is German. This may be due to decompounding issues, but the lower results are clear in both Monolingual and Bilingual runs where either the source topics or the target data is German.

5 Conclusions

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. In addition, we used only the title and desc elements of topics this year, and also we did not use automatic expansion of toponyms in the topic texts. Since this was done explicitly in some of the topic narratives we may have missed possible improvements by not using the entire topic. In previous years it has been apparent that implicit or explicit toponym inclusion in queries, as might be

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



expected, leads to better performance when compared to using titles and descriptions alone in retrieval.

Because we used a virtually identical processing approach (except for translation) this year as we used for some of our runs submitted for GeoCLEF 2006, we build Table 4 examine the differences. Overall, we did see some improvements in results. However, the submitted 2006 results used decompounding for German, which would appear to be the primary cause of our declining monolingual and bilingual scores for German, although the translation software may also be at fault. Otherwise, our bilingual results this year are largely due to the effectiveness of our new translation software. We used the Spanish topic statements provided for bilingual Spanish to English, German, and Portuguese, and saw results that look quite good for English and Portuguese, with the exception again being German. We will be interested to see how these scores compare to the various other approaches used in GeoCLEF this year.

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, 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.