Probabilistic Text Retrieval for NTCIR9 GeoTime

Ray R. Larson
School of Information
University of California, Berkeley
Berkeley, California, USA, 94720-4600
ray@ischool.berkeley.edu

ABSTRACT

For the NTCIR-9 Workshop UC Berkeley participated only in the GeoTime track. For our initial experiments we used only the Logistic Regression ranking with blind feedback approach that we also used in NTCIR-8. We participated in both English and Japanese monolingual and bilingual search tasks. For all Japanese topics we preprocessed the text using the ChaSen morphological analyzer for term segmentation. For these submitted runs we did not do any special purpose geographic or temporal processing. This brief paper describes the submitted runs and the methods used for them.

Keywords: Logistic Regression, Probabilistic Retrieval.

1. INTRODUCTION

The experimental GeoTime track for NTCIR explores the use of both time and place as elements in many of the searches performed in both IR evaluations and in day-today use of search engines for the WWW. The use of geographic elements in searching has been previously explored in the GeoCLEF evaluations for European languages, but this is the first attempt to do similar evaluation for Asian languages, with the added complexity of time constraints and temporal elements. For this first GeoTime evaluation we decided to use a set of text-based approaches without explicit geographic or temporal processing. We used, essentially, the same search tools and methods described in our IR4QA paper in this volume detailed descriptions of the algorithms used and our approach to blind or pseudo relevance feedback can be found there [9]. Our document ranking algorithm is a probability model based using the technique of logistic regression [4]. For all of our runs we used the TREC2 logistic regression model, described below, with blind or pseudo relevance feedback. In this paper we describe the submissions for this track and consider how they might be improved.

2. DATABASE AND INDEXING

The database for GeoTime consisted of both English and Japanese newspaper articles for the same time periods. The collections included English articles from the New York Times (2002-2005), Xinhua New Service (1998-2001), the Korea Times (1998-2001) and Mainichi English version (1998-2001). The Japanese collection consisted of stories from the Mainichi newpapers for the same time period (1998-2005). This combination of stories was expected to provide coverage in both English and Japanese for the entire period. For the English indexing process we used the Cheshire version of the Porter stemmer and a stoplist that we had used previously for English language databases. During the indexing process for Japanese all of the data from the Mainichi newspaper database was segmented using the ChaSen segmentation software, and each segment was indexed as a “word”. In addition a Japanese stoplist used in earlier NTCIR tracks was used to eliminate common words. Segmentation actually involved multiple steps since the UTF-8 documents had to be tranformed to EUC encoding for segmentation and then back to UTF-8 for storage in the database and indexes.

A number of separate indexes were created for each language, although the only index used in our submitted runs for NTCIR-9 was an index that contained all of the words (or segmented tokens for Japanese) from the entire record. This approach was the same that we used in the NTCIR-8 GeoTime track. Although different data sources were involved in the indexing process, they shared the same basic XML structure across the various sources, with minor variations. This allowed us to treat all of the English collections as if they were a single collection. Similarly, minor variations between the 1998-2001 and 2002-2005 sets of Mainichi in Japanese did not hinder treating the Japanese collection as a single collection.

Given the GeoTime task, one of the problematic features of the many of the collections is their lack of actual explicit story dates. For example, the New York Times collection contains the date for most stories only as substring of the document ID. The same pattern occurs in Xinhua stories, although partial dates (month and day) are sometimes included in the DATELINE field. The English Mainichi and Korea Times, on the other hand include an explicit DATE field (although in differing formats).

3. RETRIEVAL APPROACH

Note that much of this section is based on one that appears in our papers from CLEF participation[8, 7].

The basic form and variables of the *Logistic Regression* (LR) algorithm used for all of our submissions were originally developed by Cooper, et al. [4]. 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:

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

Of course, this last transformation is not actually necessary since the log odds could also be used directly to rank the results, but we do it in the cheshire system so that the result of any operation is a probability value for each item retrieved.

3.1 TREC2 Logistic Regression Algorithm

For NTCIR9 GeoTime 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:

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

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.

When 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 $c_4 = 0.0929$.

3.2 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 (in terms of mean average precision) 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 [10].

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

Table 1: Submitted GeoTime Runs

RunID	Type	MAP	Mean Q	Mean nDCG@10	Mean nDCG@100	Mean nDCG@1000
BRKLY-EN-EN-01-D	EN⇒EN	0.4066	0.4246	0.4931	0.5013	0.6012
BRKLY-EN-EN-01-DN	EN⇒EN	0.4495	0.4713	0.5690	0.5538	0.6588
BRKLY-JA-EN-01-D	JA⇒EN	0.3967	0.4081	0.4737	0.4739	0.5593
BRKLY-JA-EN-01-DN	JA⇒EN	0.4874	0.5035	0.6072	0.5950	0.6891
BRKLY-EN-JA-01-D	EN⇒JA	0.2398	0.2550	0.3124	0.3326	0.4211
BRKLY-EN-JA-01-DN	EN⇒JA	0.3081	0.3214	0.3733	0.4250	0.5151
BRKLY-JA-JA-01-D	JA⇒JA	0.2475	0.2640	0.3250	0.3492	0.4157
BRKLY-JA-JA-01-DN	JA⇒JA	0.3716	0.3836	0.4362	0.4844	0.5696

for combinations of (assumed) relevance and whether or not the term is, or is not in a document. Table 2 shows this contingency table.

Table 2: 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:

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

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 3 and the ranked results returned as the response for that topic.

4. SUBMISSIONS AND RESULTS FOR OFFICIAL RUNS

Table 3 shows the results for our official submitted runs for the GeoTime task. In examining the 3 table, some rather unusual results are apparent again this year. First, and most striking, is that once again we find a case where our cross-language runs (the BRKLY-JA-EN-01-DN JA⇒EN run) actually performed better that the corresponding monolingual run (BRKLY-EN-EN-01-DN EN⇒EN). The opposite is usually observed in cross-language retrieval, where the bilingual almost always lags the monolingual in performance. The more typical behavior is shown for monolingual Japanese and English to Japanese runs where the monolingual runs outperform the bilingual runs.

In all cases translation from English to Japanese or from Japanese to English was performed using the Google Translate service. Each of the original topics (which included both English and Japanese descriptions and narratives) was split into separate English-only and Japanese-only topics. Because Google Translate will not operate on XML files directly, but would operate on HTML, we first substituted the XML markup in the files with HTML then performed the translations and converted the HTML back to the original XML markup.

The Japanese topics (either original or translated) were segmented into "words" separated by blanks using the ChaSen segmenting tool. This tool was also used for segmenting the database before indexing. Because the version of ChaSen that we used required the text to be in EUC-JP encoding, we used iconv to convert encodings from UTF-8 to EUC-JP before segmenting and back again afterwards. All of the conversions were implemented as scripts.

Compared to our results last year, we did not perform very well in the Japanese monolingual or bilingual tasks. We suspect that this may have been due to our failure to perform one of these conversion steps in normalization and segmentation of the Japanese text. Because the segmentation tool that we are using (ChaSen) is oriented towards plain text instead of XML marked-up text, the *Romanji* characters and punctuation of tags are treated as separate segments and spaces are inserted. The same applies to numbers and latin letters including dates occurring in the text. We have a script to repair the XML markup after segmentation, but last year we also included another process to repair latin strings and numbers in the text. This later step was accidently skipped this time. Since segments are treated like words in English, this meant that each individual letter and number was treated as words, and possibly expanded during blind feedback. We suspect this may have had a significant detrimental impact on our Japanese results.

All of our submitted runs for the GeoTime track used probabilistic retrieval using TREC2 logistic regression algorithm described in detail above. Each of our submitted runs used pseudo or blind relevance feedback along with the TREC2 algorithm. For each runid in table 3 those with DN at the end of the name used both the DESCRIPTION and NARRATIVE elements of the topics, and those with D alone used the DESCRIPTION only. As the scores in table 3 show, using both the description and narrative elements along with blind feedback gives the best results for these collections.

We submitted 2 bilingual runs and 2 monolingual for each

Table 3: Comparing NTCIR8 and NTCIR9 GeoTime Results

Type	NTCIR8 MAP	NTCIR9 MAP	Diff
EN⇒EN D	0.3615	0.4066	0.0451
EN⇒EN DN	0.4045	0.4495	0.0450
JA⇒EN D	0.3759	0.3967	0.0208
JA⇒EN DN	0.4158	0.4874	0.0716
EN⇒JA D	0.3458	0.2398	-0.1060
EN⇒JA DN	0.3619	0.3081	-0.0538
JA⇒JA D	0.4143	0.2475	-0.1668
JA⇒JA DN	0.4277	0.3716	-0.0561

language as our official entries, one with description only, and the other with both description and narrative. The following information and the information on performance measures in Table 3 is presented in the GeoTime overview paper in this volume [5]. The three effectiveness metrics for evaluating the GeoTime runs: Mean Average Precision (MAP), Q-measure (Mean Q) and a version normalised Discounted Cumulative Gain (Mean nDCG) described in the overview paper[5]. The best performing English run submitted by Berkeley was BRKLY-JA-EN-01-DN, which used the DESCRIPTION and NARRATIVE topic text in Japanese, translated to English. The next best performing (BRKLY-EN-EN-01-DN) used the same algorithm and blind feedback approach, but used the original English topic text instead of the translated Japanese. In all cases using the Narrative provided an obvious boost in performance compared to using description alone.

A slightly different pattern of results is seen in table 3 for our Japanese submissions, where the Japanese monolingual runs outperformed the English to Japanese translation runs.

Overall our English runs ranked a bit above average (when compared to all other submitted English runs). It is worth pointing out that all of our submitted runs were fully automatic with no manual query expansion or modifications. As indicated above, our Japanese runs did not fare as well this year relative to other submissions, with our description-only runs being some of the worst-performing runs submitted.

5. CONCLUSION

This paper has described Berkeley's submissions to GeoTime task. We hope, time permitting, to conduct a number of further experiments with the data and relevance judgements. We plan to apply our missing repair step following segmentation and see if it actually improves our performance with these topics.

Because these submissions were intended to confirm a baseline set last year for comparison with methods using special geographic and temporal processing of the texts, we did not use any such methods for NTCIR-9. However, we found quite different results for comparable approaches between last year and this year. We also suspect that the methods (sometimes manual) used by other groups might have skewed the averages. Although the absolute values for most of our scores improved, we ended up ranked lower compared to other systems. This may be a suggestion that all other systems have improved relative to our attempt to stay stable.

We fully intend to exploit some of special indexing tools developed for the Cheshire system in the future that can take advantage of geographic proximity and other approaches for retrieval in the future.

6. REFERENCES

[1] A. Chen. Multilingual information retrieval using english and chinese queries. In C. Peters, M. Braschler, J. Gonzalo, and M. 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] A. Chen. *Cross-Language Retrieval Experiments at CLEF 2002*, pages 28–48. Springer (LNCS #2785), 2003.

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

[4] W. S. Cooper, F. C. Gey, and D. 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.

[5] F. Gey, R. Larson, N. Kando, J. Machado, and T. Sakai. NTCIR-GeoTime overview: Evaluating geographic and temporal search. In *Proceedings of the NTCIR-8 Workshop, Tokyo, June 2010*, pages 0–0, 2010.

[6] R. 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] R. 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, Sept. 2008.

[8] R. 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, Sept. 2008.

[9] R. R. Larson. Logistic regression for ir4qa. In *Proceedings of the NTCIR-8 Workshop, Tokyo, June 2010*, pages 0–0, 2010.

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