Logistic Regression for IR4QA

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

ABSTRACT

For NTCIR-8 Workshop UC Berkeley participated in IR4QA (Information Retrieval for Question Answering) as well as the GeoTime track. For IR4QA we only did Japanese monolingual search and English to Japanese bilingual search. Our focus was thus primarily upon Japanese topic search against the Japanese News document collection as in past NTCIR participations. We preprocessed the text using the ChaSen morphological analyzer for term segmentation. We utilized a timetested logistic regression algorithm for document ranking coupled with blind feedback. The results were satisfactory, ranking second among IR4QA overall submissions for Japanese.

Keywords: Logistic Regression, Probabilistic Retrieval.

1. INTRODUCTION

UC Berkeley has participated in all eight NTCIR workshops, concentrating primarily on the Cross-Language Information Retrieval Tasks. In NTCIR-3 we also participated in the Patent Retrieval task [3]. For the NTCIR Workshops NTCIR-4[6], NTCIR-5 [7] and NTCIR-6 [8] tasks, we limited our participation to a portion of the Bilingual task, specifically this search between Japanese and Chinese languages. For NTCIR-7 we participated in both the Patent Mining task[9] and the IR4QA task[13]. Our document ranking algorithm is a probability model based using the technique of logistic regression [5] In this paper we describe the Cheshire implementation of the "TREC2" logistic regression algorithm with blind (or pseudo) feedback. Then we will discuss the submissions for this track and how they might be improved.

2. THE RETRIEVAL ALGORITHMS

Note that much of this section is based on one that appears in our papers from CLEF participation[12, 11].

The basic form and variables of the *Logistic Regression* (LR) algorithm used for all of our submissions were 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:

$$\log O(R \mid Q, D) = b_0 + \sum_{i=1}^{S} b_i s_i \quad (1)$$

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:

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

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.

2.1 TREC2 Logistic Regression Algorithm

For IR4QA 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[4]. The formal definition of the TREC2 Logistic Regression algorithm used is:

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

NTCIR-8 June, 2010, Tokyo
ACM 0-89791-88-6/97/05.

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 $c_4 = 0.0929$.

2.2 Okapi BM25 Algorithm

The version of the Okapi BM25 algorithm used in these experiments is based on the description of the algorithm in Robertson [15], and in TREC notebook proceedings [16]. As with the LR algorithm, we have adapted the Okapi BM25 algorithm to deal with document components (including full documents):

$$\sum_{j=1}^{|Q_c|} w^{(1)} \frac{(k_1+1)tf_j}{K+tf_j} \frac{(k_3+1)qtf_j}{k_3+qtf_j} \quad (4)$$

Where (in addition to the variables already defined):

K is $k_1((1-b)+b\cdot dl/avcl)$

k_1, b **and** k_3 are parameters (1.5, 0.45 and 500, respectively, were used),

$avcl$ is the average component length measured in bytes

$w^{(1)}$ is the Robertson-Sparck Jones weight:

$$w^{(1)} = \log \frac{(\frac{r+0.5}{R-r+0.5})}{(\frac{n_{t_j}-r+0.5}{N-n_{t_j}-R-r+0.5})}$$

r is the number of relevant components of a given type that contain a given term,

R is the total number of relevant components of a given type for the query.

Our current implementation uses only the *a priori* version (i.e., without relevance information) of the Robertson-Sparck Jones weights, and therefore the $w^{(1)}$ value is effectively just an IDF weighting.

2.3 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 [10]. The blind feedback algorithm is based on the probabilistic term relevance weighting formula developed by Robertson and Sparck Jones [14].

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.

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:

$$w_t = log \frac{\frac{R_t}{R-R_t}}{\frac{N_t-R_t}{N-N_t-R+R_t}} \quad (5)$$

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.

3. RESULTS FOR NTCIR-8 OFFICIAL RUNS

Table 2: Submitted IR4QA Runs

RunID	Type	MAP	Mean Q	Mean nDCG
BRKLY-EN-JA-01-DN	EN⇒JA	0.36	0.38	0.59
BRKLY-EN-JA-02-T	EN⇒JA	0.35	0.36	0.54
BRKLY-JA-JA-01-DN	JA⇒JA	0.43	0.45	0.65
BRKLY-JA-JA-02-T	JA⇒JA	0.41	0.43	0.63
BRKLY-JA-JA-03-DN	JA⇒JA	0.16	0.16	0.31
BRKLY-JA-JA-04-DN	JA⇒JA	0.32	0.34	0.58
BRKLY-JA-JA-05-T	JA⇒JA	0.30	0.32	0.54

Table 2 shows the results for our official submitted runs for the IR4QA task. Among the Monolingual Japanese runs (type = JA⇒JA) the runs used different combinations of the algorithms described above. During the indexing process all of the data from the Mainichi newspaper database was segmented using the ChaSen segmentation software[1], and each segment was indexed as a "word". In addition a stoplist used in earlier NTCIR tracks was used to eliminate common words. For query processing each of the topic texts (QUESTION and NARRATIVE) were segmented using ChaSen for Monolingual runs, and the Japanese text generated by Google Translate[2] from the English text was segmented the same way for bilingual runs. Segmentation actually involved multiple steps since the UTF-8 topics had to be tranformed to EUC encoding for segmentation and then back to UTF-8 for matching with the database (stored as UTF-8).

We submitted 2 Bilingual runs (EN⇒JA) and 5 monolingual JA runs for our official entry. The following information and the information on performance measures in Table 2 is presented in the IR4QA overview paper in this volume [17]. The three effectiveness metrics for evaluating the IR4QA runs: Mean Average Precision (MAP), Q-measure (Mean Q) and a version of normalised Discounted Cumulative Gain (Mean nDCG) described in the overview paper[17]. The results shown here are from the deduplicated relevance data. The best performing run submitted by Berkeley was BRKLY-JA-JA-01-DN, which used probabilistic retrieval based on logistic regression (the TREC2 Algorithm above) with blind feedback on the QUESTION and NARRATIVE topic text. The next best performing (BRKLY-JA-JA-02-T) used the same algorithm and blind feedback approach, but used only the QUESTION text from the topic. BRKLY-JA-JA-03-DN, the worst performing of our monolingual runs, used the Okapi BM25 algorithm with no blind feedback. Interestingly another of the participating groups also used the Okapi algorithm and the results reported in the overview paper[17] are virtually identical for their runs and ours. The mid-performing runs BRKLY-JA-JA-04-DN and BRKLY-JA-JA-05-T use the logistic regression algorithm, but omit the blind feedback step for both the QUESTION and NARRATIVE topic text and the QUESTION alone, respectively.

[1] http://chasen-legacy.sourceforge.jp/
[2] http://translate.google.com/#

4. 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. Experiments in cross-language and patent retrieval at NTCIR-3 workshop. In *Proceedings of the Third NTCIR Workshop on research in Information Retrieval, Automatic Text Summarization and Question Answering, Tokyo, October 2002*, pages 173–182, 2002.

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

[5] 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.

[6] F. C. Gey. Chinese and korean topic search of japanese news collections. In *Proceedings of the Fourth NTCIR Workshop, Tokyo, June 2004*, 2004.

[7] F. C. Gey. How similar are chinese and japanese for cross-language information retrieval. In *Proceedings of the Fifth NTCIR Workshop, Tokyo, December 2005*, pages 171–174, 2005.

[8] F. C. Gey. Search between chinese and japanese text collections. In *Proceedings of the Sixth NTCIR Workshop, Tokyo, May 2007*, pages 173–182, 2007.

[9] F. C. Gey and R. R. Larson. Patent mining: A baseline approach. In *Proceedings of the NTCIR-7 Workshop Meeting, Tokyo, December 2008*, pages 358–361, 2008.

[10] 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.

[11] 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.

[12] 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.

[13] R. R. Larson and F. C. Gey. High baseline japanese information retrieval for question answering. In *Proceedings of the NTCIR-7 Workshop Meeting, Tokyo, December 2008*, pages 165–166, 2008.

[14] 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.

[15] S. E. Robertson and S. Walker. On relevance weights with little relevance information. In *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 16–24. ACM Press, 1997.

[16] S. E. Robertson, S. Walker, and M. M. Hancock-Beauliee. OKAPI at TREC-7: ad hoc, filtering, vlc and interactive track. In *Text Retrieval Conference (TREC-7), Nov. 9-1 1998 (Notebook)*, pages 152–164, 1998.

[17] T. Sakai, H. Shima, N. Kando, R. Song, C.-J. Lin, T. Mitamura, and M. Sugimoto. Overview of ntcir-8 aclia ir4qa. In *Proceedings of the NTCIR-8 Workshop, Tokyo, June 2010*, pages 000–000, 2010. This volume.