Text Retrieval Baseline for NTCIR-GeoTime

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

ABSTRACT

For the NTCIR-8 Workshop UC Berkeley participated in the GeoTime track and the IR4QA. For the GeoTime track we did both English and Japanese with both cross-language combinations. For the Japanese and translated English texts, we preprocessed the text using the ChaSen morphological analyzer for term segmentation. For GeoTime we used a time-tested logistic regression algorithm for document ranking coupled with blind feedback for most runs. 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.
NOTE THIS IS A SHORT DRAFT VERSION OF THIS PAPER

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-to-day 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 [4]. Our document ranking algorithm is a probability model based using the technique of logistic regression [1]. For all of our runs we used the TREC2 logistic regression model described in [4], both with and without 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 the New York Times and the Mainichi newpapers for the same time period. The papers were hoped to have common coverage of events that took place from the beginning of 2002 to the end of 2004 (as it turned out this was not always the case). 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-8 was an index that contained all of the words (or segmented tokens for Japanese) from the entire record.

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

3. SUBMISSIONS AND RESULTS FOR OFFICIAL RUNS

Table 1: Submitted GeoTime Runs

RunID	Type	MAP	Mean Q	Mean nDCG	B F
BRKLY-EN-EN-02-DN	EN⇒EN	0.40	0.42	0.61	Y
BRKLY-EN-EN-03-D	EN⇒EN	0.36	0.38	0.56	Y
BRKLY-EN-EN-04-DN	EN⇒EN	0.34	0.36	0.58	N
BRKLY-JA-JA-01-DN	JA⇒JA	0.25	0.26	0.40	Y
BRKLY-JA-JA-02-D	JA⇒JA	0.17	0.18	0.30	Y
BRKLY-JA-JA-03-DN	JA⇒JA	0.19	0.2	0.36	N
BRKLY-EN-JA-01-DN	EN⇒JA	0.33	0.34	0.53	Y
BRKLY-EN-JA-02-D	EN⇒JA	0.30	0.32	0.49	Y
BRKLY-JA-EN-01-DN	JA⇒EN	0.42	0.43	0.62	Y
BRKLY-JA-EN-02-D	JA⇒EN	0.38	0.39	0.56	Y

Table 1 shows the results for our official submitted runs for the GeoTime task. In examining the 1 table, some rather unusual results are apparent. First, and most striking, is that our cross-language runs (those with types of JA⇒EN or EN⇒JA) actually performed better that the corresponding monolingual runs (types JA⇒JA and EN⇒EN). Specifically for each pair of runs where the topic elements used and retrieval method was the same the bilingual runs outperformed

the monolingual runs, i.e.: BRKLY-EN-JA-01-DN outperforms BRKLY-JA-JA-01-DN, BRKLY-EN-JA-02-D outperforms BRKLY-JA-JA-02-D, BRKLY-JA-EN-01-DN outperforms BRKLY-EN-EN-02-DN and BRKLY-JA-EN-02-D outperforms BRKLY-EN-EN-03-D. This is exactly the opposite of what is usually observed in cross-language retrieval, where the bilingual almost always lags the monolingual in performance (and yes, we did doublecheck the submissions to be sure they didn't get switched).

for the full version of this paper we plan to do significance analyses of the differences.

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.

All of our submitted runs for the GeoTime track used probabilistic retrieval using TREC2 logistic regression algorithm described in detail our IR4QA paper [4]. Those runs with a "Y" in the BF column in table 1 used pseudo or blind relevance feedback along with the TREC2 algorithm, while those with "N" did not. For each runid in table 1 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 1 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 3 monolingual for each language as our official entries. The following information and the information on performance measures in Table 1 is presented in the GeoTime overview paper in this volume [2]. 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[2]. The best performing English run submitted by Berkeley was BRKLY-JA-EN-01-DN, which used probabilistic retrieval based on logistic regression (the TREC2 Algorithm above) with blind feedback on the DESCRIPTION and NARRATIVE topic text. The next best performing (BRKLY-EN-EN-02-DN) used the same algorithm and blind feedback approach, but used the original English topic text instead of the translated Japanese. BRKLY-EN-EN-04-DN, the worst performing of our English monolingual runs, omitted the blind feedback step during retrieval. A slightly different pattern of results is seen in table 1 for our Japanese submissions, with the translated English topics outperforming the native Japanese, but for Japanese the monolingual entry without blind feedback outperformed the DESCRIPTION-only queries.

4. 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 also hope to do some significance testing on the differences seen in the results. In the future we hope to try some additional segmentation approaches, particularly using the Jumon segmenter as an alternative to ChaSen for Japanese. Because these submissions were intended to form a baseline for comparison with methods using special geographic and temporal processing of the texts, we did not use any such methods for NTCIR-8. We plan, however to exploit some of special indexing tools developed for the Cheshire system in the future.

5. REFERENCES

[1] 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.

[2] 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.

[3] R. R. Larson. A fusion approach to XML structured document retrieval. *Information Retrieval*, 8:601–629, 2005.

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