Modification of Vocabulary-based Re-ranking for Geographic and Temporal Searching at NTCIR GeoTime Task

Kazuaki Kishida, Ikuko Matsushita
School of Library and Information Science, Keio University
2-15-45 Mita, Minato-ku,Tokyo,
JAPAN
kz_kishida@z8.keio.jp

ABSTRACT
This paper reports on experiments in the NTCIR-9 GeoTime task performed by a research group at the School of Library and Information Science in Keio University (KOLIS), which tried to explore techniques for searching a Japanese document collection for requests on geographic and temporal information. A special component of re-ranking for enhancing performance of geographic and temporal searches was added to the KOLIS system, in which standard Okapi BM25 and probabilistic pseudo-relevance feedback (PRF) were implemented. That is, at the first stage, a list of documents relevant to a given topic was specified by standard IR techniques, and at the second stage, the list was re-ranked after scores of documents which included geographic and temporal terms were increased. In particular, each score of documents including a syntactic pattern "geographic or temporal term + で" was augmented for improving search performance where "で" is a functional word meaning "at" or "in". In the old version of re-ranking used in the former NTCIR-8 workshop, only frequency of occurrence of geographic and temporal terms was taken into consideration. In this experiment of Japanese monolingual (JA-JA) retrieval and English to Japanese bilingual (EN-JA) retrieval, the search runs using jointly the re-ranking based on the syntactic pattern and PRF showed the highest performance. However, the ranking based on the syntactic pattern could not bring explicit improvement in comparison with the ranking technique at NTCIR-8.

Categories and Subject Descriptors
H.3.3 [**Information Systems**]: Information Search and Retrieval – retrieval models, search process.

General Terms
Experimentation; Performance; Measurement

Keywords
Japanese Monolingual Information Retrieval; Cross-lingual Information Retrieval; Geographic Information Retrieval

1. INTRODUCTION
The research group at School of Library and Information Science in Keio University (KOLIS) has participated in the GeoTime task at NTCIR-8 workshop, which challenged to enhance effectiveness of searching for geographic or temporal information [1]. In NTICR-8 workshop, simple vocabulary-based re-ranking was employed for increasing scores of documents containing intensively geographic and temporal terms listed in a special dictionary. A practical advantage of the vocabulary re-ranking is to implement it easily into the existing search systems. However, its effectiveness was not enough in the experiment at NTCIR-8 workshop although the re-ranking slightly improved performance [2].

The KOLIS group again participates in the GeoTime task at NTCIR-9 workshop [3], and attempts to modify slightly the vocabulary-based re-ranking technique by identifying a particular syntactic pattern, "geographic or temporal term + で" where "で" is a Japanese functional word indicating "at" or "in". It is expected that this syntactic pattern works as a mark representing that the document contains important geographic or temporal information. This paper reports an experimental result of the modification in Japanese monolingual search runs (JA-JA runs) and English to Japanese bilingual search runs (EN-JA runs) (i.e., the KOLIS group employs only Japanese document sets).

The KOLIS system basically consists of only a plain search engine based on a standard Okapi BM25. In addition, at NTCIR-9 workshop, a special component of re-ranking documents that were selected according to Okapi BM25 scoring is incorporated for increasing scores of documents that contains the syntactic pattern "geographic or temporal term + で" whereas only the different number of geographic and temporal terms were counted in the previous NTCIR-8 workshop. The method based on syntactic pattern remains still to be simple, and has also an advantage when implementing it into an existing system. Unfortunately, the method could not show clearly effectiveness in the experiment at NTCR-9 workshop. That is, in JA-JA runs, the syntactic pattern method outperformed very slightly the simple term counting, but in EN-JA runs, an opposite result was obtained.

2. RE-RANKING BASED ON SPECIFIC SUB-VOCABULARY
2.1 Two-stage Searching
In the NTCIR GeoTime task, documents including geographic or temporal information relevant to the search topics have to be

ranked higher in output lists. For example, the topic "GeoTime-0001" at NTCIR-8 is that "the user wants to know when and in what city the children's author Astrid Lindgren died" in which date and place information is assumed to be asked by the end-user. Since the target documents in Japanese used for the NTCIR Geo-Time task is a large set of news articles (for NTCIR-9 workshop, the Mainichi 1998-2005 news collection [3]), two-stage searching would be a realistic strategy. That is, at the first stage, a subset of documents relevant to the general subject of each topic is tried to specify, and at the second stage, documents including geographic and/or temporal information are identified from the subset by more detailed analysis.

While it is enough to apply a conventional IR technique at the first stage, a special technique can be applied for detecting documents that contains useful geographic or temporal information at the second stage. In the second stage, a complicated method based on machine learning theory may be used, but at the previous NTCIR-8 workshop, the KOLIS group adopted a simple vocabulary-based method counting the number of geographic and temporal terms identified by using a special dictionary [2].

2.2 Re-ranking by Term Counting

In the experiment at NTCIR-8 workshop, a geographic dictionary in the ChaSen system that is a well-known Japanese morphological analyzer [4] was employed. That is, Noun.place file of IPADIC ver 2.6.3 was incorporated into the KOLIS system as a special dictionary, and 'geographic terms' were operationally defined as those appearing in the Noun.place file. At the second stage, after the different numbers of the 'geographic terms' that occur actually in 1000 documents specified at the first stage were counted, each document score was modified such that

$$v'_i = v_i \times \left(1.0 + 0.5 \times \frac{x_i}{\max_{k=1,\ldots,1000} x_k} \right) \quad (1)$$

where

v_i : Original document score of *i*-th document in the output,

v'_i : Modified document score of *i*-th document in the output, and

x_i : The different number of geographic terms in *i*-th document in the output.

The final output was sorted in descending order of the modified document score. It should be noted Japanese representations indicating a specific year (00 to 99 years, i.e., "00 年" to "99 年") and a month (January to December, i.e., "1 月" to "12 月") were compulsorily added into the Noun.place in this experiment. Therefore, the re-ranking was executed based on not only geographic terms but also temporal representations.

2.3 Re-ranking by Syntactic Pattern

After NTCIR-8 workshop, a member of the KOLIS group checked carefully 100 top-ranked documents by the term counting method by each search topic, and reached to a conclusion that a syntactic pattern "geographic or temporal term + で" can work as an effective mark to identify relevant documents provided by the organizers of NTCIR-8 GeoTime task. It is possible that the conclusion is theoretically valid because "で" is a Japanese functional word indicating explicitly place or location such as "in" or "at" in English, and the term counting method at NTCIR-8 workshop ignores this information and treats all geographic and temporal terms equivalently.

In NTCIR-9 workshop, the syntactic pattern was counted at the second stage, and each document score was modified such that

$$v'_i = v_i \times \left(1.0 + 0.5 \times \frac{\tilde{x}_i + 3.0 z_i}{\max_{k=1,\ldots,1000} (\tilde{x}_k + 3.0 z_k)} \right) \quad (2)$$

where

$\tilde{x}_i$: The total number of occurrences of geographic and temporal terms in *i*-th document in the output,

z_i : The total number of occurrences of the grammatical patter "geographic terms or temporal terms + で" in *i*-th document in the output, and

the value 3.0 was empirically determined through checking performance of search runs on NTCIR-8 GeoTime test collection.

3. IR System

3.1 Indexing

In this experiment, Japanese texts of search topics and of documents were segmented based on a hybrid indexing technique, in which all word segments identified by

- character-based overlapped bi-gram technique, and
- longest matching with a machine-readable dictionary

were adopted as index terms. The indexing technique is the same with that at NTCIR-8 workshop (for detail, see [2]).

3.2 Document Scoring and Pseudo-relevance Feedback

Like the experiment at NTCIR-8 workshop, a standard Okapi BM25 [5] was again used for computing document scores at the first stage, and a standard pseudo-relevance feedback (PRF) technique based on a probabilistic term weighting [6] was applied in some search runs (more specifically, in our system, 10 new terms which have the highest term weights among those in top-ranked 30 documents are added to the set of original query terms).

3.3 Bilingual Searching

For English to Japanese (EN-JA) bilingual searching, the text of each search topic was simply entered into machine translation (MT) systems provided by Yahoo! Japan [7], Excite Japan [8] and Google [9] (Google was not used at NTCIR-8 workshop). Translation results from all the MT services for each search topic were straightforwardly concatenated and treated as a set of sentences representing the topic in Japanese language. After that, search runs were executed in the same manner with JA-JA monolingual searches (see also [2]).

4. Experiment

4.1 Submitted Runs

Table 1 shows an outline of 10 search runs submitted officially to NTCIR-9 GeoTime organizers. The baseline searches are KOLIS-JA-JA-D-04 (for Japanese monolingual search) and KOLIS-EN-JA-D-04 (for English-Japanese bilingual search) in which any re-ranking and PRF were not applied.

Re-ranking and PRF techniques were additionally applied to the baseline searches as follows.

- KOLIS-JA-JA-D-01 and KOLIS-EN-JA-D-01: Modified re-ranking based on syntactic pattern in Equation (2) (NTCIR-9 version).
- KOLIS-JA-JA-D-02 and KOLIS-EN-JA-D-02: Re-ranking based on only term counting in Equation (1) (old NTCIR-8 version)
- KOLIS-JA-JA-D-03 and KOLIS-EN-JA-D-03: PRF was added to baseline search.
- KOLIS-JA-JA-D-05 and KOLIS-EN-JA-D-05: both PRF and re-ranking based on syntactic pattern were added. More precisely, after identifying 1000 documents by PRF, re-ranking technique using Equation (2) was applied in these runs.

In all the above search runs, only <description> filed was used as search topic.

Table 1. Search runs submitted from KOLIS group

Run Type	ID	Topic field	Re-Ranking	PRF
JA-JA	KOLIS-JA-JA-D-01	D	Yes (New)	No
	KOLIS-JA-JA-D-02	D	Yes (Old)	No
	KOLIS-JA-JA-D-03	D	No	Yes
	KOLIS-JA-JA-D-04	D	No	No
	KOLIS-JA-JA-D-05	D	Yes (New)	Yes
EN-JA	KOLIS-EN-JA-D-01	D	Yes (New)	No
	KOLIS-EN-JA-D-02	D	Yes (Old)	No
	KOLIS-EN-JA-D-03	D	No	Yes
	KOLIS-EN-JA-D-04	D	No	No
	KOLIS-EN-JA-D-05	D	Yes (New)	Yes

4.2 Results and Discussions

4.2.1 Basic statistics

The Japanese document collection of NTCIR-9 GeoTime task (i.e., The Mainichi 1998-2005 news collection) consists of 797,700 records, from which 3,289,998 distinct terms were identified by the indexing method of KOLIS system. The average document length was 378.51.

4.2.2 Effect of re-ranking and PRF

Table 2 shows scores of mean average precision (AP), mean Q-measure (Q) and mean nDCG@1000 for 25 search topics (these scores are officially provided by the NTCIR GeoTime organizers).

As the scores indicate, it turns out that search runs using jointly re-ranking and PRF achieved the highest performance in both JA-JA and EN-JA tasks (i.e., KOLIS-JA-JA-D-05 and KOLIS-EN-JA-D-05). This is the same finding with that in NTCIR-8 workshop.

In comparison between new version (-01) and old version (-02) of re-ranking techniques, new version outperforms slightly old version in JA-JA runs. On the other hand, a reverse result was obtained in EN-JA runs. Unfortunately, it seems that modification of re-ranking technique using syntactic pattern did not enhance explicitly effectiveness in this experiment.

Table 2 Performance of search runs (official results)

Run Type	ID	MAP	Q	nDCG @1000
JA-JA	KOLIS-JA-JA-D-01	0.3860	0.4180	0.6111
	KOLIS-JA-JA-D-02	0.3815	0.4178	0.6042
	KOLIS-JA-JA-D-03	0.3996	0.4279	0.6027
	KOLIS-JA-JA-D-04	0.3502	0.3822	0.5715
	KOLIS-JA-JA-D-05	0.4227	0.4540	0.6294
EN-JA	KOLIS-EN-JA-D-01	0.2799	0.3165	0.4950
	KOLIS-EN-JA-D-02	0.2833	0.3212	0.5029
	KOLIS-EN-JA-D-03	0.2749	0.3065	0.4840
	KOLIS-EN-JA-D-04	0.2679	0.3022	0.4842
	KOLIS-EN-JA-D-05	0.2837	0.3175	0.4971

Also, in comparison with re-ranking (-01, -02) and PRF (-03) methods, the result in JA-JA runs is different from that in EN-JA runs. That is, re-ranking outperforms PRF in EN-JA runs, but not in JA-JA runs. This experiment did not show effectiveness of re-ranking in comparison with PRF.

Therefore, it can be concluded from the experiment that search performance is better in the following order;

JA-JA: Baseline < Re-ranking (old) < Re-ranking (new) < PRF < (PRF + Re-ranking).

EN-JA: Baseline < PRF < Re-ranking (new) < Re-ranking (old) < PRF < (PRF + Re-ranking)

4.2.3 Performance of bilingual IR

In comparison of performance between bilingual (EN-JA) and monolingual (JA-JA) searches, monolingual runs outperforms naturally bilingual ones (see Table 2). MAP of the best bilingual runs (KOLIS-EN-JA-D-05) amounts to about 67% of the best monolingual runs (KOLIS-JA-JA-D-05). In NTCIR-8, performance of EN-JA runs was more closed to that of JA-JA runs (MAP score of the best bilingual run is about 97% of the best monolingual run in NTCIR-8).

5. CONCLUDING REMARKS

In GeoTime task at NTCIR-9 workshop, KOLIS group tried to modify the own re-ranking technique by checking syntactic pattern "geographic or temporal term + で" in each document. Unfortunately, although this technique showed higher performance than baseline run, the improvement is not enough in comparison with the re-ranking technique used in the former NTICR-8 workshop.

6. ACKNOWLEDGMENTS

The authors thank to the organizers for planning and managing of the NTCIR-9 GeoTime task.

7. REFERENCES

[1] Gey, F., Larson, R., Kando, N., Machado, J., and Sakai, T. 2010. NTCIR-GeoTime overview: evaluating geographic and temporal search. In Proceedings of NTCIR-8 (Tokyo, Japan, June 15-18, 2010).

[2] Kishida, K. Vocabulary-based re-ranking for geographic and temporal searching at NTCIR GeoTime task. In Proceedings of NTCIR-8 (Tokyo, Japan, June 15-18, 2010)

[3] Gey, F., Larson, R., Machado, J., Yoshioka, M. NTCIR9-GeoTime overview – evaluating geographic and temporal search: Round 2. In Proceedings of NTCIR-9 (Tokyo, Japan, December 6-9, 2011).

[4] http://chasen.aist-nara.ac.jp/chasen/distribution.html.en

[5] Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M., Gatford, M. 1995. Okapi at TREC-3. In Overview of the Third Text REtrieval Conference (TREC-3). National Institute of Standards and Technology, Gaithersburg.

[6] Robertson, S. E. and Sparck Jones, K. 1994. Simple, Proven Approaches to Text Retrieval, Technical Report No.356, Computer Laboratory, University of Cambridge.

[7] http://honyaku.yahoo.co.jp/

[8] http://www.excite.co.jp/world/

[9] http://translate.google.co.jp/