NTCIR-6 CLIR Experiments at Osaka Kyoiku University — Term Expansion Using Online Dictionaries and Weighting Score by Term Variety —

Takashi SATO
Osaka Kyoiku University
4-698-1 Asahigaoka, Kashiwara, Osaka, Japan
sato@cc.osaka-kyoiku.ac.jp

Abstract

This paper describes experimental results of J-J subtask of NTCIR-6 CLIR. We expanded query term using online dictionaries in a WEB. It was effective for some topics of which average precision was low. Probabilistic model were employed for scoring, and we modified this score multiplying by the number of varieties of query terms, also. In most cases this works well. Query term reduction should be considered if this modified scoring fails.

Keywords: *gram-based index, term expansion, term variety, NTCIR*

1 Introduction

We participated one of traditional NTCIR task, CLIR. We made two gram-based indices, namely indices for HEADLINE and TEXT tag extracted from test collection for J-J subtask. Since gram-based indices are able to index all strings in target text, words that are not found in dictionaries, are also indexed. We used words in TITLE and DESC tag of search topics as queries. Then we expanded query term using free online dictionaries in a WEB. It was effective for some topics of which average precision was low. Probability model were employed for scoring, and we modified this score multiplying by the number of varieties of query terms, also. In most cases this worked well. Query term reduction should be considered if this modified scoring fails.

2 Indexing

We made two indices (HEADLINE and TEXT index) as inverted files of *n*-grams for each of 1st and 2nd stage corpus of J-J subtask. While the length of gram *n* is varying from gram by gram, grams are coded in fixed byte (6 byte in the task)[1]-[3]. Corpus for 1st stage is Mainichi 2000-2001 and Yomiuri 2000-2001 (858,400 documents). Mainichi 1998-1999 and Yomiuri 1998-1999 (596,058 documents) are added for 2nd stage.

Table 1 shows the size of corpus, extracted tag fields and two indices, which are made from HEADLINE and TEXT tag field. Index size overhead against extracted tag fields is 159% for 1st stage and 166% for 2nd stage added. Table 2 shows time to make indices. Computer used is an ATX compatible machine (CPU: Pentium4 1.6GHz, Memory: 512MB).

3 Term Extraction and Expansion

Query terms are extracted from TITLE and DESC tag fields in J-J subtask topics. Each compound word are segmented in words, and all combinations of these words are also made. After our submission of runs, we tried to expand terms manually using definition part of online

dictionary in a WEB (such as Wikipedia[4] and Yahoo dictionaries[5]) because we noticed that some technical terms have important synonyms, which are neither included in a topic nor best rank documents retrieved even if we use pseudo-relevance feedback. These expansions may be done automatically using namely part such as '()' or '/' in dictionaries.

Table 1. Size of corpus, tag fields and indices

stage	1st	2nd added
corpus size	1.00GB	777MB
<HEADLINE> tag	46.3MB	32.7MB
<TEXT> tag	854MB	657MB
<HEADLINE> index	110MB	80.8MB
<TEXT> index	1.48GB	1.17GB

Table 2. Time to make indices

stage	1st	2nd added
<HEADLINE>	1.54min	1.09min
<TEXT>	26.8min	20.6min
total	28.3min	21.7min

4 Ranking

We retrieved query terms obtained by section **3** from HEADLINE and TEXT index. Then we ranked documents using probabilistic model[6]. In our formal submitted run, we prepare another run in which each document score is multiplied by term variety factor (TVF) i.e. the number of query term appeared in the document (t_a) divided by the number of query terms for a topic (t_t). For example the number of query terms for a topic is t_t=5 and the number of query terms appeared in a document is t_a=3 out of 5. Then score of the document for the topic is multiplied by 0.6(=3/5). Scoring documents for two indices, we merged the score by simple addition of document score of both indices.

5 Results

We submitted 5 runs for 1st stage. Table 3 shows the combination of query term set (TITLE and DESC) for HEADLINE and TEXT index. 'OKSAT-J-J-' is abbreviated in run-id column. Last two runs i.e. D-04 and T-05 are scored multiplying by term variety factor (TVF in the table) as described in **4**.

Table 3. Submitted runs for 1st stage (OKSAT, J-J)

run-id	HEADLINE	TEXT	TVF
D-01	DESC	DESC	no
TD-02	TITLE	DESC	no
T-03	TITLE	TITLE	no
D-04	DESC	DESC	yes
T-05	TITLE	TITLE	yes

Figure 1 shows relationship between topic and average precision of relax evaluation of D-01 and D-04. In this figure topics are re-ordered by their average precision in descendent order. Average of average precision over evaluated topics of D01 run is 0.240, and that of D04 run is 0.268. In most topics, D04 (TVF runs) are better then D01 (normal one).

We made two post-submission runs D01' and D04'. These are term-arranged version of D-01 and D04. We expanded terms using definition part of online dictionary as described in **3**. On the other hand we reduce the number of terms for topics whose average precision of D04 (TVF

version) is lower than that of D01 (normal version). Figure 2 shows average precision of D01' and D04'. Topics are re-ordered again since average precision changed. Average of average precision of D01' is 0.302, and that of D04' is 0.327.

6 Discussions

Comparing D01 with D04 run, we observe that TVF multiplication is effective. As for term expansion using online dictionary is effective in some topics. The followings are success example.

Topic#020: "西暦 2000 年問題" => "Y2K"

Topic#070: "逆エルニーニョ現象"
=> "ラニーニャ現象"

Right side term of '=>' is expanded from left side term. Addition of these synonyms is very effective. On the other hand, the followings are failure example.

Topic#043: "デリバティブ"
=> "先物取引", "オプション取引",
"スワップ取引"

Right side terms are at lower rank of left term. Expanded terms may be too detail for general newspaper article in this case.

We reduce the number of terms for topics whose average precision of TVF version (D04) is lower than that of normal probabilistic version (D01). More concretely, we delete top popular terms in corpus for these topics. Table 4 shows this effect of topic #024 and #050. In this table D01' is query term reduction run of D01 and D04' is that of D04 respectively. From this table we observe that query term reduction may be worth considering if TVF version is not good.

7 Conclusions

We experimented term expansion using online dictionaries. It was effective for some topics of which average precision was low. We also tried to weight score by query term variety factor (TVF). In most cases this worked well. Query term reduction should be considered if TVF scoring fails.

Table 4. Reduction of query terms

topic#	query terms	run ID	average precision
24	5	D01	0.179
		D04	0.126
	2	D01'	0.307
		D04'	0.282
50	16	D01	0.521
		D04	0.349
	6	D01'	0.605
		D04'	0.732

References

[1] Sato, T., Fast full text search with free word using TS-file, *Proc. 19th ACM SIGIR Conf.*, p.342 (1996).

[2] Sato, T., Fast full test retrieval using gram based tree structure, *Proc. ICCPOL '97*, Vol.~2, pp. 572-577 (1997).

[3] Sato, T. *et al.*, Gram based full test search system and its application, IPSJ SIG Notes, 98-DBS-114-2 (1998).

[4] http://ja.wikipedia.org

[5] http://dic.yahoo.co.jp

[6] Robertson, S.E. and Walker, S., Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval, *Proc. 17th Int. Conf. Research and Development in Information Retrieval*, pp. 232-241 (1994).

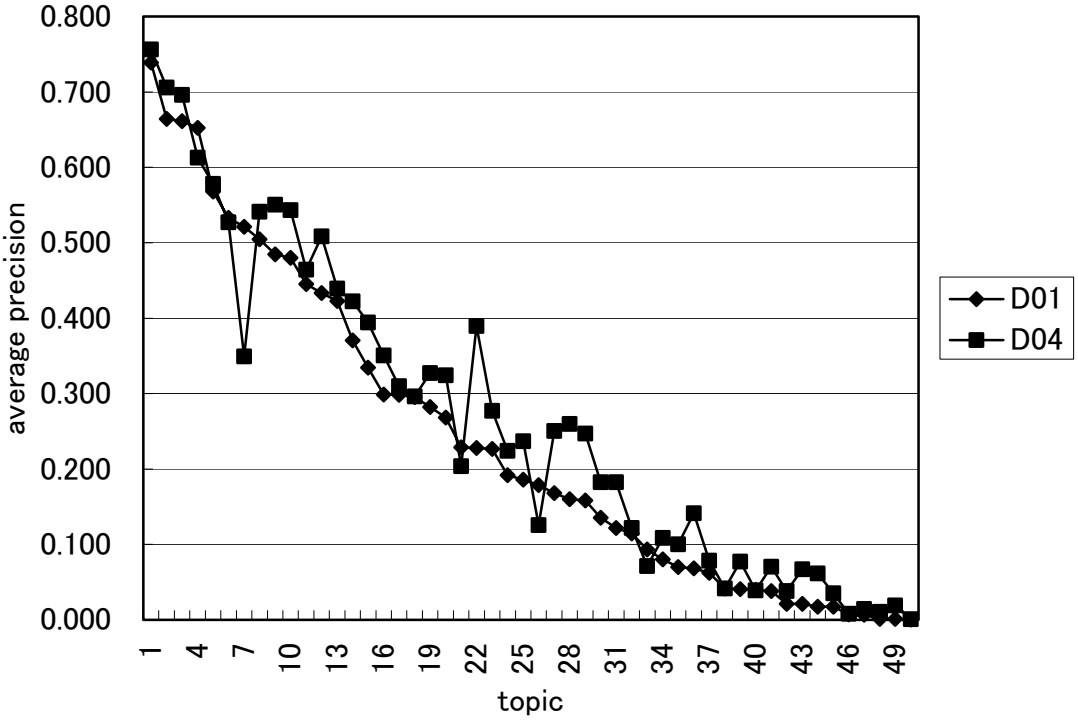

Figure 1. Average precision of submitted run D01 and D04

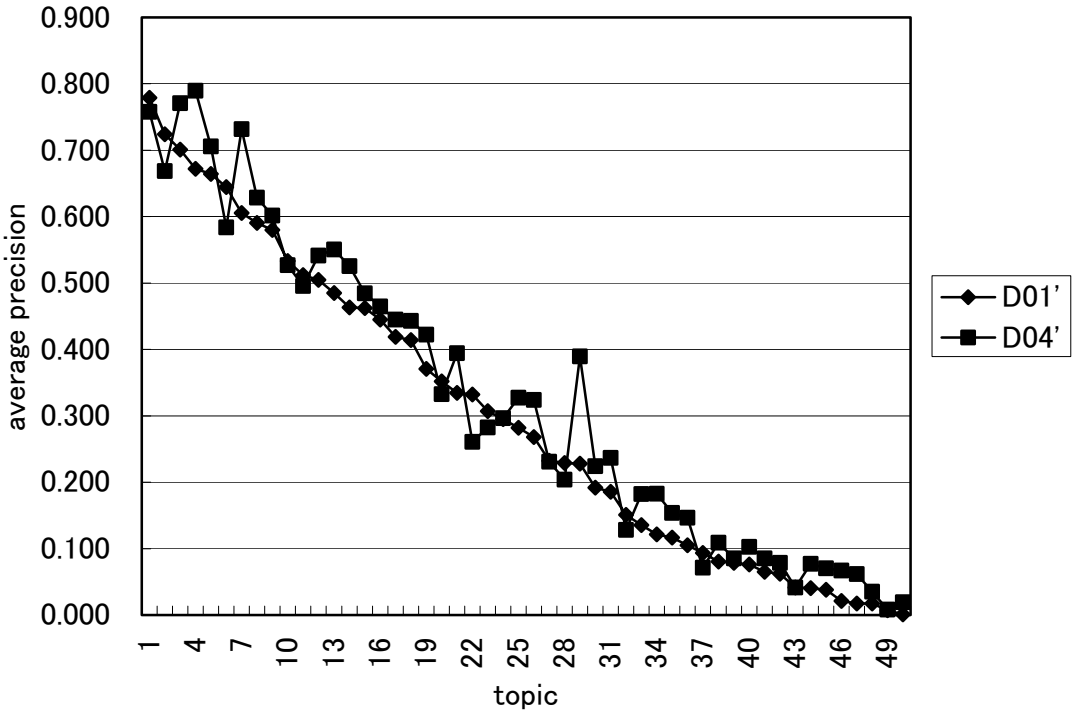

Figure 2. Average precision of post-submission run D01' and D04'