NTCIR-9 GeoTime at Osaka Kyoiku University - Toward Automatic Extraction of Place/Time Terms -

SATO, Takashi
Information Processing Center
Osaka Kyoiku University
4-698-1 Asahiga-oka
Kashiwara, Osaka, JAPAN
+81-72-978-3823
sato@cc.osaka-kyoiku.ac.jp

ABSTRACT

Our approach to NTCIR-9 Geotime was to obtain place/time information about topics from Wikipedia and Google using query terms extracted from topics. Adding this information to query terms, we retrieved documents using <TEXT> tag index and scored them. In addition, we compared <DATE> tag of searched documents with time information, weighted the score value of documents retrieved, and ranked them. Although the automation of extraction of place/time remains for future research, the validity of the method was confirmed from the comparison of evaluation results with runs which do not use these place/time information.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval – *Information filtering, Query formulation, Retrieval models, Search process.*

General Terms

Experimentation, Performance, Measurement.

Keywords

Information Retrieval, Index, Place/Time term.

Team Name: OKSAT

External Resources Used: Wikipedia, Google

1. INTRODUCTION

We have participated in NTCIR9 GeoTime for the second time following our participation in NTCIR8 GeoTime. Last time, we used Who's Who and postal code etc., for automatic information retrieval without human intervention. After our having combined query terms and the WEB search which were extracted from topic, we specified place/time as much as possible this time. After this information was added to query terms, we searched the corpus.

Our WEB search included Wikipedia and Google. For Wikipedia search, GikiCLEF [1] the CLEF 2009 track provided a precedent. Although we search specifically for place/time, these searches are included in named entity search [2, 3] in general.

From the evaluation results of a total 10 runs in English/Japanese, the effectiveness of adding place/time information from WEB search and retrieving corpus was confirmed.

2. INDEXING

For each English (EN) and Japanese (JA) corpus, we extracted the text surrounded by the <TEXT> tag and created the index for search. Since the index is based on *n*-gram [4], the morphological analysis was unnecessary as the case of our NTCIR-8 GeoTime.

Table 1 shows the size of TEXT indices, their overhead compared with original data, and creation time. The reason why the overhead and creation time of English index are bigger than that of Japanese is the characters constitute English documents are one byte whereas those of Japanese documents are one and two bytes (mostly two bytes). Since the index entries of the number of characters are created in *n*-gram coding, compared with the portion of a one byte character, as for the portion of a two bytes character, the number of entries are one half.

Table 1. Statistics of TEXT Indices

	English	Japanese
size(MB)	4,636	1,536
overhead(%)	202	151
time(min.)	31.5	7.6

In addition, we extracted <DOCNO> and <DATE> tag. Those sizes are 14.25MB for English <DOCNO>, 9.89MB for Japanese <DOCNO>, 5.56MB for English <DATE>, and 8.57MB for Japanese <DATE>.

3. RETRIEVAL

3.1 Outline of Retrieval

We made the following four different searches.

[a] Keyword Search of TEXT tag
[b] Place Search of TEXT tag
[c] Time Search of TEXT tag
[d] Time Search of DATE tag

The procedure is as follows.

Preparation

(1) From <NARRTIVE> tag of each topic, we extracted query terms. (automatic)

(1)' From <DESCRIPTION> tag of each topic, we extracted query terms. (automatic)
(2) Retrieving Wikipedia and Google by query terms, we get place/time information. (automatic + manual)
(2)' From the output of (2), we extracted time information only. (automatic)

Retrieval and Scoring

(3), (3)* Retrieving <TEXT> tag index using place/time of (1) and (2), we scored documents retrieved. (difference of (3) and (3)* is described later.)
(3)' Retrieving <TEXT> tag index using place/time of (1)' and (2), we scored documents retrieved.
(3)" Retrieving <TEXT> tag index using (1), we scored documents retrieved.
(4) Retrieving <DATE> tag using time of (2)', we set time multiplier.
(5) We multiplied the score of (3), (3)', and (3)" by multiplier of (4).

The scoring of (3), (3)', and (3)" is simplified probabilistic model by tf-idf [5]. The multiplier of (4) which is a function of the difference between two date ((2)' - <DATE>) is shown in Table 2.

Table 2. Difference of Days versus Multiplier

Difference of Days	Multiplier
0-2	2.0
3-4	1.6
5-7	1.4
8-19	1.2
others	1.0

Our group submitted five runs for each of English and Japanese corpus. Runs which have same number (01, 02, 03, 04, and 05) in run name were processed in the same manner. The retrieving process of each run was the combination of the above procedures (1) to (5) as shown in Table 3.

Table 3. Combination of procedures for submitted runs

Run Name	Method
OKSAT-{EN-EN|JA-JA}-01-DN	(1) (2) (3) (4) (5)
OKSAT-{EN-EN|JA-JA}-02-DN	(1) (2) (3)* (4) (5)
OKSAT-{EN-EN|JA-JA}-03-D	(1)' (2) (3)' (4) (5)
OKSAT-{EN-EN|JA-JA}-04-DN	(1) (2) (3)
OKSAT-{EN-EN|JA-JA}-05-DN	(1) (3)"

We have to notify that the treatment of text length of (3)* in Table3 is different from those of others ((3), (3)', and (3)"). For documents, the length of which is more than a particular fixed length (400byte for English, 200byte for Japanese), tf was divided by square root of (document length / fixed length) for (3)*, however, tf was divided by log of (document length / fixed length) for others. In addition, since processing of (1), (2), and (3) is common, they were executed only once.

Upper three in Table 3 were runs which used place/time terms extracted from Wikipedia and Google. On the other hand, lower two were runs which did not use place and/or time terms. In other words, they were prepared for the baseline model.

Since our corpus consists of newspaper articles, query terms about date were modified. For English newspapers, the date of less than one week from article date (<DATE> tag) is referred by the day of the week. In both English and Japanese newspaper, month (and year) was omitted for the date of the same month (and year) as article date.

3.2 Topic by Topic (in Retrieval)

In this section we elaborate on the retrieval process of some example topics for methods (1) and (2) in section **3.1** above.

<TOPIC ID="GeoTime-0026">
<![CDATA[Where and when did the space shuttle Columbia disaster take place?]]>

GoeTime-0026 is a comparatively easy query.

(1) Query terms “space shuttle Columbia disaster” etc. were extracted.
(2) We retrieved Wikipedia using (1), and got the page describing the accident.
Title: Space Shuttle Columbia disaster
URL:http://en.wikipedia.org/wiki/Space_Shuttle_Columbia_disaster

The first paragraph is quoted below.
[The Space Shuttle Columbia disaster occurred on February 1, 2003, when shortly before it was scheduled to conclude its 28th mission, STS-107, the Space Shuttle Columbia disintegrated over Texas during re-entry into the Earth's atmosphere, resulting in the death of all seven crew members. Debris from Columbia fell to Earth in Texas along a path stretching from Trophy Club to Tyler, as well as into parts of Louisiana.]

We extracted the following place/time.

Place: Texas, Louisiana
Time (Date): February 1, 2003

Since we did not have a suitable geographical dictionary, we extracted “Texas” and “Louisiana” ***manually***. About Time (Date), we extracted automatically using the regular expression.

Time was automatically changed into “Saturday” (the day of the week of February 1, 2003) for <TEXT> retrieval, and “2003-02-01” for <DATE>.

Although the above process was an example of English topic, it becomes as follows for the equivalent Japanese topic.

<TOPIC ID="GeoTime-0026">
<![CDATA[いつ、どこで、スペースシャトルコロンビアの事故は起こった？]]>

(1) Query term: スペースシャトルコロンビア 事故
(2) Title: コロンビア号空中分解事故
URL:http://ja.wikipedia.org/wiki/%E3%82%B3%E3%83%AD%E3%83%B3%E3%83%93%E3%82%A2%E5%8F%B7%E7%A9%BA%E4%B8%AD%E5%88%86%E8%A7%A3%E4%BA%8B%E6%95%85
[コロンビア号空中分解事故（コロンビアごうくうちゅうぶんかいじこ）は、2003 年 2 月 1 日、アメリカ合衆国の宇宙船スペースシャトル「コロンビア号」が大気圏に再突入する際、テキサス州上空で空中分解し、7 名の宇宙飛行士が

犠牲になった事故である。コロンビアは、その 28 回目の飛行である STS-107 を終え、地球に帰還する直前であった。]

Place:テキサス州
Time(Date):2003 年 2 月 1 日

About place, even if there was no geographical dictionary, automation extraction of "Texas" was carried out using a suffix "州(state)". Time is automatically changed into "1 日" (day part of 2003 年 2 月 1 日) for <TEXT> retrieval.

<TOPIC ID="GeoTime-0027">
<![CDATA[When was the last flight of Concorde and where did it land?]]>
<![CDATA[コンコルドが最後に飛行したのはいつですか？また、どこが目的地でしたか？]]>

The page of "Concorde" was retrieved from Wikipedia, however, because most information was not related to the last flight, it was difficult to extract exact place/time information. On the other hand, searching Google by words "コンコルド(Concorde) 最終(last) 飛行(fright)", the top ranked page of the retrieved list was as follows.

(1) Query terms: コンコルド, 最終, 飛行
(2) Title:ンコルド最終便
URL:
http://www2g.biglobe.ne.jp/aviation/concorde031026.html
「コンコルドの最後の飛行がおこなわれた。10 月 24 日、ニューヨークからの定期便が午後 4 時過ぎロンドン・ヒースロウ空港に着陸したのがそれであった。空港の周辺には何百とか何千もの人びとが集まって、最後の着陸進入を見守ったという。」

Place:ニューヨーク, ロンドン. ヒースロウ空港
Time:10 月 24 日

Concerning place, it is not difficult to identify "ニューヨーク(New York)" and "ロンドン(London)" using a city database. Extraction of "ヒースロウ空港(Heathrow Airport)" is additionally possible by searching for a keyword, followed by the literal string "airport". About time, adding year 2003 from date of web page written (2003.10.27), 2003-10-24" was obtained for <DATE>.

<TOPIC ID="GeoTime-0037">
<![CDATA[What fatal accident occurred near (geographical coordinates 5° 52′ 12″ N 5° 45′ 00″ E / 5.870°N 5.750°E / 5.870; 5.750), which killed hundreds of people, and when did it occur?]]>
<![CDATA[北緯 5 度 52 分 12 秒東経 5 度 45 分の近くで起きた数百人の死亡者を出した事故は、どのような事故ですか？また、それはいつ起きましたか？]]>

This topic required spatial reasoning to look up places near the geographic coordinates. We retrieved from Google maps by the geographic coordinates given, we then obtained manually the name of nearby city and the name of the country. Adding these place names to the query terms, we retrieved <TEXT> tag index. If we could have used a city database which included geographic coordinates, automatic retrieval might have been possible.

<TOPIC ID="GeoTime-0046">
<![CDATA[Where and when did presidential debate between Bush and Kerry hold?]]>
<![CDATA[ブッシュとケリーの大統領選の討論会は、いつ、どこで、行われましたか？]]>

When there were two or more possible candidate retrieval days (three days in this case), we utilized the highest multiplier calculated by Table 2.

3.3 Corpus by Corpus Comparison

The English corpus of NTCIR9 GeoTime task consists of four newspapers. In processing topics of this task, many long and explanatory articles were searched with New York Times. On the other hand, short articles, such as world news in a week, have often hit in Xinhua English. Scoring which takes account of the length and type of articles might lead to better retrieval results.

Ranges of year are different by corpus. In processing to retrieve incident (not scheduled) topics, the corpus (newspaper) published before incident year does not need to be searched.

Time differences of the country in which newspapers are published should be considered when <DATE> tags are referred to. For example, an incident in U.S. becomes newspaper article published in Asia from the next day because of time difference.

4 RESULTS

4.1 Run by Run

Figure 1 shows average precision (AP) topic by topic of submitted English runs. Figure 2 shows that of submitted Japanese runs. In this table run names are simplified by using only those middle parts of the run name. Thus {EN|JA}-{01|02|03|04|05} stands for OKSAT-{EN-EN|JA-JA}-{01|02|03|04|05}-{DN|D}. For example EN-01 stand for OKSAT-EN-EN-01-DN.

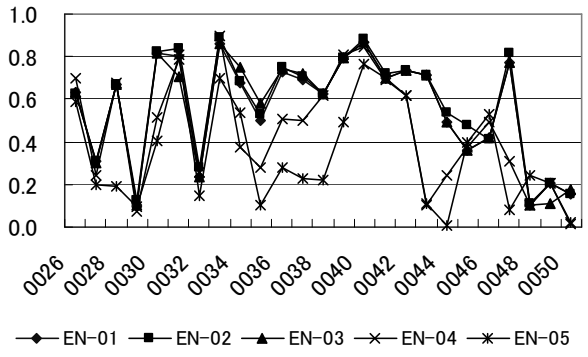

Figure 1. Topic ID vs. AP of Submitted English Runs

The mean average precision (MAP) of OKSAT-EN-EN-01-DN and OKSAT-EN-EN-02-DN, which retrieved TEXT and DATE

tags of the corpus by place/time terms from the internet, was good to the same extent. Although they extracted query terms from DESCRIPTION and NARRATIVE tags, the MAP of OKSAT-EN-EN-03-D, which extracted query terms from DESCRIPTION only, was good too. The MAP of OKSAT-EN-EN-04-DN and OKSAT-EN-EN-05-DN was lower than MAP of the above three runs. We thus conclude that DATE tag comparison with the time information searched from the internet was effective. From the results showing the MAP of OKSAT-EN-EN-04-DN better than that of OKSAT-EN-EN-05-DN (although the difference was not so large), we think that it was effective to have pinpointed the place using place information from the internet.

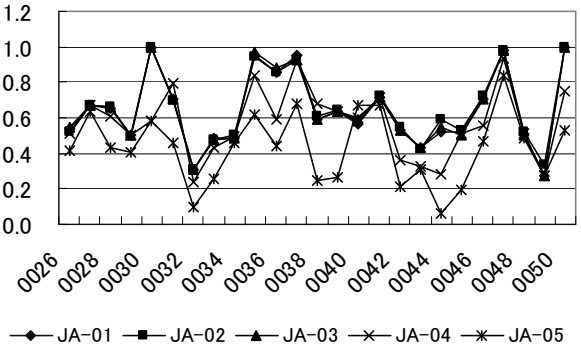

Figure 2. Topic ID vs. AP of Submitted Japanese Runs

The MAP of Japanese runs was better than that of English, although their tendencies from run to run were similar. The MAP of the top 3 runs of English runs was OKSAT-EN-EN-02-DN, OKSAT-EN-EN-03-DN, and OKSAT-EN-EN-01-DN at the descending order of MAP. Whereas that of Japanese runs was OKSAT-JA-JA-02-DN, OKSAT-JA-JA-01-DN, and OKSAT-JA-JA-03-DN. However, the difference between runs was too small and that we could not order runs by MAP.

4.2 Topic by topic (Results)

There are two types of topic about time (date). One is *incident type*, that is, its time is not expected in advance. For example GeoTime-0026 ([Where and when did the space shuttle Columbia disaster take place?]) and GeoTime-0047 ([A cable train fire in a European country killed 155 people. When and in which country?]) are this type. The other type of topic is *scheduled type*, that is, time of topic is known in advance. For example GeoTime-0029 ([When was the euro put in circulation and which three member states of the eurozone by that time declined its use?]) and Geotime-0041 ([When was control of the Panama Canal returned to Panama?]). For the incident type our DATE search works well. The scheduled type is not performed as well. The date multiplier of Table 2 works well because the incident type articles are written after usually near the day of the incident in newspaper, on the other hand, the scheduled type articles are not so.

4.3 English vs. Japanese

In relation to comparison of English and Japanese runs, the relation between OKSAT-EN-EN-01-DN and OKSAT-JA-JA-01-DN is shown in Figure 3 as an example.

In some topics, average precision of English and Japanese differed substantially. For example, average precision of GeoTime-0029, GeoTime-0048, and Geoime-0050 in Japanese is 0.5 and substantially higher, whereas in English it is from 0.1 to 0.15. It seems that it originates in the difference in the relation between topic and corpus by English and Japanese.

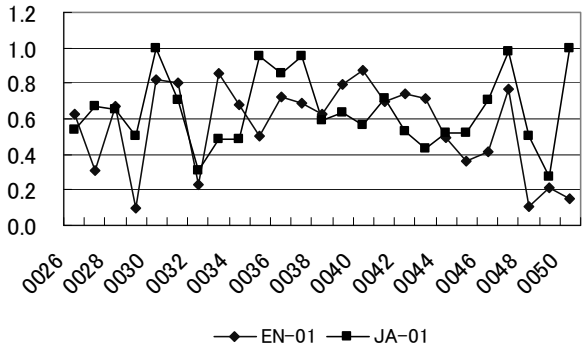

Figure 3. English vs. Japanese

5 CONCLUSIONS

We submitted five EN-EN and JA-JA runs for the NTCIR-9 GeoTime task.

Compared with the data which thinks instancy is important in newspaper data, the data from Wikipedia etc. which are entered into a database in the form where the past information was arranged tends to acquire the same information about potential suitable place/times.

We obtained place/time information about topics from Wikipedia and Google using query terms extracted from topics. Providing this additional information to query terms, we retrieved documents using <TEXT> tag index and scored them. Moreover, we compared <DATE> tag of searched documents with time information, weighted the score value of documents retrieved, and ranked them.

Although the subject of automation of extraction of place/time remains in general, the validity of the methods of we proposed was confirmed from the comparison of evaluation results with runs which do not use these place/time information.

6 REFERENCES

[1] Santos, D. and Cabral, L., GikiCLEF: Crosscultural issues in an international setting: asking non-English-centered questions to Wikipedia, In *CLEF 2009 Working Notes*, http://www.clef-campaign.org/2009/working_notes/Santos-paperCLEF2009.pdf. September (2009), 21pp.

[2] Balog, K., and Serdyukov, P., Overview of the TREC 2010 Entity Track, In *Proceedings of the 19th Text REtrieval Conference* (TREC 2010), http://trec.nist.gov/pubs/trec19/papers/bauhaus.univ.SESSION.rev.pdf (2011).

[3] Stanford Named Entity Recognizer (NER), http://nlp.stanford.edu/software/CRF-NER.shtml

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

[5] Robertson, S.E. and Walker, S., Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval, In *Proceedings of 17th International Conference Research and Development in Information Retrieval*, pp. 232-241 (1994).

[6] Gey, F., Larson R., Machado, J., and, Yoshioka, M., NTCIR9-GeoTime Overview – Evaluating Geographic and Temporal Search: Round 2, In *Proceedings of NTCIR Workshop 9*, Tokyo, JAPAN, December (2011).