NTCIR-9 GeoTime at Osaka Kyoiku University

— Toward Automatic Extraction of Place/Time Terms —

Takashi SATO
sato@cc.osaka-kyoiku.ac.jp
(Information Processing Center, Osaka Kyoiku University)

[1] Outline

- Obtained place/time information about topics from the Internet using query terms extracted from topics.
- Retrieved documents using <TEXT> tag index and scored them.
- 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.

[2] Indexing

- TEXT indices were made from <TEXT> tag part of the corpus.

Statistics of TEXT Indices

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

•<DOCNO> and <DATE> part are extracted also.

[3] Retrieval

- We made the following four different searches.

  [a] △Keyword Search of TEXT tag

  (Including Morphological analysis, Word Filtering, Word Expansion)

  [b] ○Place Search of TEXT tag

  [c] ○Time Search of TEXT tag

  [d] ◎ Time Search of DATE tag

△：Baseline, ○：effective, ◎：more effective

- 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 (1) or (1)', we get place/time information. (automatic + manual)
  (2)' From the output of (2), we extracted time information only. (automatic)

- Retrieval and Scoring (automatic)

  (3), (3)* Retrieving <TEXT> tag index using place/time of (1) and (2), we scored documents retrieved. (difference of (3) and (3)* is handling of document length.)
  (3)' Retrieving <TEXT> tag index using place/time of (1)' and (2), we scored documents retrieved.
  (3)'' Retrieving <TEXT> tag index using only (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 multiplier of (4) which is a function of the difference between two date (<DATE> - (2)‘) is shown in table below.

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

Difference of Days versus Multiplier

Larger multiplier is set when the article date(<DATE>) is near and after (2)’ (date of topic)

[4] Submitted Runs

- Process of each run was the combination of the above procedures (1) to (5) as shown in table below.

Combination of procedures for submitted runs

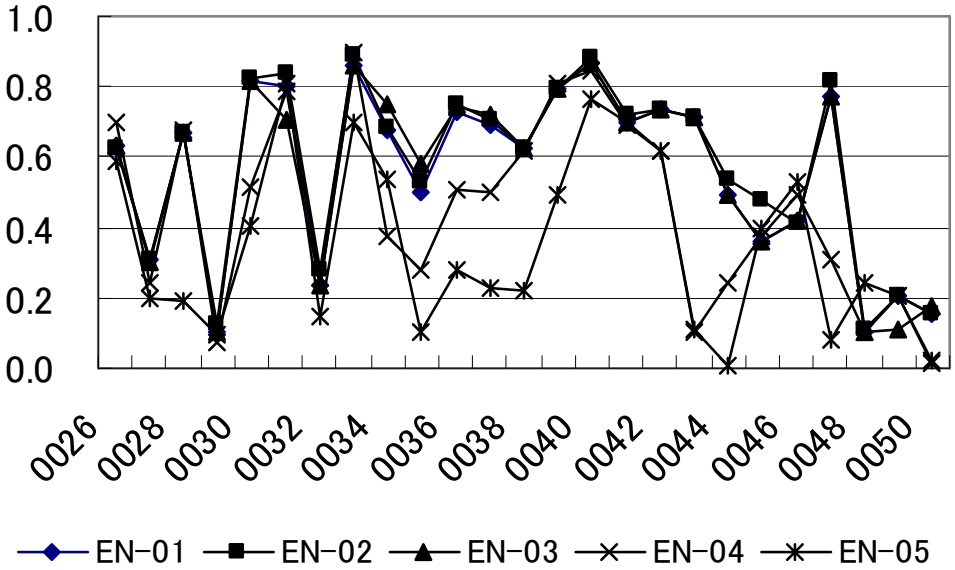
Run Name	Method
OKSAT-{EN-EN|JA-JA}-01-DN	(1) (2) (3) (4) (5)
OKSAT-{EN-EN|JA-JA}-02-DN	(1) (2) (3)* (4) (5)
OKSAT-{EN-EN|JA-JA}-03-D	(1)' (2) (3)' (4) (5)
OKSAT-{EN-EN|JA-JA}-04-DN	(1) (2) (3)
OKSAT-{EN-EN|JA-JA}-05-DN	(1) (3)"

Upper three in table were runs which used place/time terms extracted from Wikipedia and Google and performed <DATE> tag search. On the other hand, lower two were runs which did not perform <DATE> tag search. The last line is the baseline model.

[5] Example: GeoTime-0026

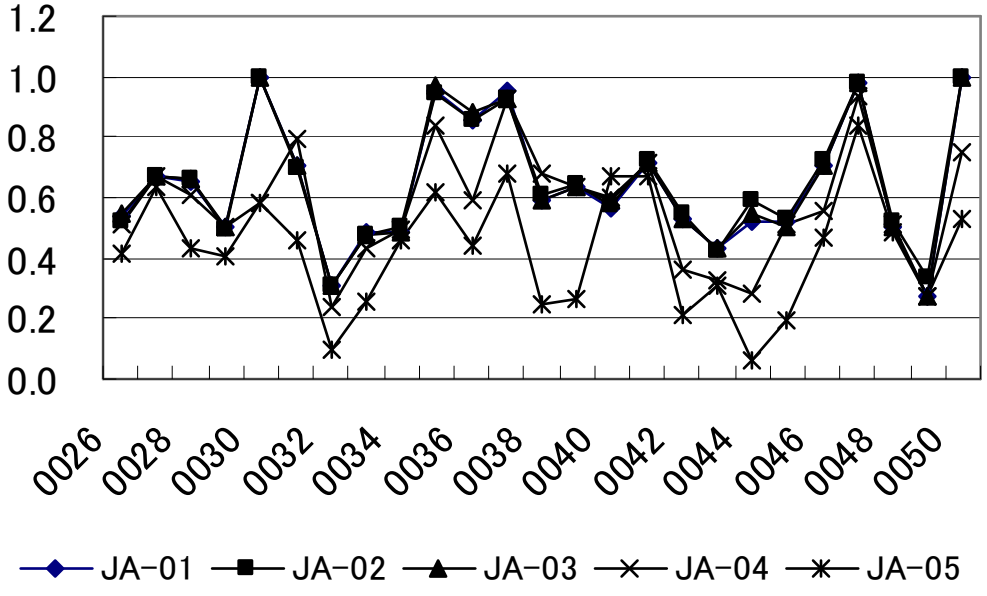
- <![CDATA[Where and when did the space shuttle Columbia disaster take place?]]>
- (1) Query terms “space shuttle Columbia disaster” etc. were extracted.
- (2) We retrieved Wikipedia using (1), and got the page describing the accident.
  - Title: Space Shuttle Columbia disaster
  - URL:http://en.wikipedia.org/wiki/Space_Shuttle_Columbia_disaster
  - 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
- 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>.

[6] Results [Run by Run]



Topic ID vs. AP of Submitted English Runs

The mean average precision (MAP) of EN-01, EN-02 and EN-03 was good to the same extent.
The MAP of EN-04 and EN-05 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 AP of EN-04 better than that of EN-05, we think that it was effective to have pinpointed the place/time using place/time information from the internet.



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 EN-02, EN-03, and EN-01 at the descending order of MAP.
Whereas that of Japanese runs was JA-02, JA-01, and JA-03.

{EN|JA}-{01|02|03|04|05} stands for OKSAT-{EN-EN|JA-JA}-{01|02|03|04|05}-{DN|D}
e.g. EN-01 stands for OKSAT-EN-EN-01-DN.

[7] 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. On the other hand, 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.

[8] Successes

- <DATE> tag search was very good.
  - RUN: EN-{01,02,03} ⇔ EN-{04,05}
    MAP: {0.53,0.55,0.53} ⇔ {0.43, 0.33}
  - RUN: JA-{01,02,03} ⇔ JA-{04,05}
    MAP: {0.64,0,65,0.64} ⇔ {0.57,0.42}
- <TEXT> tag search with place/time information was good.
  - RUN: EN-04 ⇔ EN-05
    MAP: 0.43 ⇔ 0.33
  - RUN: JA-04 ⇔ JA-05
    MAP: 0.57 ⇔ 0.42

[9] Adjustment for Date Search

- 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 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.
- 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 the time difference.

[10] Post-submission Experiments

- Succeeded in automation of place/time extraction for 11 topics (44% of task topics) of JA. Programs are written in Perl.
- Procedure

(1) Search Google using words extracted from topic and get top ten pages.

(2) Counting the number of times of appearance of each word in these pages, and sort words in descending order of its number.

(3) About place information, referring our place name database, get first match of (2).

(4) About time information, using regular expression for date, get first match of (2).

[11] 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. 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.
- <DATE> tag search works well.
- 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 .