ABRIR at NTCIR-9 at GeoTime Task Usage of Wikipedia and GeoNames for Handling Named Entity Information

Masaharu Yoshioka
Graduate School of Information Science and Technology, Hokkaido University
N14 W9, Kita-ku, Sapporo-shi
Hokkaido Japan
yoshioka@ist.hokudai.ac.jp

ABSTRACT
In the previous NTCIR8-GeoTime task, ABRIR (Appropriate Boolean query Reformulation for Information Retrieval) proved to be one of the most effective systems for retrieving documents with Geographic and Temporal constraints. However, failure analysis showed that the identification of named entities and relationships between these entities and the query is important in improving the quality of the system. In this paper, we propose to use Wikipedia and GeoNames as resources for extracting knowledge about named entities. We also modify our system to use such information.

Categories and Subject Descriptors
H3.3 [**Information Systems**]: Information Search and Retrieval

General Terms
Information Retrieval

Keywords
Named Entity, Wikipedia, GeoNames, Query formation, Question and Answering

1. INTRODUCTION
The focus of the NTCIR-GeoTime task is on search with Geographic and Temporal constraints[2]. At the last NTCIR-GeoTime task, we proposed a method to construct Boolean queries that focuses on named entities and variation of verbs by using ABRIR (Appropriate Boolean query Reformulation for Information Retrieval)[7].

ABRIR (Appropriate Boolean query Reformulation for Information Retrieval) was one of the most effective systems for retrieving to search in NTCIR8-GeoTime task. However, there were several topics where ABRIR performed worse than baseline system, due to problem which arose in handling named entities.

Therefore, in this paper, we propose to use Wikipedia [1]and GeoNames [2]as resources to extract named entity information and ABRIR is modified to utilize such information.

[1]http://en.wikipedia.org/
[2]http://www.geonames.org/

2. ABRIR(APPROPRIATE BOOLEAN QUERY REFORMULATION FOR INFORMATION RETRIEVAL)
ABRIR is an IR system which has following features for the combination of probabilistic and Boolean IR model.

1. Reformulation of a Boolean query
   The system compares an initial Boolean query and pseudo-relevant documents and modifies the query to increase the number of documents that satisfies the query.

2. Calculate score based on the results of probabilistic and IR model
   Basic documents scores are calculated by using probabilistic IR model. A penalty is applied for score of documents that do not satisfies the given Boolean query.

2.1 Reformulation of a Boolean query
In ABRIR, the Boolean query is constructed based on comparing the initial query and terms in pseudo relevant documents. Since we assume verbs and named entities are important in finding relevant documents, we use an appropriate list of synonyms and variations of Japanese katakana description for named entities.

For the verbs, the EDR electronic dictionary, developed by Japan Electronic Dictionary Research Institute, Ltd. [3] is used for finding synonyms. In this dictionary, each verb has one or more semantic id(s). All verbs that share one or more semantic id(s) with the original verb are candidate synonyms.

For the named entities written in Japanese katakana, the following rules are used for generating variations of the description.

1. Remove "ー" from the original term
2. Remove small katakana (e.g., "アイウエオヤユヨワカケツ") from the original term
3. Replace small katakana (e.g., "アイウエオヤユヨワカケツ") to large katakana (e.g., "アイウエオヤユヨワカケツ")

By applying this generation rule to the term “ヘップバーン” (Hepburn), three candidates (“ヘップバン”, “ヘプバーン” “ヘツプバーン”) are generated.

Figure 1 shows procedures for constructing query and retrieval in ABRIR.

1. Remove question part of the query
   Question part of the query (e.g., “のはいつですか？” (when)) is trimmed from the original query.

2. Morphological analysis and NE tagging
   Almost same index terms extraction system is used for extract initial terms. There are two difference in this extraction process.

   - Extraction of verb
   - Identification of named entities
     CaboCha[4] is used for identify named entities.

3. Generation of synonym and variation list
   The system generates the synonym list for verbs and variation list for named entity.

4. Initial retrieval
   Probabilistic IR model is used to obtain pseudo relevant documents. We use top 3 ranked documents as pseudo relevant ones.

5. Construction of Initial Boolean query
   There are three types of terms in query; NE, verb, and other. The system compares query terms and pseudo relevant documents in following manner.

   - Named entity

     Since the system generates variation list of given NE automatically, most of the terms are meaningless. Therefore the system compares the variational description list and terms in the documents and remove terms that do not exist in the documents. For example, when there are two documents that contain “ヘップバーン” and one document that contains “ヘプバーン”, the system constructs OR description (“ヘップバーン” or “ヘプバーン”) for “ヘップバーン”.

   - Verb

     When all pseudo relevant documents contain one or more synonyms of the verb, these documents are sufficient for generating the synonym list for the final Boolean query. In this case, synonyms that exist in the documents are used for Boolean query. For example, when two documents contain “亡くなる” (die) and one document contains “死ぬ” (die), AND elements are modified as (“亡くなる” or “死ぬ”).

     When there is one or more document(s) that do not contain any synonyms, the system generates the new query by replacing the verb with its synonym list and conducts secondary retrieval. By using new top three pseudo relevant documents, the system selects synonyms that exist in the documents are used for Boolean query.

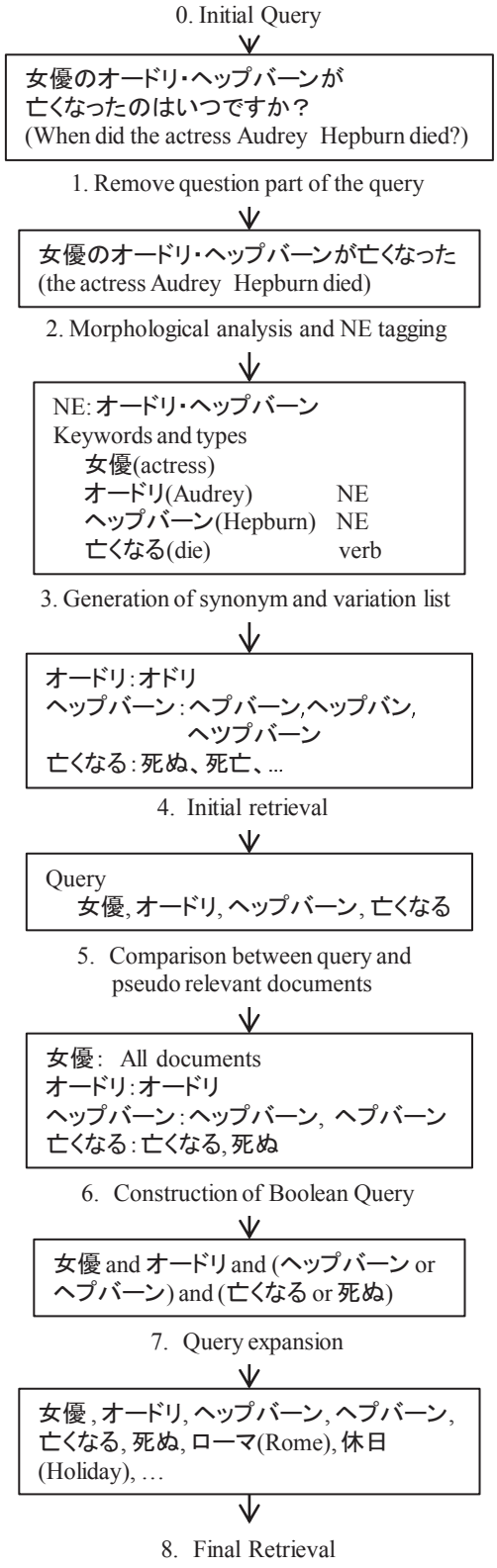

Figure 1: Procedures for Constructing Query and Retrieval in ABRIR.

- Other terms in initial query
  When other terms in the initial query exist in all pseudo relevant documents, These terms are used as AND elements of the final query.

6. Construction of Boolean query
   A set of synonyms, named entity variation lists, and terms in all pseudo relevant documents are joined by the AND operator for construction of the final Boolean query.

2.2 Modification of the Score Based on the Boolean Query

Probabilistic IR model of ABRIR is almost equivalent to Okapi BM25 with pseudo-relevance feedback and query expansion and implemented by using the Generic Engine for Transposable Association (GETA) tool [3].

The probabilistic IR model for ABRIR used the BM25 weighting formula to calculate the score of each document:

$$\sum_{T \in Q} w^{(1)} \frac{(k_1+1)tf}{K+tf} \frac{(k_3+1)qtf}{k_3+qtf} \quad (1)$$

$w^{(1)}$ is the weight of a (phrasal) term T, which is a term or a phrasal term in query Q, and is calculated using Robertson-Sparck Jones weights:

$$w^{(1)} = log \frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/(N-n-R+r+0.5)} \quad (2)$$

where N is the count of all documents in the database, n is the count of all documents containing T, R is the given number of relevant documents, and r is the count of all relevant documents containing T. In addition, tf and qtf are the number of occurrences of T in a document and in a query, respectively, and k_1, k_3 and K are control parameters.

For handling phrasal terms, we introduced a parameter $c (0 \leq c \leq 1)$ that is used for counting the phrasal terms in a query, where qtf is incremented by c rather than one when a phrasal term is found.

For the query expansion, we used Rocchio-type feedback [6]:

$$qtf = \alpha qtf_0 + (1-\alpha) \frac{\sum_{i=1}^{R} qtf_i}{R} \quad (3)$$

where qtf_0 and qtf_i are the number of times T appears in the query and in relevant document i, respectively.

ABRIR at NTCIR-8 used the five top-ranked documents for pseudo-relevance feedback and selected the 5 different terms with the highest mutual information content between a relevant document set and a term.

Because we assume that documents that do not satisfy the Boolean query may be less appropriate than documents that do satisfy the query, we subtract a penalty score from documents that do not satisfy the Boolean query.

[3] http://geta.ex.nii.ac.jp/

We apply the penalty based on the importance of the word. For a probabilistic IR model, we used the BM25 weighting formula to calculate the score of each document (Equation 1). In this equation, $w^{(1)} \frac{(k_3+1)qtf}{k_3+qtf}$ shows the importance of the word in the query. We use a control parameter β to calculate the penalty score.

$$Penalty(T) = \beta * w^{(1)} \frac{(k_3+1)qtf}{k_3+qtf} \quad (4)$$

For the OR operator, we use the highest penalty from all the OR terms as the overall penalty. In addition, since we assume named entities are more important, we set $Penalty(T) = 1000000$ for them.

3. ABRIR AT NTCIR-9

3.1 Usage of Wikipedia and GeoNames for Handling Named Entity Information

In most of the queries, ABRIR works more effectively than baseline system, but there are some difficult topics. One of the difficult topic types is one that deals with the relationship between location names. For example, topic 14 includes named entity term “アフリカ” (Africa). However, the relevant documents has name of the African country “コンゴ民主共和国” (Democratic Republic of the Congo) instead of “アフリカ”. In order to deal with the relationship between these two geographic entities, we need to have knowledge about geographic entities.

Another issue is related to the quality of named entity extraction and identification of a named entity. For example, previous system could not identify “アメリカ”(America in Katakana) and “米国”(America in Kanji). It is better to have a good named entity extraction system which supports identification of named entity in different variation of description.

Therefore, in this round of GeoTime, we decided to use Wikipedia and GeoNames as resources to extract named entity information.

GeoNames is a geographical database that contains over 7 millions of geographical location with name, country code, administration area code, type of location. This is a good resource to find out the part whole relationship between geographical names. For example, by using the information for “Los Angeles”, we can understand it is a part of “California State” and “United States of America.” However, because most of the entries in GeoNames do not have Japanese description(19,966 out of 7,660,238), it is difficult to use it for Japanese retrieval task.

For this problem, we have already proposed a method to use Wikipedia to obtain appropriate Japanese translations [5]. In this method, at first, we found correspondence between Wikipedia page and entry in GeoNames by using name matching and Wikipedia category information to identify the country. Based on this corresponding information and a language link between English Wikipedia and Japanese Wikipedia, we can add 41,580 that covers most of the ge-

ographical name for first level administrative division (e.g, State, Province,...).

For the named entity extraction, we utilized Wikipedia and DBPedia [4]. DBPedia is a community effort to extract structural information from Wikipedia and it constructs the DBPedia ontology that classify the type of Wikipedia page into several categories. However, since this DBPedia ontology is not constructed for Japanese Wikipedia, we construct initial page classification by using language links. After making language links, we expand the entry by using the similarity of using template and categories in Japanese Wikipedia. In our experiments, we used named entity related categories (e.g., Person, Organization, Place, and Infrastructure) to make a list for the system. In this system, we also use redirect links for

identifying the same entity in different description. For example, since "アメリカ軍" (American army) has a redirect link from "米軍" (American army in Kanji), we can normalize the named entity information of "米軍" to "アメリカ軍."

All of the named entity information is encoded as dictionaries for MeCab [5]. By using this dictionary, we can extract named entity information from articles and queries.

3.2 Modification of ABRIR to use Named Entity Information

ABRIR at NTCIR-8 only used named entity information during the query processing phase. By using this system architecture, it is not easy to use normalized information about named entity. Therefore, we also made another named entity databases for each article. In this database, each article has information about named entity, country and time that are described in the article. The following describes methods to extract such information from one article.

1. Extract named entity information by using Wikipedia information
   MeCab and dictionary with Wikipedia entry are used for extracting named entity information. Extracted results is normalized by using redirect link information.

2. Extract location name by using GeoNames
   MeCab and dictionary with GeoNames entry are used for extracting geographical location name. From this information, we also extract candidate name of the country associated with GeoNames entry.

3. Extract time information
   CaboCha is used for identify time related information. Extracted information is normalized by using the date that the article was published. For example, yesterday for the article published at "2, May, 2002" means "1, May, 2002." Finally, we stored information about time for the information about year, month and day. In this instance, we stored information "Year 2002", "Month 5" and "Day 1".

[4]http://www.dbpedia.org/
[5]http://mecab.sourceforge.net/

This named entity database is used for calculating the penalty. ABRIR at NTCIR-8 applied the penalty for a named entity. However, since comparison between named entity in each article (document) and the query is done by using different index schema, it is difficult to calculate penalty in manner of NTCIR-8 GeoTime.

In this research, we decide to use penalty calculation for every term

by using the original penalty function (equation 4) only. After calculating the score by using BM25 (equation 1) and penalty function, we apply the penalty based on the comparison between the named entity information in the query and the named entity database.

Followings are method to calculate penalty for each categories of data in the database

- Location
  Penalty of location is calculated by country level. When the query has location name "日本 (Japan)," then we will check the existence of country name for each article. It means we don't care the existence of term "日本 (Japan)" in the documents. When the article has location name "東京 (Tokyo)" that is a town in Japan, this documents satisfies the condition of location Boolean query. In addition, we also construct database for representing list of countries (e.g., Asia, Europe, Middle East, and so on). This list is constructed based on the information in the Wikipedia and checked as or set for the Boolean comparison. For example, since "Asia" contains "Japan," all documents that has location information in Japan satisfies the condition of Boolean query about "Asia."

- Named entity and time
  We check the existence of the entry by comparing the information in the named entity database and one identified in the query.

In this experiment, since time information is not so reliable as the other information, we set the penalty for location and named entity to be 1000 and penalty for time to 300.

3.3 Parameters for ABRIR

After NTCIR-8, we conducted experimental analysis on the parameter settings for ABRIR [8]. In this analysis, we found following issues.

1. Expansion of verbs may produce deteriorated results.
   When the verb has crucial role in question, verb expansion may improve the results. However there are many cases that verb expansion produces inferior results.

2. Large number of query expansion terms improve results.
   For the query that have many relevant documents, query expansion works well to find out varieties of documents.

3. Quality of the relevant documents are important.
   When there is no or small numbers of documents that related to the given query, the retrieval results tends to drift away from the initial query.

Based on this discussion, we decided to modify the number of pseudo relevant documents and query expansion terms.

In addition to that, we compared two different runs: using verb expansion and not.

In addition, in order to improve the quality of relevant documents, we also try to include documents based on the initial Boolean query for improvement of the quality of the relevant documents.

3.4 Retrieval Procedure

Based on the previous discussion, our retrieval procedure is modified as Follows:

1. Remove the question part of the query
   This procedure is same as previous one.

2. Dealing with geographical coordinates
   When the query has a description about geographical coordinates, the system searches the nearest first-order administrative division (such as a state in the United States) by using GeoNames database. Name of the country for this division is replaced from coordinates description.

3. Morphological analysis and NE tagging
   We extract verbs. Named entity is extracted by using same method for indexing the articles.

4. Generation of synonym and variation list
   The system generates a synonym list for verbs. In this system variation list is constructed for all Katakana terms.

5. Initial retrieval
   There are two strategies to select pseudo relevance documents. One is same as the NTCIR-8 procedure – use probabilistic IR model to finding out five pseudo relevant documents (*Prob*). The other is to use the probabilistic IR model with named entity penalty calculation (*UsePenalty*). The system then selects two pseudo relevant documents from them. Then the system adds three additional documents by using probabilistic IR model only. When there are overwrap(s) between top three documents and the documents selected using penalty calculation, those documents are removed from the list and top three documents are selected for pseudo relevant documents.

6. Construction of Initial Boolean query
   The system compares query terms and pseudo relevant documents in following manner.

   - Katakana terms The system generates variation list of given katakana terms automatically for generating candidate or list for the given katakana term.
   - Verb When the system uses verb expansion, the system uses same procedurs in NTCIR-8. We use Japanese WordNet[1] instead of EDR.
   - Other Terms in initial query
     When other terms in initial query exist in all pseudo relevant documents, these terms are used as AND elements of the final query.

7. Construction of Boolean query
   A set of synonyms, named entity variation lists, and terms in all pseudo relevant documents are joined by AND operator for constructing the final Boolean query.

8. Query expansion by using pseudo relevant documents
   The system selects 300 different terms with the highest mutual information content between a relevant document set and a term. The system also add keywords in Boolean query as expansion terms.

9. Final retrieval
   Based on the final query, final retrieval is conducted by using probabilistic IR model. We apply the penalty based on the importance of the word by using equation 4 and comparison between the named entity information in the query and the named entity database.

4. EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Experimental Set Up

Following are the parameters for the submitted results. Most of the parameters are same as NTCIR-8. We use $k_1 = 1, k_3 = 7, K = \frac{dl}{avdl}, c = 0.3, \alpha = 0.7$ for probabilistic IR model. Here, dl is the length of a document (the number of terms and phrasal terms) and $avdl$ is the average length of all documents.

We also use $\beta = 3$ for penalty calculation. By using this formalization, there are many documents with minus scores. Therefore we just recalculate the score values that retains the order of all document scores.

Following is a description of the submitted runs.

HU-KB-JA-JA-01-D Boolean operators are used for penalty calculation. Verb synonym list is used for Boolean query construction. *UsePenalty* strategy is used for selecting pseudo relevant documents. Named entity information is also used for penalty calculation.

HU-KB-JA-JA-03-D This run does not use verb synonym list. Other settings are same as HU-KB-JA-JA-01-D.

HU-KB-JA-JA-04-D This run does not use penalty calculation for named entity. Other settings are same as HU-KB-JA-JA-01-D.

HU-KB-JA-JA-05-D This run uses *Prob* strategy for selecting pseudo relevant documents. Other settings are same as HU-KB-JA-JA-01-D.

4.2 Discussion about Experimental Results

Table 1 shows the evaluation measure for each submitted run.

Table 1: Evaluation measure for each submitted run

	01-D	03-D	04-D	05-D
AP	0.4385	0.4490	0.4108	0.4363
nDCG	0.6298	0.6630	0.6085	0.6273
Q	0.4666	0.4804	0.4458	0.4648

4.2.1 Overall failure analysis

First, we would like to discuss the overall performance related to the topics. Since most of the settings are similar in these runs, we don't have so much difference about performance in general.

For the following topics, our system had difficulties (AP<0.2) in finding relevant documents and we analyse of the reason for this problem.

Topic28:Arrest of Washington sniper The system finds out relevant documents about sniper, but it fails to select documents for arrest.

Topic 30:Steve Fossett landing of aircraft The system finds out relevant documents about Steve Fossett's landing of the balloon. It is difficult to distinguish balloon case and aircraft case by using the description.

Topic 32: Cable car crush Since the Japanese description about cable car "ロープウェー" is different from the description in the newspaper"ロープウエー". Due to this problem, system can not find good pseudo relevant documents.

Topic 33: Murder by arsenic poisoning Since the Japanese description about arsenic "砒素" is different from the description in the newspaper"ヒ素". Due to this problem, system can not find good pseudo relevant documents.

Topic 37: Accident near geological coordinates The system finds out relevant documents for the accident in Nigeria based on the GeoNames information. However, since our system does not calculate the distance, the system can not distinguish the difference among relevant one and others.

Topic 43: New England Patriots last win The system finds out relevant documents about New England Patriots win at Super Bowl. However system can not sort out the results based on the time order.

Topic 45:European Central Bank The system can not recognize European Central Bank as named entity and the Japanese description about ECB "ヨーロッパ中央銀行" is different from the description in the newspaper" 欧州中央銀行". Due to this problem, system cannot find good pseudo relevant documents.

4.2.2 Verb expansion

From the comparison between 01-D and 03-D, we can discuss the effectiveness of verb expansion.

03-D is better than 01-D for 11 topics (29, 32, 36, 37, 43, 44, 45, 46, 47, 48, 49) and is worse for 6 topics (30, 33, 34, 35, 39, 42).

A typical example for a good case and a bad case is as follows.

Good case(AP:0.2072→0.3524) Topic42:Death of king: Most important verb is "亡くなる"(death). This is the case that we are expected.

Bad case(AP:0.8386→0.5765) :Topic48:International Criminal Court: Most important verbs are "投票"(vote) and "発効"(adopt). Those verbs are special terms for this case. It is no need to expand these verbs. In addition, expansion for other verbs may deteriorate the results.

4.2.3 Penalty for named entity

From the comparison between 01-D and 04-D, we can discuss the effectiveness of applying the penalty for a named entity.

04-D is better than 01-D for 1 topics (50) and is worse for 13 topics (26,32,34,35,36,37,38,40,42,44,45,47,49)

Good case(AP:0.0402→0.4547) Topic44:South American Earthquake: By using penalty for location, the system can effectively select relevant articles that includes geographical name of South America.

Bad case(AP:0.7→0.5123) :Topic50:CAFTA sign: The named entity recognition system made a mistake in extracting country name. The system extracts "中" (China) and "米" (USA) from "中米" (Central America). Due to this problem, the system fails to make rank.

4.2.4 Strategy for selecting pseudo relevant documents

From the comparison between 01-D and 05-D, we can discuss the effectiveness of our strategy for selecting pseudo relevant documents

05-D is better than 01-D for 1 topics (50) and is worse for 1 topic (44).

Good case(AP:0.0402→0.4547) Topic44:South American Earthquake: By using penalty for location, the system can select relevant articles for pseudo relevant documents

Bad case(AP:0.82→0.8228) Topic50:CAFTA sign: The named entity recognition system made a mistake for extracting country name. The system extracts "中" (China) and "米" (USA) from "中米" (Central America). Due to this problem, the system fails to find out good relevant documents and it may deteriorate the results.

4.3 Discussion

Based on the above analysis of the experimental results, the following issues remain to be addressed in improving the quality of the system.

1. Strategy for selecting pseudo relevant documents
   Based on the analysis of overall failure analysis, quality

of pseudo relevant documents is crucial. However, our proposed method is not adequate at this moment.

2. Query analysis
   When we try to use verb expansion, it is necessary to select important verbs.

3. Improvement for our named entity recognition system
   It is necessary to improve the quality of the entity recognition system. For example, there is a Wikipedia page for "European Central Bank," but we failed to select the appropriate named entity class for this page.

5. CONCLUSION

In this paper, we propose to use a named entity database constructed by using Wikipedia and GeoNames for ABRIR. We confirm that when we can find good pseudo relevant documents, the system can achieves higher performance. We also make a list of issues to solve for improving the quality of the system.

Acknowledgement

This research was partially supported by a Grant-in-Aid for Scientific Research (B) 21300029, from the Japan Society for the Promotion of Science.

6. REFERENCES

[1] F. Bond, H. Isahara, S. Fujita, K. Uchimoto, T. Kuribayashi, and K. Kanzaki. Enhancing the japanese wordnet. In *Proceedings of the 7th Workshop on Asian Language Resources*, ALR7, pages 1–8, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[2] F. Gey, R. Larson, N. Kando, J. Machado-Fisher, and M. Yoshioka. NTCIR9-GeoTime overview - evaluating geographic and temporal search: Round 2. In *Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Quesiton Answering, And Cross-Lingual Information Access*, 2011. (to appear).

[3] Japan Electronic Dictionary Research Institute, Ltd. (EDR). *EDR ELECTRONIC DICTIONARY VERSION 2.0 TECHNICAL GUIDE TR2-007*, 1998.

[4] T. Kudo and Y. Matsumoto. Japanese dependency analysis using cascaded chunking. In *CoNLL 2002: Proceedings of the 6th Conference on Natural Language Learning 2002 (COLING 2002 Post-Conference Workshops)*, pages 63–69, 2002.

[5] H. Takenaka and M. Yoshioka. Extraction of geo-spatial relationships among geographical name by using wikipedia. In *The 25th Annual Conference of the Japanese Society for Artificial Intelligence*, 2011. CD-ROM 2J3-NFC2-2.

[6] M. Uchiyama and H. Isahara. Implementation of an IR package. In *IPSJ SIGNotes, 2001-FI-63*, pages 57–64, 2001. (in Japanese).

[7] M. Yoshioka. On a combination of probabilistic and boolean ir models for geotime task. In *Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Quesiton Answering, And Cross-Lingual Information Access*, pages 154–158, 2010.

[8] M. Yoshioka. On a combination of probabilistic and boolean ir models for question answering. *IPSJ Journal*, 52(12), 2011. in Japanese (accepted).