ABRIR at NTCIR-8 GeoTime Task
Usage of Wikipedia and GeoNames for Handling Named Entity Information
Masaharu Yoshioka e-mail: yoshioka@ist.hokudai.ac.jp
Graduate School of Information Science and Technology, Hokkaido University

Motivation

ABRIR at NTCIR-8 GeoTime
Flow chart of Query Construction

1. Initial Query

2. Morphological analysis and NE tagging

3. Generation of synonym and variation list

4. Initial retrieval

5. Comparison between query and pseudo relevant documents

6. Construction of Boolean Query

7. Query expansion

8. Final Retrieval

ABRIR at NTCIR-8 GeoTime

Motivation

- Construction of Boolean query by using pseudo relevant documents
  - Named entity is more important than others
  - Verb synonym list is used for increasing coverage
  - Combination of Boolean IR model and probabilistic IR model (Okapi)
  - Penalty is applied for documents that don’t satisfy Boolean query
  - Named entity is more important than others

- Failure analysis at NTCIR-8
  - Identification of named entities and relationships between these entities and the query is important
  - Quality of pseudo relevant documents
  - Topic drift
  - Verb synonym expansion

Some verbs are not important for Boolean query construction.

Approach

Construction of Named Entity database

- Usage of Linked Open Data
- GeoNames: Large-scale database of geographic entities
- Wikipedia: Information resource for named entity

Database construction

- Geographic entities from GeoNames and Wikipedia
  - Add Japanese translation of GeoNames entry by using Wikipedia information
    - [Takenaka, et al., 2011]
  - Other named entities from Wikipedia
  - Usage of DBpedia category information to identify type of named entity (Person, Organization, Place, and Infrastructure) for making a list.
  - Redirect information is used for normalize named entity description
    - 米軍 (American army) ⇒ ミリタリーアメリカ軍 (American army)

Modification of ABRIR to use Named Entity Information

- Extra indexing for NE and time information
- Extract named entity information by using database
- MeCab is used for extraction
- Extraction of time information by using CaboCha
- Normalization by using the date that the article was published. For example, yesterday for the article published at "2 May, 2002" means "1, May, 2002."
- Calculation of penalty by using Named Entity information
  - Location: compare the country code between query and documents. Country code is generated by the name of the city, country and area such as Middle East.
  - Time: Time has penalty = 300 (other penalty = 1000)

- Parameters for ABRIR
  - Based on the experiment after NTCIR-8 GeoTime, we modify following parameters.
    - Pseudo relevant documents 5
    - Query expansion term 300
    - Dictionary for verb synonym list
    - Japanese WordNet
  - Strategy for pseudo relevant document selection
    - Probabilistic: Use top ranked documents retrieved by probabilistic IR model
    - Use Penalty: System also apply penalty calculation for initial retrieval. Top two documents are selected from the list and others from probabilistic IR model

Parameters for each run and retrieval results

<table>
<thead>
<tr>
<th>Runs</th>
<th>Boolean for NE</th>
<th>Verb Expansion</th>
<th>PR document Selection</th>
<th>AP</th>
<th>nDCG</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>JA-JA-01-D</td>
<td>Yes</td>
<td>Yes</td>
<td>Probabilistic</td>
<td>0.4385</td>
<td>0.6298</td>
<td>0.4666</td>
</tr>
<tr>
<td>JA-JA-03-D</td>
<td>Yes</td>
<td>No</td>
<td>Probabilistic</td>
<td>0.4490</td>
<td>0.6630</td>
<td>0.4804</td>
</tr>
<tr>
<td>JA-JA-04-D</td>
<td>No</td>
<td>Yes</td>
<td>Probabilistic</td>
<td>0.4108</td>
<td>0.6085</td>
<td>0.4458</td>
</tr>
<tr>
<td>JA-JA-05-D</td>
<td>Yes</td>
<td>Yes</td>
<td>Use Penalty</td>
<td>0.4363</td>
<td>0.6273</td>
<td>0.4648</td>
</tr>
</tbody>
</table>

Failure Analysis

- Type of errors in general
- Find out related documents, but it fails to select the relevant documents
  - Topic-28:Washington sniper
  - Topics: 29,30:Steve Fossett landing
- Problems to find out appropriate pseudo relevant documents
  - Topic-32:Cable car crush
  - Problems: Use top ranked documents retrieved by probabilistic IR model
- Additional function is necessary
  - Topic-37:Accident near geographic coordinates
- Type of errors
  - Boolean for NE
  - Error of NE extraction system affects the final result
  - Verb expansion
  - It is not appropriate to expand technical terms such as "adopt"
  - Pseudo relevant documents
  - Penalty approach shows the possibility to include appropriate documents, but error of NE affects the result

Conclusion

- Good pseudo relevant documents is required for higher performance.
- We need additional module to improve over all quality