ABRIR at NTCIR-9 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
  - ➢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**
    「アフリカ(Africa)」⇔「コンゴ民主共和国(Democratic Republic of the Congo)」
  - ✓**Quality of pseudo relevant documents**
    Topic drift
  - ✓**Verb synonym expansion**
    Some verbs are not important for Boolean query construction

ABRIR at NTCIR-8 GeoTime

Flow chart of Query Construction

0. Initial Query

女優のオードリ・ヘップバーンが亡くなったのはいつですか？
(When did the actress Audrey Hepburn died?)

1. Remove question part of the query

女優のオードリ・ヘップバーンが亡くなった
(the actress Audrey Hepburn died)

2. Morphological analysis and NE tagging

NE Tagger: Cabocha
Morphological Analysis: Mecab

NE: オードリ・ヘップバーン
Keywords and types
女優(actress)
オードリ(Audrey) NE
ヘップバーン(Hepburn) NE
亡くなる(die) verb

3. Generation of synonym and variation list

EDR for synonym list construction

オードリ:オドリ
ヘップバーン:ヘプバーン,ヘップバン, ヘツプバーン
亡くなる:死ぬ、死亡、…

4. Initial retrieval

Probabilistic IR model

Query
女優, オードリ, ヘップバーン, 亡くなる

5. Comparison between query and pseudo relevant documents

Usage of 3 pseudo relevant documents

女優: All documents
オードリ:オードリ
ヘップバーン:ヘップバーン, ヘプバーン
亡くなる:亡くなる, 死ぬ

6. Construction of Boolean Query

女優 and オードリ and (ヘップバーン or ヘプバーン) and (亡くなる or 死ぬ)

7. Query expansion

Usage of 5 terms with high MI from pseudo relevant documents

女優 , オードリ, ヘップバーン, ヘプバーン, 亡くなる, 死ぬ, ローマ(Rome), 休日(Holiday), …

8. Final Retrieval

Combination of Boolean and Probabilistic IR

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 identfy 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」⇒「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
  - ➢Penalty calculation by ABRIR + Peanlty calculation by using NE and time index
    - 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 = 10000)
- 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
  - ➢UsePenalty: System also apply penalty calculation for initial retrieval. Top two documents are selected from the list and others from probabilistic IR model

Experimental Results

Parameters for each run and retrieval results

Runs	Boolean for NE	Verb Expansion	PR document Selection	AP	nDCG	Q
JA-JA-01-D	Yes	Yes	Probabilistic	0.4385	0.6298	0.4666
JA-JA-03-D	Yes	No	Probabilistic	0.4490	0.6630	0.4804
JA-JA-04-D	No	Yes	Probabilistic	0.4108	0.6085	0.4458
JA-JA-05-D	Yes	Yes	UsePenalty	0.4363	0.6273	0.4648

Failure Analysis

- Type of errors in general
  - Find out related documents, but it fails to select the relevant documents
    (Topic-28:Washintong sniper,30:Steve Fosett landing)
  - Problems to find out appropriate pseudo relevant documents
    (Topic-32:Cable car crush,33:Murder by arsenic,45:European Central Bank)
  - Additional function is necessary
    (Topic-37:Accident near geographic coordinates,43:New England Patriots last win)
- 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