On a Combination of Probabilistic and Boolean IR Models for GeoTime Task

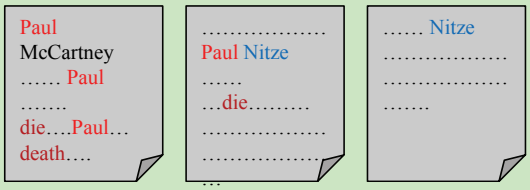
Masaharu Yoshioka e-mail: yoshioka@ist.hokudai.ac.jp
Graduate School of Information Science and Technology, Hokkaido University

Motivation

Information retrieval for Question Answering about a particular named entity

Documents that do not contain information about the entity are irrelevant.

Query: When Paul Nitze die?



Score of IR System With Partial match	○	○	×
IR for QA	×	○	×

ABRIR (Combination of Boolean IR model and Probabilistic IR model)

IR System (Probabilistic IR Model)

- Modified version of OKAPI
  - Use BM25 formula to calculate each document score

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

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

  - Term weighting for phrasal terms
  - Document score may differ according to the dictionary entry
    情報処理→ Word 情報処理
    情報科学→ Word 情報, 科学 Phrase !c情報科学
  - Discount score for phrasal index

$$qtf = c * qtf_c$$

Combination of Two IR Models

- Two approach
  - Use a Boolean IR model first and calculate score of each retrieved document by using a probabilistic model
  - Use a probabilistic IR model first and apply penalty for documents that do not satisfy a Boolean query formula
    - Penalty is calculated by using term importance in BM25

$$\beta \times w^{(1)} \times \frac{(k_3+1)qtf}{k_3+qtf}$$

β: parameter

    - Penalty is calculated for each "and" element
    - For "or" formula, use penalty of a term that has highest one among them.

Experimental Results

Runs	Boolean for NE	Boolean for others	
JA-JA-01-D	Filter	Filter	
JA-JA-02-DN	Filter	Filter	
JA-JA-03-D	Penalty	Penalty	
JA-JA-04-D	Penalty	No	
JA-JA-05-D	No	No	Baseline:Okapi

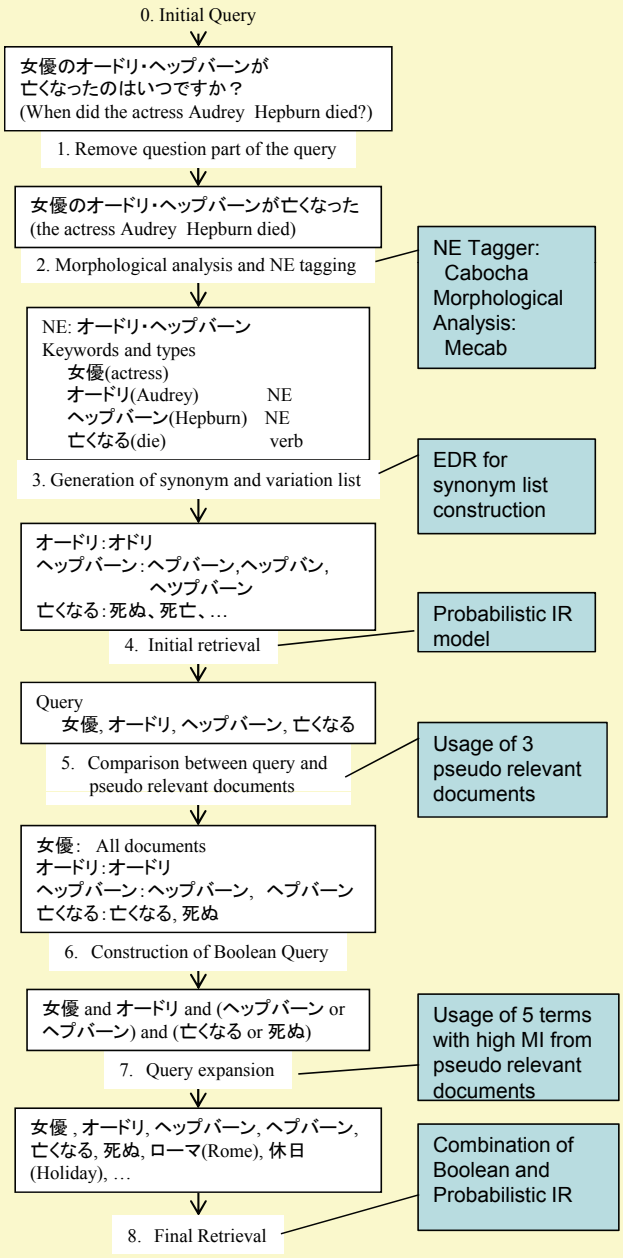
run	01-D	02-DN	03-D	04-D	05-D
AP	0.3697	0.3867	0.3719	0.3627	0.2881
nDCG	0.4117	0.4268	0.4162	0.4072	0.3282
Q	0.5710	0.5685	0.5881	0.5717	0.4993

Statistical Significance Test between 03-D and 05-D (Baseline: Okapi)
Statistically significant:
t-Test for a two-sided tests :nDCG(0.018) and Q(0.040)
Wilcoxon Signed Rank test: AP(0.0015), nDCG(0.0006) and Q(0.0024)

Modification of ABRIR for QA

- Usage of verb as index terms
  - Verb synonym set construction by using EDR
- Handling named entities
  - Identification of named entity is important for constructing a Boolean query.
  - Varieties of representation of foreign people in Japanese Katakana
- Number of relevant documents
  - There may be only a few relevant documents for a query.
- Number of query expansion terms
  - Large number of query expansion terms may cause concept drift for QA

Flow chart of Query Construction



Conclusion

- Proposal of using ABRIR as an IR system for question and answering for particular named entities.
- From the evaluation experiment, we confirm that ABRIR can make appropriate Boolean query and penalty based system outperform the baseline system (probabilistic IR model: Okapi BM25).