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

Masaharu YOSHIOKA
Hokkaido University
Motivation

- Information retrieval for Question Answering about a particular named entity
  - Documents that do not contain information about the named entity are irrelevant.

Query: When Paul Nitze die?

Score of IR System with Partial match

IR for QA
Proposed System

- Combination of Probabilistic and Boolean IR Models for QA
  - ABRIR (Appropriate Boolean query Reformulation for Information Retrieval)
    - Basic score is calculated by probabilistic IR model.
    - Documents that do not satisfy given Boolean query get penalty score.
  - Construction of appropriate Boolean Query for GeoTime QA
    - Verbs: synonym
    - Named entity: variation of description
ABRIR (a Probabilistic IR Model)

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

\[ w^{(1)} = \sum_{T \in Q} \frac{(k_1 + 1)tf_T (k_3 + 1)qtf_T}{K + tf_T k_3 + qtf_T} \]
\[ K = \frac{document \ length}{average \ document \ length} \]
\[ w^{(1)} = \log \frac{(r + 0.5)/(R - r + 0.5)}{(n - r + 0.5)/(N - n - R + r + 0.5)} \]

- 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 terms

\[ qtf = c \times qtf_c \quad qtf_c : frequency \ of \ phrase \ T \ in \ a \ query \]
\[ c : parameter \ (c \leq 1; c=0.3) \]
Relevance Feedback

- Relevance feedback
  - Pseudo-relevance feedback
  - Query expansion
    - Use terms in relevant documents as query terms
    - Rocchio-type feedback

\[ qt'f = \alpha \cdot qtf_0 + (1 - \alpha) \cdot \frac{\sum_{i=1}^{R} qtf_i}{R} \]

- \( qtf_0 \) : frequency of \( T \) in a initial query
- \( qtf_i \) : frequency of \( T \) in a \( i \)-th relevant documents
- \( R \) : the given number of relevant documents
- \( \alpha \) : parameter (\( \alpha = 0.7 \))
Problem on a Boolean IR model

- Retrieval performance of a Boolean IR model is worse than a probabilistic one
  - A Boolean query formula is expressive but is very difficult to construct appropriate one.

- Requirement for a Boolean query construction support
  - Use relevant documents for clarifying a Boolean query formula
    - Initial document retrieval without using a Boolean IR model
    - Relax a Boolean query formula by using relevant documents
Reconstruction of a Boolean Query Formula

- Relax an initial Boolean query formula to include given relevant documents as relevant one
  - Use terms that exist in all relevant documents and also exist in an initial query as a candidate to construct a relaxed Boolean query formula
  - Use an initial query for “or” formula

Initial query: \((A \text{ and } B \text{ and } (C \text{ or } D))\)

<table>
<thead>
<tr>
<th>Relevant documents</th>
<th>(A, C, E)</th>
<th>(A, C, D, E)</th>
<th>(A, B, C, E)</th>
</tr>
</thead>
</table>

Select candidate terms

- \(A\) and \(C\)
- Use initial query for “or” formula
- \(A\) and \((C \text{ or } D)\)
Combination of Probabilistic and Boolean 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:
      \[ \beta \times w^{(1)} \times \frac{(k_3 + 1)qtf}{k_3 + qtf} \]
      \( \beta \): parameter
    - Penalty is calculated for each “and” element.
    - For “or” formula, use penalty of a term that has highest one among them.
Modification of ABRIR for QA

- Usage of verb as index terms
  - QA query may include verbs that are necessary to find relevant documents.
  - Relevant documents may include synonyms instead of original verbs.

- Handling named entities
  - Identification of named entity is important for constructing a Boolean query.

- 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
Synonym Set Construction for Verbs

- Usage of Thesaurus (EDR) for synonym candidate generation.

- Usage of synonyms that exist in relevant documents are used for query expansion and a Boolean query formulation.
Identification of Named Entity

- CaboCha is used for named entity extraction for query analysis
- Characteristic of named entity description in Japanese
  - Named entity of foreign people and organization may represent by using Katakana with some variation.
  - Construction of variation candidates by using simple candidate generation rule.
  - Usage of variations that exist in relevant documents are used for query expansion and a Boolean query formulation
Query Analysis and Boolean Query Reformulation

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 keywords.
     - Extraction of verb
     - Identification of named entities

女優のオードリー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?)

NE:オードリーヘップバーン
(Audrey Hepburn)
Keywords and types
女優(actress)
オードリー(Audrey) NE
ヘップバーン(Hepburn) NE
亡くなる(die) verb
Query Analysis and Boolean Query Reformulation

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

4. Initial retrieval
   - Probabilistic IR model is used for finding out pseudo relevant documents. Top 3 ranked documents as pseudo relevant ones.

5. Comparison of pseudo relevant documents and query terms

亡くなる: 死ぬ、死亡、…
オードリー: オドリ
ヘップバーン: ヘプバーン、ヘップバン、…
女優: All documents
オードリー: オードリー
ヘップバーン: ヘプバーン、ヘップバン
亡くなる: 亡くなる、死ぬ
6. Construction of Boolean query

- There are three types of keywords in query; e.g., NE, verb, and other. The system compares query keywords and pseudo relevant documents in following manner.

  - **Named entity**
    - Candidates that exist in the pseudo relevant documents are connected with “or”.

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

Example:

- 女優: All documents
- オードリー: オードリー
- ヘップバーン: ヘップバーン、ヘップバーン
- 亡くなる: 亡くなる、死ぬ

⇒ 女優 and オードリー and (ヘップバーン or ヘップバーン)
Query Analysis and Boolean Query Reformulation

6. Construction of Boolean query
   - Verb
     • When all pseudo relevant documents contain one or more synonyms of the verb, these documents are sufficient enough for generating synonym list for final Boolean query.
     • Secondary retrieval (option)
       – When there is one or more document(s) that do not contain any synonyms, the system generates new query by replacing the verb with synonym list and conducts secondary retrieval.
Query Analysis and Boolean Query Reformulation

7. Query expansion by using pseudo relevant documents
   - The system selected the 5 different terms with the highest mutual information content between a relevant document set and a term. The system also added keywords in Boolean query as expansion terms.

8. Final retrieval
   - Final retrieval is conducted by the probabilistic IR model. The Boolean query is used for filtering out the document or calculating penalty for the document.
Experimental Result

- Parameters for the probabilistic IR model
  - $k_1 = 1, k_3 = 7, K = dl/avdl$, $c = 0.3$, $\alpha = 0.7$
  - $dl$ is the length of a document (the number of terms and phrasal terms) $avdl$ is the average length of all documents.

- Parameter for penalty calculation
  - $\beta = 3$
  - $\beta_n = 1000000$ (penalty for named entity)
Submitted Runs

<table>
<thead>
<tr>
<th>Runs</th>
<th>Type</th>
<th>Boolean for NE</th>
<th>Boolean for others</th>
</tr>
</thead>
<tbody>
<tr>
<td>HU-KB-JA-JA-01-D</td>
<td>Description</td>
<td>Filter</td>
<td>Filter</td>
</tr>
<tr>
<td>HU-KB-JA-JA-02-DN</td>
<td>Description + Narrative</td>
<td>Filter</td>
<td>Filter</td>
</tr>
<tr>
<td>HU-KB-JA-JA-03-D</td>
<td>Description</td>
<td>Penalty</td>
<td>Penalty</td>
</tr>
<tr>
<td>HU-KB-JA-JA-04-D</td>
<td>Description</td>
<td>Penalty</td>
<td>No</td>
</tr>
<tr>
<td>HU-KB-JA-JA-05-D</td>
<td>Description</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

HU-KB-JA-JA-03-D
ABRIR with Boolean penalty is the best system with Description
Evaluation Measure

- Average over all topics
  - 03-D is better than 01-D
    - Boolean penalty is better than Boolean filter
  - All runs with Boolean query outperforms baseline Okapi

<table>
<thead>
<tr>
<th>run</th>
<th>01-D</th>
<th>02-DN</th>
<th>03-D</th>
<th>04-D</th>
<th>05-D</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td>0.3697</td>
<td>0.3867</td>
<td>0.3719</td>
<td>0.3627</td>
<td>0.2881</td>
</tr>
<tr>
<td>nDCG</td>
<td>0.4117</td>
<td>0.4268</td>
<td>0.4162</td>
<td>0.4072</td>
<td>0.3282</td>
</tr>
<tr>
<td>Q</td>
<td>0.5710</td>
<td>0.5685</td>
<td>0.5881</td>
<td>0.5717</td>
<td>0.4993</td>
</tr>
</tbody>
</table>
Appropriateness of Boolean Query for Filter

- Boolean filter to restrict candidate documents less than 1000 (10 topics).
  - Inappropriate Boolean query (some relevant documents do not match the query): 3 topics (Topic: filter out/total rel 2:26/48, 14: 2/2, 25: 1/3)
Statistical Significance Test between 03-D and 05-D (Baseline: Okapi)

- t-Test with significance level of 0.05 for a two-sided tests
  - Statistically significant
    - nDCG(0.018) and Q(0.040)
  - Not Statistically significant
    - AP(0.055)

- Wilcoxon Signed Rank test
  - Statistically significant
    - AP(0.0015), nDCG(0.0006) and Q(0.0024)
AP Topic by Topic

HU-KB-JA-JA-01-D 0.775 0.269 0.334 0.811 0.711 1 0.097 0.833 0.51 0.176 0.073 0.392 0 0.072 0.497 0.007 0.25 0.176 0.073 0.21 0.008

HU-KB-JA-JA-03-D 0.775 0.258 0.334 0.811 0.711 1 0.097 0.833 0.51 0.176 0.073 0.392 0 0.072 0.497 0.007 0.25 0.176 0.073 0.21 0.008

HU-KB-JA-JA-05-D 0.479 0.306 0.039 0.246 0.674 0.624 1 0.083 0.833 0.398 0.263 0.056 0.21 0.004 0.055 0.461 0.003 0.167 0.125 0.611 0.025 0.167 0.077 0.008
Failure Analysis

- Topic 2
  - “ハリケーン” (hurricane) is recognized as named entity
  - “ハリケーン” (hurricane) without “カトリーナ” (Katrina) get similar score “カトリーナ” (Katrina) without “ハリケーン” (hurricane).

- Topic 11 and 21
  - Those topics don’t contain the named entity information. Topic 11 is a survey type question and is different from assumption of the question.

- Topic 14
  - Named entity keyword “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”.
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).
Future Works

- Consideration of different method to generate variation list for named entity.
  - Usage of Wikipedia redirect
  - Application of transliteration method
  - Part of relationship (Africa and Congo)

- Application of this approach to Web documents
  - There are more varieties of description for named entity in Web documents.