HUKB at NTCIR-12 IMine-2 task: Utilization of Query Analysis Results and Wikipedia Data for Subtopic Mining

Masaharu YOSHIOKA
Hokkaido University
Background and Objectives

■ Background
  – Subtopic mining for supporting users to find out more focused retrieved results
  – Utilization of user query log and external resource for subtopic mining

■ Objectives
  – Evaluation of external resource for subtopic mining candidates
    • query analysis results
    • Wikipedia
Approach

- Selection of subtopic candidates from external resource
  - Wikipedia and query analysis result
- Evaluation of appropriateness of the subtopic candidates
  - Use topic model for checking diversity and representativeness of the candidates
- Vertical intent analysis
  - Subtopic keyword base
  - Analysis of the type of retrieved pages
Subtopic Candidates (Wikipedia)

- Check existence of the disambiguation articles (that belongs to Wikipedia category “曖昧さ回避” (disambiguation))
- Yahoo! abstract: short description + list of chapters
- Use article title in the list for subtopic candidates e.g., Concurrent Versions System, コンビニエンスストア, CVS/ファーマシー… for CVS

Subtopic candidates (概要,歴史,地理,人口,…)
Subtopic Candidates (Query Analysis)

- Query suggestion data provided by organizers (Bing, Google, Yahoo!)
- Query analysis data: coclick and cotopic (30 candidates from each data set)
- Keywords from topic model
  - When the system fails to generate 10 subtopics from above candidates, top 10 characteristic keywords are examined for each topic in the topic model (LDA)
Selection of Subtopic Candidates

- Usage of topic model (LDA)
  - Target document
    - Initial retrieval results provided by organizers
    - Snippet based selection for two or more words query (all query keywords should exist in a given text window; calculated by snippet selection algorithm)
  - Select candidates based on its representativeness in the topic documents.
Assign Subtopic Candidates for Representative Topic of Topic Model

- Construct LDA topic model with topic size = 30
- Representativeness of the keyword for the subtopic
  - \( d_{\text{tr}_t,d_d} \): Variational Dirichlet parameters for each document (did)
  - \( D_{t_{\text{tr}_t,d_d}} = \{\text{did}|d_{\text{tr}_t,d_d} > 0\} \)
  - \( D_{r_{\text{sid}}}: \) Document set retrieved by sid

\[
F_{\text{rep}_{t_{\text{tr}_t,d_d}},\text{sid}} = \frac{|D_{t_{\text{tr}_t,d_d}} \cap D_{r_{\text{sid}}}|}{|D_{t_{\text{tr}_t,d_d}} \cup D_{r_{\text{sid}}}|} \times |D_{t_{\text{tr}_t,d_d}}|
\]
Assign Subtopic Candidates for Representative Topic of Topic Model

- Assign most representative topic for each subtopic candidate
  
  \[ \text{ Assign most representative topic for each subtopic candidate } \]

  \[ q_{\text{tr}, \text{tid}, \text{sid}}: \text{Variational Dirichlet parameters for each subtopic candidate (sid)} \]

\[
R_{\text{tid}}_{\text{sid}} = \arg\max_{\text{tid} \in \{\text{tid} | q_{\text{tr}, \text{tid}, \text{sid}} > 0\}} F_{\text{rep}}_{\text{tid}, \text{sid}} \\
R_{\text{rep}}_{\text{sid}} = F_{\text{rep}}_{R_{\text{tid}}, \text{sid}}
\]
Selection of Subtopic candidates

- Three groups for subtopic candidates
  - Candidates from Wikipedia and query analysis with higher $\text{qtr}_{\text{tid},\text{sid}} > 0.3$
  - Other candidates from Wikipedia and query analysis
  - Keywords from topic model

- Sort subtopic candidates by $\text{Rtrep}$ for each group and pick candidates from the first group
  - Exclude candidates whose representative topics are also selected or whose retrieved results are similar to the selected one.
Vertical Intent Estimation

- **Subtopic keyword base estimation**
  - Comparing subtopic candidate and original query to keywords
  - Image is verified existence of original keyword in Alt of Img tag

<table>
<thead>
<tr>
<th>Intent</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image</td>
<td>イラスト(illustration), 画像(image), 絵(picture), 似顔絵(portrait), 写真(Photo), 壁紙(Wallpaper)</td>
</tr>
<tr>
<td>News</td>
<td>news, ニュース(news), 記事(article)</td>
</tr>
<tr>
<td>Shopping</td>
<td>注文(Order), 価格(price), 値段(price), 料金(price), 相場 (market price), 費用(Cost), 安い(Cheap), 格安(Very cheap)</td>
</tr>
<tr>
<td>QA</td>
<td>質問(Question), 方法(Method)</td>
</tr>
<tr>
<td>Encyclopedia</td>
<td>内容(Contents), 意味(Meaning), 知識(Knowledge), 辞書(Dictionary), Wikipedia</td>
</tr>
<tr>
<td>Web</td>
<td>動画(Video), Youtube</td>
</tr>
</tbody>
</table>
Vertical Intent Estimation (cont.)

- Check by retrieved results
  - Shopping: existence of keywords 注文 (order), 支払 (payment), 送料 (shipping charge), 買い物カゴ (cart) in the page. (5 or more than half)
  - News: check host of url list with news site list constructed by using Open Directory and Google news (10 or more than half)

- Rests are categorized as Web
Experimental Results

- Implementation of the system
  - Tokenizer: JUMAN normalization by using utilization of 代表表記 (Normalized form)
  - Document retrieval system: groonga that supports snippet generation, phrase based retrieval
    http://groonga.org
  - LDA: LDA implemented by Prof. Blei.
    http://www.cs.princeton.edu/~blei/lda-c/
Variation of Submitted Run

- **Target document**
  - All retrieved documents provided by organizers: Q1, Q2, Q3
  - For the queries with two or more keywords, documents that don’t have all query keywords in given window (checked by snippet) are excluded: Q4, Q5

- **Subtopic candidates**
  - All candidates (Query suggestion, Query analysis, Wikipedia): Q1, Q4
  - Query suggestion + Wikipedia: Q2, Q5
  - Wikipedia only: Q3
Evaluation Results

- J-3Q (Wikipedia based candidates only) is significantly worse than others (p < 0.01)
- Utilization of Query analysis results is slightly improve the performance but it is not significant
- Target document selection is almost no effect

<table>
<thead>
<tr>
<th></th>
<th>J-1Q</th>
<th>J-2Q</th>
<th>J-3Q</th>
<th>J-4Q</th>
<th>J-5Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>I-rec@10</td>
<td>0.646</td>
<td>0.632</td>
<td>0.497</td>
<td>0.653</td>
<td>0.645</td>
</tr>
<tr>
<td>D-nDCG@10</td>
<td><strong>0.507</strong></td>
<td>0.475</td>
<td>0.368</td>
<td>0.505</td>
<td>0.470</td>
</tr>
<tr>
<td>D¥#-nDCG@10</td>
<td>0.576</td>
<td>0.553</td>
<td>0.433</td>
<td><strong>0.579</strong></td>
<td>0.557</td>
</tr>
<tr>
<td>V-Score</td>
<td><strong>0.535</strong></td>
<td>0.480</td>
<td>0.384</td>
<td><strong>0.535</strong></td>
<td>0.481</td>
</tr>
<tr>
<td>QU-Score</td>
<td>0.556</td>
<td>0.517</td>
<td>0.408</td>
<td><strong>0.557</strong></td>
<td>0.519</td>
</tr>
</tbody>
</table>
Resource Used for Subtopic Candidates (J-1Q)

- Subtopic candidates from query analysis is frequently selected for representative subtopic
- Most of subtopic candidates selected from Wikipedia is not selected as oracle subtopic candidates

<table>
<thead>
<tr>
<th>Resource</th>
<th>Selection</th>
<th>Unique selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yahoo coclick</td>
<td>260/368 (0.71)</td>
<td>145/218 (0.67)</td>
</tr>
<tr>
<td>Yahoo cotopic</td>
<td>318/447 (0.71)</td>
<td>145/217 (0.67)</td>
</tr>
<tr>
<td>Query suggestion</td>
<td>220/314 (0.70)</td>
<td>80/124 (0.65)</td>
</tr>
<tr>
<td>Wikipedia abstract</td>
<td>23/41 (0.56)</td>
<td>21/39 (0.54)</td>
</tr>
<tr>
<td>Wikipedia alternatives</td>
<td>5/15 (0.33)</td>
<td>3/10 (0.3)</td>
</tr>
<tr>
<td>Topic model</td>
<td>36/115 (0.31)</td>
<td>36/115 (0.31)</td>
</tr>
</tbody>
</table>
Discussion

- Candidates from Wikipedia is not a good one, even though those candidates seems to be reasonable subtopic.
- It may be necessary to take into account the representativeness of the subtopic.
Summary

- Query analysis results is a good resource to estimate good subtopics.
- However, candidates from Wikipedia seem to be reasonable subtopic, but it is not selected as oracle subtopics.
- Further analysis including failure analysis, effect of parameter is necessary.