KSU Systems at the NTCIR-15 Data Search Task

Taku Okamoto and Hisashi Miyamori
Kyoto Sangyo University
NTCIR-15 DAY 3: December 10 Thu. 15:10 - 15:20
Methods

1. Category search
   Narrows down the set of documents to be retrieved by category, to properly capture the scope of the query.

2. Metadata augmentation by table headers
   Augments metadata by table header information, to compensate the short document length of metadata.

3. Reranking by BERT
   Applied the BERT-based reranking, to see how much contribution it achieves for this task.
Category search

• Narrow down the documents by the categories used in Yahoo! Chiebukuro or Yahoo Answers.
  • Japanese: 10 categories, English: 23 categories

• When indexing, each document is assigned a category by a pre-built text classifier.

• When searching, the result is ranked only on the set of documents belonging to the category estimated from the given query
Category search

• Collected documents from Yahoo! Chiebukuro and Yahoo Answers.

• Trained a text classifier with the collected documents.

Table 1. Statistics for the set of documents collected to build the classifier

<table>
<thead>
<tr>
<th>Subtask</th>
<th># of categories</th>
<th>Avg # of doc/cat</th>
<th>Stddev of # of doc/cat</th>
<th>Total # of QA pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Japanese</td>
<td>10</td>
<td>149.4</td>
<td>6.28</td>
<td>1494</td>
</tr>
<tr>
<td>English</td>
<td>23</td>
<td>158</td>
<td>99.26</td>
<td>3541</td>
</tr>
</tbody>
</table>
Augmentation by table headers

• Compensates for the short document length of the metadata.

• Procedure:
  1. Tables in various formats are converted into images.
  2. Cell type, i.e. header or not, is recognized for each cell using a classifier.
  3. Texts are extracted by OCR from cells recognized as header.
Augmentation by table headers

- Compensates for the short document length of the metadata.

- Procedure:
  1. Tables in various formats are *converted into images*.
  2. *Cell type, i.e. header or not, is recognized* for each cell using a classifier.
  3. *Texts are extracted by OCR* from cells recognized as header.

→ Low accuracy of OCR
Augmentation by table headers

- Revised for extra run of Japanese subtask:
  - **Identify the header with the rule:**
    - If the number of non-empty cells in the current line increased from that in the previous line, the current line is considered as a header.
  - Same for columns

<table>
<thead>
<tr>
<th>Origin</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prefectures</td>
<td>Area code</td>
</tr>
<tr>
<td>Hokkaido</td>
<td>0000</td>
</tr>
</tbody>
</table>

Figure 1. Revised extraction of table headers for Japanese subtask
Augmentation by table headers

• Revised for extra run of English subtask:
  • Limit ourselves to PDF files
  • Obtained the entire strings from each file

<table>
<thead>
<tr>
<th>Category</th>
<th>Total $ Spent</th>
<th>% of Total Grant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemical</td>
<td>$315</td>
<td>1.00</td>
</tr>
<tr>
<td>TOTAL</td>
<td>$315</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Category  
Total $ Spent  
% of Total Grant  
Chemical  
$315  
1.00  
TOTAL  
$315  
1.00

Figure 2. Revised extraction of table headers for English subtask
Reranking by BERT

- Apply reranking by BERT to the top set of documents obtained by normal search with BM25.
- Specifically, the sentence level score inferened by BERT is combined with the normal document score according to the following equation:

\[ S_f = a \cdot S_{doc} + (1 - a) \cdot \sum_{i=1}^{n} w_i \cdot S_i \]

- \( S_f \): final doc score
- \( S_{doc} \): doc score before reranking
- \( S_i \): top i-th sentence score by BERT
- \( a, w_i \): parameters
Table 2. Evaluation result for Japanese subtask

<table>
<thead>
<tr>
<th>RUN</th>
<th>Category search</th>
<th>Table header</th>
<th>ranking</th>
<th>Text classifier for category search</th>
<th>Table header extraction</th>
<th>nDCG@10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>POS Vector training</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KSU-J-1</td>
<td>✓</td>
<td>✓</td>
<td>BM25</td>
<td>ALL TF MLP</td>
<td>OCR+CRF</td>
<td>0.391</td>
</tr>
<tr>
<td>KSU-J-3</td>
<td></td>
<td>✓</td>
<td>Bert reranking</td>
<td></td>
<td>OCR+CRF</td>
<td>0.110</td>
</tr>
<tr>
<td>KSU-J-5</td>
<td>✓</td>
<td>✓</td>
<td>BM25</td>
<td>ALL TF MLP</td>
<td></td>
<td>0.413</td>
</tr>
<tr>
<td>KSU-J-7</td>
<td></td>
<td></td>
<td>Bert reranking</td>
<td></td>
<td></td>
<td>0.110</td>
</tr>
<tr>
<td>KSU-J-EX-1</td>
<td>✓</td>
<td>✓</td>
<td>BM25</td>
<td>N+V Fasttest SVM ROW+COL</td>
<td></td>
<td>0.426</td>
</tr>
<tr>
<td>KSU-J-EX-2</td>
<td>✓</td>
<td>✓</td>
<td>BM25</td>
<td>N+V Fasttest SVM ROW</td>
<td></td>
<td>0.276</td>
</tr>
<tr>
<td>KSU-J-EX-3</td>
<td>✓</td>
<td></td>
<td>BM25</td>
<td>N+V Fasttest SVM</td>
<td></td>
<td>0.353</td>
</tr>
<tr>
<td>KSU-J-EX-6</td>
<td>✓</td>
<td>✓</td>
<td>BM25</td>
<td>N+V Fasttest LR ROW+COL</td>
<td></td>
<td>0.426</td>
</tr>
<tr>
<td>KSU-J-EX-7</td>
<td>✓</td>
<td>✓</td>
<td>BM25</td>
<td>N+V Fasttest LR ROW</td>
<td></td>
<td>0.276</td>
</tr>
<tr>
<td>KSU-J-EX-8</td>
<td>✓</td>
<td>✓</td>
<td>BM25</td>
<td>N+V Fasttest LR</td>
<td></td>
<td>0.342</td>
</tr>
</tbody>
</table>
## Result: English

### Table 3. Evaluation result for English subtask

<table>
<thead>
<tr>
<th>RUN</th>
<th>Category search</th>
<th>Table header</th>
<th>ranking</th>
<th>Text classifier for category search</th>
<th>Table header extraction</th>
<th>nDCG@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>KSU-E-2</td>
<td>✔</td>
<td>✔</td>
<td>BM25</td>
<td>ALL TF MLP OCR+CRF</td>
<td></td>
<td>0.240</td>
</tr>
<tr>
<td>KSU-E-4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>OCR+CRF</td>
<td>0.051</td>
</tr>
<tr>
<td>KSU-E-6</td>
<td>✔</td>
<td></td>
<td>BM25</td>
<td>ALL TF MLP</td>
<td></td>
<td>0.240</td>
</tr>
<tr>
<td>KSU-E-8</td>
<td></td>
<td></td>
<td>BM25</td>
<td></td>
<td></td>
<td>0.038</td>
</tr>
<tr>
<td>KSU-E-EX-4</td>
<td>✔</td>
<td>✔</td>
<td>BM25</td>
<td>ALL Fasttest SVM ALL</td>
<td></td>
<td>0.042</td>
</tr>
<tr>
<td>KSU-E-EX-5</td>
<td>✔</td>
<td></td>
<td>BM25</td>
<td>ALL Fasttest SVM</td>
<td></td>
<td>0.181</td>
</tr>
<tr>
<td>KSU-E-EX-9</td>
<td>✔</td>
<td>✔</td>
<td>BM25</td>
<td>ALL Fasttest LR ALL</td>
<td></td>
<td>0.043</td>
</tr>
<tr>
<td>KSU-E-EX-10</td>
<td>✔</td>
<td></td>
<td>BM25</td>
<td>ALL Fasttest LR</td>
<td></td>
<td>0.216</td>
</tr>
</tbody>
</table>
Discussion: Japanese

- Modification on category classifier and header extraction method successfully improved the result.
- Confirmed that headers failed to be extracted properly for some tables.
- Semantic content of the header may need to be considered.
Discussion: Japanese

• Modification on category classifier and header extraction method successfully improved the result.
• Confirmed that headers failed to be extracted properly for some tables.
• Semantic content of the header may need to be considered.

Figure 2. Example of a table where header extraction failed.
Discussion: English

• No improvement could be achieved in the extra run.
• Two reasons:
  1. Large variability of the training data
     • Ave: 158, STD: 99.26, Max: 260, Min: 20
  2. Full text were inappropriate as table headers
     • Excluding the numbers might have led to better results.
Conclusion

• We introduced three methods:
  1. Category search
  2. Metadata augmentation by table headers
  3. Reranking by BERT

• Combined method of category search and BM25 showed the highest score on NDCG@10 among all the official runs.