NII Table Linker at the NTCIR-15 Data Search Task

Re-ranking with Pre-trained Contextualized Embeddings, Data Content, Entity-centric, and Cluster-based Approaches

Phuc Nguyen, Kazutoshi Shinoda, Taku Sakamoto, Diana Andreea Petrescu, Hung Nghiep Tran, Atsuhiro Takasu, Akiko Aizawa, Hideaki Takeda

NTCIR 15 Data Search, 10 December, 2020
Outline (8 mins)

1. Introduction (1 mins)
2. Challenges (1 mins)
3. NII Table Linker (5 mins)
4. Conclusion (1 min)
1. Introduction (1 mins)

- NTCIR 15 Data Search Task
• **Task:** Ad-hoc retrieval for governmental statistical data

• **Input:** keywords

• **Output:** List of dataset ranked by a relevance score

• **Data stats:**

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Japanese</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of datasets</td>
<td>1,338,402</td>
<td>46,615</td>
</tr>
<tr>
<td>Metadata</td>
<td>4.1 GB</td>
<td>0.2 GB</td>
</tr>
<tr>
<td>Data files</td>
<td>1,338,402 (434 GB)</td>
<td>90,388 (714 GB)</td>
</tr>
<tr>
<td>Total dataset</td>
<td></td>
<td>1,2 T</td>
</tr>
</tbody>
</table>
2. Challenges (1 mins)
Data search challenges

- **Data**: Large number of datasets, and fast growth
- **Data context**: quality, trustability
- **Reusable**: machine non-readable
- **Data evolving**: data versioning, changing of publishers
- **Evaluation**: Bias relevant judgment
3. NII Table Linker (5 mins)

- Overall system

- Introduce each runs and results
NII Table Linker System

Query

Anserini

Top k ranking

Re-ranking

Results

Fine-tuned baselines

Reranking models

Fine-tuned BERT

Entity centric

Table content

Cluster-based

Runs

1

2

3

4

5

10

6

7

8

9
Fine-tuned baselines

- Run 1: Fine-tuned BM25
- Run 2: BM25+PRF
- Run 3: Fine-tuned BM25 + PRF

<table>
<thead>
<tr>
<th>Avg metric</th>
<th>English</th>
<th>Japanese</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Default</td>
<td>Fine-tuned</td>
</tr>
<tr>
<td>BM25</td>
<td>0.238</td>
<td>E1: 0.231</td>
</tr>
<tr>
<td></td>
<td>-2.94%</td>
<td></td>
</tr>
<tr>
<td>BM25+PRF</td>
<td>0.219</td>
<td>E3: 0.203</td>
</tr>
<tr>
<td></td>
<td>-7.31%</td>
<td></td>
</tr>
</tbody>
</table>

→ Fine-tuning on the standard methods:
  + English: bad
  + Japanese: Good (J1: 1st in average score)
Sequential Re-ranking with BERT (1)

• Run 4: Top 20 of BM25+PMF (default)
• Run 5: Top 20 of BM25+PRF (fine-tuned)
• Run 10: Top 100 of BM25+PRF (fine-tuned)

<table>
<thead>
<tr>
<th>Avg metric</th>
<th>English</th>
<th>Japanese</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM25+PRF (Top k)</td>
<td>E2: 0.222</td>
<td>J3: 0.435</td>
</tr>
<tr>
<td>Top k + BERT</td>
<td>E4: 0.252</td>
<td>J5: 0.430</td>
</tr>
<tr>
<td></td>
<td>+13.53%</td>
<td>-1.12%</td>
</tr>
</tbody>
</table>

→ English: Good (E-4 : 1st in average score)
Japanese: Bad
Parallel Re-ranking

- Run 6: Entity extraction
- Run 7: Date and Location
- Run 8: Table Content
- Run 9: Cluster based approach

The parallel re-ranking do not perform well

<table>
<thead>
<tr>
<th>Avg metrics</th>
<th>English</th>
<th>Japanese</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM25+PRF</td>
<td>E3: 0.209</td>
<td>J3: 0.435</td>
</tr>
<tr>
<td>Entity Centric</td>
<td>E6: 0.178 -14.6%</td>
<td>J6: 0.343 -21.13%</td>
</tr>
<tr>
<td>Date Location</td>
<td>E7: 0.192 -8.3%</td>
<td>-</td>
</tr>
<tr>
<td>Table Content</td>
<td>E8: 0.189 -9.8%</td>
<td>J8: 0.380 -12.6%</td>
</tr>
<tr>
<td>Cluster-based</td>
<td>E9: 0.221 +6%</td>
<td>J9: 0.421 -3.9%</td>
</tr>
</tbody>
</table>
Extra-runs: Embedding Search

- Only perform on English dataset
- BM25+PMF (Top 10) with fasttext
- Two runs (2, 6) with different weighting

<table>
<thead>
<tr>
<th>Avg metric</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM25+PRF</td>
<td>E2: 0.217</td>
</tr>
<tr>
<td>Embedding Search 1</td>
<td>E-Ex-2: 0.248 +14.3%</td>
</tr>
<tr>
<td>Embedding Search 2</td>
<td>E-Ex-6: 0.235 +8.3%</td>
</tr>
</tbody>
</table>
4. Conclusion (1 min)

- Contributions
- Future work
• **NII Table Linker Contributions:**
  - Preliminary work on DST
  - Study capability of standard methods
  - The four re-ranking approaches
    - pre-trained BERT
    - Entity Centric
    - Table Content
    - Cluster-based

→ English: Best performance using pre-trained BERT
→ Japanese: Best performance using fine-tuned BM25

• **Future work**
  - Data standardization
  - Data semantic annotation
  - Data content retrieval, fact checking

• **Discussion:**
  Please go to our poster at 18:00-19:30 tonight.
Ranking (Top 10) in the average of all metrics

### OFFICIAL EVALUATION

#### English

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.233</td>
<td>0.237</td>
<td>0.248</td>
<td>0.251</td>
<td>0.251</td>
<td>0.264</td>
<td>0.278</td>
<td>0.278</td>
<td>0.278</td>
<td>0.278</td>
</tr>
</tbody>
</table>

#### Japanese

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.395</td>
<td>0.409</td>
<td>0.416</td>
<td>0.445</td>
<td>0.471</td>
<td>0.482</td>
<td>0.464</td>
<td>0.473</td>
<td>0.444</td>
<td>0.483</td>
</tr>
</tbody>
</table>

### EXTRA EVALUATION

#### English

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.227</td>
<td>0.236</td>
<td>0.230</td>
<td>0.255</td>
<td>0.271</td>
<td>0.277</td>
<td>0.239</td>
<td>0.239</td>
<td>0.239</td>
<td>0.239</td>
<td></td>
</tr>
</tbody>
</table>

#### Japanese

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.393</td>
<td>0.396</td>
<td>0.426</td>
<td>0.466</td>
<td>0.480</td>
<td>0.486</td>
<td>0.296</td>
<td>0.472</td>
<td>0.482</td>
<td>0.465</td>
</tr>
</tbody>
</table>

Legend:
- nDCG@3
- nDCG@5
- nDCG@10
- nERR@3
- nERR@5
- nERR@10
- Q-measure