SLWWW at the NTCIR-16 WWW-4 Task

Yuya Ubukata, Masaki Muraoka, Sijie Tao, Tetsuya Sakai
Waseda University
Table of Contents

• Introduction
• Methodology
  • COIL
  • PARADE
  • Run details
• Results
  • NEW runs
  • REP runs
• Discussion
  • per-topic analysis
• Conclusion
Introduction

SLWWW team participated in the NTCIR-16 WWW-4 task

What we have done for the task

• Took 2 different approaches to generate NEW runs
  • COIL [1]
  • PARADE [2]
• Reproduced the KASYS run at the NTCIR-15 WWW-3 task [3]
• Performed per-topic analyses for further discussion
  • Poorly performing topics overall
  • Effect of document length
  • Comparison of COIL and BM25

Methodology
COIL (Contextualized Inverted List)

Retrieval architecture that stores representation vectors into inverted lists to perform contextualized exact matching

Cannot handle vocabulary mismatch problems

COIL introduces contextualized vector representations into the exact matching framework to incorporate the best of two systems

Loses computational efficiency

Exact lexical matching

Soft matching

Luyu Gao, Zhuyun Dai, and Jamie Callan. 2021. COIL: Revisit Exact Lexical Match in Information Retrieval with Contextualized Inverted List.
A Problem in using BERT for document ranking task
- The input length limit of 512 token
  → Unable to handle long documents
- Many approaches have been proposed to overcome this problem

PARADE aggregates passage representations to gain overall document representation

- Documents are split into a fixed number of passages
- Passage representations are computed for each passage-query pair by a pretrained transformer encoder
### Run Details

<table>
<thead>
<tr>
<th>Run name</th>
<th>Divided into chunks of 510 tokens</th>
<th>Corpus type</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLWWW-CO-NEW-2</td>
<td>✓</td>
<td>A</td>
</tr>
<tr>
<td>SLWWW-CO-NEW-3</td>
<td>✓</td>
<td>B</td>
</tr>
<tr>
<td>SLWWW-CO-NEW-4</td>
<td></td>
<td>A</td>
</tr>
</tbody>
</table>

**Reproduced run**
- SLWWW-CO-REP-1
- SLWWW-CO-NEW-5

**PARADE**
- SLWWW-CO-NEW-5

**Corpus A**
- compilation of the top 1,000 most relevant documents for each topic, extracted by BM25

**Corpus B**
- compilation of the top 10,000 most relevant documents for each topic, extracted by BM25
Results
• No statistically significant differences were observed
• SLWWW-CO-NEW-4 performed well in terms of nDCG and Q
• Splitting documents and using a larger corpus did not contribute to the search effectiveness...
• NEW runs based on COIL outperform the baseline → shows the effectiveness of contextualized representations

Results of NEW runs based on the Gold file

<table>
<thead>
<tr>
<th>Run</th>
<th>nDCG</th>
<th>Q</th>
<th>nERR</th>
<th>iRBU</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLWWW-CO-NEW-2</td>
<td>0.3398</td>
<td>0.2718</td>
<td>0.5129</td>
<td>0.7358</td>
</tr>
<tr>
<td>SLWWW-CO-NEW-3</td>
<td>0.3388</td>
<td>0.2670</td>
<td>0.5248</td>
<td>0.7368</td>
</tr>
<tr>
<td>SLWWW-CO-NEW-4</td>
<td>0.3650</td>
<td>0.2891</td>
<td>0.5052</td>
<td>0.7986</td>
</tr>
<tr>
<td>SLWWW-CO-NEW-5</td>
<td>0.3193</td>
<td>0.2538</td>
<td>0.4288</td>
<td>0.7133</td>
</tr>
<tr>
<td>baseline</td>
<td>0.3205</td>
<td>0.2473</td>
<td>0.4541</td>
<td>0.7327</td>
</tr>
</tbody>
</table>

Results of NEW runs based on the Bronze-All file

<table>
<thead>
<tr>
<th>Run</th>
<th>nDCG</th>
<th>Q</th>
<th>nERR</th>
<th>iRBU</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLWWW-CO-NEW-2</td>
<td>0.5600</td>
<td>0.5316</td>
<td>0.7330</td>
<td>0.9244</td>
</tr>
<tr>
<td>SLWWW-CO-NEW-3</td>
<td>0.5464</td>
<td>0.5137</td>
<td>0.7242</td>
<td>0.9192</td>
</tr>
<tr>
<td>SLWWW-CO-NEW-4</td>
<td>0.5750</td>
<td>0.5397</td>
<td>0.7209</td>
<td>0.9213</td>
</tr>
<tr>
<td>SLWWW-CO-NEW-5</td>
<td>0.5410</td>
<td>0.5113</td>
<td>0.6939</td>
<td>0.8888</td>
</tr>
<tr>
<td>baseline</td>
<td>0.5170</td>
<td>0.4806</td>
<td>0.6711</td>
<td>0.8920</td>
</tr>
</tbody>
</table>
Our REP run and the KASYS team’s REV run performed very similarly
→ Succeeded in reproducing the target run to some degree
→ Although we used the provided fine-tuned model ...
Topic Analysis
Highly rated documents by our runs included ...

"idf inventor"
- Different “idf”
  - India Design Forum, Intel Developer Forum, Israeli Defense Forces

"half life"
- Radioactive half-life, Biological half-life
- Download page, news article

Topics are ambiguous and have multiple intents
COIL vs. BM25

- COIL (Run 4): Lexical matching framework with contextualized representations
- BM25 (baseline): Conventional lexical matching methods

- Documents rated higher in BM25 are those that contain the words contained in the topic as they are
- Topics with poor COIL results are cases where contextual information is taken into account, which in turn leads to a discrepancy with the intent of the topic.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Content</th>
<th>Run 4</th>
<th>baseline</th>
<th>difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>214</td>
<td>inventor of the Web</td>
<td>0.7461</td>
<td>0.0568</td>
<td>0.6893</td>
</tr>
<tr>
<td>234</td>
<td>Warriors v.s. NETS 2021</td>
<td>0.5273</td>
<td>0.0000</td>
<td>0.5273</td>
</tr>
</tbody>
</table>

nDCG of the topics where Run 4 significantly outperformed the baseline

<table>
<thead>
<tr>
<th>Topic</th>
<th>Content</th>
<th>Run 4</th>
<th>baseline</th>
<th>difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>210</td>
<td>hypothermia treatment</td>
<td>0.1357</td>
<td>0.6303</td>
<td>0.4946</td>
</tr>
<tr>
<td>240</td>
<td>what is clickbait</td>
<td>0.3748</td>
<td>0.8249</td>
<td>0.4501</td>
</tr>
</tbody>
</table>

nDCG of the topics where the baseline significantly outperformed Run 4
Conclusions

• Our NEW runs outperformed the BM25 baseline
• COIL showed the effectiveness of introducing contextualized vector representations
• Splitting input documents and using a larger corpus did not improve the results
• Successfully reproduced the KASYS team’s run

Future Work

• Reproduce the KASYS team’s run using our own fine-tuned model
• Create a system that also uses the description field
Thank you for your attention