KASYS at the NTCIR-15 WWW-3 Task

Kohei Shinden, Atsuki Maruta, Makoto P. Kato
University of Tsukuba
• NTCIR-15 WWW-3 Task
  – Ad-hoc document retrieval tasks for web documents

• Proposed search model using BERT (Birch)
  – BERT has been successfully applied to a broad range of NLP tasks including document ranking tasks.
Applying a sentence-level relevance estimator learned by QA and microblog search datasets to ad-hoc document retrieval

1. The sentence-level relevance estimator is obtained by fine-tuning the pre-trained BERT model with QA and microblog search data.
2. Calculate BM25 scores and BERT scores for query and document sentences.
3. Weighted sum of the BM25 and the score of the highest BERT-score sentence in the document.
• Weighted sum of the BM25 and the score of the highest BERT-scoring sentence in the document
  – Assuming that the most relevant sentences in a document are good indicators of the document-level relevance [1]
    • $f_{BM25}(d)$: The BM25 score of document $d$
    • $f_{BERT}(p_i)$: The sentence relevance of the top $i$-th sentence obtained by BERT
    • $w_i$: The hyper-parameter $w_i$ is to be tuned with a validation set

$\begin{align*}
  f(d) &= f_{BM25}(d) + \sum_{i=1}^{k} w_i \cdot f_{BERT}(p_i)
\end{align*}$

Preliminary experiments to select datasets and hyper-parameters suitable for ranking web documents

<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Robust04</td>
<td>MS MARCO</td>
</tr>
<tr>
<td>Model MB</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Model CAR</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Model MS MARCO</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Model CAR → MB</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Model MS MARCO → MB</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

The checkmarks represent the data set used for training.

NTCIR-14 WWW-2
Test Collection (with its original qrels)
• Evaluated the prediction results of Birch models
  - Top k sentences: Uses the k-sentence with the highest BERT score for ranking

MSMARCO → MB is the best.
Thus, we submitted runs based on
MS MARCO → MB and CAR → MB.
- MSMARCO→MB is the best. The CAR→MB model also achieved similar scores.
  - The reason why MS MARCO and TREC CAR’s results are better probably because they are web documents retrieval and have a large amount of data.
- BERT is also valid for web document retrieval.

• Achieved the best performances in terms of nDCG, Q and iRBU among all the participants.

<table>
<thead>
<tr>
<th>Baseline</th>
<th>KASYS-E-CO-NEW-1</th>
<th>KASYS-E-CO-NEW-4</th>
<th>KASYS-E-CO-NEW-5</th>
</tr>
</thead>
<tbody>
<tr>
<td>nDCG</td>
<td>0.6935</td>
<td>0.7123</td>
<td>0.7959</td>
</tr>
<tr>
<td>Q</td>
<td>0.7123</td>
<td>0.7123</td>
<td>0.7959</td>
</tr>
<tr>
<td>ERR</td>
<td>0.7959</td>
<td>0.7959</td>
<td>0.9389</td>
</tr>
<tr>
<td>iRBU</td>
<td>0.9389</td>
<td>0.9389</td>
<td>0.9389</td>
</tr>
</tbody>
</table>

KASYS-E-CO-NEW-1:  
- MS MARCO→MB  
- Top 3 sentences

KASYS-E-CO-NEW-4:  
- MS MARCO→MB  
- Top 2 sentences

KASYS-E-CO-NEW-5:  
- CAR→MB  
- Top 3 sentences
• Achieved the **best performances** in terms of nDCG, Q and iRBU among all the participants.

• The effectiveness of BERT in ad hoc web document retrieval tasks was verified.

• MSMARCO→MB is the best. The CAR→MB model also achieved similar scores.

• BERT is also valid for web document retrieval.

Summary of NEW Runs
REP Runs
Replicating and reproducing the THUIR runs at the NTCIR 14 WWW-2 Task

Whether the results between models are consistent with each result.

THUIR

BM25

\land

LambdaMART (learning-to-rank model)

KASYS (ours)

BM25

\land

LambdaMART (learning-to-rank model)
Replication Procedure 1

WWW-2 and WWW-3 topics

Disney, Honda, Switch, Pokemon, Canon, Ice Age

Collection

Clueweb

Ranked by BM25 algorithm

Input

Ranked web documents

Output

1st
Tokyo Disney resort

2nd
Disney official

3rd
Disney shop

... Up to BM25

Extracted eight features

Extracting tf, idf, document length, BM25, LIR as features

Feature extracting program

Extracted eight features

Up to BM25

LambdaMART from here
MQ Track: A dataset of the relevance of a topic and a document.

Replication Procedure 2

- Extraction feature program
  - Extracted features from document
    - qid:001 1:0.2
    - qid:001 1:0.5
    - qid:001 1:0.1
    - qid:001 1:0.9

- LambdaMART
  - Input
  - Output

- MQ Track
  - Train
  - Validate

- WWW-1 test collection
  - Train
  - Validate

- Re-ranked web document
  - 1st: Disney official
  - 2nd: Disney shop
  - 3rd: Tokyo Disney resort
  - ...
• **Features for learning to rank**
  – TF, IDF, TF-IDF, document length, BM25 score, and three language-model-based IR scores

• **The differences from original paper**
  – Although THUIR extracted the features from four fields (whole document, anchor text, title, and URL), we extracted the features from only the whole document
  – Normalization is used by maximum and minimum values because the normalization of features was not described in the original paper
Preliminary Evaluation Results with Original WWW-2 qrels

- Our results is lower than original results
- LambdaMART results were above BM25 for all evaluation metrics
  - Succeeded in reproducing the run
• BM25 results were above LambdaMART for all evaluation metrics
• Failed to reproduce the run
In the original paper, LambdaMART gave better results than BM25, but on the contrary, our BM25 result was better than LambdaMART. We failed to replicate and reproduce the original paper.

Suggestions

- In web search tasks, more effective to extract features from all fields
- Better to clarify the method of normalization in a paper
NEW runs

• Achieved the best performances in terms of nDCG, Q and iRBU among all the participants

• The effectiveness of BERT in ad hoc web document retrieval tasks was verified.
  • MSMARCO→MB is the best. The CAR→MB model also achieved similar scores.
  • BERT is also valid for web document retrieval.

REP runs

• In the original paper, LambdaMART gave better results than BM25, but on the contrary, our BM25 result was better than LambdaMART
  • We failed to replicate and reproduce the original paper