Deep Relevance Ranking Using Enhanced Document-Query Interactions

Ryan McDonald, George Brokos and Ion Androutsopoulos
https://github.com/nlpaueb
Ad-hoc Retrieval - Relevance Ranking

(NL) Query → Document Ranking → Rank → Select top-N

Document Collection
(Deep) Ad-hoc Retrieval / Relevance Ranking

- Interaction-based
  - DeepMatch (Lu and Li 2013)
  - ARC-II (Hu et al. 2014)
  - MatchPyramid (Pang et al. 2016)
  - DRMM (Guo et al. 2016)
  - PACRR (Hui et al. 2017)
  - DeepRank (Pang et al. 2017)

- Relevance-based

Query-Doc term similarity matrices
Deep Relevance Matching Model (DRMM)

Guo et al. 2016

Term Score Aggregation

Term Gates

Term Score

Dense Layers

Doc-aware

Query-term Encodings

Doc-Query Interaction

Query-terms

Document-terms

Relevance Score

Term Score

Dense Layers

Fixed-width

idf

q_3
Deep Relevance Matching Model (DRMM)

Guo et al 2016

Relevance Score

Cosine Similarity
Histograms

Query

Document

A
B
C
A
B
2
3
1
Attention-Based ELement-wise (ABEL-DRMM)

Element-wise Function (e.g., Hadamard)

Attention-based doc-encoding

Differentiable End-to-end Training
Attention-Based EElement-wise (ABEL-DRMM)

Context-sensitive encoder

Attention-based doc-encoding

Element-wise Function (e.g., Hadamard)

Differentiable End-to-end Training
POooled-SImilarTy (POSIT-DRMM)

- Exact Match
- Context Insensitive (w2v)
- Context Sensitive (BiLSTM)

Query

Document

k-max

max, avg

POOLING

concatenate

MULTI-VIEW
(also applied to ABEL-DRMM)
There is more going on … (read the paper)

- Re-ranked top N documents from traditional index using BM25 score

- All models use final linear layer
  - Severyn and Moschitti 2015
  - Mohan et al. 2017

- Pairwise training with negative sampling
## Experimental Datasets

<table>
<thead>
<tr>
<th></th>
<th>BioASQ (Tsatsaronis et al. 2015)</th>
<th>TREC Robust (Voorhees 2004)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Queries</td>
<td>2,251</td>
<td>250</td>
</tr>
<tr>
<td>Train / dev / test</td>
<td>Y1-4 / Y5b1 / Y5b2-5</td>
<td>5X-CV 60/20/20</td>
</tr>
<tr>
<td>Average query length</td>
<td>9.16</td>
<td>2.74</td>
</tr>
<tr>
<td>Collection size (docs)</td>
<td>17.7M</td>
<td>528K</td>
</tr>
<tr>
<td>Average doc length</td>
<td>196.6</td>
<td>476.5</td>
</tr>
<tr>
<td>Vocabulary size</td>
<td>15.2M</td>
<td>1.4M</td>
</tr>
<tr>
<td>Average relevant docs</td>
<td>12.0</td>
<td>69.6</td>
</tr>
</tbody>
</table>
BioASQ (Year 5)

Trad IR/ML:
- BM25: 55.4
- BM25 + extra: 58.1

DL Baselines:
- PACRR: 49.1
- DRMM: 49.3
- PACRR-DRMM: 49.9

Our Models:
- ABEL-DRMM: 50.3
- ABEL-DRMM+(MV): 50.4
- POSIT-DRMM: 50.7
- POSIT-DRMM+(MV): 51.0
TREC Robust 2004

**Lin, Yang, and Lu. SIGIR Forum, December 2018, forthcoming.**
Summary

- Simple architectures can be effective for document ranking
  - Especially on NL query data sets
  - End-to-end training coupled with traditional IR signals
  - End-to-end architectures enables context-sensitive encoders

- POSIT-DRMM vs. ABEL-DRMM
  - Modeling best and top-k-average match improves accuracy

- Multi-view models promotes exact match in addition to vector match
Thanks!

https://github.com/nlpaueb/deep-relevance-ranking