Our Work

Large performance gains achieved by the BERT Cross-Encoder (CE) are not well understood particularly with respect to traditional sparse rankers.

- First, we examine how CE and BM25 rankings relate to each other for different levels of relevance (RQ1, RQ1.3).
- Second, we isolate and quantify the contribution of exact- and soft-term matching to the overall performance (RQ3, RQ4).

Experimental Setup

Model: The vanilla BERT Cross-Encoder (CE) encodes both queries and documents at the same time. Given input \( x \in \{[CLS], q_1, \ldots, q_n [SEP], d_1, \ldots, d_m [SEP]\} \), where \( q \) represents query tokens and \( d \) document tokens. The activations of the CLS token are fed to a binary classifier layer to classify a passage as relevant or non-relevant.

Data: TREC 2020 Deep Learning Track’s passage retrieval task on the MS MARCO dataset [1].


<table>
<thead>
<tr>
<th>Ranker</th>
<th>NDCG@10</th>
<th>MAP</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM25</td>
<td>0.45</td>
<td>0.22</td>
<td>0.16</td>
</tr>
<tr>
<td>Cross-Encoder</td>
<td>0.44</td>
<td>0.21</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Figure 1. Ranking differences between BERT Cross-Encoder (CE) and BM25. Origin of documents in CE ranking at different rank-ranges with respect to the initial BM25 ranking. More intuitively, each row indicates to what ratio documents stem from different rank-ranges. E.g., the top row can be read as the documents in rank 1-10 of the CE re-ranking originate 33% from rank 1-10, 41% from rank 11-100, 19% from rank 101-500 and 6.1% from rank 501-1000 in the initial BM25 ranking. The rank compositions are shown for (a) all, (b) highly relevant, (c) relevant and (d) non-relevant documents according to the NIST 2020 relevant judgments.

RQ1. How do CE and BM25 rankings vary?
- Top-10 ranks vary substantially
- CE brings up many documents from low ranks
- Items ranked high by BM25 are also ranked high by CE

RQ1.2 Does CE better rank the same documents retrieved by BM25?
- Only partially: only 40% agreement of top-10
- CE overestimates the relevance of many non-relevant documents where BM25 scored them correctly lower.

RQ1.3 Does CE better find documents missed by BM25?
- Yes, many high ranked stem from low ranks of BM25 for highly-/relevant
- Note: Some highly-/relevant heavily underestimated by CE compared to BM25

Exact Matches

To isolate and quantify the effect of *exact* matches we mask all non-query terms in the document and test zero-shot.

<table>
<thead>
<tr>
<th>input</th>
<th>NDCG@10</th>
<th>MAP</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Only Q</td>
<td>1.0</td>
<td>0.99</td>
<td>0.9</td>
</tr>
</tbody>
</table>

- Impressive performance (almost whole document is masked)
- Missed potential: performs worse than BM25

Soft Matches

To study *soft* matches we keep all query tokens and mask all others.

<table>
<thead>
<tr>
<th>input</th>
<th>NDCG@10</th>
<th>MAP</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drop Q</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
</tbody>
</table>

- Scoring on only *soft matches* performs on par with BM25.
- Note: BM25 would score random on this input
- In isolation stronger signal than exact matches

References


Read the full paper