The Impact of the Retrieval Stage in Interpolation-based Re-Ranking

Dan-Cristian Ciacu
d.c.ciacu@student.tudelft.nl

Avishek Anand
Responsible Professor

Jurek Leonhardt
Supervisor

1. Introduction

Ad-hoc retrieval. Responsible for retrieving documents that are relevant to a given query.

Retrieve and Re-Rank. Documents are retrieved using a fast retrieval system, then candidate documents are re-scored using a more expensive method.

Interpolation-based Re-Ranking. Documents are re-ranked based on the interpolation between retrieval scores and the values from re-scoring.

Fast-Forward Indexes [1]. Interpolation-based re-ranking that reduces query processing latency through index compression and early stopping.

Interpolation-based re-ranking was mostly evaluated using **simple retrieval methods**. This work explores the effect of different retrievers on various datasets in such setting.

The research question: ***What is the impact of the retrieval stage in the context of interpolation-based re-ranking?***

2. Methodology

Evaluated different retrievers on multiple datasets in an interpolation-based re-ranking setting using **TCT-ColBERT** for re-ranking.

Models. Considered sparse retrievers from three families (based on the employed term-weighting method):

- No-encoder (BM25, TF-IDF)
- Uni-encoder (DeepCT, DeepImpact)
- Bi-encoder (uniCOIL, SPLADE)

Datasets. Eight datasets originating from various domains, e.g. question-answering, web-search, or medical related, were selected.

Metrics. Ranked (Recall, Average Precision, Reciprocal Rank) and user-oriented (nDCG) metrics were used.

3. Results

Retrieval-only Performance

- SPLADE outperformed all the other retrieval models in terms of both recall and nDCG.
- Encoder-based retrievers showed statistically significant improvements in terms of nDCG@10 on 50% of the selected datasets over the no-encoder-based retrievers.
- Regarding the performance in recall, some encoder-based models show no improve-ment; in fact, they are surpassed by BM25 and TF-IDF on five datasets, as illustrated in Table 1.

Re-Ranking Performance.

- Retrieving documents using SPLADE showed substantial improvements over the other models on most datasets. Yet, on datasets with few relevant documents per query, the performance is mixed.
- On the MS MARCO Passage dataset, nDCG values became comparable across the models, showing substantial gains in the ranking quality of some models (as shown in Figure 1).
- For some datasets, re-ranking improved nDCG values, but the difference did not reach statistical significance.

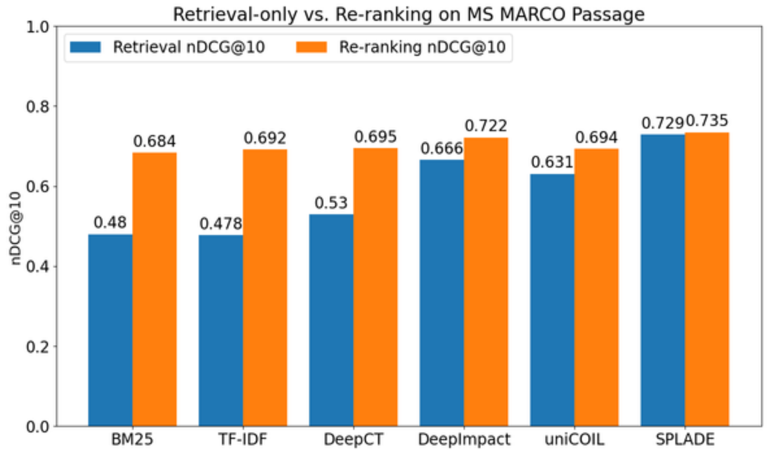

Figure 1: Comparision of nDCG@10 among different retrieval models on the MS MARCO Passage dataset, in both in retrieval-only and interpolation-based re-ranking scenarios.

Query Processing Latency. No-encoder and uni-encoder retrievers showed similar query processing times, ranging from 15ms to 30ms, with comparable ranking performance. Bi-encoder retrievers were about 3 times slower, with latencies between 45ms and 90ms (as shown in Figure 2).

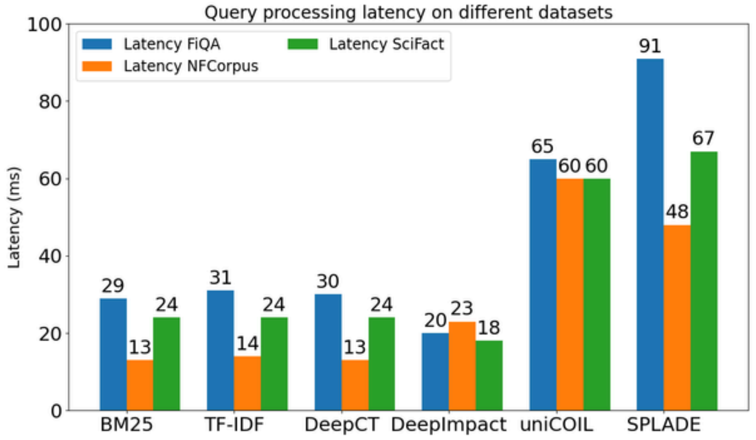

Figure 2: Comparision of query processing latency among different retrieval models on FiQA, NFCorpus and SciFact

	BM25^{1}	TF-IDF^{2}	DeepCT^{3}	DeepImpact^{4}	uniCOIL^{5}	SPLADE^{6}
FiQA	$0.774^{4,5}$	$0.769^{4,5}$	$0.773^{4,5}$	0.747	0.733	**$0.842^{1\text{-}5}$**
NFCorpus	$0.361^{3,4}$	$0.363^{3,4}$	0.351^{4}	0.325	$0.445^{1\text{-}4}$	**$0.579^{1\text{-}5}$**
Scifact	0.970	0.970	0.970	0.956	0.968	**$0.990^{1\text{-}5}$**
Quora	$0.993^{3\text{-}5}$	$0.992^{3\text{-}5}$	$0.990^{4,5}$	0.981	0.984^{4}	**$0.999^{1\text{-}5}$**
HotpotQA	$0.852^{2,3}$	0.850^{3}	0.840	$0.882^{1\text{-}3,5}$	0.850^{3}	**$0.895^{1\text{-}5}$**
DBPedia	$0.660^{4,5}$	$0.660^{4,5}$	$0.669^{4,5}$	0.627	0.611	**$0.783^{1\text{-}5}$**
Fever	0.925	0.925	$0.946^{1,2}$	$0.967^{1\text{-}3}$	$0.969^{1\text{-}4}$	**$0.972^{1\text{-}5}$**
MSMARCO	0.736	0.736	0.744	0.729	0.737	**$0.830^{1\text{-}5}$**

Table 1: Performance in R@1000 among different retrieval models on various datasets

4. Discussion

- Encoder-based retrievers tend to not generalize well when used in an out-of-domain setting.
- The interpolation-based re-ranking stage shows minimal effect when the performance gap between the simple and complex retrievers is small.
- For datasets with shorter queries, SPLADE's query tokenization technique is faster than the dimensionality reduction technique of uniCOIL.

5. Conclusions

Main findings:

- In a retrieval-only setting, SPLADE showed statistically significant improvement in terms of both recall and nDCG over all other models.
- No-encoder-based retrievers benefit from interpolation-based re-ranking, achieving comparable ranking quality to the more complex models.
- Bi-encoder retrieval models add additional overhead to query processing, increasing the latency by 3 times compared to the simpler models.

Future work:

- Re-train the term-weighting neural models on the datasets used for evaluation.
- Consider other retrievers: *TextRank*, graph-based retriever.

References

[1] Jurek Leonhardt et al. "Efficient neural ranking using forward indexes and lightweight encoders". In: ACM Transactions on Information Systems (2023).