Evaluating the Suitability of Interpolation-based Re-Ranking for Ad-Hoc Retrieval

Lucia Navarčíková

01. Introduction

- ***Ad-hoc Retrieval:*** given a query you want to retrieve relevant documents and rank them, low-latency constraints
- ***Sparse Retrieval:*** Traditional approach based on term frequency , fast and efficient, limited to exact terms -> *vocabulary mismatch problem*
- ***Dense Retrieval:*** Condensed document representation, are able to capture semantic relationships, more complex -> higher latency, expensive to compute
- ***Hybrid Retrieval:*** Compute ranking in parallel using a dense and a sparse model and combine them to obtain final ranking -> *missing document scores*
- ***Missing document score :*** sets obtained by sparse and dense retrieval are not identical, one of the document scores is missing for interpolation
- ***Interpolation-based re-ranking:*** technique where a sparse model is used to select a subset of relevant candidates and a more complex model is used to determine final ranking
- **FAST_FORWARD indexes** [1] - framework facilitating interpolation-based re-ranking utilizing dual encoder architecture

02. Research Question

How does interpolation-based re-ranking (using FF indexes) compare to dense and hybrid retrieval models in terms of ranking performance and latency?

RQ1 What is the importance of the lexical component in hybrid retrieval models and interpolation-based re-ranking, respectively?

RQ2 To what extent do missing document scores impact ranking performance in hybrid retrieval models and how can this problem be mitigated?

03. Methodology

Retrieval Approaches

- Interpolation-based Re-ranking - BM25 [2] + TCT-ColBERT [3]
- Dense Retrieval - TCT-ColBERT
- Hybrid Retrieval - BM25 + TCT-ColBERT

Missing Score Alternatives

- Average Score
- Median Score
- Zero
- Drop document

Evaluation Metrics

- RR@10
- nDCG@10
- R@100
- latency

Datasets

Dataset Name	Task	Domain	Corpus	Query
FiQA-2018	Question Answering	Finance	57638	6648
NF Corpus	Information Retrieval	Bio-Medical	3633	323
MS MARCO	Passage-Retrieval	Misc	8841823	6980

04. Results

	FiQA-2018			NF Corpus			TREC-DL-Psg'19		
	nDCG_{10}	R_{100}	RR_{10}	nDCG_{10}	R_{100}	RR_{10}	nDCG_{10}	R_{100}	RR_{10}
Interpolation									
BM25 » TCT-ColBERT	**0.316**	**0.632**	**0.385**	**0.334**	0.254	**0.538**	0.693	0.585	0.808
Dense Retrieval									
TCT-ColBERT	0.265	0.561	0.322	0.267	0.250	0.464	0.670	0.565	0.820
Hybrid Retrieval									
BM25 + TCT-ColBERT	0.313	0.627	0.379	0.330	**0.273**	0.533	**0.705**	**0.615**	**0.831**

Table 4: Ranking Performance. Retrievers use depths $k_S = 1000$ (sparse) and $k_D = 1000$ (dense) with hybrid retrieval reportes with original scores and imputing zero for missing document scores.

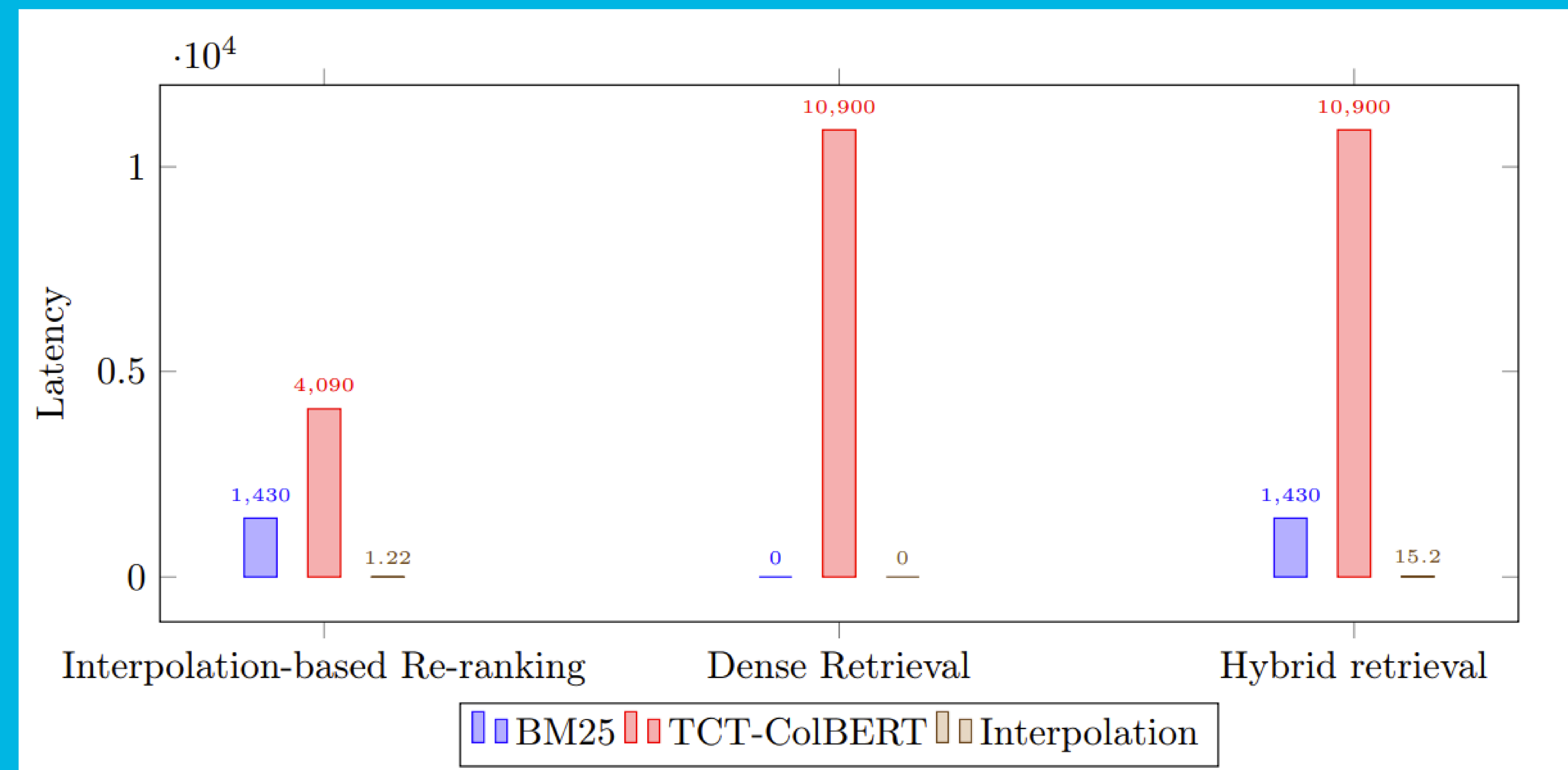
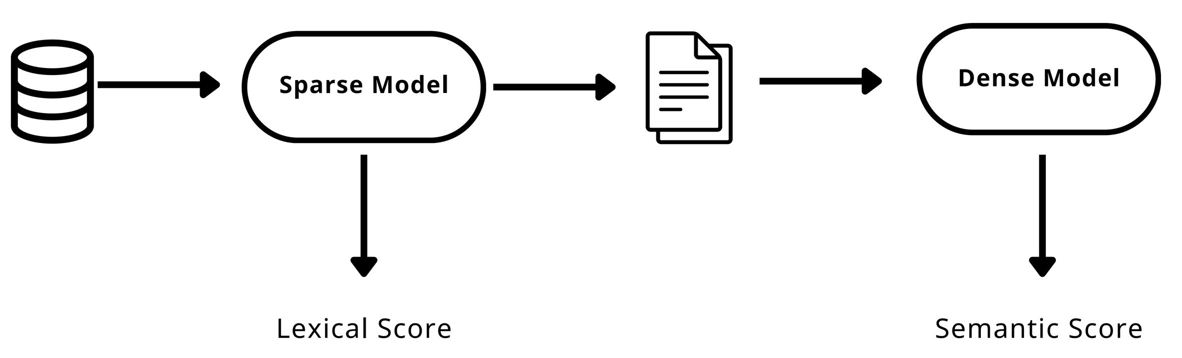


Figure 3: Latency results for 100 queries from FiQA-2018. Latency is reported in miliseconds for all stages - first-stage retrieval, re-ranking and interpolation across all retrieval approaches. Hybrid retrieval is reported for original scores with zero imputation

	FiQA-2018			NF Corpus			TREC-DL-Psg'19		
	nDCG_{10}	R_{100}	RR_{10}	nDCG_{10}	R_{100}	RR_{10}	nDCG_{10}	R_{100}	RR_{10}
Hybrid Retrieval									
↪ original scores									
↪ drop	0.313	0.625	0.379	0.329	0.243	0.535	0.691	0.566	0.808
↪ zero	0.313	**0.627**	0.379	**0.330**	0.273	0.533	**0.705**	**0.615**	0.831
↪ median	0.307	0.594	0.373	0.327	0.278	0.532	0.697	0.586	0.804
↪ average	0.306	0.590	0.372	0.326	0.279	0.529	0.693	0.577	0.797
↪ normalized scores									
↪ drop	**0.314**	0.624	**0.381**	0.329	0.243	0.535	0.688	0.561	0.821
↪ zero	0.280	0.608	0.343	0.326	0.267	**0.536**	0.655	0.585	**0.861**
↪ median	0.309	0.593	0.375	0.328	**0.280**	0.535	0.692	0.574	0.818
↪ average	0.308	0.585	0.374	0.327	0.280	0.532	0.687	0.566	0.818

Table 3: Ranking Performance for different missing score techniques for hybrid retrieval. Retrievers BM25 and TCT-ColBERT use depths $k_S = 1000$ and $k_D = 1000$.

Interpolation-based Re-Ranking



Final Score: $\phi(q,d) = \alpha \cdot \phi_S(q,d) + (1-\alpha) \cdot \phi_D(q,d)$

06. Limitations & Future Work

- Due to time constraints, there is no significance testing
- Multiple datasets from different domains would give a more clear picture of interpolation-based re-ranking and hybrid retrieval
- Possible experimentation of state-of-the-art sparse and dense models
- End-to-end pipeline experiments on larger datasets, considering index storage and leveraging lightweight-encoders

07. Conclusion

- Normalizing scores to offset scale differences brings no benefit
- Best way to deal with missing scores is to zero imputation
- Interpolation-based re-ranking outperforms other approaches on out-of-domain datasets and has lowest overall latency
- Hybrid retrieval achieves best ranking performance for ad-hoc retrieval but for double per query latency

07. References

[1] J. Leonhardt, K. Rudra, M. Khosla, A. Anand, and A. Anand, "Efficient Neural Ranking using Forward Indexes," in Proceedings of the ACM Web Conference 2022, Virtual Event, Lyon France: ACM, Apr. 2022, pp. 266–276. doi: 10.1145/3485447.3511955.

[2] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," FNT in Information Retrieval, vol. 3, no. 4, pp. 333–389, 2009, doi: 10.1561/1500000019.

[3] S.-C. Lin, J.-H. Yang, and J. Lin, "In-Batch Negatives for Knowledge Distillation with Tightly-Coupled Teachers for Dense Retrieval," in Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021), Online: Association for Computational Linguistics, 2021, pp. 163–173. doi: 10.18653/v1/2021.repl4nlp-1.17.

Author: Lucia Navarčíková Contact: L.Navarcikova@student.tudelft.nl Supervisor: Jurek Leonhardt Responsible Professor: Avishek Anand