The impact of the semantic matching within interpolation-based re-ranking

Author:
Alexandru Nistor
a.nistor@student.tudelft.nl
Responsible professor:
Avishek Anand
Supervisor:
Jurek Leonhardt

TU Delft — Delft University of Technology

1. Introduction

Ad-hoc retrieval is the process of returning a ranked list of documents from a large collection based on their relevance to a specific query.

Sparse (lexical) retrieval is represented by fast and efficient methods such as **BM25**, based on TF-IDF. However, it struggles to capture the similarity between the meanings of terms due to its reliance on exact term matching**.**

Dense retrieval addresses this challenge, by utilising low-dimensional vector representations for text. This method can capture the **semantic (meaning) similarity**, but it is inefficient in terms of resources and latency because it employs large Transformer-based language models.

2. Fast-Forward indexes pipeline

This study explores **interpolation-based re-ranking** by using the **Fast-Forward indexes** framework, which employs dual-encoders to leverage semantic matching. The two-stage document retrieval pipeline first utilizes an efficient **sparse retriever** to collect a list of **candidates**, followed by an expensive **semantic re-ranker** which sorts these documents based on the interpolated values of sparse and dense scores.

3. Scientific gap

While it is ideal for text retrieval methods to have an **outstanding ranking performance** and **low latency** in any scenario, achieving this goal is challenging.

Therefore, the aim of this research is to analyse various settings in which specific **models** demonstrate superior performance when employed within the **semantic re-ranking** phase of the **Fast-Forward indexes** pipeline, while considering **trade-offs** between **ranking accuracy** and **latency**.

4. Research questions

What is the impact of the re-ranking model?

RQ1: What is the ranking performance impact of different models during the semantic re-ranking stage?

RQ2: What is the latency impact of different models during the semantic re-ranking stage?

5. State-of-the-art models

Recent research in **general text embedding** presented state-of-the-art models that build upon **BERT-based** architectures. These models differ primarily in their training datasets and minor architectural details.

In this research, models with dimensions of **384** and **768** were explored to balance the trade-offs between **memory usage** and **ranking performance**. Our experiments featured the 768-dimensional versions of **Arctic-Embed**, **BGE**, **GTE**, **E5**, and **Nomic**, alongside the 384-dimensional *bge-small*, *arctic-embed-xs*, and *e5-small*. These models range from **23M** to **137M** parameters, allowing for flexibility in balancing between **efficiency** and **effectiveness** within the **semantic re-ranking stage**.

7. Discussion

Ranking performance impact. It is believed that the **datasets** utilized during the **supervised fine-tuning stage** significantly influence ranking results. For instance, the outstanding performance of **GTE** within **TREC-DL-PSG'19** in the web-search task can be attributed to its inclusion of the **MS MARCO** dataset in its fine-tuning stage, as opposed to **Arctic-Embed** which relies on in-house web-search datasets.

Latency impact. The analysis shows that **384-dimensional models are always faster**. This might be due to fewer computations as **smaller matrix multiplications** are employed in each layer. However, the embedding quality is reduced, leading to a **lower nDGC@10**.

6. Results

	Fast-Forward Indexes: BM25 >>										
		768-dimensional							384-dimensional		
	BM25	tct-colbert	gte-base	bge-base	arctic-m	e5-base	e5-base-pt	nomic	bge-small	arctic-xs	e5-small
TREC-DL-PSG'19	0.4795	0.6924	0.7137	0.6897	0.7042	0.6921	0.5889	0.6977	0.699	0.6924	0.7086
NFCORPUS	0.3223	0.3362	0.3649*	0.3623*	0.3619*	0.355*	0.359*	0.3582*	0.3593*	0.3408	0.3479*
HOTPOTQA	0.5128	0.6363	0.687*	0.7087*	0.7255*	0.6987*	0.6342	0.7307*	0.6873*	0.668*	0.6906*
FIQA	0.2526	0.3139	0.4755*	0.4103*	0.4241*	0.4148*	0.4169*	0.3878*	0.4084*	0.3555*	0.4038*
QUORA	0.7676	0.8464	0.8939*	0.8944*	0.8795*	0.8832*	0.8669*	0.8656*	0.893*	0.8718*	0.8734*
DBPEDIA-ENTITY	0.2744	0.4004	0.4145	0.4101	0.4443*	0.4313*	0.3898	0.4439*	0.4122	0.4071	0.4152*
FEVER	0.4273	0.6887	0.8672*	0.8058*	0.8155*	0.7528*	0.7045*	0.8171*	0.7983*	0.7608*	0.7659*
SCIFACT	0.6722	0.6901	0.7599*	0.7458*	0.7471*	0.7308*	0.7541*	0.7218*	0.7211*	0.7128*	0.7255*

Table 1: Ranking results of the Fast-Forward indexes framework on BEIR and TREC-DL benchmarks (nDCG@10). A retrieval depth of Ks= 1000 was used for the sparse retrieval. For each dataset, the best-performing model is underlined. Statistical significant differences (p $<$ 0.05) between the baseline model (**tct-colbert**) and the analysed models are reported with *.

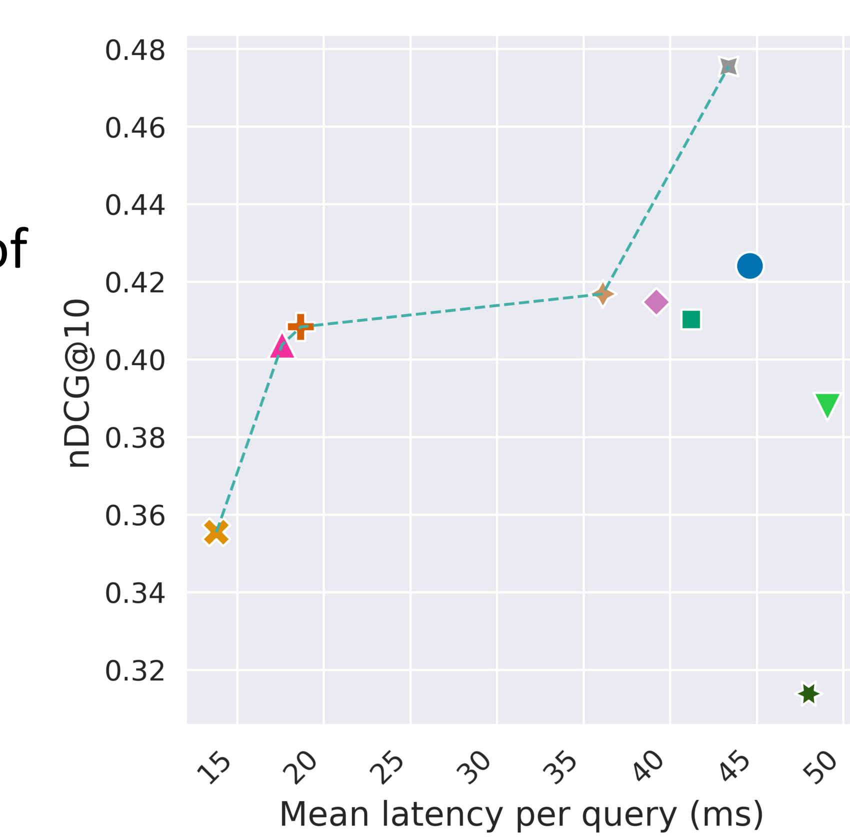

Figure 1: Latency vs. nDGC@10 on FiQA Dataset

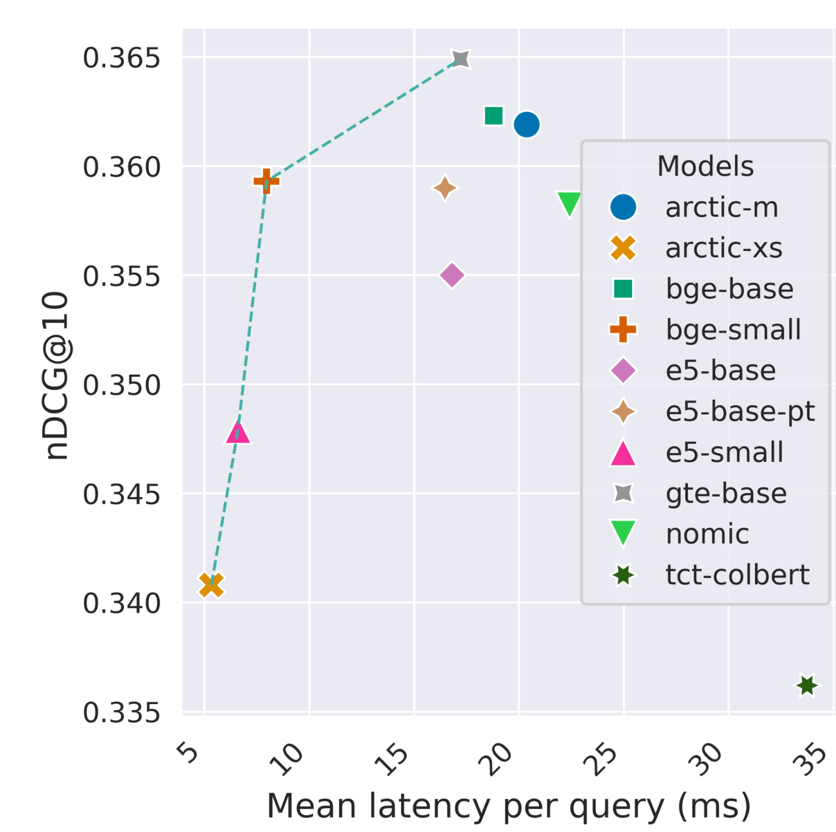

Figure 2: Latency vs. nDGC@10 on NFCorpus Dataset

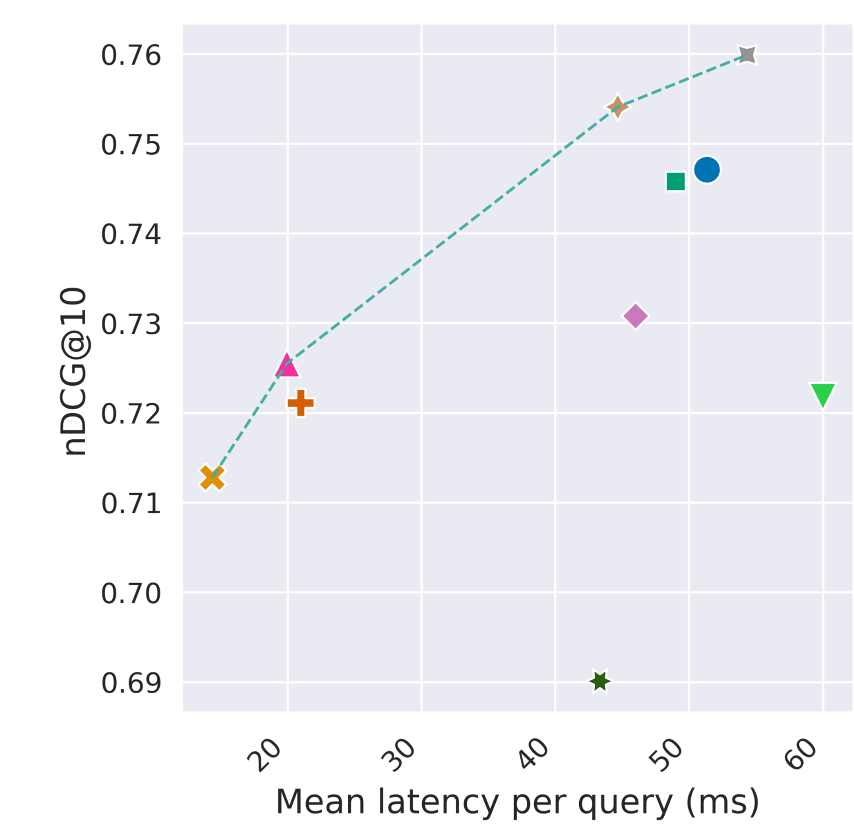

Figure 3: Latency vs. nDGC@10 on SciFact Dataset

The analysis indicates that **GTE** surpasses **BGE** and **Arctic-Embed** in ranking performance across datasets in which the **average document length** exceeds **50** words (TREC-DL-PSG'19, NF-CORPUS, FIQA, FEVER, and SCIFACT). It is hypothesized that the superior performance of GTE stems from its utilization of **mean-pooling** across token representations for text embedding, in contrast to the other two models' reliance on the **[CLS] token embedding**, which is typically used for classification tasks.

Latency is also influenced by the dataset characteristics, with NFCORPUS latency ranging from 5-20 ms and SCIFACT from 15-50 ms. We believe this is due to **the average query lengths** of 3.30 and 12.37 words, respectively.

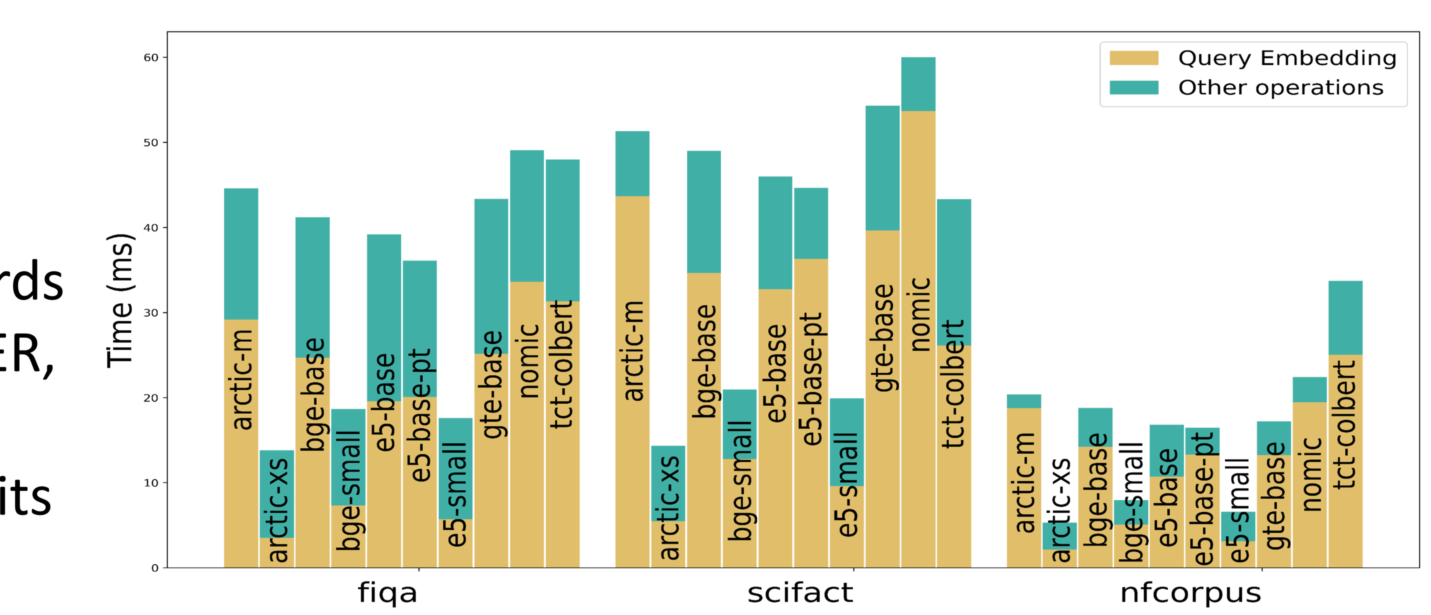

Figure 4: Breakdown of Latency per Query

8. Conclusion

Factors influencing **ranking results**: fine-tuning datasets and the vector embedding computation approach

Factors influencing **latency**: model dimensionality and average query length of each dataset

Future work could explore cross-encoders within semantic re-ranking and employ ablation studies for the fine-tuning hypothesis.