Utilizing Deep Learning for Hybrid Retrieval and Multi-Stage Text Ranking

Keertivardhan Goyal, Yash Mashru and Sanchit Satija

Abstract

This research delves into a multi-stage ranking system that combines sparse and dense retrieval techniques to enhance large-scale information retrieval. Utilizing the TREC Deep Learning Track datasets, our study integrates advanced retrieval algorithms with re-ranking models, achieving high performance metrics. An extensive ablation study highlights the superiority of techniques like Bi-Encoders and Cross-Encoders in achieving optimal ranking outcomes. The challenges and subsequent learnings provide insights into balancing precision and computational efficiency in complex systems.

Keywords

Deep Learning, Hybrid Retrieval, Multi-Stage Text Ranking, Information Retrieval, TREC Deep Learning Track, MS MARCO

1. Problem

Efficiently retrieving and ranking textual data at scale remains a fundamental challenge in information retrieval. Sparse retrieval methods, such as BM25, rely on term matching and fail to capture semantic relationships, whereas dense retrieval models like BERT offer deeper contextual understanding but are computationally intensive. This study introduces a hybrid framework that combines these approaches to optimize precision, recall, and computational efficiency.

2. Dataset

The MS MARCO dataset was central to this study. It includes one million real-world user queries and 8.8 million passages, annotated on a four-point relevance scale ranging from irrelevant to perfectly relevant. To ensure quality, filtering was applied to exclude inappropriate or ambiguous content. The dataset was split into 80% for training, 10% for development, and 10% for evaluation. This large-scale dataset provided a robust foundation for training and evaluating the models.

3. Evaluation Metrics

The evaluation employed Normalized Discounted Cumulative Gain (nDCG@10), a metric that accounts for graded relevance and ranking order. It is particularly suited for retrieval tasks where the relevance of results varies. Using the official TREC evaluation tool's Python interface, nDCG@10 ensured consistent comparisons across all experiments.

4. Methodology

The retrieval stage combined sparse methods, like BM25 with enhancements such as SPARTA, and dense approaches using pre-trained models like BERT and SBERT. FAISS indexing (HNSW and Inner Product) was integrated to improve retrieval speed and accuracy. For re-ranking, Bi-Encoders generated candidate lists, which Cross-Encoders re-ranked based on fine-grained relevance assessments. The combination of these methods formed a multi-stage pipeline aimed at maximizing performance.

Introduction to Information Retrieval (IT550), Autumn 2024, DAIICT, India

202103007@daiict.ac.in (K. Goyal); 202103045@daiict.ac.in (Y. Mashru); 202103054@daiict.ac.in (S. Satija)

5. Results

The hybrid system delivered significant improvements over baseline models. Highlights include:

- SBERT with FAISS indexing (HNSW) achieved an nDCG@10 of 0.7289.
- The Bi-Encoder and Cross-Encoder combination achieved an nDCG@10 of 0.7109.
- Elastic BM25, SPARTA, and other dense retrieval methods showed incremental improvements over traditional BM25.

The results, visualized in Figure 1, underline the effectiveness of combining sparse and dense retrieval methods.

	Model	nDCG@10
0	BM25	0.39
1	Elastic BM25	0.4768
2	SPARTA	0.5087
3	SBERT (trained with in-batch negatives)	0.6431
4	SBERT (pretrained)	0.693
5	SBERT with faiss indexing (HNSW)	0.703
6	SBERT with faiss (HNSW) indexing (retrieve) + cross-encoders (rerank)	0.7087
7	BiEncoder + Cross Encoder	0.7109
8	SBERT with faiss indexing	0.7209
9	SBERT with faiss indexing (IP)(retrieve) + cross-encoders (rerank)	0.7289

Figure 1: Performance comparison of retrieval models on nDCG@10.

Additionally, Table 1 compares the performance of our models with the TREC Deep Learning baseline runs.

Model	nDCG@10
TREC Baseline	0.4012
Elastic BM25	0.4768
SBERT with FAISS (HNSW)	0.7289
Bi-Encoder + Cross-Encoder	0.7109

Table 1
Comparison with TREC Deep Learning baseline results.

A detailed comparison with the TREC 2022 Deep Learning Track results is shown in Table 2, showcasing the competitive performance of our approach.

System	Rank	nDCG@10
UoGTr.D (Top performer)	1	0.7500
CIP.D (Second place)	2	0.7400
YorkU.D	3	0.7300
Ours (SBERT + Cross-Encoder)	-	0.7289
TREC Baseline	-	0.4012

Table 2
Comparison with top-performing systems in TREC 2022 Deep Learning Track.

6. Key Challenges and Learnings

Developing the hybrid retrieval system presented several challenges:

- **Scalability:** Managing and processing large datasets required efficient indexing and model optimization.
- **Balancing sparse and dense techniques:** Combining BM25 and BERT-based models necessitated careful parameter tuning to leverage their strengths.
- **Re-ranking complexity:** Cross-Encoder models improved precision but added computational overhead.

Key learnings include the importance of fine-tuning pre-trained models, leveraging hybrid approaches to improve performance, and using re-ranking strategies to enhance top-ranked results.

7. Conclusion

This research demonstrates the potential of hybrid frameworks in advancing large-scale text retrieval. By integrating sparse and dense retrieval methods with re-ranking strategies, the proposed system achieves superior performance metrics on the MS MARCO dataset. Future work will focus on refining the pipeline to handle even larger datasets and exploring advanced re-ranking algorithms to further enhance scalability and precision.

8. Biblography

- TREC2022, Ellen M. Voorhees and Ian Soboroff, Overview of the TREC 2022 Deep Learning Track, Proceedings of the Text Retrieval Conference (TREC 2022), https://trec.nist.gov/pubs/trec31/papers/Overview_deep.pdf
- BEIR2021, Thakur, Nandan and Reimers, Nils and Daxenberger, Johannes and Gurevych, Iryna, BEIR: A Heterogeneous Benchmark for Zero-Shot Evaluation of Information Retrieval Models, NeurIPS Proceedings, 2021, https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/65b9eea6e1cc6bb9f0cd2a47751a186f-Paper-round2.pdf
- HNSW2021, Malkov, Yu A. and Yashunin, D. A., Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, https://towardsdatascience.com/similarity-search-part-4-hierarchical-navigable-small-world-hnsw-2aad4fe87d37
- MSMARCO, Microsoft, MS MARCO Dataset Documentation, 2018, https://microsoft.github.io/msmarco/
- SmartReply, Kannan, Anjuli and Kurach, Karol and Ravi, Sujith and Tavenard, Romain, Smart Reply: Efficient Natural Language Response Suggestion for Email, arXiv preprint, 2017, https://arxiv.org/pdf/1705.00652.pdf