Hybrid Retrieval and Multi-stage Text Ranking Solution at TREC 2022 Deep Learning Track

Guangwei Xu, Yangzhao Zhang, Longhui Zhang, Dingkun Long, Pengjun Xie, Ruijie Guo
Alibaba Group
kunka.xgw,zhangyanzhao.zyz@alibaba-inc.com
pinghe.zlh,dingkun.ldk@alibaba-inc.com
pengjun.xpj,ruijie.guo@alibaba-inc.com

Abstract

Large-scale text retrieval technology has been widely used in various practical business scenarios. This paper presents our systems for the TREC 2022 Deep Learning Track. We explain the hybrid text retrieval and multi-stage text ranking method adopted in our solution. The retrieval stage combined the two structures of traditional sparse retrieval and neural dense retrieval. In the ranking stage, in addition to the full interaction-based ranking model built on large pre-trained language model, we also proposes a lightweight sub-ranking module to further enhance the final text ranking performance. Evaluation results demonstrate the effectiveness of our proposed approach. Our models achieve the **1st** and **4th** rank on the test set of passage ranking and document ranking respectively.

1 Introduction

The text retrieval task is usually divided into two sub tasks according to the length of the text: passage retrieval and document retrieval. Whether it is a Document ranking or Passage ranking task, under the setting of full ranking, it generally needs to be processed in two stages: retrieval and ranking. Among them, the retrieval stage needs to quickly find the most relevant **top-k** candidates in the entire corpus set given the input query, and then a more complex and accurate ranking stage will be performed over the **top-k** candidates thus producing the final ranked result.

For the retrieval stage, BM25 [RZ$^+$09] is a bag-of-words retrieval method that ranks a set of documents based on the query terms appearing in each document and is one of the best retrieval algorithms. Recently, various improved methods [NLE19, FLPC21] based on BM25 have been proposed in the sparse retrieval research field. With the development of deep neural network models, the performance of dense retrieval models such as DPR [KOM$^+$20], and coCondenser [GC21, GC22] have surpassed traditional methods with a large margin. For the ranking stage, since the amount of processed data is greatly reduced, a more complex model structure can be adopted. Benefit from the excellent performance of large-scale pre-trained language model (e,g. BERT [DCLT19]) on various natural language processing tasks, the ranking model is also gradually turning to the BERT model as the relevance scoring model. Specifically, the query and doc will be concatenated as the input of BERT or other pre-trained language models like RoBerta [LOG$^+$19], ERNIE [LLL$^+$22], ELEC-TRA [CLLM20], and Deberta [HLGC20].

Based on this multi-stage processing method, our main optimization methods in this evaluation are:

1. In the retrieval stage, a hybrid retrieval method that combines traditional sparse retrieval and dense retrieval model is adopted, and the ROM [LZXX22] model we proposed for retrieval tasks is used on the dense retrieval pre-training model.

2. In the ranking stage, multiple backbone networks are used to train the ranking models, and the re-ranking HLATR [ZLXX22] model we proposed before is used to further improve the final ensemble performance when merging these ranking models.

2 Methodology

From the query to the final retrieval and ranking results, our method is divided into three stages: retrieval, ranking, and HLATR re-ranking. The specific process is shown in Figure 1. The difference between the Document ranking and the Passage ranking is that document is first divided into passages to participate in the retrieval and ranking, and finally the passage with the highest score is taken as the representative of the document. The following will introduce the specific practices of these three stages.
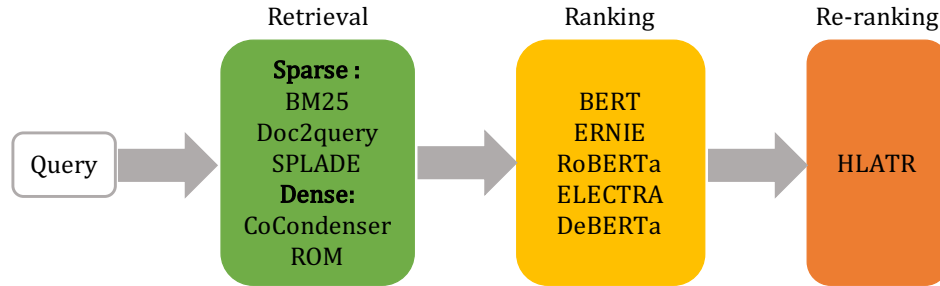

Figure 1: Model Architecture Diagram.

2.1 Retrieval stage

Sparse retrieval BM25 has a good performance as a classic efficient scoring algorithm in the retrieval stage, but it is limited by the text matching between query and doc, and has no semantic expansion ability. Doc2query and SPLADE have extended the semantics from different perspectives, which can bring significant improvement in the sparse retrieval mode. Here we use BM25, Doc2query and SPLADE at the same time to perform weighted ensemble on the final score.

Dense retrieval The retrieval stage is limited by computational efficiency. At present, the two-tower structure is basically used. The key is how to obtain a good twin-tower text semantic representation model. As a pre-training model, BERT can make full use of unlabeled corpus and can easily do fine-tuning downstream tasks, and is widely used here. For retrieval tasks, there are also many works to optimize the model structure or training loss function. The ROM model we use is optimized for the random mask in the MLM task to a weighted mask based on term weighting, which reduces the probability of the stop words that have a weak role in retrieval being masked. Such adjustments can make the final trained model more suitable for retrieval tasks.

Hybrid retrieval Although the performance of the dense retrieval method has far exceeded that of the sparse retrieval method, we found that the results of the two methods can be further improved by weighted ensemble. It may be that the focus of the two methods is different.. The sparse retrieval method pays more attention to the text literal matching information, and the dense retrieval method pays more attention to the semantic matching information, and the two can complement each other.

2.2 Ranking stage

Single-tower interaction model Compared with the retrieval stage, the amount of data processed in this stage is greatly reduced, and a single-tower model structure in which query and doc interact together can be used to model more accurately. The pre-training model of the BERT structure is also used here, and negative sampling is performed based on the results of the retrieval stage, and the rdrop method is used for the loss function.

Multi-model ensemble Here we tried a variety of backbone models for training, including BERT, RoBERTa, ERNIE, ELECTRA, and DeBERTa. In the last model trained under different parameter settings, ten models with better performance were selected, while keeping the diversity of the backbone as much as possible.
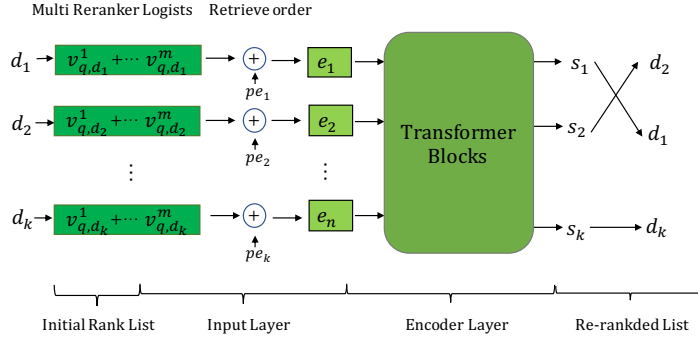
Figure 2: HLATR Re-ranking Model.

Table 1: Final results on TREC 2022 passage dataset.

Run	NDCG@10(Dev)	NDCG@10(Eval)
pass1	0.7602	0.7050
pass2	0.7568	0.7105
pass3	0.7572	**0.7184**

2.3 HLATR re-ranking stage

When training the ranking model, we found that as the candidate set participating in the ranking stage increases, the final ranking performance will decrease. In the training stage, each query only interacts with a small number of negative samples, but in the testing stage, a larger candidate set needs to be interacted with. In addition, we made a simple weighted ensemble between the ranking model stage and the retrieval stage, and found that it can bring significant improvement. Based on these two findings, we introduce the HLATR re-ranking stage, using the Transformer structure to fuse the first two stages together. The model structure is shown in Figure 2.

3 Results

3.1 Main results

The running results we finally submitted, the Passage ranking task is in Table 1, and the Document ranking task is in Table 2.

The three sets of results we submitted for the Passage ranking task were produced by different model ensemble, and finally pass3 got the best results on the test set.

For the Document ranking task, the first two sets of results we submitted tried to train the model directly based on the document data, and doc3 obtained the highest score after training the model based on the passage data as the document score. It may be due to the difference in the distribution of Dev and Eval data, the results of the first two groups dropped a lot, and only doc3 finally got normal results.

3.2 Ablation study

Since there are too many model combinations involved in the ranking model stage and the HLATR re-ranking stage, including a lot of parameter tuning work. Different data sampling and training result selection will have different performance, so the ablation experiments of these two stages are not listed here. In contrast, the retrieval stage is more valuable. We list the experimental results on the Passage ranking task in Table 3, and the experimental results on the Document ranking task in Table 4.

Whether it is Passage ranking task or Document ranking task, BM25 is a strong baseline, and the Doc2query can further strengthen the baseline. The condenser model based on the dense retrieval architecture is similar

Table 2: Final results on TREC 2022 document dataset.

Run	NDCG@10(Dev)	NDCG@10(Eval)
doc1	0.7643	0.4908
doc2	0.7631	0.4555
doc3	-	**0.7533**

Table 3: Ablation experiments result on TREC 2021 passage dataset.

Model	AP	NDCG@10	R@1000	MRR@100
BM25	0.0977	0.3977	0.5835	0.5303
Doc2query	0.1358	0.4463	0.6147	0.5076
ROM	0.2965	0.6479	0.8251	0.7865
coCondenser	0.2857	0.6520	0.8095	0.7601
BERT-large-ROM	0.2919	0.6471	0.8235	0.7799
SPLADE	0.3329	0.6844	**0.8588**	0.8289
ensemble	**0.3792**	**0.7426**	0.8185	**0.8841**

Table 4: Ablation experiments result on TREC 2021 document dataset.

Model	NDCG@10
BM25	0.4370
Doc2query	0.5703
Condenser	0.5778
ROM	0.6115
SPLADE	0.6202
coROM (trained on document)	0.5497
coCondenser (trained on document)	0.5871
coROM (trained on passage)	0.5815
ensemble	**0.7401**

to the ROM model, and SPLADE performed very well on this evaluation data set. The results of the ensemble of multiple models in the final retrieval stage are very close to the final submitted results, which shows the importance of this stage.

4   Conclusion

In this TREC 2022 Deep Learning Tack evaluation, we participated in the full ranking setting of both the Passage ranking and Document ranking tasks. Through the technology of hybrid retrieval and multi-stage ranking model, the ensemble of multiple current mainstream algorithms has achieved relatively good results. But there is still a lot of room for improvement in the ranking stage and Document ranking task.

References

[CLLM20] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020.

[DCLT19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[FLPC21] Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. Splade v2: Sparse lexical and expansion model for information retrieval. *arXiv preprint arXiv:2109.10086*, 2021.

[GC21] Luyu Gao and Jamie Callan. Condenser: a pre-training architecture for dense retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 981–993, 2021.

[GC22] Luyu Gao and Jamie Callan. Unsupervised corpus aware language model pre-training for dense passage retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2843–2853, 2022.

[HLGC20] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*, 2020.

[KOM+20] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6769–6781, 2020.

[LLL+22] Yuxiang Lu, Yiding Liu, Jiaxiang Liu, Yunsheng Shi, Zhengjie Huang, Shikun Feng Yu Sun, Hao Tian, Hua Wu, Shuaiqiang Wang, Dawei Yin, and Haifeng Wang. Ernie-search: Bridging cross-encoder with dual-encoder via self on-the-fly distillation for dense passage retrieval, 2022.

[LOG+19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[LZXX22] Dingkun Long, Yanzhao Zhang, Guangwei Xu, and Pengjun Xie. Retrieval oriented masking pre-training language model for dense passage retrieval. *arXiv preprint arXiv:2210.15133*, 2022.

[NLE19] Rodrigo Nogueira, Jimmy Lin, and AI Epistemic. From doc2query to docttttttquery. *Online preprint*, 6, 2019.

[RZ+09] Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

[ZLXX22] Yanzhao Zhang, Dingkun Long, Guangwei Xu, and Pengjun Xie. Hlatr: Enhance multi-stage text retrieval with hybrid list aware transformer reranking. *arXiv preprint arXiv:2205.10569*, 2022.