Query Term Expansion and Ranking Stability in TREC 2022 Deep Learning Track Systems

Assignee Research

June 14, 2026

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 performanc

1 Introduction

This paper examines: Hybrid Retrieval and Multi-stage Text Ranking Solution at TREC 2022 Deep Learning Track. Research question: What is the impact of query term expansion on the ranking stability of multi-stage text ranking systems in the TREC 2022 Deep Learning Track?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.9/10.

3 Results

12 papers retrieved. 16 claims extracted; 16 independently verified. Quality review score: 7.9/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The running results we finally submitted, the Passage ranking task is in Table 1, and the Document ranking task is in Ta	✓	0.36
The three sets of results we submitted for the Passage ranking task were produced by different model ensemble, and final	✓	0.38
For the Document ranking task, the first two sets of results we submitted tried to train the model directly based on the	✓	0.42
It may be due to the difference in the distribution of Dev and Eval data, the results of the first two groups dropped a	✓	0.34
Ablation study: Since there are too many model combinations involved in the ranking model stage and the HLATR re-ranking	✓	0.44
In contrast, the retrieval stage is more valuable. We list the experimental results on the Passage ranking task in Table	✓	0.38
Whether it is Passage ranking task or Document ranking task, BM25 is a strong baseline, and the Doc2query can further st	✓	0.30
The condenser model based on the dense retrieval architecture is similar.	✓	0.19
From the query to the final retrieval and ranking results, our method is divided into three stages: retrieval, ranking,	✓	0.28
The difference between the Document ranking and the Passage ranking is that document is first divided into passages to p	✓	0.31
BM25 has a good performance as a classic efficient scoring algorithm in the retrieval stage, but it is limited by the te	✓	0.33
Doc2query and SPLADE have extended the semantics from different perspectives, which can bring significant improvement in	✓	0.29
Here we use BM25, Doc2query and SPLADE at the same time to perform weighted ensemble on the final score.	✓	0.28
The retrieval stage is limited by computational efficiency. At present, the two-tower structure is basically used.	✓	0.27
The key is how to obtain a good twin-tower text semantic representation model.	✓	0.25
As a pre-training model, BERT can make full use of unlabeled corpus and can easily do fine-tuning downstream tasks.	✓	0.29

References

- http://arxiv.org/abs/2308.12039v1
- http://arxiv.org/abs/2003.11650v1
- http://arxiv.org/abs/2008.02460v1