Document Ranking with Pre-trained Sequence-to-Sequence Modal

Mehta Ayushi, Patel Pratham, Munjani Nishit and Rathod Rohit

Abstract

The primary objective of the research is to improve document ranking in Information Retrieval (IR) by leveraging pre-trained sequence-to-sequence model like T5. Traditional classification-based approaches, such as those using BERT, limit the ability to utilize latent semantic and linguistic knowledge during reranking. This research seeks to explore a generation-based ranking formulation in which relevance labels are treated as target tokens. The approach emphasizes data efficiency and performance, especially in data-scarce scenarios, to address limitations in previous methodologies.

Keywords

Document Ranking, Information Retrievel, BERT, T5,

1. Introduction

A straightforward approach to ranking is to frame the task as a classification problem. Candidate items are assigned probabilities of belonging to a target class, and these probabilities are used to rank the items. In the context of document ranking for information retrieval, given a query and a document, the system's objective is to compute the probability that the document is relevant to the query. By sorting documents based on these probability scores, the model can rank documents in the corpus according to their relevance to the query.

The contribution of this work is to adapt a pretrained sequence-to-sequence model (in our case, T5) to the task of document reranking. To our knowledge, this is a novel use of this class of models that has not been previously described in the literature. In a data-rich regime, with lots of training examples, our method can outperform a pure classification-based encoder-only approach.

To leverage sequence-to-sequence models such as T5 for ranking tasks, the probability of the output sequence (e.g., "true" or "false") generated during inference is utilized. This probability, derived from the model's output logits, reflects the model's confidence in the generated sequence. By using this probability score as a ranking signal, it becomes possible to compare and rank different documents based on their relative relevance to the query.

2. Dataset

MS MARCO Passage Dataset:

- Contains more than 8.8 million passages retrieved from Bing search engine results.
- Training data consists of approximately 500,000 query-relevant passage pairs.
- Each query has, on average, one relevant passage.
- The development and test sets include 6,900 queries each, but only the development set has publicly available relevance labels.

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

202311008 (M. Ayushi); 202311022 (P. Pratham); 202311026 (M. Nishit); 202311039 (R. Rohit)

3. Evaluation Metrics

MRR@10 (Mean Reciprocal Rank):

- Evaluates the ranking position of the first relevant document in the top 10 results.

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i} \tag{1}$$

Where:

- $|Q|$ is the total number of queries.
- rank_i is the rank position of the first relevant document for the i-th query.

4. Results

- Here are the results of the authors:

Method	MRR@10
BM25	0.184
+ BERT - Large	0.372
+ T5 - Base	0.363
+ T5 - Large	**0.383**

- Here are the results reproduced from our end:

Method	MRR@10
BM25	0.183
+ BERT - Base	0.331
+ T5 - Base	**0.356**

5. Key challenges and Learnings

1. **Key Challenges:**
   - **Reproducing Results with Different Architectures:** Matching the original paper's results for models like BM25, T5, and BERT presented challenges, especially due to slight variations in implementation and training conditions.
   - **Computational Limitations:** Training (Fine-tuning models like T5-Base on available hardware required optimizing resource usage and managing memory constraints effectively.
   - **Evaluation Challenges:** Calculating metrics like MRR@10 accurately and aligning with the authors' methods required thorough understanding and precise validation processes.
2. **Key Learnings:**
   - **Architecture Understanding:** Gained in-depth insights into advanced language models like T5 and BERT, particularly in the context of query-document relevance modeling.
   - **Reproducibility Importance:** Learned the critical need for detailed experimental setups and clear documentation to reproduce the authors' results accurately.
   - **Effective Resource Utilization:** Optimized training processes to handle large models and datasets within computational resource constraints.

This document refers to this paper [1].

References

[1] R. Nogueira, Z. Jiang, J. Lin, Document ranking with a pretrained sequence-to-sequence model, arXiv preprint arXiv:2003.06713 (2020).