A Comparative Analysis of docTTTTTquery and SPLADE++ on Passage and Document Ranking

Theresa Van[1], Hanlin Wang[2], and Hansaem Park[3]

[1]Tandon School of Engineering, New York University
[1]ttv2006@nyu.edu
[2]hw2664@nyu.edu
[3]hp2229@nyu.edu

Abstract

This paper investigates the performance of two document expansion techniques, docTTTTTquery and SPLADE++, in the context of passage ranking and document ranking inference utilizing the Microsoft Machine Reading Comprehension (MSMARCO) dataset. The study analyzes and compares the performance of docTTTTTquery and SPLADE++ across passage and document ranking tasks, providing insights into their adaptability, robustness and effectiveness.

1 Introduction

In the realm of information retrieval, the quest for models that understand and align with human semantic intent continues to be paramount. This paper delves into the intricate world of document expansion techniques, focusing on two seminal methods: docTTTTTquery and SPLADE++. Document expansion techniques are pivotal in enhancing the performance of retrieval systems by enriching documents with additional context, thereby bridging the lexical gap between query and document language. This study aims to shed light on the effectiveness of these techniques in the dual contexts of passage ranking and document ranking, two fundamental tasks in information retrieval that differ significantly in their challenges and nuances.

docTTTTTquery represents a paradigm in document expansion, leveraging deep transformer models to generate potential query terms that augment the document's representation, thus improving retrieval recall [1]. On the other hand, SPLADE++, an evolution of the Sparse Lexical and Expansion model, brings forward a method that emphasizes sparse and efficient representation, marrying the lexicon-based approach with modern deep learning advancements [2]. This paper aims to critically analyze and compare the performance of docTTTTTquery and SPLADE++ across passage and document ranking tasks, providing insights into their adaptability, effectiveness, and efficiency.

Central to this study is the utilization of the Microsoft Machine Reading Comprehension dataset, commonly known as MSMARCO [3]. This dataset is a large-scale, public benchmark specifically designed for deep learning research in information retrieval. It comprises two distinct tracks: the Passage Ranking and the Document Ranking datasets, each tailored to evaluate and advance the state-of-the-art in their respective areas. The MSMARCO Passage dataset provides a collection of passages from real-world documents, accompanied by a set of queries and relevant passage annotations, ideal for assessing the precision and effectiveness of passage ranking techniques. Meanwhile, the Document Ranking dataset expands the scope by including entire documents, offering a more comprehensive and challenging environment to test document ranking capabilities. These datasets are distinguished by their origin from genuine user queries, offering a realistic and diverse set of challenges that mirror actual information-seeking behavior. The use of MSMARCO as a benchmark in this study allows for

a rigorous, standardized comparison of document expansion techniques, providing insights into how these methods translate into real-world efficacy in both passage and document retrieval contexts.

While these technologies represent the cutting edge of retrieval and expansion techniques, this study also acknowledges the dynamic nature of the field. The pursuit of more effective, efficient, and equitable information retrieval systems is a continuing challenge, driving ongoing research and development.

2 Related Work

The research presented by Weller(2023)[4] delves into the complexities of generative query and document expansion. This study uniquely introduces HyDE, a novel approach where a language model is guided by specific tasks to create documents responsive to particular queries. The researchers build upon existing methods, extending their application to new datasets not previously examined. Additionally, the paper rigorously investigates the shortcomings of the HyDE methodology, offering an in-depth analysis of the various elements that influence the effectiveness of generative query and document expansion strategies.

Complementing the advancements in generative query and document expansion, the study by Yu et al. (2023) [5] introduces "Fusion-in-T5" (FiT5), a approach that amalgamates disparate document ranking signals to optimize information retrieval. Capitalizing on the T5-base model, FiT5 incorporates global attention modules and retrieval score features to refine the re-ranking stage, setting a new precedent in the field. Evaluated on the MSMARCO and TREC Deep Learning Tracks of 2019 and 2020, FiT5 demonstrates its superiority by outperforming both conventional two-stage retrieve-and-rerank pipelines and more elaborate ranking architectures. The studies additionally uncover the significance of the combined ranking attributes and global attention mechanisms, noting that their absence leads to a substantial decrease in performance. This work is instrumental in clarifying the efficacy of a collaborative approach to document ranking, offering valuable insights for future explorations in the field.

3 Approach

The core of our experimental approach centered around the indexing of documents and passages using the docTTTTTquery and SPLADE++ expansion methods. The aim is to assess and compare their performance in terms of accuracy, efficiency, and relevance of the results. We leverage the high-performance computing (HPC) resources provided by New York University (NYU) to conduct an extensive analysis on the MSMARCO passages and documents dataset. The HPC's scalable resources and advanced computing environment ensured that the computational intensity of document expansion and indexing operations was managed effectively, allowing for accurate and timely experimentation.

Table 1: SLURM Job Script Parameters

Parameter	**Value**
`Nodes`	1
`Number of tasks per node`	1
`CPUs per haitask`	8
`Memory`	128GB
`GPU`	1 (2 for SPLADE++)

Key to our experimental setup was PyTerrier, an IR platform built on Python that contains an extensive collection of plugins and integrations with various IR tools that made it an ideal choice for this study. We utilized PyTerrier for its comprehensive retrieval functionalities, including indexing, retrieval, and its evaluation components, which were crucial for handling the MSMARCO datasets and implementing the docTTTTTquery and SPLADE++ methods.

As the PyTerrier libraries were initially configured for compatibility with passage ranking, extra preprocessing steps were taken with the MSMARCO documents dataset. A `text` attribute was introduced for each document which consisted of concatenating the document's title and body into one unified text block. The document expansion techniques were then applied to the `text` attribute.

This ensures that the documents' representations reflect both the summarization effect of the title and the detailed exposition of the body.

To validate and benchmark the document expansion techniques employed in this study, topics and queries from the Text Retrieval Conference's Deep Learning Track (TREC DL) 2020 were incorporated. This dataset is renowned for its challenging and diverse set of queries and relevance judgments, providing a robust standard for evaluating information retrieval systems. The TREC DL 2020 queries and topics were used to simulate a wide range of information-seeking scenarios, offering a rigorous assessment of how docTTTTTquery and SPLADE++ perform in a competitive environment.

Upon indexing the expanded documents and passages, the retrieval phase was conducted using BM25. In our experiments, BM25 served as the retrieval model, interfacing with the indexed content to fetch and rank relevant documents and passages based on the TREC DL 2020 queries.

4 Experiments

4.1 Data

The MSMARCO dataset includes both a passage dataset and a document dataset. The passage dataset consists of 8.8 million passages and 1 million queries, while the document dataset comprises 3.2 million documents and 367,013 queries. The main difference between the two datasets is the length of the text, with passages being shorter than documents. Running the MSMARCO document dataset on models such as doc2Query, doc2Query–, and Splade may produce different results compared to running the passage data due to the varying text lengths and the potential impact on the effectiveness of document expansion. For instance, the document expansion model may be more effective in capturing a wider range of relevant terms from longer documents, while the passage dataset may require a different approach due to the shorter length of the passages. According to Nogueira(2019)[6], document expansion is often more effective than query expansion, as a document typically contains more signals than a query due to its longer length. Therefore, the choice of dataset and the models used for expansion should be carefully considered based on the specific characteristics of the data.

4.2 Evaluation method

In evaluating the performance of information retrieval systems, three key metrics are commonly employed: Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), and Precision at 10 (P@10). MAP offers a comprehensive view, measuring the average precision across all relevant documents for each query, thereby focusing on the overall effectiveness of the retrieval across the entire set of queries. MRR, on the other hand, emphasizes the placement of the first relevant document by averaging the reciprocal ranks of the first relevant item for each query, thus highlighting the system's ability to retrieve a relevant document at the earliest. P@10 provides a focused metric, evaluating the precision within the top 10 search results, making it particularly relevant for scenarios where immediate relevance is paramount. These metrics, while overlapping in their assessment of retrieval effectiveness, offer distinct perspectives - MAP assessing overall retrieval quality, MRR focusing on initial result relevance, and P@10 evaluating top-tier precision.

4.3 Results

Table 2: Evaluation on MSMARCO passages dataset and documents dataset. Source code can be found at `https://github.com/WSE-Document-Expansion/Evaluation-docT5query-SPLADE`

	MAP	**MRR**	**P@10**
docTTTTTquery (passages)	0.432987	**0.887963**	0.642593
docTTTTTquery (documents)	**0.412033**	**0.87702**	**0.568889**
SPLADE++ (passages)	**0.437593**	0.876411	**0.696296**
SPLADE++ (documents)	0.280139	0.829798	0.455556

5 Analysis

Our analysis based on the MSMARCO dataset evaluation demonstrates varying strengths between docTTTTTquery and SPLADE++. In passage retrieval, SPLADE++ achieves higher MAP and P@10 scores, which indicates superiority in ranking precision and relevance within the top 10 results. However, docTTTTTquery outperforms SPLADE++ in MRR for passage retrieval, suggesting it is more effective at returning a relevant passage as the top result.

When it comes to document retrieval, docTTTTTquery surpasses SPLADE++ in all metrics: it secures a MAP, MRR, and P@10, confirming its overall efficacy in ranking relevant documents and identifying pertinent results quickly. These insights underscore the nuanced performance differences between the models, emphasizing the importance of model selection based on specific retrieval needs, whether it is the precision of the first result or consistent accuracy across the top-ranked results.

Both docTTTTTquery and SPLADE++ have demonstrated strong performance on the MSMARCO passage dataset, but their effectiveness tends to diminish when applied to the more complex document ranking tasks. This discrepancy can be attributed to several factors, including the increased length and complexity of documents compared to passages, the potential for greater vocabulary mismatch in documents, and the models' optimization for passage retrieval tasks. Additionally, the characteristics of the dataset, such as query-document relevance judgments and the distribution of relevant information within the documents, can also influence the models' performance. Therefore, while these models have proven effective in passage ranking, their performance in document ranking tasks remains an area for further exploration and improvement.

6 Future Work

Given the observed performance discrepancies between passage and document retrieval tasks, future work should focus on improving the effectiveness of these models in document ranking. This could involve exploring methods to better handle the increased length and complexity of documents, as well as addressing potential vocabulary mismatches. Additionally, further research could investigate how to optimize these models for different types of datasets, taking into account factors such as query-document relevance judgments and the distribution of relevant information within the documents. It would also be beneficial to conduct more comprehensive evaluations using different datasets and performance metrics to gain a deeper understanding of these models' strengths and weaknesses. Lastly, as the field of information retrieval continues to evolve, it will be important to explore the integration of these models with newer techniques and technologies to further enhance their performance.

7 Conclusion

Our experiment has revealed nuanced performance differences between the docTTTTTquery and SPLADE++ models in both passage and document retrieval tasks. While SPLADE++ excels in precision and relevance within the top 10 results for passage retrieval, docTTTTTquery outperforms it in Mean Reciprocal Rank (MRR), indicating its superior ability to return a relevant passage as the top result. In document retrieval, docTTTTTquery surpasses SPLADE++ in all metrics, demonstrating its overall efficacy in ranking relevant documents and identifying pertinent results quickly. However, both models show a decrease in effectiveness when transitioning from passage to document ranking tasks, highlighting the need for further research and optimization in this area.

References

[1] Rodrigo Nogueira and Jimmy Lin. From doc2query to doctttttquery. 12 2019.

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

[3] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. Ms marco: A human generated machine reading comprehension dataset, 2018.

[4] Orion Weller, Kyle Lo, David Wadden, Dawn J Lawrie, Benjamin Van Durme, Arman Cohan, and Luca Soldaini. When do generative query and document expansions fail? a comprehensive study across methods, retrievers, and datasets. *arXiv preprint arXiv:2309.08541*, 2023.

[5] Shi Yu, Chenghao Fan, Chenyan Xiong, David Jin, Zhiyuan Liu, and Zhenghao Liu. Fusion-in-t5: Unifying document ranking signals for improved information retrieval, 2023.

[6] Rodrigo Nogueira and Kyunghyun Cho. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375*, 2019.