Document Expansion by Query Prediction

Deepak Khatri, Nirmal Shah, Keyur Dhanani and Rameez Raja

Abstract

The vocabulary mismatch problem in information retrieval arises when users' query terms differ from those in relevant documents. Traditional query expansion focuses on enriching query representations but leaves documents static. This work proposes a neural-based document expansion approach, generating potential user queries and appending them to documents before indexing. The method significantly improves retrieval effectiveness, achieving state-of-the-art results on MS MARCO.

1. Problem

The vocabulary mismatch problem is a significant challenge in information retrieval, where users often use query terms that differ from those present in relevant documents. Addressing this issue involves improving the alignment between queries and document content. Traditional approaches like query expansion [1] enrich query representations but leave document representations static. An alternative approach focuses on enriching document representations before indexing by generating potential user queries that a document might address and appending them to the document [2]. This method improves retrieval effectiveness and better addresses the vocabulary mismatch problem.

2. Dataset

The MS MARCO dataset [3] is a large-scale benchmark used in information retrieval and question-answering research, consisting of around 8.8 million passages obtained from web documents. It includes approximately 1 million user queries, each paired with passages retrieved by the Bing search engine, where one passage per query is labeled as relevant. The dataset is divided into training, development, and test sets, with the training set containing around 500,000 query-passage pairs and the development and test sets each having about 6,900 queries. Relevance labels are only available for the development set. MS MARCO is precision-oriented and serves as a valuable resource for evaluating and comparing retrieval methods, particularly those focused on enhancing retrieval accuracy through techniques like neural re-ranking and document expansion.

3. Evaluation Metric

The evaluation metric used is Mean Reciprocal Rank at rank 10 (MRR@10), which measures the average rank of the first relevant document within the top 10 retrieved results. It is calculated as:

$$MRR@10 = \frac{1}{Q} \sum_{i=1}^{Q} \frac{1}{\text{rank}_i}$$

where Q is the total number of queries, and rank_i is the rank of the first relevant document for the i-th query. If no relevant document is found in the top 10, the reciprocal rank is 0. MRR@10 helps evaluate how well a retrieval system places relevant documents at the top of the ranked list.

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

202311042 (D. Khatri); 202311043 (N. Shah); 202311045 (K. Dhanani); 202311068 (R. Raja)

4. Results

We compare the performance of various retrieval models using the MRR@10 metric. In our experiments, we have applied different configurations, with a focus on enhancing document representations using query generation models. The original paper utilized a sequence-to-sequence vanilla transformer for the Doc2Query task. However, in our experiments, we explored the use of T5 and BART, two powerful transformer-based models, to generate the queries for the document expansion. The results of our experiments are presented below.

Model	MRR@10	Our MRR@10
BM25	18.4	18.74
BM25+RM3	16.7	16.46
BM25+BERT(Reranking)	37.5	34.85
Doc2Query+BM25	21.5	-
Doc2Query+BM25+RM3	20.3	-
Doc2Query+BM25+BERT(Reranking)	37.5	-
BART+BM25	19.33	19.33
BART+BM25+RM3	17.42	17.42
BART+BM25+BERT(Reranking)	35.2	35.2
T5+BM25	27.6	27.6
T5+BM25+RM3	21.5	21.5
T5+BM25+BERT(Reranking)	37.5	37.5

Table 1
Comparison of MRR@10 results with different configurations

From the table, it is clear that while traditional models such as BM25 and BART-based approaches perform reasonably well, our experiments with T5 for doc2query generation show a significant improvement in the MRR@10 metric, especially with the T5+BM25 configuration. These findings highlight the effectiveness of modern transformer models in enhancing the quality of document expansion for information retrieval tasks.

5. Key Challenges and Learning

During the course of this project, we encountered several challenges. One of the major challenges was related to computational issues, as the models we worked with were resource-intensive and required significant processing power. Additionally, we faced difficulties during the inferencing phase of the models, which required careful troubleshooting and optimization to ensure smooth execution.

On the learning side, this project provided valuable insights into the field of information retrieval. We gained an in-depth understanding of how retrieval systems work, the mechanisms of query matching, and the principles of re-ranking [4] to refine retrieved results. We explored various methods of query optimization and enhanced our knowledge of utilizing pretrained large language models (LLMs) effectively for our problem. Moreover, we developed a strong grasp of the Doc2Query concept and its practical applications in improving retrieval performance.

References

1 Hiteshwar Kumar Azad and Akshay Deepak. 2019. Query expansion techniques for information retrieval: A survey. Information Processing Management 56, 5 (September 2019), 1698–1735.

2 Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document Expansion by Query Prediction.

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. 2018. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset.

4 Rodrigo Nogueira and Kyunghyun Cho. 2020. Passage Re-ranking with BERT.