ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230385

Diversified Prior Knowledge Enhanced General Language Model for Biomedical Information Retrieval

Yizheng Huang[a] **and Jimmy X. Huang**[b;*]

[a, b]Information Retrieval and Knowledge Management Research Lab, York University

Abstract. General language models have shown success in various information retrieval (IR) tasks, but their effectiveness is limited in the biomedical domain due to the specialized and complex nature of biomedical data. However, training domain-specific models is challenging and costly due to the limited availability of annotated data. To address these issues, we propose the Diversified Prior Knowledge Enhanced General Language Model (DPK-GLM) framework, which integrates domain knowledge with general language models for improved performance in biomedical IR. Our two-stage retrieval framework comprises a Knowledge-based Query Expansion method for enriching biomedical knowledge, an Aspect-based Filter for identifying highly-relevant documents, and a Diversity-based Score Reweighting method for re-ranking retrieved documents. Experimental results on public biomedical IR datasets show significant improvement, demonstrating the effectiveness of the proposed methods.

1 Introduction

General language models have demonstrated impressive capabilities in various information retrieval (IR) tasks [30, 13]. A notable example is the Bidirectional Encoder Representations from Transformers (BERT) [3], which has emerged as a standard component for developing task-specific IR models. Existing general models predominantly focus on the web domain. For instance, the original BERT model was trained on Wikipedia and BookCorpus, and subsequent work has mainly focused on large-scale pre-training on larger texts crawled from the internet. However, the efficacy of these models in the biomedical domain is impeded by considerable challenges. Biomedical data is characterized by its specialized and intricate nature, consisting of professional terminology and domain-specific concepts that general language models cannot fully grasp. Moreover, a biomedical IR system requires capturing the relationships between a user's query intent and the concepts in biomedical documents, which poses a significant challenge for general models. As a result, training domain-specific models is considered the primary method to improve the accuracy and relevance of search results within the biomedical field.

Previous research indicates that pre-training on domain-specific text can yield advantages over general language models [20, 10, 22]. However, their training process is often difficult and costly due to the scarcity of high-quality annotated data, especially for niche subdomains or uncommon diseases. In addition, the capabilities of these specialized models are still limited by their training datasets. If the user's query concerns a rare disease, the IR system may fail to accurately retrieve high-quality results, as the disease lies outside the scope of the system's learning. Therefore, a feasible alternative is to choose a cheaper but effective strategy, combining domain knowledge with general language models to enhance comprehension of biomedical data.

To achieve this purpose, two challenges need to be addressed in the biomedical IR system: diversity and accuracy. Consider a biomedical scientist searching the literature with a query such as "What is the role of PrnP in mad cow disease?". Ideally, the IR system should locate content that shares aspects with the query in documents, including related topics, such as "PrnP" and "mad cow disease". In reality, however, the search may more likely retrieve documents where the subjects partially align with the query aspects (e.g., the same gene but a different disease). Such documents could still be relevant if the matched aspects are deemed more critical than the unmatched ones, as the scientist judges. In these scenarios, the relevance judgment criteria can be characterized as diversity, meaning the IR system should return documents featuring a diverse range of entities covering various query-related aspects, such as genes, proteins, diseases, and mutations. Diversity measures whether the retrieved documents offer a comprehensive overview of the topic.

Additionally, the accuracy of the returned documents is vital, as the user aims to extract all highly-relevant documents. The primary issue with accuracy lies in the biomedical domain's unique terminology (e.g., "PrnP", the prion protein), which exhibits a considerable degree of lexical variation and ambiguity (e.g., "CD230" is synonymous with "PrnP", cluster of differentiation 230). Consequently, the ability to accurately capture biomedical terminology is essential for biomedical IR systems.

Employing prior knowledge (or external knowledge) has proven advantageous in addressing the aforementioned challenges. Several studies have investigated the incorporation of prior knowledge sources, such as biomedical ontologies, databases, and knowledge graphs, to enhance performance in biomedical IR systems [29, 11]. By introducing domain-specific knowledge, like the Medical Subject Headings (MeSH) and the Unified Medical Language System (UMLS), biomedical ontologies can enhance the accuracy and coverage of terminology recognition and relation extraction. Leveraging PubMed as a knowledge source can aid in retrieving relevant documents and enable the exploration of related aspects. Furthermore, using knowledge graph-based methods allows for capturing complex relationships between biomedical concepts. Their work demonstrates that integrating prior knowledge can significantly improve the performance of biomedical IR systems.

* Corresponding Author. Email: jhuang@yorku.ca

In this paper, we propose a framework called Diversified Prior Knowledge Enhanced General Language Model (DPK-GLM) as a cost-effective approach for merging domain knowledge with general language models to improve their performance in biomedical IR. Our framework consists of a two-stage retrieval framework with three key components: a Knowledge-based Query Expansion method to enrich biomedical knowledge, an Aspect-based Filter for identifying highly-relevant documents, and a Diversity-based Score Reweighting method for re-ranking retrieved documents. Our experimental design adopts two pre-trained general language models, BERT and RoBERTa [21], as baseline models. For comparative purposes, we also employ two pre-trained domain-specific language models, namely BioBERT [20] and ClinicalBERT [10]. The results derived from experiments conducted on publicly accessible biomedical IR datasets, in conjunction with an ablation study, manifest significant performance enhancements attributable to our proposed approaches.

2 Related Work

This section examines related works in two areas: pre-trained language models and biomedical information retrieval.

2.1 Pre-trained Language Models

Recent years have witnessed significant progress in pre-trained language models (PLMs) for information retrieval. General models such as BERT and T5 [24] have achieved impressive performance in various IR tasks, including recommendation, query generation, and document ranking. Several studies have explored the effectiveness of PLMs in biomedical domain tasks [2], such as named entity recognition (NER), relation extraction (RE), and question answering (QA). For instance, BioBERT, a domain-specific PLM tailored for biomedical text, exhibits state-of-the-art results in multiple biomedical text mining tasks (e.g., disease NER) and serves as a popular backbone in a wide range of biomedical IR tasks.

Considering that general language models are not optimized for biomedical data, training domain-specific models is a sensible choice. However, the performance of these models is also constrained by the scope of their training data. Lisa et al. [19] observed that publicly available domain-specific models such as BioBERT experience a significant performance decline when evaluated on a newly annotated COVID-19 preprint dataset. Ji et al. [15] revealed that ClinicalBERT performed worse than classic BM25 [26] on the National Center for Biotechnology Information (NCBI) disease corpus for the biomedical entity normalization task. Moreover, Wei et al. [28] developed an ensemble approach, combining convolutional neural networks (CNN) with long short-term memory (LSTM) networks to manage semantic syntax features, attaining better results compared to transformer models for the bio-concept disambiguation task. These studies demonstrate that tasks in the biomedical domain are considerably more complex than those involving general domain knowledge.

2.2 Biomedical Information Retrieval

Biomedical IR has traditionally relied on term-matching algorithms such as TF-IDF and BM25, which search for documents containing terms mentioned in the query. However, these methods struggle with biomedical terminology variation [14, 23]. To address this issue, several studies have explored the use of domain-specific knowledge bases to enhance biomedical IR systems. Koopman et al. [18] proposed a graph inference model that obtained domain knowledge from SNOMED CT to tackle the semantic gap problem. Goodwin et al. [4] utilized multiple knowledge bases, such as MeSH and UMLS, to build a unified knowledge graph for topic analysis and expansion. Jin et al. [16] expanded queries using a list of weighted synonyms extracted from the National Library of Medicine (NLM) API to achieve high recall in baseline retrieval.

Other studies have focused on different strategies. Rybinski et al. [17] employed the Divergence from Randomness (DFR) method to boost performance in the initial ranking step for the biomedical literature search. Soldaini et al. [27] proposed a convolutional neural model to reduce clinical notes' noise for medical literature retrieval. KERS [1] was designed as an article recommendation system to support decision-making in medical treatments for cancer patients.

All the approaches mentioned above depend on fine-tuning or retraining pre-trained models on domain-specific data, which can be expensive and time-consuming. In contrast, our framework offers a cheaper yet efficient alternative that can be easily applied to existing general language models to enhance their performance in biomedical IR.

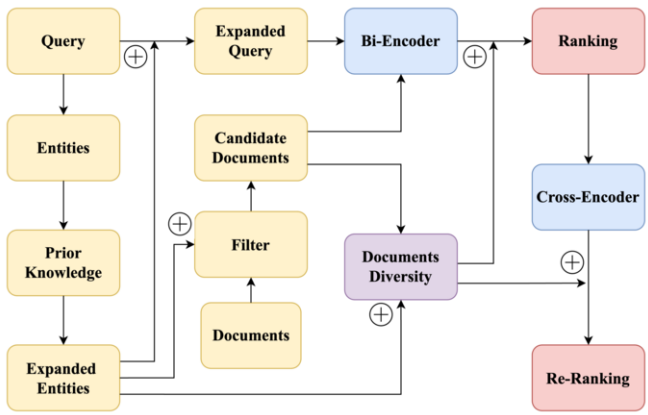


Figure 1: The architecture of our proposed framework.

3 Methodology

Our proposed DPK-GLM framework is a two-stage retrieval framework consisting of three components: a Knowledge-based Query Expansion method, an Aspect-based Filter, and a Diversity-based Score Reweighting method, as shown in Figure 1. The following sections will introduce the details of each component.

Table 1: Examples of the extracted entities.

Query	Entities
What is the role of *PrnP* in *mad cow disease*?	PrnP, mad cow disease
What is the role of *IDE* in *Alzheimer's disease*?	IDE, Alzheimer's disease
Which [GENES] involved in *NFkappaB* signaling regulate *iNOS*?	NFkappaB, iNOS

3.1 Knowledge-based Query Expansion

The diversity of search results in biomedical IR is characterized by the range of query-related aspects covered in the output ranking list. It ensures that the retrieved documents provide a comprehensive overview of the query and meet the user's information needs. Prior knowledge from respected sources such as MeSH, UMLS, and NCBI is being incorporated to enhance the diversity of query aspects.

A query q is a series of terms $q = \{t_1, e_2, ..., t_6, e_7, ..., t_n\}$, where n represents the number of terms, and a term related to biomedical

aspects is referred to as an entity e. Including more diversified entities in a query implies a higher level of diversity and broader aspect coverage, which can help users find all potentially relevant information. We utilize SpaCy [9] to extract entities $E = \{e_2, ..., e_7, ...\}$ from the query. Table 1 shows examples of the extracted entities. The extracted entities are then expanded with their synonyms and descriptions $E' = \{e'_2, e''_2, ..., e'_7, e''_7, ...\}$ from prior knowledge sources. For instance, in a query "What is the role of PrnP in mad cow disease?", the description entity of "PrnP" is prion protein, and its synonym entities, such as "ASCR", "AltPrP", "CD230", can be found in MeSH and NCBI Gene databases. As for the diversified entities of mad cow disease, it encompasses various other information: Bovine spongiform encephalopathy (BSE), neurodegenerative disease, variant Creutzfeldt-Jakob disease (vCJD), etc. These aspect-related entities are not included in the original query but are highly relevant to the query.

While incorporating prior knowledge can alleviate the challenges of lexical variant and diversified aspects problems, the biomedical domain faces an additional issue of multiple out-of-vocabulary terminology representations. This situation can be represented as $E'_v = \{e_2^{v_1}, e_2^{v_2}, ..., e_7^{v_1}, e_7^{v_2}, ...\}$, where v indicates different representations. For example, due to varying writing habits among researchers, the entity "TGF-beta1" can be represented as "TGF-betaI" and "TGF-β1".

Inspired by [12], Break-point and Replacement methods were implemented for further query expansion. Break-point indicates a specific location in a string where the space can split the string into two parts. For example, the entity "TGF-beta1" with two break-points can be transformed to "TGF-beta 1" and "TGFbeta-1". On the other hand, Replacement refers to a substring within a string that can be swapped with another string while preserving the semantic meaning of the original expression. For instance, the entity "TGF-beta1" with the number "1" can be substituted with "TGF-betaI".

In this way, the expanded query is the union of the original entities with all its extended diverse aspects, including synonyms, descriptions, and various terminology representations. The output can be formulated below:

$$q_{exp} = q \cup E' \cup E'_v = \{t_1, e_2, e'_2, ..., e_2^{v_1}, ..., e_7, e'_7, ..., e_7^{v_1}, ..., t_n\}$$

3.2 Aspect-based Filter

General language models trained on large-scale datasets are often biased towards the training domain for optimal performance. Consequently, achieving high-quality ranking results in the biomedical domain without fine-tuning or retraining can be challenging. Intuitively, if enhancing the performance of the general model proves difficult, filtering out irrelevant documents can be beneficial, as the remaining documents are more likely to be relevant. Furthermore, the reduced range of candidate documents results in lower computational and time costs, making it a practical and efficient solution.

Leveraging prior knowledge, the expanded query encompasses all highly-related aspects of the original query. Based on this point, we removed all documents devoid of any aspects and acquired a smaller set of candidate documents, which are more likely to be relevant to the query. However, this approach carries risks, as it is possible that some documents may contain valuable information that is not explicitly stated and could be mistakenly filtered out. In this case, a reasonable guess is that the retrieval results will be negatively affected due to the absence of some relevant documents. To verify the hypothesis, we conduct experiments to determine whether this concern is necessary or not, and the details of the experiments can be found in Section 4.

3.3 Two-stage Ranking

In a two-stage IR system, the initial ranking plays a critical role, as the performance of the final result heavily depends on it. Many works concern efficiency, using traditional retrieval algorithms such as BM25 to obtain initial ranking results, and then applying fine-tuned pre-trained language models for re-ranking to improve accuracy [16, 5]. Since the Aspect-based Filter can narrow the scope of documents and enhance retrieval efficiency, we utilize the general language model in the initial ranking to maximize its capabilities.

The general language model generates embeddings for the query and the document, and the ranking results of the IR system are obtained by computing the similarity between their embeddings. Intuitively, we expect to utilize the semantic understanding ability of the language model to strengthen retrieval accuracy. For this purpose, our two-stage ranking approach employs two encoder types: Bi-encoder and Cross-encoder [25].

The Bi-encoder is designed to encode the query and document independently, allowing for pre-computation and caching of document features. Its high efficiency makes it suitable for the initial ranking task. In contrast, the Cross-encoder takes a question-answer pair as input, passing both the query and document simultaneously to the neural network and leveraging cross-attention to yield better results. While this model attains higher retrieval accuracy, it is less efficient and is primarily designed for re-ranking tasks.

To strike a balance between effectiveness and efficiency, our two-stage ranking approach leverages the strengths of these two models. The Bi-encoder is employed in the initial ranking stage. As the number of initially retrieved documents is much smaller than the original document set, the efficiency of the Cross-encoder is boosted for the re-ranking stage.

3.4 Diversity-based Score Reweighting

The similarity score of the query and document embeddings determines the ranking results of a language model-based IR system. However, the similarity score is not always accurate, as it is influenced by the quality of the embeddings, particularly when using general language models in a specific domain. We propose a Diversity-based Score Reweighting method to address this issue to adjust the ranking results.

Diversity represents the degree to which a document covers different aspects of a query. The expanded query encompasses all the diversified aspects of the original query, and the documents that cover more aspects are more likely to be relevant. The occurrence of multiple different entities in a document indicates that it covers several aspects related to the query topic, which is used to compute the diversity score.

The proposed Diversity-based Score Reweighting method linearly combines the similarity and diversity scores to improve results by balancing their weights. After each ranking stage, the ranking results are re-sorted based on this rule.

We denote diversity as $\mathcal{V}$, representing the number of entities that exist in a document, and $\mathcal{S}$ as the similarity score of a document with respect to a given query. The reweighting process, $\mathcal{S}_{re}$, can be formulated as follows:

$$\mathcal{S}_{re} = \alpha\zeta \cdot \mathcal{V} + (1 - \alpha)\mathcal{S}$$

where α is a hyperparameter that controls the balance between diversity and similarity score, and ζ is a hyperparameter designed to control the weight of diversity.

4 Experiments

In this section, we present our experimental studies. Section 4.1 introduces the datasets and evaluation metrics employed in the experiments. Section 4.2 outlines the experimental settings and baselines, while Section 4.3 discusses the results of the proposed methods. Lastly, Section 4.4 features an ablation study to analyze the effectiveness of each component of the proposed methods.

4.1 Datasets and Evaluation Metrics

Datasets We conducted experiments on public biomedical IR datasets: TREC 2006&2007 Genomics Track (TREC-GENO) [8, 6]. The document collection of TREC-GENO comprises a full-text biomedical corpus containing 162,259 documents from 49 genomics-related journals indexed by MEDLINE. A total of 64 official topics from the biomedical domain were used as queries, with 28 topics being specific and 36 topics being abstract. These official topics were manually created by biomedical domain experts and formatted in question-answering style. The following are examples of the queries:

- **Specific Query**: "What is the role of IDE in Alzheimer's disease?"
- **Specific Query**: "How does Nurr-77 delete T cells before they migrate to the spleen or lymph nodes and how does this impact autoimmunity?"
- **Abstract Query**: "What [BIOLOGICAL SUBSTANCES] have been used to measure toxicity in response to zoledronic acid?"
- **Abstract Query**: "What [CELL OR TISSUE TYPES] express receptor binding sites for vasoactive intestinal peptide (VIP) on their cell surface?"

There are three main reasons for choosing TREC-GENO as the dataset for our experiments. First, the dataset lacks annotations and has only 63 queries with corresponding ground truth (golden standard), making it difficult to improve the performance of pre-trained language models by fine-tuning on this dataset. Second, there are two different query types in the dataset. Their differences significantly impact our proposed methods, and we will discuss this in detail in the following sections. Third, the dataset has official assessment metrics, which are used to evaluate aspect-level performance.

Evaluation Metrics There are four official evaluation metrics, all of which are variants of mean average precision (MAP):

- **Document MAP**: This metric calculates the average of precision values obtained after retrieving each relevant document.
- **Aspect MAP**: This metric aims to evaluate retrieval performance in terms of the diversity of search results.
- **Passage MAP**: This metric measures individual precision scores at the passage level, where a passage can be considered as a paragraph in a document.
- **Passage2 MAP**: This is an alternative Passage MAP that uses a different passage segmentation method to identify the shortest relevant passage.

For easy distinction, we consider the TREC-GENO as two subtask sets based on query types, namely Specific and Abstract. In addition, we introduce NDCG as an additional evaluation criterion to verify our methods from multiple perspectives. Statistical significances are tested by the two-tailed t-test with a significance level of 0.05.

4.2 Experimental Settings and Baselines

Experimental Settings In our two-stage IR framework, we employ two general language models, BERT and RoBERTa, as the Bi-encoder for the initial ranking, and SentenceBERT (SBERT) [25] as the re-ranker model since it has a pre-trained Cross-encoder model. To investigate whether language models pre-trained on domain-specific data without fine-tuning can yield better results on a new dataset compared to general language models, we select BioBERT and ClinicalBERT as additional Bi-encoder models for comparison. For the initial search, we extract the top 2000 documents, while in the re-ranking process, we only need the top 1000 as the final result. In addition, we set the hyperparameters α to 0.2 and ζ to 0.5 for comparative experiments, as our framework performs best under this setting.

Baselines To explore the performance of the language-model-based retrieval system without fine-tuning compared to traditional retrieval methods, we adopt the official TREC-GENO runs as the extra baselines [7]. These official runs include the Min, Median, Mean, and Max results, and our comparison experiments exclude their Min results.

4.3 Results and Analysis

Table 2 presents the results of our approach under the official evaluation metrics. Our approach achieves the best results on both the TREC-GENO Specific and the TREC-GENO Abstract. Without fine-tuning, neither general language models (BERT and RoBERTa) nor domain-specific models (BioBERT and ClinicalBERT) can deliver good results on the new dataset. Their performance is even worse than traditional information retrieval methods. Pre-trained language models tend to overfit to their training domain, resulting in decreased performance on downstream tasks when the target domain significantly differs. This also demonstrates that even BioBERT and ClinicalBERT, pre-trained on biomedical data, struggle with tasks beyond their training data due to the complexity of the biomedical domain. Currently, no large-scale annotated biomedical datasets are available to train a highly generalizable language model, which confirms our approach's feasibility.

The proposed framework demonstrates significant improvements in both the Document MAP and the Aspect MAP, aligning with our expectations. Our proposed methods primarily focus on enhancing document relevance and aspect diversity of the queries. Simultaneously, improvements in document matching also benefit passage matching. It is worth noting that our framework shows a substantial improvement in Passage-level MAP compared to the baseline model but still falls short of the best traditional retrieval method. This gap is acceptable since we did not specifically optimize passage matching in this paper. On the other hand, all two-stage search results outperform their single-stage counterparts only marginally. This suggests that although the two-stage approach can provide improvements, the overall effectiveness is limited by the accuracy of the initial ranking produced by the language models. However, the performance of different language models in the initial ranking varies considerably. We can see that BERT, as a general language model, outperforms the

Table 2: Experiment results of our DPK-GLM and baselines on the TREC-GENO Specific & Abstract under the official metrics. The superscript "*" means the method is significantly better than the best baseline.

	TREC-GENO Specific				TREC-GENO Abstract			
Methods	Document	Aspect	Passage	Passage2	Document	Aspect	Passage	Passage2
TREC Median	0.3083	0.1581	0.0316	0.0345	0.1897	0.1311	0.0565	0.0377
TREC Mean	0.2887	0.1643	0.0347	0.0392	0.1862	0.1326	0.0560	0.0398
TREC Max	0.5439	0.4411	0.1012	0.1486	0.3286	0.2631	0.0976	0.1148
BioBERT	0.2832	0.1014	0.0473	0.0515	0.2368	0.1326	0.0571	0.0506
BioBERT+SBERT	0.2954	0.1189	0.0510	0.0523	0.2398	0.1393	0.0602	0.0611
ClinicalBERT	0.2584	0.0888	0.0408	0.0452	0.2034	0.0968	0.0469	0.0458
ClinicalBERT+SBERT	0.2662	0.0971	0.0459	0.0482	0.2114	0.1006	0.0502	0.0491
BERT	0.2903	0.1012	0.0491	0.0521	0.2083	0.1172	0.0647	0.0595
BERT+SBERT	0.3011	0.1008	0.0511	0.0533	0.2144	0.1252	0.0690	0.0604
DPK-GLM-BERT+SBERT	**0.5771***	**0.4702***	0.0874	0.1081	**0.4551***	**0.4278***	0.0755	0.0858
RoBERTa	0.2953	0.1134	0.0504	0.0566	0.2159	0.1224	0.0597	0.0433
RoBERTa+SBERT	0.3078	0.1242	0.0545	0.0596	0.2212	0.1338	0.0625	0.0656
DPK-GLM-RoBERTa+SBERT	**0.5854***	**0.4733***	0.0882	0.1089	**0.4573***	**0.4356***	0.0761	0.0863

Table 3: Experiment results of our DPK-GLM and baselines on the TREC-GENO Specific & Abstract under the NDCG metrics. The superscript "*" means the method is significantly better than the best baseline.

	TREC-GENO Specific			TREC-GENO Abstract		
Methods	NDCG@5	NDCG@10	NDCG@20	NDCG@5	NDCG@10	NDCG@20
BioBERT	0.1804	0.1765	0.1733	0.1474	0.1707	0.1913
BioBERT+SBERT	0.1881	0.1794	0.1763	0.1518	0.1816	0.2002
ClinicalBERT	0.1380	0.1437	0.1326	0.1269	0.1310	0.1280
ClinicalBERT+SBERT	0.1466	0.1482	0.1415	0.1379	0.1411	0.1376
BERT	0.1037	0.1234	0.1216	0.1461	0.1339	0.1293
BERT+SBERT	0.1110	0.1352	0.1288	0.1550	0.1424	0.1353
DPK-GLM-BERT+SBERT	**0.3241***	**0.3163***	**0.3115***	**0.3365***	**0.3232***	**0.3151***
RoBERTa	0.1081	0.1201	0.1255	0.1433	0.1667	0.1338
RoBERTa+SBERT	0.1205	0.1321	0.1355	0.1520	0.1742	0.1405
DPK-GLM-RoBERTa+SBERT	**0.3415***	**0.3358***	**0.3228***	**0.3635***	**0.3840***	**0.3882***

domain-specific model ClinicalBERT in all metrics in Table 2. Language model-based IR systems heavily rely on the quality of generated embeddings, and the strength of the semantic understanding ability determines the performance of the ranking results. Our experiments show that these models struggle with the domain transfer problem for biomedical IR tasks. Nevertheless, our DPK-GLM framework successfully mitigates this issue without requiring fine-tuning.

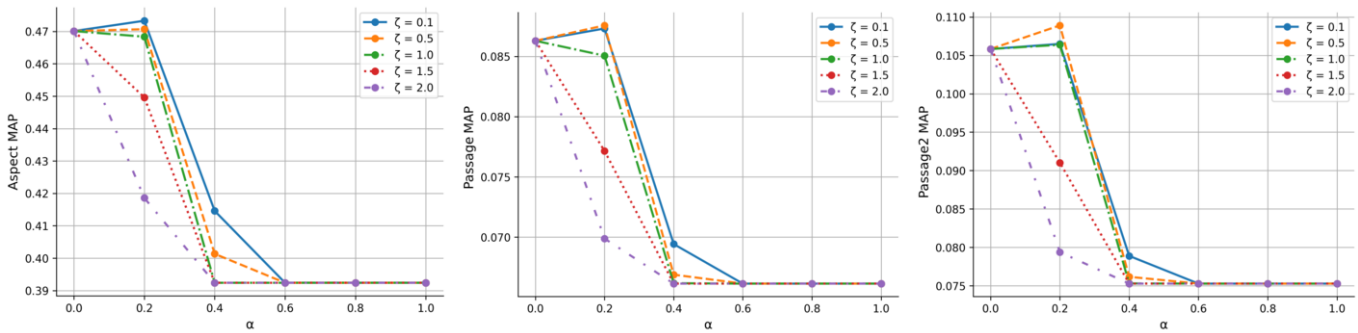
Table 3 presents the performance of our framework in terms of the NDCG metric. Similar to Table 2, our framework significantly outperforms the baselines. Notably, our approach achieves better performance on the TREC-GENO Abstract compared to the TREC-GENO Specific under the NDCG metric. In addition, as shown in Table 2, our best approach, DPK-GLM-RoBERTa+SBERT, achieves a remarkably higher Aspect MAP on the Abstract task than the best traditional retrieval method. However, this phenomenon is not observed in the Specific task. It suggests that pre-trained language models are more likely to capture abstract semantic information but struggle with understanding specific terms that they have not learned before. Even though the model's training data may not include biomedical terms, such as the gene "TGF-beta1", the model can easily learn common terms like GENE, MUTATION, and CELL. This learning ability is well-reflected in the TREC-GENO Abstract.

Through Table 5, we can address the earlier hypothesis: Will filtering out some valuable documents affect the results? By comparing the relevant documents in the filtered candidate documents with the ground truth, we can see that only a small portion of relevant documents were missed after filtering, indicating that our Filter performs well. In addition, the quality of candidate documents in the Abstract is lower than that in the Specific, which is attributed to the difficulty in extracting relevant entities from an abstract query. Nevertheless, when considering the experimental results of Table 2 and Table 3, even with the absence of a small number of relevant documents, our method can still achieve good results.

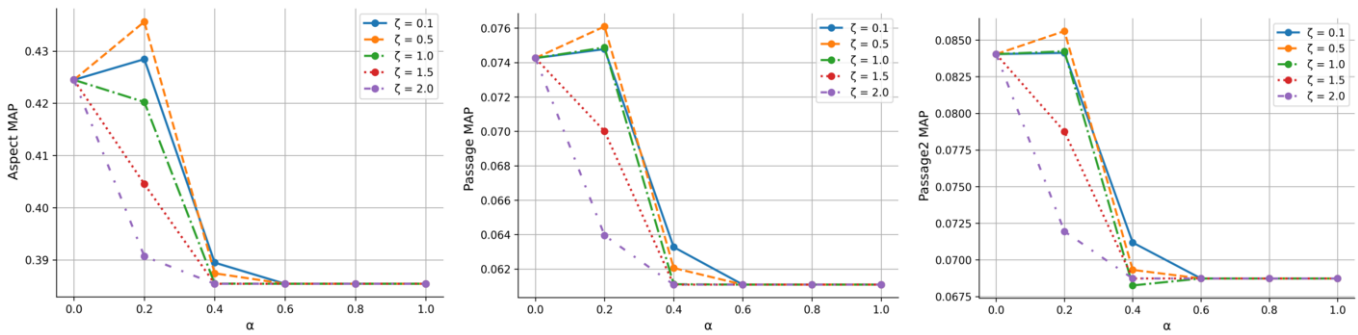
Figure 2 shows the experiment ranking results under different hyperparameters. We did not test the performance variation of the Document MAP because it measures whether the retrieval results contain relevant documents, regardless of their rank. A high diversity score means the search target has a high potential of relevance to the query, which means the semantic meanings captured by the general learning model are more relevant to the query. However, when α becomes large, the effectiveness of the ranking score decreases due to the overweighting of diversity. Only focusing on diversity will lead to poor results. Striking a balance between diversity and the semantic ranking score is crucial for optimal performance.

4.4 Ablation Study

To further investigate the effectiveness of our approach, we conducted an ablation study on the TREC-GENO dataset. We conducted a comparative analysis by individually removing the Knowledge-based Query Expansion method, the Aspect-based Filter, and the



(a) Aspect MAP in Specific (b) Passage MAP in Specific (c) Passage2 MAP in Specific



(d) Aspect MAP in Abstract (e) Passage MAP in Abstract (f) Passage2 MAP in Abstract

Figure 2: The performance of Re-ranking under different α and ζ of DPK-GLM-RoBERTa.

Table 4: The Ablation Study of the DPK-GLM framework on the TREC-GENO Specific & Abstract tasks under the official evaluation metrics.

	TREC-GENO Specific				TREC-GENO Abstract			
Methods	Document	Aspect	Passage	Passage2	Document	Aspect	Passage	Passage2
DPK-GLM-BERT+SBERT	0.5771	0.4702	0.0874	0.1081	0.4551	0.4278	0.0755	0.0858
DPK-GLM-BERT+SBERT w/o QE	0.5844	0.4811	0.0905	0.1101	0.4808	0.4435	0.0814	0.0890
DPK-GLM-BERT+SBERT w/o Filter	0.3305	0.1214	0.0632	0.0603	0.2458	0.1428	0.0712	0.0688
DPK-GLM-BERT+SBERT w/o Div	0.5698	0.4556	0.0858	0.1054	0.4503	0.4155	0.0741	0.0826
DPK-GLM-RoBERTa+SBERT	0.5854	0.4733	0.0882	0.1089	0.4573	0.4356	0.0761	0.0863
DPK-GLM-RoBERTa+SBERT w/o QE	0.5922	0.4845	0.0922	0.0945	0.4714	0.4480	0.0820	0.0901
DPK-GLM-RoBERTa+SBERT w/o Filter	0.3290	0.1455	0.0611	0.0650	0.2441	0.1582	0.0702	0.0686
DPK-GLM-RoBERTa+SBERT w/o Div	0.5774	0.4700	0.0861	0.1058	0.4495	0.4249	0.0743	0.0842

Table 5: Comparison between the candidate relevant documents screened by the Aspect-based Filter and the ground truth relevant documents in the Gold Standard.

TREC-GENO	Gold Standard	Filter	Accuracy
Specific	997	973	**0.9759**
Abstract	2490	2323	**0.9329**

Diversity-based Score Reweighting method from our framework, and juxtaposed the results with those obtained from the complete framework.

Table 4 shows the results of the ablation study. The results reveal that the Knowledge-based Filter plays a crucial role in the ranking framework. Without the Knowledge-based Filter, the performance of our approach drops notably. This observation provides an alternative perspective on why filtering out some valuable documents does not significantly impact search results negatively. More than the filtered documents, the limited ability of general language models to capture domain-specific semantics poses the most practical challenge to search efficiency, leading to document retrieval failures. Our framework leverages prior knowledge to narrow down the text scope and increase the likelihood of finding relevant documents. As demonstrated by the experimental results, our method substantially mitigates the insufficient semantic understanding abilities of general language models in the biomedical domain.

An interesting observation from our ablation study is that our method showed improved performance when the Knowledge-based Query Expansion was removed. It suggests that Query Expansion might have a negative impact on the IR system's performance. This finding aligns with the phenomenon observed in the TREC-GENO Abstract, where an abstract query is more conducive to retrieval systems finding relevant documents. Entities from prior knowledge bases may not have appeared in the training data of the general language model, potentially leading to the model misunderstanding the query. We refer to this phenomenon as "topic redundancy". An expanded query containing too many overly specific topics may result in inaccurate search results or a small number of returned results. Therefore, when constructing a query, it is crucial to avoid overly specific topics and instead select broader topics to obtain more comprehensive and accurate results.

5 Conclusions and Future Work

In this paper, we present DPK-GLM, a novel two-stage ranking framework for general language models in biomedical IR. Our approach initially utilizes a Knowledge-based Query Expansion method to enrich the queries with domain-specific entities extracted from prior knowledge bases. Next, we employ an Aspect-based Filter to remove irrelevant documents, increasing the possibility of finding relevant ones. Finally, we propose a Diversity-based Score Reweighting method to re-sort the original ranking results by combining diversity scores with similarity scores. We evaluate DPK-GLM using BERT and RoBERTa on the public biomedical IR dataset, achieving remarkable performance and demonstrating the framework's effectiveness in improving biomedical IR performance. In the future, we plan to undertake a comprehensive study and analysis of the phenomenon of "topic redundancy". We also plan to evaluate it on more medical datasets, such as TREC-PM and TREC-COVID.

Acknowledgements

We sincerely appreciate all reviewers' comments, which helped improve this paper considerably. This research is supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada and the York Research Chairs (YRC) program.

References

[1] Mohammad Mehdi Afsar, Trafford Crump, and Behrouz Far, 'An exploration on-demand article recommender system for cancer patients information provisioning', in *The International FLAIRS Conference Proceedings*, volume 34, (2021).

[2] Souradip Chakraborty, Ekaba Bisong, Shweta Bhatt, Thomas Wagner, Riley Elliott, and Francesco Mosconi, 'Biomedbert: A pre-trained biomedical language model for qa and ir', in *Proceedings of the 28th International Conference on Computational Linguistics*, (2020).

[3] 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*, pp. 4171–4186. Association for Computational Linguistics, (2019).

[4] Travis R Goodwin, Michael A Skinner, and Sanda M Harabagiu, 'Utd hltri at TREC 2017: Precision medicine track.', in *Proceedings of The 26th Text REtrieval Conference (TREC 2017)*, (2017).

[5] Jiafeng Guo, Yixing Fan, Liang Pang, Liu Yang, Qingyao Ai, Hamed Zamani, Chen Wu, W Bruce Croft, and Xueqi Cheng, 'A deep look into neural ranking models for information retrieval', *Information Processing & Management*, **57**(6), 102067, (2020).

[6] William Hersh, Aaron Cohen, Lynn Ruslen, and Phoebe Roberts, 'TREC 2007 genomics track overview', in *Proceedings of The Sixteenth Text REtrieval Conference (TREC 2007)*, (2007).

[7] William Hersh and Ellen Voorhees, 'TREC genomics special issue overview', *Information Retrieval*, **12**, 1–15, (2009).

[8] William R Hersh, Aaron M Cohen, Phoebe M Roberts, and Hari Krishna Rekapalli, 'TREC 2006 genomics track overview.', in *Proceedings of The Fifteenth Text REtrieval Conference (TREC 2006)*, (2006).

[9] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, et al. spacy: Industrial-strength natural language processing in python, 2020.

[10] Kexin Huang, Jaan Altosaar, and R. Ranganath, 'Clinicalbert: Modeling clinical notes and predicting hospital readmission', *ARXIV.ORG*, (2019).

[11] Xiangji Huang and Qinmin Hu, 'A bayesian learning approach to promoting diversity in ranking for biomedical information retrieval', in *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009*, pp. 307–314. ACM, (2009).

[12] Xiangji Huang, Ming Zhong, and Luo Si, 'York University at TREC 2005: Genomics track', in *Proceedings of The Fourteenth Text REtrieval Conference (TREC 2005)*, (2005).

[13] Yizheng Huang and Jimmy Huang, 'York university at TREC 2021: Deep Learning Track', in *Proceedings of The 30th Text REtrieval Conference (TREC 2021)*, (2021).

[14] Nicholas C Ide, Russell F Loane, and Dina Demner-Fushman, 'Essie: a concept-based search engine for structured biomedical text', *Journal of the American Medical Informatics Association*, **14**(3), 253–263, (2007).

[15] Zongcheng Ji, Qiang Wei, and Hua Xu, 'Bert-based ranking for biomedical entity normalization', *AMIA Summits on Translational Science Proceedings*, **2020**, 269, (2020).

[16] Qiao Jin, Chuanqi Tan, Mosha Chen, Ming Yan, Songfang Huang, Ningyu Zhang, and Xiaozhong Liu, 'Aliababa damo academy at TREC precision medicine 2020: State-of-the-art evidence retriever for precision medicine with expert-in-the-loop active learning.', in *Proceedings of The 29th Text REtrieval Conference (TREC 2020)*, (2020).

[17] Maciej Rybinski Sarvnaz Karimi, 'Csiromed at TREC precision medicine 2020', in *Proceedings of The 30th Text REtrieval Conference (TREC 2021)*, (2021).

[18] Bevan Koopman, Guido Zuccon, Peter Bruza, Laurianne Sitbon, and Michael Lawley, 'Information retrieval as semantic inference: A graph inference model applied to medical search', *Information Retrieval Journal*, **19**, 6–37, (2016).

[19] Lisa Kühnel and Juliane Fluck, 'We are not ready yet: limitations of state-of-the-art disease named entity recognizers', *Journal of Biomedical Semantics*, **13**(1), 26, (2022).

[20] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang, 'Biobert: a pre-trained biomedical language representation model for biomedical text mining', *Bioinformatics*, **36**(4), 1234–1240, (2020).

[21] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov, 'Roberta: A robustly optimized bert pretraining approach', *ARXIV.ORG*, (2019).

[22] Man Luo, Arindam Mitra, Tejas Gokhale, and Chitta Baral, 'Improving biomedical information retrieval with neural retrievers', in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 11038–11046, (2022).

[23] Sérgio Matos, Joel P Arrais, Joao Maia-Rodrigues, and José Luis Oliveira, 'Concept-based query expansion for retrieving gene related publications from medline', *BMC bioinformatics*, **11**, 1–9, (2010).

[24] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu, 'Exploring the limits of transfer learning with a unified text-to-text transformer', *The Journal of Machine Learning Research*, **21**(1), 5485–5551, (2020).

[25] Nils Reimers and Iryna Gurevych, 'Sentence-bert: Sentence embeddings using siamese bert-networks', in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, (2019).

[26] Stephen E. Robertson, Steve Walker, Micheline Hancock-Beaulieu, Mike Gatford, and A. Payne, 'Okapi at TREC-4', in *Proceedings of The Fourth Text REtrieval Conference, TREC 1995, Gaithersburg, Maryland, USA, November 1-3, 1995*, volume 500-236 of *NIST Special Publication*. National Institute of Standards and Technology, (1995).

[27] Luca Soldaini, Andrew Yates, and Nazli Goharian, 'Denoising clinical notes for medical literature retrieval with convolutional neural model', in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 2307–2310, (2017).

[28] Chih-Hsuan Wei, Kyubum Lee, Robert Leaman, and Zhiyong Lu, 'Biomedical mention disambiguation using a deep learning approach', in *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pp. 307–313, (2019).

[29] Xiaoshi Yin, Jimmy Xiangji Huang, Xiaofeng Zhou, and Zhoujun Li, 'A survival modeling approach to biomedical search result diversification using wikipedia', in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pp. 901–902, (2010).

[30] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma, 'An analysis of bert in document ranking', in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1941–1944, (2020).