ENHANCING GENERAL LANGUAGE MODELS FOR BIOMEDICAL TEXT RETRIEVAL VIA DIVERSIFIED PRIOR KNOWLEDGEL

YIZHENG HUANG

A THESIS SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

MASTERS OF ARTS

GRADUATE PROGRAM IN INFORMATION SYSTEMS AND TECHNOLOGY
YORK UNIVERSITY
TORONTO, ONTARIO

AUGUST 2023

Abstract

General language models have become instrumental in various information retrieval (IR) tasks but often falter when applied to the specialized and intricate nature of biomedical data. The complexity of biomedical terminology, the need for precise matching, and the limited availability of annotated data make training domain-specific models both challenging and costly. Addressing these unique challenges requires an innovative approach.

In this thesis, we introduce the Diversified Prior Knowledge Enhanced General Language Model (DPK-GLM) framework, a novel solution designed to bridge the gap between general language models and the specific demands of biomedical IR. The DPK-GLM framework integrates domain knowledge into general language models, enriching their understanding of biomedical information and thereby enhancing their performance in this specialized domain.

The DPK-GLM framework comprises three main components that synergistically work together to enhance the retrieval process. The first component is the Knowledge-based Query Expansion method, which leverages authoritative biomedical databases to infuse queries with domain-specific entities and knowledge. This expansion allows the model to recognize and respond to a wider range of biomedical concepts, providing more relevant results. The second component of the framework is the Aspect-based Filter. This method acts as a precision tool, filtering through documents to identify those that are highly relevant to the query's diversified aspects. The third and final component is the Diversity-based Score Reweighting method. Building on the work of the previous two components, this method re-ranks the

filtered documents, combining similarity and diversity scores to achieve a balanced and comprehensive ranking.

Experimental evaluation of the DPK-GLM framework on public biomedical IR datasets reveals a marked improvement in retrieval performance. The results validate the framework's ability to effectively handle the complex landscape of biomedical information, providing more accurate and contextually rich responses to queries.

Acknowledgements

I would like first to express my profound gratitude to my supervisor, Professor Jimmy Huang, whose guidance, support, and expertise have been instrumental in shaping this thesis. His relentless pursuit of excellence and insightful feedback has helped me to explore new avenues of thought and reach greater academic heights. His dedication to my development has truly made this work possible, and I am forever grateful for his trust and encouragement.

I also want to extend my appreciation to the committee members, Professor Xiaohui Yu and Professor Vijay Mago, for their guidance during my Master's defense. Their support and feedback have been invaluable.

I wish to thank Daoming Wan as well, whose suggestions have significantly improved my paper writing. His experience and collaboration have been essential in enhancing the quality of my work.

Last, but by no means least, I must acknowledge the unwavering support of my loving wife. Her encouragement, patience, and love have sustained me throughout this journey.

Table of Contents

Abstract **ii**

Acknowledgements **iv**

Table of Contents **v**

List of Tables **viii**

List of Figures **x**

1 Introduction **1**

1.1 Motivation 1

1.2 Contributions 3

1.3 Thesis Organization 6

2 Background and Literature Review **8**

2.1 Tradictional Information Retrieval 10

2.1.1 Boolean Model 11

2.1.2 Vector Space Model 13

2.1.3 Language Model 17

2.1.4 Probabilistic Model 19

2.2 Machine learning in Information Retrieval 22

2.2.1 Learning to Rank . . . 22
2.2.2 Deep Learning in Information Retrieval . . . 30
2.2.3 Biomedical Information Retrieval . . . 38

3 Limitations of General Language Models 41
3.1 Deep Learning in Passage Ranking . . . 42
3.2 The Deep Learning Track on TREC . . . 43
3.3 Combined Methodology . . . 45
3.3.1 Fuse Framework . . . 46
3.3.2 Experiment Results and Analysis . . . 49
3.4 Substitution Strategy . . . 51
3.4.1 Relation Extraction . . . 55
3.4.2 Experiment Results and Analysis . . . 58
3.5 Summary . . . 59

4 Diversified Prior Knowledge Enhanced General Language Model for Biomedical Text Retrieval 61
4.1 Framework Architecture . . . 62
4.2 Knowledge-based Query Expansion . . . 62
4.3 Aspect-based Filter . . . 66
4.4 Two-stage Ranking . . . 68
4.5 Diversity-based Score Reweighting . . . 71

5 Experimental Settings 73
5.1 Datasets . . . 73
5.2 Evaluation Metrics . . . 76
5.3 Settings . . . 79

6 Experimental Results and Analysis **82**

6.1 Experimental Results 82

6.2 Ablation Study 88

7 Conclusions and Future Work **90**

7.1 Conclusions 90

7.2 Future work 92

Bibliography **92**

A Published Papers **105**

B TREC Genomics 2006 Queries **106**

C TREC Genomics 2007 Queries **108**

D Samples of The TREC Genomics Dataset **111**

List of Tables

2.1 An example of Binary Keyword Indexing Library. The binary values (1 or 0) indicate the presence or absence of the corresponding keyword in each document. 11

2.2 An example of the labeling strategy of the classification pointwise approach. 25

2.3 Document retrieval runs of TREC Deep Learning Track 2019. RR (MS) is based on MS MARCO labels. All other metrics are based on NIST labels. . 32

3.1 Mathematical Explanations of Fuse Methods 49

3.2 Fuse Runs Descriptions of The TREC 2021 Deep Learning Track 49

3.3 The Comparison Results of Various Fuse Runs and Baselines in The TREC 2021 Deep Learning Track 50

3.4 The Comparison of Deep Learning (NDCG@10: 0.2962) and BM25 (NDCG@10: 0.5815) in Query 190623: "for what is david w. taylor known" 52

3.5 BM25 was observed to outperform the BERT model for queries containing special entities from YorkU's results of the TREC 2021 Deep Learning Track with the evaluation of NDCG@10. 53

3.6 Different performance of BM25 based on different queries from YorkU's results of the TREC 2021 Deep Learning Track with the evaluation of NDCG@10. . 56

3.7 Substitution Runs Descriptions of The TREC 2022 Deep Learning Track . . 58

3.8 The Comparison Results with Different Substitution Runs and Baselines in The TREC 2022 Deep Learning Track. 58

4.1 Examples of The Extracted Entities. 65

6.1 Comparison between the candidate relevant documents screened by the Aspect-based Filter and the ground truth relevant documents in the Gold Standard. 83

6.2 Experiment results of our DPK-GLM and baselines on the TREC-GENO Specific & Abstract under official metrics. The superscript "*" means the method is significantly better than the best baseline. 84

6.3 Experiment results of our DPK-GLM and baselines on the TREC-GENO Specific & Abstract under NDCG metrics. The superscript "*" means the method is significantly better than the best baseline. 86

6.4 The Ablation Study of the DPK-GLM framework on the TREC-GENO Specific & Abstract tasks under the official evaluation metrics. 89

B.1 The TREC Genomics 2006 Queries (TREC-GENO Specific) 106

C.1 The TREC Genomics 2007 Queries (TREC-GENO Abstract) 108

List of Figures

2.1 The main process of general text retrieval. 9
2.2 The "AND" concept of the Boolean Model. 13
2.3 An example of a 2-dimensional Vector Space Model. 16
2.4 An example of the Language Model. 17
2.5 The general framework of Learning to Rank. 23
2.6 An example of a classification pointwise approach. 24
2.7 The labeling strategy of the Pairwise approach. 26
2.8 The 2-Dimensional representation of text embedding in information retrieval. 33
2.9 Two distinct models CBOW and Skip-gram of Word2Vec. 34
2.10 Overall pre-training and fine-tuning procedures for BERT. 36
2.11 The input representation in BERT is a combination of three types of embeddings: token embeddings, segmentation embeddings, and position embeddings. 36

3.1 The Comparison of BM25, YorkU21a and CombSUM_rescal in The TREC 2021 Deep Learning Track 51

4.1 The architecture of our proposed framework. 63

6.1 The performance of Re-ranking under different α and ζ of DPK-GLM-RoBERTa. 87

Chapter 1

Introduction

1.1 Motivation

In the era of digital transformation, the Internet has evolved into a vast information repository, becoming an indispensable resource for all industries, including the biomedical field. The emergence of the Internet and the subsequent data explosion have entirely changed how we acquire and use information. Today, the Internet is not just a communication or entertainment tool. It has become an indispensable part of our daily lives, significantly impacting our learning and decision-making processes. It has changed the way we interact with the world, providing us with a wealth of information at our fingertips. Especially in the medical field, with the advent of the big data era, its paradigm has also changed.

The Internet allows people to acquire basic medical knowledge, such as disease symptoms, treatment methods, and preventive measures, at their fingertips. This convenient way of obtaining information enhances individuals' abilities, enabling them to make wise decisions about their health and well-being. The emergence of online medical Q&A platforms has further changed the interaction between patients and doctors, making remote consultations possible, thus saving time and resources and ensuring patient privacy. The shift to digital

medicine not only makes medical information more accessible but also makes medical services more patient-centered. In addition, with the help of computerized biomedical information retrieval systems, routine decision-making tasks that are repetitive and prone to human error can be effectively managed. These systems not only improve the quality of clinical services but also significantly reduce costs. They simplify the information retrieval process, making obtaining the necessary information easier for medical practitioners. This highlights the importance of developing advanced computer-assisted medical information retrieval systems. The potential benefits of such systems are enormous, from improving patient care to more effectively utilizing resources.

For medical practitioners, the Internet is a treasure trove of authoritative literature and the latest research results. The ability to access this rich information is crucial because every decision a doctor makes will have a significant impact on the patient's treatment outcome. Medical practitioners often turn to the Internet for inspiration and reference materials when faced with challenging cases. Currently, PubMed alone contains about 35 million entries. The exponential growth of biomedical literature has resulted in massive biological data, meaning that manually reading articles from bibliographic databases is no longer realistic. The information explosion requires automatic information extraction tools to extract knowledge from unstructured text and store it in structured knowledge bases to organize existing knowledge and efficiently discover new knowledge. However, a typical query can return hundreds to thousands of documents, and the rapid increase in biomedical literature makes it difficult for medical researchers, clinical doctors, medical service personnel, and the general public to find the biomedical information they need. Information overload can lead to inefficient and delayed decision-making, which can seriously impact patient treatment. Although there are many high-quality biomedical literature databases, such as PubMed, they often cannot meet doctors' and researchers' complex and precise requirements. These databases mainly rely on keyword searches and struggle to provide answers for more complex clinical queries. They cannot handle and understand the context of the query, resulting in a

mismatch between the information sought and the information provided. This is where deep learning comes into play.

In this information-rich time, we can obtain a large amount of data to train high-performance deep learning models. Deep learning models can enhance biomedical information retrieval systems, providing more accurate, context-specific results. This simplifies the information retrieval process and enables doctors to make wiser decisions or allows researchers to quickly locate the information they need to improve their work efficiency. Deep learning models can analyze and interpret complex medical data, including unstructured data such as medical records and clinical notes. They can understand the context of the query, providing more relevant, more accurate results. This can significantly improve the efficiency and effectiveness of biomedical information retrieval, making it a valuable tool for medical practitioners.

In conclusion, developing deep learning-enhanced biomedical information retrieval systems is a technological advancement and a necessary requirement in today's data-driven medical and research environment. Such systems will completely change how medical practitioners acquire and use information, paving the way for a more efficient and effective healthcare system. It will change the way we understand and use medical information, bringing better patient care and better medical outcomes. The potential benefits of such a system are enormous, using the power of deep learning to enhance biomedical information retrieval, thereby ushering in a new era of digital healthcare.

1.2 Contributions

General language models have demonstrated impressive capabilities in various information retrieval (IR) tasks [1]. A notable example is the Bidirectional Encoder Representations from Transformers (BERT) [2], 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 [3, 4, 5]. However, their training process is often difficult and costly due to the scarcity of high-quality annotated data, especially for niche sub-domains 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, accurately capturing biomedical terminology is essential for biomedical IR systems.

Employing prior knowledge (or external knowledge) has proven advantageous in addressing the challenges mentioned above. Several studies have investigated incorporating prior knowledge sources, such as biomedical ontologies, databases, and knowledge graphs, to enhance performance in biomedical IR systems [6, 7]. 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.

In this thesis, we propose a framework called the 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 [8]. 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 [9], as baseline models. For comparative purposes, we also employ two pre-trained domain-specific language models, namely BioBERT [3] and ClinicalBERT [4]. The results from experiments conducted on publicly accessible biomedical IR datasets and an ablation study manifest significant performance enhancements attributable to our proposed approaches.

1.3 Thesis Organization

Chapter 2: Background and Literature Review This chapter introduces the research background and presents a comprehensive literature review. It discusses both traditional information retrieval methods and machine learning-based information retrieval methods.

Chapter 3: Limitations of General Language Models This chapter explores the limitations of general language models by presenting experiments and findings from our participation in the TREC 2021 and 2022 Deep Learning Tracks.

Chapter 4: Diversified Prior Knowledge Enhanced General Language Model for Biomedical Text Retrieval In this chapter, the DPK-GLM approach proposed in the thesis is introduced. It explains three essential components: the Knowledge-based Query Expansion method, Aspect-based Filter, and Diversity-based Score Reweighting method. The motivations and purposes behind employing these methods are thoroughly discussed.

Chapter 5: Experimental Settings This chapter presents the experimental environment used in the thesis and provides insights into the experimental datasets, evaluation methods, baselines, and implementation details.

Chapter 6: Experimental Results and Analysis The experimental results of the proposed approach are discussed in this chapter. It includes the outcomes of ablation experiments, which further analyze the role and performance of each framework component.

Chapter 7: Conclusions and Future Work This final chapter concludes the master's thesis, situating the research work in a broader context. It also outlines potential prospects and directions for further investigation.

Chapter 2

Background and Literature Review

Information Retrieval (IR) plays a crucial role in daily life and is extensively used in practical applications, including web search, question-answering systems, medical assistants, and chatbots. The primary objective of IR is to identify the user's query needs and locate relevant information in the whole corpus. Typically, an IR system returns multiple relevant documents, which are then ranked based on their relevance to the user's query to provide the final results.

Figure 2.1 illustrates the fundamental process of text retrieval, which comprises two main stages: the initial retrieval stage and the ranking stage. In the initial retrieval stage, the retrieval system preprocesses the large-scale corpus and establishes an index. Subsequently, it employs the retrieval model to obtain relevant documents, forming the candidate document set. Moving on to the ranking stage, ranking methods are applied to sort the candidate document set. This sorting places documents with higher relevance at the forefront and those with lower relevant or unrelated content at the end. This ranking aims to ensure that users can quickly find useful information among the top-ranked documents. It is worth noting that the ranking stage can be repeated multiple times, further optimizing the relevance of the documents.

With the advent of artificial intelligence (AI), information retrieval can broadly be

categorized into two main approaches: traditional retrieval methods and machine learning-based retrieval methods. Traditional retrieval methods rely on well-established techniques and algorithms, while machine learning-based retrieval methods leverage AI models to enhance the accuracy and effectiveness of the retrieval process.

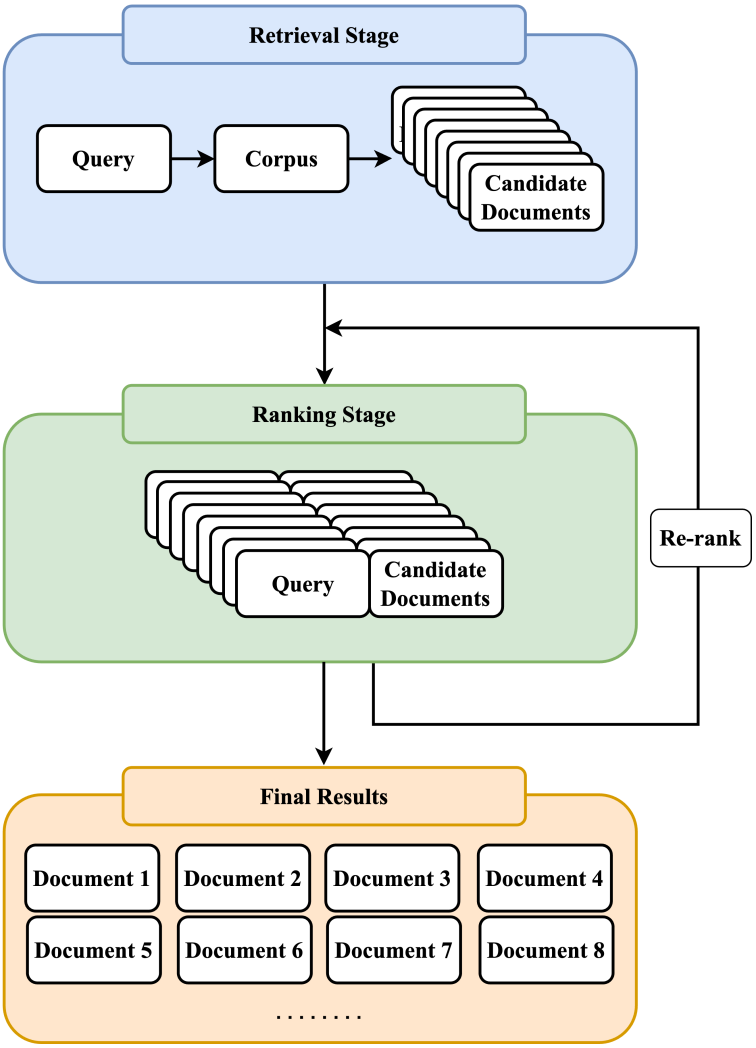


Figure 2.1: The main process of general text retrieval.

2.1 Tradictional Information Retrieval

Traditional information retrieval centers around models and algorithms designed to retrieve and rank documents from a corpus based on their relevance to the user's query [10, 11, 12]. The core concept of traditional IR revolves around search and storage. Once the retrieval system extracts relevant information, it organizes and stores the data in a structured manner to accommodate the user's future needs. An illustrative example of this can be observed in widely used search engines, where the user's query is parsed into several keywords. Subsequently, the search engine extracts highly relevant documents based on these keywords from its pre-established indexed document library and then presents the final ranked results.

Traditional information retrieval models can be broadly classified into four categories:

- Boolean Model: This model uses Boolean operators (AND, OR, NOT) to combine query terms and retrieve documents that match the query.
- Language Model: Language Models treat documents as a probability distribution of terms and compare the probability of generating the query given a document.
- Vector Space Model: This model represents documents and queries as vectors in a high-dimensional space and measures their similarity.
- Probabilistic Model: Probabilistic Models estimate the probability of a document being relevant to a query based on statistical measures.

In the context of academic research and industry application, these traditional IR models have been fundamental in advancing the field of information retrieval and continue to play a vital role in various applications and research endeavors.

2.1.1 Boolean Model

In seeking documents related to a query within a corpus, the most straightforward method is iterating through all documents and examining whether the query's keywords are present in each document [13]. While this approach may prove efficient for simple queries on a small-scale corpus, it becomes impractical in real-world scenarios where IR systems encounter more complex user requirements and vast volumes of data to be searched. A better alternative is employing a binary keyword indexing library, as shown in Table 2.1. This library operates as a matrix and enables the system to swiftly respond to queries seeking documents related to specific topics, such as finding documents related to bears, wolves, and monkeys (bear `AND` wolf `AND` monkey):

$$11001 \texttt{ AND } 11010 \texttt{ AND } 11001 = 11000$$

Table 2.1: An example of Binary Keyword Indexing Library. The binary values (1 or 0) indicate the presence or absence of the corresponding keyword in each document.

	Document 1	Document 2	Document 3	Document 4	Document 5	...
bear	1	1	0	0	1	...
wolf	1	1	0	1	0	...
tiger	0	0	1	0	0	...
monkey	1	1	0	0	1	...
...	...	...	...	...	...	...

The retrieval method that employs binary judgments based on the presence or absence of keywords in a document is commonly referred to as the Boolean Model. As one of the earliest and foundational models in the field of Information Retrieval, the Boolean Model plays a crucial role in efficiently organizing and retrieving relevant information.

As depicted in Figure 2.2, the Boolean Model operates through a straightforward set-based approach, considering each document as just a set of words. The figure illustrates two circles, one representing documents that contain the term "bear" and the other representing

documents that contain the term "forest". The intersection of these two circles represents the documents that would be retrieved by a Boolean query formulated as "bear AND forest". In the Boolean Model, the queries are constructed using logical operators such as AND, OR, and NOT, allowing users to perform more sophisticated searches to specify their information needs precisely.

The Boolean Model is valuable for its simplicity and effectiveness in certain situations, especially when dealing with queries requiring strict inclusion or exclusion criteria [14]. However, simplicity also entails poor performance. The Boolean Model has three main problems:

- Lack of Weighted Relevance: The Boolean Model does not consider the weight or importance of individual keywords in a query. Consequently, it cannot distinguish the degree of relevance of each document, leading to an inability to rank the returned results based on their relevance to the query.

- Limited by Boolean Operations: The model relies on Boolean operations to construct queries. As a result, it faces challenges in handling more complex situations that require flexible combinations of search terms.

- Inability to Support Partial Matches: The Boolean Model solely performs complete matches, meaning it cannot support partial matches. This limitation restricts its ability to control the number of documents returned, potentially returning redundant or insufficient results.

These problems led to the development of other Information Retrieval models, like the Vector Space and Probabilistic Models, catering to different retrieval requirements and challenges.

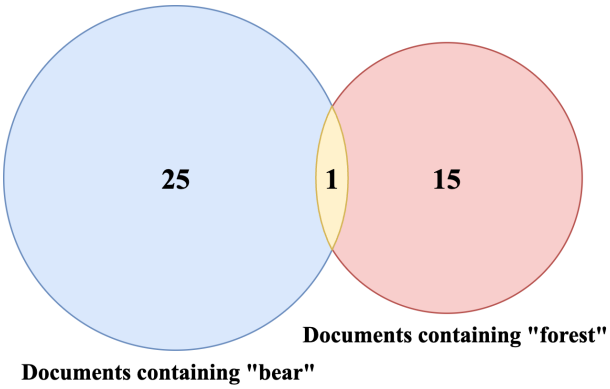


Figure 2.2: The "AND" concept of the Boolean Model.

2.1.2 Vector Space Model

In a realistic search engine scenario, the vast number of matching documents makes it impractical to manually filter out the most relevant ones. Thus, ranking becomes a fundamental and essential capability of a search engine. The ranking process involves calculating the relevance score of each matching document to a given query, aiding in the prioritization of search results.

Intuitively, a document that mentions a query's terms more frequently is likely to be more relevant to that query and, therefore, should be assigned a higher score. To achieve this, a weighting scheme called *term frequency* is employed. In the term frequency scheme, the relevance score between a query q and a document d is computed based on the weight of each query term t in document d. This weight is denoted as $tf(t,\ d)$, with the subscripts representing the term and the document, respectively.

The term frequency $tf(t,\ d)$ represents the count of occurrences $f(t, d)$ of term t in document d, and it can be simply formulated as follows:

$$tf(t, d) = f(t, d)$$

An alternative and widely used formula for *term frequency* is given by:

$$tf(t,d) = 1 + log(f(t,d))$$

where the value of 1 serves as a smoothing function, and the logarithmic function helps moderate the impact of extreme term frequencies.

By capturing the frequency of each query term within the document, the *term frequency* weighting scheme provides a simple yet effective method to rank documents based on their relevance to the query. Documents with higher term frequencies for the query terms will receive higher relevance scores, indicating a stronger alignment with the user's information needs.

Considering the frequency of terms solely has obvious drawbacks, as not all terms in a document hold equal value. Common words like "to" and "and" abound, yet they bear limited relevance to the overall meaning. However, the above approach lacks the ability to discern the importance of individual terms when assessing query relevance. For instance, when performing a web search for the "Japan earthquake," retrieved articles may lean towards discussing Japan in general, thus ignoring the significance of the earthquake in the given context. Consequently, a mechanism needs to be introduced to mitigate the impact of excessively frequent terms that may negatively influence the relevance judgment.

A straightforward approach is to reduce the weight of high-frequency terms in a document. If a term appears frequently across multiple documents, such as "can" and "try," it becomes worthless and provides limited discriminatory power. To address this, the concept of *document frequency* (df) was employed, representing the total number of occurrences of a term within the entire corpus.

To calculate the weight of *document frequency* df, the *inverse document frequency* (idf) is introduced as follows:

$$idf(t) = log\frac{N}{df(t,d)}$$

where $idf(t)$ denotes the inverse document frequency weight of term t, while N indicates the total number of documents within the corpus.

The widely recognized weighting scheme that combines *term frequency* and *inverse document frequency* is the *tf-idf* (*term frequency-inverse document frequency*) measure [15, 16], formulated as follows:

$$tf\text{-}idf(t, d) = tf(t, d) \times idf(t)$$

This formula demonstrates that the *tf-idf* score is at its highest when a t occurs multiple times within a small number of documents, thereby conferring a strong distinguish word to those particular documents. Conversely, the *tf-idf* score decreases when the t occurs fewer times in a document or appears across numerous documents, resulting in a low relevance.

The *tf-idf* effectively balances the importance of a term within a specific document (*term frequency*) against its significance within the entire corpus (*inverse document frequency*). As a result, *tf-idf* provides a powerful means of assessing the relevance of a document to a given query, promoting accurate and contextually relevant search results for users. This widely adopted weighting scheme plays a pivotal role in modern Information Retrieval systems and contributes significantly to their overall effectiveness and utility.

With *tf-idf*, both documents and queries can be represented as vectors composed of terms, as shown below:

$$\vec{d_i} = (t_{i,1}, t_{i,2}, \ldots, t_{i,n})$$

$$\vec{q} = (t_{q,1}, t_{q,2}, \ldots, t_{q,n})$$

where $\vec{d_i}$ denotes the vector for each document, $\vec{q}$ represents the vector for the query, t corresponds to the weight of each term, and n represents the total number of terms in the corpus.

The representation of documents or queries as vectors, known as the Vector Space Model, was introduced by Gerard Salton in the 1970s and has since played a pivotal role in the development of modern search engines. This model forms the foundation for various

information retrieval tasks, including document ranking, classification, and clustering.

In the Vector Space Model, each term is utilized to represent the weight of a specific dimension, resulting in an n-dimensional vector that represents the features or topic of the document. While the dimension of a document vector can be pretty high in reality, Figure 2.3 illustrates this principle using a two-dimensional vector as an example for clarity.

To determine the relevance of a document to a query, computing the cosine of the angle between the document vector and the query vector is a common choice. A smaller angle between their vector spaces indicates greater similarity. The cosine similarity is formulated as follows:

$$\cos(\vec{q}, \vec{d_i}) = \frac{\vec{q} \cdot \vec{d_i}}{|\vec{q}||\vec{d_i}|}$$

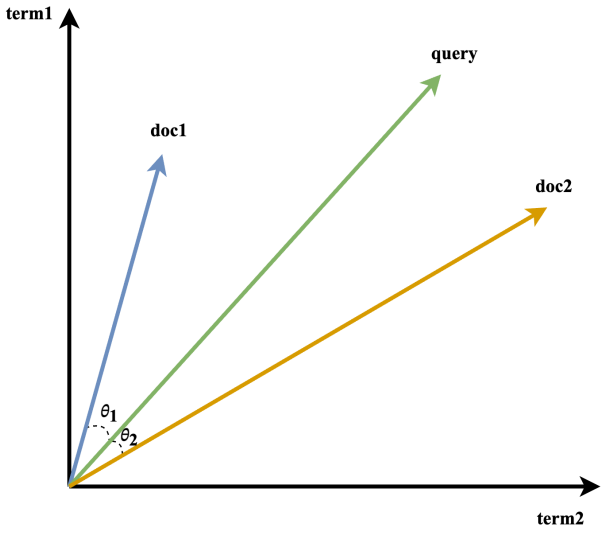


Figure 2.3: An example of a 2-dimensional Vector Space Model.

Despite its success, the Vector Space Model has limitations. It assumes independence between terms (the so-called "Bag-of-Words" assumption), ignoring the context and order of words. It also suffers from the curse of dimensionality, as the dimension of the vector space is equal to the number of unique terms in the document collection. These limitations have led to more sophisticated models, such as Probabilistic Models.

2.1.3 Language Model

Language Models in Information Retrieval are a probabilistic framework estimating the likelihood of a query given a document to rank documents based on these probabilities [17]. The concept of Language Models comes from Natural Language Processing (NLP) and Computational Linguistics, where they are used to predict the probability of a word or a sequence of words. In the context of IR, however, Language Models are used to estimate the probability of a query (a sequence of words) given a document, as shown in Figure 2.4.

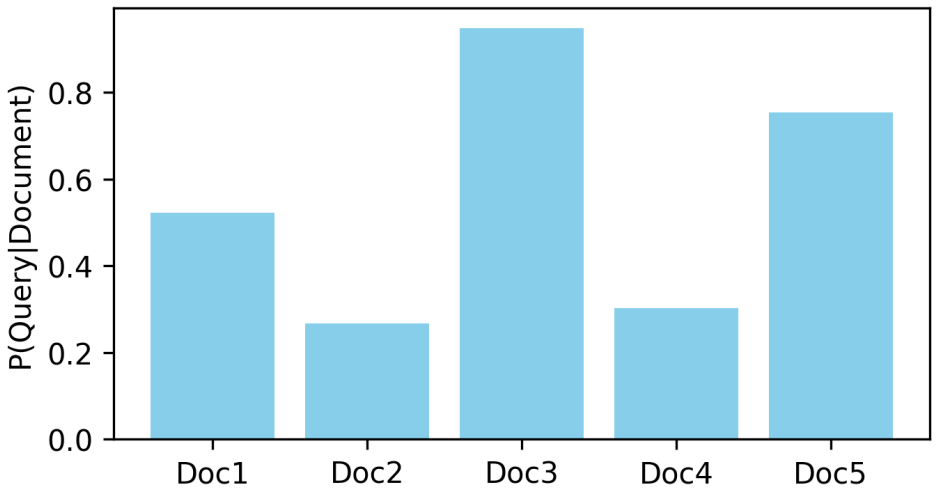


Figure 2.4: An example of the Language Model.

The simplest form of a Language Model in IR is the unigram Language Model, which treats each word in a document as an independent entity. This is similar to the Bag-of-Words model, a representation of text data where the order of words is disregarded, and each document is represented as a set (or "bag") of its words, disregarding grammar and word order but keeping multiplicity. The unigram Language Model shares this property, as it also ignores the order of words and treats each word independently. Therefore, the unigram Language Model is a probabilistic extension of the Bag-of-Words model, as it considers the presence of words and their frequency of occurrence, which is used to estimate probabilities.

The process can be formulated by:

$$P(t|d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

where $f_{t,d}$ is the frequency of term t in document d, and the denominator is the sum of frequencies of all terms in document d.

To handle the problem of zero probabilities in the Language Model, a common technique is Laplace smoothing (add-one smoothing) [18]:

$$P_{\mathscr{L}}(t|d) = \frac{f_{t,d} + 1}{\sum_{t' \in d}(f_{t',d} + 1)}$$

Higher-order n-gram models can be used in the Language Model to capture the context of words in a document. These n-gram Language Models can capture more context and provide a more accurate text representation at the cost of increased computational complexity and data sparsity issues.

For a bigram model (2-gram), the probability of a word given the previous word is:

$$P(t_i|t_{i-1}) = \frac{f_{t_{i-1},t_i}}{f_{t_{i-1}}}$$

where f_{t_{i-1},t_i} is the frequency of the bigram t_{i-1}, t_i, and $f_{t_{i-1}}$ is the frequency of the word t_{i-1}.

For a trigram model (3-gram), the probability of a word given the previous two words is:

$$P(t_i|t_{i-2}, t_{i-1}) = \frac{f_{t_{i-2},t_{i-1},t_i}}{f_{t_{i-2},t_{i-1}}}$$

where f_{t_{i-2},t_{i-1},t_i} is the frequency of the trigram t_{i-2}, t_{i-1}, t_i, and $f_{t_{i-2},t_{i-1}}$ is the frequency of the bigram t_{i-2}, t_{i-1}.

The connection between the Language Model and the Vector Space Model lies in their shared goal of representing documents to determine their relevance to a query. The Vector

Space Model represents documents as vectors in a high-dimensional space, where each dimension corresponds to a unique term, and its value represents the term's weight in the document. On the other hand, the Language Model represents documents as probability distributions over all terms. The relevance of a document to a query is assessed by the cosine similarity of their vectors in the Vector Space Model and the likelihood of the query according to the document's probability distribution in the Language Model.

In conclusion, the Language Model in IR can be seen as a probabilistic extension of the Bag-of-Words model and an alternative to the Vector Space Model, offering a different perspective on the information retrieval problem. By incorporating the context of words through n-gram models, Language Models can provide a more accurate and context-aware representation of documents, enhancing the effectiveness of information retrieval systems.

2.1.4 Probabilistic Model

The Probabilistic Model is a foundational information retrieval approach that leverages probability theory principles to rank documents in response to user queries. The primary concept behind this model is to rank documents based on the likelihood that a document is relevant to a given query.

The Binary Independence Model (BIM), the most straightforward and most well-known Probabilistic Model, was first proposed by Robertson and Sparck Jones in their paper, "Relevance Weighting of Search Terms", published in 1976 [19]. The BIM makes two major assumptions: (1) terms are binary variables (i.e., a term is either present or absent in a document), and (2) terms are independent of each other. The BIM employs term frequency in relevant and non-relevant document sets to estimate the probabilities necessary for ranking.

Over time, researchers have proposed various extensions and modifications to enhance the original Probabilistic Model's limitations and performance. Among these advancements, one of the most significant is the development of the BM25 ranking algorithm, also known as

Okapi BM25 [20].

BM25 is a Bag-of-Words retrieval algorithm that ranks a set of documents based on the appearance of query terms in each document, regardless of their order within the document. As a Probabilistic Model, BM25 extends the original BIM by incorporating *term frequency* and *document frequency* into the relevance score calculation. Unlike BIM, which treats term occurrences in a binary manner, BM25 accounts for the frequency of terms in a document, resulting in a more precise relevance score.

The BM25 ranking algorithm includes two parameters, k_1 and b, which control the scaling factor for term frequency and length normalization, respectively. Fine-tuning these parameters optimizes BM25's performance on specific tasks or datasets. The BM25 formula is as follows:

$$\text{score}(D, Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{(k_1 + 1) \cdot f(q_i, D)}{k_1 \cdot ((1 - b) + b \cdot \frac{|D|}{avgdl}) + f(q_i, D)}$$

where:

D is the document under consideration.

Q is the query that consists of n unique words $q_1, q_2, ..., q_n$.

$f(q_i, D)$ is the frequency of term q_i in document D.

$|D|$ is the length of the document D in words.

$avgdl$ is the average document length in the corpus from which documents are drawn.

k_1 and b are free parameters, usually chosen, in the absence of advanced optimization, as $k_1 = 2.0$ and $b = 0.75$.

$\text{IDF}(q_i)$ is the *inverse document frequency* weight of the query term q_i. It is usually computed as:

$$\text{IDF}(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5}$$

where N is the total number of documents in the corpus, and $n(q_i)$ is the number of documents containing q_i.

BM25 has been shown to outperform other ranking algorithms on various tasks and datasets, and it forms the basis of many modern search engines. However, it also has limitations:

- Assumption of Term Independence: BM25, like many traditional information retrieval models, assumes that terms are independent of each other. This means that it does not consider the context in which words appear or the relationships between words.
- Lack of Semantic Understanding: BM25 is a statistical model that relies on term frequency and document frequency. It does not understand the semantic meaning of words. For example, it would treat synonyms as entirely different words and would not recognize that two documents are similar if they use different words to express the same concept.
- Parameter Tuning: BM25 has two parameters (k_1 and b) that need to be tuned based on the specific dataset. There is no universally optimal setting for these parameters, and tuning them requires a validation set or extensive experimentation.
- Binary Relevance Assumption: BM25 operates under the binary relevance assumption, i.e., a document is either relevant or irrelevant to a query. In reality, relevance is often a matter of degree.
- Lack of Term Weighting: While BM25 considers term frequency and inverse document frequency, it does not consider other potential term weights, such as term importance or specificity.

Despite these limitations, BM25's versatility and effectiveness have made it a widely-used and highly-regarded Probabilistic Model in Information Retrieval. As a robust and reliable ranking function, it often serves as a strong baseline for various information retrieval tasks, enabling researchers and practitioners to benchmark their own models against its performance.

The Probabilistic Model, from its initial formulation as the Binary Independence Model to its modern incarnation as the BM25 algorithm, has played a crucial role in the development of information retrieval systems. Its focus on estimating the probability of relevance of documents provides a theoretically grounded and practically effective approach to the problem of ranking documents in response to a user query.

2.2 Machine learning in Information Retrieval

The above presentation has several categorizations but they are all keyword-based retrieval models. And the guidelines for their categorization are based on how the model measures the keywords. Traditional retrieval models are widely used because of their simplicity and ease of implementation. But this keyword-based mindset ignores or makes it technically challenging to solve semantic problems of words, such as synonyms, polysemy, etc. In today's information explosion, more factors need to be taken into account to achieve better ranking results, and this complex situation is difficult to be solved by some algorithm or human intervention. Therefore, it is intuitive to employ machine learning to enhance retrieval performance.

The intuition to employ machine learning to enhance retrieval performance has led to the exploration of Learning to Rank and BERT. These approaches differ from traditional methods because they understand semantic information and require less manual manipulation, relying instead on extensive training data for automatic learning and optimization. The applications of Learning to Rank and BERT in IR reflect a response to the evolving complexities of information needs, offering a pathway to more advanced and effective retrieval systems.

2.2.1 Learning to Rank

Learning to Rank (LTR) [21] is a subfield of machine learning that focuses on developing algorithms and techniques to construct ranking models for information retrieval systems

automatically. LTR aims to rank items (such as documents, products, or images) to maximize some measure of relevance or utility.

The concept of LTR emerged in the early 2000s, realizing that traditional information retrieval techniques, such as the Vector Space Model or Probabilistic Models like BM25, could be significantly improved by leveraging machine learning. The main advantage of LTR over traditional methods is its ability to learn from data and improve over time, as well as its capacity to handle complex, non-linear ranking functions and to incorporate a wide variety of features into the ranking process. Figure 2.5 is a general framework of LTR.

LTR methods can be broadly categorized into pointwise, pairwise, and listwise approaches.

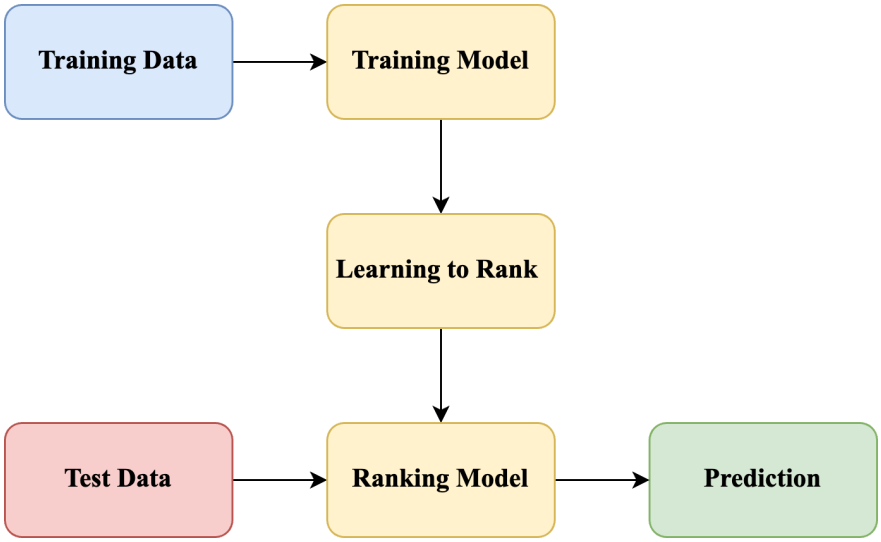


Figure 2.5: The general framework of Learning to Rank.

Pointwise In the pointwise approach, each document is considered independently of others. The goal is to predict a score or a class for each document, and then documents are ranked based on these scores. This approach transforms the ranking problem into a regression or classification problem [22]. An example of a pointwise approach is a linear regression model.

Given a document d and a query q, the relevance score s can be predicted as follows:

$$s(d, q) = w^T x$$

where x is the feature vector of the document-query pair, and w is the weight vector learned by the model.

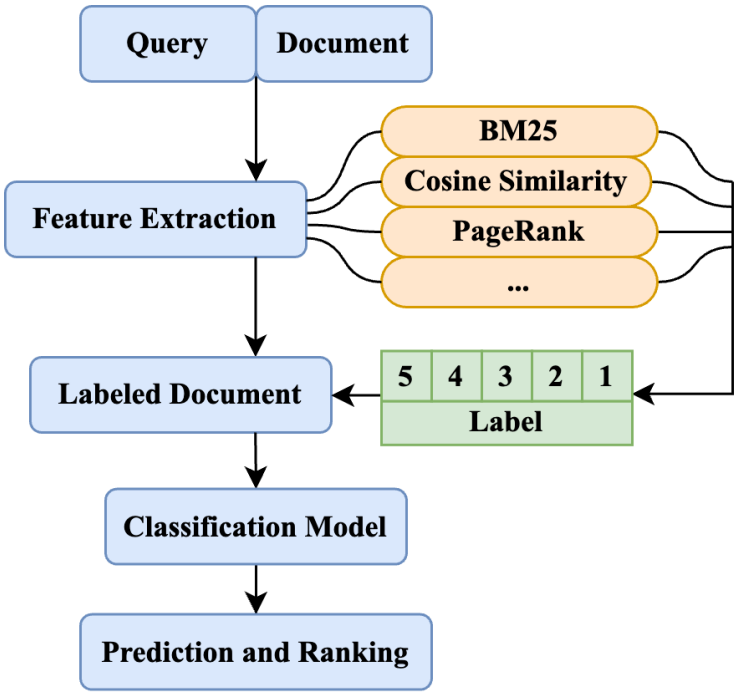


Figure 2.6: An example of a classification pointwise approach.

The pointwise approach can be divided into three categories: Regression-based, Ordinal Regression-based, and Classification-based algorithms. Taking multi-category classification as an example, the whole process is divided into three parts, as shown in Figure 2.6:

1. Feature Extraction: For each document-query pair, a feature vector is extracted. These features can include various types of information, such as term frequency, document length, or more complex features like PageRank or query-document similarity measures.

2. Model Training: A classification model is trained on the feature vectors. The target variable for training is the relevance score or class assigned to each document-query pair. This transforms the ranking problem into a classification problem. As shown in Table 1, we consider two features: the BM25 score and the cosine similarity. Since the relevance of the query to documents is multivariate, the labels are categorized into five levels, representing cases of perfect match, partial match, and complete mismatch. By doing so, each document is associated with a specific classification label. Various machine learning algorithms, such as linear regression, decision trees, or neural networks, can be employed in this stage.

3. Prediction and Ranking: The model predicts a new query's relevance score for each document. The documents are then ranked based on these predicted scores.

Table 2.2: An example of the labeling strategy of the classification pointwise approach.

Query	Document ID	BM25	Cosine Similairy	Label
apple	1	0.42	0.40	3
apple	2	0.70	0.65	4
apple	3	0.08	0.06	1
apple	4	0.21	0.19	2

The pointwise approach is conceptually simple and easy to understand. However, the main criticism of the pointwise approach is that it ignores the ranking structure of the data. It treats each document independently and does not consider the relative order of documents, which is a crucial aspect of ranking tasks. Furthermore, the loss functions typically used in regression or classification (like mean squared error or cross-entropy) may not be appropriate for ranking tasks. Unlike direct ranking optimization methods, these loss functions do not explicitly optimize the quality of the overall ranking.

Pairwise In the pairwise approach, the relative order of pairs of documents is considered. The goal is to learn a ranking function that minimizes the number of inversions in ranking.

An inversion is a pair of documents where the higher-ranked document has a lower relevance score than the lower-ranked document [23]. As shown in Figure 2.7, the pairwise method generates training instances by comparing the relevance labels of different document pairs for a specific query and its corresponding set of documents. For each pair of documents(d_i, d_j), if document d_i is considered more relevant than document d_j, it is assigned a positive label $+1$. Conversely, if document d_i is considered less relevant than document d_j, it is assigned a negative label -1.

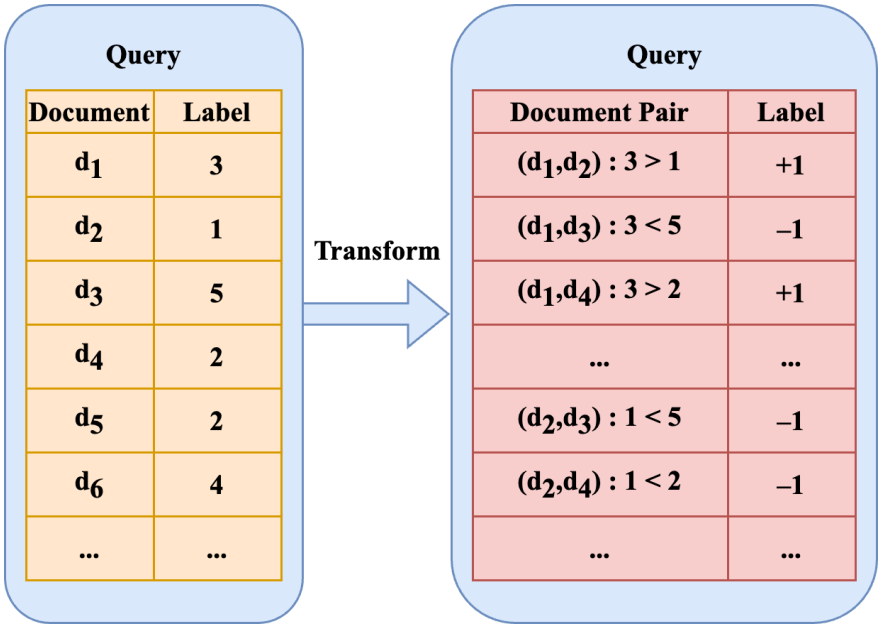


Figure 2.7: The labeling strategy of the Pairwise approach.

An example of a pairwise approach is RankNet [24]. It is based on using a neural network to model the probability that a document is more relevant than another, given a particular query. The model is trained by minimizing the cross-entropy loss between the predicted and true probabilities. The true probabilities are assumed to follow the logistic function, which is a common choice for binary classification problems. The model output is a score for each document, and the documents are ranked by their scores.

Given a pair of documents d_i and d_j, the probability that d_i should be ranked higher than

d_j is formulated as follows:

$$P(d_i > d_j) = \frac{1}{1 + e^{-\sigma(s_i - s_j)}}$$

where s_i and s_j are the predicted scores for d_i and d_j, and σ is a parameter controlling the steepness of the sigmoid function.

Through the construction of pairwise training instances, a binary classifier is taught to identify the distinguishing boundary between pairs of documents. This process significantly enhances the classifier's ability to rank documents effectively by a given query. The primary objective of the classifier is to accurately predict which document within each pair holds greater relevance, thus encapsulating the relative order of documents within a ranked list.

Although the pairwise method provides certain improvements over the pointwise approach, it has inherent challenges and limitations. These are mainly due to the rapid increase in the number of document pairs and the potential for cascading errors during the annotation process.

As the quantity of documents expands, there is a quadratic increase in the number of document pairs, leading to a substantial growth in the size of the training data set. This growth creates an imbalance between the number of queries and the number of document pairs. This disproportion can disrupt the training process and potentially cause the classifier to favor certain document pairs, thus detrimentally affecting the overall ranking performance.

Furthermore, the increased number of document pairs heightens the risk of annotation errors. Suppose a single document pair is incorrectly labeled during the training stage. In that case, this error can spread to multiple other document pairs due to the inherent interdependencies among documents in a ranked list. This domino effect can result in incorrect ranking decisions and compromise the overall effectiveness of the classifier.

Listwise In the listwise approach, the entire list of documents is considered. The goal is to minimize the difference between predicted and actual rankings. The goal is to learn a ranking

function that optimizes a list-level loss function, such as Normalized Discounted Cumulative Gain (NDCG) or Mean Average Precision (MAP) [25, 26].

An example of a listwise approach is LambdaRank [27], an extension of RankNet. In LambdaRank, the lost function is defined as the sum of the costs for all pairs of documents, where the cost for a pair of documents is a function of their predicted scores and true relevance grades. The lost function can be written as follows:

$$L = \sum_{i,j:y_i>y_j} C_{ij}$$

where y_i and y_j are the true relevance grades of documents i and j, and C_{ij} is the cost for pair (i, j). The cost C_{ij} is typically defined as a function of the difference between the predicted scores for the two documents and their true relevance difference:

$$C_{ij} = -\lambda_{ij} \log(1 + e^{s_i - s_j})$$

where s_i and s_j are the predicted scores for documents i and j, and λ_{ij} is a weight that depends on the true relevance difference $y_i - y_j$ and the ranking metric.

Another example is LambdaMART [28], which combines the ideas of LambdaRank and MART (Multiple Additive Regression Trees) [29]. The model is an ensemble of decision trees, and the predicted score for a document is the sum of the predictions of all the trees. The model can be written as follows:

$$f(x) = \sum_{k=1}^{K} T_k(x)$$

where x is the feature vector for a document, $T_k(x)$ is the prediction of the k-th tree for x, and K is the number of trees.

The trees are trained iteratively, where at each iteration, the tree that reduces the loss the most is added to the model. The loss function is defined as the sum of the losses for all

documents, where the loss for a document is a function of its predicted score and its true relevance grade:

$$L(y, f(x)) = \sum_{i=1}^{n} l(y_i, f(x_i)) - \lambda_i f(x_i)$$

where y is the true ranking, $f(x)$ is the predicted ranking, y_i is the true relevance grade of document i, l is a loss function (such as squared error or logistic loss), λ is the lambda for document i. The sum is over all documents in the training data.

The "Lambda" in LambdaRank and LambdaMART refers to the use of lambda functions as a way to compute the gradients for the boosting algorithm. These lambda functions are derived from the evaluation metrics that the ranking model tries to optimize. The process has two main steps:

1. For each pair of documents, compute the change in the ranking metric (NDCG or MAP) that would result from swapping the two documents in the ranking. The exact formula for computing the lambdas depends on the ranking metric. For NDCG, the formula considers the difference in relevance grades between the two documents and the positions of the documents in the ranking. The formula also considers the number of relevant documents ranked above each for MAP.

2. Use these lambdas as the gradients for learning. In LambdaRank, these gradients are used to update the neural network weights. In LambdaMART, these gradients are used to fit the decision trees.

Unlike the pairwise approach, which takes object pairs as instances in learning, the listwise approach uses lists of objects as instances. This makes it more aligned with the ultimate goal of ranking: producing a list of items in the correct order. However, the listwise approach also has its disadvantages. One of the main challenges is defining a suitable loss function that can accurately measure the difference between the predicted and actual rankings. This is a non-trivial task due to the inherent complexity of ranking problems. Furthermore, the

listwise approach can be computationally expensive, especially with large lists.

Over the years, numerous LTR algorithms have been proposed, including RankNet, ListNet, and RankBoost. These algorithms have been successfully applied in various domains, including web search, recommendation systems, natural language processing, and bioinformatics.

Despite its success, LTR also has its challenges. One of the main challenges is the lack of labeled training data, as obtaining relevant judgments is often expensive and time-consuming. Another challenge is the dynamic nature of the ranking problem, where the relevance of items can change over time due to changes in user preferences, item availability, or other factors.

2.2.2 Deep Learning in Information Retrieval

Deep Learning has substantially revolutionized the domain of Information Retrieval over the past decade [30, 31, 32, 33]. It has significantly contributed to constructing large-scale search engines such as Google and Bing and more focused applications, including online shopping searches, chatbots, and recommendation systems. The relevance of search or recommendation outcomes plays a pivotal role in the business performance across these domains; thus, enhancing search quality by even a marginal percentage can culminate in substantial profit augmentation.

Deep Learning, characterized by its capacity to learn intricate representations from raw data, has been successfully integrated into all IR principal tasks. The success of deep learning in these applications is attributed to its proficiency in accurately learning distributed representations (such as vector representations) of natural language expressions (such as sentences) and efficiently employing these representations in associated tasks. Table 2.3 is TREC Deep Learning Track 2019's document retrieval runs [34], showing deep learning-based models significantly outperformed traditional IR models.

Incorporating deep learning methods in IR has opened the door to potential advancements in state-of-the-art performances, paralleling the breakthroughs seen in other domains such as

Computer Vision, Speech Recognition, and Machine Translation. Foundational deep learning techniques, including word embedding, recurrent neural networks (RNN) [35, 36], and convolutional neural networks (CNN) [37], have been instrumental in propelling progress within IR. These techniques have found application in fundamental IR challenges, encompassing text matching, classification, and document ranking. This section focuses on two essential techniques: Embedding and BERT.

Embedding Embedding, with its extensive range of applications, is often considered the backbone of deep learning. The primary role of embedding is to "represent" an object using a low-dimensional dense vector. The object in question could be a word, an image, or even a specific entity. The term "representation" here does not denote transformation or equivalence but rather the utilization of a vector to reflect the characteristics of the object and the use of the distance between vectors to denote the similarity between these objects. Figure 2.8 provides a simplified two-dimensional visualization of how embedding operates in text retrieval (in a real-world scenario, this process occurs in a high-dimensional space):

- Input: The system takes in a query, ranging from a sentence to a paragraph or an entire document. Using an embedding model, this query is then transformed into a representation (vector).
- Embedding Space: The vector representation of the query is placed within a high-dimensional embedding space. Each dimension within this space corresponds to a different feature of the text (e.g., syntax, semantics, context). A combination of these features determines the positioning of the vector within this space.
- Comparison: In retrieving documents, the IR system compares the vector representation of the query with the vector representations of the documents in the corpus. This comparison often involves calculating the cosine similarity between the vectors.

Table 2.3: Document retrieval runs of TREC Deep Learning Track 2019. RR (MS) is based on MS MARCO labels. All other metrics are based on NIST labels.

run	group	subtask	RR (MS)	RR	NDCG@10	NCG@100	AP
idst_bert_v3	IDST	fullrank	0.4866	0.9612	0.7257	0.5800	0.3137
idst_bert_r1	IDST	rerank	0.4889	0.9729	0.7189	0.5179	0.2915
idst_bert_v2	IDST	fullrank	0.4865	0.9612	0.7181	0.5947	0.3157
idst_bert_v1	IDST	fullrank	0.4874	0.9729	0.7175	0.5820	0.3119
idst_bert_r2	IDST	rerank	0.4734	0.9729	0.7135	0.5179	0.2910
bm25exp_marcomb	h2oloo	fullrank	0.3518	0.8992	0.6456	0.6367	0.3190
TUW19-d3-re	TU-Vienna	rerank	0.4014	0.9457	0.6443	0.5179	0.2709
ucas_runid1	UCAS	rerank	0.4422	0.9109	0.6437	0.5179	0.2642
ucas_runid3	UCAS	rerank	0.4353	0.8992	0.6418	0.5179	0.2677
bm25_marcomb	h2oloo	fullrank	0.3591	0.9128	0.6403	0.6356	0.3229
bm25exp_marco	h2oloo	fullrank	0.3610	0.9031	0.6399	0.6191	0.3030
ucas_runid2	UCAS	rerank	0.4315	0.9496	0.6350	0.5179	0.2526
TUW19-d2-re	TU-Vienna	rerank	0.3154	0.9147	0.6053	0.5179	0.2391
uogTrDNN6LM	uogTr	fullrank	0.3187	0.8729	0.6046	0.5093	0.2488
TUW19-d1-re	TU-Vienna	rerank	0.3616	0.8915	0.5930	0.5179	0.2524
ms ensemble	Microsoft	fullrank	0.3725	0.8760	0.5784	0.4841	0.2369
srchvrs_run1	srchvrs	fullrank	0.3065	0.8715	0.5609	0.5599	0.2645
TUW19-d2-f	TU-Vienna	fullrank	0.2886	0.8711	0.5596	0.4103	0.2050
TUW19-d3-f	TU-Vienna	fullrank	0.3735	0.8929	0.5576	0.3045	0.1843
dct_tp_bm25e 2	CMU	fullrank	0.3402	0.8718	0.5544	0.4979	0.2244
srchvrs_run2	srchvrs	fullrank	0.3038	0.8715	0.5529	0.5572	0.2615
bm25tuned_rm3	BASELINE	fullrank	0.3396	0.8074	0.5485	0.5590	0.2700
dct_qp_bm25e	CMU	fullrank	0.3585	0.8915	0.5435	0.4924	0.2228
dct_tp_bm25e	CMU	fullrank	0.3530	0.8638	0.5424	0.4786	0.2098
uogTrDSSQE5LM	uogTr	fullrank	0.3264	0.8895	0.5386	0.1839	0.1085
TUW19-d1-f	TU-Vienna	fullrank	0.3190	0.8465	0.5383	0.2951	0.1647
ms_duet	Microsoft	rerank	0.2758	0.8101	0.5330	0.5179	0.2291
uogTrDSS6pLM	uogTr	fullrank	0.2803	0.8895	0.5323	0.1868	0.1129
bm25tuned_prf	BASELINE	fullrank	0.3176	0.8005	0.5281	0.5576	0.2759
bm25tuned_ax	BASELINE	fullrank	0.2889	0.7492	0.5245	0.5835	0.2816
bm25base	BASELINE	fullrank	0.2949	0.8046	0.5190	0.5170	0.2443
bm25base_rm3	BASELINE	fullrank	0.2405	0.7714	0.5169	0.5546	0.2772
runid1	CCNU_IRGroup	rerank	0.3058	0.7811	0.5164	0.5179	0.2366
bm25tuned	BASELINE	fullrank	0.2930	0.8872	0.5140	0.5262	0.2318
bm25base_prf	BASELINE	fullrank	0.2717	0.7774	0.5106	0.5303	0.2542
baseline	BITEM_DL	fullrank	0.2795	0.8037	0.4823	0.5114	0.2168
bm25base_ax	BASELINE	fullrank	0.2677	0.7424	0.4730	0.5148	0.2452

- Output: The system ranks the documents based on their similarity to the query and presents the most relevant ones.

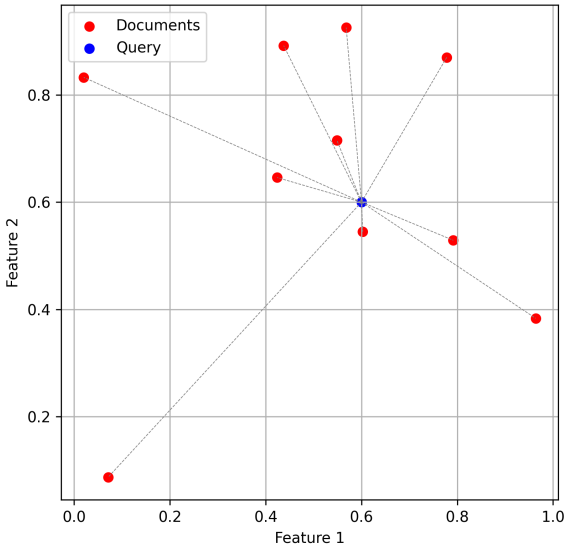


Figure 2.8: The 2-Dimensional representation of text embedding in information retrieval.

To explain the embedding technique in more detail, we take Word2Vec [12], one of the most popular methods for learning word vector representations, as an example. Word2Vec has two distinct models: the Continuous Bag of Words (CBOW) and Skip-gram.

The CBOW model assumes that a word is most closely related to its adjacent words, implying that a word is determined by its surrounding context. Conversely, the Skip-gram model posits that a word determines its adjacent words.

As depicted in Figure 2.9, given a sentence s consisting of n words $w_1, w_2, ..., w_t, ..., w_n$, the input for the Skip-gram model is the word w_t, and the output comprises words adjacent to w_t. In contrast, the CBOW model works in a reverse manner.

The Word2Vec model is trained by defining an objective function and optimizing this

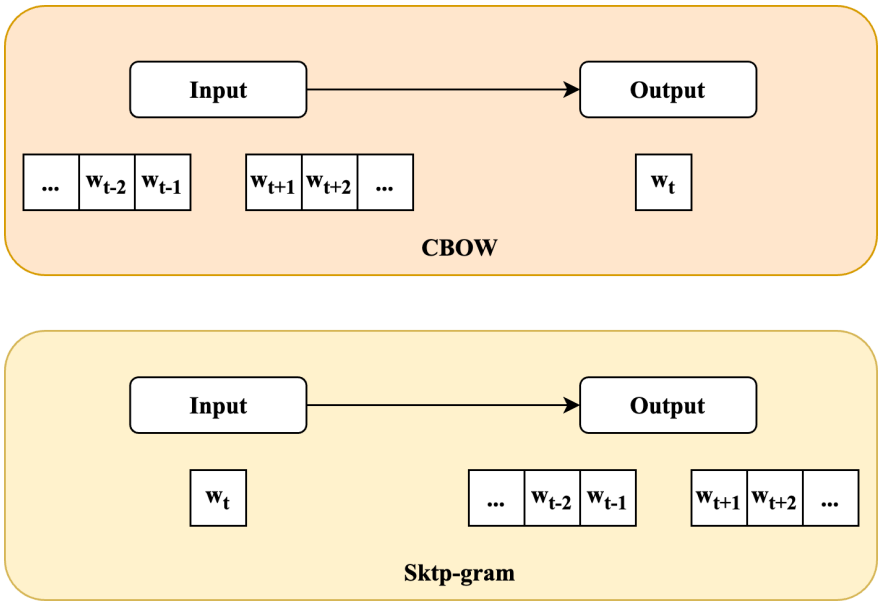


Figure 2.9: Two distinct models CBOW and Skip-gram of Word2Vec.

function. For Skip-gram, the objective function is to maximize the average log probability:

$$J(\theta) = \frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0} \log p(w_{t+j}|w_t)$$

where c is the size of the training context (which can be a function of the center word w_t), w_{t+j} is a context word, and w_t is a center word. The probability $p(w_{t+j}|w_t)$ is calculated using the softmax function:

$$p(w_{t+j}|w_t) = \frac{\exp(v'_{w_{t+j}} \cdot v_{w_t})}{\sum_{w=1}^{W} \exp(v'_w \cdot v_{w_t})}$$

where v_w and v'_w are the "input" and "output" vector representations of w, and W is the total number of words in the vocabulary.

For CBOW, the objective function is slightly different. The model predicts the current

word based on the context:

$$J(\theta) = \frac{1}{T}\sum_{t=1}^{T} \log p(w_t|w_{t-c}, ..., w_{t-1}, w_{t+1}, ..., w_{t+c})$$

where $w_{t-c}, ..., w_{t-1}, w_{t+1}, ..., w_{t+c}$ represent the context words of the target word w_t.

In terms of training, the model typically uses stochastic gradient descent and backpropagation to update the weights (i.e., the vector representations of words) and minimize the loss function.

Embeddings in Information Retrieval offer a powerful tool for understanding and processing text data. Their ability to capture semantic similarity means that words or documents with similar meanings are located close to each other in the vector space. This proximity allows for more effective semantic search and matching, enhancing the precision of IR systems. The vector space of embeddings also supports meaningful operations, such as addition and subtraction, which can be used to discover semantic relationships and analogies, thereby enriching the semantic understanding of IR systems. Furthermore, embeddings provide a form of dimensionality reduction, transforming high-dimensional data (like a large vocabulary) into a more manageable, lower-dimensional space. This transformation helps to mitigate the curse of dimensionality, improving the efficiency and scalability of IR systems. Lastly, the transfer learning capabilities of embeddings mean that models trained on one task or domain can be applied to another.

BERT BERT (Bidirectional Encoder Representations from Transformers) was introduced in a paper by Jacob Devlin and his colleagues from Google AI Language in 2018. It represented a significant breakthrough in Natural Language Processing (NLP). Before BERT, models like Word2Vec or GloVe [38] provided context-free models where each word was represented by a fixed embedding, regardless of the sentence context. BERT, on the other hand, is a deeply bidirectional model. BERT can access context from past and future directions by being

bidirectional, which was impossible in previous models. Figure 2.10 is the overall pre-training and fine-tuning procedures for BERT, and Figure 2.11 is the input representation for BERT [2].

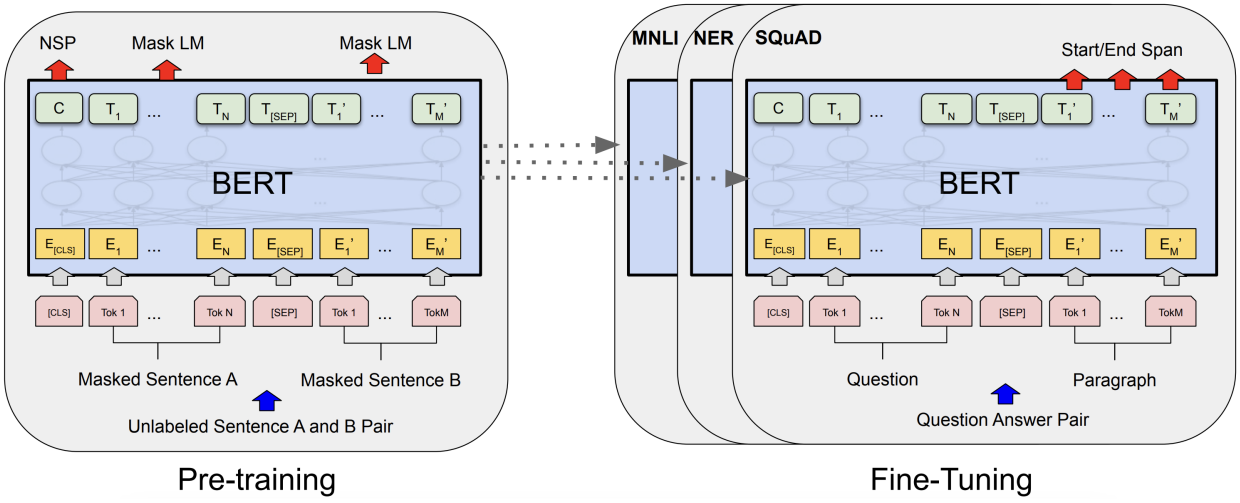


Figure 2.10: Overall pre-training and fine-tuning procedures for BERT.

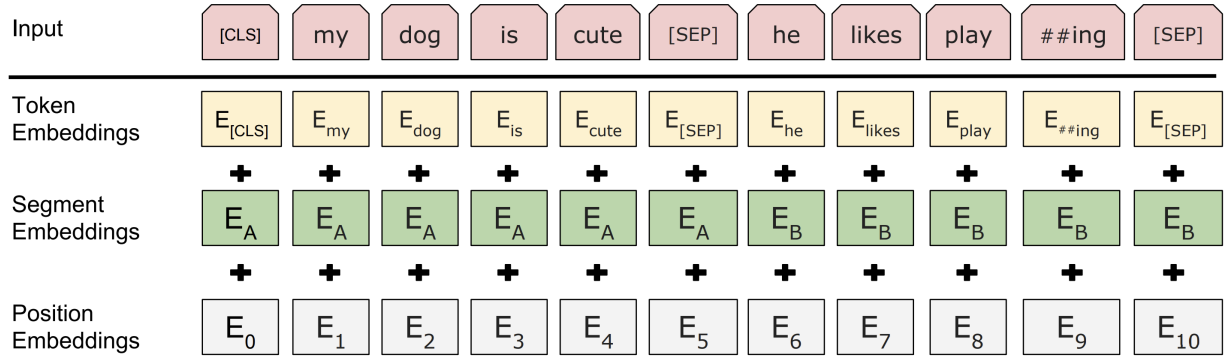


Figure 2.11: The input representation in BERT is a combination of three types of embeddings: token embeddings, segmentation embeddings, and position embeddings.

One of the core ideas of BERT is Transformer architecture. The Transformer model replaces recurrent layers with self-attention mechanisms, which weigh input vectors in the

embedding space according to their semantic similarity. This allows for parallel processing of sequences and better handling of long-term dependencies.

BERT also introduces the concept of "positional encoding," which gives the model information about the relative position of the words in a sentence. This is important because the self-attention mechanism in the Transformer model does not have any inherent notion of word order.

Another important aspect is the use of "subword tokenization". BERT uses WordPiece tokenization, where a word may be broken into multiple subwords. This allows the model to handle out-of-vocabulary words and introduce an element of character-level information into the model.

BERT uses two training strategies: Masked Language Model (MLM) and Next Sentence Prediction (NSP). In the MLM task, some percentage of the input words are masked randomly, and the model tries to predict them based on the context provided by the non-masked words. In the NSP task, the model is trained to understand the sentence's relationship, predicting whether one sentence follows another.

A key formula in BERT is the attention score in the self-attention mechanism, which is calculated as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q, K, and V represent the query, key, and value vectors, respectively, and d_k is the dimensionality of the key vectors.

BERT has been utilized effectively in various aspects of Information Retrieval, significantly enhancing the performance and accuracy of several tasks. One of the key applications is in document ranking. Traditional IR systems rank documents based on keyword matches. However, this approach can lead to less relevant documents being ranked higher if they contain more instances of the query keywords. BERT's contextual understanding enables it

to re-rank these documents, providing a more relevant list. BERT has also been employed in the task of query expansion. In traditional query expansion, synonyms or related words are added to the query to retrieve more documents. However, this can sometimes lead to query drift, where the expanded query retrieves irrelevant documents. BERT can mitigate this by generating contextually related terms for a query, thereby enhancing retrieval performance.

Furthermore, BERT has significantly improved the capabilities of semantic search [39, 40]. In traditional keyword-based search, if the exact query phrasing is not present in a document, the document might not be retrieved, even if it is relevant. With its understanding of the context and semantics of queries, BERT allows for more accurate retrieval of documents, even when the exact wording is not used in the query.

Lastly, Cross-lingual Information Retrieval (CLIR) [41, 42] is a field where BERT has shown promise. In CLIR, the query and the documents are in different languages. BERT's language-agnostic representation of text allows it to bridge the semantic gap between different languages, enabling the retrieval of information written in a language different from the query.

2.2.3 Biomedical Information Retrieval

Recent years have indeed been transformative for the field of Biomedical Information Retrieval (IR), with the emergence and success of pre-trained language models (PLMs) such as BERT and T5 [43]. These general models have shown commendable achievements in various IR tasks, transcending traditional methods and opening up new possibilities in areas like recommendation, query generation, and document ranking.

In the context of biomedical domain tasks, the landscape becomes more nuanced. Models like BioBERT have set benchmarks in tasks like disease named entity recognition (NER) and serve as essential tools in the field [44]. The distinct advantage of domain-specific PLMs like BioBERT lies in their optimization of biomedical data's intricate nature. These models address complexities that are beyond the reach of general language models.

Since 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. [45] 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. Even ClinicalBERT performed worse than the classic BM25 algorithm on certain tasks [46], highlighting the challenges and complexities specific to the biomedical domain.

The diversity of approaches in biomedical IR further underscores the complexity of the field. Wei et al. devised an ensemble approach combining convolutional neural networks (CNN) with long short-term memory (LSTM) networks, outperforming transformer models in specific tasks [47]. Such innovative combinations indicate an ongoing search for the ideal balance between computational models to capture the multifaceted nature of biomedical information.

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 [48, 49]. To address this issue, several studies have explored the use of domain-specific knowledge bases to enhance biomedical IR systems. Koopman et al. [50] proposed a graph inference model that obtained domain knowledge from SNOMED CT to tackle the semantic gap problem. Goodwin et al. [51] utilized multiple knowledge bases, such as MeSH and UMLS, to build a unified knowledge graph for topic analysis and expansion. Jin et al. [52] 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. [53] employed the Divergence from Randomness (DFR) method to boost performance in the initial ranking step for the biomedical literature search. Soldaini et al. [54] proposed a convolutional neural model to reduce clinical notes' noise for medical literature retrieval. KERS [55] 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.

Chapter 3

Limitations of General Language Models

General Language Models (LMs) have seen widespread adoption in various fields of Information Retrieval (IR), such as chatbots, machine translation, recommendation systems, document ranking, and relation extraction. However, despite their broad applications and successes, they do not come without limitations.

A primary challenge associated with these models is their inability to understand context beyond their training data. Consequently, they grapple with tasks that require knowledge or understanding beyond the scope of their training. Another limitation lies in their inability to interact with the user in a meaningful way, often leading to a lack of personalization and relevance in their responses [56, 57].

In the face of ambiguous queries, these models often struggle, potentially producing incorrect or irrelevant results. Such limitations, coupled with the challenges previously discussed, amplify when dealing with low-resource languages, potentially limiting their applicability in multilingual settings [58, 59].

Furthermore, these models are computationally intensive, demanding significant resources for both training and inference. The dependency on substantial training data to perform effectively is another hurdle, a condition that may not always be feasible or available. Finally,

the challenge of interpretability cannot be overlooked, as these models, often considered "black boxes", provide little insight into their decision-making processes [60].

This chapter, therefore, delves into the limitations of general language models within the classical passage ranking task in light of the experiments and findings from the Deep Learning Track of TREC.

3.1 Deep Learning in Passage Ranking

Passage ranking is a fundamental task in Information Retrieval (IR). It involves identifying and ranking the most relevant sections or passages from a document or a set of documents in response to a user's query.

The task of passage ranking is crucial in various applications, including question answering, document summarization, and reading comprehension. It is beneficial when dealing with long documents where the relevant information for a particular query is contained in specific sections or passages rather than the entire document.

The main challenge in passage ranking is to accurately identify the relevance of a passage to a given query. This involves understanding the semantics of both the query and the passage, and determining the degree of match between them. Various techniques have been proposed to tackle this challenge, including traditional keyword-based methods and more recent machine learning-based approaches.

In deep learning, passage ranking has been approached as a supervised learning problem, where models are trained on labeled data consisting of query-passage pairs and relevance judgments. More recently, with the advent of transformer-based models like BERT, significant improvements have been made in the field of passage ranking.

The following sections present the methodology and experimental results we used to participate in the 2021 and 2022 Deep Learning Track on The Text Retrieval Conference (TREC) passage ranking task [61, 62]. These results demonstrate that general deep learning

models have limitations even in their training domain. In addition, given the costly and time-intensive nature of the training process, it is necessary to explore strategies for cost-effective optimization.

3.2 The Deep Learning Track on TREC

The Text Retrieval Conference (TREC) Deep Learning Track is set to embark on a new phase by introducing a larger, cleaner corpus MS MARCO [63] that unifies the passage and document datasets. The Deep Learning Track centers on studying information retrieval in a large training data regime, where the number of training queries with at least one positive label is in the hundreds of thousands. This reflects real-world scenarios such as training based on click logs.

The MS MARCO (Microsoft Machine Reading Comprehension) dataset is an extensive resource designed to facilitate progress in artificial intelligence, particularly in the sphere of deep learning for search. Comprising real Bing search queries, the dataset encompasses a range of subsets, each focusing on different facets of deep learning in search, such as document ranking, passage ranking, and relation extraction.

In 2020 [64], the document ranking dataset was predicated on source documents, which encompassed passages utilized in the passage task. The corpus comprised 3.2 million documents and a training set with 367,013 queries. Each training query was linked from a positive passage ID to the corresponding document ID in the corpus under the presumption that a document yielding a relevant passage is typically relevant. This passage dataset was based on the public MS MARCO dataset, with a distinct set of test queries and a more detailed evaluation of the quality of passage rankings by relevant judges.

In 2021 [65] and 2022 [66], the MS MARCO dataset underwent substantial enhancements to render the document and passage data larger, cleaner, and more realistic. The document dataset expanded to be 3.7 times larger than its previous version, and the passage dataset

grew to be 15.6 times larger. The new dataset contains an average of 11.6 passages per document, selected using an algorithm that identifies the most promising passage candidates in a query-independent manner. This is a significant increase from the old dataset, which had 2.8 passages per document. The new dataset also features a known passage-document mapping, encouraging participants to consider how passage information may be used in document ranking and document information may be used in passage ranking.

The Track focuses on document ranking and passage ranking, aiming to study the effectiveness of various methods when a large amount of training data is available. Participants are encouraged to submit standard approaches and baselines, implement newer approaches, and try hybrid approaches enabled by the new document and passage corpus with passage-document mapping.

The Deep Learning Track has two tasks: passage ranking and document ranking, and two subtasks in each case: full ranking and reranking. Participants can submit up to three runs for each of the subtasks. Each task uses a large human-generated set of training labels from the MS MARCO dataset. The two tasks use the same test queries and the same form of training data, with usually one positive training document/passage per training query.

The Track also encourages participants to study the efficacy of transfer learning methods and use external corpora for large-scale language model pretraining or adapt algorithms built for one Track task to the other. This allows participants to study a variety of transfer learning strategies.

The document ranking task focuses on full ranking and top-100 reranking. In the full ranking subtask, participants are expected to rank documents based on their relevance to the question, where documents can be retrieved from the full document collection provided. In the reranking subtask, participants are provided with an initial ranking of 100 documents from a simple IR system, and they are expected to rerank the documents in terms of their relevance to the question.

The passage ranking task also has full ranking and reranking subtasks. In the full ranking

subtask, given a question, participants are expected to rank passages from the full collection in terms of their likelihood of containing an answer to the question. In the reranking subtask, participants are provided with an initial ranking of 100 passages, and they are expected to rerank these passages based on their likelihood of containing an answer to the question.

A set of test queries is provided as the official evaluation set, where NIST assessors will judge a subset. In addition to the main evaluation using the NIST labels and NDCG, sparse labels for the test queries, which already exist as part of the MS MARCO dataset, are also available. This allows calculating a secondary metric, Mean Reciprocal Rank (MRR). For the full-ranking setting, NDCG is also computed to evaluate the performance of the candidate generation stage.

3.3 Combined Methodology

Deep learning models like Transformers and BERT have exhibited significant potential across numerous information retrieval tasks. They excel in comprehending the semantic meaning of the text and capturing complex contextual relationships between words. This makes them particularly effective for semantic search, wherein the aim is to locate documents that are semantically related to a query, even if they do not share exact keywords.

Conversely, traditional retrieval models like BM25 are predicated on exact keyword matching. They cannot understand the semantic meaning of text and cannot capture contextual relationships between words. However, they prove highly effective in locating documents that share exact keywords with a query. This makes them especially suitable for the exact match search, where the objective is to find documents that contain the exact keywords specified in a query.

In practical applications, both the semantic and exact match search can be helpful, contingent upon the specific task and the user's information requirements. Consequently, it would be instructive to compare the performance of deep learning models and BM25 in

real-world scenarios. Such a comparison could yield valuable insights into the strengths and limitations of deep learning models in information retrieval.

With this aim, we participated in the passage ranking task in the 2021 and 2022 Deep Learning Track at TREC. The Deep Learning Track provided real queries and a corpus from the Bing search engine, offering a robust platform to investigate the problem in question.

In 2021 [61], the proposed methodology involved applying linear fusion methods to combine the results of BM25 and deep learning models. The objective was to capitalize on the strengths of both models: the semantic comprehension afforded by deep learning models and the exact match capabilities of BM25. It was hypothesized that integrating both models' results could yield superior performance compared to using either model in isolation.

In 2022 [62], the experiment was taken a step further. Rather than relying on the results of deep learning models, results less proficient than their BM25 counterparts were directly substituted. This approach provided a unique perspective on understanding the limitations of deep learning models in exact match search. It was anticipated that this strategy would offer valuable insights into the types of queries and documents where BM25 outperforms deep learning models and assist in pinpointing areas where deep learning models could be further enhanced.

3.3.1 Fuse Framework

The fuse framework has three main stages. The initial stage involved the application of pre-trained models. Specifically, Sentence-BERT [67], to generate embeddings for queries and documents. The msmarco-MiniLM-L-6-v3 model was chosen for this task, given its computational efficiency. The Semantic search was then conducted to retrieve the top 100 most relevant passages as the candidate set from each JSONL file. The candidate set was a submitted run as YorkU21b. Two pre-trained cross-encoder models, ms-marco-MiniLM-L-12-v2 and ms-marco-MiniLM-L-6-v2, were used to re-ranking these candidate sets. The

outcomes of these three rankings were combined through a voting mechanism, with the top 100 most relevant passages for each query selected as the final result. This constituted the submitted run, YorkU21a. An additional submission, YorkU21c, was made for comparison purposes, which was re-ranked solely by ms-marco-MiniLM-L-6-v2.

The second stage of the study centered around traditional information retrieval methods. Anserini [68], a toolkit for replicable information retrieval research, was employed to secure the BM25 results with default parameters.

In the final stage, the results of YorkU21a and BM25 were integrated. This phase was critical as it allowed for leveraging the strengths of deep learning (semantic understanding) and traditional retrieval methods (exact keyword matching) and combining these methodologies to yield a more robust and effective retrieval system.

Linear Fuse The linear fuse approaches follow [69]'s work, and they proposed six different methods: CombMAX, CombMIN, CombSUM, CombANZ, CombMNZ, and CombMED. Each method has its focus, and its mathematical explanations can be seen in Table 3.1, where s_b stands for the scores of BM25, s_d represents the scores of deep learning models, and N_{non} shows the number of non-zero scores retrieved candidates. CombMIN is designed to lessen the probability that non-relevant passages would be highly ranked, while CombMAX is made to reduce the number of relevant passages that receive low rankings. And CombMED uses the median value to combine the advantages of CombMIN and CombMAX. CombANZ is the average of the candidates with non-zero scores to avoid the influence of a single or no candidate; CombMNZ will give passages retrieved via multiple approaches more weight; CombSUM is the simplest way of summing up the scores of the same candidates.

Normalization Since the use of different retrieval systems will have different weighting methods and thus produce quite different ranges of similarity values, it is necessary to apply normalization methods to the different retrieval results. We used three different methods for

Normalization: Rescaling, Mean Normalization, and Z-score Normalization.

Rescaling, also known as min-max normalization, is the simplest way to reduce the range of data to [0, 1] or [-1, 1], where the general formula for Normalization of [0, 1] is as follows:

$$s' = \frac{s - \min(s)}{\max(s) - \min(s)}$$

where s is an original score of ranking, s' is the normalized ranking score.

Mean Normalization is another way to normalize data. It works by calculating and subtracting the mean value of each data. A common practice is to divide this value by the range or standard deviation, and the formula is as follows:

$$s' = \frac{s - \bar{s}}{\max(s) - \min(s)}$$

where $\bar{s}$ is the mean of scores.

When the same process uses standard deviations as denominators, the process is called Standardization, also known as Z-score Normalization. Z-score Normalization is a widely used method for Normalization in many machine learning algorithms. This method gives the mean and standard deviation of the original data to standardize the data. The processed data conforms to the standard normal distribution, that is, the mean is 0, and the standard deviation is 1, and the formula is as follows:

$$s' = \frac{s - \mu}{\sigma}$$

where μ is the mean of all sample data and σ is the standard deviation of all sample data.

Table 3.1: Mathematical Explanations of Fuse Methods

Methods	**Formula**
CombMAX	$max(s_b, s_d)$
CombMIN	$min(s_b, s_d)$
CombMED	$med(s_b, s_d)$
CombSUM	$sum(s_b, s_d)$
CombANZ	$sum(s_b, s_d)/N_{non}$
CombMNZ	$sum(s_b, s_d) * N_{non}$

3.3.2 Experiment Results and Analysis

Since YorkU21a yielded the best result for deep learning retrieval under the NDCG metric, this run was solely utilized in combination with BM25. The experiment runs are denoted as YorkU21a, BM25, CombMAX, CombMIN, CombANZ, CombMNZ, CombMED, CombSUM_rescal, CombSUM_zscore, and CombSUM_mean. These run descriptions are detailed in Table 3.2.

Table 3.2: Fuse Runs Descriptions of The TREC 2021 Deep Learning Track

Runs	**Description**
BM25	Use BM25 algorithm to obtain a baseline run by Anserini.
YorkU21a	Use Sentence-Bert to obtain the best re-rank result.
YorkU21b	Use Sentence-Bert to obtain the initial rank result.
YorkU21c	Use Sentence-Bert to obtain the alternative re-rank result.
CombMAX	Use CombMAX to combine the Yorku21a and BM25 by Rescaling.
CombMIN	Use CombMIN to combine the Yorku21a and BM25 by Rescaling.
CombMED	Use CombMED to combine the Yorku21a and BM25 by Rescaling.
CombANZ	Use CombANZ to combine the Yorku21a and BM25 by Rescaling.
CombMNZ	Use CombMNZ to combine the Yorku21a and BM25 by Rescaling.
CombSUM_rescal	Use CombSUM to combine the Yorku21a and BM25 by Rescaling.
CombSUM_zscore	Use CombSUM to combine the Yorku21a and BM25 by Z-score.
CombSUM_mean	Use CombSUM to combine the Yorku21a and BM25 by Mean.

The results of the full passage ranking runs are exhibited in Table 3.3. The NIST evaluation provided a four-degree level: “0” for irrelevant, “1” for related, “2” for highly relevant, and “3” for perfectly relevant. A judgment level of “1” for passages implies that the passage was

Table 3.3: The Comparison Results of Various Fuse Runs and Baselines in The TREC 2021 Deep Learning Track

Runs	**MAP**	**P@10**	**NDCG@10**	**NDCG@100**
BM25	0.1357	0.3547	0.4458	0.3913
CombMAX	0.2602	0.5097	0.5948	0.5223
CombMIN	0.2447	0.4941	0.5761	0.5077
CombMED	0.2587	0.5073	0.5911	0.5197
CombANZ	0.2786	0.5301	0.6199	0.5459
CombMNZ	0.2814	0.5321	0.6245	0.5511
CombSUM_rescal	**0.2917**	**0.5491**	**0.6370**	**0.5654**
CombSUM_mean	0.2230	0.5283	0.6122	0.4691
CombSUM_zscore	0.2602	0.4906	0.5875	0.5379
YorkU21a	**0.3309**	**0.6151**	**0.6965**	**0.5779**
YorkU21b	0.2032	0.4792	0.5694	0.4112
YorkU21c	0.3323	0.6189	0.6930	0.5413

related to the query but did not actually answer it. Thus, a "1" is not considered relevant for measures that use binary relevance judgments. Except for the calculation of NDCG, which requires all levels, the remainder of the evaluation calculation should not consider level "1".

The results suggest that the deep learning model delivers the most optimal outcomes. However, the choice of normalization methods does influence these results, with Rescaling emerging as the top performer. For a clearer comparison between deep learning, BM25, and their combined methods, refer to Figure 3.1. This figure indicates that when there's a significant performance difference between deep learning and BM25, the combined method tends to lag behind deep learning. On the other hand, when both methods exhibit comparable performance, the combined result tends to slightly edge out deep learning.

It's noteworthy that even in situations where deep learning considerably surpasses BM25, there are still certain queries where BM25 reigns supreme. Table 3.4 delves into an analysis of one such instance. The keyword-based BM25 model appears more adept at locating relevant passages when the query contains specialized entities, such as individual names. In contrast, the semantic-based deep learning model often identifies passages that only partially match

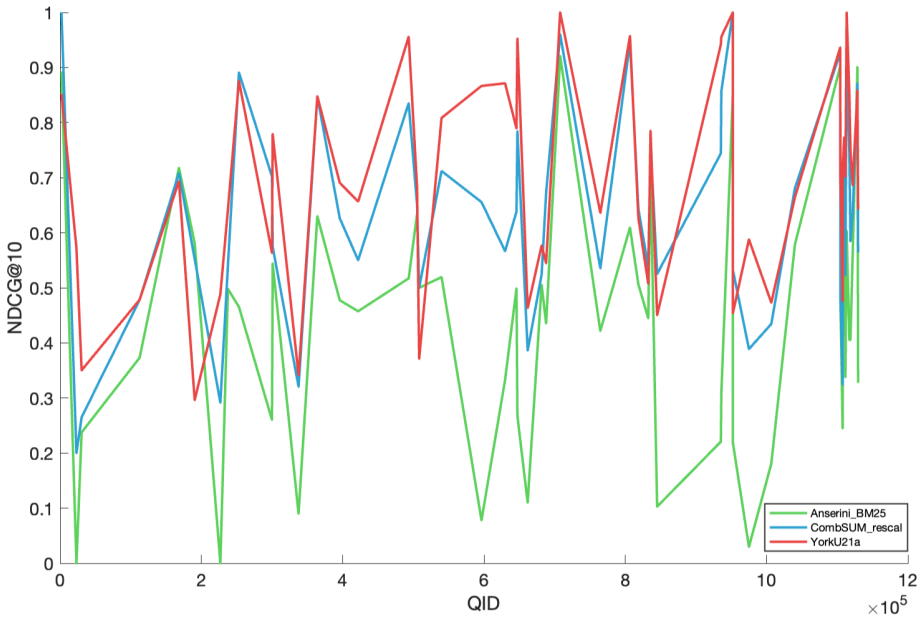


Figure 3.1: The Comparison of BM25, YorkU21a and CombSUM_rescal in The TREC 2021 Deep Learning Track

the keywords, rather than capturing exact matches. As an example from Table 3.4, the deep learning model treats "David Taylor" and "David W. Taylor" as identical, even though they are distinct.

3.4 Substitution Strategy

Recent advancements in deep learning-based dense passage retrieval [70] have demonstrated significant superiority over traditional retrieval techniques such as TF-IDF and BM25 [20] on well-established question-answering and information retrieval (IR) datasets. These dense models are trained to utilize annotated datasets, with several outperforming BM25 with as few as 1000 supervised examples. This is achieved by employing BERT [2, 71] as a pre-trained model with fine-tuning. This signifies high potential for practical applications and suggests a possible replacement for traditional retrieval methods. However, our 2021's

Table 3.4: The Comparison of Deep Learning (NDCG@10: 0.2962) and BM25 (NDCG@10: 0.5815) in Query 190623: "for what is david w. taylor known"

Runs	Rank	Related	Passage
YorkU21a	1	3	Rear Adm. **David W. Taylor**. Rear Admiral David Watson Taylor, USN (March 4, 1864 - July 28, 1940) was a naval architect and engineer of the United States Navy. He served during World War I as Chief Constructor of the Navy, and Chief of the Bureau of Construction and Repair. **Taylor** is best known as the man who constructed the first experimental towing tank ever built in the United States.
YorkU21a	2	0	World Champ **David taylor** the magic man. World Champ. **David taylor** the magic man. **David Taylor**, widely known as The Magic Man, is a 4x NCAA All-American, 4x BIG 10 Champion, and a 2x NCAA Champion – and he just getting started. Having wrapped up his NCAA career in March of 2014, David is just getting started on his international career and ultimately, his quest for Gold in Tokyo, 2020.
YorkU21a	3	0	**Taylor** is best known for his contributions to microhistory, exemplified in his William Cooper's Town: Power and Persuasion on the Frontier of the Early American Republic (1996). Using court records, land records, letters and diaries, Taylor reconstructed the background of founder William Cooper from Burlington, New Jersey, and the economic, political and social history related to the land speculation, founding and settlement of Cooperstown, New York, after the American Revolutionary War.
BM25	1	2	History. The facility was previously known as the **David W. Taylor** Naval Ship Research and Development Center; it was renamed David Taylor Research Center (DTRC) in 1987 and later became the Carderock Division of the Naval Surface Warfare Center in 1992.
BM25	2	2	David S. Taylor, CEO of Procter and Gamble. David Taylor (Wisconsin judge), American jurist and legislator. **David W. Taylor**, U.S. Navy admiral and engineer. David Taylor (banker), banker. David Taylor (veterinarian), television presenter on animal subjects.
BM25	3	3	Rear Adm. **David W. Taylor**. Rear Admiral David Watson Taylor, USN (March 4, 1864 - July 28, 1940) was a naval architect and engineer of the United States Navy. He served during World War I as Chief Constructor of the Navy, and Chief of the Bureau of Construction and Repair. Taylor is best known as the man who constructed the first experimental towing tank ever built in the United States.

Table 3.5: BM25 was observed to outperform the BERT model for queries containing special entities from YorkU's results of the TREC 2021 Deep Learning Track with the evaluation of NDCG@10.

Query ID	**Question**	**BM25**	**YorkU21a**
168329	does light intensity or concentration of carbon dioxide have a higher rate of photosynthesis	0.7179	0.6937
190623	for what is david w. taylor known	0.5812	0.2962
508292	symptoms of neuroma	0.5000	0.3708
1128632	is levothyroxine likely to cause weight loss or weight gain	0.9009	0.8582

study demonstrated that dense retrieval models have not yet developed sufficient capabilities to supplant traditional methods entirely, and it explored some of the inherent limitations that continue to afflict dense retrievers.

Building on the research conducted last year, this study primarily delves into methods to further optimize the integration of BM25 and deep learning models. The prior research revealed that while the deep learning model surpassed the traditional retrieval model for a majority of queries, BM25 showcased remarkable prowess with specific queries, especially those involving human names. As depicted in Table 3.5 from the earlier study, BM25 consistently outperformed the BERT model when handling queries containing specialized entities like individual names, places, organizations, and unique terms. Such entities, which are commonly found in real-world queries, possess distinguishing features: they typically have limited synonyms and refer explicitly to specific items. These characteristics render BM25 especially proficient in managing such queries. On the flip side, deep learning models, relying on word embeddings, can sometimes mistake similar words for synonyms, leading to the retrieval of irrelevant passages. Given these findings, the current study continued to focus on the potential benefits of combining the deep learning model with BM25. The objective is to harness BM25's capabilities for particular queries while leveraging the deep learning model's strengths for the remainder.

Before the advent of dense retrieval models, traditional retrieval techniques such as TF-IDF and BM25 dominated the landscape of information retrieval systems [72, 73, 74, 75]. These methods are characterized by relying on mathematical models to describe the retrieval process and using weighted term matching between queries and passages to determine similarity. Unlike deep learning models, traditional retrieval models do not necessitate training on labeled datasets. While traditional retrieval models excel at lexical matching, they fall short in capturing synonymy and intricate semantic relationships.

In contrast, dense retrievers utilize pre-trained language models, such as BERT, to compute similarity through embeddings learned from labeled datasets. These models typically employ two encoders: one for the query and one for the passage. The downstream tasks are then fine-tuned based on these pre-trained models. Both queries and passages are represented as word embeddings, and the top passages with the highest similarity scores to the query are returned and ranked accordingly.

While dense models perform admirably within their training domains, their ability to adapt to unfamiliar domains remains a point of concern. Thakur et al. [76] introduced a zero-shot benchmark named BEIR, demonstrating that dense retrieval models did not perform as well as BM25 in most of their datasets. Lewis et al. [77] found that the model had a tendency to memorize the training data, which results from the substantial overlap between the training and test sets and the propensity of deep learning models to overfit the training data for optimal performance. Chen et al. [78] developed demonstrating entity disambiguation proficiency, and they found that the models performed significantly worse on rare entities than on common entities. Sciavolino's findings [79] were consistent with this, indicating that the performance of dense retrieval models needs improvement in generalizability, particularly regarding rare entities. Their study concluded that integrating BM25 and deep learning models is viable.

Our approach involves a simple yet effective way to find queries suitable for retrieval through BM25. For these queries, the ranking results produced by the deep learning model

are then substituted with the results generated by BM25. Details of this approach will be presented in the following subsections. In this year's participation, we presented two outcomes: the dense retrieval YorkU22a and the non-dense retrieval YorkU22b.

3.4.1 Relation Extraction

Techniques for relationship extraction can be broadly divided into traditional and neural network-based methods. Traditional methodologies encompass manual, unsupervised, and supervised approaches. The manual approach [80] incorporates linguistic knowledge to build a linguistic model grounded in words, syntax, or semantics. This model subsequently matches preprocessed sentences and establishes the corresponding linguistic relation. On the other hand, the unsupervised methodology [81] aims to discover semantic relations by extracting entities and their contexts and grouping them based on similarities in contextual information. The supervised approach perceives relationship extraction as a classification problem and employs data-based features [82], derived from entity context semantics or syntax, to train classifiers for each related entity. In the testing phase, the classifier can identify the relation of a new entity if its features align. However, this method necessitates high-quality features. An alternative approach, as proposed by Mooney et al. [83], is to design a kernel function for classification.

Recent advancements in relation extraction techniques have led to the prevalent use of neural network-based methods. Liu et al. [84] pioneered applying the Convolutional Neural Network (CNN) model to relation extraction, transforming sentences into word embeddings through a synonym dictionary and other linguistic features. The model's output is the relationship classification probability between entities. Xu et al. [85] aimed to augment the semantic aspect of the methodology by using the shortest dependency path (SDP) and incorporating the central part of the sentence as input while omitting irrelevant words for improved accuracy. Zhang et al. [86] posited that relation extraction requires

Table 3.6: Different performance of BM25 based on different queries from YorkU's results of the TREC 2021 Deep Learning Track with the evaluation of NDCG@10.

Query ID	**Question**	**BM25**	**YorkU21a**
1129560	accounting definition of building improvements	0.3274	0.6423
168329	does light intensity or concentration of carbon dioxide have a higher rate of photosynthesis	0.7179	0.6937
225975	how does my baby get submitted for medicaid after birth	0.0000	0.4885

comprehensive and continuous information from all words in the sentence and utilized bi-directional long short-term memory networks (BLSTM) for sentence-level representation and feature enhancement. Zhou et al. [87] proposed applying an attention mechanism for BLSTM to extract essential features from the data without relying on external resources. However, additional knowledge or resources, such as knowledge graphs, may be necessary to represent the sentence comprehensively. Ji et al. [88] introduced the concept of relational vectors from knowledge graphs to represent the features of relationships. Qin et al. [89] focused on enhancing the dataset's quality by employing reinforcement learning to filter mislabeled sentences and reduce data noise, thus forming a new high-confidence training dataset that improves the performance of the trained model.

The results of the previous year's experiments suggest that BM25 is effective for queries containing special entities. As shown in Table 3.6, the query "168329" ("does light intensity or concentration of carbon dioxide have a higher rate of photosynthesis") is a semantically complex query that would typically be better suited to deep learning models. However, BM25 yields superior results. Despite the lack of a semantic relationship between the words in the query, BM25 can still identify relevant documents by the presence of several special entities, such as "light intensity", "carbon dioxide", and "photosynthesis". Notably, the entity "photosynthesis" is a term with limited synonyms, and BM25 tends to perform well in the presence of such entities.

On the other hand, the query "225975" ("how does my baby get submitted for medicaid after birth") yields an NDCG score of 0.00 in the top-10 hits for BM25 due to the absence of special entities. Therefore, we infer that entities such as "photosynthesis", "david w. taylor", "neuroma", and "levothyroxine" are strong entities, and BM25 tends to perform better with queries that include these strong entities.

Based on this analysis, the goal of relation extraction is set as the identification of strong entities. These entities can be classified into several categories, including names of individuals, locations, organizations, and specialized terms such as medical. In other words, these strong entities are the core keywords in the sentence.

We employ the YAKE model [90], a lightweight unsupervised automatic keyword extraction method that relies on statistical text features to select the most relevant keywords in the text. Additionally, we utilize Gensim and Rake-NLTK (Rapid Automatic Keyword Extraction algorithm with the NLTK toolkit) to identify strong entities. The extracted entities are deemed correct if their results match those of the YAKE model. Otherwise, the entity is discarded.

Two submissions were made in this year's full-passage ranking task: YorkU22a and YorkU22b. YorkU22b represents a non-dense retrieval approach and serves as the first stage of the dense retrieval process, YorkU22a. As this year's work builds upon last year's efforts, the same dense retrieval architecture is adopted. The Sentence-BERT model was utilized as the pre-training model, and a Bi-Encoder was applied to generate the dense retrieval's initial ranking result, YorkU22b. After performing relation extraction on the query, queries containing strong entities were identified, and their BM25 ranking results were computed using Anserini. The results for strong entity queries in YorkU22b were then replaced with their corresponding BM25 results, thereby forming a new initial ranking. Finally, YorkU22a was obtained using a Cross-Encoder, with the scope of documents constrained to those retrieved in the preceding ranking.

Table 3.7: Substitution Runs Descriptions of The TREC 2022 Deep Learning Track

Runs	**Description**
BM25	The BM25 baseline.
YorkU22b	The first ranking obtained from SBERT.
YorkU22b+BM25	The first ranking combined with BM25.
YorkU22a	The re-ranking based on the first ranking combined with BM25.
YorkU22a−BM25	The re-ranking based on the first ranking without combining BM25.
YorkU22a+BM25	The optimal result combines re-ranking and BM25.

Table 3.8: The Comparison Results with Different Substitution Runs and Baselines in The TREC 2022 Deep Learning Track.

Runs	**MAP@100**	**P@10**	**NDCG@10**	**NDCG@100**
BM25	0.0325	0.1421	0.2692	0.2133
YorkU22b	0.1130	0.3947	0.5076	0.3408
YorkU22b+BM25	0.1162	0.4066	0.5181	0.3471
YorkU22a−BM25	0.1989	0.5288	0.6003	0.4587
YorkU22a	0.2003	0.5316	0.6089	0.4610
YorkU22a+BM25	0.2011	0.5303	0.6144	0.4628

3.4.2 Experiment Results and Analysis

The experimental runs are denoted as BM25, YorkU22a, YorkU22a-BM25, YorkU22a+BM25, YorkU22b, and YorkU22b+BM25. These run descriptions are detailed in Table 3.7.

The experimental results for all runs are presented in Table 3.8. It can be observed that the traditional BM25 model, serving as the baseline, demonstrates significantly lower performance than the other models. However, in the initial-stage ranking, substituting the results of the deep learning model for strong entity queries with their BM25 results leads to performance enhancement. The re-ranking results of the dense retrieval that underwent this process also show improvement. An exploration into merging the re-ranked outcomes with BM25 was also conducted. Notably, even after re-ranking, YorkU22a's outcomes remain subpar relative to BM25. Direct substitutions with BM25 outcomes enhance the retrieval efficacy, as evidenced in the experimental result denoted as YorkU22a+BM25. This suggests

that the deep learning model fails to completely capture all the information gained from BM25 in the initial ranking, indicating a potential for optimization in the results after replacing the queries with strong entities.

In conclusion, while deep learning-based information retrieval models are generally acknowledged to outperform traditional BM25 models, BM25 remains a prevalent information retrieval algorithm due to its mathematical explanation of the retrieval process to a certain extent, which is lacking in deep learning models. On the other hand, training deep learning models is resource-intensive and requires large amounts of training data, often unavailable in fields such as biology and medicine. Utilizing a small-scale pre-trained model combined with traditional retrieval methods is feasible to improve performance. The proposed approach attempts to address this issue by employing relation extraction to identify strong entity queries and combining the results of the deep learning model with the BM25 results of such queries to achieve superior results.

3.5 Summary

The experiments conducted in the 2021 and 2022 Deep Learning Track have demonstrated the strengths and weaknesses of general language models in classical passage ranking tasks. While these models offer considerable advancements over traditional retrieval models like BM25, especially in recognizing semantic information, challenges remain.

Deep learning models are adept at fuzzy matching, allowing them to go beyond mere keyword matching to capture the underlying semantics of a query. This ability has significantly improved retrieval system performance in various domains.

Despite their capabilities, general language models have been found to underperform in certain specific tasks. Person name matching, a prevalent requirement in real search scenarios, is an example of where these models may fall short. Typically, pre-trained models like SBERT must be fine-tuned to adapt to new domains, an operation that can be resource and time-

intensive. Contrastingly, the BM25 algorithm offers a more straightforward approach, capable of being deployed across various scenarios without additional customization or overhead. Though lacking the advanced semantic recognition of deep learning models, BM25 remains a viable option in specific contexts.

The biomedical field presents unique challenges, such as lexical variants, which traditional models find challenging to handle. The scarcity of large annotated training datasets further complicates the development of domain-specific language models.

Given these observations, there is a growing interest in exploring ways to enhance the performance of general language models for specific domains without imposing high costs. The next chapter proposes a low-cost but effective approach to enhancing general language models for biomedical IR tasks.

Chapter 4

Diversified Prior Knowledge Enhanced General Language Model for Biomedical Text Retrieval

Biomedical Text Retrieval is a specialized field that focuses on searching and retrieving biomedical information from text sources, including scientific literature, clinical records, genomic data, etc. This area has gained significant attention due to the exponential growth of biomedical data and the critical need to access relevant information for research, clinical decision-making, drug discovery, and public health policy.

Unlike general IR, biomedical IR deals with highly specialized and complex information. The terminology is often domain-specific, and the relationships between concepts can be intricate. This complexity requires advanced techniques to understand the information's context, semantics, and structure. Traditional IR methods, based on term-matching, are often inadequate in this specialized field, suffering from failure modes, especially when dealing with terms that have different meanings in various contexts or when essential semantics from the question are not considered during retrieval.

Recent advancements in machine learning and natural language processing have opened new avenues for addressing these challenges. Developing domain-specific models like BioBERT has shown promising results in tasks like named entity recognition and relation extraction. Furthermore, integrating domain knowledge bases like MeSH has effectively bridged the semantic gap, providing more context-aware retrieval.

However, fine-tuning or retraining pre-trained models on domain-specific data remains a substantial challenge even with these advancements. The associated costs of time and resources can be prohibitive, particularly in the rapidly evolving field of biomedical research, where quick access to relevant information is paramount.

Our framework [8] emerges as an alternative solution, offering a cost-effective enhancement to existing general language models for biomedical IR . It builds on the strengths of existing pre-trained general models and addresses their limitations by incorporating a multi-stage approach. By leveraging both general language models and domain-specific knowledge, this framework adapts to the unique characteristics of biomedical text retrieval, providing a flexible and efficient solution.

4.1 Framework Architecture

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 4.1. The following sections will introduce the details of each component.

4.2 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

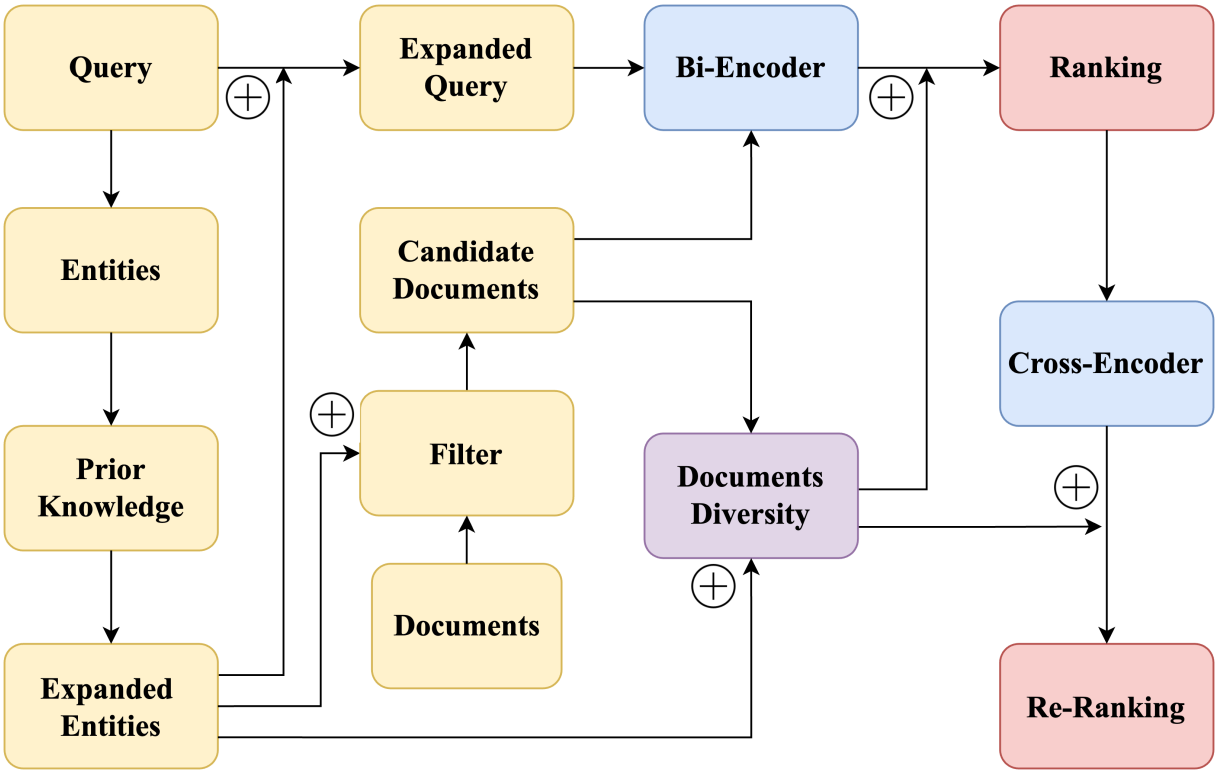


Figure 4.1: The architecture of our proposed framework.

comprehensive overview of the query and meet the user's information needs. To enrich query aspect diversity, we introduce prior knowledge from MeSH, UMLS, and NCBI.

MeSH (Medical Subject Headings) MeSH is a controlled vocabulary used by the National Library of Medicine (NLM) to index articles in MEDLINE and other databases. It consists of a hierarchical structure of medical terms, allowing for the classification and retrieval of biomedical literature according to specific subject headings.

UMLS (Unified Medical Language System) UMLS is a compendium of controlled vocabularies developed by the NLM. It integrates and standardizes various biomedical and healthcare terminologies, enabling the consistent representation of concepts across different

systems and applications. UMLS aims to facilitate information retrieval, data integration, and semantic interoperability within the biomedical domain.

NCBI (National Center for Biotechnology Information) NCBI is a leading research institution that provides access to various biological databases, tools, and resources. It hosts repositories such as GenBank for genetic sequences and PubMed for biomedical literature. NCBI's resources can be leveraged to obtain detailed information about genes, proteins, diseases, and other biological entities.

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, a term related to biomedical aspects is referred to as an entity e, and all expanded information are aspects of this entity.

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 to extract entities from the query:

$$E = \{e_2, ..., e_7, ...\}$$

Table 4.1 shows examples of the extracted entities. The extracted entities are then expanded with their synonyms and descriptions from prior knowledge sources:

$$E' = \{e'_2, e''_2, ..., e'_7, e''_7, ...\}$$

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.

Table 4.1: 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
What is the role of *MMS2* in cancer?	MMS2
What is the role of *Nurr-77* in *Parkinson's disease*?	Nurr-77, Parkinson's disease
How does *BARD1* regulate *BRCA1* activity?	BARD1, BRCA1
How does *p53* affect *apoptosis*?	p53, apoptosis
What [MUTATIONS] in the *Raf* gene are associated with cancer?	Raf
What [GENES] are involved in *insect segmentation*?	insect segmentation
Which [GENES] involved in *NFkappaB* signaling regulate *iNOS*?	NFkappaB, iNOS
What [TOXICITIES] are associated with *etidronate*?	etidronate
What [GENES] regulate *puberty* in humans?	puberty
What [TUMOR TYPES] are found in *zebrafish*?	zebrafish
Which [PATHWAYS] are mediated by *CD44*?	CD44

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 [91], 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, diversified aspects, including synonyms, descriptions, and various terminology representations. The output can be formulated below:

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

The Knowledge-based Query Expansion approach offers a sophisticated and tailored strategy for enhancing the quality and diversity of search results in biomedical IR. By thoughtfully integrating prior knowledge and addressing the unique characteristics of the biomedical domain, this approach contributes to a more effective and nuanced retrieval process. Whether applied to research, clinical practice, or other specialized contexts, the principles and methods outlined in this section hold promise for advancing the state-of-the-art in biomedical information retrieval.

4.3 Aspect-based Filter

General Language Models trained on large-scale datasets are often biased toward 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 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 lack 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 Chapter 6.

General Language Models, particularly those trained on extensive and diverse datasets, often demonstrate an inclination toward the specific domains present within their training data. This bias can lead to challenges when attempting to apply these models to specialized fields such as the biomedical domain, especially without domain-specific fine-tuning or retraining.

One intuitive approach to addressing this challenge is to filter out irrelevant documents from the candidate set. By focusing on a reduced range of documents that are more likely to be relevant, this method aims to enhance the quality of the ranking results. Moreover, limiting the scope of consideration to a smaller set of candidates reduces computational and time costs, offering a more practical and efficient solution.

The filtering process is guided by expanding the original query to encompass all highly-related aspects or facets. By identifying and targeting these specific aspects, documents that do not contain any identified aspects can be excluded from consideration. The remaining documents, which align more closely with the expanded query, are presumed to have a higher likelihood of relevance.

Potential Risks Despite its apparent benefits, this filtering approach is not without risks. Specifically, there may be documents containing implicit or unstated information that is relevant to the query but not captured by the identified aspects. Such documents could

be erroneously filtered out, leading to a loss of valuable information. This potential pitfall raises concerns that the retrieval results might be negatively affected by the absence of these relevant documents.

A series of experiments are conducted to assess the validity of these concerns and explore the efficacy of the filtering approach. These experiments are designed to probe the relationship between the filtering process, the expanded query, and the final ranking results. The objective is to determine whether the potential drawbacks of the filtering approach are substantiated in practice and whether the benefits outweigh the risks. The comprehensive details of these experiments, including results and interpretations, are presented in Chapter 6.

In summary, the described approach represents a strategic attempt to enhance the performance of General Language Models in the biomedical domain by narrowing the focus to a more relevant subset of documents. While promising in theory, the approach must be carefully implemented and evaluated to ensure it does not inadvertently exclude pertinent information. The experimental analysis serves as a critical component in this process, providing empirical insights into the method's effectiveness and potential limitations.

4.4 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 [52, 92]. 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 [67].

Bi-encoder and Cross-encoder Bi-encoders and Cross-encoders are methods used in natural language processing, specifically in transformer-based models like BERT, to encode text data. They are mainly used in tasks like sentence or document similarity, information retrieval, and question-answering systems.

A Bi-encoder model processes two inputs independently through a transformer encoder, then computes the final representations (embeddings) separately. These representations are usually compared for similarity using dot product or cosine similarity. Bi-encoders are models that independently encode each piece of text input. They are typically used for tasks where the relationship between two pieces of text does not depend on each other.

The working process of a Bi-encoder involves the following steps:

1. The input sequences (two pieces of text) are independently passed through the transformer encoder.

2. The encoder generates an embedding (vector representation) for each input sequence.

3. The similarity between these embeddings is calculated using a measure like a cosine similarity or dot product.

Let's denote the two input sequences as A and B, and their corresponding embeddings as e_A and e_B. The similarity is then calculated as follows:

Cosine similarity:

$$\text{similarity} = \frac{e_A \cdot e_B}{\|e_A\| \, \|e_B\|}$$

Dot product:

$$\text{similarity} = e_A \cdot e_B$$

A Cross-encoder, on the other hand, takes two inputs, combines them into a single input, and then processes this combined input through the transformer encoder to compute a single representation that encapsulates information from both inputs. Cross-encoders are models that encode pieces of text together. They are used when the relationship between two pieces of text needs to be understood in the context of each other.

The working process of a Cross-encoder involves the following steps:

1. The input sequences (two pieces of text) are combined into a single sequence, often with a special token (like "[SEP]" in BERT) to indicate the boundary between the two sequences.

2. This combined sequence is passed through the transformer encoder.

3. The encoder generates a single embedding representing the relationship between the two inputs.

4. This single embedding is typically passed through a classification layer to compute a score or category representing the relationship between the inputs.

Let's denote the combined input sequence as C and its embedding as e_C. If the model is being used to compute a similarity score, this score is often calculated using a dense layer with a single output unit:

$$\text{score} = W \cdot e_C + b$$

where W and b are the weights and biases of the dense layer.

Bi-encoders are often used in large-scale information retrieval systems. Each document in the database can be independently encoded into an embedding, and these embeddings can be stored for efficient nearest neighbor search. When a new query comes in, it is encoded into an embedding and compared with the stored embeddings to find the most similar documents.

Cross-encoders can provide more precise results by considering the query and document relationship. However, they are less efficient because they require combining each query with

each document and passing this combined input through the model. As a result, they are often used in a re-ranking step, where the top results from a more efficient method (like a Bi-encoder or keyword search) are re-ranked for precision.

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.

4.5 Diversity-based Score Reweighting

The similarity score in a Language Model-based Information Retrieval system, used to determine results ranking, can sometimes fall short of accuracy. This inaccuracy is especially prevalent when general Language Models are applied within a specific domain, and the quality of the embeddings affects the similarity score. A Diversity-based Score Reweighting method has been proposed to tackle this challenge, aiming to fine-tune the ranking results by considering the content's diversity.

In information retrieval, diversity signifies the extent to which a document encompasses various aspects of a query. When a query is expanded to include the diversified aspects of the original query, documents that address a broader spectrum of these aspects are more inclined to be relevant. The concept of diversity becomes crucial in instances where a query may have multiple interpretations or require information from several perspectives. The presence of diverse entities within a document serves as an indicator of its coverage of multiple aspects related to the query topic. This is then utilized to calculate a diversity score, capturing the richness and breadth of the content.

The Diversity-based Score Reweighting method is designed to combine the similarity score with the diversity score. By linearly combining these two metrics, the method seeks to enrich the ranking results, optimizing them through a delicate balance of weights. The ranking

results undergo continuous refinement, and after each ranking stage, they are re-sorted according to the combined score. This iterative process aims to ensure that the final ranked list provides a comprehensive and well-rounded response to the query, enhancing relevance and coverage.

Mathematically, the reweighting process can be expressed through the formulation:

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

where $\mathcal{V}$ denotes the diversity, symbolized by the number of unique entities within a document, and $\mathcal{S}$ symbolizes the similarity score about a specific query. The reweighted score, $\mathcal{S}_{re}$, is controlled by two hyperparameters: α modulates the balance between diversity and similarity score; ζ is designed to manage the weight of diversity.

The integration of diversity into the ranking mechanism acknowledges the multifaceted nature of queries and the importance of providing a well-rounded response. This approach not only improves the relevance of the retrieved documents but also enhances the robustness of the IR system, particularly when dealing with complex queries or specialized domains. By considering both similarity and diversity, the proposed method offers a more nuanced and adaptive way to rank documents, paving the way for more intelligent and context-aware information retrieval systems.

Chapter 5

Experimental Settings

5.1 Datasets

Experiments were conducted on public biomedical information retrieval datasets, specifically the TREC 2006 & 2007 Genomics Track (TREC-GENO) [93, 94]. These datasets are part of the Text REtrieval Conference (TREC) series and have been instrumental in promoting research in information retrieval. Utilizing these datasets leveraged the challenges and benchmarks within the genomics domain, allowing for evaluating and comparing various information retrieval techniques and models.

TREC-GENO TREC Genomics represents a specialized track within the TREC, targeted at the unique and complex challenges associated with the field of genomics. As a branch of molecular biology, genomics centers on the comprehensive study of genomes, encompassing the structure, function, evolution, and mapping of the entire genetic material within a cell or organism.

The initiation of the TREC Genomics Track responded to an emergent demand for advanced text retrieval and information extraction techniques within genomics. The rapid growth in biological data necessitated the development of refined tools to enable scientists to

manage, interpret, and assess this wealth of information.

In 2006, the challenges posed by the track included navigating the intricacy of biological terminology, resolving ambiguity in gene names, and accommodating the requirement to retrieve information at various levels of granularity, such as passages, documents, and aspects. The same document collection was retained in the subsequent 2007 Genomics Track, but modifications were made to the relevance criteria and evaluation methodologies. The emphasis continued to be on retrieving pertinent information across multiple dimensions, reflecting the diverse and multifaceted information needs inherent to the field of genomics.

The TREC-GENO document collection consists of a full-text biomedical corpus containing 162,259 documents from 49 genomics-related journals indexed by MEDLINE. The dataset includes 64 official topics from the biomedical domain, which were used as queries for evaluation. These topics are divided into 28 specific and 36 abstract categories, reflecting various aspects of genomics research. The official topics were manually crafted by biomedical domain experts and are presented in a question-answering format, thereby providing a realistic and challenging testbed for information retrieval methods in genomics.

Query Examples The queries are manually constructed by biomedical domain experts in a question-answering format, providing an authentic and complex challenge for information retrieval systems. Some examples include:

- **Specific Queries**:
  - What is the role of IDE in Alzheimer's disease?
  - How does Nurr-77 delete T cells before they migrate to the spleen or lymph nodes and how does this impact autoimmunity?
  - What is the role of Transforming growth factor-beta1 (TGF-beta1) in cerebral amyloid angiopathy (CAA)?

- **Abstract Queries**:
  - What [BIOLOGICAL SUBSTANCES] have been used to measure toxicity in response to zoledronic acid?
  - What [CELL OR TISSUE TYPES] express receptor binding sites for vasoactive intestinal peptide (VIP) on their cell surface?
  - What [GENES] when altered in the host genome improve solubility of heterologously expressed proteins?

Reasons for Selection TREC-GENO is chosen as the dataset for the experiments due to several distinct attributes:

- **Challenging Environment**: The lack of annotations and the limited number of queries with corresponding ground truth create a demanding environment for fine-tuning pre-trained Language Models.
- **Query Diversity**: The inclusion of two different query types (specific and abstract) adds complexity and variability to the dataset, influencing the application and evaluation of the proposed methods.
- **Standardized Evaluation**: The availability of official assessment metrics allows for consistent and standardized evaluation, particularly at the aspect level, aligning with the multidimensional information needs of genomics research.

Overall, the TREC-GENO dataset provides a comprehensive and rigorous testbed for information retrieval methods, reflecting the real-world challenges and intricacies of the genomics domain. Its diverse set of queries, ranging from specific questions to abstract concepts, facilitates in-depth analysis of the performance and capabilities of the retrieval system.

5.2 Evaluation Metrics

Within TREC Genomics 2006 and 2007, four distinct types of MAPs were assessed: Passage MAP, Passage2 MAP, Document MAP, and Aspect MAP. All of them are based on the Mean Average Precision (MAP) metric.

Mean Average Precision Mean Average Precision (MAP) is a widely used metric in Information Retrieval for evaluating the effectiveness of retrieval systems. It provides a single-figure measure of quality across recall levels.

In the context of IR, precision is the proportion of retrieved documents that are relevant, and recall is the proportion of relevant documents that are retrieved. Average Precision (AP) is the average of the precision values obtained for each relevant document retrieved up to the position of that document in the ranked list.

For example, if a system retrieves five documents, where the 1st and 4th documents are relevant, then the AP is $(1/1 + 2/4)/2 = 0.75$. The denominator is the total number of relevant documents, and the numerator is the sum of the precision at each point a relevant document is retrieved.

MAP is the mean (average) of the AP values for all queries or all users. It provides an overall measure of the effectiveness of a retrieval system across all queries or users. MAP is particularly useful when the relevant documents for each query are known, as in the case of benchmark datasets used in IR research.

One of the advantages of MAP is that it takes into account both the precision and recall of the retrieval system, and it also considers the ranking of the retrieved documents. A higher MAP value indicates a more effective retrieval system. However, one limitation of MAP is that it assumes all relevant documents are equally relevant, which may not be accurate in practice.

TREC-GENO Official Metrics

- **Document MAP**: It is a classical measure used to gauge the effectiveness of an information retrieval system at the document level. It computes the mean average precision across all retrieved documents, providing a comprehensive view of the system's performance. This measure considers each document as a whole unit. It is useful in scenarios where the relevance of information is evaluated at the document level, such as in document classification or ranking tasks.

- **Aspect MAP**: It is employed to evaluate the effectiveness of an information retrieval system in retrieving different aspects or facets of the information needed. An aspect can be perceived as a unique piece of information related to a query. For instance, for a query about a disease, different aspects could include symptoms, causes, treatments, etc. The Aspect MAP calculates the mean average precision across all aspects, thus offering a measure of the system's capacity to retrieve a diverse range of information related to the query. This is particularly beneficial in complex information needs where multiple pieces of information are required to address the query comprehensively.

- **Passage MAP**: It is used to evaluate the effectiveness of an information retrieval system at the passage level. A passage refers to a specific document subsection, such as a paragraph or a set of contiguous sentences. The Passage MAP calculates the mean average precision across all retrieved passages, providing a detailed view of the system's performance. This measure is handy when relevant information is embedded within specific sections of documents rather than distributed across the entire document.

- **Passage2 MAP**: It is a variant of the Passage MAP. Instead of treating each passage as an individual unit, the Passage2 MAP considers each character within a passage as a separate document. This method offers an even finer level of granularity, enabling thorough analysis of the system's precision at the character level. Such precision can

be crucial in tasks such as named entity recognition or other intricate information extraction tasks.

For easy distinction, the TREC-GENO was considered as two sub-task sets based on query types: Specific and Abstract. In addition, the NDCG metric was employed as an additional evaluation criterion to verify our methods from multiple perspectives.

Normalized Discounted Cumulative Gain Normalized Discounted Cumulative Gain (NDCG) is a popular metric used in Information Retrieval to evaluate the effectiveness of ranking systems, especially when the relevance of retrieved documents is graded (i.e., not binary).

The "gain" in NDCG refers to the relevance of a retrieved document. The "discounted" part implies that this gain is reduced logarithmically proportional to the position of the result. This means that highly relevant documents appearing lower in a search result list are penalized compared to those appearing higher. The "cumulative" part means the metric considers the entire list of retrieved documents up to a certain rank position.

The formula for DCG (Discounted Cumulative Gain) at a position p in the ranked list is:

$$\mathrm{DCG_p} = \sum_{i=1}^{p} \frac{rel_i}{\log_2(i+1)} = \mathrm{rel_1} + \sum_{i=2}^{p} \frac{rel_i}{\log_2(i+1)}$$

where rel_i is the relevance of the result at position i.

However, DCG values depend on the actual number of relevant documents and can vary across different queries or users. To make the metric more robust and comparable across different queries or users, DCG is often normalized by the Ideal DCG (IDCG), which is the DCG value of the ideal (i.e., best possible) ranking. This gives us the NDCG:

$$\mathrm{NDCG}_p = \mathrm{DCG}_p / \mathrm{IDCG}_p$$

NDCG values range from 0 to 1, with 1 being the best possible value. NDCG is particularly useful when the relevance of documents is a graded measure (e.g., on a scale from 1 to 3), rather than a binary measure (relevant or not relevant). It is widely used in fields like web search, recommender systems, and other ranking problems where the position of a relevant item in the ranked list matters.

Statistical Significances Statistical significances are tested by the two-tailed t-test with a significance level of 0.05.

5.3 Settings

A key component of the two-stage IR framework is the establishment of appropriate baselines for comparison and evaluation. Several models and methods are selected to provide a comprehensive understanding of the system's performance:

- **General Language Models:** BERT and RoBERTa are used as the Bi-encoder models for the initial ranking phase. These general language models provide a broad understanding of semantic relationships and serve as a standard against which other models can be compared.
- **Domain-Specific Models:** BioBERT and ClinicalBERT are included as additional Bi-encoder models for the initial ranking phase. Pre-trained on domain-specific data, these models are selected to investigate the potential benefits of specialized training in the biomedical context.
- **Re-ranking with SentenceBERT:** SentenceBERT (SBERT) is used as the re-ranker model in the second stage of the framework. Its pre-trained Cross-encoder model allows for efficient re-ranking, potentially enhancing the precision of the retrieval results.

- **Traditional Retrieval Methods:** The official TREC-GENO runs, including the Min, Median, Mean, and Max results (excluding the Min results in our experiments), are adopted as extra baselines. This allows for comparing the proposed language-model-based retrieval system and conventional retrieval approaches, thereby providing insights into the relative strengths and weaknesses of different methodologies.

By combining general language models, domain-specific models, and traditional retrieval methods, the baselines provide a multi-faceted perspective on the performance of the proposed two-stage IR framework. This comprehensive comparison ensures that the evaluation captures various aspects of retrieval effectiveness, offering valuable insights for further refinement and optimization of the system.

Implementation Details The experiments were carried out on Compute Canada's Cedar cluster, leveraging an NVIDIA P100 Volta GPU and three cores of the Intel E5-2650 v4 Broadwell CPU clocked at 2.2GHz.

The implementation of the framework involves a two-stage process: initial search and re-ranking. During the initial search, the top 2000 documents are extracted from the corpus based on their reweighting scores with respect to the given query. In the subsequent re-ranking process, these documents are further refined, and the top 1000 are retained as the final result. This two-step procedure allows for efficiently narrowing down relevant documents while ensuring that the diversity aspect is properly considered.

The selection of hyperparameters is a critical aspect of this approach, necessitating a careful balance to ensure that both semantic and diversity scores are appropriately weighted. The hyperparameter α plays a pivotal role in balancing the semantic and diversity scores, and it is confined to values between 0 and 1. A value close to 0 emphasizes the semantic score, while a value close to 1 puts more weight on the diversity score. The other hyperparameter, ζ, is used to control the weight assigned to the diversity score. It is also confined to the range between 0 and 1. A large value of ζ could potentially overpower the semantic score, rendering

it meaningless. Therefore, careful tuning of ζ is necessary to maintain the harmony between the two scores.

The hyperparameters were set to $\alpha = 0.2$ and $\zeta = 0.5$ in the conducted experiments. This particular configuration was found to yield the best performance within the framework, achieving an optimal balance between semantic relevance and diversity.

It is important to note that the choice of hyperparameters may vary depending on the specific nature of the corpus, queries, and the intended application of the retrieval system. Therefore, while the mentioned values provide a starting point, fine-tuning and experimentation might be required to adapt the method to different scenarios or domains.

Chapter 6

Experimental Results and Analysis

6.1 Experimental Results

Table 6.2 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 the queries' document relevance and aspect diversity. Simultaneously,

Table 6.1: 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**

improvements in document matching also benefit passage matching. 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 thesis.

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. As a general Language Model, we can see that BERT outperforms the domain-specific model ClinicalBERT in all metrics in Table 6.2. 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 6.2: Experiment results of our DPK-GLM and baselines on the TREC-GENO Specific & Abstract under official metrics. The superscript "*" means the method is significantly better than the best baseline.

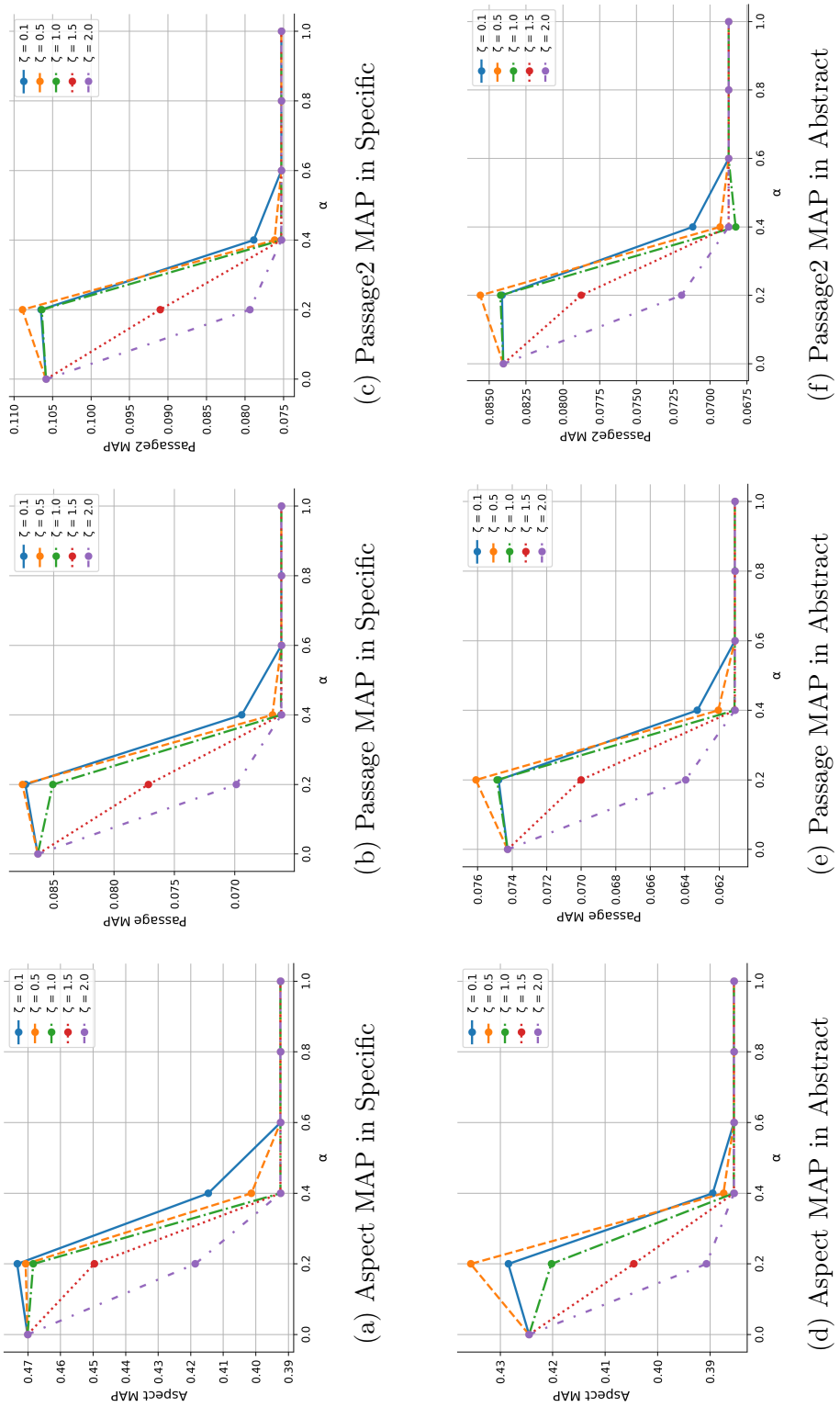
	TREC-GENO Specific				TREC-GENO Abstract			
Methods	Document	Aspect	Passage	Passage2	Document	Aspect	Passage	Passage2
TREC Median	0.3083	0.1581	0.0316	0.0345	0.1897	0.1311	0.0565	0.0377
TREC Mean	0.2887	0.1643	0.0347	0.0392	0.1862	0.1326	0.0560	0.0398
TREC Max	0.5439	0.4411	0.1012	0.1486	0.3286	0.2631	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 6.3 presents the performance of our framework in terms of the NDCG metric. Similar to Table 6.2, our framework significantly outperforms the baselines. Notably, our approach performs better on the TREC-GENO Abstract than the TREC-GENO Specific under the NDCG metric. In addition, as shown in Table 6.2, our best approach, DPK-GLM-RoBERTa+SBERT, achieves a remarkably higher Aspect MAP on the Abstract task compared to 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 6.1, we can address the hypothesis raised earlier: 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 was missed after filtering, which indicates that our Filter has a good performance. 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 6.2 and Table 6.3, even with the absence of a small number of relevant documents, our method can still achieve good results.

Table 6.3: Experiment results of our DPK-GLM and baselines on the TREC-GENO Specific & Abstract under 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***



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

6.2 Ablation Study

To further investigate the effectiveness of our approach, we conducted an ablation study on the TREC-GENO dataset. The Knowledge-based Query Expansion method, the Aspect-based Filter, and the Diversity-based Score Reweighting method from the framework are independently removed, and the results are compared with the full framework.

Table 6.4 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 the text scope and increase the likelihood of finding relevant documents. As the experimental results demonstrate, our method 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 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 general Language Model training data, 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.

Table 6.4: 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

Chapter 7

Conclusions and Future Work

7.1 Conclusions

Deep learning models have undeniably changed the Information Retrieval (IR) field, introducing innovative techniques and methodologies that have significantly enhanced the efficiency and accuracy of search systems. By leveraging complex neural network architectures, deep learning models have enabled the extraction of intricate patterns and relationships within data, leading to more accurate and context-aware retrieval results.

However, as we have explored, these models are not without their challenges. The need for extensive labeled data, computational resources, and careful hyperparameter tuning can pose barriers to their implementation. The lack of interpretability and challenges in transferability across different domains or tasks further complicate their application in diverse IR scenarios.

This thesis introduces a novel framework, DPK-GLM, designed to address these challenges and improve the effectiveness of general Language Models in biomedical IR. The proposed framework leverages domain knowledge and incorporates a series of strategic steps to enhance retrieval.

In the first stage, DPK-GLM employs a Knowledge-based Query Expansion method.

Using prior knowledge from biomedical databases such as MeSH, UMLS, and NCBI, queries are enriched with domain-specific entities. This expansion enables the model to recognize more diverse and intricate relationships within the query, making it more responsive to the multifaceted nature of biomedical research questions.

Next, the framework introduces an Aspect-based Filter to remove documents that are irrelevant to the query. This focused filtering process ensures that only documents with a high likelihood of relevance are retained, reducing computational costs and increasing the precision of the search results.

The final component of DPK-GLM is the Diversity-based Score Reweighting method. This approach re-sorts the initial ranking results by linearly combining diversity and similarity scores. The resulting ranking reflects both the semantic relationship between the query and the documents and the diverse aspects covered by the documents. By balancing these two dimensions, the model provides a more comprehensive response to the query.

DPK-GLM was evaluated using popular models like BERT and RoBERTa on the public biomedical IR dataset, TREC-GENO. The results demonstrate a significant improvement in retrieval performance, validating the framework's effectiveness. The innovative combination of query expansion, filtering, and reweighting techniques proved to be a powerful strategy for addressing the unique challenges of biomedical IR.

The proposed DPK-GLM framework represents a feasible strategy in applying general Language Models to biomedical IR. Integrating domain-specific knowledge and employing a multi-stage approach successfully navigates the complexities of biomedical queries, providing more accurate and relevant results. While deep learning models have brought unprecedented capabilities to the field of IR, the DPK-GLM framework illustrates that there is still room for innovation and refinement, especially in specialized domains.

7.2 Future work

Future work may explore further optimizations and applications of this framework, potentially extending its benefits to other datasets and specialized fields. A particular area of interest could be fine-tuning the Knowledge-based Query Expansion to avoid topic redundancy and ensure that expanded queries maintain relevance and accuracy. There could also be opportunities for enhancing the Aspect-based Filter by exploring hybrid filtering approaches or alternative methods that might lead to more precise document selection.

An interesting extension might be adapting DPK-GLM for domains like legal studies, which struggle with similar complexities in queries and domain-specific knowledge. Furthermore, as Generative AI and models like GPT-4 and Med-PaLM gain traction, their compatibility with DPK-GLM could be an avenue worth treading. Techniques like contrastive learning could be integrated to improve the model's ability to distinguish between similar yet different concepts, adding a layer of refinement to the query results [95]. Reinforcement learning approaches might enable the system to adapt and optimize its performance over time, learning from user interactions and feedback [96]. The use of Graph Neural Networks (GNNs) could help in capturing the complex relationships between legal or medical terms and concepts, enhancing the search capabilities. Another promising direction could be the exploration of our method on new datasets like TREC-PM [97] and MedQuAD [98], which would provide valuable insights into the system's adaptability and effectiveness across different medical sub-domains. Incorporating the ICD disease classification code could also add another layer of specificity to the searches. Automated mechanisms for annotations can further streamline the labeling process, making the framework more accessible.

Lastly, assessing DPK-GLM's scalability for larger datasets and its potential optimizations can lead to a more resource-efficient model apt for real-world implementations. The continued exploration in these directions could make search engines more intelligent and user-centered, contributing to the ongoing innovation in the field of IR, especially in specialized domains.

Bibliography

[1] Jingtao Zhan et al. "An analysis of BERT in document ranking". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2020, pp. 1941–1944.

[2] Jacob Devlin et al. "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*. Association for Computational Linguistics, 2019, pp. 4171–4186.

[3] Jinhyuk Lee et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". In: *Bioinformatics* 36.4 (2020), pp. 1234–1240.

[4] Kexin Huang, Jaan Altosaar, and R. Ranganath. "ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission". In: *ARXIV.ORG* (2019).

[5] Man Luo et al. "Improving biomedical information retrieval with neural retrievers". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 10. 2022, pp. 11038–11046.

[6] Xiaoshi Yin et al. "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*. 2010, pp. 901–902.

[7] 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*. ACM, 2009, pp. 307–314.

[8] Yizheng Huang and Jimmy Huang. "Diversified Prior Knowledge Enhanced General Language Model for Biomedical Information Retrieval". In: *ECAI 2023*. IOS Press, 2023.

[9] Yinhan Liu et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *ARXIV.ORG* (2019).

[10] Yang Liu et al. "ARSA: a sentiment-aware model for predicting sales performance using blogs". In: *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007*. Ed. by Wessel Kraaij et al. ACM, 2007, pp. 607–614.

[11] Yang Liu et al. "Modeling and Predicting the Helpfulness of Online Reviews". In: *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), December 15-19, 2008, Pisa, Italy*. IEEE Computer Society, 2008, pp. 443–452.

[12] Tomas Mikolov et al. "Efficient Estimation of Word Representations in Vector Space". In: *International Conference On Learning Representations* (2013).

[13] Gerard Salton, Edward A Fox, and Harry Wu. "Extended boolean information retrieval". In: *Communications of the ACM* 26.11 (1983), pp. 1022–1036.

[14] Gloria Bordogna and Gabriella Pasi. "A fuzzy linguistic approach generalizing boolean information retrieval: A model and its evaluation". In: *Journal of the American Society for Information Science* 44.2 (1993), pp. 70–82.

[15] Akiko Aizawa. "An information-theoretic perspective of tf–idf measures". In: *Information Processing & Management* 39.1 (2003), pp. 45–65.

[16] Juan Ramos et al. "Using tf-idf to determine word relevance in document queries". In: *Proceedings of the first instructional conference on machine learning*. Vol. 242. 1. Citeseer. 2003, pp. 29–48.

[17] Fei Song and W Bruce Croft. "A general language model for information retrieval". In: *Proceedings of the eighth international conference on Information and knowledge management*. 1999, pp. 316–321.

[18] Chengxiang Zhai and John Lafferty. "A study of smoothing methods for language models applied to ad hoc information retrieval". In: *ACM SIGIR Forum*. Vol. 51. 2. ACM New York, NY, USA. 2017, pp. 268–276.

[19] Stephen E Robertson and K Sparck Jones. "Relevance weighting of search terms". In: *Journal of the American Society for Information science* 27.3 (1976), pp. 129–146.

[20] Stephen E. Robertson et al. "Okapi at TREC-4". In: *Proceedings of The Fourth Text REtrieval Conference, TREC 1995, Gaithersburg, Maryland, USA, November 1-3, 1995*. Vol. 500-236. NIST Special Publication. National Institute of Standards and Technology, 1995.

[21] Tie-Yan Liu et al. "Learning to rank for information retrieval". In: *Foundations and Trends® in Information Retrieval* 3.3 (2009), pp. 225–331.

[22] Ping Li, Qiang Wu, and Christopher Burges. "Mcrank: Learning to rank using multiple classification and gradient boosting". In: *Advances in neural information processing systems* 20 (2007).

[23] Yoav Freund et al. "An efficient boosting algorithm for combining preferences". In: *Journal of machine learning research* 4.Nov (2003), pp. 933–969.

[24] Chris Burges et al. "Learning to rank using gradient descent". In: *Proceedings of the 22nd international conference on Machine learning*. 2005, pp. 89–96.

[25] Hamed Valizadegan et al. "Learning to rank by optimizing ndcg measure". In: *Advances in neural information processing systems* 22 (2009).

[26] Yining Wang et al. "A theoretical analysis of NDCG type ranking measures". In: *Conference on learning theory*. PMLR. 2013, pp. 25–54.

[27] Christopher Burges, Robert Ragno, and Quoc Le. "Learning to rank with nonsmooth cost functions". In: *Advances in neural information processing systems* 19 (2006).

[28] Christopher JC Burges. "From ranknet to lambdarank to lambdamart: An overview". In: *Learning* 11.23-581 (2010), p. 81.

[29] Qiang Wu et al. "Adapting boosting for information retrieval measures". In: *Information Retrieval* 13 (2010), pp. 254–270.

[30] Hamid Palangi et al. "Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.4 (2016), pp. 694–707.

[31] Liang Pang et al. "Deeprank: A new deep architecture for relevance ranking in information retrieval". In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 2017, pp. 257–266.

[32] Omar Khattab and Matei Zaharia. "Colbert: Efficient and effective passage search via contextualized late interaction over bert". In: *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 2020, pp. 39–48.

[33] Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. *Pretrained transformers for text ranking: Bert and beyond*. Springer Nature, 2022.

[34] Nick Craswell et al. "Overview of the TREC 2019 deep learning track". In: *Proceedings of The 28th Text REtrieval Conference (TREC 2019)*. NIST. TREC, 2020.

[35] A. Sherstinsky. "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network". In: *Physica A: Statistical Mechanics And Its Applications* (2018). DOI: 10.1016/j.physd.2019.132306.

[36] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[37] Keiron O'Shea and Ryan Nash. "An Introduction to Convolutional Neural Networks". In: *arXiv preprint arXiv: 1511.08458* (2015).

[38] Jeffrey Pennington, Richard Socher, and Christopher D Manning. "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543.

[39] Wei Yang, Haotian Zhang, and Jimmy Lin. "Simple Applications of BERT for Ad Hoc Document Retrieval". In: *arXiv preprint arXiv: 1903.10972* (2019).

[40] Sean MacAvaney et al. "CEDR: Contextualized embeddings for document ranking". In: *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*. 2019, pp. 1101–1104.

[41] Shijie Wu and Mark Dredze. "Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT". In: *Conference on Empirical Methods in Natural Language Processing* (2019).

[42] Karthikeyan K et al. "Cross-Lingual Ability of Multilingual BERT: An Empirical Study". In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL: https://openreview.net/forum?id=HJeT3yrtDr.

[43] Colin Raffel et al. "Exploring the limits of transfer learning with a unified text-to-text transformer". In: *The Journal of Machine Learning Research* 21.1 (2020), pp. 5485–5551.

[44] Souradip Chakraborty et al. "BioMedBERT: A pre-trained biomedical language model for QA and IR". In: *Proceedings of the 28th International Conference on Computational Linguistics*. 2020.

[45] Lisa Kühnel and Juliane Fluck. "We are not ready yet: limitations of state-of-the-art disease named entity recognizers". In: *Journal of Biomedical Semantics* 13.1 (2022), p. 26.

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

[47] Chih-Hsuan Wei et al. "Biomedical mention disambiguation using a deep learning approach". In: *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. 2019, pp. 307–313.

[48] Nicholas C Ide, Russell F Loane, and Dina Demner-Fushman. "Essie: a concept-based search engine for structured biomedical text". In: *Journal of the American Medical Informatics Association* 14.3 (2007), pp. 253–263.

[49] Sérgio Matos et al. "Concept-based query expansion for retrieving gene related publications from MEDLINE". In: *BMC bioinformatics* 11 (2010), pp. 1–9.

[50] Bevan Koopman et al. "Information retrieval as semantic inference: A graph inference model applied to medical search". In: *Information Retrieval Journal* 19 (2016), pp. 6–37.

[51] 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.

[52] Qiao Jin et al. "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.

[53] Maciej Rybinski Sarvnaz Karimi. "CSIROmed at TREC Precision Medicine 2020". In: *Proceedings of The 30th Text REtrieval Conference (TREC 2021)*. 2021.

[54] 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*. 2017, pp. 2307–2310.

[55] 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*. Vol. 34. 2021.

[56] Hamed Zamani and W Bruce Croft. "Estimating embedding vectors for queries". In: *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*. 2016, pp. 123–132.

[57] Saizheng Zhang et al. "Personalizing Dialogue Agents: I have a dog, do you have pets too?" In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018, pp. 2204–2213.

[58] Ying Luo, Fengshun Xiao, and Hai Zhao. "Hierarchical contextualized representation for named entity recognition". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 05. 2020, pp. 8441–8448.

[59] Yonghui Wu et al. "Google's neural machine translation system: Bridging the gap between human and machine translation". In: *arXiv preprint arXiv:1609.08144* (2016).

[60] Anh Nguyen, Jason Yosinski, and Jeff Clune. "Understanding neural networks via feature visualization: A survey". In: *Explainable AI: interpreting, explaining and visualizing deep learning* (2019), pp. 55–76.

[61] Yizheng Huang and Jimmy Huang. "York University at TREC 2021: Deep Learning Track". In: *Proceedings of The 30th Text REtrieval Conference (TREC 2021)*. 2021.

[62] Yizheng Huang and Jimmy Huang. "York University at TREC 2022: Deep Learning Track". In: *Proceedings of The 31st Text REtrieval Conference (TREC 2022)*. 2022.

[63] Daniel Fernando Campos et al. "MS MARCO: A Human Generated MAchine Reading COmprehension Dataset". In: *COCO@NIPS* (2016).

[64] Nick Craswell et al. "Overview of the TREC 2020 deep learning track". In: *Proceedings of The 29th Text REtrieval Conference (TREC 2020)*. TREC, 2020.

[65] Nick Craswell et al. "Overview of the TREC 2021 deep learning track". In: *Proceedings of The 30th Text REtrieval Conference (TREC 2021)*. NIST. TREC, 2021.

[66] Nick Craswell et al. "Overview of the TREC 2022 deep learning track". In: *Proceedings of The 31st Text REtrieval Conference (TREC 2022)*. NIST. TREC, 2022.

[67] 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)*. 2019, pp. 3982–3992.

[68] Peilin Yang, Hui Fang, and Jimmy Lin. "Anserini: Enabling the use of lucene for information retrieval research". In: *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*. 2017, pp. 1253–1256.

[69] Edward A Fox and Joseph A Shaw. "Combination of multiple searches". In: *NIST special publication SP* 243 (1994).

[70] Vladimir Karpukhin et al. "Dense Passage Retrieval for Open-Domain Question Answering". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 6769–6781. DOI: 10.18653/v1/2020.emnlp-main.550. URL: https://aclanthology.org/2020.emnlp-main.550.

[71] Md. Tahmid Rahman Laskar, Jimmy Xiangji Huang, and Enamul Hoque. "Contextualized Embeddings based Transformer Encoder for Sentence Similarity Modeling in Answer Selection Task". In: *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*. Ed. by Nicoletta Calzolari et al. European Language Resources Association, 2020, pp. 5505–5514. URL: https://aclanthology.org/2020.lrec-1.676/.

[72] Ben He, Jimmy Xiangji Huang, and Xiaofeng Zhou. "Modeling term proximity for probabilistic information retrieval models". In: *Inf. Sci.* 181.14 (2011), pp. 3017–3031. DOI: 10.1016/j.ins.2011.03.007. URL: https://doi.org/10.1016/j.ins.2011.03.007.

[73] Xiangji Huang, Ming Zhong, and Luo Si. "York University at TREC 2005: Genomics Track". In: *Proceedings of the Fourteenth Text REtrieval Conference, TREC 2005, Gaithersburg, Maryland, USA, November 15-18, 2005*. Ed. by Ellen M. Voorhees and Lori P. Buckland. Vol. 500-266. NIST Special Publication. National Institute of Standards and Technology (NIST), 2005. URL: http://trec.nist.gov/pubs/trec14/papers/yorku-huang2.geo.pdf.

[74] Xiangji Huang et al. "Applying Machine Learning to Text Segmentation for Information Retrieval". In: *Inf. Retr.* 6.3-4 (2003), pp. 333–362. DOI: 10.1023/A:1026028229881. URL: https://doi.org/10.1023/A:1026028229881.

[75] Jiashu Zhao, Jimmy Xiangji Huang, and Ben He. "CRTER: using cross terms to enhance probabilistic information retrieval". In: *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*. Ed. by Wei-Ying Ma et al. ACM, 2011, pp. 155–164. DOI: 10.1145/2009916.2009941. URL: https://doi.org/10.1145/2009916.2009941.

[76] Nandan Thakur et al. "BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models". In: *Neurips Datasets And Benchmarks* (2021).

[77] Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. "Question and Answer Test-Train Overlap in Open-Domain Question Answering Datasets". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, Apr. 2021, pp. 1000–1008. DOI: 10.18653/v1/2021.eacl-main.86. URL: https://aclanthology.org/2021.eacl-main.86.

[78] Anthony Chen et al. "Evaluating Entity Disambiguation and the Role of Popularity in Retrieval-Based NLP". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 4472–4485. DOI: 10.18653/v1/2021.acl-long.345. URL: https://aclanthology.org/2021.acl-long.345.

[79] Christopher Sciavolino et al. "Simple Entity-Centric Questions Challenge Dense Retrievers". In: *2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*. Association for Computational Linguistics (ACL). 2021, pp. 6138–6148.

[80] Marti A Hearst. "Automatic acquisition of hyponyms from large text corpora". In: *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*. 1992.

[81] Takaaki Hasegawa, Satoshi Sekine, and Ralph Grishman. "Discovering relations among named entities from large corpora". In: *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*. 2004, pp. 415–422.

[82] Bryan Rink and Sanda Harabagiu. "Utd: Classifying semantic relations by combining lexical and semantic resources". In: *Proceedings of the 5th international workshop on semantic evaluation*. 2010, pp. 256–259.

[83] Raymond Mooney and Razvan Bunescu. "Subsequence kernels for relation extraction". In: *Advances in neural information processing systems* 18 (2005).

[84] ChunYang Liu et al. "Convolution neural network for relation extraction". In: *Proceedings of the Advanced Data Mining and Applications: 9th International Conference, ADMA 2013, Hangzhou, China, December 14-16, 2013, Proceedings, Part II 9*. Springer. 2013, pp. 231–242.

[85] Kun Xu et al. "Semantic Relation Classification via Convolutional Neural Networks with Simple Negative Sampling". In: *Conference On Empirical Methods In Natural Language Processing* (2015). DOI: 10.18653/v1/D15-1062.

[86] Shu Zhang et al. "Bidirectional long short-term memory networks for relation classification". In: *Proceedings of the 29th Pacific Asia conference on language, information and computation*. 2015, pp. 73–78.

[87] Peng Zhou et al. "Attention-based bidirectional long short-term memory networks for relation classification". In: *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*. 2016, pp. 207–212.

[88] Guoliang Ji et al. "Distant supervision for relation extraction with sentence-level attention and entity descriptions". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 31. 1. 2017.

[89] Pengda Qin, Weiran Xu, and William Yang Wang. "Robust Distant Supervision Relation Extraction via Deep Reinforcement Learning". In: *Annual Meeting Of The Association For Computational Linguistics* (2018). DOI: 10.18653/v1/P18-1199.

[90] Ricardo Campos et al. "YAKE! Keyword extraction from single documents using multiple local features". In: *Information Sciences* 509 (2020), pp. 257–289.

[91] 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.

[92] Jiafeng Guo et al. "A deep look into neural ranking models for information retrieval". In: *Information Processing & Management* 57.6 (2020), p. 102067.

[93] William R Hersh et al. "TREC 2006 genomics track overview." In: *Proceedings of The Fifteenth Text REtrieval Conference (TREC 2006)*. 2006.

[94] William Hersh et al. "TREC 2007 genomics track overview". In: *Proceedings of The Sixteenth Text REtrieval Conference (TREC 2007)*. 2007.

[95] Lianghao Xia et al. "Hypergraph Contrastive Collaborative Filtering". In: *Proceedings of The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*. Ed. by Enrique Amigó et al. ACM, 2022, pp. 70–79.

[96] Pengfei Wang et al. "KERL: A Knowledge-Guided Reinforcement Learning Model for Sequential Recommendation". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, China, July 25-30, 2020*. Ed. by Jimmy X. Huang et al. ACM, 2020, pp. 209–218.

[97] Kirk Roberts et al. "Overview of the TREC 2020 precision medicine track". In: *Proceedings of The 29th Text REtrieval Conference (TREC 2020)*. 2020.

[98] Asma Ben Abacha and Dina Demner-Fushman. "A Question-Entailment Approach to Question Answering". In: *BMC Bioinform.* 20.1 (2019), 511:1–511:23.

Appendix A

Published Papers

1. Yizheng Huang and Jimmy Huang. "Diversified Prior Knowledge Enhanced General Language Model for Biomedical Information Retrieval". In: *ECAI 2023*. IOS Press, 2023.

2. Yizheng Huang. "A Noise-enhanced Fuse Model for Passage Ranking". In: *2022 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. IEEE, 2022.

3. Yizheng Huang and Li Zeng. "Multiple Linear Combination Approaches for Information Search in Ranking". In: *2022 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. IEEE, 2022.

4. Yizheng Huang and Jimmy Huang. "York University at TREC 2022: Deep Learning Track". In: *Proceedings of The 31st Text REtrieval Conference (TREC 2022)*. NIST. 2022.

5. Yizheng Huang and Jimmy Huang. "York University at TREC 2021: Deep Learning Track". In: *Proceedings of The 30th Text REtrieval Conference (TREC 2021)*. NIST. 2021.

Appendix B

TREC Genomics 2006 Queries

Table B.1: The TREC Genomics 2006 Queries (TREC-GENO Specific)

ID	Query
160	What is the role of PrnP in mad cow disease?
161	What is the role of IDE in Alzheimer's disease
162	What is the role of MMS2 in cancer?
163	What is the role of APC (adenomatous polyposis coli) in colon cancer?
164	What is the role of Nurr-77 in Parkinson's disease?
165	How do Cathepsin D (CTSD) and apolipoprotein E (ApoE) interactions contribute to Alzheimer's disease?
166	What is the role of Transforming growth factor-beta1 (TGF-beta1) in cerebral amyloid angiopathy (CAA)?
167	How does nucleoside diphosphate kinase (NM23) contribute to tumor progression?
168	How does BARD1 regulate BRCA1 activity?
169	How does APC (adenomatous polyposis coli) protein affect actin assembly
170	How does COP2 contribute to CFTR export from the endoplasmic reticulum?

171	How does Nurr-77 delete T cells before they migrate to the spleen or lymph nodes and how does this impact autoimmunity?
172	How does p53 affect apoptosis?
173	How do alpha7 nicotinic receptor subunits affect ethanol metabolism?
174	How does BRCA1 ubiquitinating activity contribute to cancer?
175	How does L2 interact with L1 to form HPV11 viral capsids?
176	How does Sec61-mediated CFTR degradation contribute to cystic fibrosis?
177	How do Bop-Pes interactions affect cell growth?
178	How do interactions between insulin-like GFs and the insulin receptor affect skin biology?
179	How do interactions between HNF4 and COUP-TF1 suppress liver function?
180	How do Ret-GDNF interactions affect liver development?
181	How do mutations in the Huntingtin gene affect Huntington's disease?
182	How do mutations in Sonic Hedgehog genes affect developmental disorders?
183	How do mutations in the NM23 gene affect tracheal development?
184	How do mutations in the Pes gene affect cell growth?
185	How do mutations in the hypocretin receptor 2 gene affect narcolepsy?
186	How do mutations in the Presenilin-1 gene affect Alzheimer's disease?
187	How do familial hemiplegic migraine type 1 (FHM1) gene mutations affect calcium ion influx in hippocampal neurons?

Appendix C

TREC Genomics 2007 Queries

Table C.1: The TREC Genomics 2007 Queries (TREC-GENO Abstract)

ID	Query
200	What serum [PROTEINS] change expression in association with high disease activity in lupus?
201	What [MUTATIONS] in the Raf gene are associated with cancer?
202	What [DRUGS] are associated with lysosomal abnormalities in the nervous system?
203	What [CELL OR TISSUE TYPES] express receptor binding sites for vasoactive intestinal peptide (VIP) on their cell surface?
204	What nervous system [CELL OR TISSUE TYPES] synthesize neurosteroids in the brain?
205	What [SIGNS OR SYMPTOMS] of anxiety disorder are related to coronary artery disease?
206	What [TOXICITIES] are associated with zoledronic acid?
207	What [TOXICITIES] are associated with etidronate?

208	What [BIOLOGICAL SUBSTANCES] have been used to measure toxicity in response to zoledronic acid?
209	What [BIOLOGICAL SUBSTANCES] have been used to measure toxicity in response to etidronate?
210	What [MOLECULAR FUNCTIONS] are attributed to glycan modification?
211	What [ANTIBODIES] have been used to detect protein PSD-95?
212	What [GENES] are involved in insect segmentation?
213	What [GENES] are involved in Drosophila neuroblast development?
214	What [GENES] are involved axon guidance in C.elegans?
215	What [PROTEINS] are involved in actin polymerization in smooth muscle?
216	What [GENES] regulate puberty in humans?
217	What [PROTEINS] in rats perform functions different from those of their human homologs?
218	What [GENES] are implicated in regulating alcohol preference?
219	In what [DISEASES] of brain development do centrosomal genes play a role?
220	What [PROTEINS] are involved in the activation or recognition mechanism for PmrD?
221	Which [PATHWAYS] are mediated by CD44?
222	What [MOLECULAR FUNCTIONS] is LITAF involved in?
223	Which anaerobic bacterial [STRAINS] are resistant to Vancomycin?
224	What [GENES] are involved in the melanogenesis of human lung cancers?
225	What [BIOLOGICAL SUBSTANCES] induce clpQ expression?
226	What [PROTEINS] make up the murine signal recognition particle?
227	What [GENES] are induced by LPS in diabetic mice?
228	What [GENES] when altered in the host genome improve solubility of heterologously expressed proteins?

229	What [SIGNS OR SYMPTOMS] are caused by human parvovirus infection?
230	What [PATHWAYS] are involved in Ewing's sarcoma?
231	What [TUMOR TYPES] are found in zebrafish?
232	What [DRUGS] inhibit HIV type 1 infection?
233	What viral [GENES] affect membrane fusion during HIV infection?
234	What [GENES] make up the NFkappaB signaling pathway?
235	Which [GENES] involved in NFkappaB signaling regulate iNOS?

Appendix D

Samples of The TREC Genomics Dataset

```
<html>
<body>
<H2>
John Snow and Modern-Day Environmental Epidemiology
</H2>
<STRONG>
</NOBR><NOBR>Dale P. Sandler<SUP></SUP></NOBR>
</STRONG><P>
<FONT SIZE=-1>
From the Epidemiology Branch, National Institute of Environmental Health Sciences, P.O. Box 12233&#151;Mail Drop A3-05, 111 T. W. Alexander Drive,
      Research Triangle Park, NC 27709.
</FONT><P>
What does an anecdote about John Snow have to do with modern-day<SUP> </SUP>epidemiology? And why use it to introduce an issue of the <I>Journal</I
      ><SUP> </SUP>highlighting the challenges of studying disease risks associated<SUP> </SUP>with low dose environmental exposures?<SUP> </SUP><P
      >
In this issue, Lilienfeld describes John Snow giving expert-witness<SUP> </SUP>testimony on behalf of industry (1<A HREF="#B1"><IMG BORDER=1 WIDTH
      =8 HEIGHT=7 ALT="Go"
SRC="/icons/fig-down.gif"></A>). Besides being interesting<SUP> </SUP>on a historical basis, this incident raises several issues that<SUP> </SUP>
      are pertinent today. Lilienfeld's paper and the accompanying<SUP> </SUP>commentary by Vandenbroucke (2<A HREF="#B2"><IMG BORDER=1 WIDTH=8
      HEIGHT=7 ALT="Go"
SRC="/icons/fig-down.gif"></A>) deal directly or indirectly<SUP> </SUP>with the role and responsibilities of expert witnesses, the<SUP> </SUP>
      extrapolation of data on health effects from high dose exposures<SUP> </SUP>to low dose exposures, the importance of epidemiology to the<SUP>
       </SUP>development of public health policy, the current debates on<SUP> </SUP>environmental justice (3<A HREF="#B3"><IMG BORDER=1 WIDTH=8
      HEIGHT=7 ALT="Go"
SRC="/icons/fig-down.gif"></A>), and the use of the precautionary<SUP> </SUP>principle (4<A HREF="#B4"><IMG BORDER=1 WIDTH=8 HEIGHT=7 ALT="Go"
SRC="/icons/fig-down.gif"></A>) in standard-setting. Furthermore, if faced with<SUP> </SUP>an issue similar to that faced by Snow&#151;namely,
      local residents'<SUP> </SUP>being worried about health consequences associated with emanations<SUP> </SUP>from factories&#151;would modern-
      day environmental epidemiologists<SUP> </SUP>be any better positioned to carry out appropriate studies and<SUP> </SUP>reach sound conclusions
      ?<SUP> </SUP><P>
Snow can be seen at once as victim and perpetrator of sins that<SUP> </SUP>are common in epidemiology in general and in environmental epidemiology<
      SUP> </SUP>in particular. Was Snow victimized by the medical establishment,<SUP> </SUP>including <I>The Lancet</I>, for expressing views that
       were not commonly<SUP> </SUP>held by the scientists of the day? Were his peers outraged because<SUP> </SUP>of the reactionary social
      position he was taking (as suggested<SUP> </SUP>by Vandenbroucke)? On the other hand, was he as guilty as proponents<SUP> </SUP>of the miasma
       theory for trying to apply his theory of disease<SUP> </SUP>transmission to all situations without allowing for the possibility<SUP> </SUP>
```

of multiple disease pathways? Did he fall into the trap of equatingthe absence of data with an absence of effect?<P>
When Snow contended that emanations from the bone-boiling factorieswere not causing ill health in the community at large, he invokedarguments that are often raised when unexpected health effectsare encountered following supposed low dose exposures. One argumentis that such health effects are implausible given what we knowabout high dose exposures. In this instance, Snow noted thatthe factory workers were not dying and therefore health effectsin the community at large were not plausible. A related argumentis that, even if workers are dying or suffering other healtheffects, because of the distance from the exposure source, theexposure levels in the community are probably too low to plausiblyaffect health.<P>
Health effects of low dose exposures are often seen as implausible,even in the face of accumulated consistent evidence. Such argumentshave frequently been invoked in environmental epidemiology.Examples of low dose exposures that have been deemed implausiblecontributors to disease risk based on what is known about highdose exposures include passive smoking, residential radon exposure,childhood lead exposure, electromagnetic fields, and residencenear nuclear facilities. If one begins with a fixed idea ofwhat is plausible, arguments regarding susceptible subgroups,inverse dose rate, hormesis, multiple pathways, multifactoretiologies, and complex exposures (e.g., the different constituentsof sidestream and mainstream smoke) are untenable.<P>
But how do we know that the factory workers were not dying orsuffering other ill effects? Snow cited no studies. All toooften the absence of data is argued as proof of no effect. Thisissue becomes especially difficult when regulatory decisionsare being made. In the absence of evidence, can something beconsidered safe? While science is important, it is ultimatelysocial forces, as much as science, that guide regulators indecision-making.<P>
Snow's statements and the questions that were put to him callto mind some of the fundamental difficulties inherent in environmentalepidemiology. Today, there are numerous examples of residentswho live near potential environmental hazards claiming healtheffects that can never be proven beyond a reasonable doubt.Although the "gold standard" is an unbiased risk estimate withprecise confidence limits, studies focused on overt health effectsare invariably underpowered because of the small numbers ofresidents in the neighborhoods of interest. Other creative approachesto assessment of subclinical health effects are more costlyand difficult to implement, but even these studies are oftentoo small for conclusive results. Yet, what is the right thingto do? If we wait for strong scientific evidence before we act—ifwe require proof that workers are dying or evidence of overtillness in the community—have we waited too long? Fewclusters are ever resolved with the identification of a causallink between some localized exposure and disease. While manyapparent clusters may be artifacts, what is the real cost ofthe true hazards that cannot be proven? These were the issuesfacing Parliament when Snow testified on behalf of industry.<P>
What is the role of the epidemiologist in this quagmire? InSnow's London, the living conditions of people near the factorieswere likely to have been dismal. There were no doubt residentswho perceived their symptoms as being related to the smells—smellsthat, if nothing else, impacted the quality of life. Policy-makersmust balance "doing the right thing" with regard to human sufferingand quality of life with the financial costs of doing so. Epidemiologycan only go so far in providing the answers.<P>
It is this political and social tug-of-war that makes environmentalepidemiology especially difficult. On the one hand, there arewell—funded industries with a financial stake in the outcomeof such research. As Vandenbroucke notes (2<A HREF="#B2"><IMG BORDER=1 WIDTH=8 HEIGHT=7 ALT="Go"
SRC="/icons/fig-down.gif"></A>), these industriesoften are in a position to exploit the many weaknesses thatepidemiologists are trained to identify in their own studiesand in the work of others to cast potentially damaging resultsin a more favorable light. On the other hand, there are environmentalgroups committed to proving that a particular environmentalexposure can be linked to a variety of personal complaints;these groups may be motivated by the possibility of effectingsocial change through science or by the prospect of receivingneeded medical attention or financial compensation. Those whoattempt to work in this arena often find themselves and theirresearch attacked from all directions.<P>
Environmental epidemiology is difficult to conduct today forother reasons as well. Adequate tools with which to measureand quantify exposures are lacking. Studies are often unableto detect meaningful effects because exposures are low, infrequent,or difficult to measure with certainty. How many investigatorsare willing to tackle this problem? In the case of the bone-boilingfactories, would research linking questionnaire data on symptomsto factory releases be believed? Would a study relating distancefrom the factory to disease be sufficient evidence of effect?What health effects would be plausible based on known biologicmechanisms? How well could those effects be measured, and couldthey be measured objectively? Is there a biomarker of exposure?If a biomarker exists, does it measure relevant past exposures?Is the measure unaffected by current health status—particularlythe disease under study?<P>
In addition to Lilienfeld's historical report and Vandenbroucke'scommentary, this issue of the <I>Journal</I> features papers that illustratevarious aspects of the difficulties faced in studying healtheffects of environmental exposures. Several of

these includeinnovative attempts to improve the quality of such research.<P>
The paper by Viel et al. (5<A HREF="#B5"><IMG BORDER=1 WIDTH=8 HEIGHT=7 ALT="Go"
SRC="/icons/fig-down.gif"></A>) may come closest to what many maythink of as environmental epidemiology. The authors have examinedthe spatial distribution of soft tissue sarcomas and non-Hodgkin'slymphomas around an incinerator with high dioxin emissions.Their results are suggestive but need to be followed by studiesincorporating more rigorous exposure assessment—perhapsa biologic measure of exposure such as that used in the studyof polychlorinated biphenyls and breast cancer reported by Zhenget al. (6<A HREF="#B6"><IMG BORDER=1 WIDTH=8 HEIGHT=7 ALT="Go"
SRC="/icons/fig-down.gif"></A>). Other studies described in this issue used a varietyof approaches to exposure assessment. Rondeau et al. (7<A HREF="#B7"><IMG BORDER=1 WIDTH=8 HEIGHT=7 ALT="Go"
SRC="/icons/fig-down.gif"></A>) linkedestimates of levels of aluminum and silica in drinking waterto risks of dementia and Alzheimer's disease. Laden et al. (8<A HREF="#B8"><IMG BORDER=1 WIDTH=8 HEIGHT=7 ALT="Go"
SRC="/icons/fig-down.gif"></A>)used questionnaire data on use of electric blankets to estimateexposure to electromagnetic fields, and Gustavsson et al. (9<A HREF="#B9"><IMG BORDER=1 WIDTH=8 HEIGHT=7 ALT="Go"
SRC="/icons/fig-down.gif"></A>)used questionnaire data and expert assessment by industrialhygienists to classify environmental and occupational exposures.Radiation workers are one of the few groups for which historicalrecords of personal exposure typically are available. Dupree-Elliset al. (10<A HREF="#B10"><IMG BORDER=1 WIDTH=8 HEIGHT=7 ALT="Go"
SRC="/icons/fig-down.gif"></A>) took advantage of such records to estimate cumulativeexternal radiation exposure.<P>
Several of the papers evaluate methods for assessing exposure.For example, Oglesby et al. (11<A HREF="#B11"><IMG BORDER=1 WIDTH=8 HEIGHT=7 ALT="Go"
SRC="/icons/fig-down.gif"></A>) average individual-level annoyancescores to estimate community-level exposure to air pollution.The authors propose that this measure better accounts for exposurevariability than data from fixed-site monitoring stations. Thisis an interesting twist in a field where much work is basedon linking data from monitoring stations with population-levelmortality statistics. The measure seems to be easy to operationalize,and it correlates well with monitoring station data, althoughits ultimate utility may be limited. The real gold standard—amore precise direct measure of individual exposure, rather thananother indirect measure—is what is needed. Hwang et al.(12<A HREF="#B12"><IMG BORDER=1 WIDTH=8 HEIGHT=7 ALT="Go"
SRC="/icons/fig-down.gif"></A>) propose an alternative modeling approach whereby air pollutionmonitoring station data are used to ascribe exposures to individualswith and without school absences due to respiratory disease.Auvinen et al. (13<A HREF="#B13"><IMG BORDER=1 WIDTH=8 HEIGHT=7 ALT="Go"
SRC="/icons/fig-down.gif"></A>) compare several possible methods for measuringand classifying exposure to electromagnetic fields. This isa topic that has been hurt by the lack of consensus on the bestand most appropriate exposure measure, and results tend to varyfor studies employing different exposure metrics. The paperby Karagas et al. (14<A HREF="#B14"><IMG BORDER=1 WIDTH=8 HEIGHT=7 ALT="Go"
SRC="/icons/fig-down.gif"></A>) attempts to link a biologic measure,arsenic in toenails, with an environmental measure of arsenicin water. The toenail measure is likely to reflect total bodyburden, but it appears to correlate with water only when waterlevels are high. This presents an interesting regulatory dilemma.The best epidemiologic research may be based on a direct measureof body burden such as levels in toenails, whereas it is waterlevels that need to be regulated. Studies of toenail arseniclevels may not shed direct light on the link between water levelsand disease.<P>
As these papers demonstrate, technological advances are makingpossible a wide range of new study designs and strategies tobetter assess both exposures and outcomes. Although progresshas been made, research in environmental epidemiology is farfrom perfect. As epidemiologists face pressures and criticismsfrom industry, regulatory bodies, and other scientific disciplines,it is important to not lose sight of the lessons from John Snow.<P>
<FONT SIZE=+1><STRONG>NOTES</STRONG></FONT><P>
<A NAME=""><!-- *null* --></A>
Reprint requests to Dr. Dale P. Sandler at this address (e-mail:sander{at}niehs.nih.gov<script type="text/javascript"><!--
var u = "sander", d = "niehs.nih.gov"; *document.getElementById*("em0")*.innerHTML* = '<a href="mailto:' + u + '@' + d + '">' + u + '@' + d + '<\/a>'//--></script>).<P>
<FONT SIZE=+1><STRONG>REFERENCES</STRONG></FONT><P>
<OL COMPACT>
<A NAME="B1"><!-- *null* --></A>
<LI VALUE=1>
Lilienfeld DE. John Snow: the first hired gun? Am J Epidemiol 2000;152:4–9.<!-- *HIGHWIRE* ID="152:1:1:1" --><A HREF="/cgi/ijlink?linkType=ABST&journalCode=amjepid&resid=152/1/4" ><nobr>[Abstract/Free Full Text]</nobr></A><!-- */HIGHWIRE* -->

<A NAME="B2"><!-- *null* --></A>
<LI VALUE=2>
Vandenbroucke JP. Invited commentary: the testimony of Dr. Snow. Am J Epidemiol 2000;152:10–12.<!-- *HIGHWIRE* ID="152:1:1:2" --><A HREF="/cgi/ijlink?linkType=FULL&journalCode=amjepid&resid=152/1/10" ><nobr>[Free Full Text]</nobr></A><!-- */HIGHWIRE* -->
<A NAME="B3"><!-- *null* --></A>
<LI VALUE=3>
Foreman CH Jr. The promise and peril of environmental justice. Washington, DC: Brookings Institute, 1998.<!-- *HIGHWIRE* ID="152:1:1:3" --><!-- */HIGHWIRE* -->
<A NAME="B4"><!-- *null* --></A>
<LI VALUE=4>
Horton R. The <I>new</I> new public health of risk and radical engagement. (Editorial). Lancet 1998;352:251.<!-- *HIGHWIRE* ID="152:1:1:4" --><A HREF="/cgi/external_ref?access_num=000074974500003&link_type=ISI" >[ISI]</A><A HREF="/cgi/external_ref?access_num=9690400&link_type=MED" >[Medline]</A><!-- */HIGHWIRE* -->
<A NAME="B5"><!-- *null* --></A>
<LI VALUE=5>
Viel J-F, Arveux P, Baverel J, et al. Soft-tissue sarcoma and non-Hodgkin's lymphoma clusters around a municipal solid waste incinerator with high dioxin emission levels. Am J Epidemiol 2000;152:13–19.<!-- *HIGHWIRE* ID="152:1:1:5" --><A HREF="/cgi/ijlink?linkType=ABST&journalCode=amjepid&resid=152/1/13" ><nobr>[Abstract/Free Full Text]</nobr></A><!-- */HIGHWIRE* -->
<A NAME="B6"><!-- *null* --></A>
<LI VALUE=6>
Zheng T, Holford TR, Tessari J, et al. Breast cancer risk associated with congeners of polychlorinated biphenyls. Am J Epidemiol 2000;152:50–8.<!-- *HIGHWIRE* ID="152:1:1:6" --><A HREF="/cgi/ijlink?linkType=ABST&journalCode=amjepid&resid=152/1/50" ><nobr>[Abstract/Free Full Text]</nobr></A><!-- */HIGHWIRE* -->
<A NAME="B7"><!-- *null* --></A>
<LI VALUE=7>
Rondeau V, Commenges D, Jacqmin-Gadda H, et al. Relation between aluminum concentrations in drinking water and Alzheimer's disease: an 8-year follow-up study. Am J Epidemiol 2000;152:59–66.<!-- *HIGHWIRE* ID="152:1:1:7" --><A HREF="/cgi/ijlink?linkType=ABST&journalCode=amjepid&resid=152/1/59" ><nobr>[Abstract/Free Full Text]</nobr></A><!-- */HIGHWIRE* -->
<A NAME="B8"><!-- *null* --></A>
<LI VALUE=8>
Laden F, Neas LM, Tolbert PE, et al. Electric blanket use and breast cancer in the Nurses' Health Study. Am J Epidemiol 2000;152:41–9.<!-- *HIGHWIRE* ID="152:1:1:8" --><A HREF="/cgi/ijlink?linkType=ABST&journalCode=amjepid&resid=152/1/41" ><nobr>[Abstract/Free Full Text]</nobr></A><!-- */HIGHWIRE* -->
<A NAME="B9"><!-- *null* --></A>
<LI VALUE=9>
Gustavsson P, Jakobsson R, Nyberg F, et al. Occupational exposure and lung cancer risk: a population-based case-referent study in Sweden. Am J Epidemiol 2000;152:32–40.<!-- *HIGHWIRE* ID="152:1:1:9" --><A HREF="/cgi/ijlink?linkType=ABST&journalCode=amjepid&resid=152/1/32" ><nobr>[Abstract/Free Full Text]</nobr></A><!-- */HIGHWIRE* -->
<A NAME="B10"><!-- *null* --></A>
<LI VALUE=10>
Dupree-Ellis E, Watkins J, Ingle JN, et al. External radiation exposure and mortality in a cohort of uranium processing workers. Am J Epidemiol 2000;152:91–5.<!-- *HIGHWIRE* ID="152:1:1:10" --><A HREF="/cgi/ijlink?linkType=ABST&journalCode=amjepid&resid=152/1/91" ><nobr>[Abstract/Free Full Text]</nobr></A><!-- */HIGHWIRE* -->
<A NAME="B11"><!-- *null* --></A>
<LI VALUE=11>
Oglesby L, Künzli N, Monn C, et al. Validity of annoyance scores for estimation of long term air pollution exposure in epidemiologic studies: The Swiss Study on Air Pollution and Lung Diseases in Adults (SAPALDIA). Am J Epidemiol 2000;152:75–83.<!-- *HIGHWIRE* ID="152:1:1:11" --><A HREF="/cgi/ijlink?linkType=ABST&journalCode=amjepid&resid=152/1/75" ><nobr>[Abstract/Free Full Text]</nobr></A><!-- */HIGHWIRE* -->
<A NAME="B12"><!-- *null* --></A>
<LI VALUE=12>
Hwang J-S, Chen Y-J, Wang J-D, et al. Subject-domain approach to the study of air pollution effects on schoolchildren's illness absence. Am J Epidemiol 2000;152:67–74.<!-- *HIGHWIRE* ID="152:1:1:12" --><A HREF="/cgi/ijlink?linkType=ABST&journalCode=amjepid&resid=152/1/67" ><nobr>

```
      [Abstract/<font COLOR="CC0000">Free</font> Full Text]</nobr></A><!-- /HIGHWIRE -->
<A NAME="B13"><!-- null --></A>
<LI VALUE=13>
 Auvinen A, Linet MS, Hatch EE, et al. Extremely low-frequency magnetic fields and childhood acute lymphoblastic leukemia: an exploratory analysis
      of alternative exposure metrics. Am J Epidemiol 2000;152:20&#150;31.<!-- HIGHWIRE ID="152:1:1:13" --><A HREF="/cgi/ijlink?linkType=ABST&
      journalCode=amjepid&resid=152/1/20" ><nobr>[Abstract/<font COLOR="CC0000">Free</font> Full Text]</nobr></A><!-- /HIGHWIRE -->
<A NAME="B14"><!-- null --></A>
<LI VALUE=14>
 Karagas MR, Tosteson TD, Blum J, et al. Measurement of low levels of arsenic exposure: a comparison of water and toenail concentrations. Am J
      Epidemiol 2000;152:84&#150;90.<!-- HIGHWIRE ID="152:1:1:14" --><A HREF="/cgi/ijlink?linkType=ABST&journalCode=amjepid&resid=152/1/84" ><nobr>
      [Abstract/<font COLOR="CC0000">Free</font> Full Text]</nobr></A><!-- /HIGHWIRE -->
</OL>
<FONT SIZE=-1><EM>Received for publication March 17, 2000.</EM></FONT>
<FONT SIZE=-1><EM>Accepted for publication March 29, 2000.</EM></FONT><P>
<BR CLEAR=ALL>
<A NAME="related"><!-- null --></A>
<P>
<STRONG><FONT SIZE="+1" FACE="verdana,arial,helvetica,sans-serif">Related articles in Am. J. Epidemiol.:</FONT></STRONG>
<P>
<DL>
  <FONT SIZE="-1" FACE="verdana,arial,helvetica,sans-serif">
  <DT><STRONG>John Snow: The First Hired Gun?</STRONG>
  <DD>David E. Lilienfeld<BR>
Am. J. Epidemiol. 2000 152: 4-9.
  <NOBR>
  <A HREF="/cgi/content/abstract/152/1/4">[Abstract]</A>
  <A HREF="/cgi/content/full/152/1/4">[FREE Full Text]</A>
   </NOBR>
  <P>
  </FONT><FONT SIZE="-1" FACE="verdana,arial,helvetica,sans-serif">
  <DT><STRONG>Invited Commentary: The Testimony of Dr. Snow</STRONG>
  <DD>Jan P. Vandenbroucke<BR>
Am. J. Epidemiol. 2000 152: 10-12.
  <NOBR>
  <A HREF="/cgi/content/extract/152/1/10">[Extract]</A>
  <A HREF="/cgi/content/full/152/1/10">[FREE Full Text]</A>
   </NOBR>
  <P>
  </FONT>
</DL>
  <FONT FACE="">
  </FONT>
<BR CLEAR=ALL>
<BR CLEAR=ALL>
<A NAME="otherarticles"><!-- null --></A>
<P>
</body></html>
```

Listing D.1: A sample from American Journal of Epidemiology

```
<html>
<body>
<H2>
<FONT COLOR=A70716 SIZE=-1>PHYSIOLOGY FORUM</FONT><BR>
```

```
Prologue
</H2>
<STRONG>
<NOWRAP>
Jeffrey
Pessin, <I>Editor</I>
</NOWRAP>
</STRONG>
<P>
American Journal of Physiology:, Endocrinology and Metabolism
<P>
<TXT>
<A NAME="SECO"><!-- comment for mosaic --></A>
<BR CLEAR=ALL>
<A NAME="Article"><!-- null --></A>
<TABLE WIDTH=100% BGCOLOR=E1E1E1 CELLPADDING=0 CELLSPACING=0>
<TR>
<TD ALIGN=LEFT VALIGN=MIDDLE WIDTH=5% BGCOLOR=FFFFFF> <IMG WIDTH=10 HEIGHT=21 HSPACE=5 ALT="" SRC="/icons/toc/rarrow.gif"></TD>
<TH ALIGN=LEFT VALIGN=MIDDLE WIDTH=95%><FONT SIZE=+2>  ARTICLE</FONT></TH>
</TR>
</TABLE>
<TABLE ALIGN=RIGHT CELLPADDING=5 BORDER NOWRAP><R>
<TH ALIGN=LEFT>
<FONT SIZE=-1>
<A HREF="#Top"><IMG WIDTH=11 HEIGHT=9 BORDER=0 HSPACE=5 SRC="/icons/toc/uarrow.gif">TOP</A>
<BR><IMG WIDTH=11 HEIGHT=9 BORDER=0 HSPACE=5 SRC="/icons/toc/dot.gif"><FONT COLOR=46
4C53>ARTICLE</FONT>
</FONT></TH>
</TR></TABLE>
<P>It is increasingly apparent that the individual metabolic processes
responsible for the maintenance of the normal physiological<SUP> </SUP>state are
highly integrated, complex, and multifaceted events.<SUP> </SUP>Because of the
inherent difficulties in understanding the interrelationships<SUP> </SUP>between
these processes, the precise molecular mechanisms are<SUP> </SUP>often difficult
to solve. Occasionally, a situation arises where<SUP> </SUP>a debate on a
particular issue ensues as either a series of papers,<SUP> </SUP>letters to the
editor, or both. This discussion can be scientifically<SUP> </SUP>stimulating,
providing a means to expand upon different points<SUP> </SUP>of view. However, in
certain circumstances this debate becomes<SUP> </SUP>prolonged, without any clear
resolution or development of new<SUP> </SUP>approaches to address the experimental
differences and/or conclusions<SUP> </SUP>drawn. To resolve such a debate, perhaps
to broaden the discussion,<SUP> </SUP>or to further stimulate interest in an
important area, the <I>American<SUP> </SUP>Journal of Physiology: Endocrinology
and Metabolism</I> is establishing<SUP> </SUP>a new venue for this type of<SUP> </SUP>discourse.
</P>
<P>When an Editor identifies an important area of controversy that both
merits and needs a thorough analysis of this type, the<SUP> </SUP>Editor will
initially invite a discussion of the topic by an individual<SUP> </SUP>or group of
investigators to provide an impartial and careful<SUP> </SUP>analysis of the
problem. This <font size=-1>INVITED DISCUSSION</font> may undergo some<SUP> </SUP>external
review and will then be published as the first part of<SUP> </SUP>the Physiology
Forum. The Editor can then encourage further discussion<SUP> </SUP>of the topic by
inviting additional <font size=-1>COMMENTARY</font> from the investigators<SUP> </SUP>directly involved in the original debate as well as other
      concerned<SUP> </SUP>investigators. The commentaries should focus on the published<SUP> </SUP><font size=-1>INVITED DISCUSSION</font> and
```

```
      other relevant papers, thereby
adding depth<SUP> </SUP>to the debate. These manuscripts will be published in a
subsequent<SUP> </SUP>issue of the Journal in the Physiology Forum under the
heading<SUP> </SUP>of <font size=-1>COMMENTARY.</font> We hope that this additional format
will provide<SUP> </SUP>an open forum for the pursuit of difficult discussions
that may<SUP> </SUP>point a way toward the resolution of an apparently intractable<SUP> </SUP>problem.
</P>
<P>The first of this series concerns the important issue of the
measurement of gluconeogenesis in vivo. Specifically, over the<SUP> </SUP>past
several years a great deal of controversy has arisen over<SUP> </SUP>the data and
interpretation based upon the use of [U-<SUP>13</SUP>C]glucose as a
tracer. This method has recently been discussed<SUP> </SUP>by both Drs. John Tayek
and Joseph Katz and Dr. Bernard Landau<SUP> </SUP>and his colleagues. Both groups
of investigators have introduced<SUP> </SUP>formulas for the determination of
gluconeogenic rates based upon<SUP> </SUP>the analysis of glucose and lactate
isotopomers derived from [U-<SUP>13</SUP>C]glucose. These analyses
substantially differ from each other<SUP> </SUP>in their estimates and from
physiologically expected results.<SUP> </SUP>To address these issues, we have
solicited the first <font size=-1>INVITED DISCUSSION</font><SUP> </SUP>from Drs. Jerry
Radziuk and Paul Lee. In the following issue,<SUP> </SUP>we will publish separate
commentaries from Dr. Joanne Kelleher,<SUP> </SUP>Drs. Katz and Tayek, and Dr.
Landau. We hope that this new publication<SUP> </SUP>venue will encourage
investigators in other areas to make use<SUP> </SUP>of the Physiology Forum to
discuss difficult and complex experimental<SUP> </SUP>issues.
</P>
</TXT>
<!-- Stray figures and tables -->
<!-- Appendices -->
<!-- Notes -->
<!-- Acknowledgements -->
<!-- Footnotes -->
<!-- Reprint -->
<!-- Corresp. Address -->
<!-- Received & Accepted -->
<!-- Bibliography -->
<P><HR>
<FONT SIZE=-1>Am J Physiol Endocrinol Metab 277(2):E197-E198<BR>
0002-9513/99 $5.00
<A HREF="/misc/terms.shtml"><B>Copyright &#169; 1999 the American Physiological Society</B></A><BR>
<!-- /COPYRIGHT -->
<!--

   Pages created by the Electronic Press Engine from

    Atypon Systems, Inc.

   Visit http://www.atypon.com/
-->
</vardef>
<BR CLEAR=ALL>
  <FONT FACE="">
  </FONT>
<BR CLEAR=ALL>
<BR CLEAR=ALL>
```

```
<BR CLEAR=ALL>
    <table class="content_box_outer_table"
    >
      <tr>
      <td>
<!-- beginning of inner table -->
  <table class="content_box_inner_table">
<!-- citation -->
<tr><td class="content_box_title_highlight" colspan="2">This Article</td></tr>
  <tr><td class="content_box_space_between_sections" colspan="2"><img alt="" width="200" height="1" src="/icons/spacer.gif"></td></tr>
    <tr><td width="4" class="content_box_arrow" valign="top"><img alt="" width="4" height="11" border="0" src="/icons/shared/misc/arrowTtrim.gif"><
      /td><td class="content_box_item">
    <strong>
          <a href="/cgi/reprint/277/2/E197">
    <strong>Full Text</strong> (PDF)</a>
    </strong>
    </td></tr>
    <tr><td width="4" class="content_box_arrow" valign="top"><img alt="" width="4" height="11" border="0" src="/icons/shared/misc/arrowTtrim.gif"><
      /td><td class="content_box_item">
        <strong><a href="/cgi/alerts/ctalert?alertType=citedby&addAlert=cited_by&saveAlert=no&cited_by_criteria_resid=ajpendo;277/2/E197&
      return_type=article&return_url=http%3A%2F%2Fajpendo.physiology.org%2Fcgi%2Fcontent%2Ffull%2F277%2F2%2FE197">
        Alert me when this article is cited</a></strong>
      </td></tr>
    <tr><td width="4" class="content_box_arrow" valign="top"><img alt="" width="4" height="11" border="0" src="/icons/shared/misc/arrowTtrim.gif"><
      /td><td class="content_box_item">
        <strong><a href="/cgi/alerts/ctalert?alertType=correction&addAlert=correction&saveAlert=no&correction_criteria_value=277/2/E197&return_type
      =article&return_url=http%3A%2F%2Fajpendo.physiology.org%2Fcgi%2Fcontent%2Ffull%2F277%2F2%2FE197">
        Alert me if a correction is posted</a></strong>
      </td></tr>
<tr><td class="content_box_space_between_sections" colspan="2"><img alt="" width="200" height="1" src="/icons/spacer.gif"></td></tr><tr><td class="
      content_box_title" colspan="2">Services</td></tr>
  <tr><td class="content_box_space_between_sections" colspan="2"><img alt="" width="200" height="1" src="/icons/spacer.gif"></td></tr>
    <tr><td width="4" class="content_box_arrow" valign="top"><img alt="" width="4" height="11" border="0" src="/icons/shared/misc/arrowTtrim.gif"><
      /td><td class="content_box_item">
        <strong><a href="/cgi/mailafriend?url=http%3A%2F%2Fajpendo.physiology.org%2Fcgi%2Fcontent%2Ffull%2F277%2F2%2FE197&title=Prologue">
        Email this article to a friend</a></strong>
      </td></tr>
    <tr><td width="4" class="content_box_arrow" valign="top"><img alt="" width="4" height="11" border="0" src="/icons/shared/misc/arrowTtrim.gif"><
      /td><td class="content_box_item">
    <strong>
    <a href="/cgi/search?qbe=ajpendo;277/2/E197&journalcode=ajpendo&minscore=5000">
    Similar articles in this journal</a>
    </strong>
    </td></tr>
    <tr><td width="4" class="content_box_arrow" valign="top"><img alt="" width="4" height="11" border="0" src="/icons/shared/misc/arrowTtrim.gif"><
      /td><td class="content_box_item">
        <strong><a href="/cgi/external_ref?access_num=10444412&link_type=MED_NBRS">
        Similar articles in PubMed</a></strong>
      </td></tr>
    <tr><td width="4" class="content_box_arrow" valign="top"><img alt="" width="4" height="11" border="0" src="/icons/shared/misc/arrowTtrim.gif"><
      /td><td class="content_box_item">
        <strong><a href="/cgi/alerts/etoc">
        Alert me to new issues of the journal</a></strong>
      </td></tr>
```

```
    <tr><td width="4" class="content_box_arrow" valign="top"><img alt="" width="4" height="11" border="0" src="/icons/shared/misc/arrowTtrim.gif"><
      /td><td class="content_box_item">
      <strong><a href="/cgi/citmgr?gca=ajpendo;277/2/E197">
      Download to citation manager</a></strong>
      </td></tr>
<tr><td class="content_box_space_between_sections" colspan="2"><img alt="" width="200" height="1" src="/icons/spacer.gif"></td></tr><tr><td class="
      content_box_title" colspan="2">Google Scholar</td></tr>
  <tr><td class="content_box_space_between_sections" colspan="2"><img alt="" width="200" height="1" src="/icons/spacer.gif"></td></tr>

    <tr><td width="4" class="content_box_arrow" valign="top"><img alt="" width="4" height="11" border="0" src="/icons/shared/misc/arrowTtrim.gif"><
      /td><td class="content_box_item">
              <strong><a target="_blank" href="http://scholar.google.com/scholar?q=%22author%3AJ.+author%3APessin%22">
              Articles by Pessin, J.</a></strong>
            </td></tr>
    <tr><td width="4" class="content_box_arrow" valign="top"><img alt="" width="4" height="11" border="0" src="/icons/shared/misc/arrowTtrim.gif"><
      /td><td class="content_box_item">
              <strong><a target="_blank" href="/cgi/external_ref?access_num=
            http://ajpendo.physiology.org

            /cgi/content/full/277/2/E197

          &link_type=GOOGLESCHOLAR">Articles citing this Article</a></strong>
            </td></tr>
<tr><td class="content_box_space_between_sections" colspan="2"><img alt="" width="200" height="1" src="/icons/spacer.gif"></td></tr><tr><td class="
      content_box_title" colspan="2">PubMed</td></tr>
  <tr><td class="content_box_space_between_sections" colspan="2"><img alt="" width="200" height="1" src="/icons/spacer.gif"></td></tr>
    <tr><td width="4" class="content_box_arrow" valign="top"><img alt="" width="4" height="11" border="0" src="/icons/shared/misc/arrowTtrim.gif"><
      /td><td class="content_box_item">
        PubMed Citation</a></strong>
      </td></tr>
    <tr><td width="4" class="content_box_arrow" valign="top"><img alt="" width="4" height="11" border="0" src="/icons/shared/misc/arrowTtrim.gif"><
      /td><td class="content_box_item">
              <strong><a target="_blank" href="/cgi/external_ref?access_num=Pessin+J&link_type=AUTHORSEARCH">
              Articles by Pessin, J.</a></strong>
            </td></tr>
  </td></tr></table>
</td></tr></table>
<BR CLEAR=ALL>
  <P>
  <HR NOSHADE ALIGN=LEFT WIDTH=450>
  <TABLE CELLPADDING=2 CELLSPACING=2 WIDTH=450>
  <TR>
    <TD ALIGN=CENTER VALIGN=TOP BGCOLOR=003399 NOWRAP><A TARGET="_top" STYLE="text-decoration: none" HREF="/"><FONT SIZE=-2 COLOR=FFFFFF FACE="
      arial,helvetica">HOME</FONT></A></TD>
      <TD ALIGN=CENTER VALIGN=TOP BGCOLOR=003399 NOWRAP><A TARGET="_top" STYLE="text-decoration: none" HREF="/help/"><FONT SIZE=-2 COLOR=FFFFFF
      FACE="arial,helvetica">HELP</FONT></A></FONT></TD>
      <TD ALIGN=CENTER VALIGN=TOP BGCOLOR=003399 NOWRAP><A TARGET="_top" STYLE="text-decoration: none" HREF="/cgi/feedback"><FONT SIZE=-2 COLOR=
      FFFFFF FACE="arial,helvetica">FEEDBACK</FONT></A></TD>
      <TD ALIGN=CENTER VALIGN=TOP BGCOLOR=003399 NOWRAP><A TARGET="_top" STYLE="text-decoration: none" HREF="/subscriptions"><FONT SIZE=-2 COLOR=
      FFFFFF FACE="arial,helvetica">SUBSCRIPTIONS</FONT></A></TD>
      <TD ALIGN=CENTER VALIGN=TOP BGCOLOR=003399 NOWRAP><A TARGET="_top" STYLE="text-decoration: none" HREF="/contents-by-date.0.shtml"><FONT SIZE
      =-2 COLOR=FFFFFF FACE="arial,helvetica">ARCHIVE</FONT></A></TD>
      <TD ALIGN=CENTER VALIGN=TOP BGCOLOR=003399 NOWRAP><A TARGET="_top" STYLE="text-decoration: none" HREF="/search.dtl"><FONT SIZE=-2 COLOR=
      FFFFFF FACE="arial,helvetica">SEARCH</FONT></A></TD>
```

```
     <TD ALIGN=CENTER VALIGN=TOP BGCOLOR=003399 NOWRAP><A TARGET="_top" STYLE="text-decoration: none" HREF="/content/vol277/issue2/"><FONT SIZE=-2
      COLOR=FFFFFF FACE="arial,helvetica">TABLE OF CONTENTS</FONT></A></TD>
  </TR>
  </TABLE>
       <TABLE CELLPADDING=2 CELLSPACING=2 WIDTH=
500
>
       <TR>
               <TD ALIGN=CENTER VALIGN=TOP BGCOLOR=003399 NOWRAP><A STYLE="text-decoration: none" HREF="http://www.physiology.org"><FONT SIZE=-2
     COLOR=FFFFFF FACE="arial,helvetica">Visit Other APS Journals Online</A></TD>
       </TR>
       </TABLE>
</BODY>
</HTML>
```

Listing D.2: A sample from American Journal of Physiology - Endocrinology And Metabolism