UNIVERSITY OF PIRAEUS

SCHOOL OF INFORMATICS AND COMMUNICATIONS

DEPARTMENT OF DIGITAL SYSTEMS

POSTGRADUATE PROGRAM

INFORMATION SYSTEMS & SERVICES

Enhancing Biomedical Question Answering Systems for COVID-19

MSc Thesis

By Ioannis-Andreas Philippas

Supervisor: Christos Doulkeridis

February, 2024

Abstract

In the domain of biomedical research and service, the retrieval of relevant information from diverse data sources remains a critical challenge. Traditional Information Retrieval (IR) systems often struggle with the complexity and the specificity of the biomedical domain. The COVID-19 pandemic has underscored the critical need for robust biomedical Question Answering (QA) systems capable of rapidly retrieving accurate and relevant information from validated biomedical literature sources. This thesis proposes an innovative approach that integrates dense neural networks with traditional IR methods, to enhance the performance of biomedical QA systems, with primary focus on addressing COVID-19-related inquiries.

At its core, the system utilizes dense models, such as transformer-based architectures like BERT (Bidirectional Encoder Representations from Transformers), known for their ability to capture semantic relationships and context in textual data. These models are trained on large-scale biomedical corpora to develop a deep understanding of domain-specific language and terminology. Additionally, the integration of traditional IR methods like BM25 complements the dense model IR infrastructure by providing an efficient and effective mechanism for initial document retrieval based on keyword matching and statistical relevance scoring. Combining these two approaches, the proposed system aims to enhance the accuracy, relevance and efficiency of biomedical QA tasks, particularly in the context of COVID-19.

The system proposed in this thesis, incorporates a reader module trained on both biomedical and general QA datasets. This module, leverages techniques from machine reading comprehension, further refines retrieved documents to extract precise answers to user queries.

The proposed QA system is supported by a web application, offering users a friendly interface for querying biomedical-related inquiries. The back-end system orchestrates various components to efficiently retrieve documents stored in a specific vector database, rank their relevance, and extract or generate potential answers. These answers are then pre-

sented to users through a user-friendly interface. Additionally, users have the flexibility to customize system parameters via the user interface, enhancing the system's usability.

By adapting advances neural networks such as BERT and Transformer-based models in biomedical domain, the system exhibited an increase in metrics over traditional and zero-shot methods. This thesis underscore the potential of dense models and QA systems to revolutionize biomedical IR, offering promising directions for future research and practical applications in enhancing the accessibility of critical biomedical knowledge.

Subject Area

Information Retrieval and Question Answering, within Biomedical domain.

Keywords

Natural Language Process, Question Answering, Information Retrieval, COVID-19, Transformers, BERT, SBERT, Generative Pseudo Labeling

Acknowledgements

I would like to thank my Supervisor professor Mr. Christos Doulkeridis for the provided boost during a difficult period of my life, as for his precious guidance, indications and help. I want to thank my family and my close ones for their patience, support and courage they provided me during the writing of this interesting research. Finally, I would like to dedicated this effort to my father who passed away recently, a true hero of his times, who he taught me to always push my limits.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

In December 2019, a new type of coronavirus, later identified as Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), triggered a series of unusual respiratory illnesses in Wuhan, Hubei Province, China finally causing pandemic. These respiratory illnesses were labeled as COVID-19 [1]. According to the World Health Organization's statistics, from the beginning of the pandemic until February 2024, the global death toll has reached 7,035,337 individuals [2]. Furthermore, a total of 774,771,942 people have been reported to have contracted the virus, approximately 9.55% of the total population [2].

Due to the unprecedented nature of this pandemic, governments around the world and by extension healthcare systems, have had to rapidly adapt to high demands and organize strategies to address this crisis. As a result, healthcare workers, which are in the front line of this battle, are now required to handle increased workloads in order to successfully manage these changing circumstances. Besides, it was a situation of which they had no prior experience since they had never encountered similar incidents before.

Since the emergence of COVID-19, there has been a notable surge in the volume of data related to this topic [3]. The scientific community quickly came together to address the crisis and create preventive measures to stop it from happening again. On the one hand, the AI community took steps towards finding automated detection of COVID-19 through Computed Tomography (CT) scans and X-ray images and on the other hand, epidemiologists and mathematicians by creating complex models to track the virus. Furthermore, many scientists turned to collecting scientific articles related to this topic creating a comprehensive repository of literature that encourages the creation of knowledge-based systems [4].

At the same time, the rapid increase in computing power has laid the groundwork for the development of fast QA systems that meet the needs of domain specialists and assist them in terms of accuracy and productivity. Such systems, should be capable of answering

both simple and more complex questions posed by clinicians in order to facilitate them in the field. It is imperative to note that the reliability of the sources from which information is retrieved and processed is crucial.

The main objective of the QA system presented by this thesis is to construct a time efficient, low cost accurate web infrastructure that can be easily deployed and updated in any health compartment. This system, allows users to ask biomedical questions and receive pertinent answers sourced from authoritative literature. The system can either directly extract answers from relevant texts or synthesize responses based on the information found in those texts. Additionally, the system can provide relevant passages without marking the exact answer. The users also have the option to configure the number of relevant passages to retrieve according to their preferences.

The proposed system can significantly assist healthcare workers in multiple ways. Given the rapid evolution of the pandemic and the continuous emergence of new data, such a system can provide instant, reliable, and up-to-date information, crucial for healthcare decision-making.

1.2 Question Answering Systems in Biomedical Field

The complexity of medical data, coupled with the countless sources and continual advancements, often acts as a barrier for clinicians in making decisions. Additionally, the individualized nature of patient care further complicates their tasks, requiring tailored approaches for each case. Last but not least, medicine is divided into many fields, each with its own unique knowledge and practices, which also adds complexity to healthcare decision-making process.

Biomedical QA systems seeks to support healthcare professionals by providing them with relevant information to address their queries effectively. Such systems, can address medical data complexities efficiently as they can parse extensive medical literature repositories in order to answer questions posed by clinicians. Natural Language Processing (NLP) techniques are able to handle large amounts of data and extract meaningful insights. The

fact that they are capable to retrieve information rapidly, enable them to act as a valuable and disconnected piece of their work.

Despite the possible advantages of QA systems in the biomedical field, there are some issues that need to be addressed. The explanation shortage is a common challenge when it comes to this field. There is imperative need for the clinicians to receive explanations for the responses and not only provide precise answers (e.g. "yes" or "no") [5]. Moreover, the fact that science is continuously evolving adds another risk to these systems [5]. This implies that may have incorporated data (e.g. scientific literature) whose results and practices are now out of dated or, even worse, have been contradicted by more recent research. Consequently, the system could potentially mislead the user. Another challenge arises from the scarcity of datasets for training compared to general domain QA datasets. While most general domain datasets are easier to find, biomedical ones are more complex and larger and require careful expert annotation which is more expensive and extensive work hours are necessary [6].

1.3 Structure

The master thesis begins with an introductory section 1 that provides a concise overview of the subject and outlines objectives of the research.

Subsequently, section 2 dices into the theoretical concepts essential for comprehending the development of a QA system. This includes an in-depth exploration of the components comprising a QA system, and a presentation of Deep Learning (DL) NLP concepts.

Section 3 provide a comprehensive overview of QA systems, describing their components and functionalities. This section serves as foundational resource for understanding the theoretical foundations of QA systems, facilitating deeper comprehension of their inner functionalities and mechanisms.

Moving forward, section 4 describes the methodologies undertaken in training and developing a hybrid QA system. A description of the dataset utilized in the research, ac-

companied by a presentation of the techniques employed to tailor NLP models to the specialized biomedical domain are provided in the same section.

Furthermore, section 5 presents and analyzes the outcomes of the evaluation process. In this section benchmark datasets are described, as the evaluation metrics employed during the model evaluations.

Section 6 provides a description of the real-world QA application, developed based on the models, which were trained for the purpose of this research.

Finally, the section 7 quotes conclusions and denotes observed drawbacks. In this section, future work and possible enhancements are mentioned, aiming to augment the proposed QA system accuracy.

2 Theoretical Background and Literature Review

In this section, key concepts and advancements in NLP with a focus on transformer models are explored. It begins by discussing fundamental NLP techniques such as Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) for text representation. Next, it covers word embeddings like Word2Vec and GloVe, which enhance semantic understanding in NLP tasks. The document then introduces transformers, emphasizing their innovative architecture and components like the attention mechanism. It highlights notable transformer models such as BERT and its variants, showcasing their impact on various NLP applications and tasks. In the second part of this section, the reader is informed about the literature which this thesis was based on

2.1 Natural Language Processing

2.1.1 Overview

NLP is a subfield of Artificial Intelligence (AI). It's primary task is to enable computers interact with humans using natural language. It is obvious that computers cannot understand human language in the same way humans do. Through the utilization of text numerical representation by transforming text data into format that can be processed mathematically, the interaction between humans and computers is achieved.

2.1.2 Numerical Representation of Text

Bag-of-Words (BoW) [7] and Term Frequency-Inverse Document Frequency TF-IDF [8] are fundamental NLP techniques, both of which maps a vector to each word. In the case of BoW, it counts the frequency of each word in the document. This method has limitations, as it cannot understand the semantic meaning of each word in the given context. On the other hand, TF-IDF, measures the importance of each word in a given context. It calculates TF and IDF. TF measures how frequent is a word in a document and IDF measures the uniqueness or rareness of a word in a document. This component penalizes words that appear frequently across many documents and gives higher weight to words that are more unique to a particular document.

One of the most widely used algorithms that utilize the aforementioned methods is BM25 algorithm. BM25 (Best Match 25) serves as a ranking algorithm introduced in 1994 by Robertson and Walker in their paper "*The Probabilistic Relevance Framework: BM25 and Beyond*" [9]. BM25 calculates the relevance score of a document to a given query. Its implementation utilizes TF-IDF and document length. Regarding TF, it incorporates a saturation function for term frequency. When a word appears multiple times in a document, each additional occurrence does not count as much as the first one. This prevents long documents from being favored too much in search results only because they use the same word over and over. It also employs a saturation function in IDF. This function ensures that uncommon words do not play a huge role in determining the importance of a document. This action prevents them from having too much influence on search results. Moreover, BM25 normalizes the document by comparing it with the average document length. This ensures that larger documents will not have bigger impact on search results that smaller ones. Moreover, the algorithm has two tunable parameters: *k1* and *b*, which adjusts how quickly the importance of a term diminishes as it appears more in a document and controls how much document length affects the score, respectively.

BM25 score is given by the following formula:

$$\text{BM25}(D,Q) = \sum \big(\text{IDF}(q) \times \frac{\text{TF}(q,D) \times (k1+1)}{\text{TF}(q,D) + k1 \times \big(1 - b + b \times \big(\frac{|D|}{\text{avgdl}}\big)\big)}\big) \tag{1}$$

where:

- D is the document.
- Q is the query.
- q is a term in the query.
- IDF(q) is the inverse document frequency of term q.
- TF(q,D) is the term frequency of term q in document D
- *k1* and *b* are constants that can be adjusted to tune the BM25 formula.
- |D| is the length of document D
- avgdl is the average document length in the corpus.

Overall, BM25 is a straightforward yet effective algorithm that can handle large documents as it normalizes the length. However, its performance rely on parameter selection and it does not take into account semantics or the relationship between words within the same document, assuming them to be independent of each other.

2.1.3 Word Embeddings

The aforementioned methods of representing words in vector space have limitations stemming from their inability to comprehend semantic associations between words within a context. Word embeddings mitigate the shortcomings of BoW and TF-IDF by offering approaches to represent words in vector space while leveraging their semantics. In this framework, similar words are assigned similar or identical representations within the vector space, enabling more context-aware language processing.

Word2Vec and GloVe are two popular techniques for generating word embeddings. Word2Vec [10], pioneered by researchers at Google, learns distributed representations of words in a continuous vector space, where words with similar meanings are located closer together. Furthermore, it is trained using large text corpora and typically results in dense word embeddings of fixed dimensions. GloVe [11] firstly introduced by researchers at Stanford University. Unlike Word2Vec, which focuses on predicting word co-occurrences, GloVe directly learns word embeddings by optimizing a global objective function that captures the ratio of co-occurrence probabilities between words. It leverages global statistics of word co-occurrence across the entire corpus to produce word embeddings that reflect both local and global semantic relationships.

2.2 Transformers

Transformers are models that utilize word embeddings. Embeddings serve as the initial input to transformers, representing each word or token in a continuous vector space. The concept of transformers was first introduced in the paper titled "*Attention Is All You Need*" [12] in 2014 which presents the transformer architecture and its core components. Among these components is the encoder-decoder architecture. According to the paper, transformers are simpler, more effective compared to Recurrent Neural Networks (RNNs)

and Long Short-Term Memory (LSTMs). The attention mechanism, which enables the model to focus on various segments of the input sequence while producing the output sequence, is the primary innovation of the this architecture. The model can better manage variable-length input and output sequences and capture long-range dependencies thanks to this attention approach. With the transformer, the authors achieved state-of-the-art results in various NLP tasks, including machine translation and language modeling, and it has since become a cornerstone in the field of DL.

2.2.1 Attention Mechanism

The attention mechanism is a fundamental concept regarding transformers. This mechanism guides the model's focus to different segments of the sequence, aiding in concept comprehension. It provides a mathematical representation of the text that captures the underlying concepts, allowing the model to understand the context better. By dynamically allocating attention, the mechanism ensures that the model prioritizes relevant information, leading to a more nuanced understanding of the sequence's concepts. For each embedding is obtained query, key and value vectors, denoted as Q, K and V, respectively. Matrices W_Q, W_K and W_V are used to derive the query, key, and value vectors, respectively. These matrices are model's learnable parameters, meaning that they are updated during training process through backpropagation algorithm. In the scaled dot-product attention mechanism, the query (Q), key (K) and value (V) vectors are computed using linear transformations of the input embeddings:

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V \tag{2}$$

where X represents the input embeddings and W_Q, W_K and W_V are the learnable weight matrices for the query, key and value, respectively. Once the query, key, and value vectors are obtained, the attention scores (A) are computed as follows:

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{3}$$

where d_k represents the dimensionality of the key vectors. The output of the attention mechanism is computed by the weighted sum of the value vectors using the attention scores:

$$\text{Attention}(Q, K, V) = AV \tag{4}$$

This operation results in a new set of context-aware embeddings that capture the semantic relationships between the original tokens in the input sequence.

2.2.2 Encoder

Figure 1 illustrates a simplified version of the transformer architecture. The left part presents the encoder module which consists of multiple sub-modules. The process initiates with the input embeddings. After that, the attention mechanism processes the input embeddings. A normalization layer is applied after attention mechanism stabilizing the learning process by normalizing the outputs of the previous layer. The output from the normalization layer is then passed through a feed-forward neural network. Another normalization layer is applied after the feed-forward network. This sequence of attention, normalization, and feed-forward layers can be stacked multiple times to form a deep encoder network.

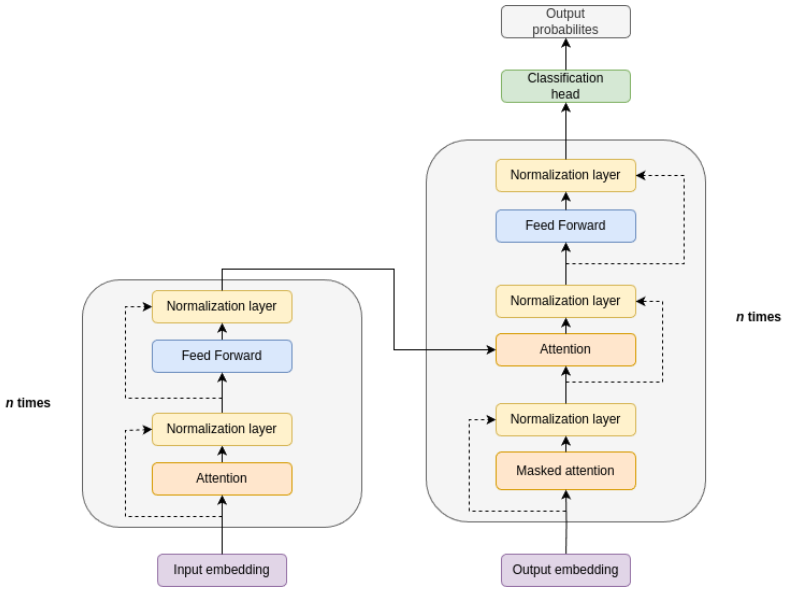

Figure 1: Simplified transformer representation

2.2.3 Decoder

The right part of the Figure 1 illustrates decoder module which can be replicated multiple times within a transformer model architecture. The decoder starts by taking output embeddings, which are processed by a masked attention mechanism. This masking ensures predictions are made based on already produced tokens, maintaining an auto-regressive nature. The decoder then uses a normalization layer, after which it employs another attention mechanism that draws information from the encoder's outputs. Another round of normalization precedes the feed-forward network in the decoder. Finally, after a last normalization step, the output passes through a classification head that predicts the probabilities of the next tokens in the sequence.

2.3 Notable Transformer Models

In this section, some of the most notable transformer models are introduced that have significantly impacted the field of NLP.

2.3.1 BERT (Bidirectional Encoder Representations from Transformers)

BERT, or Bidirectional Encoder Representations from Transformers, stands as a revolutionary model within the realm of NLP, pioneered by Google AI researchers in 2018. Described in the paper "*BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*" [13] this model introduces an innovative pre-training methodology tailored for enhancing language comprehension tasks. Built upon the Transformer model, BERT utilizes self-attention mechanisms to capture word dependencies within sentences. Unlike prior models, BERT adopts a bidirectional approach by pre-training on a large corpus using masked language modeling (MLM) and next sentence prediction (NSP) tasks. In MLM, random tokens are masked, and the model predicts these tokens based on context, allowing BERT to learn bidirectional representations of words. NSP involves predicting whether pairs of sentences are consecutive, aiding in understanding sentence relationships. BERT achieves state-of-the-art performance on various NLP tasks, such as QA and sentiment analysis, by fine-tuning the pre-trained model on task-specific datasets. Its ability to capture contextual information bidirectionally eliminates the need for task-specific architectures. Furthermore, BERT introduces innovative pre-training techniques, including

large-scale corpus usage, dynamic masking, and next sentence prediction. Variants like BERT-base and BERT-large offer different model sizes and parameters.

2.3.2 Variants of BERT

- RoBERTa: RoBERTa [14], developed by Facebook AI in 2019, is an optimized variant of BERT. It enhances pre-training strategies by using a larger corpus, dynamic masking, and longer sequences during training. Unlike BERT, RoBERTa removes the next sentence prediction task and focuses solely on masked language modeling. These optimizations lead to significant performance improvements across various NLP tasks, making RoBERTa a widely adopted model in the field.
- BIOBERT: While it is based on BERT, BIOBERT [15] is specifically designed and trained for biomedical text mining and NLP tasks. It utilizes the same architecture and pre-training objectives as BERT but is pre-trained on biomedical text data to better understand and process domain-specific language. In summary, BIOBERT is a specialized variant of BERT tailored for biomedical NLP applications, aiming to address the unique challenges and requirements of analyzing biomedical text data.

2.3.3 Sentence-BERT (SBERT)

Sentence-BERT (SBERT) [16] represents a significant advancement in NLP by addressing the challenge of capturing semantic similarity between sentences. Developed by researchers at the UKP Lab, SBERT introduces modifications to the BERT architecture to enhance its effectiveness in generating fixed-length sentence embeddings. SBERT adopts a siamese network architecture, where two identical BERT models share weights. This architecture allows SBERT to compare pairs of sentences and generate embeddings that capture semantic similarities or dissimilarities between them. Unlike traditional BERT models, SBERT is trained using a contrastive loss function. This loss function encourages semantically similar sentence pairs to have embeddings that are close together in the embedding space, while dissimilar pairs are pushed further apart. By optimizing this training objective, SBERT learns to generate discriminative embeddings that encode semantic information effectively. BERT, like BERT, supports fine-tuning on downstream tasks using task-specific datasets. Leveraging transfer learning from pre-trained SBERT models en-

ables users to achieve high performance on various NLP tasks with minimal task-specific training data.

2.4 Literature Review

In recent years, AI in biomedical domain has witnessed significant advancements, particularly in the development of complex QA systems. The onset of COVID-19 pandemic has further demanded the need for such systems, as healthcare professional and researchers seeking updated information related to the virus. This literature review presents some insights, on deep network and statistical architectures utilized in this thesis, the development of the biomedical datasets and the evolution of QA systems within the biomedical domain.

Robertson and Zaragoza in 2009 contributed to the field of information retrieval by introducing BM25 algorithm [9]. The article outlines a probabilistic approach to ranking documents based on their relevance to a given query, aiming to overcome limitations of traditional vector models. BM25, a variant of the probabilistic model, incorporates term frequency and document length normalization to calculate document relevance scores, offering robust performance across a wide range of retrieval tasks.

Efficient Estimation of Word Representations in Vector Space is a landmark paper in NLP, introduced by Tomas Mikolov and his colleagues at Google in 2013 [10]. This paper introduced the Word2Vec model, which has become a cornerstone in NLP and ML. The primary contribution of this paper is the development of two algorithms for training distributed representation of words: Continuous Bag-of-Words (CBOW) and Skip-gram. These algorithms learn dense vector representations of words, where similar words are represented by vectors that are close together in the vector space.One of the key advantages of Word2Vec is its efficiency. It can be trained on very large corpora of text data in a reasonable amount of time. This efficiency was achieved through techniques such as hierarchical softmax and negative sampling, which enable faster training without sacrificing much in terms of accuracy.

Vaswani et al., in 2017 published the paper *Attention is All You Need* [12] which was pivotal in the field of NLP, particularly in the field of neural network architectures. The model proposed in this paper leverages self-attention mechanisms to effectively capture long-range dependencies in input sequences, making it highly efficient for parallelization and capable capturing contextual information across sequences of varying lengths. Its significance lies in its ability to achieve state-of-the-art performance on various tasks reducing training-time.

In 2018 Devlin et al., presented a groundbreaking approach to language representation learning through bidirectional transformers with the paper *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* [13]. BERT (Bidirectional Encoder Representations from Transformers) is pre-trained on large-scale text-corpora using masked language modeling and next sentence prediction tasks, enabling it to capture deep contextual representations of words and sentences. BERT has demonstrated remarkable performance across a wide range of NLP tasks, inluding QA, sentiment analysis and Named Entity Recognition. Its effectiveness stems from the ability to encode both leftand right context simultaneously, capturing semantic relationships within text.

Nogueira and Cho presented an innovative approach to passage re-ranking within the context of information retrieval. In their paper *Passage Reranking with BERT* [17], they proposed a state-of-the-art model, to enhance the relevance and accuracy of passage retrieval. By fine-tuning BERT on passage re-ranking tasks. The proposed cross encoder model was figuring as the the leader of the MS MARCO passage retrieval task, surpassing the next model by 27% in terms of MRR@10.

Reimers and Gurevych on 2019, introduced a significant advancement in the realm of sentence embeddings. In their paper *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks* [16] proposed a novel approach by training BERT-based models specifically for generating sentence embeddings by employing Siamese BERT-network architecture. By fine-tuning BERT on Siamese network setup, Sentence-BERT effectively learns to transform sentences onto vectors in a continuous vector space where semantically similar sentences are closer together. This was a great achievement in context of computation power. Comparing sentences using BERT is expensive in computation and time terms.

Reimers method enables the generation of high-quality sentence embeddings that capture semantic information. Sentence-BERT models are suitable for tasks like semantic textual similarity, paraphrase identification and clustering.

Similar to Sentence-BERT architecture, another bi-encoder model was introduced by Facebook in 2020 [18]. This work introduces Dense Passage Retriever (DPR), a dense retrieval approach that utilizes dense representations of passages to retrieve relevant information from large-scale document collections. The difference with Sentence-BERT models, is the utilisation of separate encoders, for query and context vector representation. DPR models has demonstrated remarkable effectiveness for information retrieval in general domains.

A novel approach to train dense retrievers for IR tasks in NLP was presented by Hofstätter et al. in *Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling*. The authors proposed Topic Aware Sampling Balance (TAS-B), a technique designed to address the challenge of topic imbalance in retriever training data. By dynamically adjusting the sampling strategy during training, TAS-B ensures that retriever is exposed to diverse range of topics, improving its capability to retrieve relevant documents across various domains. The paper demonstrates the superiority in performance of the proposed trained model over traditional models, retrieving related documents. The TAS-B model outperform BM25 method by 44% in terms of NDCG@10, and 11% more than the docT5query model. The benchmark dataset used for the evaluation of the models was TREC-Deep Learning Track query sets.

The BioBERT model, introduced in the paper *BioBERT: a pre-trained biomedical language representation model for biomedical text mining* [15], a pre-trained language representation model tailored specifically for biomedical applications. Developed by Lee et al. in 2019, BioBERT is a BERT based model pre-trained on a massive corpus of biomedical literature, including PubMed abstracts and PMC full-text articles. To evaluate the performance of BioBERT, Lee et al. conducted extensive experiments across various biomedical text mining tasks, demonstrating superior performance compared to general-purpose language models like BERT and domain-specific baselines. The BioBERT model fine-tuned for QA task demonstrated an MRR improvement of 12.24% over a generic BERT model.

Pranav Rajpurkar et al. in 2016, presented a significant advancement in the field of NLP by introducing the Stanford Question Answering Dataset (SQuAD) [19]. This dataset provides over 100,000 question-answer pair in a specific format, sourced from Wikipedia articles. The dataset is serving as a training dataset to build machine comprehension models, and evaluate them. A second dataset was curated two years later, incorporating in the dataset questions that could not be answered from given context [20].

In July of 2020 after the eruption of the COVID-19 pandemic, a critical resource was created to consolidate COVID-19 related knowledge in a single dataset. Allen Institute for AI with the help of other organizations released a massive dataset named CORD-19 [21] comprised of scholarly articles, pre-prints and other research materials related to COVID-19 and coronaviruses.

Another critical resource related to COVID-19 based on CORD-19 dataset was created and introduced in 2020. TREC-COVID [22] is a specialized test collection designed to evaluate information retrieval system's performance in retrieving relevant scientific literature related to COVID-19. The dataset comprises a curated set of documents and relevance judgments to 50 questions created by biomedical experts.

In the study *"COVID-QA: A Question Answering Dataset for COVID-19"* by Timo Möller et al. [23] COVID-QA dataset is presented. The dataset consists of 2,019 question-answer pairs related to COVID-19, annotated by biomedical experts , related to COVID-19. The study demonstrates that training a RoBERTa model on this domain-specific dataset results in significant performance gains, underlining the importance of specialized datasets for improving QA systems in biomedical domain.

Nandan Thakur et al. introduced BEIR a groundbreaking benchmark of information retrieval models across diverse domains and languages [24]. By providing a standardized evaluation framework, BEIR enables researchers to assess the generalization capability and effectiveness of their information retrieval models. The accompanied paper outlines the effectiveness of a hybrid IR model, which utilizes the IR power of keyword searching with a BM25 model, alongside a powerful dense cross-encoder. This model outperformed other models in most domain specific datasets.

Generative Pseudo Labeling which this thesis relies on was introduced in 2022 by Kexin Wang et al. The paper [25] presents a novel approach to unsupervised domain adaptation within the context of dense retrieval. It leverages generative models to create queries for an unlabeled dataset and with the assistance of powerful cross-encoders relation between generated queries and context are annotated. GPL enables effective adaptation of dense retrieval models to target specific domains without requiring labeled target domain data. In this work a model which was trained with GPL and TSDAE [26] method outperformed othe dense and keyword-searching applications in terms of *NDCG@10*.

In a related study by Shaina Raza, Brian Schwartz and Laura C. Rosella entitled *"CoQUAD: a COVID-19 question answering dataset system, facilitating research, benchmarking, and practice"* [27], they introduce CoQUAD, a QA system designed to help researchers and practitioners quickly find answers to COVID-19 related questions. The system is built on a Transformer-based architecture and utilizes a gold-standard dataset of 150 question-answer pairs prepared by public health domain scientists. The MPNet model, fine-tuned on the gold-standard dataset, is used within QA pipeline to improve answer retrieval accuracy.

The COBERT model which firstly proposed in the paper entitled *"COBERT: COVID-19 Question Answering System Using BERT"* [28] by Jafar A. Alzubi, Rachna Jain, Anubhav Singh, Pritee Parwekar and Meenu Gupta, employs a sophisticated approach that integrates domain-specific knowledge from various biomedical datasets, including COVID-19 research articles, to fine-tune a BERT model for enhanced understanding and processing of medical queries. A critical contribution of the paper is its curated dataset, derived from vast biomedical and COVID-19 specific sources, which it was utilized to train COBERT model. The paper outlines the model's performance in accurately answering complex medical questions.

In their paper entitled *"List-wise learning to rank biomedical question-answer pairs with deep ranking recursive autoencoders"* by Yan, Yan AND Zhang, Bo-Wen and Li, Xu-Feng and Liu, Zhenhan, an innovative approach to improve the retrieval of relevant answers in biomedical QA systems is proposed.

Paper [29] proposes an innovative approach to improve the retrieval of relevant answers in biomedical QA systems. By employing Recursive Neural Networks (RNNs) for learning semantic representations of question-answer relations and introducing an loss function tailored for ranking, this study aims to enhance the accuracy of snippet answer retrieval. The approach outperformed other challengers of the BioASQ tasks.

3 Question Answering Systems

QA systems are computer-based systems whose main purpose is to process and answer questions posed by the user in natural language [30]. Early research on QA in AI dates back to the 1960s. The QA Track in Text Retrieval Conference (TREC) evaluations in 1999 laid the groundwork for extensive community research about Information Retrieval/Information Extraction (IR/IE) [31]. The unpredictable growth of available data and the rise in processing power contribute to the improvement of such systems.

QA systems can be divided into two main categories based on the source from which the answers are retrieved from:

- Open QA Systems: Often call open-book, the answers to questions is achieved by giving the model the context and then utilizing algorithms to retrieve relevant information.
- Closed QA Systems: Often called closed-book, the context is not being given explicitly as the model has already incorporated the knowledge into its internal parameters.

In this thesis, open QA approach is implemented and the core components are: document store, retriever, reader and generator. First, a typical flow of such a system will be described and then each part of this system will be described in detail.

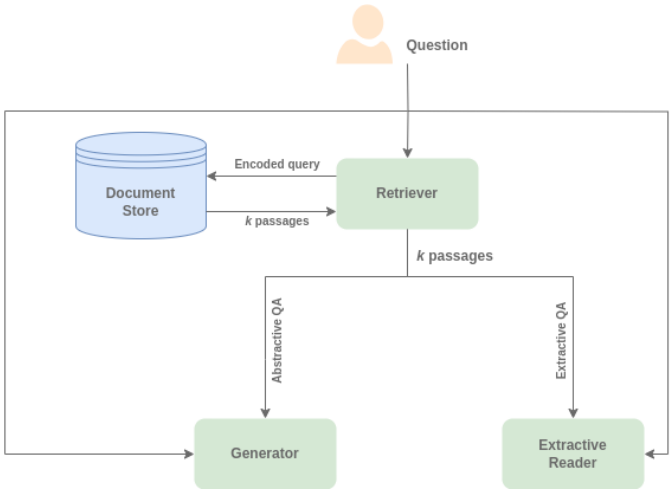

Figure 2: Architecture of an Open Domain Question Answering (Open QA) System

In Figure 2 the workflow of Open QA System in presented. Initially, a user poses a question and the system with the help of a retriever model, encodes the query. Then the retriever is responsible for passing the encoded query to the document store, which fetches the top k related passages that holds the answer. Upon retrieval, there are dual paths for generating the final answer: the Abstractive QA path and the Extractive QA path. The Abstractive QA employs a generator, which can create answers not limited to the exact words found in the text, effectively generating a response that synthesizes information from the k passages. In contrast, the Extractive QA utilizes an Extractive Reader model that pinpoints and extracts the precise text span from the provided passages that directly answers the query. With the first approach the system creates human-like responses, while in the second approach the responses are source-based.

3.1 Document Store

In these systems, document store or *knowledge base* frequently utilizes vector databases. A vector database stores data in a high-dimensional vectorized form. This vectorized version of each document arises from the importance of each term in the document and in the whole corpus. This representation facilitates efficient semantic similarity calculations and enables rapid retrieval of documents that closely match the query's semantic context.

3.2 Retriever

The retriever component plays a crucial role in IR systems by identifying the top k passages likely to contain the answer to the user's query. It calculates relevance by applying a similarity metric to assess how closely a specific document aligns with a given query. Given the computational expense of processing all available context, only a subset of relevant information is passed to the reader component [13]. This subset is determined by the hyperparameter k, which configures the amount of context fed into the reader.

3.3 Reader

Its main function is to process the passages or documents that the retriever has found in order to extract relevant data or responses to the user's inquiry. The reader receives these passages once the retriever has identified the top k passages most likely to contain the

desired information. By utilizing a range of NLP methods and Machine Learning (ML) models, including Neural Networks, the reader is able to understand the passages' content and identify important entities, events, or concepts that are referenced. After that, it responds to user questions by matching their inquiries with details found in the sections to provide a precise answer.

3.4 Generator

A common feature of QA systems is the generator component, which is essential for generating answers to user inquiries. In contrast to the reader, who takes answers straight out of passages that have been retrieved, the generator must create new content that includes responses or answers from beginning. It utilizes many natural language generation techniques, such as template-based, rule-based, and more sophisticated DL models. This module processes the user's query and relevant context such as retrieved passages or extra information when it receives it. Then, it predicts the word order that will best answer the user's question or offer the required information, and it generates a response accordingly.

4 Biomedical Domain Adapted Question Answering System

This chapter presents an in-depth exploration of the proposed QA system. At its core, the system is structured into two distinct subsystems: the IR system and the Reader/-Generator as it is shown in Figure 3. The IR system is tasked with conducting semantic searches within a designated corpus, matching a given query with stored data, and presenting a set of n relevant passages. The Reader/Generator system operates either in an extractive capacity, endeavoring to identify pertinent answers within a given context c in response to a query q, or in a generative capacity, aiming to formulate a coherent response to a prompted query by synthesizing information from context passages provided by the IR system.

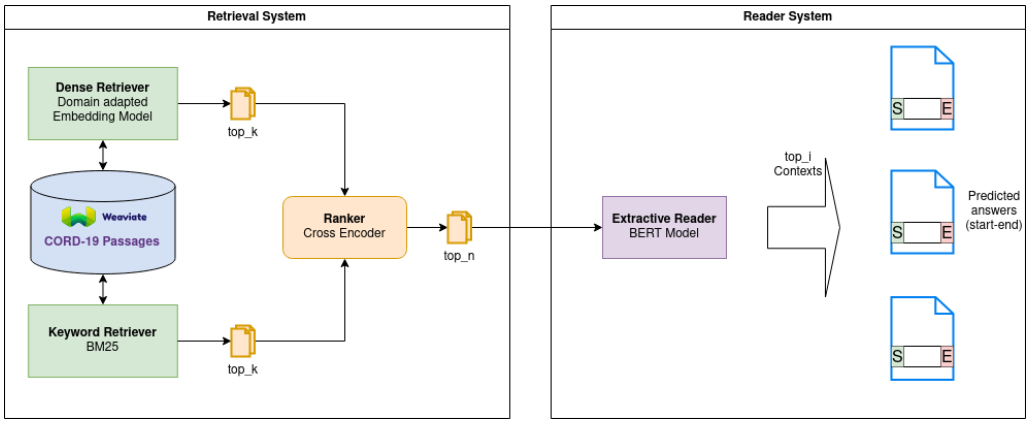

Figure 3: QA system

Numerous studies and academic papers have demonstrated that out-of-the-box performance of dense models often fall short of expectations when applied within specific domains such as biomedicine and finance due to differences in domain-specific terminology and linguistic structures. Notably, conventional methods like BM25 (referenced as [9]) outperform dense models in biomedical benchmarks.

Domain adaptation technique bridging the gap between general-domain pre-training and domain specific fine-tuning, giving the capability to dense models to leverage rich semantic information and relations within specific domain literature. In this thesis Generative Pseudo Labeling (GPL) methodology is employed (as described in [25]) to train dense retrieval models, thereby tailoring them to the biomedical domain. This approach resulted in notable enhancements to the performance of the retriever model, as assessed through

evaluation on a biomedical IR benchmark dataset (TREC-COVID). Subsequently, the adapted model was utilized to index the target domain corpus—a segmented version of the CORD-19 dataset (referenced as [21])—and store it within the Weaviate vector database (referenced as [32]) in the form of embeddings. In regard to the extractive reader model, a BERT model (referenced as [13]) was utilized initially fine-tuned for QA tasks using the SQuAD dataset (referenced as [19]), further refining its performance on a COVID-19 specific QA dataset (COVID-QA) (referenced as [23]). For the generative aspect of the reader, the FLAN T5 model (referenced as [33]) was employed, which has undergone fine-tuning across a diverse array of tasks.

The core QA system, encompassing both the IR and Reader components, is served through a Python-based Web API. This API is engineered utilizing the FastAPI framework. The User Interface (UI) is developed utilizing React. Notably, the entire system infrastructure operates within a containerized environment, employing Docker for deployment and management.

4.1 Information Retrieval

In this section the pipeline of the development of the IR system is presented, beginning with the data processing. The trained method (GPL) to adapt the dense models to the CORD-19 dataset is described, followed by indexing the corpus into the vector database. Finally, the TREC-COVID evaluation IR benchmark dataset [22] is presented.

4.1.1 Corpus and Data Preparation

The presented system is based on providing information according to the CORD-19 corpus [21], nevertheless the system is not limited to the specific dataset. The system provides utilities which can be used to fetch queried PubMed articles in available formats, convert them to the appropriate form and finally use them to train the model. Those articles can be also be a member of the information corpus of the dataset when indexed and uploaded in the vector database.

CORD-19. COVID-19 Open Research Dataset (CORD-19) [21] is a corpus which contains over a million scholarly articles, half of them in full text version about coronaviruses

COVID-19 and SARS-Cov-2. The dataset contains both meta information in a csv (comma separated values) format and the text of the corpus article split in individual json files. The baseline model has a maximum sequence length constraint of 512 tokens, which roughly equates to 350-400 words. Due to this limitation, passages have partitioned into chunks of 200 words, ensuring continuity in sentence structure to prevent automatic truncation of overflow words by the model. Moreover, passages containing specific COVID-19 related keywords were selectively collected, thereby facilitating the model's adaptation to the COVID thematic context. Sentences were not converted to lowercase and stop-words were not removed, as such pre-processing actions contravene the specifications under which the model was pre-trained, potentially compromising evaluation performance. The domain adaptation dataset contains just 50,000 chunked passages, as the relevant research [25] indicate increase in model performance within the initial few thousand training steps. Given time limitation for training and evaluation, as well as limitations imposed by the hardware infrastructure employed for this experiment, this dataset size is adequate for achieving meaningful results.

BioASQ. BioASQ [34] is a benchmarking activity focused on biomedical semantic indexing and QA. It provides a framework for evaluating the performance of systems in the context of retrieving and answering biomedical questions using highly specialized datasets. The BioASQ challenge is divided into several tasks, notably including Task A (large-scale biomedical semantic indexing) and Task B (biomedical semantic QA), each targeting different aspects of biomedical information processing.

For the COVID-19 domain adaptation process the BioASQ eleventh revision task A dataset was utilized. It is formatted in a json line file which each line represent a single article. Specifically, the abstract and title of the article were used, ignoring the rest of the metadata provided. The identical pre-processing pipeline described for the CORD-19 dataset is followed to finally obtain 50,000 passages.

The process method of the datasets described above is just an initial transformation. Further dataset manipulation is described during the explanation of the training method in a later section.

4.1.2 Domain Adaptation with GPL

Dense retriever models have obtained state-of-the-art results on datasets which contained a large number of training set elements from diverse topics, shown in Figure 4. This is not the case though when it comes to perform in datasets focused on specific domains such as biomedical field.

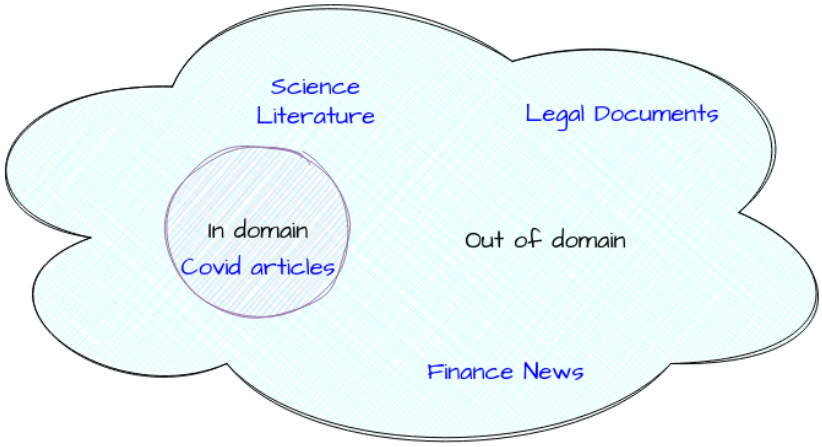

Figure 4: Domains

BEIR IR benchmark framework [24] reveals in its research, that sparse method like BM25 outperformed most of the single evaluated dense models (in terms of Recall@100). Domain Adaptation tries to adapt models to private or public specific text domain without the need of labeled training data, which is very limited. There are several methods proposed for domain adaptation such Masked Language Modeling, TSDAE [26] and GPL. The proposed system is based on the GPL method [25] which overcome the heavy computational overhead which the other methods introduce. The main difference of the training system is that GPL can be applied on an already fine-tuned model, thus the other method proposes extensive pre-training on target domain and then fine tune it on labeled data. Figure 5 illustrates the difference between GPL method and adaptive pre-training.

In the context of biomedical IR, GPL [25] offers a compelling solution for leveraging large volumes of unlabeled biomedical text data to enhance the performance of dense retrieval models. GPL is a rather complex method which augment IR performance in

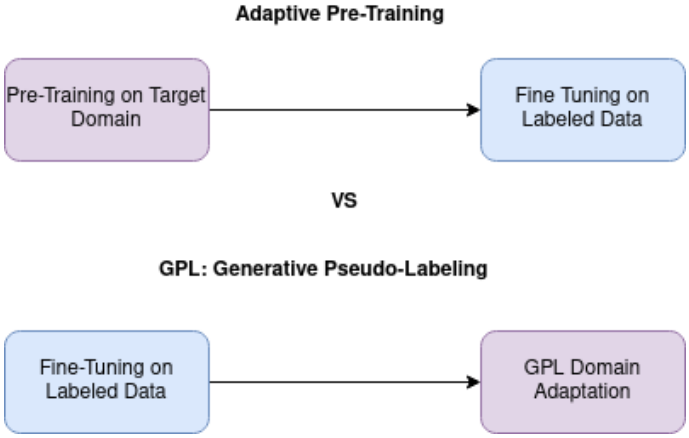

Figure 5: Domain Adaptation

domain specific datasets. The method consists three data process steps and a fine tuning process in the end.

A brief description of the GPL method is provided first, with further discussion of method steps to follow later in this section. The three data preparation steps are:

1. **Query Generation** using a T5 encoder-decoder [35]. The pre-trained T5 model generates questions for the CORD-19 training corpus.
2. **Negative mining** of negative passages with a baseline pre-trained model. Negative passages are defined as the passages not related to the provided query q (irrelevant passages Q^-)
3. **Pseudo Labeling**. A cross-encoder model is utilized to define similarity scores for retrieved pairs.

The last step of the method is to train/fine-tune a bi-encoder model by optimizing a margin MSE loss function.

Query generation is the process which generates possible search queries to an input document passage. The training dataset selected for the domain adaptation process consists 50,000 passages of the CORD-19 dataset and 50,000 BioASQ v2022 task a abstract passages. Each training passage average about 200 words. The CORD-19 passages met an inclusion criteria based on the matching of COVID-19 related keywords (SARS.CoV.2, COVID.19, 2019.nCoV, covid, corona, coronavirus, SARS). No exclusion criteria was made

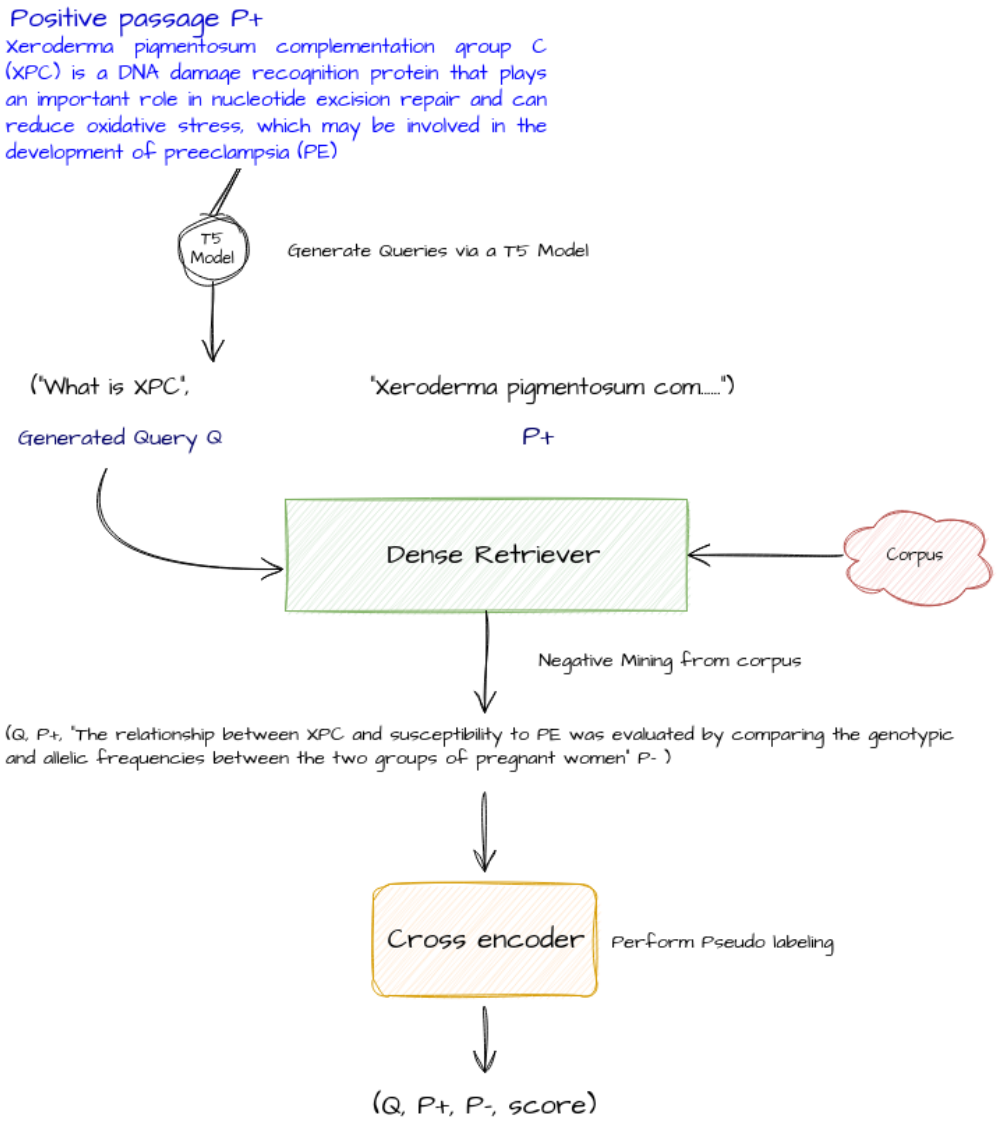

Figure 6: GPL

for the BioASQ dataset. The objective was to adapt the model to encompass general biomedical terminology.

DocT5query T5 based model was used to generate question for the pre-processed biomedical corpus. In the experiment conducted *query-gen-msmarco-t5-base-v1* model was selected which is a pre-trained T5 model. This model has been trained on MS MARCO Passage Dataset [36]. MS MARCO comprises about a million questions from Bing's search query log and the answering (relevant) passages and can be used for both training and as benchmark dataset for IR. It must be pointed out that the MS MARCO dataset does not contain any information about COVID-19.

For the training purpose of our experiment, 3 queries are generated for each passage. Each element of the new derived dataset contain a generated query Q and the relevant passage P^+. The dataset generated in this step contains $3 \times (2 \times 50,000) = 300,000$ pairs of (Q, P^+). The process is presented in the Figure 7

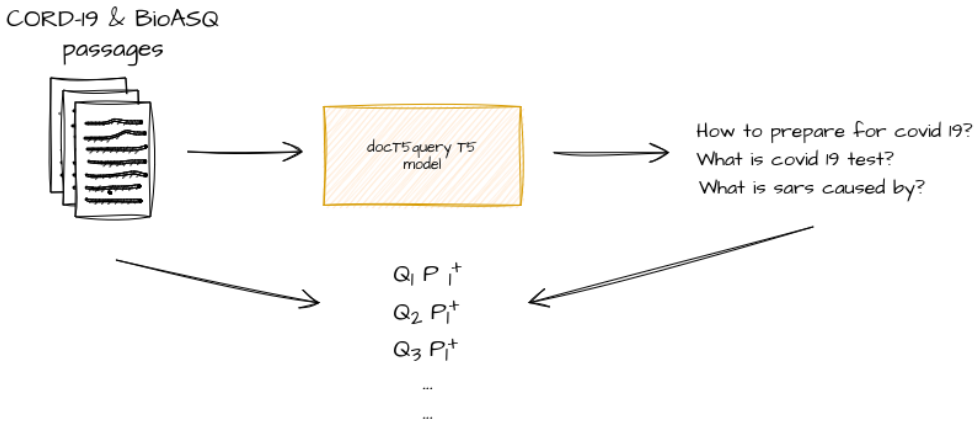

Figure 7: Question Generation with T5

Generated queries with pre-trained docT5query models are far from being perfect. In domain datasets where the vocabulary and the terms are domain specific and might not been seen during the model pre-training, can add noise to the generated queries. In fact, the query generation (GenQ) method for unsupervised training of model retrievers was introduced before GPL. GenQ is sensible to random and nonsensical queries, and the performance of the system rely on the quality of the synthetic queries. In contrast, GPL is not affected so much by the noise of the synthetic queries as a last data preparation step a cross-encoder is labeling the similarity of the pairs (generated query and passages) thus labeling a pair containing a noisy query as irrelevant.

Negative Mining. The dataset derived from the preceding step comprises pairs consisting of a generated query alongside the corresponding passage (Q, P^+). However, fine-tuning the baseline model solely on these pairs presents a potential drawback: the model may struggle to differentiate between genuinely relevant passages and those that bear similarity but are ultimately irrelevant to the query. To address this concern, negative passages are incorporated into a new dataset. By refining the loss function, the pre-trained model learns to distinguish the genuine positive passage P^+ from similar but irrelevant to the query passages P_i^-.

The negative mining process is an autonomous module of the experiment and can be configured in many ways. In the experiment, a combination of a sparse BM25 model and a pre-trained dense retriever were utilized. It is important to note that the pre-trained dense model that was utilized for negative mining has not been pre-trained on COVID-19 related terms. The integration of BM25 alongside the dense retriever, as demonstrated by BEIR research [24], has shown promising results in enhancing the retrieval of biomedical-related passages, particularly in datasets such as TREC-COVID and BioASQ Task b. *Msmarco-MiniLM-L-6-v3* was selected as a dense retriever, a bi-encoder sentence-transformer model [16]. The bi-encoder consists of 2 encoders, one for the query Q and one for the passage P. During training, the model tries to maximize the similarity score between the query Q and the relevant passage P^+ with the help of a contrastive loss function. At inference time calculates the similarity score between the query Q and each element of a document set D considering the document with the highest similarity score as the most relevant.

The document set of the training dataset is mapped to vector embeddings with the bi-encoder. The vector embeddings are then upserted to the Weaviate vector database [32] which facilitates retrieval process during the negative mining. Weaviate also offers out-of-the-box BM25 retrieval services making it perfect candidate for our training and production corpus storage database. The vector database is queried, requesting 50 relevant passages with each method (BM25, bi-encoder). From the database results, the golden passage is excluded that was originally used to generate the queries in the previous step (referred to as the positive passage). Subsequently, a new dataset is constructed wherein each training instance comprises a triplet: a query Q_i, the corresponding positive passage P_i^+, and a negative passage P_{ij}^-. This process is illustrated in Figure 8.

Pseudo-labeling with cross encoder model: In simple terms, a cross encoder is a neural network architecture which receives two input sequence in the input and produce a single output representing the similarity between the two inputs Figure 9.

Pseudo labeling is the final step of training data preparation. A cross encoder model was utilized to generate similarity scores from positive $sim(Q, P^+)$ and negative pairs $sim(Q, P^+)$. The objective is to train our baseline model (SBERT bi-encoder) using the margin Mean Squared Error (MSE) loss function. This function requires the margin

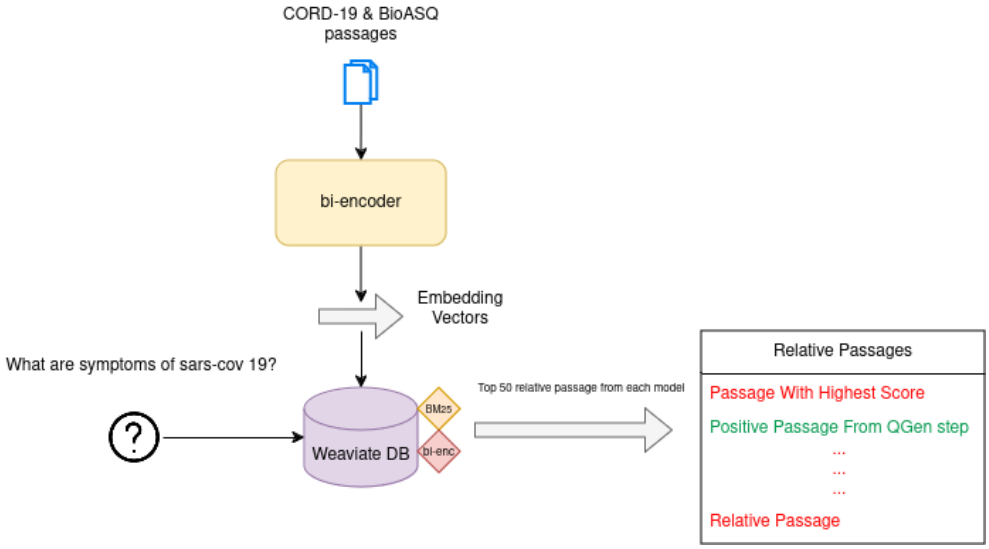

Figure 8: Negative Mining

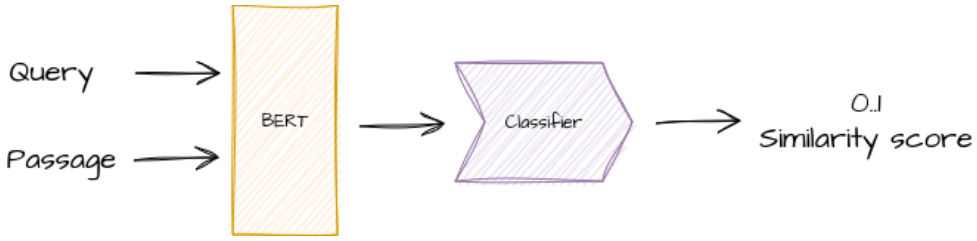

Figure 9: Cross Encoder

between the two scores calculated above. The margin is defined as $margin = sim(Q, P^+) - sim(Q, P^-)$ [37]. The training elements of the dataset produced in this step has the form of $(Q_i, P_i^+, P_i^-, Margin_\iota)$.

Training the bi-encoder. The last step is a traditional way of bi-encoder training. The training is achieved by optimizing the margin MSE loss function 5

$$L_{marginMSE}(\theta) = -\frac{1}{M}\sum_{i=0}^{M-1} |\hat{\delta}_i - \delta_\iota|^2 \tag{5}$$

- M is the number of elements of the dataset.
- δ_i represents the true label of the margin between the i_th pairs
- $\hat{\delta}_i$ represents the predicted value of the the margin of the i_th pairs

The bi-encoder is used to encode the positive and negative pairs (Q_i, P_i^+) and (Q_i, P_i^-), respectively. Similarity between the items of the pair is computed via the dot product of

their vectors. By subtracting the two similarities, concludes to the predicted margin δ_i as presented in Equation 6. Predicted margin value $\hat{\delta}_i$ is computed by the cross-encoder via pseudo labeling in the previous step.

$$\delta_i = sim(Q, P^+) - sim(Q, P^-) \tag{6}$$

The baseline model selected for the final step (training) of the GPL domain adaptation model is a distilbert-based model initially been trained on MS MARCO dataset. The model then was ported to sentence-transformer architecture as a bi-encoder *marco-distilbert-base-tas-b*. The model was evaluated with the MS MARCO benchmark and shown better performance among its competitors in terms of $nDCG$@10 [38]. The model was trained for 100,000 training steps for 1 epoch due to time limitation. Hardware constraints of the GPU (AMD RX 6700, 8GM RAM), imposed a batch of 8 during the training. The data process pipeline produced a dataset by primarily providing all the unique query-positive passage combination with one negative (random) passage. If the length dataset was surpassed by the training step the training generator provided subsequent negative passages. This way the model could be trained on a wider area of biomedical terms.

4.1.3 Alternative IR training methods

In the experiment, different methods were used to train the baseline models. Ultimately, it was concluded that GPL was the most effected architecture for the core of the proposed system. This section presents alternative methods implemented to enhance IR performance.

Dense Passage Retrieval (DPR). Dense retrieval methods have emerged as a powerful approach in IR systems, offering a significant leap over traditional sparse retrieval techniques, such as TF-IDF and BM25. Unlike sparse retrieval, which relies on exact keyword matches between query and document terms, dense retrieval methods utilize DL models to encode both queries and documents into continuous vector spaces. These methods leverage the semantic meaning of text, enabling the retrieval system to understand and match the intent behind queries and the information within documents more effectively.

The first dense custom model was implemented based on the dense passage model proposed by Vladimir Karpukhin [18]. In this paper the proposed model outperformed BM25 systems by 9% - 15% in terms of top-20 passage retrieval accuracy, in various open-domain QA benchmark datasets (NQ, TriviaQA, WQ, TREC, SQuAD). The baseline model did not performed well during the evaluation of our domain specific benchmark dataset TREC COVID [22].

To increase the quality of IR, a training process of the DPR model was conducted. DPR are trained in a specific way as is indicated in the paper. Training the model's encoders, involves building a vector space where relevant elements of the question-passage pairs will be closer (more similar) than the elements of the irrelevant pairs. In order to train the model effectively, it was imperative to curate a dataset where each element adhered to a specific format:

$$L(q, p^+, p_1^-, \ldots, p_n^-) \tag{7}$$

where q is the query, p^+ is the relevant passage and p_n^- are the n irrelevant passages for this query. The overall goal is to optimize the loss function as the negative log likelihood of the positive passage (8).

$$-\log\left(e^{\text{sim}(q_i,p_i^+)}\right) + \frac{e^{\text{sim}(q_i,p_i^+)}}{e^{\text{sim}(q_i,p_i^+)} + \sum_{j=1}^{n} e^{\text{sim}(q_i,p_{i,j}^-)}} \tag{8}$$

The BioASQ Task B dataset [34] provides potential answers and corresponding contexts for each query. To align with the specifications of the DPR training process, several pre-processing steps are undertaken. Initially, the dataset is filtered to include only factoid and list type responses, facilitating the conversion to the SQuAD format for subsequent evaluation. Subsequently, the filtered dataset is converted into SQuAD format and up-load it into an ElasticSearch database for efficient data mining. The next step involves labeling the first relevant context as the positive context. However, identifying negative contexts poses a challenge. Building on insights from Karpukhin et al. [18], which suggest that incorporating more negative samples can enhance model performance, the following strategy is adopted for negative context retrieval. For each question, the BM25 algorithm [9] is employed to retrieve up to 30 relevant documents. Subsequently, "hard negatives"

are designated as those documents that do not pertain to the relevant (positive) contexts associated with the given question. The selection of the latest BioASQ dataset as the train dataset was made due to the inclusion of COVID-19 terms in its data.

Sentence Embeddings using Siamese BERT-Networks. The evaluation of the DPR model demonstrated that the biomedical domain adaptation was poor. Our research lead us to dense retrieval methods based on Siamese BERT-Networks [16]. Sentence-BERT (SBERT) is an adaptation of the BERT network [13] that incorporates siamese and triplet networks to generate semantically meaningful sentence embeddings. This modification enhances BERT's capabilities, allowing it to address tasks that were previously inaccessible. These tasks include large-scale semantic similarity comparison, clustering, and semantic search for IR. BERT [13] becomes very slow as the corpus get bigger due to the use of cross encoder which computes the similarity between two sentences (here query and passage). To retrieve the most relevant passage from the corpus C to a given query q, n computations must be performed where n is the length of the corpus C. In contrast, SBERT aims to derive fixed-size vectors for sentence inputs. Consequently, all equivalent vector embeddings of corpus passages can be computed once and stored in a vector database. During inference, only the query needs encoding, facilitating retrieval of the k most similar passages. In Figure 10 the architecture of SBERT model is presented.

Fine tuning a bi-encoder is a straightforward process and it can achieved by optimizing MNRL (Multiple Negatives Ranking Loss) loss function presented in the paper entitled *"Efficient Natural Language Response Suggestion for Smart Reply"* [39]. The model tries to create a clear distinction between positive and negative passages concerning a query. During training, the model tries to increase the distance in a vector space between positive and negative passage vectors in relation to a query vector as it is shown in Figure 11.

The MNRL method require a dataset in the form of triplets. The triplets include a query Q, a positive passage P^+ and a negative one P^-. The product model of this model behaves rather poorly as it is shown later in the evaluation section. This is due to false negative passages selection. The negative passages was selected via BM25 query, declaring all related passages for a query as negatives except the actual positive.

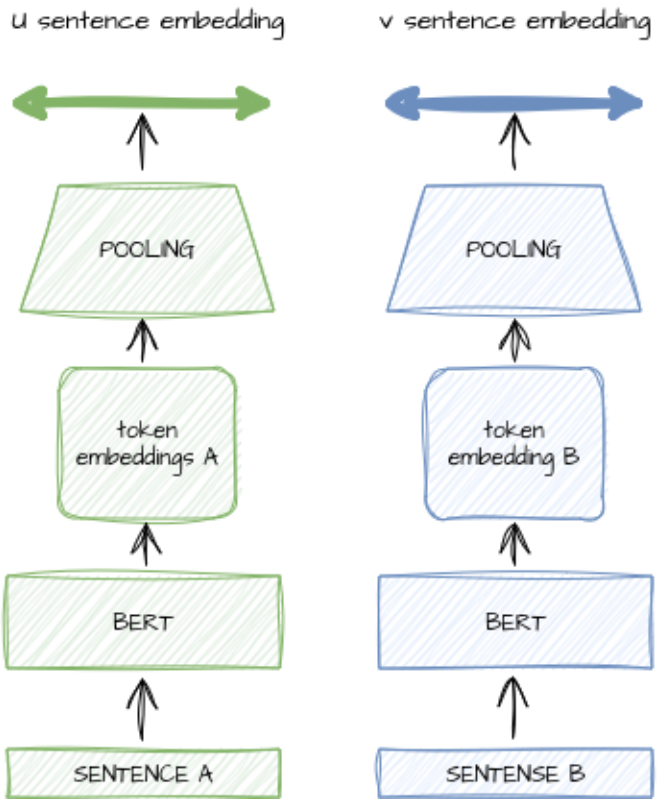

Figure 10: Bi-Encoder

4.1.4 Proposed IR Architecture

Following the evaluation research, it is deduced that the SBERT model, trained utilizing the GPL method, exhibited superior adaptability within the biomedical domain. All passages from the CORD-19 corpus were encoded using the proposed adapted bi-encoder and subsequently integrated into a Weaviate vector database for future retrieval purposes.

Weaviate is a vector database built in the GO language, designed to store both vectors and objects, thus enabling the combination of vector searching and structured queries. It also provides out-of-the-box BM25, ideal for standalone or hybrid retrieval architectures. Weaviate stores data both on disk and memory. Disk storage is used for persistence and data durability in case of an emergency event occurs (shutdown, restart). Data are also cached in memory for faster access and retrieval, optimizing the overall performance of the system.

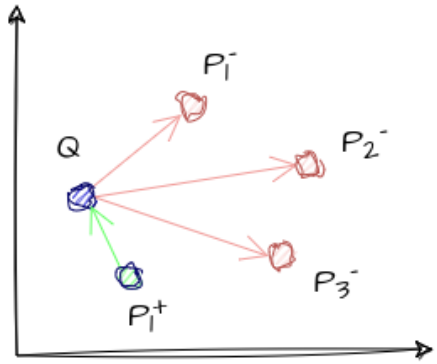

Figure 11: query-passages relation

Indexing corpus passages into the mentioned database was a time-consuming procedure. Due to limitation in both time and hardware resources, only 2,867,150 of 11,843,565 passages were successfully encoded and uploaded. Each vector is of length of 768 (embedding length). The procedure required approximately 9 hours to complete, encompassing both encoding and uploading the passages into the database.

An evaluation of a dense-sparse hybrid retrieval system was conducted, incorporating our GPL trained model alongside the BM25 retrieval model. A pre-trained cross-encoder on MS-MARCO dataset was added, reranking retrieved passages as the last step of the retrieval pipeline, taking advantage of the demonstrated power of the model. The architectures of cross encoders and their use cases have been previously discussed. Observations indicated a notable enhancement in the IR system's performance in terms of $NDCG$@10. The proposed IR system is shown on Figure 3.

4.2 Reader System

In this section, the reader system of the proposed architecture is presented. The reader model comprises two components: a pre-trained transformer-based model and a classification head. The transformer model identifies semantic relations between the query and the provided context, facilitating the extraction of relevant information. The classification head determines the possibility of token being the start or the end of a possible answer. Through this dual-functionality mechanism the reader model effectively identifies and extract possible answers included in provided contexts, given a query.

The foundation of the approach lies in leveraging state-of-the art transformer-based BERT models. BERT bidirectional architecture allows it to capture contextual information effectively, make in it well-suited for natural language understanding tasks. For baseline models, BERT-based models pre-trained on generic datasets such as BookCorpus and English Wikipedia are utilized. The architecture of BERT model is presented in Figure 12.

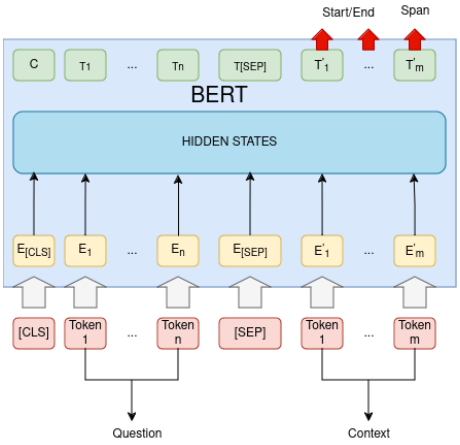

Figure 12: Fine-tuning BERT for QA

To adapt the pre-trained BERT-based models for QA tasks, fine-tuning was performed on the SQuADv2 (Stanford Question Answering Dataset) dataset. SQuADv2 consists of real-world questions posed by crowdworkers on a diverse set of Wikipedia articles, along with the corresponding passage spans containing the answers. The Version 2 of this dataset also includes query-passages pairs where they cannot be answered. By fine-tuning the model on squad format, model's parameters are optimized to minimise the loss function, which measures the discrepancy between the predicted and the ground truth answers. This process adapts the model to QA tasks, refining its ability to extract accurate answers from given contexts.

Due to the critical need for reliable QA systems in the biomedical domain, further refinement was conducted on the fine-tuned BERT-model (on SQuAD dataset) using the COVID-QA dataset [23]. This dataset comprises 2,019 question/answer pairs annotated by biomedical experts on scientific articles related to the COVID-19 pandemic, sourced from reputable sources such as medical forums, research articles and healthcare guidelines. By fine-tuning the model on the COVID-QA dataset, the aim was to enhance the model's

performance specifically in the biomedical domain. The dataset was already in SQuAD format which facilitated the pre-processing step.

Transformer based models encounter a specific challenge when dealing with long-text passages. Notably, datasets such as SQuAD and COVID-QA, utilized as our training data, encompass numerous answering contexts exceeding 300 words in length. BERT models employed in our training pipeline are constrained by a maximum length limit of 512 tokens, corresponding down to approximately 300-350 words. This discrepancy necessitates careful consideration in pre-processing strategies to effectively handle longer passages while adhering to the model's input constraints. It is noteworthy to mention that the Haystack framework [40], which is used to train the reader models, and to build the proposed QA system, incorporates pre-processing tools tailored to address challenges associated with lengthy text passages. Specifically, Haystack framework provides truncation, padding and chunking methods, enabling seamless processing of our data with reduced code complexity.

Our exploration of related work within the biomedical domain led us to uncover the BioBERT [41] model, a significance advancement suited specifically for biomedical tasks. BioBERT is pre-trained on large-scale biomedical corpora, including PubMed abstracts and PubMed Central full-text articles, enabling it to capture domain-specific knowledge and terminologies essential for biomedical applications. BioBERT offers versatility through fine-tuning, allowing customization for various objectives such as entity extraction and QA.

During evaluation process, our fine-tuned model based on BioBERT demonstrated the most promising results. Our proposed reader model leverages BioBERT, initially fine-tuned on the SQuAD dataset, followed by further refinement the COVID-QA dataset. This sequential fine-tuning process enhances the model's understanding of both general QA tasks and specialized biomedical inquiries.

4.3 Generative Reader System

As part of a comprehensive approach to build a biomedical QA system, the utilization of a T5-FLAN (Text-to-Text Transformer) generative model [33] is introduced as a reader component. The T5-FLAN model is based on encoder-decoder transformer architecture,

renowned for its effectiveness in natural language understanding and generation tasks. T5-FLAN employing text-to-text transfer learning, wherein both inputs and outputs are representing text sequences. This feature allows the model to handle various tasks including QA, summarization and translation.

In our biomedical QA system, the T5-FLAN generative model serves as an alternative to the extractive model as a reader component. It is responsible for generating answers to user queries based on retrieved passages from the IR system. In our system.

The T5-FLAN model is utilized at inference as a zero-shot model without undergoing further fine-tuning. The decision not to fine-tune the FLAN model was primarily influenced by the significant computational resources required to accommodate its large size. Despite recognizing potential benefits of fine-tuning hardware limitations, rendered this solution unfeasible.

5 Results

In the following section, the results of the experiments performed during the building of the proposed QA system, which was described in Section 4 are presented. This experiment analysis presents the efficacy and the robustness of our solution in accurately retrieving relevant information and providing precise answers to user queries related to the biomedical domain. Based on innovative methodologies and established metrics, an examination of the proposed system's performance across biomedical benchmarking datasets is performed, aiming to highlight its strength, limitations, and potential enhancement.

5.1 Hardware

One of the goals of this experiment is to develop a product that can be trained at low cost, perform well, and be easily deployed within any healthcare unit. The goal is to empower healthcare units, especially those with limited resources, financially and in terms of hardware, by developing a system that is cost-effective, robust, and accessible. Preprocessing steps of the datasets utilized in this experiment, fine-tuning transformer models

except GPL domain adaptation, indexing and database management and the development were performed in a personal PC with the following specs:

- Ryzen 9 6000 series, 8 high performance cores, 4.9GHz max boost clocking.
- 40GB RAM Memory.
- AMD Radeon™ RX 6700S, with 28 compute units and 8GB of Memory.

Unfortunately, the training phase of the bi-encoder component in the GPL method faced obstacles within the specified infrastructure due to memory constraints. Consequently, the decision was made to train the bi-encoder transformer model in Kaggle's infrastructure, leveraging the computational power of a P100 GPU with 16GB of memory. This shift allowed the smooth execution of the process, overcoming the memory constraints imposed by the previous infrastructure. The model was successfully trained over 50,000 steps, utilizing a batch size of 8 to maximize efficiency and optimize performance.

5.2 Metrics

In the results section, separate evaluations were conducted for the QA system, with one focusing on the IR component and the other on the reader component. To ensure a comprehensive assessment, well-known benchmark datasets relevant to the study domain (biomedical) were utilized. Each evaluation method employed distinct metrics tailored to the specific requirements of the method.

5.2.1 Information Retrieval Metrics

CappedRecall@k **Score**. Recall in K is the fraction of the relevant items returned by the model successfully (true positives), within the set of the relevant items that exists in the entire dataset 9. K refers to the number of the items retrieved by the IR system. As for the IR task, recall reflects the ability of the model to correctly return all the relevant documents according to a query.

$$Recall@K = \frac{|\{\text{relevant documents}\} \cap \{\text{K retrieved documents}\}|}{|\{\text{relevant documents}\}|} \tag{9}$$

TREC-COVID dataset averaging 300 relevant passages per query. Some queries can contain up to 500 relevant passages. The $Recall@K$ defined above 9 would be deceptive for the dataset, miss-evaluating the model. Consider that a query Q_i included in the TREC-COVID dataset, contains 500 relevant passages $Passages$. By retrieving all the relevant passages, the $Recall@10$ score for this query would be 0.02, which is quite low even though the model retrieves successfully 10 relevant documents. To correct this issue, the capped recall [24] was introduced. $CapedRecall@K$ is defined as the number of the relevant documents returned by the IR system, within the minimum of K and the total number of relevant documents 10.

$$CappedRecall@K = \frac{|\{\text{relevant documents}\} \cap \{\text{K retrieved documents}\}|}{|\min\{\text{K}, \text{relevant documents}\}|} \tag{10}$$

MRR **Score**. Mean Reciprocal Rank in contrast with $R@K$ defined above, is an order-aware metric. MRR takes into account the rank or order of the relevant documents. Specifically, MRR measures the mean reciprocal rank of the relevant documents, considering their descending order of importance. MRR is calculated based on multiple queries as it is shown below:

$$\text{MRR@K} = \frac{1}{Q} \sum_{q=1}^{Q} \frac{1}{\text{rank}_q} \tag{11}$$

where:

- Q denotes the total number of queries
- $rank_i$ is the rank of the first relevant result.

While MMR provides insights into the rank of the first relevant retrieved item, it may not capture the overall performance of the IR system, when dealing with multiple relevant documents.

$NDCG@K$ **Score**. $NDCG@K$ stands for Normalized Discounted Cumulative Gain at K. It is a powerful tool in the hands of the analyst. It is an order-aware metric that is derived for simpler one. Unlike traditional metrics such as $Recall$ and MRR, considers both the relevance and the position order of retrieved documents. In short, $NDCG$ ensures

that higher ranks and highly relevant documents receive greater weights, providing a clear picture of the retrieval capacity of the IR system.

$NDCG@K$ sums up the discounted gains of the relevant documents up to rank K, normalized by the ideal $DCG@K$ value. The discount factor is determined by the logarithm of the position of the document in the returned list, reflecting the attenuation of importance moving down the list.

$$\text{NDCG@K} = \frac{1}{Z} \sum_{i=1}^{K} \frac{2^{rel_i} - 1}{\log_2(i+1)} \tag{12}$$

where:

- K represent the rank cutoff.
- rel_i represent the relevance of the document at $rank_i$.
- Z is the formalization factor. It is the ideal $DCG@K$ value for the given set of queries.

5.2.2 Reader Metrics

Exact match EM **Score** is ta binary metric that measures whether a model predicted the answer exactly. If the model answer prediction matches the ground truth answer , EM is 1, otherwise is 0. Overall, EM is the proportion of instances where the model's answer prediction matches exactly the correct answer provided by the dataset as it is shown below:

$$EM = \begin{cases} 1 & \text{if predicted answer} = \text{ground truth answer} \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

$F1$ **Score** is a metric less strict than EM. It measures word overlap between model's predicted answer and the labeled. It considers two sub-metrics: precision and recall. Precision is the common words (predicted, labeled) to the total words of the predicted answer. Recall is the ratio of the number of common words to the number of words in ground truth. The math formula is as follows:

$$F1 = 2 * \frac{precision * recall}{precision + recall} \tag{14}$$

5.3 IR System Evaluation

In this section, the experiments conducted to evaluate the methodologies and implementation described in Section 4 are presented. Each component of the system was evaluated separately, with appropriate benchmark datasets representative of real-world scenarios chosen to ensure the robustness and reliability of the findings. Specific metrics tailored to describe the performance of each component were utilized in the evaluation. The results were compared with state-of-the-art models and methodologies to assess the performance of the proposed system.

5.3.1 Datasets and knowledge source

In the thesis, as outlined in the methodology described in Section 4, two biomedical benchmark datasets, BioASQ and TREC-COVID, were employed for evaluation purposes. BEIR is a widely used specific dataset format in the IR task, on which evaluation experiments were based. Typically, this format comprises three main components: a file containing the queries, each identified with a unique ID, a corpus file containing the passages, and a relation file containing the query-passage relation labeled by the score of the relation.

BioASQ Task 9a & 9b. BioASQ Task 9a v.2020 is a collection of 14,913,939 English article abstracts sourced from PubMed within the biomedical domain [42]. Document chunking was not performed during pre-processing. The dataset is consisted from abstracts, each spanning approximately 250-300 words in length. This length falls within the processing capabilities of our GPL model, as well as most and most of the experimental models. To transform BioASQ dataset to the BEIR format, BioASQ Task 9b 2021 dataset was processed, comprising 3,742 queries encompassing various type of answers such as factoids, list answers and yes/no. All labeled passages were considered as relevant, denoted by a binary label of 1. It is important to note that GPL guidelines [25] recommend reducing the corpus size to one million for more efficient evaluation; however, the reduction

was not opted for, and the original corpus size was used following BEIR instructions [24], allowing direct comparisons with BEIR benchmark for the BioASQ dataset.

TREC-COVID. TREC-COVID dataset is a collection of documents and queries curated specifically for research related to the COVID-19 pandemic. The primary objective of this dataset is to evaluate the efficacy of a retrieval model operating within the biomedical domain, with a specific focus on the COVID-19 topic.

To evaluate the models, the round 5 document and judgments were utilized, relying on version 2020-07-16 of the CORD-19 dataset as a corpus. It contains 50 topics/queries, each labeled to corpus passages as follows: 0 unrelated, 1 partially relevant, 2 fully relevant. In our evaluation experiment, all passages containing only the title of the article were eliminated, as instructed by the GPL method [25]. The final evaluation dataset consisted of 129,192 passages, 50 queries, and 21,541 query-passage relations. Additionally, all relations labeled as zero were excluded.

CORD-19. CORD-19 is a comprehensive and valuable dataset, particularly in the biomedical domain. The final update, which performed on 2022-06-02, includes more than 1,000,000 scholarity articles, focused on COVID-19, SARS-CoV-2, and coronaviruses in general. It enables researchers to extract crucial insights, identify trends, and develop innovative solutions in the fight against COVID-19. As a trusted repository of knowledge, it serves as an initial knowledge base of our proposed QA system.

5.3.2 IR Evaluation Results on TREC-COVID

The evaluation of our proposed QA system in biomedical domain, with a specific emphasis on COVID-19, was conducting using the BEIR benchmark framework, which is formed of various IR datasets and tools, offering a robust robust platform for our model performance evaluation experiments. BEIR includes three datasets within the biomedical domain: BioASQ, TREC-COVID and NFCorpus. However, our primary attention was focused to the TREC-COVID dataset. Specifically, metrics such as $NDCG@k$ and $Recall@k$ was employed to evaluate the IR system performance. Given the large size of the

BioASQ corpus dataset, the evaluation process was limited to the final proposed model, comparing results with the original BEIR benchmark [24].

Table 1: IR Evaluation on TREC-COVID dataset using $nDCG@K$. Bold indicate the best result.

Model	**nDCG@1**	**nDCG@10**	**nDCG@100**
BM25	0.81	0.65	0.45
msmarco-distilbert-base-tas-b	**0.82**	0.68	0.48
DPR base	0.22	0.23	0.14
DPR fine-tuned	0.30	0.24	0.15
msmarco-distilbert-base-tas-b + CE	0.76	0.72	0.46
BM25 + CE	0.74	0.70	0.50
GPL on CORD19 full-articles	0.76	0.72	0.52
GPL on BIOASQ + CORD19	0.77	0.64	0.46
GPL + BM25 + CE	0.75	**0.73**	**0.55**

Table 1 presents an evaluation of various IR models on the TREC-COVID dataset, measured used the $nDCG@K$ metric, where higher values indicate better retrieval performance. Retrieving relevant documents at the top of the ranking list (just one element), it is described with $nDCG@1$, where the base SBERT model *msmarco-distilbert-base-tas-b* performed better with a score of 0.82. The *msmarco-distilbert-base-tas-b*, is a DistillBert-based model, trained with Balanced Topic Aware Sampling [43] on the MSMARCO dataset [38], then ported to SBERT [16] architecture.

On the other hand, models like DPR, both in its base and fine-tuned forms, perform comparatively poorer, indicating potential limitations in their retrieval capabilities on this dataset. As it was already demonstrated by GPL paper [25], BM25 perform rather well on biomedical domain benchmark such as BioASQ and TREC-COVID.

The evaluation results highlight the significant enhancement of the IR system by the GPL method, particularly when consider $k = 10$ and $k = 100$. These parameters represent crucial retrieval scenarios where users often expect results within many passages. Using GPL as a hybrid system alongside BM25, complemented by a cross-encoder ranker, showcased the system's superior performance, particularly in retrieving 10 and 100 passages.

This hybrid approach demonstrates the strength og both traditional probabilistic retrieval methods like BM25 and advanced transformer-based models like SBERT.

Regarding the calculation of *Recall@k* with the BEIR tool, recall is calculated utilizing all relevant documents as the denominator. In scenarios where the number of relevant documents for each query varies, as TREC-COVID dataset has an average of 500 relevant documents per passages, the recall scores can lead to false performance conclusions. To address this issue, as BEIR suggests, is to compute the capped recall metric. This metric adjusts the calculation of *Recall@k* by considering the minimum between the number of relevant documents available and the specified value of k.

The capped recall metric, in this experiment behaves similarly to the *nDCG@k* metric, providing insights into how well a model can retrieve truly relevant passages related to queries. The hybrid proposed system demonstrates superior performance compared to other experimental models, particularly for $k = 10$ and $k = 100$. The computed *Recall@k* values are demonstrated in the following Table 2.

Table 2: IR Evaluation on TREC-COVID dataset using *Recall@K*. Bold indicate the best result.

Model	**R@1**	**R@10**	**R@100**
BM25	0.81	0.65	0.45
msmarco-distilbert-base-tas-b	**0.88**	0.68	0.45
DPR base	0.32	0.25	0.14
DPR fine-tuned	0.32	0.25	0.14
msmarco-distilbert-base-tas-b + CE	0.84	0.75	0.46
BM25 + CE	0.80	0.75	0.50
GPL on CORD19 full-articles	0.82	0.77	0.51
GPL on BIOASQ + CORD19	0.82	0.67	0.41
GPL + BM25 + CE	0.78	**0.78**	**0.58**

5.3.3 IR Evaluation Results on BioASQ

Permission constraints preventing researches to load BioASQ dataset, with the corresponding format. Instead, they must be authorized by BioASQ source to legally download the dataset. The dataset was processed for evaluation, as instructed by BEIR [24] and GPL [25] papers. The original corpus size from 14,913,939 documents to 1,000,000, due to hardware constraints. The same constraints also impact the evaluation process on

BioASQ dataset, preventing the assessment to every individual experimental model. The evaluation focused solely on the GPL model trained on both CORD-19 full-text articles and BioASQ abstracts. The results are compared with those of GPL original paper [25]. Comparisons are demonstrated on Table 3.

Table 3: *nDCG@k* performance of several models, included our trained model, on BioASQ dataset

Model	**nDCG@10**
BM25	0.47
BM25 + CE	0.52
msmarco-distilbert-base-tas-b	0.34
GPL on BIOASQ + CORD19	**0.59**

5.4 Extractive Reader Evaluation

In this section, the evaluation performed on several fine-tuned experimental models based on BERT architecture [13] is presented. Evaluation is performed on two biomedical related QA datasets.

Covid-QA [23] is a SQuAD [19] formatted dataset, comprised by 2,019 question-answers pairs, annotated by biomedical experts on 147 scientific articles extracted from CORD-19 dataset. It differs from SQuAD dataset in terms of context passage length, where in CORD-19 the average token per passage is 6,118.5 versus 153.2 in SQuAD. The context passages within the corpus exhibits an average word count of 4,889.2, while question and answers presents modest lengths, with average of 9.2 and 13.9 words.

BioASQ QA dataset [34] is a biomedical benchmark dataset including questions in the English language with golden answers, annotated by biomedical experts. Moreover, the dataset contains metadata such as ideal answers and ontology keywords which are not used in this experiment. The version 7b of this dataset was selected during the experiment. BioASQ dataset contains multiple type of answering. The evaluation performed solely on the factoid type answer set. The dataset is transformed according the SQuAD format for easier training and evaluation. The corpus passages related to the answers of the dataset are parsed through a custom web scraping pipeline, querying the E-utilities PubMed API [44].

To ensure efficient processing without compromising comprehension, lengthy contexts was partitioned into passages, each one consisting of 256 tokens-the maximum sequence length possible indicated by the baseline models. Furthermore, a passage stride of 128 tokens was performed.

In the experimental setup with the benchmark datasets mentioned above, three distinct baseline models were employed. RoBERTa [14], BERT [13] and BioBERT [15]. These models are variants of the Transformer architecture, each offering unique advantages for NLP tasks. A notable difference between BERT-based models (BERT, BioBERT) and RoBERTa, is in the handling of masked tokens during training. In BERT, the mask is predetermined during pre-processing and remains static through training, whereas in RoBERTa, the mask label is dynamically changed during each epoch of training. BioBERT is a specialized variant of BERT model. It has been pre-trained on domain-specific biomedical-data, which enables it to capture domain-specific knowledge and terminology effectively.

To prepare the baseline models for QA tasks, they were fine-tuned using the SQuADv2 [19] dataset. By fine-tuning with SQuAD, the aim was to equip the models with the capability to accurately extract answer spans from passages. The model was trained on the SQuAD dataset for 10 epochs using a learning rate of 0.00001. The evaluation performed on the dataset's test split is shown below.

Table 4: Reader models fine-tuned on SQuADv2, QA evaluation

Model	**F1%**	**EM%**
bert-base-cased-squad	77.45	75.21
roberta-base-squad	**84.24**	**82.12**
biobert v1.1-cased-squad	52.11	48.94

As a next experiment, the performance of SQuAD fine-tuned models on previously unencountered biomedical datasets was investigated. Specifically, the efficacy of these models was investigated on two distinct datasets: COVID-QA and BioASQ. Each dataset presents unique challenges. The COVID-QA dataset, structured in the SQuAD format, comprises extensive passages spanning thousands of words. Conversely, the BioASQ dataset diverges in nature, featuring factoid-style questions, which demand concise, direct, and short an-

swers. The evaluation scores of the SQuAD fine-tuned models are presented in Table 5.

Table 5: QA reader models, Evaluation on biomedical datasets

	COVID-QA		**BioASQ 7b**	
Dataset	**F1%**	**EM%**	**F1%**	**EM%**
bert-base-cased-squad	26.54	15.10	21.60	15.70
roberta-base-squad	29.15	**18.31**	20.85	16.95
biobert v1.1 base-cased-squad	27.61	16.83	27.38	23.23
biobert v1.1 large-cased-squad	**32.88**	**18.31**	**37.43**	**31.16**

It is surprising to observe that the BioBERT model v1.1, which was pre-trained on PubMED articles targeting the biomedical domain, performs poorly on the COVID-1QA dataset. This behaviour may be attributed to the fact that the BioBERT model was trained before the COVID-19 pandemic, thus it may lack specific knowledge related to the pandemic. Instead, the model focuses on general biomedical information available in the pre-pandemic literature.

To address this issue, fine-tuning the models on COVID-19 related data was considered. As the training dataset, the trained split CORD-19 dataset was utilized. The models were trained for 5 epochs and then evaluated on the biomedical datasets. Due to hardware and time constraints, the base version of the BioBERT model was selected for further fine-tuning. Results are presented in the following table 6.

Table 6: QA reader models fine-tuned on COVID-QA , Evaluation on 2 biomedical datasets

	COVID-QA		**BioASQ 7b**	
Dataset	**F1%**	**EM%**	**F1%**	**EM%**
bert-base-cased-squad-covidqa	34.11	20.14	21.50	13.64
roberta-base-squad-covidqa	53.42	26.73	31.96	20.70
biobert v1.1 base-cased-squad-covidqa	**55.03**	**37.87**	**40.12**	**31.17**

The *roberta-base-squad-covidqa* markedly outperforms the BERT-based model on COVID-QA dataset. However, its performance on the BioASQ 7b dataset, while improved compared to BERT, suggests only a moderate superiority. The *biobert v1.1 base-cased-squad-covidqa* model demonstrated better performance, compared to *roberta-base-squad-covidqa* on the COVID-19 dataset, proving its effectiveness in understanding and processing pan-

demic related information. The *biobert v1.1 base-cased-squad-covid* model outperforms the other models on the BioASQ 7b dataset. This underscores the model's enhanced capability to generalize across different biomedical domains, likely due to its pre-training on a large biomedical corpus.

In the next phase of the experiment, the SQuAD v2 trained models were refined by conducting fine-tuning on the BIOASQ 7b factoid subset for 5 epochs. Following this fine-tuning process, the model's performance across the two biomedical benchmark datasets, COVID-QA and BioASQ 7b factoid subset, was meticulously evaluated. The results are presented in Table 7.

Table 7: QA reader models fine-tuned on BioASQ 7b factoids, Evaluation on 2 biomedical datasets

	COVID-QA		**BioASQ 7b**	
Dataset	**F1%**	**EM%**	**F1%**	**EM%**
bert-base-cased-squad-bioasq	27.13	16.89	37.13	30.60
roberta-base-squad-bioasq	**30.25**	**17.82**	43.67	40.77
biobert v1.1 base-cased-squad-bioasq	28.16	16.35	**56.96**	**45.10**

In the final phase of the reader evaluation experiment, an in-depth assessment of the BioBERT QA model's performance was conducted. Initially trained on the COVID-QA dataset, followed by fine-tuning on the BioASQ 7b dataset, unsatisfactory outcomes were encountered. The subsequent training adversely affected the model's ability to answer COVID-19 related questions, as it appears to prioritize adapting to general biomedical factoid answers. To address this issue, a new dataset was created. The dataset is a shuffled and merged version of COVID-QA and BioASQ 7b factoid subset. The BioBERT SQuAD model was trained for five epochs with the modified dataset. The results were notably promising, as the model demonstrated adaptability on both COVID-19 related questions and general biomedical queries as it is shown on Table 8.

Table 8: BioBERT initially fine-tuned on SQuAD v2 and then on the biomedical dataset consisting of COVID-QA set and BioASQ 7b, Evaluation on 2 biomedical datasets

Dataset	**F1%**	**EM%**
COVID-QA	51.90	27.50
BioASQ 7b factoids	52.16	48.55

6 Covid WISE. A Covid QA Application

After conducting extensive experiments and evaluations, the most appropriate models for both retrieval and reading within our QA Web application have been carefully selected. Our approach involves indexing a significant portion of the CORD19 corpus (2,867,150 passages) into embeddings within the Weaviate vector database [32], using a biomedical adapted retriever model trained on both COVID-19 and general biomedical corpus. Additionally, the Weaviate database offers BM25 indexing, enhancing the retrieval capabilities of the server. Every indexed passage was chunked into approximately 200-word segments, to be processed correctly by all involved models in the pipeline.

The users of our system have the flexibility to choose between different retrieval modes. They can opt for solely dense retrieval or a hybrid approach combining BM25 lexical search. Furthermore, users can specify the number of articles they wish to retrieve, tailoring the search results to their requirements. Our application provides users with multiple options for interacting with the retrieved information. They can navigate through passages, explore relevant content, or attempt to find or generate answers based on the retrieved passages. These features aim to assist and guide users in efficiently obtaining the information they seek.

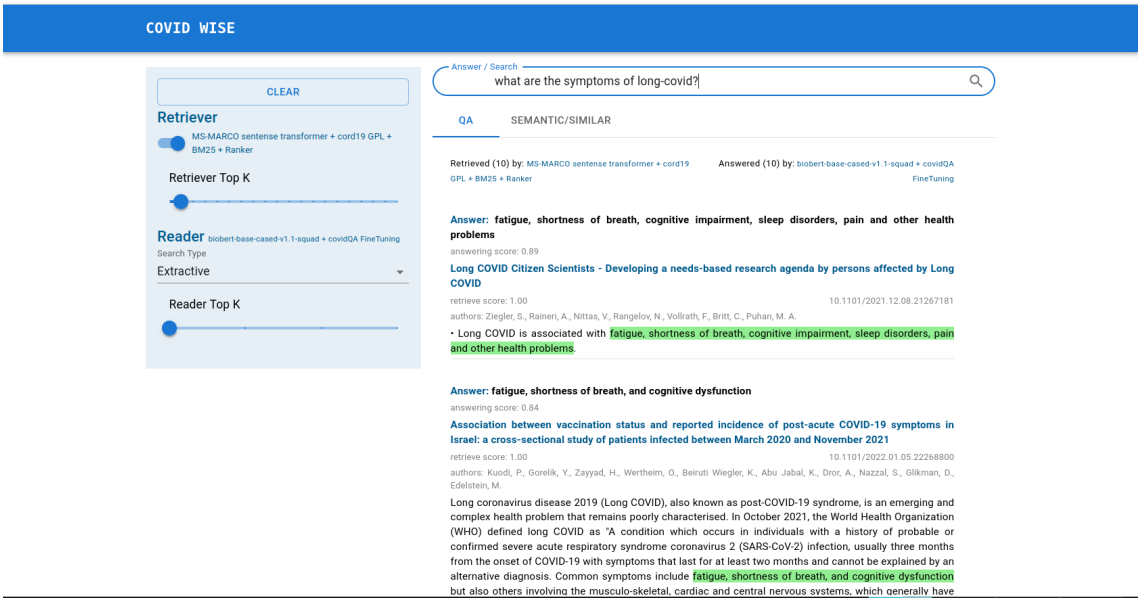

Figure 13: Covid WISE QA Application

6.1 Querying Covid WISE

The evaluation process demonstrated good performance of the proposed QA system. The system's performance was tested as components and not as an individual system. Sometimes QA systems perform well during evaluation using benchmark datasets, but behaving rather bad into real-life scenarios.

6.1.1 What do we know about COVID-19 risk factors?

This subsection demonstrates a real user query and the system's behaviour in the two available architectures. Figures 14 15 16 display the results generated by the QA application, utilizing the hybrid IR model as a retriever and the COVID-QA fine-tuned BioBERT model as a reader. Figure 14 presents the most relevant results retrieved by the hybrid IR model, followed by subsequent figures displaying subsequent relevant outputs in descending order of relevance.

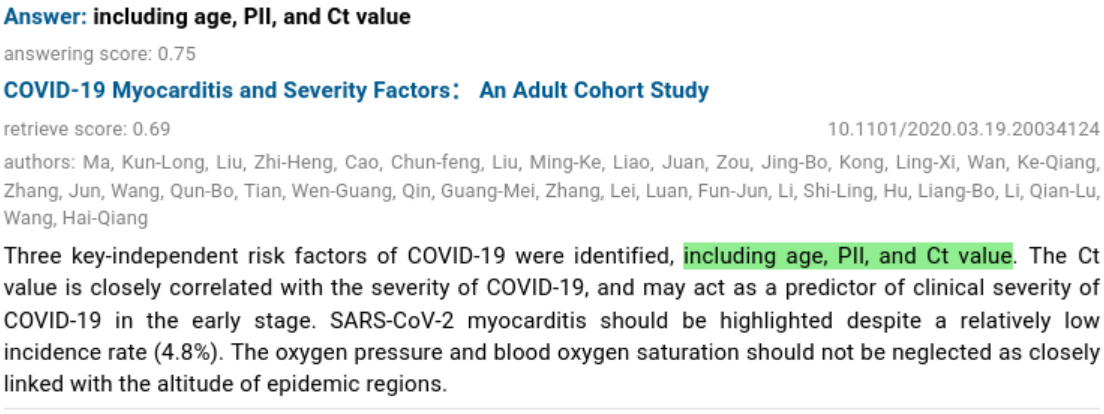

Figure 14: 1st most relative answer to question "What do we know about COVID-19 risk factors?"

The system's generative answer using the T5-FLAN model utilizing the hybrid retriever is presented at 17

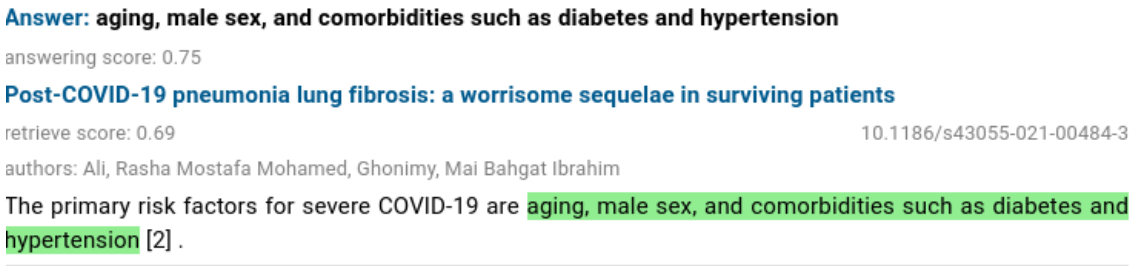

Figure 15: 2nd most relative answer to question "What do we know about COVID-19 risk factors?"

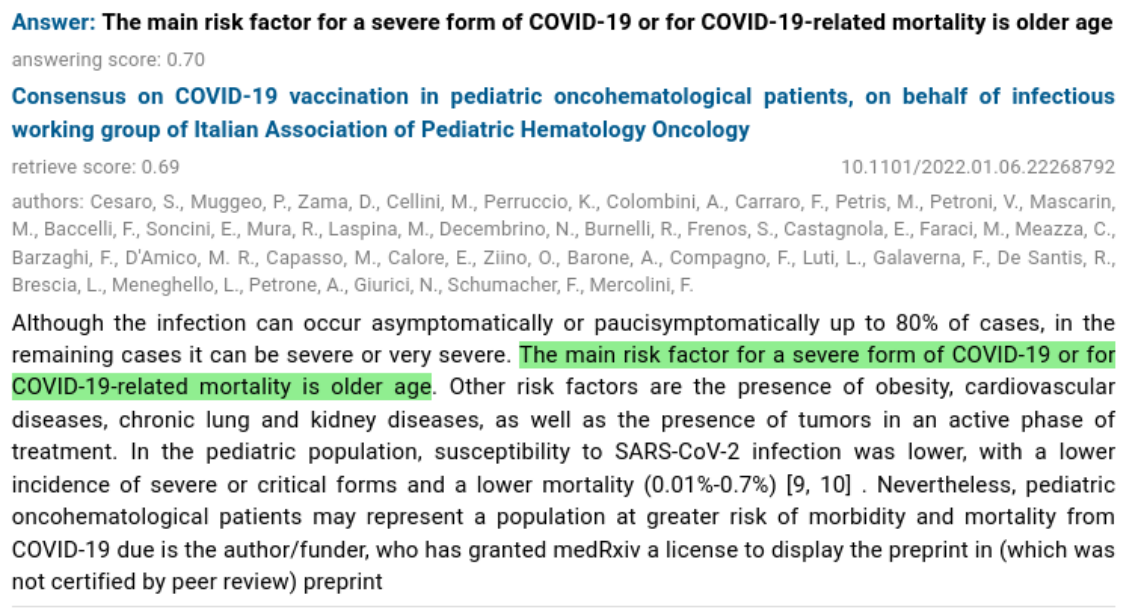

Figure 16: 3nd most relative answer to question "What do we know about COVID-19 risk factors?"

6.1.2 Application Infrastructure

The QA pipeline within our system was developed using the Haystack framework [40], which simplifies the development of complex NLP tasks such as QA, by providing powerful tools. QA components, can be built and evaluated separately and finally utilize them to a single pipeline with appropriate configuration environment variables. The QA system is backed by FastAPI, a powerful yet simple Python Web framework. FastAPI is widely used in AI services due to its efficiency and simplicity, enabling seamless integration with the rest of our application. The User Interface (UI) of our application is designed using React, a Javascript framework created by Facebook. React is a framework for building

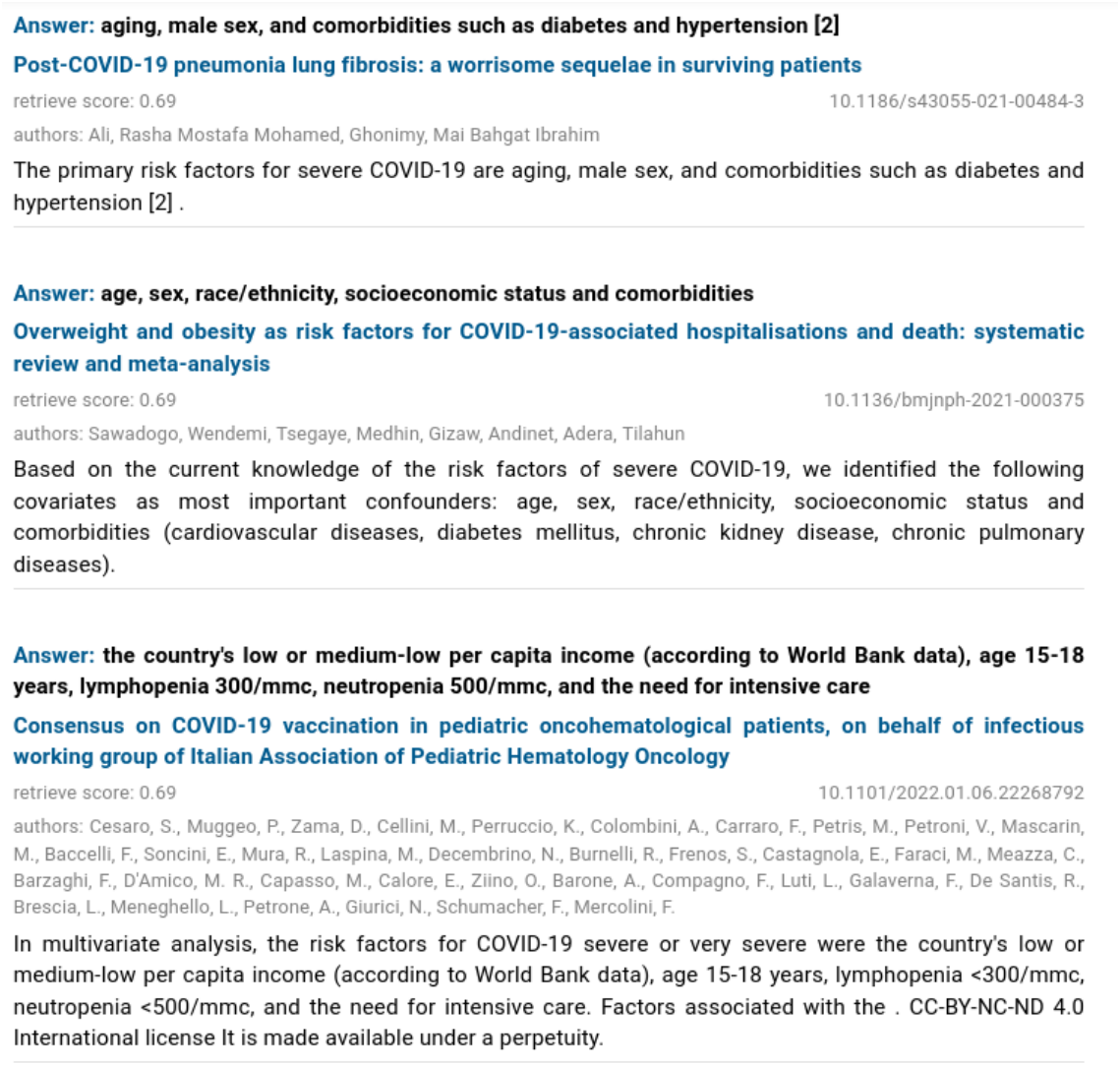

Figure 17: Generated Answers

dynamic and responsive interfaces, enhancing the overall user experience of our QA web application

7 Conclusions and future work

In this thesis, a robust QA system was developed and evaluated with a focus on COVID-19-related and biomedical topics. Despite constraints such as limited time and processing power, the development of a system that can aid medical professionals and researchers in addressing complex biomedical inquiries was demonstrated. The aim of this project is to facilitate knowledge acquisition and informed decision-making in the face of this global health crisis.

This study has focused on the development and evaluation of a biomedical QA system tailored specifically for addressing inquiries related to COVID-19, with a potential to be adapted on several topics. The system employs a hybrid retriever, combining a domain-adapted [25] dense retrieval model based on Siamese BERT Networks [16], with a BM25 lexical retriever [9], to enhance the retrieval of relevant information from biomedical literature.

The performance evaluation of the hybrid retriever on the TREC-COVID dataset [22], as depicted in Table 1, demonstrates its effectiveness in retrieving pertinent information. Particularly noteworthy is the competitive performance of the bi-encoder retrieval model, which was adapted to the biomedical domain with focus on COVID-19, using domain adaptation techniques with GPL [25]. Despite the strength of BM25 in biomedical data retrieval, the proposed dense model's performance is close, indicating its capability to capture domain-specific information effectively.

Furthermore, incorporating a cross-encoder model [13] as a ranker. Cross-encoder models have consistently demonstrated superior performance compared to other architectures in various IR tasks [24]. In our study, the cross-encoder model is employed as a ranker on documents fetched by the hybrid retriever, thereby enhancing its overall performance, as evaluated by $nDCG@10$ presented in Table 1. Utilizing the proven strength of cross-encoders, the disadvantage of those models on comparing a large number of documents is considered, as the input of the proposed cross-encoder is the top-n documents already fetched by our retriever, thereby reducing the computational need.

Regarding the extractive reader component, the performance of three transformer-based BERT models was investigated: BERT [13], RoBERTa [14], and BioBERT [15]. Specifically opting for the cased versions of these models, the aim was to retain case sensitivity for enhanced linguistic understanding. The base version of those models was selected for fine-tuning due to time and hardware constraints. Among the models considered, BioBERT, which had been pre-trained on PubMed articles, was selected as the primary focus. To enhance its adaptability to a broader range of biomedical inquiries, particularly COVID-19 topics, BioBERT was trained with a merged version of datasets consisting of COVID-QA and BioASQ 7b training datasets. As demonstrated in Table 8, the model performed well on each test dataset individually.

The development of the proposed biomedical QA system was not without its challenges. Numerous difficulties were encountered along the way, with time constraints and hardware limitations posing significant hurdles. These constraints imposed exploration of suitable methodologies to construct a system capable of truly assisting biomedical professionals.

Further improvements and customization opportunities exists on the proposed system. While the retrieval model was trained for domain adaptation using the full text of a small subset of the CORD-19 and BioASQ Task a corpus, the GPL method, which the system is based on, utilized abstract passages from the TREC-COVID corpus. Exploring the potential impact of training the bi-encoder on a larger subset of the dataset, shuffled with biomedical PubMed articles covering various topics, could be particularly interesting. This approach could enhance the model's adaptability and performance across diverse biomedical domains. The model has trained for a limited number of training steps, approximately 50,000 steps. However, there remains an opportunity for further investigation by extending the training duration with additional training steps and epochs. The exploration would provide insights on how the model evolves over time and weather it continues to improve with prolonged training.

One of the crucial steps of the GPL method [25], that the proposed retrieval was based on, involves question generation for the unlabeled target domain corpus, leveraging a T5 question generation model. It is worth noting that this model may lack exposure to COVID-19 related contexts and biomedical passages in general, potentially impacting

the quality of generated questions. To address this limitation and enhance the efficacy of the model, an approach would be training the T5 model biomedical generative text, with the help of BioASQ datasets. By exposing the model to biomedical contexts and terminology, it can develop a deeper understanding of relevant topics and generate more accurate and contextually relevant questions, simulating real biomedical queries performed by professionals.

Utilizing our robust retrieval component as a hard negative miner represents a promising enhancement to the GPL method. Given the retrieval system's strong performance on biomedical data, as demonstrated by its successful evaluation on reliable datasets, it could suggest related passages to questions during the mining step of the GPL method. This could be augment the effectiveness of negative sampling, outperforming conventional bi-encoders in identifying negative examples.

Expanding the proposed system to support different languages presents an opportunity for broadening its applicability impact. Transformer-based models like BERT and SBERT have been ported to many languages. One potential approach involves training the retriever component on a targeted biomedical corpus consisting of books, notes, prescriptions etc, in a different language. This expansion not only enhance accessibility for non-English speakers, but also facilitates cross-linguistic biomedical research and collaboration.

In this study only three BERT-based models were trained on size-limited biomedical training datasets. A further study on additional transformer-based models could be performed, especially on those pre-trained models on large biomedical corpora, across multiple diverse training datasets.

This study was concentrated on transformer-based models such as BERT and SBERT, and the effectiveness of larger models like GPT, LLaMA and Galactica was not investigated. However, with the advent of these powerful language models, a new era in NLP is emerging, especially within the biomedical domain. There is a significant potential to integrate those models into the proposed biomedical QA system, by combining them with the adapted retriever within a Retrieval-Augmented Generation (RAG) architecture [45].

Bibliography

[1] Koichi Yuki, Miho Fujiogi, and Sophia Koutsogiannaki. Covid-19 pathophysiology: A review. *Clinical Immunology*, 215:108427, 2020.

[2] World Health Organization. WHO COVID-19 Dashboard, 2024. Accessed: 25 February 2024.

[3] Andre Esteva, Anuprit Kale, Romain Paulus, Kazuma Hashimoto, Wenpeng Yin, Dragomir Radev, and Richard Socher. Covid-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization. *npj Digital Medicine*, 4:68, 2021.

[4] Junaid Shuja, Eisa Alanazi, Waleed Alasmary, and Abdulaziz Alashaikh. Covid-19 open source data sets: A comprehensive survey. *medRxiv*, 7 2020. Preprint. Not peer-reviewed.

[5] Qiao Jin, Zheng Yuan, Guangzhi Xiong, Qianlan Yu, Huaiyuan Ying, Chuanqi Tan, Mosha Chen, Songfang Huang, Xiaozhong Liu, and Sheng Yu. Biomedical question answering: A survey of approaches and challenges. *ACM Computing Surveys*, 55(2):35, 2022. Published on January 18, 2022.

[6] Zakaria Kaddari, Youssef Mellah, Jamal Berrich, Toumi Bouchentouf, and Mohammed G. Belkasmi. Biomedical question answering: A survey of methods and datasets. In *2020 Fourth International Conference On Intelligent Computing in Data Sciences (ICDS)*, pages 1–8, 2020.

[7] Zellig S. Harris. Distributional structure. *WORD*, 10(2-3):146–162, 1954. Published online: 04 Dec 2015.

[8] K. Jones. A statistical interpretation of term specificity in retrieval. *Journal of Documentation*, 60:493–502, 01 2004.

[9] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

[10] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26:3111–3119, 2013.

[11] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[14] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[15] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, September 2019.

[16] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

[17] Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with bert, 2020.

[18] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. Dense passage retrieval for open-domain question answering, 2020.

[19] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text, 2016.

[20] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad, 2018.

[21] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex Wade, Kuansan Wang, Nancy Xin Ru Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. Cord-19: The covid-19 open research dataset, 2020.

[22] Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. Trec-covid: Constructing a pandemic information retrieval test collection, 2020.

[23] Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. COVID-QA: A question answering dataset for COVID-19. In Karin Verspoor, Kevin Bretonnel Cohen, Mark Dredze, Emilio Ferrara, Jonathan May, Robert Munro, Cecile Paris, and Byron Wallace, editors, *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, July 2020. Association for Computational Linguistics.

[24] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models, 2021.

[25] Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. Gpl: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval, 2022.

[26] Kexin Wang, Nils Reimers, and Iryna Gurevych. Tsdae: Using transformer-based sequential denoising auto-encoder for unsupervised sentence embedding learning, 2021.

[27] Brian Schwartz Shaina Raza and Laura C. Rosella. Coquad: a covid-19 question answering dataset system, facilitating research, benchmarking, and practice. *BMC Bioinformatics*, 23, 2022.

[28] Jafar A Alzubi, Rachna Jain, Anubhav Singh, Pritee Parwekar, and Meenu Gupta. COBERT: COVID-19 question answering system using BERT. *Arab. J. Sci. Eng.*, 48(8):1–11, June 2021.

[29] Yan Yan, Bo-Wen Zhang, Xu-Feng Li, and Zhenhan Liu. List-wise learning to rank biomedical question-answer pairs with deep ranking recursive autoencoders. *PLOS ONE*, 15(11):1–19, 11 2020.

[30] Eduard Hovy, Laurie Gerber, Ulf Hermjakob, Michael Junk, and Chin-Yew Lin. Question answering in webclopedia. *Information Sciences Institute, University of Southern California*, 2002.

[31] Sofia J. Athenikos and Hyoil Han. Biomedical question answering: A survey. *Computer Methods and Programs in Biomedicine*, 99(1):1–24, 2010.

[32] Weaviate Vector Database. Weaviate is an open source, AI-native vector database that helps developers create intuitive and reliable AI-powered applications.

[33] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022.

[34] George Tsatsaronis, Michael Schroeder, Georgios Paliouras, Yannis Almirantis, Ion Androutsopoulos, Eric Gaussier, Patrick Gallinari, Thierry Artieres, Michael R Alvers, Matthias Zschunke, et al. Bioasq: A challenge on large-scale biomedical semantic indexing and question answering. In *2012 AAAI Fall Symposium Series*, 2012.

[35] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023.

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

[37] Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. Improving efficient neural ranking models with cross-architecture knowledge distillation, 2021.

[38] MS MARCO sentence-transformers. MS MARCO sentence-transformers.

[39] Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun hsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. Efficient natural language response suggestion for smart reply, 2017.

[40] haysttack by deepset. Open-source LLM framework to build production-ready applications.

[41] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 09 2019.

[42] Anastasios Nentidis, Georgios Katsimpras, Eirini Vandorou, Anastasia Krithara, Luis Gasco, Martin Krallinger, and Georgios Paliouras. *Overview of BioASQ 2021: The Ninth BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering*, page 239–263. Springer International Publishing, 2021.

[43] Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. Efficiently teaching an effective dense retriever with balanced topic aware sampling, 2021.

[44] Eric Sayers, PhD. A General Introduction to the E-utilities.

[45] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.