Interpreting Patient Case Descriptions with Biomedical Language Models

A thesis submitted in partial fulfilment

of the requirement for the degree of Doctor of Philosophy

Israa Ali Alghanmi

January 2023

Cardiff University
School of Computer Science & Informatics

Abstract

The advent of pre-trained language models (LMs) has enabled unprecedented advances in the Natural Language Processing (NLP) field. In this respect, various specialised LMs for the biomedical domain have been introduced, and similar to their general purpose counterparts, these models have achieved state-of-the-art results in many biomedical NLP tasks. Accordingly, it can be assumed that they can perform medical reasoning. However, given the challenging nature of the biomedical domain and the scarcity of labelled data, it is still not fully understood what type of knowledge these models encapsulate and how they can be enhanced further. This research seeks to address these questions, with a focus on the task of interpreting patient case descriptions, which provides the means to investigate the model's ability to perform medical reasoning. In general, this task is concerned with inferring a diagnosis or recommending a treatment from a text fragment describing a set of symptoms accompanied by other information. Therefore, we started by probing pre-trained language models. For this purpose, we constructed a benchmark that is derived from an existing dataset (MedNLI). Following that, to improve the performance of LMs, we used a distant supervision strategy to identify cases that are similar to a given one. We then showed that using such similar cases can lead to better results than other strategies for augmenting the input to the LM. As a final contribution, we studied the possibility of fine-tuning biomedical LMs on PubMed abstracts that correspond to case reports. In particular, we proposed a self-supervision task which mimics the downstream tasks of inferring diagnoses and recommending treatments. The findings in this thesis indicate that the performance of the considered biomedical LMs can be improved by using methods that go beyond relying on additional manually annotated datasets.

Acknowledgements

First and foremost, I thank Allah for granting me the means, patience, determination and countless blessings to complete this journey. Secondly, it is hard to overstate my gratitude and profound appreciation to my supervisor, Professor Steven Schockaert, for his exceptional guidance and support throughout the course of this work. The discussions during our weekly meetings have consistently broadened my perspective. I am deeply indebted to him for his advice, generosity with time, and thorough feedback. It was a privilege to work under his supervision. I would also like to wholeheartedly thank and express my gratitude to my co-supervisor, Dr Luis Espinosa-Anke, for his help, valuable input and dedicated time during my PhD.

My heartfelt appreciation and thanks to my beloved family for their continuous support in pursuing my goals. I would like to extend special thanks to my dear mother. There are no adequate words to express my gratitude for her love, prayers and encouragement that helped me overcome many difficult times during my studies.

Lastly, I sincerely thank the University of Jeddah, the Saudi Arabian Ministry of Education, and the Saudi Arabia Cultural Bureau in London for funding my study and for their support throughout these years.

Contents

List of Publications

The work introduced in this thesis is based on the following publications:

1. Israa Alghanmi, Luis Espinosa Anke, and Steven Schockaert. 2021. Probing Pre-Trained Language Models for Disease Knowledge. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3023–3033, Online. Association for Computational Linguistics. [5]

2. Israa Alghanmi, Luis Espinosa-Anke, and Steven Schockaert. 2022. Interpreting Patient Descriptions using Distantly Supervised Similar Case Retrieval. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22). Association for Computing Machinery, New York, NY, USA, 460–470. [6]

3. Israa Alghanmi, Luis Espinosa-Anke, and Steven Schockaert. 2022. Self-Supervised Intermediate Fine-Tuning of Biomedical Language Models for Interpreting Patient Case Descriptions. In Proceedings of the 29th International Conference on Computational Linguistics, pages 1432–1441, Gyeongju, Republic of Korea. International Committee on Computational Linguistics. [7]

List of Figures

List of Tables

List of Acronyms

NLP Natural Language Processing

LM Language Model

BERT Bidirectional Encoder Representations from Transformers

GloVe Global Vectors for Word Representation

CE Cross-Encoder

TSDAE Tranformer-based Denoising AutoEncoder

FT Fine-tuning

NLI Natural Language Inference

QA Question Answering

MLM Masked Language Modelling

NSP Next Sentence Prediction

Chapter 1

Introduction

1.1 Background and Motivation

Natural language processing (NLP) has witnessed, in recent years, a major successful shift in how text representations are learned. This fundamental step lies at the heart of NLP, enabling language to be processed by computers. Lately, a shift from traditional word representation methods, in which words are represented by low-dimensional vectors, to the paradigm of pre-training and then fine-tuning, which is led by pre-trained language models (LMs), boosted the performance of many NLP tasks. Pre-trained LMs are deep neural networks that attempt to initially learn general-purpose rather than task-specific word representations. Specifically, the representations are first learned from a massive amount of text (i.e. the pre-training phase), then tweaked for the target task (i.e. the fine-tuning phase). As a result, NLP systems could be built with less training data for a particular task and with much higher accuracy. While having pre-trained word vectors is possible with some earlier word representation methods, the underlying idea behind the recently established models is to produce contextual word vectors (i.e. each word is represented considering the surrounding context).

With such progress, pre-trained LMs such as BERT (Bidirectional Encoder Representations from Transformers) [41] are currently the de-facto architecture for solving most NLP tasks, and their prevalence in general language understanding tasks is today indisputable [224, 225]. This comes as no surprise since pre-trained LMs overcome several limitations associated with traditional word vectors. In addition to the pre-trained LM's ability to learn context-dependent representations in contrast to the traditional static vectors, there are several reasons why they are preferred over standard neural architectures. One important (and perhaps less obvious) reason is that LMs capture a substantial amount of world knowledge. For instance, several authors have found that LMs are able to answer questions without having access to external resources [165, 178], or that they exhibit commonsense knowledge [50, 39]. Therefore, in parallel with such advancement, new questions and challenges emerged, including the need to investigate what type of knowledge such models capture, what are the limitations, what are the possible ways to inject external knowledge, and how to adapt them to different languages and domains. Domain-specific NLP has it is own challenges due to, for example, the specific vocabulary and text genres (e.g. tweets or medical reports) while having less available (raw and annotated) data in contrast to the general domain.

One of the most widely studied specialised/domain-specific areas in NLP is the biomedical domain, which has seen substantial work produced in terms of specialised LMs. Several versions of BERT [41] were adapted to support biomedical NLP including ClinicalBERT [9], SciBERT [23], BioBERT [104] and PubMedBERT [61]. The potential that NLP brings to the biomedical domain is enormous. Increasingly, electronic health records (EHR) are being adopted to hold patients' records. EHR is a digital health record system that organizes and stores patients health information, facilitating convenient access to the medical history of patients, including diagnoses, lab results, treatments, discharge summaries, among others. It consist of two primary types of textual data: structured data, which refers to information organized within predefined fields, and unstructured data, which consists of free text entries. A prime example of unstructured data found within EHR is discharge summaries. These summaries are prepared by physicians to provide a brief and concise report of the patient's condition during their current hospitalisation. It typically encompasses information about the patient's medical history (e.g past surgical procedures), chief complaint representing the main symptoms and signs of the medical condition (e.g shortness of breath), the performed procedures (e.g. surgery) and tests (e.g. complete blood count), medications (e.g aspirin) and follow-up care recommendations (e.g lifestyle modifications). Aside from medical records, other sources of biomedical text, such as scientific articles, are still underused and their full potential is yet to be uncovered. Among these articles, there are case reports that provide detailed descriptions of clinical cases about individual patients, offering real-world examples that aid in clinical decision-making. Clinical decision-making refers to the reasoning process through which healthcare professionals employ their knowledge and skills to diagnose and treat patients.

The automation of analysing, processing and interpreting patients records will potentially contribute valuable insights. Manually filled structured data could overburden the possibility of finding and observing hidden patterns while also limiting the aspects in which such data are viewed, not to mention being tedious and time-consuming. Exploiting the rich information available in the unstructured text will eventually lead to better healthcare systems benefiting the patients in the first place as well as the healthcare providers. There are several use cases and applications in which such systems might be useful. Examples include automating the extraction of structured data, which saves time and effort. Furthermore, it enables the development of evidence-based decision support systems and enhances patient outcomes, more generally, such as finding potential diagnoses or treatments. Moreover, such systems could have positive impacts on aiding drug research, discovering side effects, and recognising unknown symptoms, among others.

Although it has great potential, the biomedical domain poses a number of particular challenges for NLP. For instance, there is limited availability of unstructured text in the form of clinical notes. The reason behind this could be related to many factors, including privacy concerns. The de-identification and prepossessing of such text before being considered ready to train publicly shared models is essential, which is laborious and expensive. Beyond that, the high cost of annotating large biomedical datasets serves as another obstacle towards effectively training neural models, otherwise these models might suffer from generalization issues at inference time. Along with data scarcity, biomedical NLP models still face difficulties with such domain-intensive textual data due to it is complexity. Moreover, the variation of medical records writing styles from different healthcare providers, the use of negation and the extensive use of abbreviations and acronyms, all present challenges for this domain.

One of the useful use cases for biomedical NLP is to make inferences about patient case descriptions, which is the focus of this thesis. A patient case description typically contains information to present a specific clinical case, such as the age, gender, medical history, and current symptoms of some patient, along with physical examinations and lab results. The task of interpreting patient case descriptions can be defined as follows: Given a patient case description of the symptoms displayed by a patient, possibly in combination with other relevant factors such as age, gender or medical history, we may want to infer a diagnosis or identify recommended medications. An example of a patient case description from MedQA [82], which is one of the evaluation datasets, is shown below:

> *"A 16-year-old female high school student is brought to the physician by her parents for her repeated behavioral problems at home and school during the past 10 months. Her teachers describe her behavior as uncooperative and disruptive as she persistently refuses to answer questions, insults her teachers, and annoys her classmates on a daily basis. At home, her parents try to address her frequent violations of curfew, but attempts at discussing the issue often result in their daughter losing her temper and screaming at her parents. Her grades have deteriorated over the past year. She has no history of psychiatric illness. On questioning, the patient refuses to answer and frequently disrupts the physician's conversation with the parents."*

As standard language models (LMs) are able to make various factual and commonsense inferences [165, 39, 258], one might expect these biomedical LMs to be similarly capable of tasks such as inferring diagnoses from symptoms. However, it is not yet clear whether LMs indeed have or actually lack sufficient knowledge to solve such tasks, which makes the evaluation of their capabilities an important and natural direction. As biomedical LMs have proven successful in capturing the meaning of specialised terminology [9, 23, 104, 61], the main question is whether they also have medical reasoning capabilities, e.g. for predicting a likely diagnosis from a given patient case description. This is highly challenging, even for biomedical language models, because many pieces of information may need to be combined to find the right answer, and often some degree of clinical judgment is needed.

To address these challenges, as a first step, we need to understand the weaknesses and limitations of pre-trained biomedical LMs and investigate what aspects of knowledge these models capture or fail to capture. Thus, in this thesis, we first aim to analyse and evaluate the disease knowledge captured by biomedical LMs in a fine-grained way. In general, various methods have been proposed lately to analyse pre-trained LMs' capabilities and knowledge, reflecting the importance of having better expectations about their applicability and what areas need to be improved. However, given the challenging nature of the biomedical domain and, more specifically, the task of interpreting a patient description, we would still expect them to have a rather poor performance. This assumption is due to the fact that such a task requires complex reasoning to draw a conclusion based on different pieces of text. Consequently, this highlights the necessity to develop methods for enriching and enhancing their performance.

To meet the need for improving the ability of biomedical LMs to interpret patient case descriptions and considering the challenges regarding the limited availability of annotated datasets, we propose two strategies that go beyond fully supervised learning. In particular, in our first method, we employ a nearest neighbour strategy in which we seek to find similar patient cases to a given patient case description. To identify these similar patient cases, we rely on a language model that is fine-tuned in a distantly supervised way. In the end, we judge the likelihood of a specific diagnosis or treatment based on the similarity score between them. In the second approach, we construct datasets which are annotated in a self-supervised way and then use as intermediate tasks for fine-tuning the LMs. Intermediate fine-tuning is a technique where models are first fine-tuned on some task before the final fine-tuning on the evaluation task. Particularly, here we intend to exploit the freely available case reports mentioned in the abstracts of scientific articles for the intermediate fine-tuning.

1.2 Hypothesis and Research Questions

The main hypothesis in this thesis is as follows:

Existing biomedical LMs still struggle when it comes to interpreting patient case descriptions, which can partly be explained by the limited amounts of relevant annotated data. We hypothesize that the development of strategies that obviate the need for manual labelling can at least partially alleviate this issue, allowing biomedical LMs to interpret patient case descriptions with higher accuracy.

In order to verify this hypothesis, we aim to answer the following research questions:

Research Question 1: What kinds of medical knowledge do pre-trained LMs capture? More specifically, are these models capable of performing medical reasoning such as linking symptoms to diseases, or treatments to diseases?

Research Question 2: Is it possible to use nearest neighbour strategies for enhancing the LM's interpretation of patient case descriptions (i.e. relying on similar patient cases to drive the predictions)? Can we construct distantly supervised datasets to compensate for the lack of annotated datasets to train the model on identifying similar patient cases?

Research Question 3: How, and to what extent, can we obtain self-supervised datasets that serve as intermediate tasks for fine-tuning the LMs?

1.3 Contributions

Our primary aim is to enhance the knowledge of pre-trained LMs and boost their performance in interpreting patient case descriptions through the use of approaches that go beyond fine-tuning with only manually labelled datasets. To achieve this aim, the following contributions are made:

1. We introduce a new probing method to analyse to what extent different language models capture knowledge about diseases in a fine-grained way. In particular, we split positive examples of available datasets (MedNLI [184] and MEDIQA-NLI [2]) into two main categories to test the LM's medical and terminological knowledge. We further divide the medical knowledge category into four sub-categories that represent the type of reasoning that is needed. Specifically, the sub-categories include linking symptoms to diseases, treatments to diseases, procedures to diseases, and tests to diseases. We then evaluate, in isolation, the LM's knowledge about each disease in each one of these categories. In other words, we propose training-test splits per disease. This is to prevent the LMs from learning about the target disease from the training data. After that, we generate negative examples by corrupting the positive examples in an adversarial way. The results suggested that LMs performance is better with examples requiring terminological knowledge. We also show that the performance of the individual LMs differs across the different categories. This work was published in [5].

2. We develop a simple yet effective nearest neighbour strategy to enhance the performance of biomedical LMs in interpreting patient case descriptions. Since we want the model to infer whether the patient case description entails a hypothesis of interest (e.g. diagnosis), we start by retrieving a set of text passages mentioning this hypothesis. Next, we applied our model to get the most similar retrieved passage to the given patient description. Then based on the similarity score between the top retrieved passage and the given patient case description,

we will determine whether the patient case description entails the hypothesis. To train the model for similar patient cases, we need to overcome the lack of annotated datasets. Therefore, we proposed constructing the dataset in a distantly supervised way. This work was published in [6].

3. We propose a set of self-supervised intermediate fine-tuning tasks to boost the performance of biomedical LMs. We accomplish this primarily by exploiting case reports found in the literature while aiming to infuse knowledge by targeting specific types of medical concepts (i.e. diseases or treatments) using different strategies. The results show how the different medical concepts and strategies eventually influence the model performance, and the arbitrary use of the medical concept could sometimes even drive the performance down. This work was published in [7].

1.4 Thesis Structure

The remainder of the thesis is organised as follows:

- Chapter 2 - Background and Related Work - provides a general overview of text representation models and the journey that led to the introduction of pre-trained LMs. Furthermore, this chapter lists the details of some of the available biomedical LMs and discusses the related work which investigates the knowledge encoded in such models and methods to enrich them. Along with that, some of the different downstream tasks and applications related to this thesis are defined. In addition, this chapter gives a brief explanation of the different supervision strategies such as distant and self-supervision.
- Chapter 3 - Datasets and Resources - presents the details of the datasets that were considered in this work for evaluating our proposed approaches. Additionally,

this chapter covers the various resources and tools that were utilised to carry out the experiments.

- Chapter 4 - Probing Pre-Trained Language Models for Disease Knowledge - presents the proposed method for probing and investigating the capabilities of several LMs in the biomedical domain. In particular, this chapter proposes to fine-tune the LMs based on various categories, aiming to analyse to what extent the LMs capture knowledge about the considered diseases and how they differ across categories.

- Chapter 5 - Interpreting Patient Case Descriptions using Distantly Supervised Similar Case Retrieval - describes a model that relies on identifying similar cases using external resources to drive the model predictions, in which we ultimately aim to enhance the LM performance over standard fine-tuning (i.e without the use of external knowledge). A description of the considered pipeline to construct distantly supervised datasets to train this model is also presented in this chapter.

- Chapter 6 - Self-Supervised Intermediate Fine-Tuning of Biomedical Language Models for Interpreting Patient Case Descriptions - provides a set of strategies to construct self-supervised datasets to be used as intermediate fine-tuning tasks, while comparing their performance. This chapter also introduces various analysis methods driven by the concept of the proposed strategies and shows the results of ablation experiments.

- Chapter 7 - Conclusions and Future Work- concludes the thesis by providing summaries of the results and how those met our initial aims. In addition, this chapter presents the suggested directions for future work.

1.5 Summary

In this chapter, we started with a brief overview of pre-trained LMs in general, the opportunities which NLP brings to the biomedical domain, the challenges faced by biomedical language models, and our considered problem with motivations behind the proposed methods. In addition, we discussed the hypothesis and the research questions. Finally, we listed the contributions and the structure of this thesis. The following chapter will give thorough background information about LMs with a review of the relevant work in the literature.

Chapter 2

Background and Related Work

2.1 Introduction

This chapter provides background information for the work done in this thesis. As we intend to improve the capabilities of the existing biomedical pre-trained LMs, this chapter reviews the relevant research on this topic. We start with general background about traditional neural word representation models in Section 2.2. Next, we focus on the state-of-art pre-trained language models in Section 2.3. After that, in Section 2.4, we present the details of commonly used biomedical LMs with the followed approaches to train them and work comparing their performance. In Section 2.5, we review research on assessing the knowledge encoded within the LMs, for the general and biomedical domains. Subsequently, the methods and strategies to enhance LMs knowledge and performance are discussed in Section 2.6. In Section 2.7 we review works on analysing patient case descriptions. In Section 2.8, we list and briefly define a range of downstream NLP applications used in this thesis. Then in Section 2.9, we discuss the common supervision strategies generally used in machine learning with a narrowed focus on the approaches used to go beyond fully supervised learning. Finally, we summarise the chapter in section 2.10.

2.2 Word Representations

To solve NLP tasks, as a preliminary step, text data is converted into meaningful numerical representations (i.e. vectors of numbers), which are then used by machine learning models. Vector representations of words are also known as word embeddings. The quality of such representations and how well they capture words' semantic and syntactic properties play an essential role in the performance of NLP systems. Accordingly, works on how to construct these vector representations evolved over the years. Most early text representation methods primarily rely on the position of the word within the given text ignoring many other linguistic properties. For instance, one-hot encodings construct a Boolean vector for each word in a sentence (i.e. the value of 1 is assigned for the presence of a given word and 0 for it is absence). Obviously, such a simple and straightforward representation technique is accompanied by various drawbacks. A key limitation is a lack of semantic, syntactic, and relational information being captured in terms of word meanings. More sophisticated word representation methods have been proposed to overcome the drawbacks of the earlier methods. Generally, employing neural approaches brought substantial advancements by automatically learning dense vectors, which facilitates identifying useful features while avoiding the heavy feature engineering for solving each NLP task. In this section, we briefly describe existing text representations models predated the state-of-the-art and go through the journey that led to the introduction of transformer-based pre-trained LMs.

Traditional word embeddings models, such as Word2vec [134] and GloVe [160], aim to learn dense low-dimensional vector representations of words relatively in a way that similar words are clustered closer to each other in a vector space [89]. Creating the vectors in this manner intends to capture semantic and syntactic relationships between different words. Although such initialisation strategies led to better performing deep learning models, however as a shortfall, these models are context-independent. In other words, each word is assigned a fixed vector, irrespective of whether the word has different senses [227, 89]. For instance, the word "season" would always have the exact numerical representation regardless of whether the intended meaning is "a time of the year" or "adding spices to the food". Specifically, each word is mapped to a single vectorised representation that is averaged across different meanings (depending on which appeared in the training text corpus). As a consequence, the obtained word vector might not precisely represent the target word as intended.

To address this limitation, contextualized word embedding models were introduced with the purpose of capturing the sense of a word as expressed in a specific context rather than a generalized one. Particularly, unlike representing each word with a single static vector, the core idea behind them is to consider the context in which the word is used and the ability to establish more than one representation per word [195]. A straightforward strategy is to consider the weights of previous words occurring in a text sequence preceding the target word [210]. These are then provided as informative cues to construct the word representation. Some of the earliest context-dependent models are TagML [162], CoVe [128], and Context2Vec [129]. However, apart from the fact that each has its own weaknesses, all simply consider the output of the model's final layer as the encoded-word representation, which bounds their success [147].

ELMo (Embeddings from Language Model)[163] was proposed to address some of the limitations of the aforementioned models. First, it computes the word representation based on the average of all layers, which makes the representations deeper. Moreover, it scans the text sequence independently from left-to-right and right-to-left (i.e. in a bi-directional way) using a language modelling objective and a two-layer Bi-LSTM. Thus, the use of language modelling objectives has been shown to be more successful than prior techniques in advancing some NLP challenging tasks. Despite such success, the fairly small training dataset sizes act as an obstacle against revealing the full capabilities. Deep neural models need a large amount of training data in order to achieve reasonable performance, and while text is abundant, annotations are expensive. On that account, the necessity increased for using freely available text in an unsupervised way. In this regard, ULMFiT (Universal Language Model Fine-tuning) [70] proposes to first pre-train a language model on a large general corpus and then allows the fine-tuning of the encoded representations with respect to a target task. Yet, such models relied on LSTM, which suffers from the lack of capturing longer range dependencies [232]. On the grounds of this, the combined use of self-attention mechanism [217], and language models pre-training was introduced, particularly with BERT (Bidirectional Encoder Representations from Transformers) [41] achieving state-of-the-art results on a broad set of NLP tasks. The following section will explain it in greater detail.

2.3 Transformer-based Pre-trained Language Models

Late 2018 witnessed a breakthrough technique introduced by [41], shifting the NLP field to a new paradigm. In particular, since the introduction of BERT, the dominant approach for tackling most NLP problems has been confined to two main steps: the unsupervised pre-training of such models on a large text corpus and then easily adapting them to downstream tasks through fine-tuning on annotated datasets. Pretraining has been shown to be an effective strategy for alleviating some of the main issues associated with earlier neural models, enabling the model to learn universal features before learning task-specific features. Therefore, it offers better initialisation and generalisation for different downstream tasks while reducing the need for annotating large datasets or training the model from scratch for each task.

BERT builds upon various complex concepts that exist in the literature and inherits the idea of training a bi-directional language model from ELMo. Instead of using LSTM as in other earlier models, BERT uses the encoder part from the Transformer architecture. Transformers [217] were initially introduced with encoder-decoder configuration for solving sequence-to-sequence tasks such as machine translation. The main idea is to use attention mechanism and feedforward layers instead of LSTMs or other types of Recurrent Neural Networks. Rather than sequentially scanning the input sequence (i.e. left-to-right or right-to-left), attention looks at the entire sequence at once and then determines which parts are important to a given target word [125, 52]. This, in particular, helps in learning deep representations by encoding contextual embeddings that are derived from the context of both directions concurrently instead of either left or right [149, 229]. As explained earlier, each word could have multiple representations based on the context in which that word is mentioned.

The standard language model training technique is to predict the next word, given the first part of a sentence. However, since BERT uses self-attention, it randomly masks words within the sequence using [MASK] token to be predicted during the model training. This masked language modelling (MLM) objective makes training a language model compatible with the attention mechanism, which forces the model to look into the entire context rather than the next word. More specifically, 15% of the input sequence is masked, and the model's objective is to predict those masked tokens by relying on the unmasked ones. Beside MLM, BERT trains the model with the next sentence prediction (NSP) objective as well. By using NSP, the model learns to classify whether a sentence is the following sentence of a given sentence. This objective grants the model the capability of learning not only the relations between words but also between sentences. This is particularly beneficial for sentence-pair tasks such as natural language inference and question answering [64].

Advancements in computational power enabled the pre-training of large transformer-based language models using a large amount of data (English Wikipedia and books in the case of BERT), which also led to an increasing number of model parameters. BERT comes in two main variants, BERT base and BERT large. The former has 110 million trainable parameters and is organised in 12 layers, and the latter has 340 million trainable parameters with 24 layers. The maximum input length supported by the model is 512 tokens, where a token could represent a word, subword or even a single character. To convert given input sequences into vector representations using BERT, we first need to tokenize the inputs. Each resulted token is then checked against BERT vocabulary. In the case when a word does not exist in the vocabulary, the BERT tokenizer splits the word into pieces (i.e. subwords or even characters). In particular, the number of pieces depends on the largest subword tokens that can be found in the BERT vocabulary, which in an extreme case could mean that the word is split into individual characters. All sub-word tokens following the first token are preceded by "##". Since BERT produces contextualised embeddings, unlike models such as word2vec, it usually takes a text sequence as an input (as word order matters) rather than individual words.

To effectively perform various NLP tasks, BERT uses special tokens within each input sequence. First, it uses [CLS] as the first token for any input sequence and [SEP] as the last token. [CLS] is often used as a special token for classification, representing the entire input (i.e. sentence). The final hidden state of this token is used for tasks that require sentence-level reasoning. A common alternative to using the [CLS] token is averaging the final layer representations of all tokens from the given input [175, 212, 47]. For sentence pair tasks, to distinguish the first from the second sentence, the [SEP] token is also used between the two sequences.

2.4 Pre-trained LMs for Biomedical Text

General-purpose pre-trained LMs (i.e. trained on general domain corpora such as Wikipedia, news, and books), have achieved remarkable results. Following this success, variants of such models have been specifically adapted to various domains. Most relevant for this thesis, several pre-trained LMs have been released for the biomedical domain. The unsupervised pre-training on biomedical corpora overcomes many previous limitations, including the need for a large annotated dataset to train the model. Additionally, domain-specific pertaining mitigate the shift issue in word distributions, which in turn contributes to substantial gains in performance.

Predominately, for domain-specific variants, two pre-training paradigms are adopted. The first is to pre-train the LM from scratch on some biomedical text. The other is to initialise the LM with an already pre-trained LM (either on a general or specific domain corpus), then further pre-train it using an unlabeled biomedical text corpus. In the former case, the model vocabulary includes specific domain terminology. Therefore, each word contained in the vocabulary is considered as a single token by the model tokenizer, as opposed to a model that is pre-trained first on the general domain, in which some domain words might split into pieces (sub-tokens) [88]. This would also result in a shorter input length sequence, which is more efficient. The latter case is particularly beneficial for domains in which specialised available text is limited. Although the biomedical domain has an abundance of available texts (e.g. in the form of scientific papers in PubMed), both pre-training methods have been proposed and advanced many biomedical tasks. Examples of biomedical corpus include the abstracts of scientific articles, their full texts, electronic health records, and Wikipedia medical articles.

Many variants have become readily available for the biomedical domain to cope with the fast progress in the field, exploring different pre-training options. In other words, these models differ from each other mostly in the pre-training corpora rather than architectural features. Below is the commonly used list of the released variants:

BioBERT Lee et al. [103] proposed a model based on BERT$_{\text{base}}$-cased [41], which they further trained on biomedical corpora. More specifically, BioBERT was further pre-trained on the full text of PMC articles along with PubMed abstracts.

ClinicalBERT Alsentzer et al. [9] introduced four variants based on BERT model. The models are initialised from either BERT or BioBERT and then further pre-trained on either the full version of MIMIC-III notes or only the discharge summaries.

SciBERT Beltagy et al. [23] introduced a BERT model variant that was trained from scratch on approximately 1.14M scientific papers from semantic scholar. More specifically, 82% of the training corpus were biomedical articles, while the rest were mainly computer science articles. The full text of the papers was used for training.

PubMedBERT Gu et al. [61] released newer BERT model variants that were trained from scratch on either only PubMed abstracts or on both the abstracts and the full text of PubMed articles. PubMedBERT is the first model that was purely pre-trained on biomedical corpus rather than mixing with other domains such as SciBERT or initialising with a general domain model such as BioBERT and ClinicalBERT. Thereby, this model contains in its vocabulary more medical terms than the previous models.

Several authors have analyzed the performance of these models and the impact of considering different types of biomedical corpora. For instance, Peng et al. [159] proposed an evaluation framework for biomedical language understanding (BLUE) tasks. They obtained the best results with a BERT model variant that was pre-trained on PubMed abstracts and MIMIC-III clinical notes. Later, Gu et al. [61] introduced BLURB (Biomedical Language Understanding & Reasoning Benchmark), a collection of existing biomedical datasets representing a set of biomedical tasks such as named entity recognition and relation extraction. Another large-scale evaluation of pre-trained biomedical LMs has been carried out by Lewis et al. [106]. Such evaluation works are needed to be carried out on a regular basis to find the best-performing model for a given task, especially in this domain where information is updated frequently. To the best of our knowledge, no research has been found that evaluated the performance of the different LMs for interpreting patient case descriptions. Beyond that, there have been a number of works in the form of surveys summarise the existing pre-trained LMs in the biomedical domain [88, 226]

2.5 Knowledge Encoded in LMs

There is a rapidly growing body of work that is focused on analysing what knowledge is captured by pre-trained LMs. For example, the syntactic knowledge in LMs has been extensively evaluated [74, 69, 115, 211, 236, 57], as well as semantic knowledge [223, 49, 222, 28, 100, 257]. Other types of knowledge, such as factual and commonsense knowledge, have been studied as well [164, 179, 111, 258, 153]. A recurring challenge in such analyses is to separate the knowledge that is already captured by a pre-trained model from the knowledge that it may acquire during a task-specific fine-tuning step. A common solution to address this is to focus on zero-shot performance, i.e. to focus on tasks that require no fine-tuning, such as filling in a blank [39, 208].

As an alternative strategy, Talmor et al. [208] proposes to analyse the performance of models that were fine-tuned on a small training set. Other work has focused on extracting structured knowledge from pre-trained LMs. Early approaches involved manually designing suitable prompts for extracting particular types of relations [165]. However, several authors have proposed strategies that automatically construct such prompts [30, 192, 81]. Finally, Bosselut et al. [29] proposed to fine-tune LMs on knowledge graph triples, with the aim of then using the model to generate new triples. To some extent, these LMs have indeed been shown to implicitly encode different types of knowledge in their parameters. This encouraged several authors to work towards distilling the observed knowledge in an attempt to augment the promising features of the LMs with the accessibility that is found in knowledge bases (KBs), mostly with the above-mentioned prompts-based methods. Beyond encoded knowledge, other works focused on understanding biases resulting from pretraining such models on a vast amount of data [17].

To evaluate the biomedical knowledge that is captured in pre-trained LMs, as opposed to acquired during training, Jin et al. [83] freeze the transformer layers during training. They find that when biomedical LMs are thus used as fixed feature extractors, BioELMo outperforms BioBERT. As several studies proposed the use of pre-trained LMs' factual knowledge as an alternative for knowledge bases, Sung et al. [203] evaluate the applicability of such an approach for biomedical LMs and investigate whether such models have sufficient knowledge. Their results showed that the considered biomedical LMs are still not ready to serve as KB using the proposed methods, which needs more research attention. In a similar direction, Meng et al. [132] proposed a new probing technique where they aim to overcome one of these domain challenges. In particular, many entities are encoded in a multi-token fashion (i.e. more than a single token). For example, the "Renal artery stenosis" condition would be split into at least three tokens, and in prompting-based methods, each token would be replaced by [MASK], which means such methods might not be effective. Therefore, they rely on a nearest neighbour strategy to find the most similar entity to the target entity being queried. Another line of work suggested the use of adversarial tests [14] to evaluate the LM's robustness against spelling errors and synonyms. Particularly, for named entity recognition tasks, they randomly modify some words by adding noise (e.g. changing one character within the word) or swapping with synonyms. The results showed that the performance dropped significantly. To date, analysing and probing medical reasoning capabilities within biomedical LMs are still under-studied.

2.6 Enhancing LMs

Various strategies have been proposed for improving the amount of knowledge that is captured by transformer-based pre-trained LMs. One common approach is to rely on some kind of knowledge infusion while training the model [255] or during the fine-tuning phase [116, 53]. The latter is more efficient and less expensive as there is no need to re-train the model from scratch with the additional pre-training data. Investigating such ways will ultimately facilitate updating or exploiting other forms of knowledge, for example, using the available knowledge graphs. This is especially important for knowledge-intensive tasks, where relying on only task-specific training data might lead to suboptimal results. Also, this is helpful for domains with inadequate training data or small datasets. In the following subsections, we will group the common strategies into categories with an emphasis on works for the biomedical domain, noting that such methods are not mutually exclusive.

2.6.1 Continual Pre-training

One of the most straightforward approaches is to further train the LM on more data using the same training objectives (i.e. MLM and NSP). Starting with an existing LM and then further pre-train it using a new corpus which could be particularly beneficial for a target domain, downstream task, or for the sake of updating the encoded knowledge. A notable example of such sequential pre-training in the biomedical domain is ClinicalBERT. ClinicalBERT was initialised from a previously pre-trained LM, i.e. BioBERT, and further pre-trained on clinical notes, not to mention that even BioBERT itself was initialised from the standard BERT and additionally pre-trained on scientific articles. However, it has been argued that training from scratch, directly on the domain-specific data, yields better results than the additional pre-training in which the training start from general domain data [61].

2.6.2 Pre-training Objectives

Several works have been exploring the use of different pre-training objectives aside from the original ones proposed with the standard BERT (i.e. random MLM and NSP) either as an alternative or for the additional pre-training. In general, such an approach has been shown as an effective way to infuse knowledge with respect to some tasks or domains. Tailoring the objectives to the specific task while pre-training the model might offer more significant advantages to the performance, and this could be the case even when some LMs were already pre-trained on the same data but with the standard objectives. For the biomedical domain, He et al. [68] proposed a pre-training objective that aims to infuse disease knowledge by exploiting the structure of Wikipedia pages about diseases. Yuan et al. [247] pre-trained a language model with entity extraction and linking objectives based on UMLS [27]. Similarly, Michalopoulos et al. [133] incorporated semantic type embeddings into the pre-training phase while also taking into consideration the synonyms of the predicated token for the masked language modelling objective. On the other hand, Wijesiriwardene et al. [239] replaced the standard next sentence prediction objective (NSP) by predicting the synonyms of the medical entities. In [161], the authors proposed to fine-tune a biomedical language model by using a masked language modelling objective which is modified such that only biomedical concepts are masked.

2.6.3 Intermediate Task Fine-tuning

The standard paradigm in NLP at the moment is to fine-tune a pre-trained LM, such as BERT [40], on task-specific training data. However, it has been observed that adding an intermediate step, where the LM is first fine-tuned on a different task, for which training data is more abundant, can be highly beneficial [166, 167, 152, 156, 169]. Several works have investigated the role of intermediate tasks, in particular with the aim of analysing when and why results improve [170, 36]. While already available datasets are usually used for this approach, Vu et al. [221] showed that even synthesizing the training data can enhance the results. For the biomedical domain, one strategy has been to rely on transfer learning from general-domain tasks. For instance, Soni and Roberts [197] use general-domain question answering for intermediate training to improve a clinical question answering system. Another strategy has been to rely on different but related tasks, such as pre-training on natural language inference to develop a question answering system [75].

2.6.4 Augmenting Input with Unstructured Text

Generally, some methods augment the model input with knowledge expressed in textual form. For instance, Lu et al. [119] used definitions of UMLS concepts for this purpose. While this improved the results, their evaluation was based on static general-purpose word vectors and an LSTM-based model. The usefulness of their strategy in combination with biomedical LMs has not been extensively explored. More generally, however, there is some evidence that the effectiveness of augmenting the input with textual knowledge is limited in the biomedical domain. For instance, Sushil et al. [204] evaluated the effect of such augmentation strategies and failed to obtain any statistically significant improvements for MedNLI [183], a well-known benchmark for Natural Language Inference (NLI) in the biomedical domain. A more detailed description of this dataset will be given in Chapter 3, Section 3.2.

2.6.5 Structured Knowledge

Beyond unstructured text, other valuable sources of knowledge, such as knowledge graphs and tabular data, potentially offer stronger capabilities to pre-trained LMs. Therefore, several studies have been examining how to utilise such sources in order to complement the knowledge encoded within the LMs. In particular, works on how to incorporate knowledge graphs attracted the attention of many authors. Knowledge graphs basically structure the data in the form of triples which mostly represent the world or commonsense knowledge. Each triple consists of three components: subject, relation and object. The subject and object are represented as nodes and the relation as an edge to connect them. Examples of well-known knowledge graphs are ConceptNet [198] and WikiData [220]. Injecting such distinct forms of knowledge would require some kind of pre-processing to transform them in a way that is compatible with pre-trained LMs, especially if that is intended to be during the pre-training phase. One common approach is to convert the triples into a textual format and then apply the masked language modelling objective [3].

For the biomedical domain, several authors have proposed techniques for infusing the knowledge from biomedical knowledge graphs (e.g. UMLS knowledge graph) into LMs [66, 76]. Zhang et al. [250] used structured knowledge about entities and their relations for pre-training. As the large size of knowledge graphs is one of the main challenges associated with using them, Meng et al. [131] introduced a method for infusing knowledge from a large biomedical knowledge graph through partitioning such a graph into smaller sub-graphs.

2.6.6 Static Vector Representations

Instead of improving the language model itself, some authors have also explored the possibility of combining contextualised embeddings with static vector representations such as word2vec [4]. The efficiency of training and using static vectors motivates work towards incorporating them. For the biomedical domain, static representations of biomedical concepts are readily available. Sharma et al. [191] used UMLS knowledge graph embeddings to improve the BioELMo model (i.e. a biomedical version of ELMo model) while Chang et al. [35] combined a BERT-based model with SNOMED CT knowledge graph embeddings.

2.7 Analysing Patient Case Descriptions

A considerable amount of literature has considered the analysis of patient case descriptions. These works are primarily focused on outcome predictions such as the length of stay at the hospital [215, 77], re-admission [72, 58, 120, 77], automated ICD coding for procedures or diagnosis [107, 142, 123], and mortality prediction [235, 215, 43]. For example, Naik et al. [145] studied enhancing the predictions for in-hospital mortality, length of stay and the need for ventilation using PubMed abstracts. Patients' clinical records have also been utilised for clinical trial recruitment [54, 252]. Apart from that, studies about predicting the risk of a patient having a particular disease in the future or by the next hospital visit have been also conducted [121, 118, 136, 110, 249, 158]. Different lines of research have been explored for diagnosis prediction. Some studies have only targeted the diagnosis of specific diseases such as pneumonia [140, 141] and liver disease [113]. Boag et al. [26] evaluated different static text representations techniques, such as Word2Vec, for predicting the diagnosis of a particular set of diseases, including sepsis and coronary artery disease. Others have examined predicting a patient's diagnosis based on that patient's previous visits and the associated longitudinal records [158, 109]. In some works on predicting ICD diagnostic code, the disease might be already mentioned in the text, and the task is to map it to it is corresponding code. However, and most closely related to our work, van Aken et al. [215] proposed to use admission notes from MIMIC-III, in which the discharge diagnosis is not mentioned, to evaluate and enhance the performance of different LMs for ICD code prediction. In particular, they further pre-train BioBERT on the next sentence prediction task considering patients' notes and scientific articles as the first sentence and the outcome as the second sentence. Wang et al. [234] argued that models need to be trained on predicting the diagnosis without requiring patients' historical visit data, which is particularly useful in practice as the patient sequential visits records might not always be available. Therefore, they used the clinical notes of a single visit to predict the diagnosis. Many methods for predicting the diagnosis relied on graph neural networks [109, 233]. Another common task is medication recommendation, in which several works relied on the patients' longitudinal data. For example, An et al. [12] and Su et al. [200] studied medicines prediction based on historical patient data. On the other hand, Shang et al. [188] used the single visit data to pre-train BERT and then combined the representation of graph neural networks with BERT for predicting the medication for patients with multiple visits.

2.8 Downstream Tasks and Applications

In the previous section, we already discussed the task of analysing patient case descriptions, which plays a central role in this thesis. However, there are a number of additional biomedical NLP tasks that will also be used, either directly or indirectly. In this section, we present a quick overview of these tasks.

Natural Language Inference (NLI) Natural language inference (NLI) is the task of determining the relation between pairs of text. In particular, given two sequences of text, the task is to decide whether the first text sequence (i.e. the premise) entails the second (i.e. the hypothesis), contradicts, or the relationship between them is neutral [98, 19]. In other words, this task is concerned with predicting whether the hypothesis is true, false or undetermined given the premise. For example:

Premise: The girl is holding a juice.

Hypothesis:	Relation:
The girl is holding a drink.	Entailment
The girl is empty-handed.	Contradiction
The girl is holding a fruit juice.	Neutral

NLI is also known as recognizing textual entailment (RTE)[155, 209]. A simplified version of NLI consists in only determining whether the hypothesis can be inferred from the premise or not (i.e. a binary version) [97, 138]. In general, NLI is a central task in natural language understanding, which is closely related to several applications, including question answering and information retrieval.

Question Answering (QA) QA is a common NLP task, which is concerned with providing answers to questions written in natural language. A number of different variants exist, which depend on how the question and answer are formatted, including multiple-choice question answering (MCQA) and extractive question answering. In the former case, a question is given together with a list of possible answers, and the model is expected to select the correct answer [96, 146, 48]. In the latter case (extractive QA), a context is provided along with the question, and the model is expected to extract the answer span from this text [193, 174, 44]. Another variant of QA systems is abstractive QA, in which the model is expected to directly generate the answer [193, 44].

Semantic Textual Similarity The aim of this task is to measure the semantic textual similarity between two pieces of text either by providing the degree of similarity within a defined range or by simply classifying it as similar or not [20, 231, 202]. In other words, the task is concerned with how close is the meaning of one text to another. This task is useful for several applications, such as paraphrase or plagiarism identification, question answering systems or in specific domains such as healthcare. For example, in the biomedical domain, models for measuring semantic textual similarity could be used to find similar patient cases.

Information Extraction (IE) Information extraction mainly enables the automatic extraction of useful, structured or semi-structured data from unstructured text [79, 60, 90]. It is often used to address a number of NLP tasks, such as named entity recognition (NER), which is the task that aims to extract and classify entities from raw text [46]. Another example is relation extraction (RE), which is concerned with identifying and characterizing semantic relations between two entities from text [80].

2.9 Supervision Strategies

Supervised and unsupervised learning are the two standard approaches used within machine learning [85, 25]. In supervised learning, the model is trained with input data along with the expected output (i.e. label) [218]. Model performance is then measured by how accurately it maps previously unseen input data to the target label. Often the data are manually labelled by human annotators. Supervised learning is used for tasks such as classification and enables straightforward evaluation of the model performance [85]. On the contrary, with unsupervised learning, there is no need to provide the model with predefined labels. Unlabelled inputs are therefore used for training the model to find patterns and define common characteristics in order to analyze or cluster the data [218]. To evaluate and validate the model, experts might need to manually examine the results, which is subjective and time-consuming. Thus, the main distinction among these strategies is that one uses annotated data while the other does not [25]. In between supervised and unsupervised learning, several approaches have been introduced to overcome the cost of manually labelling the data while maintaining the advantages of supervised learning as much as possible. We discuss in the following two such approaches that are used in this thesis, namely distant and self-supervised learning.

Distant Supervision

Distant supervision aims to ease the process of obtaining the labels. In particular, the idea is to automatically annotate a set of unlabelled data using other external data sources such as knowledge bases or dictionaries [135, 148]. An important advantage of such an approach lies in its practicality, especially when the needed training data for a particular task is unavailable, hard, or expensive to obtain. It can also be helpful in augmenting already available training data at low-cost [127, 201, 230]. On the other hand, this automated way of annotating could inevitably lead to noisy, incomplete, wrong, and low-quality labels.

In general, distant supervision has been successfully applied in several tasks such as relation extraction [135, 254], named entity recognition [130], information retrieval [186], sentiment analysis [56, 171, 205], temporal recognition and normalization [206], question answering [256], and learning cross-lingual embeddings [33]. In the biomedical domain, Fu et al. [51] proposed the use of distant supervision as a preliminary step to minimize the load on human annotators for the task of assessing suicidal risk while [99] aimed to generate a distantly supervised dataset using heuristics to identify the phenotypes of depression from clinical notes. Taewijit et al. [207] automatically labels the relation between a drug and an event to capture adverse drug reactions by exploiting knowledge bases. Pattisapu et al. [157] extracted training examples for normalizing medical concepts from patient discussion forums using text embeddings models to identify the most similar phrases to a given medical concept. Furthermore, several works introduced new datasets constructed fully or partly in a distantly supervised way, such as ChemDisGene [248], TBGA [124], and MedDistant19 [11] for biomedical relation extraction. Also, a multimodal dataset named MELINDA [240] has been proposed for classifying biomedical experiment methods. As mentioned earlier, one of the shortcomings of this approximate approach is that it could result in noisy and imprecise labels. Therefore, as a remedial solution, another line of research has been proposing methods to de-noise the generated labels and filter which information to consider [177, 185, 108, 114, 251].

Self-Supervision

In self-supervised approaches, the labels are obtained by leveraging the data itself without the need for manual labelling or the use of external data sources [148, 172]. In other words, the learning process is derived from the structure of the input data itself through automated label generation. For instance, target labels can be extracted from data using a specific rule. One of the major advantages of self-supervised learning is that we can expand training data to massive amounts without the need for any form of external annotations or resources, as the labels are already part of the data. This, in fact, could allow the model to better generalize to more diverse examples which might be less represented with hand-labelled datasets. As a downside, however, using a greater amount of data would demand more computational power. A classic example of self-supervised models in NLP is BERT. More specifically, BERT employs MLM and NPS objectives for the self-supervised pre-training, advancing the field to new levels. Nevertheless, several works proposed alterations to those self-supervised objectives boosting the performance further, which we already discussed in Section 2.6.

2.10 Summary

In this chapter, we presented background information on different text representation methods, focusing in particular on pre-trained language models. We then explored the existing biomedical LMs. After that, we reviewed the literature on analysing knowledge captured by such models. Furthermore, we discussed works and general strategies that have been proposed in the literature to enhance standard language models. We then described some downstream NLP applications and the supervision techniques adopted in this thesis. In the next chapter, we will discuss the datasets, resources and tools that we have used in this thesis to perform the experiments and evaluate the proposed approaches.

Chapter 3

Datasets and Resources

3.1 Introduction

This chapter outlines the considered datasets, the textual and structured resources, and the tools employed in this thesis. As a general approach, we used binary textual entailment as the main task to train the proposed models and assess the strategies. In this task, the relation between two text fragments holds if one can be inferred from the other. Thus, we initially targeted two downstream tasks: Natural Language Inference (NLI) and Multiple-Choice Question Answering (MCQA), which we then recast as a binary textual entailment task. We primarily considered three available datasets in the biomedical domain: MedNLI, MedQA and HeadQA. The choice of each dataset was based on whether it contains, mainly or partly, patient case descriptions, which is the focus of this thesis. Additionally, we used various types of resources to carry out some of the experiments or to analyse the results, such as external textual data, named entity recognition tools and static embeddings of medical concepts.

Section 3.2 lists the details and defines the tasks along with the statistics for each dataset (i.e. the size of the training, validation and test sets). After that, section 3.3 presents the resources which we used to perform or analyse the experiments. In section 3.4, we provide an overview and describe the tools that were used for the preprocessing step of the unstructured text. Finally, section 3.5 summarises the chapter.

3.2 Evaluation Datasets

When it comes to properly evaluating the generalization capabilities of a strategy for a particular task (e.g. by having a set of evaluation datasets), the biomedical domain faces a challenge due to the lack of readily available datasets. Therefore, throughout this thesis, we attempt to compensate for that by adapting existing datasets to the considered problem, if needed. In this section, we list and describe these existing datasets.

3.2.1 MedNLI

MedNLI [184] is a clinical natural language inference (NLI) dataset that requires reasoning over domain-specific knowledge. The NLI task has been described in detail in Chapter 2. Since this task is concerned with predicting the relation between two pieces of text, it is in line with the settings that we consider to tackle the problem of interpreting patient case descriptions. Namely, the patient case description represents the premise and the diagnosis, for instance, represents the hypothesis. In other words, given a premise, the task aim is to determine if the given hypothesis could be inferred. Due to the restricted access, Table 3.1 shows examples that are slightly modified and similar in spirit to MedNLI examples. This dataset contains a total of 14,049 sentence pairs, particularly 11,232 for training, 1,395 for validation, and 1,422 for testing. The premise is derived from MIMIC-III v1.3 notes [86], which is described in 3.3.3. Specifically, snippets from the "Past Medical History" section are used for constructing the premises from these notes. Each premise is repeated three times for each hypothesis relation, whereas the hypothesis for each premise is written by expert annotators.

3.2.2 MEDIQA-NLI

MEDIQA shared challenge [2] introduced a set of tasks including natural language inference (NLI), Recognizing Question Entailment (RQE), and Question Answering

Premise	**Hypothesis**	**Gold Label**
Last night labs were drawn and she was found to have a Hgb of 7.2 per report	She is anemic	Entailment
	Her hgb is within normal limits	Contradiction
	She is fatigued	Neutral
A 54-year-old man with end-stage renal disease secondary to type 1 diabetes was admitted for a kidney transplant.	He is on insulin	Entailment
	He has normal renal function	Contradiction
	He has diabetic neuropathy	Neutral

Table 3.1: Examples approximating MedNLI instances, along with gold labels.

(QA) for the medical domain to encourage more research in these areas. We are particularly interested in the NLI shared task, which offers a new test set for the medical natural language inference while using the same training data as MedNLI. This test set consists of 405 premise and hypothesis pairs, and it follows the same annotations schemes as MedNLI, where each pair is labelled with either entailment, neutral, or contradiction. The premises are also driven from MIMIC-III notes, and the same annotators were asked to generate the hypotheses.

3.2.3 MedQA

MedQA [82] is a multiple-choice question answering dataset that is derived from medical exams where the questions require complex medical reasoning. MedQA is a publicly available dataset that is offered in three languages. Specifically, it covers English, traditional and simplified Chinese. In this thesis, we use the USMLE variant, which is the English version of the dataset. The total number of questions in the USMLE version is 12,723, where the training set consists of 10,178 questions, the validation of 1,272 questions, and 1,273 questions for testing. For each question, there are four answer candidates. The questions mainly represent patient case descriptions. Table 3.2 shows examples from this dataset.

Question	Answer candidates
A 38-year-old woman comes to the emergency department because of progressive headache, blurry vision, and nausea for 1 day. Four days ago, she was diagnosed with a right middle ear infection. She appears lethargic. Her temperature is 39.1°C (102.3°F), and blood pressure is 148/95 mm Hg. Ophthalmologic examination shows bilateral swelling of the optic disc. The corneal reflex in the right eye is absent. Sensation to touch is reduced on the upper right side of the face. Serum studies show increased concentrations of fibrin degradation products. Which of the following is the most likely diagnosis?	**(A) Cerebral venous thrombosis** **(B)** Hypertensive emergency **(C)** Subarachnoid hemorrhage **(D)** Viral meningitis
A 35-year-old man comes to the physician because of itchy, watery eyes for the past week. He has also been sneezing multiple times a day during this period. He had a similar episode 1 year ago around springtime. He has iron deficiency anemia and ankylosing spondylitis. Current medications include ferrous sulfate, artificial tear drops, and indomethacin. He works as an elementary school teacher. His vital signs are within normal limits. Visual acuity is 20/20 without correction. Physical examination shows bilateral conjunctival injection with watery discharge. The pupils are 3 mm, equal, and reactive to light. Examination of the anterior chamber of the eye is unremarkable. Which of the following is the most appropriate treatment?	**(A)** Erythromycin ointment **(B) Ketotifen eye drops** **(C)** Warm compresses **(D)** Fluorometholone eye drops

Table 3.2: Examples of questions from MedQA, along with the answer candidates. The correct answer is shown in bold.

3.2.4 Head-QA

Head-QA [219] is a multiple-choice question answering dataset that covers questions about different areas within the healthcare domain, such as medicine, psychology and biology. The language of this dataset is originally Spanish, but an English version is provided as well. We use the English version. Some questions correspond to patient case descriptions, but the majority of questions are about recalling specific factual knowledge. In addition, some questions in this dataset require interpreting images (i.e. some images are provided along with the question in order to drive the answer from both). As this is beyond the scope of this thesis, we discard all questions involving images in our experiments. This resulted in a total number of 2589 questions for training, 1336 for validation, and 2675 for testing. Table 3.3 presents a number of examples from this dataset.

Question	**Answer candidates**
The cardiolipin phospholipid is abundant in the membrane:	**(A) Internal mitochondrial** **(B)** External mitochondrial **(C)** Plasma. **(D)** Lysosomal
Jos, a 61-year-old man with obesity, sleep apnea, admitted to the ICU five days ago after abdominal surgery. Since the operation has not received sedatives for sleep, it is scheduled analgesia but fails to be effective with the consequent prolonged discomfort, in addition the environment is overstimulating and refers not to have a restful sleep. During the last 24 hours he has begun to manifest anxiety, increased sensitivity to pain, agitation that leads him to retire the night mask for sleep apnea, irritability and is even starting with hallucinations and an episode of aggression. Point out the present diagnostic label:	**(A)** Despair **(B) Deprivation of sleep** **(C)** Willingness to improve sleep **(D)** Sleep pattern disorder

Table 3.3: Examples of questions from HeadQA, along with the answer candidates. The correct answer is shown in bold.

3.3 Resources

As interpreting a patient case description is a domain-intensive task that requires a substantial amount of biomedical knowledge, we utilise a number of external resources to implement or analyse our work. In this section, we describe these resources in further detail.

3.3.1 UMLS Metathesaurus

The Unified Medical Language System (UMLS) metathesaurus [27] is a large biomedical vocabularies repository consisting of millions of medical concepts, their relations and semantic types (i.e. disease, drug, etc.), mainly incorporated from a set of existing ontologies systems. UMLS is provided by the US National Library of Medicine, and it basically unifies access to each medical concept from the different terminologies. It assigns a unique identifier to each concept, namely a concept unique identifier (CUI). Furthermore, it intends to facilitate the mapping between the different terminology systems such as SNOMED-CT, MeSH, and ICD-10. The UMLS metathesaurus allows for a broader usage for all these terminologies and could be used to build or enhance many biomedical applications. The UMLS metathesaurus is updated in a regular basis. We utilise the UMLS metathesaurus 2020AA full version release in this thesis. First, we use it to extract the medical concepts from the text in Chapters 4 and 6. We also use it to access other terminology hierarchies in Chapters 4 and 6.

3.3.2 SNOMED-CT Terminology

The SNOMED Clinical Terms (SNOMED-CT) [45] is a multilingual clinical vocabulary that enables the standardization of medical concepts across different languages. Therefore, it helps to maintain consistency and minimize the disparity when recording and sharing healthcare data such as patient records. SNOMED-CT is a terminological taxonomy that generally consists of medical terms, synonyms, and relationships. It places the medical concepts in hierarchies through the "is a" relationship, ranging from the specific to the most general concepts. Each concept is associated with a unique identifier. We use SNOMED-CT in this thesis to access the medical concept hierarchies, particularly in Chapters 4 and 6.

3.3.3 MIMIC-III

The Medical Information Mart for Intensive Care database (MIMIC-III) [86] is a large-scale, freely available set of de-identified information records about 53,423 patients admissions to the critical care units (ICU). Precisely, this dataset covers the period from 2001 to 2012 at the Beth Israel Deaconess Medical Center in Boston. MIMIC-III mainly consists of relational tables, specifically a total of 26 tables, which include both unstructured texts in the form of clinical notes and discharge summaries and structured data such as vital signs, lab results and medications. Access to MIMIC-III is restricted to approved users. We particularly use, in this thesis, the discharge summaries in Chapter 5.

3.3.4 WikiMed and PubMedDS Datasets

WikiMed and PubMedDS [216] are two publicly available datasets that were automatically constructed and proposed for medical entity linking tasks. WikiMed contains 393,618 Wikipedia articles (being those that mention some UMLS concept), while PubMedDS contains 13,197,430 PubMed abstracts. Despite the fact that these two datasets were initially introduced for entity linking, which is not our target task, we instead use them to enhance the LMs' interpretation of patient case descriptions. In particular, we are primarily interested in the included set of articles and abstracts relevant to the medical domain for developing our proposed methods. We use WikiMed in Chapter 5 and PubMedDS in Chapters 5 and 6.

3.3.5 Cui2vec Embeddings

Pre-trained cui2vec embeddings [22] is a publicly available pre-trained clinical concept embeddings. It is concerned with clinical concepts (i.e. CUI codes from UMLS) rather than words, and it is learned using the word2vec algorithm. In particular, cui2vec utilise a massive set of medical data from multiple sources, including insurance claims, PubMed articles, and clinical notes. To learn such embeddings, all mentioned concepts in these sources are mapped to their corresponding CUI codes as a pre-processing step. That resulted in learning 500-dimensional embeddings for a total of 108,477 medical concepts. We use cui2vec in Chapters 4 and 6.

3.4 Tools

To efficiently access or pre-process some of the above-mentioned resources, we need to use specific tools that were designed for this purpose. This section lists the tools that were used in this thesis.

3.4.1 MetaMap

MetaMap [16] is a commonly used biomedical named entity recognition tool that was developed by the US National Library of Medicine (NLM) to extract medical concepts from unstructured text. It is written in the Java programming language and basically extracts and then maps the medical concepts to their corresponding CUI codes (i.e. UMLS unique identifiers) ranked by their relevance. To extract the concepts, MetaMap might take a considerable amount of time to annotate the text due to the number of processing steps which the text needs to undergo. These steps consist of but are not limited to, tokenization, part-of-speech tagging, lexical lookup, syntactic analysis, locating then mapping medical concepts (by relying on UMLS metathesaurus), and finally, word sense disambiguation. MetaMap is a command-line tool that offers a number of format options to generate output files, including XML file format. The output file contains the identified phrases from a given text mentioning the medical terms, the candidate concepts with both the exact match and preferred concepts names, their CUIs, scores, semantic types, etc. After acquiring the output file, parsing the data is needed to be eventually used for the intended purpose. The download of MetaMap is available for authorized users only. We utilise MetaMap in Chapter 4.

3.4.2 QuickUMLS

QuickUMLS [196] is a biomedical named entity recognition tool for Python, which is used to extract medical concepts from free text. The purpose of this tool is similar to MetaMap, but it provides more efficient extraction suitable for large-scale data (faster by up to 135 times) while outperforming MetaMap on a number of evaluation benchmarks or achieving comparable performance. QuickUMLS maps the extracted medical concepts to the UMLS CUI identifiers by employing approximate string matching (i.e. locating fuzzy string patterns rather than an exact match) in an unsupervised way. This tool also provides the spans in which the medical terms appear, similarity scores and the UMLS semantic types of these medical concepts (e.g. disorders, drugs, etc.). To set up and use QuickUMLS, an installation of UMLS using the MetaMorphoSys tool is required beforehand. We use QuickUMLS in Chapters 5 and 6.

3.4.3 PyMedTermino

PyMedTermino [101] is a Python package that offers convenient and fast access to the various terminologies in UMLS and facilitates the mapping between them. For example, it can be used to map the CUI code of a medical term to its corresponding ICD-10 or SNOMED-CT terminology identifiers. In general, this tool provides an easy interface to approach the different medical terms within different terminologies systems, their synonyms, and their relations. The terminologies aren't included in the

API itself but rather require a separate download from the terminologies' sources due to the restricted access. We use the second version of PyMedTermino (PyMedTermino2) in Chapters 4 and 6.

3.5 Summary

In this chapter, we covered the details of the considered datasets to evaluate our proposed methods. Besides that, we described the tools and the external resources that were needed to perform the experiments. In the next chapter, we will explain how we utilised MedNLI dataset to drive the analysis of the LMs' capabilities while adapting it to the target problem.

Chapter 4

Probing Pre-Trained Language Models for Disease Knowledge

4.1 Introduction

Several biomedical LMs have enabled impressive results on various reading comprehension benchmarks for the medical domain, such as MedNLI [184] and MEDIQA-NLI [2] for Natural Language Inference (NLI), and PubMedQA [84] for QA. As an example, Wu et al. [241] achieved an accuracy of 98% on MEDIQA-NLI, which might suggest that medical NLI is essentially a solved problem. This would be exciting, as medical NLI intuitively requires a wealth of medical knowledge, much of which is not available in structured form.

However, a closer inspection of MedNLI, the most well-known medical NLI benchmark, reveals three important limitations, namely: (1) only few test instances actually require *medical disease* knowledge, with instances that (only) require terminological and lexical knowledge (e.g. understanding acronyms or paraphrases) being more prevalent; (2) training and test examples often cover the same diseases, and thus it cannot be determined whether good performance comes from the capabilities of the pre-trained LM itself, or from the fact that the model can exploit similarities between training and test examples; and (3) hypothesis-only baselines perform rather well on MedNLI, which shows that this benchmark has artefacts that can be exploited, similarly to general-purpose NLI benchmarks [168].

In this chapter, we aim to analyze to what extent pre-trained LMs are able to perform medical reasoning in a systematic way. More specifically, we focus on *disease knowledge*, which encompasses, for instance, the ability to link symptoms to diseases, or treatments to diseases. To this end, we propose DisKnE (Disease Knowledge Evaluation), a new benchmark for evaluating biomedical LMs. DisKnE is derived from MedNLI [184] and is organized into two top-level categories, which cover instances requiring medical and terminological knowledge, respectively. The medical category is furthermore divided into four sub-categories, depending on the type of medical knowledge that is required. Our proposed DisKnE dataset explicitly addresses the three limitations listed above and thus constitutes a more reliable testbed for evaluating the disease knowledge captured by biomedical LMs. We empirically analyse the performance of existing biomedical LMs, as well as the standard BERT model, on the proposed benchmark.

The remainder of this chapter is organised as follows. In Section 4.2 we review the related work to this chapter. In Section 4.3 we describe the process of constructing DisKnE. Subsequently, in Section 4.4, we thoroughly examine our experimental results. Lastly, in Section 4.5, we conclude the chapter by summarising our findings.

4.2 Related Work

There is a growing interest in designing probing tasks and analyzing what knowledge is captured by pre-trained LMs, which are now common across the NLP landscape, e.g., for word and sentence-level semantics [154, 38]. Most closely related to our work, Kearns et al. [92] presented an approach in which they categorise each sentence pair according to the tense and focus (e.g. medication, diseases, procedures, location) of the hypothesis, with the aim of providing a detailed examination of MEDIQA-NLI. Based on this categorization, they compare the performance of Enhanced Sequential

Inference Model (ESIM) using ClinicalBERT, Embeddings of Semantic Predications (ESP), and cui2vec. However, their analysis was limited to the MEDIQA-NLI test set, whereas we include entailment examples from the entire MedNLI and MEDIQA-NLI datasets. Moreover, we focus specifically on the ability of LMs to distinguish between closely related diseases, and we move away from the NLI setting to avoid training-test leakage and artefacts.

Adversarial NLI Several Natural Language Inference (NLI) benchmarks have been found to contain artefacts that can be exploited by NLP systems to perform well without actually solving the intended task [168, 62]. In particular, it has been found that strong results can often be achieved by only looking at the hypothesis of a (premise, hypothesis) pair. In response to this finding, several strategies for creating harder NLI benchmarks have been proposed. One established approach is to create adversarial stress tests [144, 55, 18], in which synthetically generated examples are created to specifically test for phenomena that are known to confuse NLI models. This may, for instance, involve the use of WordNet to obtain nearly identical premise and hypothesis sentences, in which one word is replaced by an antonym or co-hyponym. In this work, we rely on a somewhat similar strategy, using UMLS to replace diseases in hypotheses. As another strategy to obtain hard NLI datasets, Nie et al. [150] used human annotators to iteratively construct examples that are incorrectly labelled by a strong baseline model. While the aforementioned works are concerned with open-domain NLI, some work on creating adversarial datasets for the biomedical domain has also been carried out. In particular, Araujo et al. [13] studied the robustness of systems for biomedical named entity recognition and semantic text similarity, by introducing misspellings and swapping disease names by synonyms. To the best of our knowledge, no adversarial NLI datasets for the biomedical domain have yet been proposed.

Category	# inst.	Premise	Hypothesis
Symptoms → Disease	112	The patient developed neck pain while training with increasing substernal heaviness and left arm pain together with sweating.	The patient has symptoms of acute coronary syndrome
Treatments → Disease	60	The patient started on Mucinex and Robitussin.	The patient has sinus disease
Tests → Disease	116	Cardiac enzymes recorded CK 363, CK-MB 33, TropI 6.78	The patient has cardiac ischemia
		A large R hemisphere ICH was revealed when the patient had head CT	The patient has an aneurysm
Procedures → Disease	70	Bloody fluid was removed by pericardiocentesis	The patient has hemopericardium.
Terminological	259	The patient has urinary tract infection	The patient has a UTI
		The patient has high blood pressure	Hypertension
		Transfusions in the past could be the cause of the patient having hepatitis C	The patient has hepatitis C

Table 4.1: Considered categories of disease-focused entailment pairs.

4.3 Dataset Construction

In this section, we describe the process we followed for constructing DisKnE. As we explain in more detail in Section 4.3.1, this process involved filtering the entailment instances from the MedNLI and MEDIQA-NLI datasets, to select those in which the hypothesis expresses that the patient has (or is likely to have) a particular target disease. These instances were then manually categorized based on the type of knowledge that is needed for recognizing the validity of the entailment. Section 4.3.2 discusses our strategy for generating negative examples, which were obtained in an adversarial way, by replacing diseases occurring in entailment examples with similar ones. Details of the resulting training-test splits are provided in Section 4.3.3. In a final step, we canonicalize the hypotheses of all examples, as explained in Section 4.3.4. Note that the benchmark we propose consists of binary classification problems (i.e. predicting entailment or not), rather than the standard ternary NLI setting (i.e. predicting entailment, neutral, or contradiction), which is motivated by the fact that natural contradiction examples are hard to find when focusing on disease knowledge. Table 4.2 presents the key statistics data associated with DisKnE.

	Medical	**Terminological**
Examples	4133	2639
Diseases	47	24
Balance of positive and negative (P/N)	1/10	1/10

Table 4.2: Key statistics data for DisKnE.

4.3.1 Selecting Entailment Pairs

We started from the set of all entailment pairs (i.e. premise-hypothesis pairs labelled with the *entailment* category) from the full MedNLI and MEDIQA-NLI datasets. We used MetaMap to find those pairs whose hypothesis mentions the name of a disease, and to retrieve the UMLS CUI (Concept Unique Identifier) code corresponding to that disease.

We then manually identified those pairs, among the ones whose hypothesis mentions a disease, in which the hypothesis specifically expresses that the patient has that disease. For instance, in this step, a number of instances were removed in which the hypothesis expresses that the patient does not have the disease. The remaining cases were manually assigned to categories that reflect the type of disease knowledge that is needed to identify that the hypothesis is entailed by the premise. To clarify, our initial step involved employing MetaMap to filter out a specific subset of entailment pairs, specifically those where the hypothesis makes reference to a disease. Following that, the author performed a manual categorisation, allocating each filtered set of examples into distinct categories. Overall, the differentiation between each category was straightforward, and this was largely facilitated by the use of keywords in the hypothesis or the premise that provided some indication of the corresponding category. For instance, keywords like "started on" and "treated with" were indicative of the treatment category. The considered categories are described in Table 4.1, which also shows the number of (positive) examples we obtained and illustrative examples [1].

The primary distinction we make is between examples that need medical knowledge and those that need terminological knowledge. The former category is divided into four sub-categories, depending on the type of inference that is needed. First, we have the *symptoms-to-disease* category, containing examples where the premise describes the signs or symptoms exhibited by the patient, and the hypothesis mentions the corresponding diagnosis. Second, we have the *treatments-to-disease* category, where the premise instead describe medications (or other treatments followed by the patient). The third category, *tests-to-disease*, involves instances where the premise describes lab tests and diagnostic tools such as X-rays, CT scans and MRI. Finally, the *procedures-to-disease* category has instances where the premise describes surgeries and therapeutic procedures that the patient underwent. In the *terminological* category, the disease is mentioned in both the premise and hypothesis, either as an abbreviation, a synonym or within a rephrased sentence.

4.3.2 Generating Examples

The process outlined in Section 4.3.1 only provides us with positive examples. Unfortunately, MedNLI and MEDIQA-NLI contain only few negative examples (i.e. in-

[1] For data protection reasons, we only provide synthetic examples, which are different from but similar in spirit to those from the original MedNLI dataset.

stances of the *neutral* or *contradiction* categories) in which the hypothesis expresses that the patient has some disease. For this reason, rather than selecting negative examples from these datasets, we generate negative examples by corrupting the positive examples.

In particular, to generate negative examples, we replace the disease X from a given positive example by other diseases $Y_1, ..., Y_n$ that are similar to X, but not ancestors or descendants of X in SNOMED CT [45]. To identify similar diseases, we have relied on cui2vec [22], a pre-trained clinical concept embedding that was learned from a combination of insurance claims, clinical notes and biomedical journal articles. Apart from the requirement that the diseases $Y_1, ..., Y_n$ should be similar to X, it is also important that they are sufficiently common diseases, as including unusual diseases would make the corresponding negative examples too easy to detect. For this reason, we only consider the diseases that occur in the hypothesis of other positive examples as candidates for the negative examples. Specifically, among these set of candidate diseases, we selected the $n = 10$ most similar ones to X, which were not descendants or ancestors of X in SNOMED CT (as ancestors and descendants would not necessarily invalidate the entailment). This resulted in a total of 4133 examples requiring medical knowledge and 2639 examples requiring terminological knowledge.

4.3.3 Training-Test Splits

Because our focus is on evaluating the knowledge captured by pre-trained language models, we want to avoid overlap in the set of diseases in the training and test splits. In other words, if the model is able to correctly identify positive examples for a target disease X, this should be a reflection of the knowledge about X in the pre-trained model, rather than knowledge that it acquired during training. However, any single split into training and test diseases would leave us with a relatively small dataset. For this reason, we consider each disease X in isolation. Let $\mathcal{E}$ be the set of all positive examples, obtained using the process from Section 4.3.1. Furthermore, we write $\mathcal{E}_X$

for the set of those examples from $\mathcal{E}$ in which the target disease in the hypothesis is X. Finally, we write *neg*(X) for the set $\{Y_1, ..., Y_n\}$ of associated diseases that was selected to construct negative examples, following the process from Section 4.3.2.

For each target disease X, we define a corresponding test set *Test*$_X$ and training set *Train*$_X$ as follows. *Test*$_X$ contains all the positive examples from $\mathcal{E}_X$. Moreover, for each $e \in \mathcal{E}_X$ and each $Y \in$ *neg*(X) we add a negative example $e_{X \to Y}$ to *Test*$_X$ which is obtained by replacing the occurrence of X by Y. If the word before the occurrence of X is *a* or *an*, we modify it depending on whether Y starts with a vowel or consonant. The positive in *Train*$_X$ consist of all examples from $\mathcal{E}$ in which X is not mentioned. The negative in *Train*$_X$ consist of all examples that are not considered for *Test*$_X$.

Note that in *Train*$_X$, we also remove examples in which these diseases are only mentioned in the premise. Furthermore, we check for occurrences of all the synonyms of these diseases that are listed in UMLS. The process of creating the training and test set for a given target disease X is illustrated in Figure 4.1.

4.3.4 Canonicalization

We noticed that the way in which a given hypothesis expresses that “the patient has disease X” is correlated with the type of the disease. For this reason, as a final step, we canonicalize the hypotheses in the dataset. Specifically, we replace each hypothesis by the name of the corresponding disease X. Several hypotheses in the dataset already have this form. By converting the other hypotheses in this format, we eliminate any artefacts that are present in their specific formulation.

4.4 Experiments

We experimentally compare a number of pre-trained biomedical LMs on our proposed DisKnE benchmark. In Section 4.4.1, we first describe the considered LMs and the

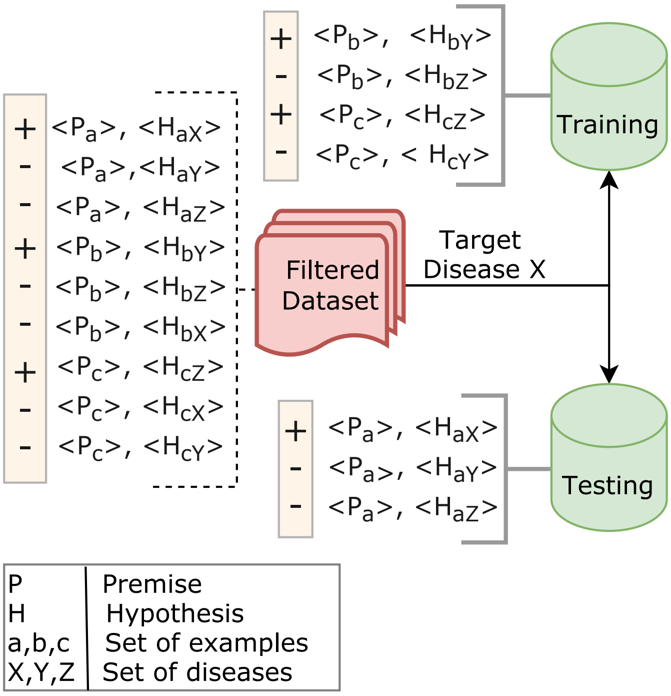

Figure 4.1: Illustration of training-test splitting process.

experimental setup. The main results are subsequently presented in Section 4.4.2. This is followed by a discussion in Section 4.4.3.

4.4.1 Experimental Setup

Pre-trained LMs. To understand to what extent the pretraining data of a LM affects its performance on our fine-grained evaluation of disease knowledge, we specifically used BERT$_{\text{base}}$, BioBERT [103] ClinicalBERT [9], and SciBERT [23].

Training Details. For fine-tuning, model hyper-parameters were the same across all BERT variants such as the random seeds, batch size and the learning rate. In this study, we fix the the learning rate at 2e-5, batch size of 8 and we set the maximum number of epochs to 8 with the use of early stopping. We used 10% of the training set as validation split.

Evaluation Protocol. We analyze the results per disease and per category in terms of F1 score for the positive class, reporting results for all diseases that have at least two positive examples for the considered category. To this end, for each disease X, we start from its corresponding training-test split, which was constructed as explained in Section 4.3.3. To show the results for a particular category, we remove from the test set all the examples that do not belong to that category.

4.4.2 Results

The main results are shown in Tables 4.3–4.7. A number of clear observations can be made. First, the results for the terminological category are substantially higher than the results for the other categories, which suggests that the masked language modelling objective, which is used as the main pre-training task in all the considered LMs, may not be ideally suited for learning medical knowledge. Second, recall that the main difference between the considered biomedical LMs comes from the corpora that were used for pre-training them. As the results for the terminological category (Table 4.7) reveal, the inclusion of domain-specific corpora does not seem to benefit their ability to model biomedical terminology, as similar results for this category are obtained with the standard BERT model, which was pre-trained on Wikipedia and a corpus of books and movie scripts. For the *Symptoms* $\rightarrow$ *Disease* category, we see that ClinicalBERT outperforms the other biomedical LMs, although the standard BERT model actually achieves the best performance overall. The results suggest that ClinicalBERT is better at distinguishing between relatively rare diseases, but that the focus on encyclopedic text benefits BERT for more common diseases. Intuitively, we can indeed expect that the encyclopedic style of Wikipedia focuses more on symptoms of diseases than scientific articles, which might focus more on treatments, procedures and diagnostic tests. This is also in accordance with the findings from He et al. [67], who obtained promising results with a disease-centric LM pre-training task that relies on Wikipedia. On the *Procedures* $\rightarrow$ *Disease* and *Tests* $\rightarrow$ *Disease* categories, we can

	ClinicalBERT	**BioBERT**	**SciBERT**	**BERT**
coronary atherosclerosis	0	0	29	10
chf	67	67	67	67
acs	04	33	0	05
stroke	80	56	90	90
heart disease	80	87	93	100
myocardial infarction	0	0	19	0
heart failure	0	0	22	0
urinary tract infection	100	100	67	100
disorder of lung	89	97	97	100
cirrhosis of liver	0	11	0	0
hyperglycemic disorder	27	13	22	0
pneumonia	89	93	67	100
neurological disease	67	67	80	67
respiratory failure	87	70	22	43
pulmonary edema	74	25	0	50
ami	0	0	0	0
deep vein thrombosis	47	48	50	48
acute cardiac ischemia	0	45	17	72
uri	78	45	67	83
cholangitis	22	22	33	22
atherosclerosis	66	0	67	0
Macro-average	$46_{\pm 3.0}$	$42_{\pm 7.3}$	$43_{\pm 3.1}$	$46_{\pm 3.4}$
Weighted average	$49_{\pm 3.1}$	$47_{\pm 6.0}$	$49_{\pm 2.7}$	$51_{\pm 2.7}$

Table 4.3: Results for the *Symptoms* → *Disease* category in terms of F1 (%) averaged over three runs. Standard deviations (over the three runs) of the macro and weighted average are also reported.

	ClinicalBERT	**BioBERT**	**SciBERT**	**BERT**
chf	55	55	53	55
acs	12	19	0	0
hypertensive disorder	55	67	54	22
heart disease	45	22	0	89
urinary tract infection	100	100	100	100
disorder of lung	82	89	100	93
hyperglycemic disorder	100	69	87	69
pneumonia	60	67	78	57
anemia	17	17	45	22
renal insufficiency	69	89	67	72
pulmonary infection	82	77	89	83
copd	45	67	61	39
hyperlipidemia	59	61	61	55
Macro-average	$60_{\pm 6.1}$	$61_{\pm 1.4}$	$61_{\pm 3.8}$	$58_{\pm 1.6}$
Weighted average	$51_{\pm 5.3}$	$54_{\pm 1.6}$	$51_{\pm 1.7}$	$45_{\pm 2.4}$

Table 4.4: Results for the *Treatments* → *Disease* category in terms of F1 (%) averaged over three runs. Standard deviations (over the three runs) of the macro and weighted average are also reported.

	ClinicalBERT	**BioBERT**	**SciBERT**	**BERT**
coronary atherosclerosis	0	0	0	0
chf	52	55	52	55
acs	0	22	0	0
stroke	87	87	95	77
hypertensive disorder	09	26	45	21
myocardial infarction	28	0	30	14
heart failure	0	55	40	0
urinary tract infection	87	90	59	90
hyperglycemic disorder	81	10	68	33
pneumonia	100	100	89	89
anemia	0	0	24	0
aortic valve stenosis	11	24	0	27
syst. inflam. resp. syndr.	76	64	80	80
acute renal failure syndr.	0	0	0	22
chronic renal insufficiency	0	0	0	0
kidney disease	22	0	45	0
ischemia	93	100	93	100
Macro-average	38 $_{\pm 2.4}$	37$_{\pm 1.6}$	42$_{\pm 3.1}$	36 $_{\pm 5.0}$
Weighted average	31 $_{\pm 2.6}$	32$_{\pm 1.2}$	37$_{\pm 1.5}$	31 $_{\pm 3.7}$

Table 4.5: Results for the *Tests* $\rightarrow$ *Disease* category in terms of F1 (%) averaged over three runs. Standard deviations (over the three runs) of the macro and weighted average are also reported.

	ClinicalBERT	**BioBERT**	**SciBERT**	**BERT**
coronary atherosclerosis	0	0	16	0
heart disease	83	74	84	84
heart failure	33	33	50	0
cirrhosis of liver	0	0	0	0
end stage renal disease	37	29	70	79
respiratory failure	58	27	57	27
renal insufficiency	100	100	93	100
cardiac arrest	100	100	93	100
disorder of resp. syst.	76	80	80	71
peripheral vascular dis.	0	0	78	0
Macro-average	$49_{\pm 3.2}$	$44_{\pm 5.9}$	$62_{\pm 3.9}$	$46_{\pm 5.0}$
Weighted average	$40_{\pm 3.3}$	$36_{\pm 7.4}$	$55_{\pm 5.6}$	$44_{\pm 4.6}$

Table 4.6: Results for the *Procedures* → *Disease* category in terms of F1 (%) averaged over three runs. Standard deviations (over the three runs) of the macro and weighted average are also reported.

	ClinicalBERT	**BioBERT**	**SciBERT**	**BERT**
anemia	95	100	100	93
aortic valve stenosis	100	100	93	100
carotid artery stenosis	50	50	60	50
coronary atherosclerosis	79	79	76	79
type 2 diabetes mellitus	67	56	64	61
gerd	0	0	0	0
cardiac arrest	95	97	92	97
heart disease	100	100	93	80
heart failure	100	100	100	100
chf	19	37	35	36
hyperglycemic disorder	57	63	80	57
hypertensive disorder	84	87	90	84
acute renal failure synd.	67	67	58	61
end-stage renal disease	77	77	78	70
disorder of lung	89	76	70	52
copd	100	100	97	100
myocardial infarction	24	25	25	21
pancreatitis	33	0	22	33
pleural effusion	80	100	100	80
pneumonia	89	93	89	66
pulmonary edema	87	82	56	76
stroke	81	100	71	100
urinary tract infection	78	77	78	77
aaa	100	96	100	100
Macro-average	$73_{\pm 2.7}$	$73_{\pm 0.4}$	$72_{\pm 2.5}$	$70_{\pm 3.2}$
Weighted average	$74_{\pm 1.8}$	$76_{\pm 1.4}$	$75_{\pm 1.3}$	$72_{\pm 3.0}$

Table 4.7: Results for the terminological category in terms of F1 ($\%$) averaged over three runs. Standard deviations (over the three runs) of the macro and weighted average are also reported.

see that SciBERT achieves the best results, with a particularly wide margin on the *Procedures* → *Disease* category. Finally, for the *Treatments* → *Disease* category, the relatively poor performance of BERT stands out, which conforms with the aforementioned intuition that scientific articles put more emphasis on procedures, treatments and tests. BioBERT achieves the best results, although the performance of the other biomedical LMs is quite similar.

4.4.3 Discussion

4.4.3.1 Which LM model?

Several published works have found ClinicalBERT to outperform the other considered biomedical LMs on biomedical NLP tasks [9, 92, 65]. In our results, however, SciBERT achieves the most consistent performance, clearly outperforming ClinicalBERT on the *Procedures* → *Disease* and *Test* → *Disease* categories, while performing similar to ClinicalBERT on the remaining categories.

However, rather than providing a blanket recommendation for SciBERT, our fine-grained analysis highlights the fact that different models have different strengths. The most surprising finding, in this respect, is the performance of the standard BERT model, which achieves the best results on the *Symptoms* → *Disease* category and performs comparably to BioBERT on several other categories (with *Treatments* → *Disease* being a notable exception).

4.4.3.2 Dataset Artefacts

As already reported by Romanov and Shivade [184], the original MedNLI dataset has a number of annotation artefacts, which mean that hypothesis-only baselines can perform well. In our dataset, we tried to address this by only using entailment examples, and creating negative examples by corrupting these. However, without canonicalizing the

		Standard		**Hyp. only**	
		full	**can**	**full**	**can**
MACRO	*Symptoms → Dis.*	$48_{\pm 0.7}$	$46_{\pm 3.0}$	$47_{\pm 4.9}$	$23_{\pm 0.5}$
	Treatments → Dis.	$64_{\pm 4.7}$	$60_{\pm 6.1}$	$65_{\pm 2.5}$	$29_{\pm 2.1}$
	Tests → Dis.	$41_{\pm 1.7}$	$38_{\pm 2.4}$	$44_{\pm 2.3}$	$18_{\pm 2.0}$
	Procedures → Dis.	$59_{\pm 4.9}$	$49_{\pm 3.2}$	$52_{\pm 2.6}$	$19_{\pm 3.0}$
	Terminological	$71_{\pm 2.3}$	$73_{\pm 2.7}$	$39_{\pm 1.3}$	$25_{\pm 0.4}$
WEIGHTED	*Symptoms → Dis.*	$54_{\pm 2.9}$	$49_{\pm 3.1}$	$53_{\pm 4.7}$	$23_{\pm 1.3}$
	Treatments → Dis.	$62_{\pm 2.8}$	$51_{\pm 5.3}$	$60_{\pm 7.1}$	$24_{\pm 1.0}$
	Tests → Dis.	$37_{\pm 1.4}$	$31_{\pm 2.6}$	$42_{\pm 0.2}$	$17_{\pm 2.8}$
	Procedures → Dis.	$54_{\pm 6.2}$	$40_{\pm 3.3}$	$59_{\pm 5.1}$	$14_{\pm 2.0}$
	Terminological	$71_{\pm 1.1}$	$74_{\pm 1.8}$	$41_{\pm 2.7}$	$22_{\pm 0.4}$

Table 4.8: Comparison between a variant with the full hypothesis and the proposed canonicalized version. Results are for the ClinicalBERT model in terms of F1 (%) averaged over three runs. Standard deviations (over the three runs) of the macro and weighted average are also reported.

hypotheses, we found that hypothesis-only baselines were still performing rather well. This is shown in Table 4.8, which summarizes the results we obtained for a version of our dataset without canonicalization, i.e. where the full hypotheses are provided, and the canonicalized version, where the hypotheses were replaced by the disease name only. The table shows results for the standard ClinicalBERT model, as well as for a hypothesis-only variant, which is only given the hypothesis. As can be seen, without canonicalization, the hypothesis only baseline performs similarly to the full model, even outperforming it in a few cases, with the exception of the *Terminological* category where a clear drop in performance for the hypothesis-only baseline can be seen. In contrast, for the canonicalized version of the dataset, we can see that the hypothesis only baseline, which only gets access to the name of the disease in this case, underperforms consistently and substantially. Note that the hypothesis-only baseline still achieves a non-trivial performance in most cases, noting that an uninformed classifier that always predicts true would achieve an F1 score of 0.167. However, this simply shows that the model has learned to prefer frequent diseases over rare ones.

4.4.3.3 Adversarial Examples.

A key design choice has been to select negative examples from the diseases that are most similar to the target disease. To analyse the impact of this choice, we carried out an experiment in which negative examples were instead randomly selected. As before, we only consider diseases that are present in the dataset, and we ensure that negative examples are not ancestors or descendants of the target disease in SNOMED CT. The results are presented in Table 4.9. As expected, the results are overall higher than those from the main experiment. More surprisingly, this easier setting benefits some models more than others. The relative performance of ClinicalBERT in particular is now clearly better, with this model achieving the best results for *Symptoms → Disease*. Furthermore, the standard BERT model now clearly underperforms the biomedical LMs, except for *Procedures → Disease* where it outperforms ClinicalBERT and BioBERT.

4.5 Conclusion

In this chapter, we have proposed DisKnE, a new benchmark for analysing the extent to which biomedical LMs capture knowledge about diseases. Positive examples were obtained from MedNLI and MEDIQA-NLI, by manually identifying and categorizing hypotheses that express that the patient has some disease. Negative examples were selected to be similar to the target disease. To prevent shortcut learning, the hypotheses were canonicalized, such that models only get access to the name of the disease that is inferred. Our empirical analysis shows that existing biomedical language models particularly struggle with cases that require medical knowledge. The relative performance on the different categories suggests that different (biomedical) LMs have complementary strengths. In the next chapter, we will focus on improving the performance of these

		ClinicalBERT	**BioBERT**	**SciBERT**	**BERT**
MACRO	*Symptoms → Dis.*	$66_{\pm 4.0}$	$56_{\pm 3.2}$	$57_{\pm 5.2}$	$56_{\pm 4.1}$
	Treatments → Dis.	$69_{\pm 4.3}$	$70_{\pm 2.0}$	$76_{\pm 4.5}$	$55_{\pm 4.8}$
	Tests → Dis.	$53_{\pm 0.9}$	$49_{\pm 3.3}$	$52_{\pm 1.0}$	$47_{\pm 0.6}$
	Procedures → Dis.	$60\ _{\pm 1.8}$	$56_{\pm 0.8}$	$76_{\pm 2.6}$	$60_{\pm 4.5}$
	Terminological	$77_{\pm 0.9}$	$77_{\pm 0.6}$	$74_{\pm 0.6}$	$76_{\pm 1.0}$
WEIGHTED	*Symptoms → Dis.*	$66\ _{\pm 5.2}$	$59_{\pm 3.5}$	$59_{\pm 4.1}$	$56_{\pm 4.6}$
	Treatments → Dis.	$64_{\pm 6.2}$	$59_{\pm 3.6}$	$68_{\pm 4.8}$	$46_{\pm 3.1}$
	Tests → Dis.	$53\ _{\pm 0.6}$	$51_{\pm 2.4}$	$54_{\pm 1.6}$	$43_{\pm 4.0}$
	Procedures → Dis.	$65_{\pm 3.0}$	$58_{\pm 1.0}$	$76_{\pm 0.4}$	$67_{\pm 4.5}$
	Terminological	$76\ _{\pm 1.6}$	$77_{\pm 1.0}$	$75\ _{\pm 0.4}$	$72\ _{\pm 0.7}$

Table 4.9: Results for a variant of our benchmark, in which negative examples were selected at random, in terms of F1 (%) averaged over three runs. Standard deviations (over the three runs) of the macro and weighted average are also reported.

LMs while specifically targeting medical knowledge. In particular, we will use the DisKnE medical knowledge variant along with other datasets to evaluate our proposed approach.

Chapter 5

Interpreting Patient Descriptions using Distantly Supervised Similar Case Retrieval

5.1 Introduction

The previous chapter has shown that existing biomedical LMs often struggle with medical reasoning tasks, such as linking symptoms to diseases. Moreover, it found that the standard BERT model was remarkably competitive with specialised biomedical LMs for inferring diagnoses from patient case descriptions. Therefore, in this chapter, we intend to improve the medical reasoning abilities of these LMs. In general, many techniques have been proposed to enhance the performance of pre-trained LMs using unstructured text. Although the biomedical domain has vast volumes of unstructured text data, annotation is costly. This hinders the ability to fully explore and thereby benefit from the rich information in the biomedical literature to enhance the LMs. Thus, we also aim to address this annotated data scarcity problem with a distant supervision strategy.

One possible approach to enhance the predictions of LMs with unstructured data is by augmenting the LM model input with retrieved sentences, which could sometimes be sufficient to fill the gap with the needed knowledge. However, previous work [204] has shown the limitation of this approach for the biomedical domain. Therefore, instead of retrieving explicit medical knowledge, in this chapter, we rely on a nearest neighbour strategy.

The remainder of the chapter is organised as follows. Section 5.2 provides our motivation and the overall proposed strategy. Subsequently, Section 5.3 reviews the relevant work to this study. After that, Section 5.4 describes in detail the proposed method. Section 5.5 presents our experimental results along with considered datasets, external text corpora and analysis. Finally, Section 5.6 concludes the chapter.

5.2 Motivation & Overall Strategy

To alleviate the limitations of biomedical LMs, a natural strategy would be to augment patient case descriptions with sentences expressing relevant knowledge, which are retrieved from some text corpus. Similar strategies have already proven useful for factual and commonsense question answering [137, 87, 194]. When it comes to interpreting patient case descriptions, however, the potential of such strategies is less clear. For instance, Sushil et al. [204] used an information retrieval engine to find relevant sentences in biomedical corpora, which were then added to the premise of Natural Language Inference (NLI) instances. In experiments on MedNLI [183], they found no statistically significant improvements as a result of this augmentation strategy. While retrieved sentences can be helpful to clarify the meaning of an unusual term, or to provide specific knowledge, it is unlikely that we would find a sentence that captures the specific knowledge that is needed to infer a diagnosis, or recommend a particular treatment, from a given patient case description. Indeed, such inferences are often a matter of clinical judgement, more than the application of rule-like knowledge that could be expressed in a sentence [187, 237].

Rather than searching for sentences that directly express medical knowledge, we aim to find passages that are similar to the given patient case description itself. The underlying intuition is that such passages are likely to describe patients in similar situations, and that whatever is true for these patients is likely to be true for the patient from the given description as well. We specifically focus on passages that also mention some hypothesis of interest, e.g. an answer candidate in the context of question answering (QA). We then estimate the likelihood that this hypothesis holds based on the similarity between the given patient case description and the retrieved passages. Figure 5.1 shows an overview of the overall strategy. The use of similar cases plays an important role in clinical decision making [21, 15, 213, 139], hence the use of a nearest neighbour strategy is natural and conceptually straightforward. Moreover, the idea of retrieving similar cases is also appealing from an application perspective, as these cases can be used as supporting evidence for a given prediction. This is particularly important for the biomedical domain, where explainability and transparency are clearly paramount.

However, the success of such a nearest neighbour strategy critically hinges on our ability to identify the commonalities between different patient case descriptions in a suitable way, which is in itself a challenging problem. For instance, even if two patients experienced a similar situation, the details of their cases are likely to differ in many respects, some of which may or may not matter. Moreover, the patient case descriptions may differ in the level of detail they provide, as well as their overall writing style. To illustrate these issues, Table 5.1 shows the top passage that was retrieved by our model for a given question from the MedQA benchmark [82]. As can be seen, both patient case descriptions refer to the sudden development of unusual behaviour shortly after experiencing bereavement. Beyond this central correspondence, however, the details of the two descriptions differ substantially. Identifying relevant patient case descriptions is thus a non-trivial problem, which requires specialised clinical knowledge. Given these challenges, off-the-shelf models for estimating textual similarity are clearly insufficient for identifying relevant patient case descriptions. Moreover, to the best of our knowledge, there are no labelled datasets that can be used for training a supervised model. This makes the problem of interpreting patient case descriptions inherently different from settings such as open-domain QA, where gold annotations of

Question: A 20-year-old woman is brought in for a psychiatric consultation by her mother who is concerned because of her daughter's recent bizarre behavior. The patient's father died from lung cancer 1 week ago. Though this has been stressful for the whole family, the daughter has been hearing voices and having intrusive thoughts ever since. These voices have conversations about her and how she should have been the one to die and they encourage her to kill herself. She has not been able to concentrate at work or at school. She has no other history of medical or psychiatric illness. She denies recent use of any medication. Today, her heart rate is 90/min, respiratory rate is 17/min, blood pressure is 110/65 mm Hg, and temperature is 36.9°C (98.4°F). On physical exam, she appears gaunt and anxious. Her heart has a regular rate and rhythm and her lungs are clear to auscultation bilaterally. CMP, CBC, and TSH are normal. A urine toxicology test is negative. What is the patient's most likely diagnosis?

Answer candidate: Brief psychotic disorder

Retrieved passage: Brief psychotic disorder associated with bereavement in a patient with terminal-stage uterine cervical cancer: a case report and review of the literature. We report here a terminally ill patient with uterine cervical cancer who developed a brief psychotic disorder after bereavement following the loss of three close friends also suffering from gynecological cancer. A 49-year-old housewife, who was diagnosed as having uterine cervical cancer and was receiving palliative care was referred for psychiatric consultation because of an abrupt onset of delusions, bizarre behavior, disorganized speech, and catatonic behavior. On psychiatric examination, she showed delusional thought and catatonic behavior. Laboratory data were unremarkable, as was brain MRI. She had no history of psychiatric illness or drug or alcohol abuse. After receiving haloperidol, psychiatric symptoms disappeared, and she returned to the previous level of functioning after 3 days. The patient explained that the death of three of her friend due to gynecological cancer was shocking event for her. She focused her attention on her own fears of dying from the same disease. Brief psychotic disorder in cancer patients is rare in the literature. However, our report of brief psychotic disorder associated with bereavement may highlight possible precipitating factors, which have not been adequately emphasized in the literature to date.

Table 5.1: Example of a question from MedQA, along with the top-retrieved passage by our model for the answer candidate *brief psychotic disorder*.

relevant passages are often available and systems can rely on transfer learning from closely related tasks.

In this study, we propose a distant supervision strategy to address these challenges. We start from the intuition that interpreting patient case descriptions is easier than open-domain QA in one important aspect: the presence of a hypothesis (or answer candidate) in a context passage makes it highly likely that this passage is at least somewhat relevant, which is related to the fact that we are looking for similar cases rather than for specific knowledge. For instance, most patient case descriptions mentioning *brief psychotic disorder* would tell us something about the likelihood that this is the correct diagnosis for the question in Table 5.1. In contrast, passages mentioning *Paris* may be completely irrelevant to a question asking about the capital of France. Our central hypothesis is that this aspect of patient case descriptions can compensate for the lack of relevant supervision data for learning to identify similar cases. In particular, we propose a strategy to train a *cross-encoder* for comparing patient case descriptions, i.e. a fine-tuned language model which takes two patient case descriptions as input and estimates their degree of similarity. To this end, we generate a distantly supervised training set, by using a baseline model to rank candidate passages and relying on the assumption that such a passage is relevant if it mentions a hypothesis that can be inferred from the target patient case description. Conceptually, this is similar in spirit to distant supervision strategies for open-domain QA (see Section 5.3). A key difference, however, lies in the fact that we cannot use standard retrieval models for ranking the candidate passages. Our solution relies on the following two steps:

- We train an unsupervised text encoder on a set of patient case descriptions. This encoder is used to select an initial set of candidate passages. It has two primary advantages: (i) it allows for efficient dense retrieval of a small set of candidate passages and (ii) it can rely on some clinical knowledge of patient case descriptions because it was trained on this domain.

- The initial set of candidate passages is then ranked using a pre-trained cross-encoder. We initialise this cross-encoder from a biomedical LM and pre-train it on a standard textual similarity dataset. Despite not being trained on patient case descriptions, we anticipate that this re-ranking step will improve the effectiveness of our approach. Intuitively, an out-of-domain cross-encoder can be effective because all of the candidate passages are (at least somewhat) relevant. The model can thus focus on identifying more particular commonalities, which may not require as much clinical knowledge.

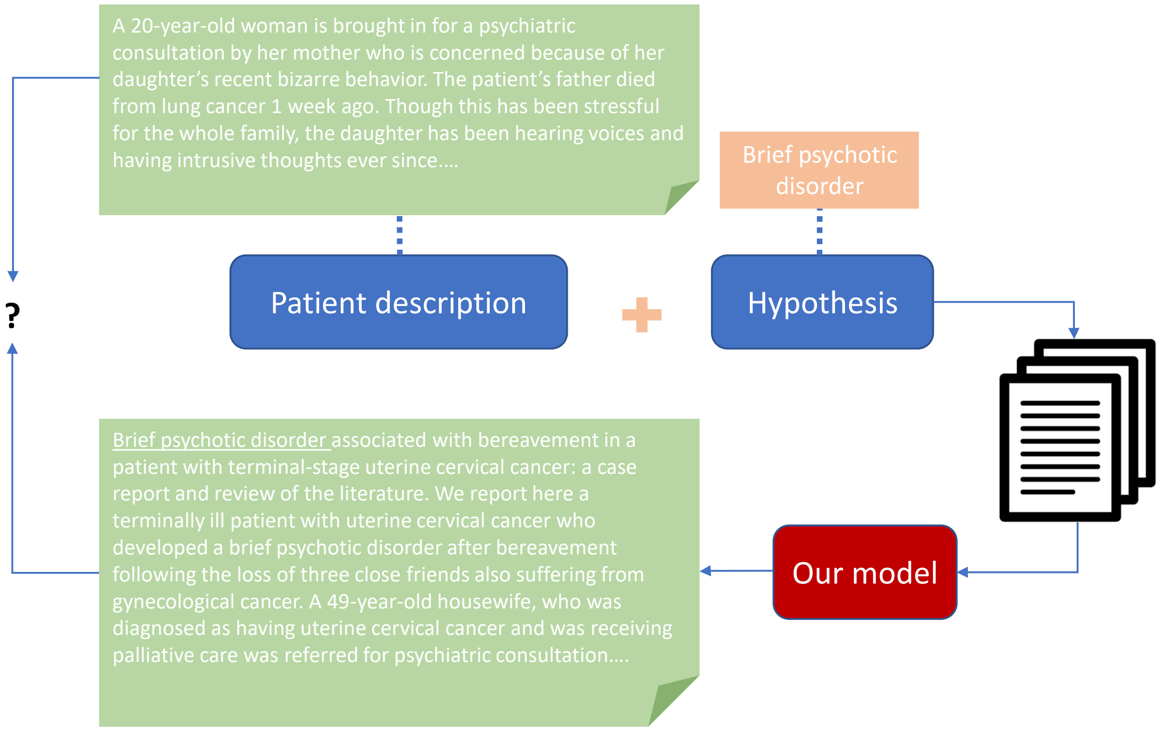
A 20-year-old woman is brought in for a psychiatric consultation by her mother who is concerned because of her daughter's recent bizarre behavior. The patient's father died from lung cancer 1 week ago. Though this has been stressful for the whole family, the daughter has been hearing voices and having intrusive thoughts ever since....
Brief psychotic disorder
?
Patient description
Hypothesis
Brief psychotic disorder associated with bereavement in a patient with terminal-stage uterine cervical cancer: a case report and review of the literature. We report here a terminally ill patient with uterine cervical cancer who developed a brief psychotic disorder after bereavement following the loss of three close friends also suffering from gynecological cancer. A 49-year-old housewife, who was diagnosed as having uterine cervical cancer and was receiving palliative care was referred for psychiatric consultation....
Our model

Figure 5.1: Overview of the overall strategy.

5.3 Related Work

Distant Supervision in IR. The application of distant supervision strategies has seen considerable success in scenarios where gold-annotated data is scarce, e.g., in open question answering or dense retrieval. Most relevant to our study, several retrieval models that combine distant supervision with BERT-based encodings have been proposed in recent years. For instance, Karpukhin et al. [91] trained a dual encoder (i.e. separate passage and question encoders) for open question answering, which uses distantly labelled question-passage pairs for those datasets where gold annotations are not available. To obtain positive examples, for a given question, they then select those passages which contain the answer and are ranked highest using BM25 [182]. They use several strategies for selecting negative passages, e.g. taking the top retrieved passages that do not mention the answer. Our model similarly obtains positive examples from top-ranked passages, but given the challenging nature of patient case descriptions, we found that relying on BM25 for generating pseudo-labels was not sufficient and that the use of a cross-encoder for the final model was essential. The use of cross-encoders for open-domain QA has also been extensively explored. However, different from our setting, most works rely on gold annotations of passage relevance [243, 173]. These gold labels are used to train the cross-encoder, which is used to generate pseudo-labels. These pseudo-labels are then in turn used for training an improved dual encoder model. In other words, these works are using a supervised cross-encoder to generate pseudo-labels, whereas our focus is on generating pseudo-labels for training the cross-encoder itself. Rather than using a cross-encoder, Khattab et al. [95] start from a pre-trained ColBERT model [94] to get an initial ranking of passages that are similar to the question. ColBERT separately encodes the passages and question, but rather than representing these text fragments as single vectors, they are represented as sequences of token-level vectors, which enables a finer-grained interaction than standard dual encoders. Given the ColBERT ranking, they assume that the top-k passages are positive examples if they contain the answer candidate and negative examples otherwise. Based on these pseudo-labels, the ColBERT model is then fine-tuned. This process is repeated a few times to iteratively improve the model. The ability to pre-train ColBERT on a relevant supervised task is crucial to this approach, however, hence a similar strategy cannot straightforwardly be applied to the setting of patient case descriptions. The aforementioned methods rely on a baseline retrieval model to generate pseudo-labels, which is also the approach we follow in this work. As an alternative, some authors have also proposed models in which the retrieval model is jointly optimised with the rest of the QA model [105, 63]. However, these approaches involve computationally intensive language model pre-training tasks, which makes them difficult to implement and analyse. More widely, distant supervision is also commonly used for span selection in open-domain QA [73] and for ad-hoc document retrieval [126], among many others.

Similar Case Retrieval. Within NLP, similar case retrieval has primary been applied to the analysis of legal cases. For instance, Westermann et al. [238] proposed a strategy for finding legal cases that are similar to a given one, which involved an initial filtering step to eliminate cases that are unlikely to be related, followed by the use of an SVM model for making the final prediction. Shao et al. [190] introduced BERT-PLI. Given a query case, they first retrieve potentially relevant cases from a corpus of legal cases using BM25. Subsequently, they use a BERT model that was fine-tuned on a legal entailment dataset. This model is applied to individual paragraphs from the query and candidate cases, with the final score being obtained by aggregating the paragraph-level interactions. Shao et al. [189] combine the features extracted from BERT-PLI with traditional bag-of-words features, and then use RankSVM to rank the considered cases. Summarizing the retrieved cases before ranking them has been investigated as well, as a strategy to deal with documents that are longer than the language model can handle [10].

Beyond the legal domain, the idea of exploiting similar cases has recently been used for question answering [176], semantic parsing [246], text generation [214] and language modelling [93] among many others. Within the biomedical domain, one relevant line of research aims to capture the similarity between different patients to predict, for example, a diagnosis or treatment [78, 143, 71], usually by learning a dense vector representation of each patient. Another related line of research has focused on linking patient records to relevant articles from the biomedical literature [181, 180].

5.4 Proposed Method

We are interested in the problem of interpreting patient case descriptions. More specifically, given a patient case description $\mathcal{D}$ and a hypothesis H, we are interested in determining whether H can be inferred from $\mathcal{D}$, i.e. whether $\mathcal{D}$ entails H. For instance, H could be a diagnosis or a recommended treatment, diagnostic test or procedure. In the example displayed in Table 5.1, the question corresponds to the patient case description $\mathcal{D}$ while the given answer candidate (i.e. *brief psychotic disorder*) corresponds to the hypothesis H.

To determine whether $\mathcal{D}$ entails H, we search for a text fragment $\mathcal{C}_H$, from a given corpus, which (i) mentions H and (ii) is as similar as possible to $\mathcal{D}$. We then use the similarity between $\mathcal{D}$ and $\mathcal{C}_H$ to assess the likelihood that H is entailed by $\mathcal{D}$. The underlying intuition is that $\mathcal{C}_H$ and $\mathcal{D}$ are both presumed to be patient case descriptions, and moreover, that the fact that H is mentioned in $\mathcal{C}_H$ means that H can be inferred from that patient case description.

Our central aim is to demonstrate the strong potential of nearest neighbour strategies for interpreting patient case descriptions, and to show how the main technical obstacles can be overcome, in particular the lack of training data for learning to recognise similar patient case descriptions. To focus the empirical analysis on these key aims, we keep our overall model as simple as possible. To this end, we rely on the following simplifying assumptions:

- We assume that there will exist relevant text fragments that literally mention the hypothesis H.
- We assume that text fragments which are similar to the patient case description $\mathcal{D}$ will themselves also be patient case descriptions.

- We take the fact that H is mentioned in the text fragment $\mathcal{C}_H$ as evidence that H applies to the patient being described.

In principle, it is possible to weaken some of these assumptions. For instance, rather than looking for literal mentions of H, we could use a medical concept normalisation method such as MetaMap [16] or QuickUMLS [196] to identify phrases with the same meaning. Similarly, rather than simply looking for passages that mention H, we could use a baseline NLI model to check whether H can be entailed from $\mathcal{C}_H$. However, such solutions may themselves introduce errors. Furthermore, as we will see, sometimes passages are retrieved that are not patient case descriptions but which nonetheless help the model to make the correct prediction. We can often think of such passages as being generic patient case descriptions, e.g. discussing how a given illness in general presents itself, hence specifically restricting the retrieved passages to actual patient case descriptions may not always be helpful. We leave a detailed study of these considerations for future work.

We next present a more detailed overview of our approach. In Section 5.4.2 we then describe our strategy for generating a distantly supervised training set, which will allow us to train the cross-encoder that sits at the heart of our model. Finally, Section 5.4.3 describes how the cross-encoder is used as part of our overall model.

5.4.1 Overview of the Nearest Neighbour Strategy

Let $\mathcal{D}$ be a patient case description and let $\mathcal{C}_H$ be a text fragment mentioning some hypothesis of interest H. We want to train a model that allows us to predict whether $\mathcal{C}_H$ is sufficiently similar to $\mathcal{D}$ to plausibly infer that H can be entailed from $\mathcal{D}$. We use a cross-encoder to this end, i.e. we fine-tune a language model to predict similarity scores, where the concatenation of $\mathcal{D}$ and $\mathcal{C}_H$ (separated by the special <*sep*> token) is used as the input. Cross-encoders are able to measure similarity in a more intricate way than strategies that rely on comparing sentence embeddings, but the latter are

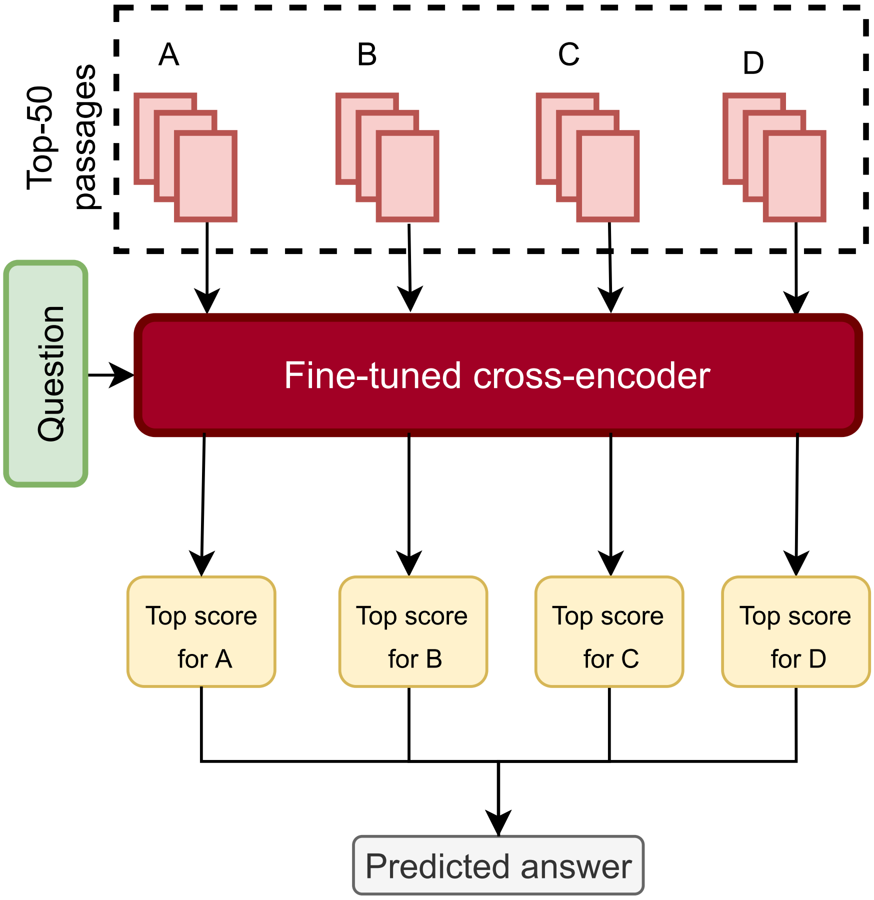

Figure 5.2: Overview of the application of our proposed model for answering multiple-choice questions.

more scalable. For this reason, in line with the standard usage of cross-encoders as re-rankers in information retrieval [37, 151, 112], we first use sentence embeddings to identify the 50 most similar text fragments containing H and then use the fine-tuned cross-encoder for identifying the most similar text fragment among these. Figure 5.2 illustrates how the overall process can be applied to multiple-choice question answering. In this case, for each of the answer candidates A, B, C, D we retrieve an initial set of 50 text fragments and then use the cross-encoder to find the single most similar document from each set. Let us call these documents $\mathcal{C}_A$, $\mathcal{C}_B$, $\mathcal{C}_C$ and $\mathcal{C}_D$. For instance, $\mathcal{C}_A$ is assumed to be the text fragment which is most similar to $\mathcal{D}$, among all those mentioning A. The model would then, for instance, predict answer candidate A if $\mathcal{C}_A$

is estimated to be more similar to $\mathcal{D}$ than $\mathcal{C}_B$, $\mathcal{C}_C$ and $\mathcal{C}_D$.

5.4.2 Obtaining Similarity Labels

We assume that we are given a set of positive examples E^+ of the form $(\mathcal{D}, H)$, where $\mathcal{D}$ is a patient case description and H is a hypothesis that can be inferred from $\mathcal{D}$. Similarly, we assume we have a set of negative examples E^- of the form $(\mathcal{D}, H)$, where H cannot be inferred from $\mathcal{D}$. For instance, in the setting of multiple-choice question answering, E^+ would be constructed from the correct answer candidates whereas E^- would be constructed from the incorrect answer candidates. Similarly, the sets E^+ and E^- can be straightforwardly obtained from NLI training data.

To allow us to train the cross-encoder, we derive a synthetic training set $\mathsf{S}^+ \cup \mathsf{S}^-$ from E^+ and E^-. This training set consists of pairs $(\mathcal{D}, \mathcal{C}_H)$, where $\mathcal{C}_H$ is a passage that was retrieved, by an unsupervised retrieval model, as one of the top-k most similar text fragments to $\mathcal{D}$ containing the hypothesis H. In particular, the set of positive examples S^+ contains those pairs $(\mathcal{D}, \mathcal{C}_H)$ for which $(\mathcal{D}, H) \in \mathsf{E}^+$, whereas S^- contains those pairs for which $(\mathcal{D}, H) \in \mathsf{E}^-$. Note how this overall strategy is somewhat reminiscent of pseudo-relevance feedback [32, 242, 31, 102], in the sense that we rely on the assumption that the top-k retrieved passages are relevant. However, rather than trying to improve a ranked list of passages, our aim is to train a cross-encoder to distinguish between passages that contain valid hypotheses and those that do not. In principle, this could be done without a retrieval model, by simply assuming that passages $\mathcal{C}_H$ are similar to $\mathcal{D}$ if and only if the hypothesis H they contain can be inferred from $\mathcal{D}$. Our purpose in restricting the training data $\mathsf{S}^+ \cup \mathsf{S}^-$ to the top-k retrieved passages is to denoise the supervision labels as much as possible.

The quality of the training set $\mathsf{S}^+ \cup \mathsf{S}^-$ crucially relies on the retrieval model that is used to select the top-k passages. To obtain these passages, we rely on a two-step process. First, an unsupervised sentence embedding model is used to select the top-50

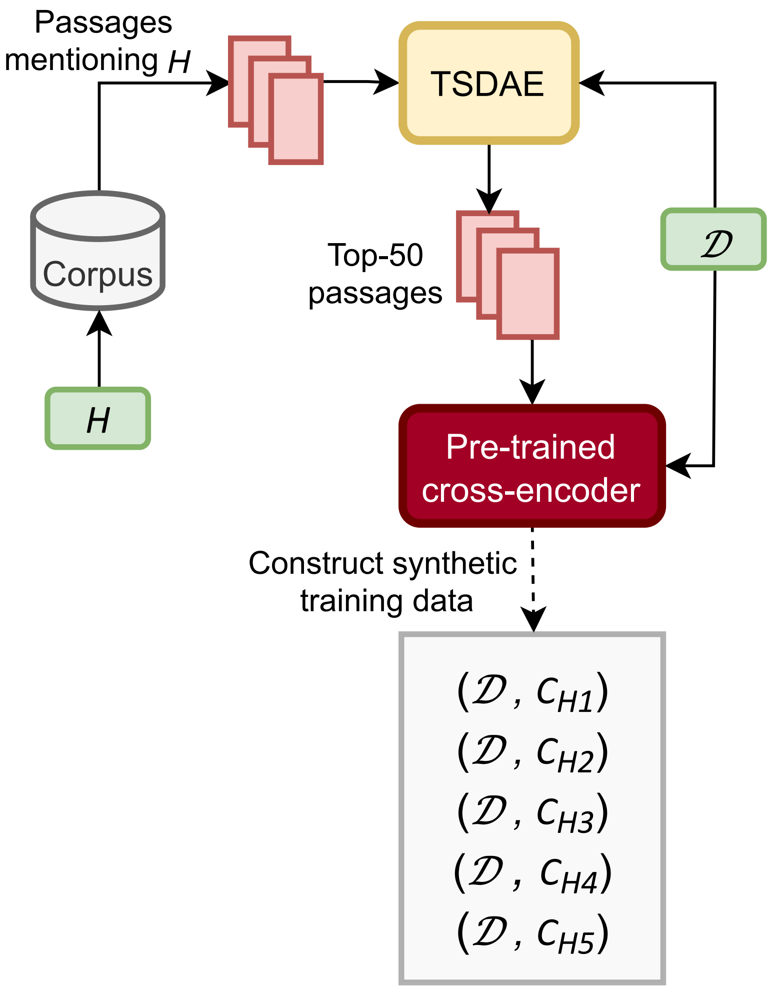

Figure 5.3: Overview of how the distantly supervised examples for training the cross-encoder are obtained (shown for $k = 5$).

most similar passages. Subsequently, we use a pre-trained cross-encoder to select the k most similar passages among these 50 (with $k < 50$). We now describe these two steps in more detail. The overall process for generating the training set $\mathsf{S}^+ \cup \mathsf{S}^-$ is illustrated in Figure 5.3 .

5.4.2.1 Initial Retrieval Step

Given a pair $(\mathcal{D}, H)$ we first use Elasticsearch [59] to retrieve all text passages mentioning H. For efficiency reasons, in our experiments we retrieve a maximum of 1000 passages. We then use an unsupervised sentence embedding model to encode each of the selected passages, as well as the patient case description $\mathcal{D}$ itself. We use these embeddings to select the 50 passages that are most similar to $\mathcal{D}$ in terms of cosine similarity. Specifically, we use the Tranformer-based Denoising AutoEncoder (TSDAE) approach [228] to train a sentence embedding model for the clinical domain.

We initialize this model from ClinicalBERT and use MIMIC-III [86] discharge summaries as input fragments for training. Due to the noisy nature of these summaries, rather than working at the sentence level, we split the documents in passages of up to 250 words, while respecting sentence boundaries.

5.4.2.2 Reranking with a Pre-Trained Cross-Encoder

We rely on a pre-trained cross-encoder to identify the most relevant passages, among the 50 that were selected based on their TSDAE embeddings. We experiment with cross-encoders that are trained on one of the following tasks:

- **Semantic Textual Similarity Benchmark (STS-B)** [34]: An open-domain benchmark where the goal is to determine the semantic relatedness between two sentences as a score from 1 to 5.
- **Recognizing Question Entailment (RQE)** [1]: Given a pair of health-related questions, this binary classification dataset aims to identify whether the answer to the second question is also a complete or partial answer to the first. The question pairs were retrieved from Frequently Asked Questions on the National Institutes of Health (NIH) websites, as well as consumer health questions collected by the National Library of Medicine.

- **HealthQA** [259]: A set of question and answer pairs annotated with relevance labels. The answers were collected from the Patient website [1] and questions were provided by human annotators.

STS-B and RQE have already been found useful for improving semantic similarity tasks, including in the clinical domain [122]. We also include HealthQA because of its structural similarity with our considered setting. Note that none of these pre-training tasks involve patient case descriptions, while STS-B is not even focused on the biomedical domain.

5.4.3 Training and Using the Cross-Encoder

We use the training set $\mathsf{S}^+ \cup \mathsf{S}^-$ to fine-tune our cross-encoder. We initialise the model with the pre-trained cross-encoder that was used for the reranking step in Section 5.4.2.2. To use the resulting model, e.g. for QA or NLI, we again use the TSDAE sentence encoder to select the top-50 most similar passages for each hypothesis of interest. We then use the fine-tuned cross-encoder to select the most similar passage. For instance, to answer a multiple-choice question, where $\mathcal{D}$ is the question and $H_1, ..., H_m$ are the possible answers, we use the fine-tuned cross-encoder to select for each candidate H_i the most relevant passage $\mathcal{C}_{H_i}$. We predict the answer candidate H_i for which the similarity between $\mathcal{D}$ and $\mathcal{C}_{H_i}$, as estimated by the fine-tuned cross-encoder, is maximal. In cases where a hypothesis H_i does not appear in the corpus at all, we simply set $\mathcal{C}_{H_i} = H_i$, i.e. we compute the similarity between $\mathcal{D}$ and H_i instead.

5.5 Experimental Results

In this section, we present our experimental analysis. Apart from assessing the overall effectiveness of our proposed strategy, we are interested in the following research

[1] https://patient.info/

questions:

- Is the use of an unsupervised sentence embedding model (i.e. TSDAE) viable as the primary retrieval strategy? Can such an approach overcome the limitations of BM25 for identifying potentially relevant cases?
- Can the use of a cross-encoder that is pre-trained on an out-of-domain task (e.g. STS-B) lead to meaningful improvements?
- How sensitive is the model to the value of k and to the chosen pre-training task for the cross-encoder? Are there any differences across different biomedical LMs and datasets?

5.5.1 Evaluation Datasets

We evaluate our method on the MedQA, DisKnE and HeadQA datasets. We include HeadQA dataset because it allows us to explore to what extent the proposed methodology can be effective in a broader setting than for interpreting patient case descriptions.

5.5.2 Corpora

The choice of the external corpus, from which the text passages are retrieved, is an important factor for the effectiveness of our method. Given the aims of this work, we focus on corpora that contain patient case descriptions. We have, in particular, used the following two corpora, both of which are widely used in biomedical NLP.

5.5.2.1 WikiMed and PubMedDS (Wiki-PubMed) [216]

We split the documents in this corpora into text passages of up to 250 words, respecting sentence boundaries. This resulted in a total of 14,582,089 passages. While this corpus covers a wide variety of documents, many PubMed abstracts correspond to patient case descriptions (i.e. the abstracts of medical case reports). This corpus thus allows us to analyse to what extent our method is able to identify patient case descriptions and to what extent it is able to exploit generic descriptions.

5.5.2.2 MIMIC-III discharge summaries [86]

To split the discharge summaries into text passages, we first split them according to the section headers and then split the resulting sections into passages of up to 250 words. This allows us to go beyond the sentence level, while keeping in mind that the concatenation of the question and a retrieved passage can be at most 512 tokens, given the limitations of the considered transformer-based language models. We obtained a total of 3,623,209 passages from 59,652 discharge summaries, although it should be noted that many of these passages are short and uninformative (e.g. the passage obtained from the admission date section). MIMIC-III has the advantage that it consists entirely of patient case descriptions. The main drawbacks are that summaries are often noisy (e.g. not always containing well-structured sentences) and that they are limited to descriptions of critical care patients. Given this latter point, MIMIC-III is particularly suitable for DisKnE, whose patient case descriptions are also taken from the MIMIC-III corpus. This allows us to experiment with a setting where the corpus contains patient case descriptions that are written in a similar style as the target description. Note, however, that the patient case descriptions from DisKnE themselves are never retrieved by our method, as the corresponding hypotheses are not mentioned in the original notes.

5.5.3 Pre-trained Language Models

We experiment with four pre-trained LMs to initialize the cross-encoder: the cased version of BERT$_{\text{base}}$ [42]; the version of ClinicalBERT [8] that was initialized from

BioBERT [104] and further pre-trained on MIMIC-III; the cased version of SciBERT [24]; and the released version of PubMedBERT [61] .

5.5.4 Baselines

We consider the following baselines.

5.5.4.1 Standard Fine-tuning (FT)

We fine-tune a pre-trained language model to predict whether a given hypothesis can be entailed from a patient case description, as in standard NLI models. Specifically, we concatenate the patient case description and the hypothesis, separated by a [SEP] token, and fine-tune this model using binary cross-entropy. We refer to this model as *BERT-FT* in the case BERT is used, and similar for the other LMs.

5.5.4.2 Definitions

We use QuickUMLS [196] to identify the UMLS CUI codes of the medical concepts mentioned in the hypothesis. We then use these CUI codes to retrieve the definition(s) of the corresponding concepts from UMLS. These definitions, if they exist, are concatenated to the hypothesis. We then fine-tune a language model on the augmented input. This follows the strategy proposed by [191] for improving LSTM-based models. We refer to this strategy as *BERT-Def*, and similar for the other LMs.

5.5.4.3 Unsupervised Retrieval

Finally, we also report results for unsupervised retrieval models. In this case, we simply compute the similarity degree between the patient case description and the most similar passage, for each hypothesis. We test this strategy with two retrieval models: (i) BM25 and (ii) dense retrieval with the TSDAE embeddings that are also used for our main model.

5.5.5 Evaluation Metrics

For MedQA and HeadQA, we solve the standard multiple-choice QA task as explained in Section 5.4.3, reporting results in terms of accuracy. In addition, we have included experiments where MedQA and HeadQA are treated as ranking tasks. We then rank all (question, answer candidate) pairs, across all questions and answer candidates, and report the results in terms of average precision (AP). This essentially allows us to assess to what extent our model is able to recognise valid hypotheses in isolation, instead of selecting the most plausible answer candidate among a small set of choices. We similarly treat DisKnE as a ranking task, rather than a binary classification task. In this case, we obtain the average precision score for each training-test split (i.e. for each of the considered diseases). The AP scores for each split are then averaged to get the overall Mean Average Precision (MAP).

5.5.6 Training Details

Across all datasets and language models, we use the same settings and hyper-parameters. For the baselines, and when pre-training and fine-tuning the cross-encoders, we set the batch size to 8, the number of epochs to 4 and the learning rate set to 2e-5. The cross-encoders are pre-trained and fine-tuned using binary cross-entropy (where similarity scores are normalised between 0 and 1 for STS-B). We use the standard training/validation/test splits, with the exception of HeadQA, where we have removed all questions involving images.

	MedQA		HeadQA		DisKnE
	AP	Acc	AP	Acc	MAP
BERT-FT	26.8	27.8	28.1	28.8	57.0
ClinicalBERT-FT	27.7	29.1	28.5	29.3	67.5
SciBERT-FT	28.6	29.2	29.5	32.8	69.2
PubMedBERT-FT	**32.8**	**35.5**	**35.4**	**39.5**	**69.7**
BERT-Def	27.8	27.7	27.9	30.4	50.5
ClinicalBERT-Def	28.2	29.5	27.8	30.2	59.3
SciBERT-Def	29.7	30.8	30.3	34.5	56.2
PubMedBERT-Def	30.1	32.9	35.2	38.3	65.2
TSDAE Wiki-PubMed	26.2	29.3	26.7	31.1	27.8
TSDAE MIMIC-III	25.0	25.1	26.0	28.3	32.7
BM25 Wiki-PubMed	25.3	26.8	25.6	25.9	22.3
BM25 MIMIC-III	25.0	23.8	25.0	23.8	22.5

Table 5.2: Baselines results for all datasets. We report DisKnE in terms of Mean Average Precision (MAP), MedQA and HeadQA in terms of Average Precision (AP) and Accuracy (Acc). The best results are shown in bold.

5.5.7 Results

The experimental results are summarized in Table 5.3 for MedQA, Table 5.4 for DisKnE and Table 5.5 for HeadQA. We write CE-k for our method, where the cross-encoder is fine-tuned using k passages per $(\mathcal{D}, H)$ pair. The baseline results are reported in Table 5.2.

For MedQA (Table 5.3), the results for Wiki-PubMed (abbreviated as Wiki-PM) clearly outperform those for MIMIC-III (abbreviated as MIM-III), which is as expected given the aforementioned limitations of MIMIC-III. Focusing on the results for Wiki-PubMed, we can see that for each of the language models, the results in Table 5.3 consistently outperform the baseline results (for these language models) in Table 5.2, across all choices of k and each of the three pre-training tasks. The results also clearly outper-

		STS-B				**RQE**				**HealthQA**			
		MIM-III		**Wiki-PM**		**MIM-III**		**Wiki-PM**		**MIM-III**		**Wiki-PM**	
		AP	**Acc**	**AP**	**Acc**	**AP**	**Acc**	**AP**	**Acc**	**AP**	**Acc**	**AP**	**Acc**
BERT	CE-1	26.7	26.8	29.9	32.4	25.3	24.1	28.5	31.5	25.8	25.3	28.9	32.2
	CE-5	25.1	25.0	**31.7**	**35.5**	25.0	26.3	31.5	33.6	28.8	27.8	27.9	29.8
	CE-10	25.3	23.4	30.5	34.0	25.1	26.9	30.8	32.9	25.5	24.6	25.2	26.7
ClinicalBERT	CE-1	25.9	25.3	33.2	35.4	27.6	28.2	30.4	32.2	25.6	25.6	31.3	34.0
	CE-5	27.8	28.8	33.4	35.4	27.9	29.4	**35.1**	**38.0**	24.9	24.7	32.9	35.5
	CE-10	25.3	26.8	31.5	33.6	26.7	27.0	31.5	36.2	25.7	23.4	32.1	37.4
SciBERT	CE-1	25.2	24.3	32.4	34.5	27.2	28.9	32.7	33.8	25.2	24.7	32.3	34.5
	CE-5	25.8	25.7	30.5	35.1	27.6	28.7	31.0	33.8	28.1	29.2	**33.0**	**37.6**
	CE-10	24.7	24.0	30.1	34.5	25.4	25.3	31.2	32.7	23.9	22.2	32.3	35.3
PubMedBERT	CE-1	30.5	32.3	**36.0**	**39.3**	24.9	26.6	32.8	35.8	27.4	28.6	34.0	**39.3**
	CE-5	29.1	30.5	33.1	35.8	26.1	26.7	31.6	37.2	26.8	26.6	34.4	36.4
	CE-10	31.2	34.8	33.8	37.7	30.8	32.6	32.8	38.0	29.5	31.3	33.4	37.3

Table 5.3: Results for MedQA in terms of Average Precision (AP) and Accuracy (Acc). The best results for each language model are shown in bold.

form the unsupervised retrieval baselines. Comparing the different language models, PubMedBERT achieves the best results. With regards to the choice of k, we find that $k = 5$ is generally the best choice, with the exception of PubMedBERT where $k = 1$ performs much better. This appears to be related to the fact that PubMedBERT itself performs better than the other LMs. In general, larger values of k leads to more, but noisier training data. Since PubMedBERT is better at selecting the most relevant paragraphs, even when using the pre-trained encoder, this problem of training data becoming noisier for larger values of k is more pronounced.

Regarding the pre-training tasks, STS-B and HealthQA lead to the best results in most cases, with the exception of ClinicalBERT. To the best of our knowledge, the best reported results in the literature at the time of the publication for MedQA are those from Meng et al. [131], where an accuracy of 38.02 was obtained for their best-performing configuration, using a large biomedical knowledge graph to augment the PubMedBERT model. This contrasts to an accuracy of 39.3 for the best-performing model in

		STS-B		RQE		HealthQA	
		MIM	**WPM**	**MIM**	**WPM**	**MIM**	**WPM**
BERT	CE-1	47.5	36.6	46.6	34.1	45.6	37.4
	CE-5	66.0	48.7	65.4	43.2	59.5	44.1
	CE-10	67.1	55.4	**70.4**	54.9	61.7	48.0
ClinicalBERT	CE-1	51.4	50.9	53.4	50.7	52.0	53.4
	CE-5	63.9	59.7	66.0	57.4	65.4	53.9
	CE-10	62.1	59.7	67.7	63.7	**67.8**	58.5
SciBERT	CE-1	60.7	45.4	54.4	46.8	58.0	50.4
	CE-5	69.6	59.6	65.4	56.4	65.6	54.2
	CE-10	**73.2**	65.1	67.3	59.5	72.8	61.9
PubMedBERT	CE-1	63.3	60.0	63.6	54.1	57.4	52.3
	CE-5	**71.6**	64.6	69.1	59.0	64.6	58.3
	CE-10	69.0	67.1	70.3	61.7	67.6	63.7

Table 5.4: Results for DisKnE in terms of Mean Average Precision (MAP). The best results for each language model are shown in bold.

Table 5.3. Since then better results have been reported as well in [245, 244]. We also report improved results in Chapter 6 [7].

For HeadQA (Table 5.5), as expected we again find that Wiki-PubMed leads to much better results than the MIMIC-III corpus. Moreover, we can again see that the use of the cross-encoder consistently leads to better results than when using the baseline fine-tuned language model, across all values of k and all pre-training tasks. The best results are again obtained with PubMedBERT. However, here we see that RQE is the most suitable pre-training task for most configurations. This can be explained by the observation that HeadQA primarily consists of factual questions, which clearly makes RQE the most closely related pre-training task. Overall, the choice of $k = 5$ generally performs best. The improvements for HeadQA are remarkable, since many of the questions in this dataset are not about patient case descriptions. To explore this further, we manually split the test set into those questions which are about patient case descrip-

		STS-B				RQE				HealthQA			
		MIM-III		Wiki-PM		MIM-III		Wiki-PM		MIM-III		Wiki-PM	
		AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc
BERT	CE-1	27.2	28.2	32.6	33.4	27.4	29.3	30.0	32.6	27.2	29.2	32.3	33.3
	CE-5	27.4	29.0	34.2	36.1	27.6	30.2	**34.9**	**38.0**	26.4	28.2	32.4	34.8
	CE-10	26.8	28.2	33.7	36.6	27.1	29.1	33.5	37.1	26.8	28.6	31.3	34.4
ClinicalBERT	CE-1	28.7	29.4	**33.8**	34.8	26.5	27.0	31.3	32.8	27.8	30.8	33.0	33.1
	CE-5	27.3	29.6	**33.8**	36.4	27.4	28.1	32.8	**36.9**	27.9	29.6	33.7	35.7
	CE-10	27.4	30.5	33.6	36.7	27.8	29.6	32.5	35.9	27.1	29.3	32.0	35.8
SciBERT	CE-1	29.9	32.2	33.9	35.7	30.3	34.2	32.8	33.0	29.0	32.9	34.4	35.0
	CE-5	29.2	29.5	**35.3**	**39.8**	28.8	32.1	33.2	37.0	28.5	31.9	33.3	35.8
	CE-10	29.3	32.4	33.1	35.6	28.5	33.2	33.4	37.6	28.8	31.5	32.4	34.9
PubMedBERT	CE-1	34.5	37.1	38.2	39.3	33.9	37.9	**38.8**	41.2	33.0	36.9	36.6	40.3
	CE-5	32.7	36.4	38.4	41.2	33.7	37.0	38.7	**42.3**	32.1	33.0	35.9	39.8
	CE-10	33.9	37.9	37.4	40.3	33.4	38.4	37.5	40.0	32.4	37.2	35.1	40.5

Table 5.5: Results for HeadQA in terms of Average Precision (AP) and Accuracy (Acc). The best results for each language model are shown in bold.

	Patient case descriptions		Other questions	
	AP	Acc	AP	Acc
SciBERT-FT	27.9	27.7	31.3	34.4
SciBERT-CE	29.6	32.4	33.8	38.2
PubMedBERT-FT	29.6	34.7	37.7	40.4
PubMedBERT-CE	32.3	35.6	38.5	41.3

Table 5.6: Analysis of HeadQA results, where test questions were split depending on whether or not they are about patient case descriptions. Results are reported in terms of average precision and accuracy.

tions (216 in total) and those which are not (2458 in total). Table 5.6 shows the results obtained for these two sets of questions, for the SciBERT-FT and PubMedBERT-FT baselines, as well as our proposed model, where we used the RQE pre-training task and $k = 5$. As we can see, our model improves the results on both sets of questions. This suggests that our proposed strategy could be beneficial for biomedical QA more generally. However, on its own, our approach is not sufficient to obtain state-of-the-art results, which rely on methods that are specifically designed to enable the kind of multi-hop reasoning that is often needed for this dataset [117].

For DisKnE (Table 5.4), as expected, the best results are obtained when MIMIC-III is used as the corpus. For this choice, our method consistently outperforms the baselines for all language models, provided that $k \geq 5$. On average, the optimal value of k is larger than what we found for MedQA and HeadQA. This suggests that identifying the most relevant passages is more challenging for this dataset.

Comparing the baseline results in Table 5.2, we can clearly see the limited usefulness of augmenting the inputs with definitions of medical concepts. For DisKnE, adding these definitions actually has a detrimental effect. For MedQA and HeadQA, the unsupervised retrieval baselines are remarkably competitive compared to the fine-tuned language models. However, in the case of DisKnE these unsupervised models substantially underperform. We can also see that TSDAE consistently outperforms BM25. This was expected, given the fact that comparing patient case descriptions intuitively requires more than surface-level matching.

Given the required computational cost, it was not feasible to obtain results for multiple runs for all configurations. We report in table 5.7 the averaged result over three runs for MedQA's best configuration (i.e. top-1 with PubMedBERT and STS-B as the pre-training task).

	AP	**Acc**
PubMedBERT-FT	33.2 $_{\pm 0.36}$	35.9 $_{\pm 0.46}$
PubMedBERT-CE	35.9 $_{\pm 0.37}$	39 $_{\pm 0.24}$

Table 5.7: Results for the PubMedBERT model in terms of Average Precision (AP) and Accuracy (Acc) for MedQA best configuration averaged over three runs. Standard deviations (over the three runs) are also reported.

	MedQA		HeadQA		DisKnE
	AP	**Acc**	**AP**	**Acc**	**MAP**
Pretrained CE	30.6	33.0	32.9	38.0	41.4
TSDAE-Selected	35.0	36.7	33.0	36.9	70.0
Full model	36.0	39.3	38.7	42.3	73.2

Table 5.8: Ablation analysis for all datasets. We report results for DisKnE in terms of Mean Average Precision (MAP), MedQA and HeadQA in terms of Average Precision (AP) and Accuracy (Acc).

5.5.8 Analysis

5.5.8.1 Ablation Results

In Table 5.8, we show results for the following simplified versions of our model.

- *Pretrained CE:* Rather than fine-tuning a cross-encoder using our distant supervision strategy, we simply use the pre-trained cross-encoder to re-rank the top-50 passages selected by TSDAE. Note that this variant of our method does not rely on the training data at all. This Pretrained CE model is illustrated in Figure 5.4
- *TSDAE-Selected:* When creating the distantly supervised training set for fine-tuning the cross-encoder, we simply choose the k highest ranked passages according to their TSDAE-embeddings, thus omitting the stage where we re-rank the candidate passages using a pre-trained cross-encoder. TSDAE-Selected is illustrated in Figure 5.5

In all cases, we used the best configurations from the main experiments (i.e. the optimal value of k and pre-training task). For MedQA and HeadQA we used Wiki-PubMed as the corpus while for DisKnE we used MIMIC-III. As can be seen in Table 5.8, the *Pretrained CE* model outperforms the unsupervised baseline retrieval models from Table 5.2. In fact, for MedQA and HeadQA, the results of this unsupervised model

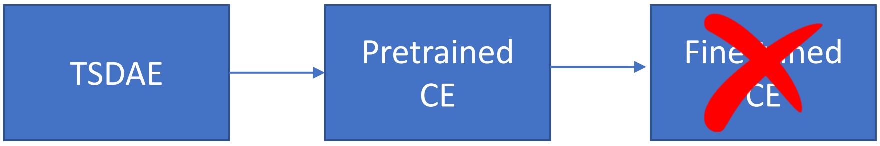

Figure 5.4: Illustration of Pretrained CE.

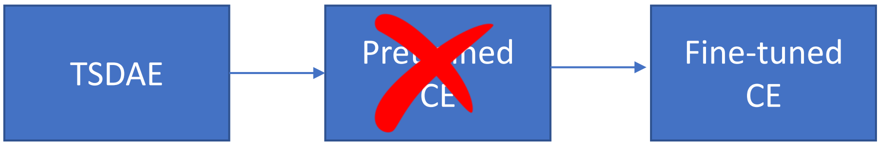

Figure 5.5: Illustration of TSDAE-Selected.

are almost in line with those of the fine-tuned PubMedBERT model. This clearly shows the usefulness of the pre-trained cross-encoder, even when it cannot be fine-tuned on task-specific data. This usefulness can furthermore be seen in the performance of *TSDAE-Selected*. While this variant performs quite well, it clearly underperforms the full model, showing the importance of the cross-encoder based re-ranking step.

Question : A 38-year-old woman comes to the physician because of difficulty falling asleep for the past 2 months. She wakes up frequently during the night and gets up earlier than desired. She experiences discomfort in her legs when lying down at night and feels the urge to move her legs. The discomfort resolves when she gets up and walks around or moves her legs. She has tried an over-the-counter sleep aid that contains diphenhydramine, which worsened her symptoms. She exercises regularly and eats a well-balanced diet. She admits that she has been under a lot of stress lately. Her brother has similar symptoms. The patient appears anxious. Physical examination shows no abnormalities. A complete blood count and iron studies are within the reference range. Which of the following is the most appropriate pharmacotherapy for this patient's symptoms?

Answer candidate: Pramipexole

Retrieved Passage: Medications used include levodopa or a dopamine agonist such as pramipexole. RLS affects an estimated 2. 5–15% of the American population. Females are more commonly affected than males and it becomes more common with age. RLS sensations range from pain or an aching in the muscles, to "an itch you can't scratch", a "buzzing sensation", an unpleasant "tickle that won't stop", a "crawling" feeling, or limbs jerking while awake. The sensations typically begin or intensify during quiet wakefulness, such as when relaxing, reading, studying, or trying to sleep. It is a "spectrum" disease with some people experiencing only a minor annoyance and others having major disruption of sleep and impairments in quality of life. The sensations—and the need to move—may return immediately after ceasing movement or at a later time. RLS may start at any age, including childhood, and is a progressive disease for some, while the symptoms may remit in others. In a survey among members of the Restless Legs Syndrome Foundation, it was found that up to 45% of patients had their first symptoms before the age of 20 years. - "An urge to move, usually due to uncomfortable sensations that occur primarily in the legs, but occasionally in the arms or elsewhere".

Table 5.9: Example of a correctly answered question from the MedQA test set in which the retrieved passage is not a patient description.

5.5.8.2 Qualitative Analysis

We manually analysed the retrieved passages for MedQA and HeadQA with Wiki-PubMed (given that MIMIC-III notes and DisKnE examples cannot be shared without access to a MedNLI license). Our main findings can be summarized as follows. First, we found that in many cases, the retrieved passages were indeed patient case descriptions. This is somewhat surprising, given that only a small fragment of Wiki-PubMed consists of patient case descriptions (which appear as abstracts of published medical case reports). Nonetheless, there are also many cases where the retrieved text passage was a generic description (e.g. from Wikipedia). Often, however, such passages can still be successfully exploited by the cross-encoder. An example illustrating such a case is presented in Table 5.9. In this example, the retrieved passage intuitively acts as a generic description of how patients experience Restless Leg Syndrome (RLS). While not referring to a particular case, such descriptions can intuitively act as prototypes of actual patient case descriptions. Table 5.10 shows other examples where the retrieved passages are still relevant, despite not corresponding to specific patient case descriptions. For instance, some of the retrieved passages explicitly list some of the symptoms described in the questions, such as in question #2. Other specific examples of questions and retrieved text passages, covering different situations, are also provided. Table 5.11 shows examples where the retrieved passage corresponds to a similar case. In Table 5.12 we show examples where the correct answer was predicted with a high confidence score, despite the fact that the usefulness of the retrieved text passage is unclear. In Table 5.13, we show an example of a question where the correct answer was not predicted correctly because the retrieved passages for several of the other answer candidates were also highly similar. Table 5.14 shows examples where the question does not correspond to a patient case description, yet where the retrieved passages are still meaningful. This shows that the effectiveness of the proposed method extends beyond retrieving similar cases.

Question #1	A 46-year-old man is brought to the emergency department for evaluation of altered mental status. He was found on the floor in front of his apartment. He is somnolent but responsive when aroused. His pulse is 64/min, respiratory rate is 15/min, and blood pressure is 120/75 mm Hg. On physical examination, an alcoholic smell and slurred speech are noted. Neurological exam shows diminished deep tendon reflexes bilaterally and an ataxic gait. His pupils are normal. Blood alcohol concentration is 0.04%. An ECG shows no abnormalities. Which of the following is the most likely cause of this patient's symptoms?
Answer	*Benzodiazepine intoxication*
Retrieved Passage	Hyporeflexia Hyporeflexia refers to below normal or absent reflexes (areflexia). It can be detected through the use of a reflex hammer. It is the opposite of hyperreflexia. Hyporeflexia is generally associated with a lower motor neuron deficit (at the alpha motor neurons from spinal cord to muscle), whereas hyperreflexia is often attributed to upper motor neuron lesions (along the long, motor tracts from the brain). The upper motor neurons are thought to inhibit the reflex arc, which is formed by sensory neurons from intrafusal fibers of muscles, lower motor neurons (including alpha and gamma motor fibers) and appurtenant interneurons. Therefore, damage to lower motor neurons will subsequently result in hyporeflexia and/or areflexia. Note that, in spinal shock, which is commonly seen in the transection of the spinal cord (Spinal cord injury), areflexia can transiently occur below the level of the lesion and can, after some time, become hyperreflexic. Furthermore, cases of severe muscle atrophy or destruction could render the muscle too weak to show any reflex and should not be confused with a neuronal cause. Hyporeflexia may have other causes, including hypothyroidism, electrolyte imbalance (e.g. excess magnesium), drug induced (e.g. the symptoms of benzodiazepine intoxication include confusion, slurred speech, ataxia, drowsiness, dyspnea, and hyporeflexia).
Question #2	A 31-year-old woman comes to the physician because of a 5-month history of intermittent flank pain. Over the past 2 years, she has had five urinary tract infections. Her blood pressure is 150/88 mm Hg. Physical examination shows bilateral, nontender upper abdominal masses. Serum studies show a urea nitrogen concentration of 29 mg/dL and a creatinine concentration of 1.4 mg/dL. Renal ultrasonography shows bilaterally enlarged kidneys with multiple parenchymal anechoic masses. Which of the following is the most likely diagnosis?
Answer	*Autosomal dominant polycystic kidney disease*
Retrieved Passage	Autosomal dominant polycystic kidney disease: presentation, complications, and prognosis. Fifty-three symptomatic adults with autosomal dominant polycystic kidney disease were studied retrospectively for a mean follow-up of 12 years (range 10 months to 33 years). Diagnosis was confirmed by either x-ray, ultrasound, laparotomy, or autopsy. Commonest presenting clinical findings were flank pain (30%), hypertension (21%), symptomatic urinary tract infection (UTI) (19%), gross hematuria (19%), and palpable masses (15%). A total of nine patients (17%) progressed to end-stage renal disease. Change in renal function measured using the reciprocal of plasma creatinine plotted against time was linear for each individual patient with a maximum functional decline of 0. 7 mg/dL/yr (slope = -0. 07). Past the age of sixty renal failure was uncommon. Easily controlled hypertension developed in 64% attended by mild retinopathy. UTIs were common (53%), often recurrent (61%), precipitated by instrumentation in 6 of 14 patients (43%), leading to death in two (33%). Renal calculi were extremely common (34%) and had no defined metabolic cause. The presence of hematuria (64%), gross or microscopic, bore no relationship to the decline in renal function. Pregnancy was normal in these patients with no increase in fetal or maternal morbidity or mortality. We conclude the following: Renal functional deterioration is linear, less than previously reported, and bears no relationship to hematuria.

Table 5.10: Examples of correctly answered questions from MedQA test set in which the retrieved passages informative in general.

Question #1	A gunshot victim is brought to the Emergency Department and appears to be in shock. You note a penetrating wound at the level of L3. Assuming the bullet remained at this level, which vascular structure might have been injured?
Answer	*Inferior vena cava*
Retrieved Passage	Right ventricular gunshot wound with retrograde embolization. There have been numerous reports concerning gunshot wounds to the heart over the years. Many reports discuss bullets that have embolized and have migrated antegrade. However, there has never been a case reported on the retrograde embolization of a bullet from the right ventricle into the inferior vena cava. This case report involves a 15-year-old boy who was accidentally shot in the chest. The bullet entered at the mid-manubrial area, and penetrated the anterior wall of the right ventricle causing a tamponade. A chest x-ray film confirmed a bullet in the right ventricle. The patient was stabilized in the emergency department, and taken to the operating room for an emergent mediastinal exploration with evacuation of pericardial tamponade and repair of the right ventricle. After the tamponade was relieved, a Trans-Esophageal Echocardiogram was performed to locate the bullet, which could not be found in the ventricle. Chest and abdominal radiography were performed to locate the bullet. X-ray films confirmed that the bullet had migrated retrograde down into the inferior vena cava. Interventional radiology and vascular surgery departments were consulted. The consensus was to snag the bullet under fluoroscopic guidance, and pull it down into the right femoral vein for easy retrieval.
Question #2	A 32-year-old woman comes to the physician because of fatigue and joint pain for the past 4 months. Examination shows erythema with scaling on both cheeks that spares the nasolabial folds and two 1-cm ulcers in the oral cavity. Further evaluation of this patient is most likely to show which of the following findings?
Answer	*Decreased lymphocyte count*
Retrieved Passage	[A case developing tetraplegia due to systemic lupus erythematosus which was remitted by a steroid]. The case reported here was a 58-year-old woman who was diagnosed as having systemic lupus erythematosus (SLE) in 1985 because she had erythema in the cheeks arthritis, a hematological abnormality (decreased white blood cell count), an immunological abnormality (LE-positive cells), and a positive result of anti-nuclear antibody test. Although the patient was once remitted by treatment with prednisolone (PSL) at 60 mg/day, and continuously received PSL at a maintenance dose of 2. 5 mg/day, she was admitted in June 1996 by our hospital with chief complaints of fever and decreased muscular strength in the four extremities. At admission, she had symmetrical tetraplegia, which was peripherally predominant and severer in the lower extremities, and hypoesthesia accompanied by numbness. She was negative for anti-phospholipid antibody and showed no abnormality in cerebrospinal fluid examination. No lesions responsible for tetraplegia were detected at brain MRI, spinal MRI, or myelography. Because fever, multiple arthralgia, an increased erythrocyte sedimentation rate, a decreased lymphocyte count, hypocomplementemia, and a high immune complex level indicated the active stage of SLE (recurrence), she was given PSL at dose increased to 60 mg/day. After about 2 months, SLE was remitted and her tetraplegia and hypoesthesia was gradually improved thereafter.

Table 5.11: Examples of correctly answered questions from MedQA test set in which the retrieved passages represent similar cases.

Question #1	A 13-year-old boy is brought to the physician because of progressive left leg pain for 2 months, which has started to interfere with his sleep. His mother has been giving him ibuprofen at night for “growing pains,” but his symptoms have not improved. One week before the pain started, the patient was hit in the thigh by a baseball, which caused his leg to become red and swollen for several days. Vital signs are within normal limits. Examination shows marked tenderness along the left mid-femur. His gait is normal. Laboratory studies show a leukocyte count of 21,000/mm3 and an ESR of 68 mm/h. An x-ray of the left lower extremity shows multiple lytic lesions in the middle third of the femur, and the surrounding cortex is covered by several layers of new bone. A biopsy of the left femur shows small round blue cells. Which of the following is the most likely diagnosis?
Answer	*Ewing sarcoma*
Retrieved Passage	Primary undifferentiated small round cell sarcoma of the deep abdominal wall with a novel variant of t(10;19) CIC-DUX4 gene fusion. We experienced a 38-year-old Japanese male with t(10;19) CIC-DUX4 -positive undifferentiated small round cell sarcoma in the deep abdominal wall. Three months before his first visit to our hospital, he noticed a mass in his right abdominal wall. Computed tomography on admission revealed a solid abdominal tumor 70×53mm in size and multiple small tumors in both lungs. The biopsy of the abdominal tumor revealed undifferentiated small round cell sarcoma, suggestive of Ewing sarcoma. Under the clinical diagnosis of Ewing-like sarcoma of the abdominal wall with multiple lung metastases, several cycles of ICE (ifosfamide, carboplatin and etoposide) therapy were performed. After the chemotherapy, the lung metastases disappeared, while the primary lesion rapidly grew. Additional VDC (vincristine, doxorubicin and cyclophosphamide) therapy was carried out without apparent effect. Although the surgical removal of the primary lesion was done, peritoneal dissemination and a huge metastatic liver tumor appeared thereafter. The patient died of disease progression two months after the surgery. The total clinical course was approximately one year, showing that the tumor was extremely aggressive. The tumor cells of the surgical specimen were positive for CD99, WT1, calretinin, INI1, ERG and Fli1 by immunohistochemistry.
Question #2	A 22-year-old man with sickle cell disease is brought to the emergency room for acute onset facial asymmetry and severe pain. He was in school when his teacher noted a drooping of his left face. His temperature is 99.9°F (37.7°C), blood pressure is 122/89 mmHg, pulse is 110/min, respirations are 19/min, and oxygen saturation is 98% on room air. Physical exam is notable for facial asymmetry and 4/5 strength in the patient’s upper and lower extremity. A CT scan of the head does not demonstrate an intracranial bleed. Which of the following is the most appropriate treatment for this patient?
Answer	*Exchange transfusion*
Retrieved Passage	Continuous monitoring of intracranial pressure in Reye’s syndrome–5 years experience. Monitoring of intracranial pressure (ICP) and efforts to keep the ICP below the critical level are vital in the treatment of Reye’s syndrome. Continuous monitoring of ICP was carried out in 21 cases of Reye’s syndrome who were at or beyond stage III at the time of admission to the Veterans General Hospital, between January 1981 and August 1986. Seventeen had ICP ranging from 15 mmHg to 67 mmHg. Three patients died, 1 in stage V with an ICP of 67 mmHg received a craniectomy, and 2 others were in stage IV with ICP’s of 66 mmHg and 25 mmHg, respectively. The fatality rate was 14% (3/21). Among 18 patients, 5 had moderate psychomotor retardation (PMR), 4 had severe PMR and 2 had mild PMR. The remaining 7 patients survived without sequelae. Blood exchange transfusion could further reduce ICP and seemed to improve neurologic outcome. Blood ammonia higher than 400 micrograms% is indicative of a bad prognosis. Hyperventilation was the most rapid and effective means of reducing moderate degrees of increased ICP. During intensive supportive care, we also found that coughing, endotracheal intubation, seizures, asynchronous respiration to an artificial respirator, suction of the airway and any painful stimulation caused further increases in ICP and worsened the situation.

Table 5.12: Examples of correctly answered questions from MedQA test set in which unclear how the retrieved passages are related.

Question	A 28-year-old woman presents following a suicide attempt 2 days ago. She says that her attempt was a result of a fight with her boyfriend and that she slit her wrists in an attempt to keep him from breaking up with her. In the past, she has had many turbulent relationships, both romantic and in her family life. Her family members describe her as being very impulsive and frequently acting to manipulate people's feelings. Since she was admitted to the hospital, she has spit at several staff members and alternated between sobbing and anger. She has no significant past medical history. The patient denies any history of smoking, alcohol use, or recreational drug use. Which of the following is the most likely diagnosis in this patient?
Answer Candidate A	*Histrionic personality disorder*
Retrieved Passage	Traumatic bonding can occur between the abuser and victim as the result of ongoing cycles of abuse in which the intermittent reinforcement of reward and punishment creates powerful emotional bonds that are resistant to change and a climate of fear. An attempt may be made to normalise, legitimise, rationalise, deny, or minimise the abusive behaviour, or blame the victim for it. Isolation, gaslighting, mind games, lying, disinformation, propaganda, destabilisation, brainwashing and divide and rule are other strategies that are often used. The victim may be plied with alcohol or drugs or deprived of sleep to help disorientate them. Certain personality types feel particularly compelled to control other people. In the study of personality psychology, certain personality disorders display characteristics involving the need to gain compliance or control over others: - Those with antisocial personality disorder tend to display glibness, giving them a grandiose sense of self-worth. Due to their callous and unemotional traits, they are well suited to con and/or manipulate others into complying with their wishes. - Those with borderline personality disorder tend to display black-and-white thinking and no sense of self-worth. - Those with histrionic personality disorder need to be the center of attention; and in turn, draw people in so they may use (and eventually dispose of) their relationship.
Answer Candidate B	***Borderline personality disorder***
Retrieved Passage	These behaviors include greater expressed negativity (e.g. criticism, blaming, demanding, and disengagement) toward romantic partners, and negative feedback seeking. Excessive reassurance seeking is also a vulnerability factor for depression. However, Marroquin (2011) proposes adaptive interpersonal emotion regulation as a mechanism of the positive effects of social support. Social interaction that diverts attention away from self-referential negative thinking and promotes cognitive reappraisal may help to alleviate depression. According to the biosocial model, individuals with borderline personality disorder develop intense emotional expression in part because they have been reinforced throughout development. For instance, a teenager with heightened emotional sensitivity is not taken seriously by her family until she threatens a suicide attempt. If her family responds with attention to extreme emotional expressions, she will learn to continue to express emotions in this way. Venting is another interpersonal emotion regulation strategy that is associated with personality disorder symptoms. Certain types of psychotherapy target interpersonal factors to improve well-being. Dialectical behavioral therapy, originally developed for individuals with borderline personality disorder, teaches clients interpersonal effectiveness, which includes a variety of skills for communicating emotions in a clear and socially acceptable manner. Assertiveness training is a behavioral intervention that teaches verbal and non-verbal assertiveness skills to inhibit anxiety.
Answer Candidate C	*Dependent personality disorder*
Retrieved Passage	Passive-Aggressive Personality Disorder was expanded somewhat as an official diagnosis in the DSM-III-R but then relegated to the appendix of DSM-IV, tentatively renamed 'Passive-Aggressive (Negativistic) Personality Disorder'. Millon devised a set of widely acknowledged subtypes for each of the DSM personality disorders: - Sadistic (psychopathic) personality disorder subtypes - Self-defeating (masochistic) personality disorder subtypes - Schizotypal personality disorder subtypes - Schizoid personality disorder subtypes - Paranoid personality disorder subtypes - Antisocial (sociopathic) personality disorder subtypes - Borderline personality disorder subtypes - Histrionic personality disorder subtypes - Narcissistic personality disorder subtypes - Dependent personality disorder subtypes - Obsessive-compulsive personality disorder subtypes - Avoidant personality disorder subtypes - Passive-aggressive (negativistic) personality disorder subtypes - Depressive personality disorder subtypes - Exuberant/Hypomanic (turbulent) personality disorder subtypes - Decompensated Personality Disorder - 1969: "Modern Psychopathology: A Biosocial Approach to Maladaptive Learning and Functioning", Saunders - 1996: (with Roger D. Davis)...
Answer Candidate D	*Narcissistic personality disorder*
Retrieved Passage	"Sometimes he manipulates even me", admitted one psychiatrist. In the end, Lewis agreed with the majority: "I always tell my graduate students that if they can find me a real, true psychopath, I'll buy them dinner", she told Nelson. "I never thought they existed . . . but I think Ted may have been one, a true psychopath, without any remorse or empathy at all". Narcissistic personality disorder (NPD) has been proposed as an alternative diagnosis in at least one subsequent retrospective analysis. On the afternoon before he was executed, Bundy granted an interview to James Dobson, a psychologist and founder of the Christian evangelical organization Focus on the Family. He used the opportunity to make new claims about violence in the media and the pornographic "roots" of his crimes. "It happened in stages, gradually", he said. "My experience with . . . pornography that deals on a violent level with sexuality, is once you become addicted to it . . . I would keep looking for more potent, more explicit, more graphic kinds of material. Until you reach a point where the pornography only goes so far . . . where you begin to wonder if maybe actually doing it would give that which is beyond just reading it or looking at it".

Table 5.13: Example of incorrectly answered question from MedQA test set in which answer candidates are different types of a particular disorder. The correct answer is in bold and the predicted underlined.

Question #1	The fibrocartilage is located in:
Answer	*Intervertebral discs*
Retrieved Passage	Intervertebral disc An intervertebral disc (or intervertebral fibrocartilage) lies between adjacent vertebrae in the vertebral column. Each disc forms a fibrocartilaginous joint (a symphysis), to allow slight movement of the vertebrae, to act as a ligament to hold the vertebrae together, and to function as a shock absorber for the spine. Intervertebral discs consist of an outer fibrous ring, the annulus fibrosus disci intervertebralis, which surrounds an inner gel-like center, the nucleus pulposus. The "annulus fibrosus" consists of several layers (laminae) of fibrocartilage made up of both type I and type II collagen. Type I is concentrated toward the edge of the ring, where it provides greater strength. The stiff laminae can withstand compressive forces. The fibrous intervertebral disc contains the "nucleus pulposus" and this helps to distribute pressure evenly across the disc. This prevents the development of stress concentrations which could cause damage to the underlying vertebrae or to their endplates. The nucleus pulposus contains loose fibers suspended in a mucoprotein gel. The nucleus of the disc acts as a shock absorber, absorbing the impact of the body's activities and keeping the two vertebrae separated. It is the remnant of the notochord. There is one disc between each pair of vertebrae, except for the first cervical segment, the "atlas".
Question #2	The linear molecule of DNA associated with proteins is:
Answer	*Eukaryotic chromosome*
Retrieved Passage	DNA sequences of telomeres maintained in yeast. Telomeres, the ends of eukaryotic chromosomes, have long been recognized as specialized structures. Their stability compared with broken ends of chromosomes suggested that they have properties which protect them from fusion, degradation or recombination. Furthermore, a linear DNA molecule such as that of a eukaryotic chromosome must have a structure at its ends which allows its complete replication, as no known DNA polymerase can initiate synthesis without a primer. At the ends of the relatively short, multi-copy linear DNA molecules found naturally in the nuclei of several lower eukaryotes, there are simple tandemly repeated sequences with, in the cases analysed, a specific array of single-strand breaks, on both DNA strands, in the distal portion of the block of repeats. In general, however, direct analysis of chromosomal termini presents problems because of their very low abundance in nuclei. To circumvent this problem, we have previously cloned a chromosomal telomere of the yeast Saccharomyces cerevisiae on a linear DNA vector molecule. Here we show that yeast chromosomal telomeres terminate in a DNA sequence consisting of tandem irregular repeats of the general form C1-3A. The same repeat units are added to the ends of Tetrahymena telomeres, in an apparently non-template-directed manner, during their replication on linear plasmids in yeast.
Question #3	They are multinucleated cells:
Answer	*Osteoclasts*
Retrieved Passage	The osteoclast generation: an in vitro and in vivo study with a genetically labelled avian monocytic cell line. Osteoclasts are multinucleate giant cells responsible for bone resorption. Osteoclast precursors are hematopoietic mononucleate cells, which give rise to osteoclasts after fusion. Nevertheless, the precise stage of differentiation where osteoclast precursors diverge from other hematopoietic lineages is still debated. We describe here both in vitro and in vivo approaches to the study of the osteoclast differentiation pathway. We used cells of the BM2 avian monocytic cell line, which are able to differentiate into macrophages both in vitro and in vivo. In order to follow the progeny of BM2 monocytes, we have derived a BM2 cell clone expressing the nlslacZ gene (BM2nlslacZ) which has still retained the main features of the parental cell line. In vitro, when BM2nlslacZ cells were triggered toward macrophages, they participated in the formation of multinucleate osteoclast-like cells as seen by their blue nuclei. Furthermore, when BM2nlslacZ cells were injected into the blood stream of chicken embryos, they could give rise to blue nucleate macrophages in the bone marrow, as well as to osteoclasts with blue nuclei in bone. Finally, we have shown that fusion of tagged mononucleate precursor cells not only occurs with other mononucleate precursor cells but also with mature multinucleate osteoclasts.

Table 5.14: Examples of correctly answered from HeadQA dataset, where questions represent general knowledge rather than patient case descriptions.

5.6 Conclusions

In this chapter, we have proposed a nearest neighbour strategy for interpreting patient case descriptions. Crucial to our solution is the use of a distantly supervised training set for fine-tuning the cross-encoder. Experimental results showed this strategy to perform well across three challenging benchmarks. Our results suggest that the lack of gold-annotated patient case descriptions can be overcome, at least to some extent, by using distant supervision strategies. We highlighted, in particular, that the setting of patient case descriptions allows us to avoid some of the usual pitfalls of distant supervision, as the presence of a disease or treatment name in two patient case descriptions provides us with reasonably reliable evidence that these descriptions are similar. Despite the fact that this method is highly effective, it requires a number of steps to reach good performance. In the next chapter, we will explore the possibility of using the considered PubMed corpus as an intermediate fine-tuning task to boost the performance of these LMs as a simpler solution.

Chapter 6

Self-Supervised Intermediate Fine-Tuning of Biomedical Language Models for Interpreting Patient Case Descriptions

6.1 Introduction

The baseline results from the previous chapter show that the basic fine-tuning performance of the state-of-the-art biomedical LMs remains rather low for benchmarks such as MedQA. However, we observed that reformulating the task of interpreting patient case descriptions in a way that relies on similar patient cases boosted the performance. We argue that this can, to some extent, be explained by the fact that interpreting patient case descriptions is a paragraph-level task, whereas the standard masked language modelling objective encourages the model to primarily focus on sentence-level context. Ideally, biomedical language models for interpreting patient case descriptions would be pre-trained on a task that involves predicting diagnoses, or other salient aspects of these patient cases. Unfortunately, beyond the training fragment of benchmarks such as MedQA, such labelled data is not readily available. As an alternative, in this chapter, we propose to generate a pseudo-labelled dataset, based on the heuristic that whenever a case description mentions a disease, it is likely (although by no means guaranteed) that this disease is a valid diagnosis, and similar for other medical concepts such as treatments. We then use this generated dataset for fine-tuning the LM on the proposed task, before finally fine-tuning it on our downstream task.

To get access to a large set of case descriptions, we rely on abstracts of published case reports. In particular, starting from a collection of PubMed abstracts, we first use a simple heuristic to identify those that are likely to correspond to case reports. Given a case report that mentions some disease, we then fine-tune the pre-trained LM on the task of predicting that disease. Note that the target disease is masked, as the task would otherwise be trivial. The pre-training task is formulated as a binary classification problem, i.e. given a patient description and a disease, is that disease the correct diagnosis (or more precisely, is it the disease that was masked). This formulation has the advantage that the input format is similar to that of multiple-choice question answering (QA) and natural language inference (NLI). Beyond diseases, we also experiment with predicting masked treatments. Similar to the usual masked language modelling objective, our pre-training task involves making predictions about masked text spans. However, due to the fact that we specifically mask diseases and treatments, we hypothesize that this will improve the model's ability to take the whole case description into account when making predictions. Nevertheless, diseases can be mentioned for two common reasons: (i) because the patient has been diagnosed with that disease, which is the case that underpins the intuition behind our proposed approach, or (ii) because the disease is relevant to the medical history of the patient. In the latter case, only a small part of the abstract may be relevant to the disease, which hampers the extent to which the model learns to focus on the case description as a whole. To address this issue, we propose to split abstracts in which multiple diseases are mentioned. As an alternative, we also propose a strategy where we consider case descriptions mentioning only one disease.

The rest of this chapter is organised as follows. Section 6.2 reviews the related work to this chapter. After that, Section 6.3 describes in detail our proposed method and each of the strategies. Subsequently, Section 6.4 lists the experiments, results, and ablation with some further strategies for analysis. Finally, Section 6.5 summarises our findings.

6.2 Related Work

More closely related to our approach, He et al. [67] propose a strategy which relies on the structure of Wikipedia to infuse knowledge about diseases. For instance, to teach the model about how diseases are treated, they rely on the fact that disease-centric Wikipedia articles tend to have a section called *Treatment*. They then combine the content of that section with a generated question-style sentence mentioning the aspect considered (i.e. treatment in this case) and a masked disease. However, rather than infusing encylopedic knowledge, our aim is to teach LMs to interpret patient case descriptions. Another related approach was introduced by Pergola et al. [161], who propose to fine-tune a biomedical language model by using a masked language modelling objective which is modified such that only biomedical concepts are masked. This approach has some similarities with our work, e.g. the idea of masking biomedical concepts as an intermediate fine-tuning task, but there are also some clear differences. First, we formulate our task as a binary classification problem, rather than masked language modelling. Moreover, we specifically target diseases and treatments, and we only mask one concept at a time (although all occurrences of that concept are masked). Finally, since we focus on paragraph-level understanding, we pay particular attention to how these input paragraphs can be selected. As we will see, each of these differences has a clear impact on the empirical results.

6.3 Proposed Method

We consider the problem of making predictions from patient case descriptions. For instance, given a description that lists symptoms and other information about the patient (e.g. gender, age, and medical history), we would like to infer the corresponding diagnosis or to recommend suitable treatments. We are specifically interested in the potential of using freely available case reports from the medical literature to improve the ability of standard biomedical LMs to make such predictions. In Section 6.3.1, we first explain our overall strategy. Subsequently, in Section 6.3.2 we describe the specific variants that we included in our analysis.

6.3.1 Overall Strategy

Our aim is to design an intermediate fine-tuning task for specialising biomedical LMs towards the task of interpreting patient case descriptions. This fine-tuning task relies on passages from PubMedDS [216], a corpus which primarily consists of abstracts from PubMed. First, we split the abstracts into passages of up to 250 words, to address the limitations on input length of BERT-based LMs. Next, we aim to identify those passages that contain a case report, describing a specific patient rather than more general findings. To this end, we rely on the simple but effective heuristic that case reports often mention the age of the patient. In particular, we select those passages that contain at least one keyword from the following list: *year-old male*, *year-old female*, *year-old boy*, *year-old girl*, *year-old woman*, *year-old man*. Let us write $\mathcal{D}$ for the resulting corpus, i.e. the set of passages that contain at least one of the aforementioned keywords.

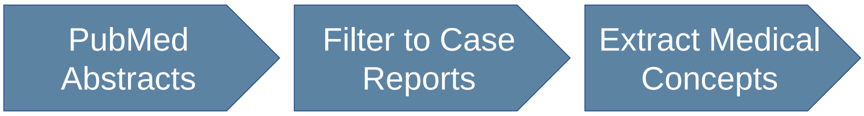

Figure 6.1: Preprocessing steps.

Subsequently, we determine which medical concepts are mentioned in the passages from $\mathcal{D}$. To this end, we use QuickUMLS [196] with UMLS-2020AA to identify both the spans and the semantic types (e.g. diseases, treatments) of the mentioned concepts.

Figure 6.1 shows a general overview of the preprocessing steps. Finally, we create positive training examples of the form (P,C), where C is a medical concept, and P is a passage from $\mathcal{D}$ in which all mentions of C have been replaced by a single <*mask*> token.

To generate negative training examples, we simply replace the medical concept C by another concept, as explained below. A given example (*passage*,*concept*) is encoded as follows: "<*cls*> *passage* <*sep*> *concept*", mimicking the input format that is typically used for question answering and natural language inference models. Figure 6.2 illustrate this process. The LM is fine-tuned on these examples using a standard cross-entropy loss.

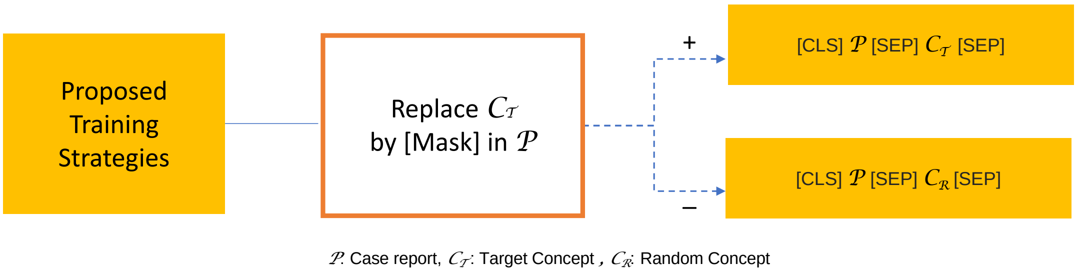
Proposed Training Strategies
Replace $C_\mathcal{T}$ by [Mask] in $\mathcal{P}$
+
[CLS] $\mathcal{P}$ [SEP] $C_\mathcal{T}$ [SEP]
–
[CLS] $\mathcal{P}$ [SEP] $C_\mathcal{R}$ [SEP]
$\mathcal{P}$: Case report, $C_\mathcal{T}$: Target Concept , $C_\mathcal{R}$: Random Concept

Figure 6.2: General overview of generating training examples process.

6.3.2 Training Strategies

We now describe the different variants that we considered. These variants primarily differ in the kinds of medical concepts that are selected as target concepts. Across all variants, we never mask the concept *disorder*, as constructing training examples from such mentions was found to be highly detrimental, given its prevalence and generic meaning. For all variants, we attempt to balance the number of positive and negative examples. Table 6.1 provides an overview of the total number of training examples arising from each of the following strategies.

	#
AnyType	1,011,482
SpecificType	
– *diseases*	160,534
– *treatments*	2,460
SplitDis	100,225
OneDis	3,310

Table 6.1: The total number of training examples for each of the intermediate fine-tuning tasks (#).

AnyType We create a positive example for every medical concept that is found (with the exception of *disorder*). Note that passages typically mention several concepts, hence this strategy allows us to derive multiple positive examples from the same passage, each time masking a different concept. To construct negative examples, we corrupt positive examples by randomly selecting a concept from those that have been identified in the corpus, regardless of the semantic type.

SpecificType In this variant, we only construct training examples from medical concepts of particular types. Specifically, we have experimented with diseases and treatments. Negative examples are constructed by replacing the target concept with another concept of the same semantic type, i.e. diseases are replaced by diseases, and treatments are replaced by treatments.

SplitDis Many passages contain more than one disease, which may confuse the model. For instance, diseases which are mentioned as part of the patient history may only be loosely related to the rest of the case report. Since our aim is to train the model to make predictions based on the whole case description, in this variant, passages containing more than one disease are split into sub-passages. In particular, when constructing a positive example for a target disease d, we select the sub-passage which begins with the first sentence in which d is mentioned, and includes all the subsequent sentences, until we reach a sentence that mentions another disease (where this final sentence is excluded from the selected sub-passage). If the target disease is mentioned in a sentence that also contains another disease, it is excluded altogether. For illustration, training examples that were obtained with the *SplitDis* strategy are presented in Table 6.2.

OneDis Instead of splitting passages mentioning more than one disease into sub-passages, as with SplitDis, here we simply discard such passages. This results in a much smaller number of positive examples, but with stronger guarantees that the disease being masked is salient. In both this and the *SplitDis* method, negative examples are obtained by using randomly selected diseases.

6.4 Experiments

In this section, we empirically analyse the different variants of the intermediate fine-tuning strategy.

Evaluation Datasets We mainly focus on two benchmarks that are specifically focused on interpreting and reasoning about patient cases. First, we use the English version of **MedQA** [82]. Results for this benchmark are reported in terms of accuracy (Acc). Second, we use **DisKnE**, in which we consider the task of ranking all test cases, according to our confidence that the given target disease is a valid diagnosis. The results are averaged across all diseases and are reported in terms of Mean Average Precision (MAP). In addition, we also consider the English version of **HeadQA** [219], as a more general healthcare-oriented QA dataset.

SpecificType-treatments	The role of [MASK] in the treatment of a patient with a pure silent pituitary somatotroph carcinoma. To describe a case of a pure silent somatotroph pituitary carcinoma. We describe a 54-year-old female with a clinically nonfunctioning pituitary macroadenoma diagnosed 15 years earlier. The patient underwent transsphenoidal surgery and no visible tumor remnant was observed for 6 years. A magnetic resonance imaging (MRI) detected the recurrence of a 1.2 × 1.5 cm macroadenoma. The patient was submitted to conventional radiotherapy (4500 cGy), and the tumor volume remained stable for 7 years. Then, an MRI revealed a slight increase in tumor size, and 2 years later, a subsequent MRI detected a very large, invasive pituitary mass. The patient was resubmitted to transsphenoidal surgery, and the histopathological examination showed diffuse positivity for growth hormone (GH). The nadir GH level during an oral glucose tolerance test was 0. 06 ng/mL, and the pre- and postoperative insulin like growth factor type I (IGF-I) levels were within the normal range. Abdominal, chest, brain, and spine MRI showed multiple small and hypervascular liver and bone lesions suggestive of metastases. Liver biopsy confirmed metastasis of GH-producing pituitary carcinoma. The patient has been treated with [MASK] and zoledronic acid for 7 months and with octreotide long-acting release (LAR) for 4 months. → *Temozolomide*
SpecificType-diseases	Intestinal cholesterol absorption inhibitor ezetimibe added to cholestyramine for sitosterolemia and xanthomatosis. Sitosterolemia is a rare, recessively inherited disorder characterized by increased absorption and delayed removal of noncholesterol sterols, which is associated with accelerated atherosclerosis, premature [MASK], hemolysis, and xanthomatosis. Treatments include low-sterol diet and bile salt-binding resins; however, these often do not reduce the xanthomatosis. We examined the effects of the intestinal cholesterol/phytosterol transporter inhibitor ezetimibe added to cholestyramine in a young female patient with sitosterolemia and associated xanthomatosis. The patient was an 11-year-old female with sitosterolemia presenting with prominent xanthomas in the subcutaneous tissue of both elbows who was receiving treatment with cholestyramine 2 g once daily. Bilateral carotid bruits were audible, and a grade II/VI systolic murmur was detected at the left upper sternal border. She also had a low platelet count of 111,000/microL. Ezetimibe 10 mg once daily was added to the patient's ongoing cholestyramine regimen, and she was evaluated for 1 year. The patient followed an unrestricted diet during the 1-year treatment period. After 1 year of treatment with ezetimibe added to ongoing cholestyramine therapy, the patient's plasma sitosterol and campesterol levels decreased by approximately 50. → *coronary artery disease*
SplitDis	After initial improvement artificial ventilation had to be be gun on day 3 because of an acute [MASK], diagnosed both clinically and radiologically. Despite additional antiviral and intensive medical treatment he died on day 11. → *respiratory distress syndrome*
	Traumatic [MASK] present diagnostic and therapeutic challenges. Owing to their fragile nature, endovascular intervention has become the first-line treatment; however, direct surgery has an advantage in certain cases. → *intracranial aneurysms*
	A fluoroscopic sniff test demonstrated diaphragmatic dysfunction and pulmonary function tests revealed [MASK] with evidence of neuromuscular etiology. → *restrictive pulmonary disease*

Table 6.2: Examples obtained with the different variants of the proposed strategies.

Setup We use four pre-trained LMs for the baselines and main experiments:

- the cased version of the standard BERT$_{\text{base}}$ [42];
- the cased version of SciBERT [24];
- the cased version of ClinicalBERT [8] that was trained on MIMIC-III while being initialized from BioBERT [104];
- the PubMedBERT model [61] that was trained from scratch on full-length PubMed articles as well as abstracts.

As a baseline, we directly fine-tune the models on the training data from the downstream task. For the other configurations, we first fine-tune the models on the proposed intermediate task.

We use the official training, validation, and test splits for each dataset, with the exception that we excluded questions with images for HeadQA.

Training Details We use the same settings and hyper-parameters for all datasets. For fine-tuning the models on the target task, we set the batch size to 8, the number of epochs to 4 and the learning rate to 2e-5. For the intermediate fine-tuning step, we again set the batch size to 8 and the learning rate to 2e-5. Regarding the number of epochs for intermediate fine-tuning, we note that the number of training examples varies greatly across the different variants. For this reason, and to mitigate the potential for catastrophic forgetting, we tuned the number of epochs, choosing from $\{2, 3, 4\}$, based on the development split of the downstream task.

Limitations Our method relies on an automated extraction tool for identifying the target medical concepts, which will inevitably lead to some noisy training examples. For example, *SplitDis* and *OneDis* rely on the assumption that we can detect all mentions of diseases in the text. More generally, regardless of performance, the predictions of a biomedical LM can clearly not be relied upon for diagnosing patients or recommending treatments in a clinical setting. Our purpose in studying these models is rather because a deeper understanding of patient records would make it possible to improve retrieval systems (e.g. suggesting relevant case reports to a clinician handling an unusual patient) or to identify hypotheses for medical research (e.g. by inducing patterns from large sets of case reports).

6.4.1 Results

Tables 6.3, 6.4 and 6.5 summarize our results. As can be seen, PubMedBERT clearly outperforms the other language models. In general, most variants of the intermediate fine-tuning tasks lead to clear improvements over the baselines. A clear and remarkable conclusion that can be observed for all benchmarks is that the type of intermediate fine-tuning data appears to be much more important than the number of training examples. For instance, the version of *SpecificType* which only uses treatments achieves the best overall results, outperforming the previous state-of-the-art for MedQA and achieving among the strongest results for both DisKnE and HeadQA. This is surprising, both because of the small number of training examples we can generate for this variant and because of the focus on diseases in DisKnE and many of the MedQA and HeadQA questions.

For MedQA, *SpecificType* with treatments outperforms the previous state-of-the-art [253] by 1.9 percentage points, despite not relying on any structured knowledge graphs. The *OneDis* variant performs well for DisKnE, despite the low number of corresponding training examples. For MedQA, *SplitDis* outperforms *SpecificType* with diseases (with the exception of BERT), which supports the idea that simply masking diseases can lead to training examples that are too noisy. While HeadQA is not particularly focused on patients case descriptions, we still see consistent improvements over the baselines with *SpecificType*, *SplitDis* and *OneDis*, although the improvements are somewhat smaller than those for MedQA and DisKnE.

We can see that our proposed strategy outperforms the baselines for each of the different language models, with the exception of SciBERT with DisKnE. However, there are some differences between the language models in terms of which variant of our method performs best. For MedQA, for instance, we can see that *SpecificType* with diseases is highly competitive for BERT and ClinicalBERT (compared to the other variants for these language models). As these are the language models that are least adapted to the considered task, we can indeed expect that more pre-training data might be needed for these models. This can explain the relative success of *SpecificType* with diseases and *SplitDis*, given that these are associated with a larger number of training examples.

Due to the computational cost that is required to report the results of several runs for all experiments, we average the results over three runs for MedQA's best configuration (i.e. PubMedBERT and *SpecificType* with treatments as the intermediate task). The average for basic fine-tuning is 35.9, with a standard deviation of 0.46. The average for *SpecificType* with treatments is 39.4, with a standard deviation of 1.08.

In general, compared to the distant supervision strategy in Chapter 5, we can observe from the results of this chapter that the improvements in Chapter 5 for BERT and ClinicalBERT are more significant. We assume the reason behind this is that the LMs were not completely reliant on the encoded knowledge but also on those external passages. This, in turn, made their performance more independent from what these LMs had already seen during pre-training. On the contrary, for DisKnE, we can observe that using PubMed abstracts in Chapter 5 did not yield better results over the standard fine-tuning, which could suggest that the proposed strategy in this chapter can somewhat mitigate the issue of the different writing styles.

6.4.2 Analysis

Table 6.6 shows the results of some variants of the *SpecificType* with diseases and *SplitDis* strategies, as explained next. We use PubMedBERT for these experiments, as

	BERT	ClinicalBERT	SciBERT	PubMedBERT
Baseline	27.8	29.1	29.2	35.5
AnyType	**28.2**	31.2	32.7	36.5
SpecificType				
– *diseases*	**28.2**	31.5	30.4	38.0
– *treatments*	27.8	31.0	**34.5**	**40.4**
SplitDis	27.7	**31.8**	33.4	38.7
OneDis	27.0	29.6	33.3	35.6

Table 6.3: Results for MedQA in terms of Accuracy.

this model achieved the best results in the main experiments. We focus on the MedQA benchmark as this is the most representative benchmark for our problem setting.

Frequent vs Rare We analyze whether there is any advantage in focusing specifically on common diseases, or conversely, in focusing on rare diseases. Table 6.6 shows the results of two variants of SpecificType, called *Most-Frequent* and *Least-Frequent*. The former only considers training examples, for the intermediate fine-tuning task, involving the 50 diseases which are most common in our corpus of case reports. Similarly, the *Least-Frequent* variant only considers the 5000 least frequent diseases. *Least-Frequent* achieves the best result, despite involving far fewer training examples than *Most-Frequent*. The results of both variants are either below or similar to those with the full set of diseases in Table 6.3.

General vs Specific Rather than selecting diseases based on their number of occurrences, here we investigate the effect of choosing diseases based on whether they are general or specific, in terms of the level at which they appear in the SNOMED CT

	BERT	ClinicalBERT	SciBERT	PubMedBERT
Baseline	57.0	67.5	**69.2**	69.7
AnyType	**64.2**	71.6	68.8	71.9
SpecificType				
– *diseases*	60.2	70.0	67.0	72.9
– *treatments*	57.5	67.5	68.3	73.6
SplitDis	58.3	**74.1**	68.1	72.2
OneDis	64.0	68.2	66.2	**74.4**

Table 6.4: Results for DisKnE in terms of Mean Average Precision (MAP).

hierarchies [199]. Specifically, for the *Most-General* variant, we only consider diseases with fewer than 5 ancestors in SNOMED CT. For the *Most-Specific* variant, we only consider diseases with at least 30 ancestors. We find that both variants of *SpecificType* perform similarly.

Similar vs Different We explore a setting in which only case reports about diseases similar to "heart disease" are provided during training. Specifically, we use cui2vec [22] to identify the 50 most similar diseases that occur at least once in our corpus of case reports. We then consider a variant of *SplitDis* where the only passages that are used are those in which *heart disease* occurs. In addition to passages in which *heart disease* occurs literally, we also include passages in which similar disease names are mentioned. For example, *heart failure*, *cardiovascular disease*, and *cerebrovascular disease* are among the considered diseases. Our aim in this experiment is to see whether training on one type of diseases is sufficient to obtain good results. Furthermore, we may also assume that because the resulting corpus only involves similar diseases, the model is forced to focus on more subtle details in the paragraphs, and might thus

	BERT	ClinicalBERT	SciBERT	PubMedBERT
Baseline	28.8	29.3	32.8	39.5
AnyType	29.3	30.0	31.7	39.1
SpecificType				
– *diseases*	29.8	30.1	34.5	**41.8**
– *treatments*	**30.3**	**31.1**	**35.7**	41.0
SplitDis	29.8	29.6	32.6	40.7
OneDis	29.7	29.8	34.0	40.8

Table 6.5: Results for HeadQA in terms of Accuracy.

improve as a result. To test this hypothesis, we also consider the variant *Least-Similar*, where we instead use the diseases that are least similar to *heart disease*. Rather than fixing the number of diseases at 50, in this case we chose the number to ensure a similar number of training examples as for *Most-Similar*. The results for both variants are below those of the standard *SplitDis* variant. However, we can see that *Most-Similar* clearly outperforms *Least-Similar*.

Adding Definitions We analyse the usefulness of UMLS definitions. Specifically, we augment the SplitDis training examples with examples of the form (def, dis), where *def* is the UMLS definition of a disease, and *dis* is the corresponding disease. Negative examples are again created by replacing the target disease with a randomly chosen other disease. The results in Table 6.6 show that adding definitions does not improve the results.

Diseases in Treatment Cases The good performance of the SpecificType variant with treatments, despite the small number of training examples we have for that setting,

			MedQA
		#	(Acc)
SpecificType	Most-Frequent	49,816	36.8
	Least-Frequent	8,466	38.0
	Most-General	7,229	36.6
	Most-Specific	8,778	36.9
	Treatment-Case-Dis	6,934	38.2
SplitDis	Most-Similar	1,858	37.7
	Least-Similar	1,870	36.7
	SplitDis+Def	105,952	37.7
	Treatment-Case-Dis	2,430	38.4

Table 6.6: Analysis results for MedQA (Accuracy). We also report the total number of training examples for each of the intermediate fine-tuning tasks (#). Results were obtained using PubMedBERT.

is one of the most surprising findings from the main experiments. Here we analyse whether this might be related to the quality of the case reports that were selected in that setting, i.e. the case reports that mention a treatment. To this end, we consider all such case reports, but instead of using the treatments as the target concepts, we instead focus on diseases. In other words, we use the SpecificType setting for diseases, but applied to the case reports that mention treatments. We also consider a variant in which the *SplitDis* setting is applied to these case reports. The results in Table 6.7, shown as *Treatment-Case-Dis*, reveal that this variant still underperforms the *SpecificCase* variant with treatments.

	MedQA
	(Acc)
– *SplitDis*	38.7
– *SpecificType: treatments*	40.4
MLM-RandomMask	
– *SplitDis*	36.4
– *SpecificType: treatments*	35.2
MLM-SpecificMask	
– *SplitDis*	37.6
– *SpecificType: treatments*	38.5
Random-Abstracts	
– *SplitDis*	38.2
– *SpecificType: treatments*	37.6
No Mask	
– *SplitDis*	36.7
– *SpecificType: treatments*	37.8
Remove-Sent (treatments)	38.9

Table 6.7: Ablation results for MedQA in terms of Accuracy. Results were obtained using PubMedBERT.

6.4.3 Ablation Experiments

In this section, we analyse the importance of a number of our design choices. We again focus on PubMedBERT and MedQA. We specifically consider the SplitDis and SpecificType with treatments, as these yielded the best results in the main experiments. The results are summarized in Table 6.7.

Masked Language Modelling We experimented with two variants of the masked language modelling (MLM) objective for the intermediate fine-tuning task. For the *MLM-RandomMask* variant, we randomly mask tokens, following the standard approach that is used for LM pre-training. For the *MLM-SpecificMask* variant, we specifically mask the tokens corresponding to diseases (for the SplitDis setting) or treatments (for the SpecificType setting). The results show that our approach outperforms both MLM strategies, while *MLM-SpecificMask* outperforms *MLM-RandomMask*.

Random Abstracts vs Case Reports We analyse the importance of specifically focusing on case reports. In the *Random-Abstracts* variant, rather than targeting abstracts which are likely to correspond to case reports, we consider a set of 60,000 randomly sampled abstracts from PubMedDS. We then use our SplitDis and SpecificType settings to construct the examples. The results in Table 6.7 show that using randomly chosen abstracts leads to worse results, compared to our standard setting.

Masking vs not Masking We consider a variant of the method in which the original passage is used, i.e. where we do not replace occurrences of the target disease with a <*mask*> token. The results in Table 6.7 clearly shows that masking is essential to achieve the best results. Nonetheless, even without masking we obtain results that are clearly better than those of the baseline (i.e. PubMedBERT without intermediate fine-tuning).

Masking vs Removing Sentences Instead of replacing the target concept with a <*mask*> token, here we remove the entire sentence in which this concept is mentioned. For this variant, called *Remove-Sent*, we only consider the SpecificType setting (with treatments), as using SplitDis would result in too few examples, given that several SplitDis examples consist of a single sentence. The results show that removing the sentence under-performs masking the concept.

6.5 Conclusions

In this chapter, we have proposed a strategy for intermediate fine-tuning of biomedical language models, to improve their ability to interpret patient case descriptions. The core of our strategy is to exploit abstracts of case reports found in the literature, as a surrogate of patient case descriptions, and to rely on the heuristic that diseases and treatments that are mentioned in such abstracts are likely to correspond to diagnoses and recommendations, respectively. Our observations suggested that the type and quality of intermediate fine-tuning data hold greater significance than the quantity of training examples. We performed a number of experiments to understand the impact of several design choices. Specifically, we evaluated the effect of two standard MLM variants instead of the binary classification settings for the intermediate fine-tuning task. In these variants, tokens are randomly masked or specifically masked based on the semantic type, such as diseases or treatments. We found that targeting specific types yielded better performance. However, both underperformed our main settings. Furthermore, we investigated the significance of focusing on case reports, showing that randomly chosen abstracts achieved worse results. Additionally, the results demonstrated that masking tokens is essential for achieving optimal results. Moreover, a comparison between masking and removing sentences revealed that masking the concept outperforms removing the entire sentence. Despite its conceptual simplicity and without the cost of manual annotation, this approach was found to lead to clear performance gains.

Chapter 7

Conclusions and Future Work

7.1 Introduction

This chapter concludes the thesis by presenting a summary of the main research contributions and key findings, coupled with a discussion of some possible future research. First, we revisit the aims, hypothesis and the undertaken research in Section 7.2. After that, we discuss the outcomes and the main findings related to the contributions, particularly by addressing each research question in Section 7.3. Subsequently, we suggest some future research directions to build on our work in Section 7.4.

7.2 Thesis Summary and Contributions

Advancing text representation techniques is one of the fundamental research areas within the NLP field, which has witnessed a boom since the introduction of the state-of-the-art pre-trained LMs. To better understand pre-trained LMs' strengths, a growing body of literature has been investigating what type of knowledge these models can encapsulate. Interestingly, it has been observed that such models capture a significant amount of world knowledge. However, the biomedical domain in general and, more specifically, the problem of interpreting patient case descriptions, which is the focus of this thesis, are understudied. This has several causes, including the scarcity of data. Therefore, in this thesis, we aimed to investigate the capabilities of pre-trained LMs to interpret patient case descriptions while also proposing approaches to enhance their performance using methods that obviate the need for additional human-labelled datasets.

The hypothesis of this thesis was presented in Chapter 1 as follows: *Existing biomedical LMs still struggle when it comes to interpreting patient case descriptions, which can partly be explained by the limited amounts of relevant annotated data. We hypothesize that the development of strategies that obviate the need for manual labelling can at least partially alleviate this issue, allowing biomedical LMs to interpret patient case descriptions with higher accuracy.* In order to examine the research hypothesis, we conducted several experiments, which are introduced in Chapter 4,5 and 6. We find that all support this hypothesis.

We started Chapter 2 by briefly reviewing the progress of neural text representation models prior to the introduction of pre-trained LMs. Subsequently, we explained the concepts associated with these LMs, and then we listed some of the common pre-trained LMs in the biomedical domain. After that, we summarised the related work to this thesis. Finally, we defined the considered downstream tasks for the proposed approaches and the employed supervision strategies. Then, in Chapter 3, we pinpointed the considered datasets for evaluating our work. Additionally, we presented the external resources and tools that were used to perform the experiments.

In Chapter 4, we introduced the Disease Knowledge Evaluation benchmark (DisKnE) to assess disease-centred knowledge captured by pre-trained LMs. In this benchmark, we obtain positive examples by organising the entailment pairs from the MedNLI and MEDIQA-NLI datasets into categories, reflecting the type of reasoning that we want to investigate. Then, we construct the negative examples in an adversarial way. We develop training-test splits that avoid leakage of disease knowledge. Lastly, we analysed the performance of several biomedical BERT variants for each category. We find that all considered models struggle with examples that require medical disease knowledge. We also observe that, without canonicalising the hypotheses, hypothesis-only baselines achieve high results in some categories. This shows that the original MedNLI dataset suffers from annotation artefacts, even within the set of entailment examples.

As per our findings from Chapter 4, biomedical language models struggle with medical reasoning tasks. Therefore, in Chapter 5, we attempted to improve the performance of these LMs for such tasks by exploiting unstructured text. However, it is rare to find sentences which express the exact type of knowledge that is needed for interpreting a given patient description. For this reason, rather than attempting to retrieve explicit medical knowledge, we instead proposed to rely on a nearest neighbour strategy to find similar patient cases.

Nevertheless, identifying similar cases is challenging, as descriptions of similar patients may superficially look rather different, among others because they often contain an abundance of irrelevant details. To address this challenge, we proposed a strategy that relies on a distantly supervised cross-encoder. In general, we retrieved text passages that are similar to the given patient description, and are thus likely to describe patients in similar situations, while also mentioning some hypothesis (e.g. a possible diagnosis of the patient). We then judged the likelihood of the hypothesis based on the similarity of the retrieved passages.

The overall pipeline for constructing the distantly supervised dataset for fine-tuning the cross-encoder is as follows: the starting point for this strategy is to query a corpus for passages that mention some hypothesis of interest (e.g. a disease, treatment, etc.). Each passage is then passed to a TSDAE model that we trained on MIMIC-III notes along with the patient description to compute the similarity score between each retrieved text passage and the given patient description. Subsequently, we used a cross-encoder, which we initialised with a pre-trained LM and then further fine-tuned based on labelled datasets for more or less related tasks. We then used this cross-encoder to re-rank the top-50 passages that resulted from the TSDAE step. Among the re-ranked 50 passages, we chose the top-k passages for each given patient description. This re-ranking step and the choice of k examples per patient description are used to construct our distantly supervised dataset. Finally, we fine-tuned our final cross-encoder with this dataset. Despite its conceptual simplicity, we found this strategy to be effective in practice.

In Chapter 6, we intended to improve the performance of LMs through a self-supervised intermediate fine-tuning strategy based on PubMed abstracts. Our solution is built on the observation that many of these abstracts are case reports, and thus essentially patient case descriptions. As a general strategy, we proposed to fine-tune biomedical language models on the task of predicting masked medical concepts from such abstracts. To achieve good results, we found that a careful selection of the target concepts is needed. For instance, strong results are obtained when only masking medical treatments. When masking diseases, the improvements over the baseline are sometimes smaller. This is surprising, given that most questions in the considered benchmarks are about diagnosing diseases. Upon closer inspection, the under-performance of strategies that rely on masking diseases appears to be related to the fact that diseases can be mentioned for reasons other than being the diagnosis. To address this issue, we proposed to split abstracts in which multiple diseases are mentioned. Despite the simplicity of the overall approach, our fine-tuning strategies enabled significant improvements for the considered evaluation datasets.

To conclude, the outcomes from the latest three chapters support the hypothesis. In particular, we found that the considered LMs do struggle when it comes to tasks that require medical reasoning, and we were able to construct datasets that enhanced their capabilities without the need for manual annotations.

7.3 Research Questions

In this section, we revisit the research questions formulated in Section 1.2 and relate each question to the work that was carried out in this thesis to answer it.

Research Question 1: *What kinds of medical knowledge do pre-trained LMs capture? More specifically, are these models capable of performing medical reasoning such as linking symptoms to diseases, or treatments to diseases?*

Our results from Chapter 4 showed that all the considered LMs struggle with NLI examples that require medical knowledge. We also found that the relative performance of the pre-trained models differs across medical categories (symptoms to diseases, treatments to diseases, etc.), where the best performance is obtained by ClinicalBERT, BioBERT, SciBERT or BERT depending on the category and experimental setting. Conversely, for examples that are based on terminological knowledge, overall performance is much higher, with relatively little difference between different pre-trained models.

Research Question 2: *Is it possible to use nearest neighbour strategies for enhancing the LM's interpretation of patient case descriptions (i.e. relying on similar patient cases to drive the predictions)? Can we construct distantly supervised datasets to compensate for the lack of annotated datasets to train the model on identifying similar patient cases?*

To answer this question, in Chapter 5, we constructed a distantly supervised dataset using two main steps, in which we ultimately aim to obtain a set of pairs representing similar patient cases. In particular, we used an unsupervised text encoder for the initial retrieval step from a considered set of passages and then a cross encoder for the re-tanking step. We relied on the possibility that a mentioned medical concept, for example, a disease, in a passage is likely to represent the diagnosis. Our experimental results show that our overall approach is highly effective, improving the performance of LMs in question answering and NLI tasks about patient descriptions.

Research Question 3: *How, and to what extent, can we obtain self-supervised datasets that serve as intermediate tasks for fine-tuning the LMs?*

We addressed this question in Chapter 6, in which we designed intermediate fine-tuning strategies tailored for the task of interpreting patient case descriptions. We specifically targeted case reports found in PubMed abstracts. The labels are obtained in a self-supervised way, specifically by extracting the mentioned medical concepts, and the task is then formulated as a binary classification. As input to the LM, we provided the case report, in which we replaced the mentioned medical concept with a <*Mask*> token, and the medical concept. This setting particularly corresponds to the construction of the positive examples. The negative examples are obtained by randomly selecting another concept. We find that this intermediate fine-tuning leads to substantial improvements in downstream tasks, even when using a biomedical LM that was already pre-trained on PubMed. To some extent, this comes from the fact that we specifically fine-tune the model on case reports. However, this in itself is not sufficient. We found that the success of this strategy crucially depends on the selection of the medical concepts to be masked.

7.4 Future Work

In terms of future research, this section highlights possible directions to further refine and investigate the capabilities of pre-trained biomedical LMs, with more focus on aspects to build on our work.

Considering other sources of knowledge. The datasets in this thesis are constructed based on free text only. However, the work could be extended to structured sources, such as knowledge graphs. As shown by Meng et al. [131], biomedical knowledge graphs can play an important role in interpreting patient descriptions. Hence, integrating such resources with the considered nearest neighbour strategy proposed in Chapter 5 is a natural direction to explore. For example, candidate answers with no retrieved passages could be replaced by KG triples (converted into text) by taking into consideration the relation between every medical concept mentioned in the case description (if any) and the answer candidate. Beyond that, the pre-trained LMs could potentially serve as a source of knowledge once they achieve sufficient performance. As we have seen in Chapter 4, the performance of the LMs differs across categories, suggesting that distilling the knowledge from a set of LMs is also a possible future direction. One way to employ such a strategy could be, for instance, through designing prompts (manually or automatically) to construct synthesis examples, in which we query a model using a piece of text, and the task is formulated as a fill-in-the-blank. These could be used as additional examples for the proposed datasets in Chapter 6.

Summarisation. Previous work [204], and our experiments in Chapter 5 and Chapter 6, suggest that augmenting the knowledge with sentence-level context might not be adequate for paragraph-level understanding tasks, namely in the case of interpreting patient case descriptions. Whereas the use of text with longer length is useful for these tasks, alternatives could be explored in further research for the sake of training more efficient models. For instance, in Chapter 5, we used the top-k retrieved passages to construct the distantly supervised dataset. A suggested future direction is to summarise such passages. This could possibly be accomplished in two ways. One is to summarise the case description itself or the retrieved passage focusing on text snippets that are mostly related to the given case and the answer candidate. The other way is to summarise a set of passages to have only the most informative sentences.

Meta-Learning. The findings of Chapter 4 suggested that the considered LMs have complementary strengths. Several LMs could have been pre-trained on different textual resources. Consequently, one pre-trained LM could possibly have superiority over the other in solving a specific patient case. This can be seen in the case of LMs that have been trained on a particular corpus, such as MIMIC-III notes. For instance, the knowledge about diseases learned from pre-training on MIMIC-III notes might be mostly related to what is commonly seen in critical care units. In our work, we attempted to improve each individual model. Nevertheless, the joint use of several pre-trained LMs is another possible area of future research. For example, this could be achieved by developing a mixture of experts (MoE) model for the biomedical domain where each expert (LM) specialises in a family of diseases or a particular text genre.

De-noising. The automation of label generation can inevitably lead to noisy instances. To further enhance the results of the proposed strategies in this thesis, several de-nosing techniques could be applied to distinguish informative from noisy examples. For instance, in Chapter 5 and Chapter 6, we mainly assume that the mention of the medical concept within a given passage (e.g. a disease or treatment) is associated with that particular patient case at that time. Yet, developing tools to specify whether the extracted disease is the diagnosis or whether the treatment is what has been prescribed for that case, may contribute to alleviating the noise issue.

Transfer Learning. Investigating and improving the ability of pre-trained LMs to perform medical reasoning, such as linking a set of symptoms to a particular disease, is helpful for several downstream tasks in the biomedical domain, such as relation classification and extraction. Thus, the work in this thesis, or specifically the task of interpreting patient case descriptions, could serve as an intermediate task to enhance the inference abilities of pre-trained LMs for other target biomedical tasks.

Differential Diagnosis. Several diseases share or have closely related symptoms, which makes it challenging for clinicians to provide a diagnosis before going through a list of potential conditions. In this thesis, we used available datasets with various diseases, many of which have distinct symptoms. Therefore, constructing a dataset that is primarily concerned with diseases and conditions sharing similar symptoms, signs or treatments to train and evaluate the models is a possible future direction. This, in turn, could help healthcare providers prioritise or eliminate the list of possible causes, saving time, cost and effort. Although in Chapter 4, we used similar diseases as negative, adversarial examples, we relied on cui2vec to identify the notion of similarity. A starting point could be to filter and organise the existing datasets utilised in this thesis in categories based on the commonality of symptoms across different patient cases.

Explainability. Interpretability and explainability are necessary for adapting machine learning models to healthcare applications. Ideally, the predictions of pre-trained LMs should only be relied upon with an explanation or evidence. Developing evidence-based strategies improves the transparency and credibility of the decision-making process. In Chapter 5, the predictions were mainly derived from computing the similarity between a given patient's case and other patients' cases. Such a retrieval method provides some degree of explainability. While there has been a surge of research in this direction, additional future work for the biomedical domain should be explored to understand the models' reasoning behind the predictions.

Multi-Modality. In several cases, text alone is inadequate to build knowledge or make a decision about a patient's case. Images such as X-rays, CT-Scans, and even pictures included in biomedical articles are sometimes essential for accurate results. Therefore, fusing both modalities (i.e. text and images) to complement the needed information is necessary for these situations. Advances in such methods would be effective at the task of interpreting patient case descriptions and many other biomedical tasks.

Rare Diseases. Using AI models in clinical settings would be particularly useful when the target disease is rare and not frequently seen by doctors. Therefore, developing methods that are less dependent on the frequent occurrence of a medical concept within the corpus is also a helpful and necessary direction for the biomedical domain. In Chapter 5, we mainly relied on the similarity between text passages, which could be an example of a strategy that somewhat contributes towards this aim. Further research could be explored to develop models and find ways that take into consideration such diseases.

Other Languages. The work in this thesis is focused on the English language. However, extending this focus to other languages can bring several benefits, particularly in the field of healthcare, which is a global concern for people from diverse cultures who speak different languages. Some non-English speaking countries have their own EHR systems, which imposes the necessity to adapt the biomedical models to such languages. Nevertheless, several challenges still remain due to various factors. One of the primary difficulties is the availability of resources in the target language. While training models from scratch has its own benefits such as maintaining cultural sensitivities and local medical practices, that may not be always feasible due to the lack of training data. One approach is to directly translate such resources from English, which has abundant data. However, such direct translation to the target language may result in the absence of cultural differences. Another approach involves developing cross-lingual models that align with regional and cultural variations, thereby avoiding biases. Additionally, generative LMs can be employed to synthesize training or evaluation data, with the involvement of human experts to validate such models.

Bibliography

[1] A. B. Abacha and D. Demner-Fushman. Recognizing question entailment for medical question answering. In *AMIA Annual Symposium Proceedings*, volume 2016, page 310. American Medical Informatics Association, 2016.

[2] A. B. Abacha, C. Shivade, and D. Demner-Fushman. Overview of the MEDIQA 2019 shared task on textual inference, question entailment and question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 370–379, 2019.

[3] O. Agarwal, H. Ge, S. Shakeri, and R. Al-Rfou. Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3554–3565, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.278. URL `https://aclanthology.org/2021.naacl-main.278`.

[4] I. Alghanmi, L. Espinosa Anke, and S. Schockaert. Combining BERT with static word embeddings for categorizing social media. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 28–33, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.wnut-1.5. URL `https://aclanthology.org/2020.wnut-1.5`.

[5] I. Alghanmi, L. Espinosa Anke, and S. Schockaert. Probing pre-trained language models for disease knowledge. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3023–3033, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.266. URL `https://aclanthology.org/2021.findings-acl.266`.

[6] I. Alghanmi, L. Espinosa-Anke, and S. Schockaert. Interpreting patient descriptions using distantly supervised similar case retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 460–470, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450387323. doi: 10.1145/3477495.3532003. URL `https://doi.org/10.1145/3477495.3532003`.

[7] I. Alghanmi, L. Espinosa-Anke, and S. Schockaert. Self-supervised intermediate fine-tuning of biomedical language models for interpreting patient case descriptions. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1432–1441, Gyeongju, Republic of Korea, Oct. 2022. International Committee on Computational Linguistics. URL `https://aclanthology.org/2022.coling-1.123`.

[8] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, and M. McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1909. URL `https://aclanthology.org/W19-1909`.

[9] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, and M. McDermott. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, 2019.

[10] S. Althammer, A. Askari, S. Verberne, and A. Hanbury. Dossier@ coliee 2021: Leveraging dense retrieval and summarization-based re-ranking for case law retrieval. *arXiv preprint arXiv:2108.03937*, 2021.

[11] S. Amin, P. Minervini, D. Chang, P. Stenetorp, and G. Neumann. MedDistant19: Towards an accurate benchmark for broad-coverage biomedical relation extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2259–2277, Gyeongju, Republic of Korea, Oct. 2022. International Committee on Computational Linguistics. URL `https://aclanthology.org/2022.coling-1.198`.

[12] Y. An, L. Zhang, H. Yang, L. Sun, B. Jin, C. Liu, R. Yu, and X. Wei. Prediction of treatment medicines with dual adaptive sequential networks. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[13] V. Araujo, A. Carvallo, C. Aspillaga, and D. Parra. On adversarial examples for biomedical NLP tasks. *arXiv:2004.11157*, 2020.

[14] V. Araujo, A. Carvallo, C. Aspillaga, C. Thorne, and D. Parra. Stress test evaluation of biomedical word embeddings. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 119–125, 2021.

[15] C. W. Arnold, S. M. El-Saden, A. A. Bui, and R. Taira. Clinical case-based retrieval using latent topic analysis. In *AMIA annual symposium proceedings*, volume 2010, page 26. American Medical Informatics Association, 2010.

[16] A. R. Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2001.

[17] A. Arora, L.-A. Kaffee, and I. Augenstein. Probing pre-trained language models for cross-cultural differences in values. *arXiv preprint arXiv:2203.13722*, 2022.

[18] C. Aspillaga, A. Carvallo, and V. Araujo. Stress test evaluation of transformer-based models in natural language understanding tasks. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1882–1894, 2020.

[19] I. Augenstein, S. Ruder, and A. Søgaard. Multi-task learning of pairwise sequence classification tasks over disparate label spaces. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1896–1906, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1172. URL `https://aclanthology.org/N18-1172`.

[20] N. Barlaug and J. A. Gulla. Neural networks for entity matching: A survey. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(3):1–37, 2021.

[21] N. Barzegar Marvasti, C. B. Akgül, B. Acar, N. Kökciyan, S. Üsküdarlı, P. Yolum, R. Türkay, and B. Bakır. Clinical experience sharing by similar case retrieval. In *Proceedings of the 1st ACM international workshop on Multimedia indexing and information retrieval for healthcare*, pages 67–74, 2013.

[22] A. L. Beam, B. Kompa, A. Schmaltz, I. Fried, G. Weber, N. Palmer, X. Shi, T. Cai, and I. S. Kohane. Clinical concept embeddings learned from massive sources of multimodal medical data. In *Pacific Symposium on Biocomputing*, volume 25, pages 295–306, 2020.

[23] I. Beltagy, K. Lo, and A. Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3613–3618, 2019.

[24] I. Beltagy, K. Lo, and A. Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Meth-*

ods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3615–3620, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1371. URL `https://aclanthology.org/D19-1371`.

[25] M. W. Berry, A. Mohamed, and B. W. Yap. *Supervised and unsupervised learning for data science*. Springer, 2019.

[26] W. Boag, D. Doss, T. Naumann, and P. Szolovits. What's in a note? unpacking predictive value in clinical note representations. *AMIA Summits on Translational Science Proceedings*, 2018:26, 2018.

[27] O. Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

[28] R. Bommasani, K. Davis, and C. Cardie. Interpreting pretrained contextualized representations via reductions to static embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781, 2020.

[29] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Çelikyilmaz, and Y. Choi. COMET: commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 4762–4779, 2019.

[30] Z. Bouraoui, J. Camacho-Collados, and S. Schockaert. Inducing relational knowledge from BERT. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 7456–7463, 2020.

[31] C. Buckley and S. Robertson. Relevance feedback track overview: Trec 2008. Technical report, MICROSOFT CORP REDMOND WA, 2008.

[32] C. Buckley, G. Salton, J. Allan, and A. Singhal. Automatic query expansion using smart: Trec 3. *NIST special publication sp*, pages 69–69, 1995.

[33] J. Camacho-Collados, Y. Doval, E. Martínez-Cámara, L. Espinosa-Anke, F. Barbieri, and S. Schockaert. Learning cross-lingual word embeddings from twitter via distant supervision. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 72–82, 2020.

[34] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada, Aug. 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2001. URL https://aclanthology.org/S17-2001.

[35] D. Chang, E. Lin, C. Brandt, R. A. Taylor, et al. Incorporating domain knowledge into language models by using graph convolutional networks for assessing semantic textual similarity: Model development and performance comparison. *JMIR Medical Informatics*, 9(11):e23101, 2021.

[36] T.-Y. Chang and C.-J. Lu. Rethinking why intermediate-task fine-tuning works. *arXiv preprint arXiv:2108.11696*, 2021.

[37] W.-C. Chang, F. X. Yu, Y.-W. Chang, Y. Yang, and S. Kumar. Pre-training tasks for embedding-based large-scale retrieval. *arXiv preprint arXiv:2002.03932*, 2020.

[38] A. Conneau, G. Kruszewski, G. Lample, L. Barrault, and M. Baroni. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*, 2018.

[39] J. Davison, J. Feldman, and A. M. Rush. Commonsense knowledge mining from pretrained models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 1173–1178, 2019.

[40] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[41] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019.

[42] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423`.

[43] I. Deznabi, M. Iyyer, and M. Fiterau. Predicting in-hospital mortality by combining clinical notes with time-series data. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4026–4031, 2021.

[44] Y. Ding, Z. Huang, R. Wang, Y. Zhang, X. Chen, Y. Ma, H. Chung, and S. C. Han. V-doc: Visual questions answers with documents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21492–21498, June 2022.

[45] K. Donnelly et al. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279, 2006.

[46] D. Downey, M. Broadhead, and O. Etzioni. Locating complex named entities in web text. In *IJCAI*, volume 7, pages 2733–2739, 2007.

[47] J. Du, M. Ott, H. Li, X. Zhou, and V. Stoyanov. General purpose text embeddings from pre-trained language models for scalable inference. In *Findings*

of the Association for Computational Linguistics: EMNLP 2020, pages 3018–3030, 2020.

[48] G. Echegoyen, A. Rodrigo, and A. Penas. Cross-lingual training for multiple-choice question answering. *Procesamiento del Lenguaje Natural*, 65:37–44, 2020.

[49] K. Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, 2019.

[50] M. Forbes, A. Holtzman, and Y. Choi. Do neural language representations learn physical commonsense? In *Proceedings of the 41th Annual Meeting of the Cognitive Science Society*, pages 1753–1759, 2019.

[51] G. Fu, C. Song, J. Li, Y. Ma, P. Chen, R. Wang, B. X. Yang, Z. Huang, et al. Distant supervision for mental health management in social media: suicide risk classification system development study. *Journal of medical internet research*, 23(8):e26119, 2021.

[52] S. Fu, D. Chen, H. He, S. Liu, S. Moon, K. J. Peterson, F. Shen, L. Wang, Y. Wang, A. Wen, et al. Clinical concept extraction: a methodology review. *Journal of biomedical informatics*, 109:103526, 2020.

[53] A. Gajbhiye, N. A. Moubayed, and S. Bradley. ExBERT: An external knowledge enhanced BERT for natural language inference. In *International Conference on Artificial Neural Networks*, pages 460–472. Springer, 2021.

[54] J. Geraci, P. Wilansky, V. de Luca, A. Roy, J. L. Kennedy, and J. Strauss. Applying deep neural networks to unstructured text notes in electronic medical records for phenotyping youth depression. *Evidence-based mental health*, 20(3):83–87, 2017.

[55] M. Glockner, V. Shwartz, and Y. Goldberg. Breaking nli systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 650–655, 2018.

[56] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009, 2009.

[57] Y. Goldberg. Assessing bert's syntactic abilities. *arXiv preprint arXiv:1901.05287*, 2019.

[58] S. N. Golmaei and X. Luo. Deepnote-gnn: predicting hospital readmission using clinical notes and patient network. In *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 1–9, 2021.

[59] C. Gormley and Z. Tong. *Elasticsearch: the definitive guide: a distributed real-time search and analytics engine*. " O'Reilly Media, Inc.", 2015.

[60] R. Grishman and J. Sterling. Information extraction and semantic constraints. In *COLING 1990 Volume 3: Papers presented to the 13th International Conference on Computational Linguistics*, 1990.

[61] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

[62] S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. Bowman, and N. A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 107–112, 2018.

[63] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang. Retrieval augmented language model pre-training. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR, 2020.

[64] X. Han, Z. Zhang, N. Ding, Y. Gu, X. Liu, Y. Huo, J. Qiu, Y. Yao, A. Zhang, L. Zhang, et al. Pre-trained models: Past, present and future. *AI Open*, 2:225–250, 2021.

[65] B. Hao, H. Zhu, and I. Paschalidis. Enhancing clinical bert embedding using a biomedical knowledge base. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 657–661, 2020.

[66] B. He, D. Zhou, J. Xiao, X. Jiang, Q. Liu, N. J. Yuan, and T. Xu. Bert-mk: Integrating graph contextualized knowledge into pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2281–2290, 2020.

[67] Y. He, Z. Zhu, Y. Zhang, Q. Chen, and J. Caverlee. Infusing disease knowledge into BERT for health question answering, medical inference and disease name recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4604–4614, 2020.

[68] Y. He, Z. Zhu, Y. Zhang, Q. Chen, and J. Caverlee. Infusing disease knowledge into BERT for health question answering, medical inference and disease name recognition. *arXiv preprint arXiv:2010.03746*, 2020.

[69] J. Hewitt and C. D. Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, 2019.

[70] J. Howard and S. Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 328–339, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1031. URL `https://aclanthology.org/P18-1031`.

[71] H.-z. Huang, X.-d. Lu, W. Guo, X.-b. Jiang, Z.-m. Yan, and S.-p. Wang. Heterogeneous information network-based patient similarity search. *Frontiers in Cell and Developmental Biology*, page 2297, 2021.

[72] K. Huang, J. Altosaar, and R. Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

[73] S. Iyer, S. Min, Y. Mehdad, and W.-t. Yih. Reconsider: Improved re-ranking using span-focused cross-attention for open domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1280–1287, 2021.

[74] G. Jawahar, B. Sagot, and D. Seddah. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[75] M. Jeong, M. Sung, G. Kim, D. Kim, W. Yoon, J. Yoo, and J. Kang. Transferability of natural language inference to biomedical question answering. *arXiv preprint arXiv:2007.00217*, 2020.

[76] K. Jha and A. Zhang. Continual knowledge infusion into pre-trained biomedical language models. *Bioinformatics*, 38(2):494–502, 2022.

[77] S. Ji and P. Marttinen. Patient outcome and zero-shot diagnosis prediction with hypernetwork-guided multitask learning. *arXiv preprint arXiv:2109.03062*, 2021.

[78] Z. Jia, X. Zeng, H. Duan, X. Lu, and H. Li. A patient-similarity-based model for diagnostic prediction. *International journal of medical informatics*, 135: 104073, 2020.

[79] J. Jiang. Information extraction from text. In *Mining text data*, pages 11–41. Springer, 2012.

[80] J. Jiang and C. Zhai. A systematic exploration of the feature space for relation extraction. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 113–120, 2007.

[81] Z. Jiang, F. F. Xu, J. Araki, and G. Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020.

[82] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

[83] Q. Jin, B. Dhingra, W. Cohen, and X. Lu. Probing biomedical embeddings from language models. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 82–89, 2019.

[84] Q. Jin, B. Dhingra, Z. Liu, W. Cohen, and X. Lu. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2567–2577, 2019.

[85] T. Jo. *Machine Learning Foundations*. Springer, 2021.

[86] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-Wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1):1–9, 2016.

[87] M. Joshi, K. Lee, Y. Luan, and K. Toutanova. Contextualized representations using textual encyclopedic knowledge. *CoRR*, abs/2004.12006, 2020. URL `https://arxiv.org/abs/2004.12006`.

[88] K. S. Kalyan, A. Rajasekharan, and S. Sangeetha. Ammu: a survey of transformer-based biomedical pretrained language models. *Journal of biomedical informatics*, page 103982, 2021.

[89] U. Kamath, J. Liu, and J. Whitaker. *Deep learning for NLP and speech recognition*, volume 84. Springer, 2019.

[90] S. Kaneko, A. Hayashi, N. Suematsu, and K. Iwata. Hierarchical hidden conditional random fields for information extraction. In *International Conference on Learning and Intelligent Optimization*, pages 191–202. Springer, 2011.

[91] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, 2020.

[92] W. Kearns, W. Lau, and J. Thomas. UW-BHI at MEDIQA 2019: An analysis of representation methods for medical natural language inference. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 500–509, 2019.

[93] U. Khandelwal, O. Levy, D. Jurafsky, L. Zettlemoyer, and M. Lewis. Generalization through memorization: Nearest neighbor language models. In *8th International Conference on Learning Representations*, 2020.

[94] O. Khattab and M. Zaharia. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48, 2020.

[95] O. Khattab, C. Potts, and M. Zaharia. Relevance-guided supervision for openqa with colbert. *Transactions of the Association for Computational Linguistics*, 9: 929–944, 2021.

[96] J. H. Kim, J. H. Lee, and K. J. Lee. A study on the issues related to building a library information system based on deep learning. In *2021 21st ACIS International Winter Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD-Winter)*, pages 287–289. IEEE, 2021.

[97] N. Kim and T. Linzen. Compositionality as directional consistency in sequential neural networks. In *Workshop on Context and Compositionality in Biological and Artificial Neural Systems*, 2019.

[98] D. Koutsomitropoulos. Validating ontology-based annotations of biomedical resources using zero-shot learning. In *The 12th International Conference on Computational Systems-Biology and Bioinformatics*, CSBio2021, page 37–43, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450385107. doi: 10.1145/3486713.3486730. URL `https://doi.org/10.1145/3486713.3486730`.

[99] B. S. A. Kshatriya, N. A. Nunez, M. G. Resendez, E. Ryu, B. J. Coombes, S. Fu, M. A. Frye, J. M. Biernacka, and Y. Wang. Neural language models with distant supervision to identify major depressive disorder from clinical notes. *arXiv preprint arXiv:2104.09644*, 2021.

[100] S. Laicher, S. Kurtyigit, D. Schlechtweg, J. Kuhn, and S. S. im Walde. Explaining and improving bert performance on lexical semantic change detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 192–202, 2021.

[101] J.-B. Lamy, A. Venot, and C. Duclos. Pymedtermino: an open-source generic api for advanced terminology services. In *Digital Healthcare Empowering Europeans*, pages 924–928. IOS Press, 2015.

[102] V. Lavrenko and W. B. Croft. Relevance-based language models: Estimation and analysis. *Croft and Lafferty [2]*, pages 1–5, 2001.

[103] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 09 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz682. URL `https://doi.org/10.1093/bioinformatics/btz682`.

[104] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[105] K. Lee, M.-W. Chang, and K. Toutanova. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, 2019.

[106] P. Lewis, M. Ott, J. Du, and V. Stoyanov. Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157, 2020.

[107] F. Li and H. Yu. Icd coding from clinical text using multi-filter residual convolutional neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8180–8187, 2020.

[108] Y. Li, G. Long, T. Shen, T. Zhou, L. Yao, H. Huo, and J. Jiang. Self-attention enhanced selective gate with entity-aware embedding for distantly supervised relation extraction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8269–8276, 2020.

[109] Y. Li, B. Qian, X. Zhang, and H. Liu. Graph neural network-based diagnosis prediction. *Big Data*, 8(5):379–390, 2020.

[110] Y. Li, S. Rao, J. R. A. Solares, A. Hassaine, R. Ramakrishnan, D. Canoy, Y. Zhu, K. Rahimi, and G. Salimi-Khorshidi. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):1–12, 2020.

[111] B. Y. Lin, S. Lee, X. Qiao, and X. Ren. Common sense beyond english: Evaluating and improving multilingual language models for commonsense reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1274–1287, 2021.

[112] J. Lin, R. Nogueira, and A. Yates. Pretrained transformers for text ranking: Bert and beyond. *Synthesis Lectures on Human Language Technologies*, 14(4): 1–325, 2021.

[113] R.-H. Lin. An intelligent model for liver disease diagnosis. *Artificial Intelligence in Medicine*, 47(1):53–62, 2009.

[114] X. Lin, T. Liu, W. Jia, and Z. Gong. Distantly supervised relation extraction using multi-layer revision network and confidence-based multi-instance learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 165–174, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.15. URL `https://aclanthology.org/2021.emnlp-main.15`.

[115] Y. Lin, Y. C. Tan, and R. Frank. Open sesame: Getting inside bert's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, 2019.

[116] W. Liu, P. Zhou, Z. Zhao, Z. Wang, Q. Ju, H. Deng, and P. Wang. K-BERT: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2901–2908, 2020.

[117] Y. Liu, S. Chowdhury, C. Zhang, C. Caragea, and P. S. Yu. Interpretable multi-step reasoning with knowledge extraction on complex healthcare question answering. *CoRR*, abs/2008.02434, 2020.

[118] H. Lu and S. Uddin. A weighted patient network-based framework for predicting chronic diseases using graph neural networks. *Scientific reports*, 11(1):1–12, 2021.

[119] M. Lu, Y. Fang, F. Yan, and M. Li. Incorporating domain knowledge into natural language inference on clinical texts. *IEEE Access*, 7:57623–57632, 2019.

[120] Q. Lu, T. H. Nguyen, and D. Dou. Predicting patient readmission risk from medical text via knowledge graph enhanced multiview graph convolution. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1990–1994, 2021.

[121] J. Luo, M. Ye, C. Xiao, and F. Ma. Hitanet: Hierarchical time-aware attention networks for risk prediction on electronic health records. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 647–656, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3403107. URL `https://doi.org/10.1145/3394486.3403107`.

[122] D. Mahajan, A. Poddar, J. J. Liang, Y.-T. Lin, J. M. Prager, P. Suryanarayanan, P. Raghavan, and C.-H. Tsou. Identification of semantically similar sentences in clinical notes: Iterative intermediate training using multi-task learning. *JMIR medical informatics*, 8(11):e22508, 2020.

[123] S. Malakouti and M. Hauskrecht. Hierarchical deep multi-task learning for classification of patient diagnoses. In *International Conference on Artificial Intelligence in Medicine*, pages 122–132. Springer, 2022.

[124] S. Marchesin and G. Silvello. Tbga: a large-scale gene-disease association dataset for biomedical relation extraction. *BMC bioinformatics*, 23(1):1–16, 2022.

[125] V. Maslej-Krešňáková, M. Sarnovskỳ, P. Butka, and K. Machová. Comparison of deep learning models and various text pre-processing techniques for the toxic comments classification. *Applied Sciences*, 10(23):8631, 2020.

[126] Y. Mass and H. Roitman. Ad-hoc document retrieval using weak-supervision with BERT and GPT2. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4191–4197, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.343. URL https://aclanthology.org/2020.emnlp-main.343.

[127] J. Mathew, S. Fakhraei, and J. L. Ambite. Biomedical named entity recognition via reference-set augmented bootstrapping. *arXiv preprint arXiv:1906.00282*, 2019.

[128] B. McCann, J. Bradbury, C. Xiong, and R. Socher. Learned in translation: Contextualized word vectors. *Advances in neural information processing systems*, 30, 2017.

[129] O. Melamud, J. Goldberger, and I. Dagan. context2vec: Learning generic context embedding with bidirectional lstm. In *Proceedings of the 20th SIGNLL conference on computational natural language learning*, pages 51–61, 2016.

[130] Y. Meng, Y. Zhang, J. Huang, X. Wang, Y. Zhang, H. Ji, and J. Han. Distantly-supervised named entity recognition with noise-robust learning and language model augmented self-training. In *Proceedings of the 2021 Conference on*

Empirical Methods in Natural Language Processing, pages 10367–10378, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.810. URL `https://aclanthology.org/2021.emnlp-main.810`.

[131] Z. Meng, F. Liu, T. H. Clark, E. Shareghi, and N. Collier. Mixture-of-partitions: Infusing large biomedical knowledge graphs into BERT. *arXiv preprint arXiv:2109.04810*, 2021.

[132] Z. Meng, F. Liu, E. Shareghi, Y. Su, C. Collins, and N. Collier. Rewire-then-probe: A contrastive recipe for probing biomedical knowledge of pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4798–4810, 2022.

[133] G. Michalopoulos, Y. Wang, H. Kaka, H. Chen, and A. Wong. Umlsbert: Clinical domain knowledge augmentation of contextual embeddings using the unified medical language system metathesaurus. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1744–1753, 2021.

[134] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *Proceedings of the 1st International Conference on Learning Representations (ICLR)*, Scottsdale, Arizona, USA, May 2013.

[135] M. Mintz, S. Bills, R. Snow, and D. Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore, Aug. 2009. Association for Computational Linguistics. URL `https://aclanthology.org/P09-1113`.

[136] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6(1):1–10, 2016.

[137] A. Mitra, P. Banerjee, K. K. Pal, S. Mishra, and C. Baral. Exploring ways to incorporate additional knowledge to improve natural language commonsense question answering. *CoRR*, abs/1909.08855, 2019. URL http://arxiv.org/abs/1909.08855.

[138] A. Mitra, C. Baral, A. Bhattacharjee, and I. Shrivastava. A generate-validate approach to answering questions about qualitative relationships. *arXiv preprint arXiv:1908.03645*, 2019.

[139] S. Montani, R. Bellazzi, L. Portinale, S. Fiocchi, and M. Stefanelli. A case-based retrieval system for diabetic patients therapy. *Proceedings of IDAMAP*, 98:64–70, 1998.

[140] C. Mugisha and I. Paik. Pneumonia outcome prediction using structured and unstructured data from ehr. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2640–2646. IEEE, 2020.

[141] C. Mugisha and I. Paik. Comparison of neural language modeling pipelines for outcome prediction from unstructured medical text notes. *IEEE Access*, 10: 16489–16498, 2022.

[142] J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, and J. Eisenstein. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1100. URL https://aclanthology.org/N18-1100.

[143] N. Naganure, N. U. Ashwin, and S. S. Kamath. Leveraging deep learning approaches for patient case similarity evaluation. In *Intelligent Data Engineering and Analytics*, pages 613–622. Springer, 2021.

[144] A. Naik, A. Ravichander, N. Sadeh, C. Rose, and G. Neubig. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, 2018.

[145] A. Naik, S. Parasa, S. Feldman, L. L. Wang, and T. Hope. Literature-augmented clinical outcome prediction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 438–453, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.33. URL `https://aclanthology.org/2022.findings-naacl.33`.

[146] N. Nangia, S. Sugawara, H. Trivedi, A. Warstadt, C. Vania, and S. Bowman. What ingredients make for an effective crowdsourcing protocol for difficult nlu data collection tasks? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1221–1235, 2021.

[147] U. Naseem, S. K. Khan, I. Razzak, and I. A. Hameed. Hybrid words representation for airlines sentiment analysis. In *Australasian Joint Conference on Artificial Intelligence*, pages 381–392. Springer, 2019.

[148] V. Nastase, S. Szpakowicz, P. Nakov, and D. Ó. Séagdha. Semantic relations between nominals. *Synthesis lectures on human language technologies*, 14(1): 1–234, 2021.

[149] A. Niam, A. Das, and S. Haque. Performance analysis and implementation of pre-trained model using transfer learning on bangla document clustering. In

Proceedings of International Conference on Fourth Industrial Revolution and Beyond 2021, pages 659–671. Springer, 2022.

[150] Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, and D. Kiela. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.441. URL `https://www.aclweb.org/anthology/2020.acl-main.441`.

[151] R. Nogueira and K. Cho. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*, 2019.

[152] B. Oğuz, K. Lakhotia, A. Gupta, P. Lewis, V. Karpukhin, A. Piktus, X. Chen, S. Riedel, W.-t. Yih, S. Gupta, et al. Domain-matched pre-training tasks for dense retrieval. *arXiv preprint arXiv:2107.13602*, 2021.

[153] O. A. Pandit and Y. Hou. Probing for bridging inference in transformer language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4153–4163, 2021.

[154] D. Paperno, G. Kruszewski, A. Lazaridou, N.-Q. Pham, R. Bernardi, S. Pezzelle, M. Baroni, G. Boleda, and R. Fernández. The lambada dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, 2016.

[155] A. Paramasivam and S. J. Nirmala. A survey on textual entailment based question answering. *Journal of King Saud University-Computer and Information Sciences*, 2021.

[156] S. Park and C. Caragea. Scientific keyphrase identification and classification by pre-trained language models intermediate task transfer learning. In *Pro-*

ceedings of the 28th International Conference on Computational Linguistics, pages 5409–5419, Barcelona, Spain (Online), Dec. 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.472. URL `https://aclanthology.org/2020.coling-main.472`.

[157] N. Pattisapu, V. Anand, S. Patil, G. Palshikar, and V. Varma. Distant supervision for medical concept normalization. *Journal of biomedical informatics*, 109: 103522, 2020.

[158] X. Peng, G. Long, T. Shen, S. Wang, and J. Jiang. Sequential diagnosis prediction with transformer and ontological representation. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 489–498. IEEE, 2021.

[159] Y. Peng, S. Yan, and Z. Lu. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, 2019.

[160] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[161] G. Pergola, E. Kochkina, L. Gui, M. Liakata, and Y. He. Boosting low-resource biomedical qa via entity-aware masking strategies. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1977–1985, 2021.

[162] M. E. Peters, W. Ammar, C. Bhagavatula, and R. Power. Semi-supervised sequence tagging with bidirectional language models. *arXiv preprint arXiv:1705.00108*, 2017.

[163] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL `https://aclanthology.org/N18-1202`.

[164] F. Petroni, T. Rocktäschel, P. Lewis, A. Bakhtin, Y. Wu, A. H. Miller, and S. Riedel. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019.

[165] F. Petroni, T. Rocktäschel, S. Riedel, P. S. H. Lewis, A. Bakhtin, Y. Wu, and A. H. Miller. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2463–2473, 2019.

[166] J. Phang, T. Févry, and S. R. Bowman. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*, 2018.

[167] J. Phang, I. Calixto, P. M. Htut, Y. Pruksachatkun, H. Liu, C. Vania, K. Kann, and S. R. Bowman. English intermediate-task training improves zero-shot cross-lingual transfer too. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 557–575, Suzhou, China, Dec. 2020. Association for Computational Linguistics. URL `https://aclanthology.org/2020.aacl-main.56`.

[168] A. Poliak, J. Naradowsky, A. Haldar, R. Rudinger, and B. Van Durme. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, 2018.

[169] C. Poth, J. Pfeiffer, A. Rücklé, and I. Gurevych. What to pre-train on? efficient intermediate task selection. *arXiv preprint arXiv:2104.08247*, 2021.

[170] Y. Pruksachatkun, J. Phang, H. Liu, P. M. Htut, X. Zhang, R. Y. Pang, C. Vania, K. Kann, and S. R. Bowman. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.467. URL `https://aclanthology.org/2020.acl-main.467`.

[171] M. Purver and S. Battersby. Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 482–491, 2012.

[172] T. Qin. *Dual Learning*. Springer, 2020.

[173] Y. Qu, Y. Ding, J. Liu, K. Liu, R. Ren, W. X. Zhao, D. Dong, H. Wu, and H. Wang. Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5835–5847, 2021.

[174] O. Ram, Y. Kirstain, J. Berant, A. Globerson, and O. Levy. Few-shot question answering by pretraining span selection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3066–3079, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.239. URL `https://aclanthology.org/2021.acl-long.239`.

[175] N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[176] D. Ribeiro, T. Hinrichs, M. Crouse, K. Forbus, M. Chang, and M. Witbrock. Predicting state changes in procedural text using analogical question answering. In *7th Annual Conference on Advances in Cognitive Systems*, 2019.

[177] S. Riedel, L. Yao, and A. McCallum. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer, 2010.

[178] A. Roberts, C. Raffel, and N. Shazeer. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 5418–5426, 2020.

[179] A. Roberts, C. Raffel, and N. Shazeer. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, 2020.

[180] K. Roberts, D. Demner-Fushman, E. M. Voorhees, and W. R. Hersh. Overview of the TREC 2016 clinical decision support track. In E. M. Voorhees and A. Ellis, editors, *Proceedings of The Twenty-Fifth Text REtrieval Conference, TREC 2016, Gaithersburg, Maryland, USA, November 15-18, 2016*, volume 500-321 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2016. URL `http://trec.nist.gov/pubs/trec25/papers/Overview-CL.pdf`.

[181] K. Roberts, M. Simpson, D. Demner-Fushman, E. Voorhees, and W. Hersh. State-of-the-art in biomedical literature retrieval for clinical cases: a survey of the trec 2014 cds track. *Information Retrieval Journal*, 19(1):113–148, 2016.

[182] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR94*, pages 232–241. Springer, 1994.

[183] A. Romanov and C. Shivade. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1187. URL `https://aclanthology.org/D18-1187`.

[184] A. Romanov and C. Shivade. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels*, pages 1586–1596, 2018.

[185] B. Roth, T. Barth, M. Wiegand, and D. Klakow. A survey of noise reduction methods for distant supervision. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 73–78, 2013.

[186] K. Rudra and A. Anand. Distant supervision in bert-based adhoc document retrieval. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2197–2200, 2020.

[187] W. B. Schwartz, G. A. Gorry, J. P. Kassirer, and A. Essig. Decision analysis and clinical judgment. *The American journal of medicine*, 55(4):459–472, 1973.

[188] J. Shang, T. Ma, C. Xiao, and J. Sun. Pre-training of graph augmented transformers for medication recommendation. In *28th International Joint Conference on Artificial Intelligence, IJCAI 2019*, pages 5953–5959. International Joint Conferences on Artificial Intelligence, 2019.

[189] Y. Shao, B. Liu, J. Mao, Y. Liu, M. Zhang, and S. Ma. Thuir@ coliee-2020: Leveraging semantic understanding and exact matching for legal case retrieval and entailment. *arXiv preprint arXiv:2012.13102*, 2020.

[190] Y. Shao, J. Mao, Y. Liu, W. Ma, K. Satoh, M. Zhang, and S. Ma. BERT-PLI: Modeling paragraph-level interactions for legal case retrieval. In *IJCAI*, pages 3501–3507, 2020.

[191] S. Sharma, B. Santra, A. Jana, S. Tokala, N. Ganguly, and P. Goyal. Incorporating domain knowledge into medical NLI using knowledge graphs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6092–6097, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1631. URL `https://aclanthology.org/D19-1631`.

[192] T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace, and S. Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*, 2020.

[193] C. Shorten, T. M. Khoshgoftaar, and B. Furht. Deep learning applications for covid-19. *Journal of big Data*, 8(1):1–54, 2021.

[194] V. Shwartz, P. West, R. L. Bras, C. Bhagavatula, and Y. Choi. Unsupervised commonsense question answering with self-talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4615–4629, 2020.

[195] N. A. Smith. Contextual word representations: Putting words into computers. *Commun. ACM*, 63(6):66–74, may 2020. ISSN 0001-0782. doi: 10.1145/3347145. URL `https://doi.org/10.1145/3347145`.

[196] L. Soldaini and N. Goharian. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, sigir*, pages 1–4, 2016.

[197] S. Soni and K. Roberts. Evaluation of dataset selection for pre-training and fine-tuning transformer language models for clinical question answering. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5532–5538, 2020.

[198] R. Speer, J. Chin, and C. Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

[199] M. Q. Stearns, C. Price, K. A. Spackman, and A. Y. Wang. Snomed clinical terms: overview of the development process and project status. In *Proceedings of the AMIA Symposium*, page 662. American Medical Informatics Association, 2001.

[200] C. Su, S. Gao, and S. Li. Gate: graph-attention augmented temporal neural network for medication recommendation. *IEEE Access*, 8:125447–125458, 2020.

[201] P. Su, G. Li, C. Wu, and K. Vijay-Shanker. Using distant supervision to augment manually annotated data for relation extraction. *PloS one*, 14(7):e0216913, 2019.

[202] J. Sun and S. Li. Domain adaptation for medical semantic textual similarity. In *2021 7th IEEE International Conference on Network Intelligence and Digital Content (IC-NIDC)*, pages 319–323, 2021. doi: 10.1109/IC-NIDC54101.2021.9660484.

[203] M. Sung, J. Lee, S. Yi, M. Jeon, S. Kim, and J. Kang. Can language models be biomedical knowledge bases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[204] M. Sushil, S. Suster, and W. Daelemans. Are we there yet? exploring clinical domain knowledge of BERT models. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 41–53, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.bionlp-1.5. URL `https://aclanthology.org/2021.bionlp-1.5`.

[205] J. Suttles and N. Ide. Distant supervision for emotion classification with discrete binary values. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 121–136. Springer, 2013.

[206] J. Tabassum, A. Ritter, and W. Xu. Tweetime: A minimally supervised method for recognizing and normalizing time expressions in twitter. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 307–318, 2016.

[207] S. Taewijit, T. Theeramunkong, and M. Ikeda. Distant supervision with transductive learning for adverse drug reaction identification from electronic medical records. *Journal of healthcare engineering*, 2017, 2017.

[208] A. Talmor, Y. Elazar, Y. Goldberg, and J. Berant. oLMpics - on what language model pre-training captures. *Trans. Assoc. Comput. Linguistics*, 8:743–758, 2020.

[209] N. Tawfik and M. Spruit. UU_TAILS at MEDIQA 2019: Learning textual entailment in the medical domain. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 493–499, Florence, Italy, Aug. 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5053. URL `https://aclanthology.org/W19-5053`.

[210] N. S. Tawfik and M. R. Spruit. Evaluating sentence representations for biomedical text: Methods and experimental results. *Journal of Biomedical Informatics*, page 103396, 2020.

[211] I. Tenney, P. Xia, B. Chen, A. Wang, A. Poliak, R. T. McCoy, N. Kim, B. Van Durme, S. R. Bowman, D. Das, et al. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*, 2019.

[212] Z. Tian, Q. Liu, M. Liu, and W. Deng. Simple flow-based contrastive learning for bert sentence representations. In *International Conference on Sensing and Imaging*, pages 265–275. Springer, 2022.

[213] S. Tsevas and D. K. Iakovidis. Fusion of multimodal temporal clinical data for the retrieval of similar patient cases. In *2011 10th International Workshop on Biomedical Engineering*, pages 1–4. IEEE, 2011.

[214] A. Upadhyay, S. Massie, and S. Clogher. Case-based approach to automated natural language generation for obituaries. In *International Conference on Case-Based Reasoning*, pages 279–294. Springer, 2020.

[215] B. van Aken, J.-M. Papaioannou, M. Mayrdorfer, K. Budde, F. Gers, and A. Loeser. Clinical outcome prediction from admission notes using self-supervised knowledge integration. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 881–893, Online, Apr. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.75. URL `https://aclanthology.org/2021.eacl-main.75`.

[216] S. Vashishth, D. Newman-Griffis, R. Joshi, R. Dutt, and C. P. Rosé. Improving broad-coverage medical entity linking with semantic type prediction and large-scale datasets. *Journal of Biomedical Informatics*, 121:103880, 2021.

[217] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[218] V. Verdhan. Supervised learning with python. *Okänd. Irland: Apress*, 2020.

[219] D. Vilares and C. Gómez-Rodríguez. HEAD-QA: A healthcare dataset for complex reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 960–966, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1092. URL `https://aclanthology.org/P19-1092`.

[220] D. Vrandečić and M. Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.

[221] T. Vu, M.-T. Luong, Q. V. Le, G. Simon, and M. Iyyer. Strata: Self-training with task augmentation for better few-shot learning. *arXiv preprint arXiv:2109.06270*, 2021.

[222] I. Vulić, S. Baker, E. M. Ponti, U. Petti, I. Leviant, K. Wing, O. Majewska, E. Bar, M. Malone, T. Poibeau, et al. Multi-simlex: A large-scale evaluation of multilingual and crosslingual lexical semantic similarity. *Computational Linguistics*, 46(4):847–897, 2020.

[223] I. Vulić, E. M. Ponti, R. Litschko, G. Glavaš, and A. Korhonen. Probing pre-trained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, 2020.

[224] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, 2018.

[225] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems*, 32, 2019.

[226] B. Wang, Q. Xie, J. Pei, P. Tiwari, Z. Li, et al. Pre-trained language models in biomedical domain: A systematic survey. *arXiv preprint arXiv:2110.05006*, 2021.

[227] C. Wang, P. Nulty, and D. Lillis. A comparative study on word embeddings in deep learning for text classification. In *Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval*, pages 37–46, 2020.

[228] K. Wang, N. Reimers, and I. Gurevych. Tsdae: Using transformer-based sequential denoising auto-encoder for unsupervised sentence embedding learning. *arXiv preprint arXiv:2104.06979*, 2021.

[229] S. Wang, W. Zhou, and C. Jiang. A survey of word embeddings based on deep learning. *Computing*, 102(3):717–740, 2020.

[230] X. D. Wang, U. Leser, and L. Weber. BEEDS: Large-scale biomedical event extraction using distant supervision and question answering. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 298–309, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.bionlp-1.28. URL `https://aclanthology.org/2022.bionlp-1.28`.

[231] Y. Wang, S. Fu, F. Shen, S. Henry, O. Uzuner, H. Liu, et al. The 2019 n2c2/ohnlp track on clinical semantic textual similarity: overview. *JMIR medical informatics*, 8(11):e23375, 2020.

[232] Y. Wang, Y. Hou, W. Che, and T. Liu. From static to dynamic word representations: a survey. *International Journal of Machine Learning and Cybernetics*, 11 (7):1611–1630, 2020.

[233] Z. Wang, R. Wen, X. Chen, S. Cao, S.-L. Huang, B. Qian, and Y. Zheng. Online disease diagnosis with inductive heterogeneous graph convolutional networks. In *Proceedings of the Web Conference 2021*, pages 3349–3358, 2021.

[234] Z. Wang, Y. Yang, R. Wen, X. Chen, S.-L. Huang, and Y. Zheng. Lifelong learning based disease diagnosis on clinical notes. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 213–224. Springer, 2021.

[235] T. Wanyan, H. Honarvar, A. Azad, Y. Ding, and B. S. Glicksberg. Deep learning with heterogeneous graph embeddings for mortality prediction from electronic health records. *Data Intelligence*, 3(3):329–339, 2021.

[236] A. Warstadt and S. R. Bowman. Can neural networks acquire a structural bias from raw linguistic data? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2020.

[237] M. W. Wartofsky. Clinical judgment, expert programs, and cognitive style: a counter-essay in the logic of diagnosis. *The Journal of medicine and philosophy*, 11(1):81–92, 1986.

[238] H. Westermann, J. Savelka, and K. Benyekhlef. Paragraph similarity scoring and fine-tuned BERT for legal information retrieval and entailment. In *JSAI International Symposium on Artificial Intelligence*, pages 269–285. Springer, 2020.

[239] T. Wijesiriwardene, V. Nguyen, G. Bajaj, H. Y. Yip, V. Javangula, Y. Mao, K. W. Fung, S. Parthasarathy, A. P. Sheth, and O. Bodenreider. Ubert: A novel language model for synonymy prediction at scale in the umls metathesaurus. *arXiv preprint arXiv:2204.12716*, 2022.

[240] T.-L. Wu, S. Singh, S. Paul, G. Burns, and N. Peng. Melinda: A multimodal dataset for biomedical experiment method classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14076–14084, 2021.

[241] Z. Wu, Y. Song, S. Huang, Y. Tian, and F. Xia. WTMED at MEDIQA 2019: A hybrid approach to biomedical natural language inference. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 415–426, 2019.

[242] J. Xu and W. B. Croft. Quary expansion using local and global document analysis. In *Acm sigir forum*, volume 51, pages 168–175. ACM New York, NY, USA, 2017.

[243] Y. Yang, N. Jin, K. Lin, M. Guo, and D. Cer. Neural retrieval for question answering with cross-attention supervised data augmentation. In *Proceedings*

of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 263–268, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.35. URL `https://aclanthology.org/2021.acl-short.35`.

[244] M. Yasunaga, A. Bosselut, H. Ren, X. Zhang, C. D. Manning, P. Liang, and J. Leskovec. Deep bidirectional language-knowledge graph pretraining. In *Neural Information Processing Systems (NeurIPS)*, 2022.

[245] M. Yasunaga, J. Leskovec, and P. Liang. LinkBERT: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8003–8016, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.551. URL `https://aclanthology.org/2022.acl-long.551`.

[246] W. Yu, X. Guo, F. Chen, T. Chang, M. Wang, and X. Wang. Similar questions correspond to similar sql queries: A case-based reasoning approach for text-to-sql translation. In *International Conference on Case-Based Reasoning*, pages 294–308. Springer, 2021.

[247] Z. Yuan, Y. Liu, C. Tan, S. Huang, and F. Huang. Improving biomedical pretrained language models with knowledge. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 180–190, 2021.

[248] D. Zhang, S. Mohan, M. Torkar, and A. McCallum. A distant supervision corpus for extracting biomedical relationships between chemicals, diseases and genes. *arXiv preprint arXiv:2204.06584*, 2022.

[249] T. Zhang, M. Chen, and A. A. Bui. Diagnostic prediction with sequence-of-sets representation learning for clinical events. In *International Conference on Artificial Intelligence in Medicine*, pages 348–358. Springer, 2020.

[250] T. Zhang, Z. Cai, C. Wang, M. Qiu, B. Yang, and X. He. SMedBERT: A knowledge-enhanced pre-trained language model with structured semantics for medical text mining. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5882–5893, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.457. URL `https://aclanthology.org/2021.acl-long.457`.

[251] W. Zhang, H. Lin, X. Han, L. Sun, H. Liu, Z. Wei, and N. Yuan. Denoising distantly supervised named entity recognition via a hypergeometric probabilistic model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14481–14488, 2021.

[252] X. Zhang, C. Xiao, L. M. Glass, and J. Sun. Deepenroll: Patient-trial matching with deep embedding and entailment prediction. In *Proceedings of The Web Conference 2020*, WWW '20, page 1029–1037, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370233. doi: 10.1145/3366423.3380181. URL `https://doi.org/10.1145/3366423.3380181`.

[253] X. Zhang, A. Bosselut, M. Yasunaga, H. Ren, P. Liang, C. D. Manning, and J. Leskovec. GreaseLM: Graph reasoning enhanced language models for question answering. *arXiv preprint arXiv:2201.08860*, 2022.

[254] Y. Zhang, H. Fei, and P. Li. Readsre: Retrieval-augmented distantly supervised relation extraction. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2257–2262, 2021.

[255] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu. Ernie: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129*, 2019.

[256] C. Zhao, C. Xiong, J. Boyd-Graber, and H. Daumé III. Distantly-supervised dense retrieval enables open-domain question answering without evidence annotation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9612–9622, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.756. URL `https://aclanthology.org/2021.emnlp-main.756`.

[257] M. Zhao, P. Dufter, Y. Yaghoobzadeh, and H. Schütze. Quantifying the contextualization of word representations with semantic class probing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1219–1234, 2020.

[258] X. Zhou, Y. Zhang, L. Cui, and D. Huang. Evaluating commonsense in pre-trained language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9733–9740, 2020.

[259] M. Zhu, A. Ahuja, W. Wei, and C. K. Reddy. A hierarchical attention retrieval model for healthcare question answering. In *The World Wide Web Conference*, pages 2472–2482, 2019.