Interpreting Patient Descriptions using Distantly Supervised Similar Case Retrieval

Israa Alghanmi
Cardiff University
United Kingdom
alghanmiia@cardiff.ac.uk

Luis Espinosa-Anke
Cardiff University
United Kingdom
espinosa-ankel@cardiff.ac.uk

Steven Schockaert
Cardiff University
United Kingdom
schockaerts1@cardiff.ac.uk

ABSTRACT

Biomedical natural language processing often involves the interpretation of patient descriptions, for instance for diagnosis or for recommending treatments. Current methods, based on biomedical language models, have been found to struggle with such tasks. Moreover, retrieval augmented strategies have only had limited success, as 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 propose to rely on a nearest neighbour strategy. First, we retrieve 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 judge the likelihood of the hypothesis based on the similarity of the retrieved passages. Identifying similar cases is challenging, however, 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 propose a strategy that relies on a distantly supervised cross-encoder. Despite its conceptual simplicity, we find this strategy to be effective in practice.

CCS CONCEPTS

• **Applied computing** → **Life and medical sciences**; • **Information systems** → **Information retrieval**; • **Computing methodologies** → **Natural language processing**.

KEYWORDS

Biomedical NLP, Similar Case Retrieval, Distant Supervision

ACM Reference Format:
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), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3477495.3532003

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SIGIR '22, July 11–15, 2022, Madrid, Spain
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3532003

1 INTRODUCTION

An important challenge for biomedical natural language processing (NLP) is to make inferences about patient descriptions. For instance, given a 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 medication. Table 1 shows an example from a multiple-choice QA evaluation dataset to illustrate this setting. To support biomedical NLP, several versions of BERT [17] that were adapted to the biomedical domain have been introduced, including ClinicalBERT [3], SciBERT [9], BioBERT [35] and PubMedBERT [21]. As standard language models (LMs) are able to make various factual and commonsense inferences [16, 49, 76], one might expect these biomedical LMs to be similarly capable of tasks such as inferring diagnoses from symptoms. Prior work, however, has shown that existing biomedical LMs often struggle with such tasks. For instance, Alghanmi et al. [2] found that the standard BERT model was remarkably competitive with specialised biomedical LMs for inferring diagnoses from patient descriptions. Meng et al. [44] furthermore introduced a probing task for evaluating the knowledge captured by biomedical LMs, which also revealed significant issues.

To alleviate the limitations of biomedical LMs, a natural strategy would be to augment patient 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 [29, 45, 60]. When it comes to interpreting patient descriptions, however, the potential of such strategies is less clear. For instance, Sushil et al. [62] 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 [55], they found no statistically significant improvements as a result

Table 1: Example of a question from MedQA, along with the answer candidates.

Question: 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?
(A) Medullary sponge kidney
(B) Simple renal cysts
(C) Autosomal dominant polycystic kidney disease
(D) Autosomal recessive polycystic kidney disease

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 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 [56, 68].

Rather than searching for sentences that directly express medical knowledge, we aim to find passages that are similar to the given patient 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 description and the retrieved passages. The use of similar cases plays an important role in clinical decision making [6, 8, 46, 63], 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 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 descriptions may differ in the level of detail they provide, as well as their overall writing style. To illustrate these issues, Table 2 shows the top passage that was retrieved by our model for a given question from the MedQA benchmark [27]. As can be seen, both patient 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 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 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 descriptions inherently different from settings such as open-domain QA, where gold annotations of relevant passages are often available and systems can rely on transfer learning from closely related tasks.

In this paper, we propose a distant supervision strategy to address these challenges. We start from the intuition that interpreting patient 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 descriptions mentioning *brief psychotic disorder* would tell us something about the likelihood that this is the correct diagnosis for the question in Table 2. 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 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 descriptions, i.e. a fine-tuned language model which takes two patient 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 description. Conceptually, this is similar in spirit to distant supervision strategies for open-domain QA (see Section 2). 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 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 descriptions because it was trained in this domain.
- The initial set of candidate passages is then ranked using a pre-trained cross-encoder. We initialise this cross-encoder

Table 2: Example of a question from MedQA, along with the top-retrieved passage by our model for the answer candidate *brief psychotic disorder*.

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.

from a biomedical LM and pre-train it on a standard textual similarity dataset. Despite not being trained on patient descriptions, we show that this re-ranking step improves 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.

Our experimental results show that our overall approach is highly effective, improving the state-of-the-art for question answering about patient descriptions [27]. [1]

2 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 paper, several retrieval models that combine distant supervision with BERT-based encodings have been proposed in recent years. For instance, Karpukhin et al. [30] 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 [54]. 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 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 [50, 71]. 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. [32] start from a pre-trained ColBERT model [33] 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 descriptions. The aforementioned methods rely on a baseline retrieval model to generate pseudo-labels, which is also the approach we follow in this paper. As an alternative, some authors have also proposed models in which the retrieval model is jointly optimised with the rest of the QA model [22, 36]. 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 [25] and for ad-hoc document retrieval [42], among many others.

Knowledge-Enhanced LMs. Various strategies have been proposed for improving the amount of knowledge that is captured by transformer-based language models. One common approach is to rely on some kind of knowledge infusion while training the model [75] or during the fine-tuning phase [19, 38]. For the biomedical domain, He et al. [23] proposed a pre-training objective that aims to infuse disease knowledge by exploiting the structure of Wikipedia pages about diseases. Yuan et al. [73] pre-trained a language model with entity extraction and linking objectives based on UMLS [11], while Zhang et al. [74] also used structured knowledge about entities and their relations for pre-training. Meng et al. [43] introduced a method for infusing knowledge from large biomedical knowledge graphs. Instead of improving the language model itself, some authors have also explored the possibility of combining contextualised embeddings with static vector representations of biomedical concepts, e.g. obtained from UMLS knowledge graph embeddings [59]. Most relevant to our work, some approaches augment questions with knowledge expressed in textual form. For instance, Lu et al. [40] 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 questions with textual knowledge is limited in the biomedical domain. For instance, Sushil et al. [62] evaluated the effect of such augmentation strategies and failed to obtain any statistically significant improvements for MedNLI [55], a well-known benchmark for Natural Language Inference (NLI) in the biomedical domain. These findings were also corroborated by our own initial analysis.

Similar Case Retrieval. Within NLP, similar case retrieval has primary been applied to the analysis of legal cases. For instance, Westermann et al. [69] 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. [58] 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. [57] 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 [5].

[1] Source code to replicate our experiments is available at: https://github.com/israa-alghanmi/PD-SimilarCase

Beyond the legal domain, the idea of exploiting similar cases has recently been used for question answering [51], semantic parsing [72], text generation [64] and language modelling [31] 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 [24, 26, 47], 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 [52, 53].

3 PROPOSED METHOD

We are interested in the problem of interpreting patient descriptions. More specifically, given a patient 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 2, the question corresponds to the patient 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 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 C_H to assess the likelihood that H is entailed by $\mathcal{D}$. The underlying intuition is that C_H and $\mathcal{D}$ are both presumed to be patient descriptions, and moreover, that the fact that H is mentioned in C_H means that H can be inferred from that patient description.

Our central aim is to demonstrate the strong potential of nearest neighbour strategies for interpreting patient 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 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 description $\mathcal{D}$ will themselves also be patient descriptions.
- We take the fact that H is mentioned in the text fragment 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 [7] or QuickUMLS [61] 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 C_H. However, such solutions may themselves introduce errors. Furthermore, as we will see, sometimes passages are retrieved that are not patient descriptions but which nonetheless help the model to make the correct prediction. We can often think of such passages as being generic patient descriptions, e.g. discussing how a given illness in general presents itself, hence specifically restricting the retrieved passages to actual patient 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 3.2 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 3.3 describes how the cross-encoder is used as part of our overall model.

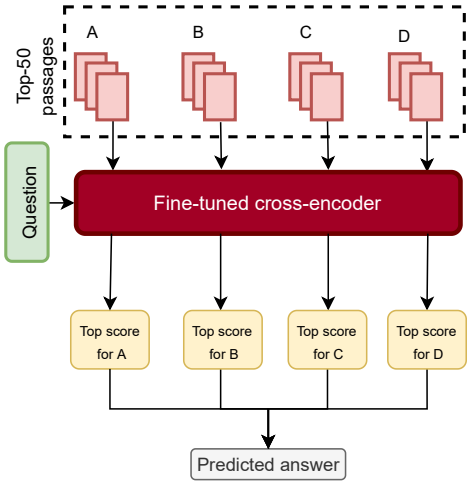

Figure 1: Overview of the application of our proposed model for answering multiple-choice questions.

3.1 Overview of the Nearest Neighbour Strategy

Let $\mathcal{D}$ be a patient description and let C_H be a text fragment mentioning some hypothesis of interest H. We want to train a model that allows us to predict whether 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 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 more scalable. For this reason, in line with the standard usage of cross-encoders as re-rankers in information retrieval [15, 37, 48], 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 1 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 C_A, C_B, C_C and C_D. For instance, 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 C_A is estimated to be more similar to $\mathcal{D}$ than C_B, C_C and C_D.

3.2 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 description and H is a hypothesis that can be inferred from $\mathcal{D}$. Similarly, we assume we have a

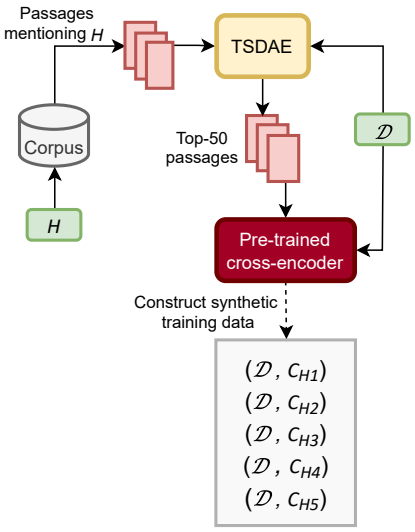

Figure 2: Overview of how the distantly supervised examples for training the cross-encoder are obtained (shown for $k = 5$).

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 $(\mathcal{D}, C_H)$, where 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 $(\mathcal{D}, C_H)$ 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 [12, 13, 34, 70], 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 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 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 2 .

3.2.1 Initial Retrieval Step. Given a pair $(\mathcal{D}, H)$ we first use Elasticsearch [20] 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 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 [67] to train a sentence embedding model for the clinical domain.

We initialize this model from ClinicalBERT and use MIMIC-III [28] 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.

3.2.2 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)** [14]: 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** [77]: A set of question and answer pairs annotated with relevance labels. The answers were collected from the Patient website[2] 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 [41]. We also include HealthQA because of its structural similarity with our considered setting. Note that none of these pre-training tasks involve patient descriptions, while STS-B is not even focused on the biomedical domain.

3.3 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 3.2.2. 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 C_{H_i}. We predict the answer candidate H_i

[2] https://patient.info/

for which the similarity between $\mathcal{D}$ and 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 $C_{H_i} = H_i$, i.e. we compute the similarity between $\mathcal{D}$ and H_i instead.

4 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 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?

4.1 Evaluation Datasets

We evaluate our method on the following datasets.

MedQA. [27]: A multiple-choice question answering dataset that is derived from medical exams. We use the USMLE variant, which is the English version of the dataset. This dataset allows for the most direct evaluation of our proposed strategy, as it specifically focuses on the problem of interpreting patient descriptions. An example of a question from this dataset can be found in Table 1.

DisKnE. [2]: A binary classification task, where instances consist of a patient description and a disease name, and the aim is to decide whether it can be inferred that the patient has the disease. This dataset was derived from MedNLI [55]. We use DisKnE in our evaluation, rather than the original MedNLI dataset, for two main reasons. First, it prevents the model from learning medical knowledge about a given disease during training, by avoiding overlap between the disease covered by the test data and the diseases covered by the training data. Specifically, it considers a separate training-test split for each disease. Reported results are averaged across these different splits. Second, DisKnE specifically focuses on those MedNLI instances that require interpreting patient descriptions, whereas MedNLI also covers instances that require terminological inferences, among others (e.g. expanding acronyms used in the patient description). We use the medical-similar version of the benchmark, where negative examples were obtained from positive examples by replacing the disease by a similar one.

Head-QA. [66]: A multiple-choice question answering dataset that covers questions about different areas within the healthcare domain, such as medicine, psychology and biology. We use the English version of the dataset. Some questions correspond to patient descriptions, but the majority are about recalling specific factual knowledge. An example of a question from this dataset is as follows:

> **Question:** The fibrocartilage is located in:
> **Answer:** Intervertebral discs

Some of the questions in this dataset require multi-modal reasoning, combining information from the question with an associate image. As this goes beyond the scope of our paper, in our experiments we have excluded all questions which have an associated image. The main reason for including this dataset is because it allows us to explore to what extent the proposed methodology can be effective in a broader setting than for interpreting patient descriptions.

4.2 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 paper, we focus on corpora that contain patient descriptions. We have, in particular, used the following two corpora, both of which are widely used in biomedical NLP.

WikiMed and PubMedDS (Wiki-PubMed) [65]. This dataset contains 393,618 Wikipedia articles (being those that mention some UMLS concept) as well as 13,197,430 PubMed abstracts. We split the documents 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 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 descriptions and to what extent it is able to exploit generic descriptions.

MIMIC-III [28]. We use the discharge summaries from MIMIC-III, which is a database of records about patients that were admitted to the critical care unit of a large hospital. 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 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 descriptions are also taken from the MIMIC-III corpus. This allows us to experiment with a setting where the corpus contains patient descriptions that are written in a similar style as the target description. Note, however, that the patient descriptions from DisKnE themselves are never retrieved by our method, as the corresponding hypotheses are not mentioned in the original notes.

4.3 Pre-trained Language Models

We experiment with four pre-trained LMs to initialize the cross-encoder: the cased version of $\text{BERT}_{\text{base}}$ [18]; the version of ClinicalBERT [4] that was initialized from BioBERT [35] and further pre-trained on MIMIC-III; the cased version of SciBERT [10], which

was trained from scratch on scientific articles; the version of PubMedBERT [21] that was trained from scratch on PubMed abstracts and full-length medical articles.

4.4 Baselines

We consider the following baselines.

Standard Fine-tuning (FT). We fine-tune a pre-trained language model to predict whether a given hypothesis can be entailed from a patient description, as in standard NLI models. Specifically, we concatenate the patient 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.

Definitions. We use QuickUMLS [61] 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 [59] for improving LSTM-based models. We refer to this strategy as *BERT-Def*, and similar for the other LMs.

Unsupervised Retrieval. Finally, we also report results for unsupervised retrieval models. In this case, we simply compute the similarity degree between the patient 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.

4.5 Evaluation Metrics

For MedQA and HeadQA, we solve the standard multiple-choice QA task as explained in Section 3.3, 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).

4.6 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.

Table 3: Results for DisKnE in terms of Mean Average Precision (MAP). The best results for each language model are shown in bold.

		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 4: 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.

	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

4.7 Results

The experimental results are summarized in Table 5 for MedQA, Table 3 for DisKnE and Table 6 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 4.

For MedQA (Table 5), 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 consistently outperform the baseline results (for these language models) in Table 4, across all choices of k and each of the three pre-training tasks. The results also clearly outperform 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

Table 5: Results for MedQA in terms of Average Precision (AP) and Accuracy (Acc). The best results for each language model are shown in bold.

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

		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

the most relevant paragraphs, even when using the pre-trained encoder, this problem of training data becomes 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 for MedQA are those from Meng et al. [43], 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 Table 5.

For HeadQA (Table 6), 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 descriptions. To explore this further, we manually split the test set into those questions which are about patient descriptions (216 in total) and those which are not (2458 in total). Table 7 shows the results obtained for these two sets of questions, for the SciBERT-FT and PubMedBERT-FT baselines, as well as our

Table 7: Analysis of HeadQA results, where test questions were split depending on whether or not they are about patient descriptions. Results are reported in terms of average precision and accuracy.

	Patient 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 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).

	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

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 [39].

For DisKnE (Table 3), 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. The only published results for DisKnE, to the best of our knowledge, are those from the original paper, where the focus was on comparing different language models, i.e. they only reported results for the standard fine-tuning baselines.

Comparing the baseline results in Table 4, 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 descriptions intuitively requires more than surface-level matching.

4.8 Analysis

Ablation Results. In Table 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.
- *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.

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 8, the *Pretrained CE* model outperforms the unsupervised baseline retrieval models in Table 4. In fact, For MedQA and HeadQA, the results of this unsupervised model 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.

Table 9: Example of a correctly answered question from the MedQA test set in which the retrieved passage is not a patient description.

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".

Qualitative Analysis. We manually analysed the retrieved passages for MedQA and HeadQA with Wiki-PubMed. Our main findings can be summarized as follows. First, we found that in many cases, the retrieved passages were indeed patient descriptions. This is somewhat surprising, given that only a small fragment of Wiki-PubMed consists of patient 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 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 descriptions.

5 CONCLUSIONS

We have proposed a nearest neighbour strategy for interpreting patient 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 descriptions can be overcome, at least to some

extent, by using distant supervision strategies. We highlighted, in particular, that the setting of patient 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 descriptions provides us with reasonably reliable evidence that these descriptions are similar. In terms of future work, a promising direction would be to design an unsupervised pre-training task which exploits this latter observation, e.g. by pre-training a cross-encoder on patient descriptions in which disease names are masked. Furthermore, as shown by Meng et al. [43], biomedical knowledge graphs can play an important role for interpreting patient descriptions, hence the integration of such resources with the considered nearest neighbour strategy is also a natural direction to explore.

REFERENCES

[1] Asma Ben Abacha and Dina Demner-Fushman. 2016. Recognizing question entailment for medical question answering. In *AMIA Annual Symposium Proceedings*, Vol. 2016. American Medical Informatics Association, 310.

[2] 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*. Association for Computational Linguistics, Online, 3023–3033. https://doi.org/10.18653/v1/2021.findings-acl.266

[3] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. 72–78.

[4] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics, Minneapolis, Minnesota, USA, 72–78. https://doi.org/10.18653/v1/W19-1909

[5] Sophia Althammer, Arian Askari, Suzan Verberne, and Allan Hanbury. 2021. DoSSIER@ COLIEE 2021: Leveraging dense retrieval and summarization-based re-ranking for case law retrieval. *arXiv preprint arXiv:2108.03937* (2021).

[6] Corey W Arnold, Suzie M El-Saden, Alex AT Bui, and Ricky Taira. 2010. Clinical case-based retrieval using latent topic analysis. In *AMIA annual symposium proceedings*, Vol. 2010. American Medical Informatics Association, 26.

[7] Alan R Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.. In *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 17.

[8] Neda Barzegar Marvasti, Ceyhun Burak Akgül, Burak Acar, Nadin Kökciyan, Suzan Üsküdarlı, Pınar Yolum, Rüstü Türkay, and Barış Bakır. 2013. Clinical experience sharing by similar case retrieval. In *Proceedings of the 1st ACM international workshop on Multimedia indexing and information retrieval for healthcare*. 67–74.

[9] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. 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*. 3613–3618.

[10] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. 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 (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3615–3620. https://doi.org/10.18653/v1/D19-1371

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

[12] Chris Buckley and Stephen Robertson. 2008. *Relevance feedback track overview: TREC 2008*. Technical Report. MICROSOFT CORP REDMOND WA.

[13] Chris Buckley, Gerard Salton, James Allan, and Amit Singhal. 1995. Automatic query expansion using SMART: TREC 3. *NIST special publication sp* (1995), 69–69.

[14] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, 1–14. https://doi.org/10.18653/v1/S17-2001

[15] Wei-Cheng Chang, Felix X Yu, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2020. Pre-training tasks for embedding-based large-scale retrieval. *arXiv preprint arXiv:2002.03932* (2020).

[16] Joe Davison, Joshua Feldman, and Alexander M Rush. 2019. Commonsense knowledge mining from pretrained models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. 1173–1178.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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*. 4171–4186.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[19] Amit Gajbhiye, Noura Al Moubayed, and Steven Bradley. 2021. ExBERT: An External Knowledge Enhanced BERT for Natural Language Inference. In *International Conference on Artificial Neural Networks*. Springer, 460–472.

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

[21] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)* 3, 1 (2021), 1–23.

[22] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Retrieval Augmented Language Model Pre-Training. In *ICML (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 3929–3938.

[23] Yun He, Ziwei Zhu, Yin Zhang, Qin Chen, and James Caverlee. 2020. Infusing disease knowledge into BERT for health question answering, medical inference and disease name recognition. *arXiv preprint arXiv:2010.03746* (2020).

[24] Hao-zhe Huang, Xu-dong Lu, Wei Guo, Xin-bo Jiang, Zhong-min Yan, and Shipeng Wang. 2021. Heterogeneous Information Network-Based Patient Similarity Search. *Frontiers in Cell and Developmental Biology* (2021), 2297.

[25] Srinivasan Iyer, Sewon Min, Yashar Mehdad, and Wen-tau Yih. 2021. 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*. 1280–1287.

[26] Zheng Jia, Xian Zeng, Huilong Duan, Xudong Lu, and Haomin Li. 2020. A patient-similarity-based model for diagnostic prediction. *International journal of medical informatics* 135 (2020), 104073.

[27] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences* 11, 14 (2021), 6421.

[28] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3, 1 (2016), 1–9.

[29] Mandar Joshi, Kenton Lee, Yi Luan, and Kristina Toutanova. 2020. Contextualized Representations Using Textual Encyclopedic Knowledge. *CoRR* abs/2004.12006 (2020). arXiv:2004.12006 https://arxiv.org/abs/2004.12006

[30] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 6769–6781.

[31] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through Memorization: Nearest Neighbor Language Models. In *8th International Conference on Learning Representations*.

[32] Omar Khattab, Christopher Potts, and Matei Zaharia. 2021. Relevance-guided supervision for openqa with colbert. *Transactions of the Association for Computational Linguistics* 9 (2021), 929–944.

[33] Omar Khattab and Matei Zaharia. 2020. 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*. 39–48.

[34] Victor Lavrenko and W Bruce Croft. 2001. Relevance-based language models: Estimation and analysis. *Croft and Lafferty [2]* (2001), 1–5.

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

[36] Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 6086–6096.

[37] Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2021. Pretrained transformers for text ranking: Bert and beyond. *Synthesis Lectures on Human Language Technologies* 14, 4 (2021), 1–325.

[38] Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-BERT: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 2901–2908.

[39] Ye Liu, Shaika Chowdhury, Chenwei Zhang, Cornelia Caragea, and Philip S. Yu. 2020. Interpretable Multi-Step Reasoning with Knowledge Extraction on

Complex Healthcare Question Answering. *CoRR* abs/2008.02434 (2020).

[40] Mingming Lu, Yu Fang, Fengqi Yan, and Maozhen Li. 2019. Incorporating domain knowledge into natural language inference on clinical texts. *IEEE Access* 7 (2019), 57623–57632.

[41] Diwakar Mahajan, Ananya Poddar, Jennifer J Liang, Yen-Ting Lin, John M Prager, Parthasarathy Suryanarayanan, Preethi Raghavan, and Ching-Huei Tsou. 2020. Identification of Semantically Similar Sentences in Clinical Notes: Iterative Intermediate Training Using Multi-Task Learning. *JMIR medical informatics* 8, 11 (2020), e22508.

[42] Yosi Mass and Haggai Roitman. 2020. 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)*. Association for Computational Linguistics, Online, 4191–4197. https://doi.org/10.18653/v1/2020.emnlp-main.343

[43] Zaiqiao Meng, Fangyu Liu, Thomas Hikaru Clark, Ehsan Shareghi, and Nigel Collier. 2021. Mixture-of-Partitions: Infusing Large Biomedical Knowledge Graphs into BERT. *arXiv preprint arXiv:2109.04810* (2021).

[44] Zaiqiao Meng, Fangyu Liu, Ehsan Shareghi, Yixuan Su, Charlotte Collins, and Nigel Collier. 2021. Rewire-then-Probe: A Contrastive Recipe for Probing Biomedical Knowledge of Pre-trained Language Models. *arXiv preprint arXiv:2110.08173* (2021).

[45] Arindam Mitra, Pratyay Banerjee, Kuntal Kumar Pal, Swaroop Mishra, and Chitta Baral. 2019. Exploring ways to incorporate additional knowledge to improve Natural Language Commonsense Question Answering. *CoRR* abs/1909.08855 (2019). arXiv:1909.08855 http://arxiv.org/abs/1909.08855

[46] Stefania Montani, Riccardo Bellazzi, Luigi Portinale, Stefano Fiocchi, and Mario Stefanelli. 1998. A case-based retrieval system for diabetic patients therapy. *Proceedings of IDAMAP* 98 (1998), 64–70.

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

[48] Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. *arXiv preprint arXiv:1901.04085* (2019).

[49] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. 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*. 2463–2473.

[50] Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. 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*. 5835–5847.

[51] Danilo Ribeiro, Thomas Hinrichs, Maxwell Crouse, Kenneth Forbus, Maria Chang, and Michael Witbrock. 2019. Predicting state changes in procedural text using analogical question answering. In *7th Annual Conference on Advances in Cognitive Systems*.

[52] Kirk Roberts, Dina Demner-Fushman, Ellen M. Voorhees, and William R. Hersh. 2016. Overview of the TREC 2016 Clinical Decision Support Track. In *Proceedings of The Twenty-Fifth Text REtrieval Conference, TREC 2016, Gaithersburg, Maryland, USA, November 15-18, 2016 (NIST Special Publication, Vol. 500-321)*, Ellen M. Voorhees and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec25/papers/Overview-CL.pdf

[53] Kirk Roberts, Matthew Simpson, Dina Demner-Fushman, Ellen Voorhees, and William Hersh. 2016. State-of-the-art in biomedical literature retrieval for clinical cases: a survey of the TREC 2014 CDS track. *Information Retrieval Journal* 19, 1 (2016), 113–148.

[54] Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94*. Springer, 232–241.

[55] Alexey Romanov and Chaitanya Shivade. 2018. Lessons from Natural Language Inference in the Clinical Domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 1586–1596. https://doi.org/10.18653/v1/D18-1187

[56] William B Schwartz, G Anthony Gorry, Jerome P Kassirer, and Alvin Essig. 1973. Decision analysis and clinical judgment. *The American journal of medicine* 55, 4 (1973), 459–472.

[57] Yunqiu Shao, Bulou Liu, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2020. THUIR@ COLIEE-2020: Leveraging Semantic Understanding and Exact Matching for Legal Case Retrieval and Entailment. *arXiv preprint arXiv:2012.13102* (2020).

[58] Yunqiu Shao, Jiaxin Mao, Yiqun Liu, Weizhi Ma, Ken Satoh, Min Zhang, and Shaoping Ma. 2020. BERT-PLI: Modeling Paragraph-Level Interactions for Legal Case Retrieval.. In *IJCAI*. 3501–3507.

[59] Soumya Sharma, Bishal Santra, Abhik Jana, Santosh Tokala, Niloy Ganguly, and Pawan Goyal. 2019. 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)*. Association for Computational Linguistics, Hong Kong, China, 6092–6097. https://doi.org/10.18653/v1/D19-1631

[60] Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised Commonsense Question Answering with Self-Talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. 4615–4629.

[61] Luca Soldaini and Nazli Goharian. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, sigir*. 1–4.

[62] Madhumita Sushil, Simon Suster, and Walter Daelemans. 2021. Are we there yet? Exploring clinical domain knowledge of BERT models. In *Proceedings of the 20th Workshop on Biomedical Language Processing*. Association for Computational Linguistics, Online, 41–53. https://doi.org/10.18653/v1/2021.bionlp-1.5

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

[64] Ashish Upadhyay, Stewart Massie, and Sean Clogher. 2020. Case-Based Approach to Automated Natural Language Generation for Obituaries. In *International Conference on Case-Based Reasoning*. Springer, 279–294.

[65] Shikhar Vashishth, Denis Newman-Griffis, Rishabh Joshi, Ritam Dutt, and Carolyn P Rosé. 2021. Improving broad-coverage medical entity linking with semantic type prediction and large-scale datasets. *Journal of Biomedical Informatics* 121 (2021), 103880.

[66] David Vilares and Carlos Gómez-Rodríguez. 2019. HEAD-QA: A Healthcare Dataset for Complex Reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 960–966. https://doi.org/10.18653/v1/P19-1092

[67] Kexin Wang, Nils Reimers, and Iryna Gurevych. 2021. TSDAE: Using Transformer-based Sequential Denoising Auto-Encoder for Unsupervised Sentence Embedding Learning. *arXiv preprint arXiv:2104.06979* (2021).

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

[69] Hannes Westermann, Jaromir Savelka, and Karim Benyekhlef. 2020. Paragraph similarity scoring and fine-tuned BERT for legal information retrieval and entailment. In *JSAI International Symposium on Artificial Intelligence*. Springer, 269–285.

[70] Jinxi Xu and W Bruce Croft. 2017. Quary expansion using local and global document analysis. In *Acm sigir forum*, Vol. 51. ACM New York, NY, USA, 168–175.

[71] Yinfei Yang, Ning Jin, Kuo Lin, Mandy Guo, and Daniel Cer. 2021. 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)*. Association for Computational Linguistics, Online, 263–268. https://doi.org/10.18653/v1/2021.acl-short.35

[72] Wei Yu, Xiaoting Guo, Fei Chen, Tao Chang, Mengzhu Wang, and Xiaodong Wang. 2021. Similar Questions Correspond to Similar SQL Queries: A Case-Based Reasoning Approach for Text-to-SQL Translation. In *International Conference on Case-Based Reasoning*. Springer, 294–308.

[73] Zheng Yuan, Yijia Liu, Chuanqi Tan, Songfang Huang, and Fei Huang. 2021. Improving Biomedical Pretrained Language Models with Knowledge. *arXiv preprint arXiv:2104.10344* (2021).

[74] Taolin Zhang, Zerui Cai, Chengyu Wang, Minghui Qiu, Bite Yang, and Xiaofeng He. 2021. 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)*. Association for Computational Linguistics, Online, 5882–5893. https://doi.org/10.18653/v1/2021.acl-long.457

[75] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129* (2019).

[76] Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating commonsense in pre-trained language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 9733–9740.

[77] Ming Zhu, Aman Ahuja, Wei Wei, and Chandan K Reddy. 2019. A hierarchical attention retrieval model for healthcare question answering. In *The World Wide Web Conference*. 2472–2482.