Infusing Disease Knowledge into BERT for Health Question Answering, Medical Inference and Disease Name Recognition

Yun He[1], Ziwei Zhu[1], Yin Zhang[1], Qin Chen[2], James Caverlee[1]
[1]Texas A&M University, College Station, USA
[2]Fudan University, Shanghai, China
{yunhe, zhuziwei, zhan13679, caverlee}@tamu.edu
qin_chen@fudan.edu.cn

Abstract

Knowledge of a disease includes information of various aspects of the disease, such as signs and symptoms, diagnosis and treatment. This disease knowledge is critical for many health-related and biomedical tasks, including consumer health question answering, medical language inference and disease name recognition. While pre-trained language models like BERT have shown success in capturing syntactic, semantic, and world knowledge from text, we find they can be further complemented by specific information like knowledge of symptoms, diagnoses, treatments, and other disease aspects. Hence, we integrate BERT with disease knowledge for improving these important tasks. Specifically, we propose a new disease knowledge infusion training procedure and evaluate it on a suite of BERT models including BERT, BioBERT, SciBERT, ClinicalBERT, BlueBERT, and ALBERT. Experiments over the three tasks show that these models can be enhanced in nearly all cases, demonstrating the viability of disease knowledge infusion. For example, accuracy of BioBERT on consumer health question answering is improved from 68.29% to 72.09%, while new SOTA results are observed in two datasets. We make our data and code freely available.[1]

1 Introduction

Human disease is "a disorder of structure or function in a human that produces specific signs or symptoms" (Oxford-English-Dictionary, 2020). Disease is one of the fundamental biological entities in biomedical research and consequently it is frequently searched for in the scientific literature (Islamaj Dogan et al., 2009) and on the internet (Brownstein et al., 2009).

Knowledge of a disease includes information about various aspects of the disease, like the signs and symptoms, diagnosis, and treatment (Saleem et al., 2012; Urnes et al., 2008; Du Jeong et al., 2017). As an example, Table 1 highlights several aspects for COVID-19. Specialized disease knowledge is critical for many health-related and biomedical natural language processing (NLP) tasks, including:

Table 1: Disease knowledge of *COVID-19* is presented from three aspects: symptoms, diagnosis and treatment (based on Wikipedia).

Disease	Aspect	Information
COVID-19	symptoms	Fever is the most common symptom, but highly variable in severity and presentation, with some older...
COVID-19	diagnosis	The standard method of testing is real-time reverse transcription polymerase chain reaction (rRT-PCR)...
COVID-19	treatment	People are managed with supportive care, which may include fluid therapy, oxygen support, and supporting...

- *Consumer health question answering* (Abacha et al., 2019) - the goal is to rank candidate passages for answering questions like "What is the diagnosis of *COVID-19*?" as shown in Figure 1a;
- *Medical language inference* (Romanov and Shivade, 2018) - the goal is to predict if a given hypothesis (description of a patient) can be inferred from a given premise (another description of the patient);
- *Disease name recognition* (Doğan et al., 2014) - the goal is to detect disease concepts in text.

For these tasks, it is critical for NLP models to capture disease knowledge, that is the semantic relations between a disease-descriptive text and its corresponding aspect and disease:

- As shown in Figure 1a, if models can semantically relate "...real-time reverse transcrip-

[1]https://github.com/heyunh2015/diseaseBERT

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 4604–4614,
November 16–20, 2020.

tion polymerase chain reaction..." (disease-descriptive text) to the diagnosis (aspect) of *COVID-19* (disease), it is easier for them to pick up the most relevant answer among the candidates.

- Likewise, as shown in Figure 1b, if models know that the premise is the symptoms (aspect) of *Aphasia* (disease) in the hypothesis, they can easily predict that it is entailment not contradiction.

- Another example is shown in Figure 1c, if models can semantically relate "CTG expansion' to the cause (aspect) of *Myotonic dystrophy* (disease), it is easier for them to detect this disease.

In a nutshell, NLP models require the disease knowledge for these disease-related tasks.

Recently, a new style of knowledge learning and leveraging has shaken NLP field with dramatic successes, enabled by BERT (Devlin et al., 2019) and its variants (Yang et al., 2019; Liu et al., 2019b; Raffel et al., 2019; Lan et al., 2020). These models capture language and world knowledge (Qiu et al., 2020; Rogers et al., 2020) in their parameters via self-supervised pre-training over large-scale unannotated data and then leverage these knowledge in further fine-tuning over downstream tasks. Moreover, many biomedical BERT models such as BioBERT (Lee et al., 2020) are proposed, which are pre-trained over biomedical corpora via a masked language model (MLM) that predicts randomly masked tokens given their context. This MLM strategy is designed to capture the semantic relations between random masked tokens and their context, but not the disease knowledge. Because the corresponding disease and aspect *might not be randomly masked or might not be mentioned at all* in the disease-descriptive text, the semantic relations between them cannot be effectively captured via MLM. Therefore, a new training strategy is required to capture this disease knowledge.

In this paper, we propose a new *disease knowledge infusion* training procedure to explicitly augment BERT-like models with the disease knowledge. The core idea is to train BERT to infer the corresponding disease and aspect from a disease-descriptive text, enabled by weakly-supervised signals from Wikipedia. Given a passage extracted from a section (normally describes an aspect) of a disease's Wikipedia article, BERT is trained to infer

Question: ...keen to learn **how to get COVID-19 diagnosed**, many thanks
Answer 1: ... real-time reverse transcription polymerase chain reaction...
Answer 2: ... diagnosis of vipoma requires demonstration of diarrhea...
Answer 3: ...affected by this disorder are not able to make lipoproteins...
Label: Answer 1 is the most relevant
Disease Knowledge: Answer 1 is the diagnosis of COVID-19

(a) Consumer Health Question Answering

Premise: She was **not able to speak, but appeared to comprehend well**
Hypothesis: Patient had **aphasia**
Label: entailment
Disease Knowledge: Premise describes the symptoms of aphasia

(b) Medical Language Inference

Text: **Myotonic dystrophy** (DM) is **caused by a CTG expansion** in the 3 untranslated region of the DM gene.
Label: Myotonic dystrophy
Disease Knowledge: the text contains the cause of Myotonic dystrophy

(c) Disease Name Recognition

Figure 1: Examples of tasks that can benefit from disease knowledge.

the title of the corresponding section (aspect name) and the title of the corresponding article (disease name). For example, in Table 1, given "...testing is real-time reverse transcription polymerase chain reaction (rRT-PCR)...", BERT is trained to infer that this passage is from the section "diagnosis" of the article "COVID-19". Moreover, because some passages do not mention the disease and aspect, we construct auxiliary sentences that contain the disease and aspect, such as "What is the diagnosis of COVID-19?" and insert this sentence at the beginning of the corresponding passage. After that, we mask the disease and aspect in the auxiliary sentence and then let BERT-like models infer them given the passage. In this way, BERT learns how to semantically relate a disease-descriptive text with its corresponding aspect and disease.

To evaluate the quality of disease knowledge infusion, we conduct experiments on a suite of BERT models – including BERT, BlueBERT, ClinicalBERT, SciBERT, BioBERT, and ALBERT – over consumer health question (CHQ) answering, medical language inference, and disease name recognition. We find that (1) these models can be enhanced in nearly all cases. For example, accuracy of BioBERT on CHQ answering is improved from 68.29% to 72.09%; and (2) our method is superior to MLM for infusing the disease knowledge. Moreover, new SOTA results are observed in two datasets. These results demonstrate the potential of disease knowledge infusion into pre-trained language models like BERT.

2 Related Work

Knowledge-Enriched BERT: Incorporating external knowledge into BERT has been shown to be effective. Such external knowledge includes world (factual) knowledge for tasks such as entity typing and relation classification (Zhang et al., 2019; Peters et al., 2019; Liu et al., 2019a; Xiong et al., 2019), sentiment knowledge for sentiment analysis (Tian et al., 2020; Yin et al., 2020), word sense knowledge for word sense disambiguation (Levine et al., 2019), commonsense knowledge for commonsense reasoning (Klein and Nabi, 2020) and sarcasm generation (Chakrabarty et al., 2020), legal knowledge for legal element extraction (Zhong et al., 2020), numerical skills for numerical reasoning (Geva et al., 2020), and coding knowledge for code generation (Xu et al., 2020).

Biomedical BERT: BERT can also be enriched with biomedical knowledge via pre-training over biomedical corpora like PubMed, as in BioBERT (Lee et al., 2020), SciBERT (Beltagy et al., 2019), ClinicalBERT (Alsentzer et al., 2019) and BlueBERT (Peng et al., 2019). These biomedical BERT models report new SOTA performance on several biomedical tasks. Disease knowledge, of course, is a subset of biomedical knowledge. However, there are two key differences between these biomedical BERT models and our work: (1) Many biomedical BERT models are pre-trained via BERT's default MLM that predicts 15% randomly masked tokens. In contrast, we propose a new training task: disease knowledge infusion, which infers the disease and aspect from the corresponding disease-descriptive text; (2) Biomedical BERT models capture the general syntactic and semantic knowledge of biomedical language, while our work is specifically designed for capturing the semantic relations between a disease-descriptive text and its corresponding aspect and disease. Experiments reported in Section 4 show that our proposed method can improve the performance of each of these biomedical BERT models, demonstrating the importance of disease knowledge infusion.

Biomedical Knowledge Integration Methods with UMLS: Previous non-BERT methods connect data of downstream tasks with knowledge bases like UMLS (Sharma et al., 2019; Romanov and Shivade, 2018). For example, they map medical concepts and semantic relationships in the data to UMLS. After that, these concepts and relationships are encoded into embeddings and incorporated into models (Sharma et al., 2019). The advantage is that they can explicitly incorporate knowledge into models. However, these methods have been outperformed by biomedical BERT models such as BioBERT in most cases.

Table 2: Eight aspects of knowledge of a disease that are considered in this work.

Aspect Name	Definition
Information	The general information of a disease.
Causes	The causes of a disease.
Symptoms	The signs and symptoms of a disease.
Diagnosis	How to test and diagnose a disease.
Treatment	How to treat and manage a disease.
Prevention	How to prevent a disease.
Pathophysiology	The physiological processes of a disease.
Transmission	The means by which a disease spread.

3 Proposed Method: Disease Knowledge Infusion Training

In this section, we propose a new training task: Disease Knowledge Infusion Training. Our goal is to integrate BERT-like pre-trained language models with disease knowledge to achieve better performance on a variety of medical domain tasks including answering health questions, medical language inference, and disease name recognition. Our approach is guided by three questions: Which diseases and aspects should we focus on? How do we infuse disease knowledge into BERT-like models? What is the objective function of this training task?

3.1 Targeting Diseases and Aspects

First, we seek a disease vocabulary that provides disease terms. Several resources include Medical Subject Headings[2] (MeSH) (Lipscomb, 2000), the National Cancer Institute thesaurus (De Coronado et al., 2004), SNOMED CT (Donnelly, 2006), and Unified Medical Language System (UMLS) (Bodenreider, 2004). Each has a different scope and design purpose, and it is an open question into which is most appropriate here. As a first step, we select MeSH, which is a comprehensive controlled vocabulary proposed by the National Library of Medicine (NLM) to index journal articles and books in the life sciences, composed of 16 branches like anatomy, organisms, and diseases. We collect all unique disease terms from the Disease (MeSH tree number C01-C26) and Mental Disorder branch (MeSH tree number F01), resulting in 5,853 total disease terms.

[2] https://meshb.nlm.nih.gov/treeView

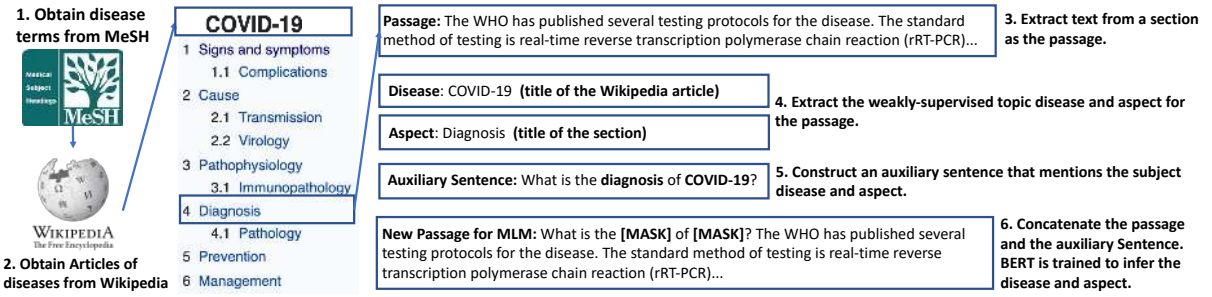

Figure 2: Disease Knowledge Infusion Training: An example with COVID-19.

Knowledge of a disease involves information about various aspects of the disease (Saleem et al., 2012; Urnes et al., 2008; Du Jeong et al., 2017). For each aspect, we focus on text alone (excluding images or other media). Following Abacha and Demner-Fushman (2019), we consider eight disease aspects as shown in Table 2.

3.2 Weakly Supervised Knowledge Infusion from Wikipedia

Given the target set of diseases and aspects, the next challenge is how to infuse knowledge of the aspects of these diseases into BERT-like models. We propose to train BERT to infer the corresponding disease and aspect from a disease-descriptive text. By minimizing the loss between the predicted disease and aspect and the original disease and aspect, the model should memorize the semantic relations between the disease-descriptive text and its corresponding disease and aspect.

A straightforward approach is to mask and predict the disease and aspect in the disease-descriptive text. However, this strategy faces two problems: (1) Given a passage extracted from disease-related papers, clinical notes, or biomedical websites, the ground-truth of its topic (i.e., disease and aspect) is difficult to identify. Medical expert annotation is time-consuming and expensive; while automatic annotation can suffer from large errors. For example, we need to recognize disease names in the passage, which is yet another challenging and still open problem in biomedical text mining (Doğan et al., 2014); (2) Diseases and aspects mentioned in a passage are not necessarily the topic words. Multiple disease names or aspect names might appear, making it difficult to determine which is the correct topic. For example, in Table 1, the symptoms of *COVID-19* also mentions *fever*[3], while the correct topic is *COVID-19*.

[3]Fever is included in the disease branch of MeSH.

Weakly-Supervised Knowledge Source: Instead of annotating an arbitrary disease-related passage, we exploit the structure of Wikipedia as a weakly-supervised signal. In many cases, each disease's Wikipedia article consists of several sections where each introduces an aspect of the disease (like diagnosis). For example, step 2 in Figure 2 shows several aspects on the Wikipedia page for *COVID-19*. By extracting the passage from each section, the title of the section (e.g., diagnosis) is the topic aspect of the passage and the title of the article is the topic disease (e.g., COVID-19). Specifically, we search Wikipedia to obtain the articles for the 5,853 target disease terms from MeSH and apply regular expressions to extract the text of the sections corresponding to the appropriate aspects. In total, we collect a disease knowledge resource consisting of 14,617 passages.[4] In fact, there are other online resources[5] with the similar structure. As a first step, we start with Wikipedia.

Auxiliary Sentences for Disease and Aspect Prediction: The second problem is that the extracted passages do not necessarily mention the corresponding disease and the aspect. For example, in Table 1, the disease name "COVID-19" does not appear in the information of its symptoms. In the disease knowledge resource, we find that only 51.4% of passages mention both the corresponding diseases and aspects. Hence, we cannot simply mask-and-predict the disease and aspect because the passage does not mention them at all.

A remedy for this problem is an auxiliary sentence that contains the corresponding disease and aspect for each passage. We use a template of question style: "What is the [*Aspect*] of [*Disease*]?" to automatically generate auxiliary sentences as shown in step 5 in Figure 2. Some examples are

[4]Note that each disease article does not necessarily have all eight target aspects.

[5]https://medlineplus.gov/skincancer.html

shown in Table 3. The advantage of this question style template is that the cloze statement of the auxiliary sentences for all aspects (except for the "information" aspect) are the same (What is the [MASK] of [MASK]?). Hence, the auxiliary sentences provide no clues (i.e., bias) for predicting the corresponding aspect.

Table 3: Examples of auxiliary sentences

Aspect Name	Auxiliary Sentence
Diagnosis	What is the diagnosis of COVID-19?
Treatment	What is the treatment of COVID-19?
Prevention	What is the prevention of COVID-19?
Transmission	What is the transmission of COVID-19?
Cloze Statement	What is the [MASK] of [MASK]?

After that, we replace the corresponding disease and aspect with the special token [MASK] in the auxiliary sentences. Then, we insert the auxiliary sentence at the beginning of its corresponding passage to form a new passage with a question-and-answer style as shown in Figure 2, where BERT is trained to predict the original tokens of the masked disease and aspect.

3.3 Training Objective and Details

Finally, we show the objective function of disease infusion training. Since most disease names are out of BERT vocabulary, the WordPiece tokenizer (Wu et al., 2016) will split these terms into sub-word tokens that exist in the vocabulary. For example, "COVID-19" will be split into 4 tokens: "co", "vid", "-" and "19". Formally, let $X = (x_1, ..., x_T)$ denote a sequence of T tokens that are split from a disease name where x_t is the t-th token. The original cross-entropy loss is to get the conditional probability of a masked token as close as possible to the 1-hot vector of the token:

$$\mathcal{L}_{disease} = -\sum_{t=1}^{T} log\ p(x_t|passage) \quad (1)$$

where $p(x_t|context)$ is a conditional probability over x_t given the corresponding passage, which can be defined as:

$$p(x_t|passage) = \frac{exp(z_t)}{\sum_{z \in \mathcal{V}} exp(z)} \quad (2)$$

where $\mathcal{V}$ is the vocabulary and z_t is the unnormalized log probability of x_t. Let $\mathbf{y}_t$ denote the embedding of token x_t from the output layer of BERT. We can estimate z_t via:

$$z_t = \mathbf{w} \cdot \mathbf{y}_t + b \quad (3)$$

where the weight $\mathbf{w}$ and bias b are learnable vectors.

Note that the vocabulary size of BERT is around 30,000 which means masked language modeling task is a 30,000 multi-class problem. The logits (like z_t) after the normalization of softmax (Equation 2) will be pretty small (the expectation of mean should be around 1/30,000=3.3*e-5), which might cause some obstacles for the learning. Therefore, we also maximize the raw logits (like z_t) before softmax normalization which might keep more useful information. Empirically, we add the reciprocal of the logits to the cross-entropy loss:

$$\mathcal{L}_{disease} = -\sum_{t=1}^{T} logp(x_t|passage) + \frac{\beta}{\sum_{t=1}^{T} z_t} \quad (4)$$

where β balances the two parts of the loss. The final objective function is combined with the loss of the disease and aspect: $\mathcal{L} = \mathcal{L}_{disease} + \mathcal{L}_{aspect}$ where $\mathcal{L}_{aspect} = -log\ p(a|passage)$ and a is the token of the aspect name. By minimizing this loss function, BERT can update its parameters to store the disease knowledge.

4 Experiments

In this section, we examine disease knowledge infusion into six BERT variants over three disease-related tasks: health question answering, medical language inference, and disease name recognition.

Reproducibility: *The code and data in this paper is released.*[6] A model is firstly initialized with the pre-trained parameters from BERT or its variants and then is further trained by disease knowledge infusion to capture the disease knowledge. We use a widely used Pytorch implementation[7] of BERT and Adam as the optimizer. We empirically set learning rate as 1e-5, batch size as 16 and β as 10. Because MeSH (5,853 disease terms) is chosen as the disease vocabulary in our experiments, as a smaller vocabulary compared with others like UMLS (540,000 disease terms), we obtain a relatively small dataset of 14,617 passages. Hence, the training of disease knowledge infusion is as fast as fine-tuning BERT over downstream datasets, which takes 2-4 epochs to enhance BERT for a better performance on downstream tasks, which will be discussed in Section 4.5. The training is performed on one single NVIDIA V100 GPU and

[6] https://github.com/heyunh2015/diseaseBERT

[7] https://github.com/huggingface/transformers

it takes about 10 minutes to complete one training epoch using BERT-base architecture. The reproducibility for fine-tuning over downstream tasks will be detailed in Section 4.2.

4.1 BERT and its Biomedical Variants

We consider six BERT models: two pre-trained over general language corpora (BERT and ALBERT) and four pre-trained over biomedical corpora (Clinical BERT, BioBERT, BlueBERT and SciBERT).

BERT (Devlin et al., 2019) is a multi-layer bidirectional Transformer encoder. Since the following biomedical versions of BERT are often based on the BERT-base architecture (12 layers and 768 hidden embedding size with 108M parameters), we choose BERT-base here for fair comparison.

ALBERT[8] (Lan et al., 2020) compresses the architecture of BERT by factorized embedding parameterization and cross-layer parameter sharing. Via this compression, ALBERT can have a substantially higher capacity than BERT, with stronger performance on many tasks. We choose the maximum version ALBERT-xxlarge (12 layers and 4096 hidden embedding size with 235M parameters).

BioBERT[9] (Lee et al., 2020) is the first BERT pre-trained on biomedical corpora. It is initialized with BERT's pre-trained parameters (108M) and then further trained over PubMed abstracts (4.5B words) and PubMed Central full-text articles (13.5B words). We choose the best version BioBERT v1.1.

ClinicalBERT[10] (Alsentzer et al., 2019) is a BERT model initialized from BioBERT v1.0 (Lee et al., 2020) and further pre-trained over approximately 2 million notes in the MIMIC-III v1.4 database of patient notes (Johnson et al., 2016). We adopt the best performing version of ClinicalBERT (108M parameters) based on discharge summaries of clinical notes: Bio-Discharge Summary BERT.

BlueBERT[11] (Peng et al., 2019) is firstly initialized from BERT (108M parameters) and further pre-trained over a biomedical corpus of PubMed abstracts and clinical notes (Johnson et al., 2016).

SciBERT[12] (Beltagy et al., 2019) is a BERT-base (108M parameters) model pre-trained on a random sample of the full text of 1.14M papers from Semantic Scholar (Ammar et al., 2018), with 18% of papers from the computer science domain and 82% from the biomedical domain.

[8] https://huggingface.co/albert-xxlarge-v2
[9] https://github.com/dmis-lab/biobert
[10] https://huggingface.co/emilyalsentzer
[11] https://github.com/ncbi-nlp/bluebert
[12] https://huggingface.co/allenai/scibert_scivocab_uncased

Table 4: Summary of Tasks and Datasets.

Datasets	Train	Dev	Test
MEDIQA-2019	208 (1, 701)[1]	25 (234)	150 (1,107)
TRECQA-2017	254 (1,969)	25 (234)	104 (839)
MEDNLI	11, 232[2]	1,395	1,422
BC5CDR-disease	4, 182[3]	4,244	4,424
NCBI	5,145	787	960

1, Questions with associated answers; 2, Pairs of premise and hypothesis; 3, Disease name mentions

4.2 Tasks

We test disease knowledge infusion over three biomedical NLP tasks. The dataset statistics are in Table 4. For fine-tuning of BERT and its variants, the batch size is selected from [16, 32] and learning rate is selected from [1e-5, 2e-5, 3e-5, 4e-5, 5e-5].

Task 1: Consumer Health Question Answering. The objective of this task is to rank candidate answers for consumer health questions.

Datasets. We consider two datasets: MEDIQA-2019 (Ben Abacha et al., 2019) and TRECQA-2017 (Abacha et al., 2017).[13] MEDIQA-2019 is based on questions submitted to the consumer health QA system CHiQA[14]. TRECQA-2017 is based on questions submitted to the National Library of Medicine. Medical experts manually re-ranked the original retrieved answers and provide *Reference Score* (1 to 11) and *Reference Rank* (4: Excellent, 3: Correct but Incomplete, 2: Related, 1: Incorrect).

Fine-tuning. MEDIQA-2019 and TRECQA-2017 are used as the fine-tuning dataset for each other. MEDIQA-2019 also contains a validation set for tuning hyper-parameters for both datasets. Following Xu et al. (2019), the task is cast as a regression problem where the target score is: $score = Reference\ Score - \frac{Reference\ Rank - 1}{m}$ where m is the number of candidate answers. Each question-answer pair is packed as a single sequence as the input for BERT. A single linear layer is on top of the output embedding of the special token [CLS] to generate the predicted score. MSE is adopted as the loss and we use Adam as the optimizer. All hyper-parameters are tuned on the validation set in terms of accuracy, where we set the batch size as 16 and learning rate as 1e-5.

[13] https://sites.google.com/view/mediqa2019
[14] https://chiqa.nlm.nih.gov/

Table 5: Experimental Results

Tasks	Consumer Health Question Answering						NLI	NER	
Datasets	MEDIQA-2019			TRCEQA-2017			MEDNLI	BC5CDR	NCBI
Metrics(%)	Accuracy	MRR	Precision	Accuracy	MRR	Precision	Accuracy	F1	F1
BERT	64.95	82.72	66.49	74.61	56.17	52.55	75.95	83.09	85.14
BERT + disease*	66.40↑	83.33↑	68.94↑	75.33↑	56.41↑	54.01↑	77.29↑	83.47↑	86.81↑
BlueBERT	65.13	81.50	67.35	74.26	48.40	52.55	82.21	85.73	87.78
BlueBERT + disease	68.47↑	81.17	71.57↑	77.59↑	50.96↑	57.62↑	83.90↑	86.30↑	87.79↑
ClinicalBERT	67.30	84.78	70.59	77.00	52.56	56.62	81.50	84.90	87.25
ClinicalBERT + disease	69.02↑	88.94↑	69.84	78.90↑	54.97↑	60.40↑	81.65↑	85.63↑	87.22
SciBERT	68.47	84.47	68.07	77.23	54.57	57.54	80.94	86.16	87.24
SciBERT + disease	73.35↑	85.44↑	76.28↑	79.02↑	56.57↑	59.57↑	82.14↑	86.34↑	88.30↑
BioBERT	68.29	83.61	72.78	77.12	49.84	57.25	81.86	85.99	87.70
BioBERT + disease	72.09↑	87.78↑	74.40↑	78.43↑	54.76↑	58.45↑	82.21↑	86.52↑	87.14
ALBERT	76.54	88.46	81.41	75.09	**58.57**	53.03	85.48	84.28	87.56
ALBERT + disease	**79.49**↑	90.00↑	**84.02**↑	**80.10**↑	57.21	**62.40**↑	**86.15**↑	84.71↑	87.69↑
SOTA*	78.00	**93.67**	81.91	77.23	54.57	57.54	84.00	**87.15**	**89.71**

* SOTA, state-of-the-art as of May 2020, to the best of our knowledge.
* " + disease" means that we train BERT via disease knowledge infusion training before fine-tuning.

SOTA. The state-of-the-art (SOTA) performance on MEDIQA-2019 is achieved by Xu et al. (2019), which is an ensemble method. Because TRECQA-2017 is fine-tuned on MEDIQA-2019, which is different from the original settings (Abacha et al., 2017) (BERT had not been proposed at that time), we use the best result of SciBERT among the BERT models as SOTA for TRECQA-2017.

Task 2: Medical Language Inference. The goal of this task is to predict whether a given hypothesis can be inferred from a given premise.

Datasets. MEDNLI (Romanov and Shivade, 2018) is a natural language inference dataset for the clinical domain.[15] For each premise (a description of a patient) selected from clinical notes (MIMIC-III), clinicians generate three hypotheses: entailment (alternate true description of the patient), contradiction (false description of the patient), and neutral (alternate description that might be true).

Fine-tuning. Following Peng et al. (2019), we pack the premise and hypothesis together into a single sentence. A linear layer is on top of the output embedding of [CLS] to generate logits. Cross-entropy loss function is adopted, and we use Adam as the optimizer. All hyper-parameters are tuned on the validation set in terms of accuracy, where we set the batch size as 32 and learning rate as 1e-5.

SOTA. To the best of our knowledge, the state-of-the-art on MEDNLI is achieved by BlueBERT, reported in Peng et al. (2019).

Task 3: Disease Name Recognition. This task is to detect disease names from free text.

Datasets. BC5CDR[16] (Wei et al., 2016) and NCBI[17] (Doğan et al., 2014) are collections of PubMed titles and abstracts. Medical experts annotate diseases mentioned in the collection. Since BC5CDR includes both chemicals and diseases, we focus on diseases in this dataset.

Fine-tuning. Following Peng et al. (2019), we cast this task as a token-level tagging (classification) problem, where each token is classified into three classes: B (beginning of a disease), I (inside of a disease) or O (out of a disease). Cross-entropy is adopted as the loss function and we use Adam as the optimizer. All hyper-parameters are tuned on the validation set in terms of F1, where we set the batch size as 32 and learning rate as 5e-5.

SOTA. The best performance is achieved by BioBERT v1.1, reported in Lee et al. (2020)[18].

4.3 Results

The experimental results are presented in Table 5. We show each original model and its disease

[15] https://physionet.org/content/mednli/1.0.0/

[16] https://github.com/ncbi-nlp/BLUE_Benchmark

[17] https://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/

[18] Although SciBERT reports a better result in NCBI, it uses a conditional random field on top of BERT, which is more complicated than the linear layer normally used in fine-tuning for BERT models including BioBERT.

knowledge infused variant (e.g,. *BERT* and *BERT + disease*). We have two main findings:

Effectiveness of Disease Infusion. First, by infusing disease knowledge via our new training regimen, we see a significant improvement in nearly all cases. For example, *ALBERT + disease* achieves 80.10% in terms of accuracy which is superior to 75.09% by ALBERT alone on TRECQA-2017. Standing on the shoulders of ALBERT, disease knowledge infusion leads to state-of-the-art results on MEDIQA-2019 and MEDNLI, to the best of our knowledge. Although BERT and ALBERT are pre-trained on all of Wikipedia, including the articles of diseases, they might not pay enough attention to the disease part since Wikipedia is so large. Hence, disease knowledge infusion that leverages the Wikipedia structure to capture the disease knowledge is a complement for BERT and ALBERT. Moreover, it is encouraging to see the improvements of disease knowledge infusion in biomedical BERT models, even though these variants are already pre-trained over large-scale biomedical corpora like PubMed with access to comprehensive disease information. This improvement demonstrates that the disease knowledge captured by our method – that is, the semantic relations between a disease-descriptive text and its corresponding aspect and disease – is different from the general linguistic knowledge in the biomedical domain captured by the randomly masked tokens prediction strategy of these biomedical BERT models. To sum up, the results show that the proposed disease knowledge infusion method can effectively complement BERT and its biomedical variants and hence improve the performance on health question answering, medical language inference, and disease name recognition.

Effectiveness of Biomedical BERT Models. We also observe that BERT models pre-trained on biomedical corpora outperform the same BERT architecture that is pre-trained on general language corpora. For example, BioBERT achieves 68.29% in terms of accuracy on MEDIQA-2019 while BERT only obtains 64.95%. This demonstrates that with the same model architecture, pre-training on biomedical corpora can capture more biomedical language knowledge that improves BERT for downstream biomedical tasks.[19]

[19]Note that our results for the biomedical BERT models in Table 5 are slightly different from the results reported in the original papers that normally only provide a search range for hyper-parameters and not the specific optimal ones.

Table 6: Ablation Study on MEDIQA-2019

Variants	Accuracy	MRR	Precision
Default	79.49	90.00	84.02
- Auxiliary Sentence	78.23	90.89	78.10
- Aspect Prediction	78.41	89.06	80.00
- Disease Prediction	72.90	85.72	79.44
15% Randomly Masked Tokens	77.06	87.33	85.18

In addition, we find that a high-capacity model like ALBERT can achieve similar performance as biomedical BERT models on TRECQA-2017, BC5CDR and NCBI, and even better performance on MEDIQA-2019 and MEDNLI. This observation might motivate new biomedical pre-trained models based on larger models like ALBERT-xxlarge.

4.4 Ablation Study

We present the results of an ablation study on MEDIQA-2019 in Table 6. Similar results are observed on other datasets but omitted here due to the space limitation. We first remove "Auxiliary Sentence". That is, we remove the auxiliary question: "What is the [*Aspect*] of [*Disease*]?" and let BERT to predict the corresponding disease and aspect in the original passage if they appear. We observe worse results in terms of accuracy and precision, which shows that the auxiliary sentence is an effective remedy for the problem that some passages do not mention their disease and aspects. We also remove aspect prediction or disease prediction in the auxiliary sentence; both lead to worse results but removing disease prediction leads to a much lower performance. This shows that it is more important for BERT to infer the disease than the aspect from the passage. We also pre-train BERT on the same corpus (the disease-related passages) as our method. Following Devlin et al. (2019), we randomly mask 15% tokens in each sentence and let BERT to predict them. As shown in "15% Randomly Masked Tokens", we observe that our proposed disease infusion training task outperforms the default masked language model in BERT. This shows that our approach that leverages the structure of Wikipedia article to enhance the disease knowledge infusion works better than simply adding more data to the training process. Specifically, via leveraging the Wikipedia structure, we could effectively mask key words like aspect names and disease names that are related to disease knowledge and hence more effective than randomly masking strategy over the simply added data.

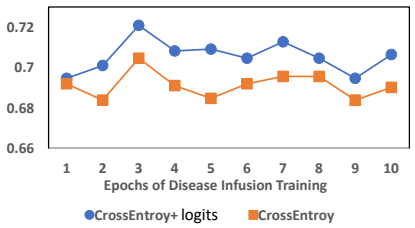

Figure 3: Learning curve of disease infusion knowledge. The y-axis is the accuracy of BERT models over MEDIQA-2019.

4.5 Learning Curve

In this section, we present the learning curve of our proposed disease infusion training task. The x-axis denotes the training epochs and the y-axis denotes the performance of BERT models that are augmented with disease infusion training at that epoch. We take BioBERT and MEDIQA-2019 as examples; similar results are obtained in other models over other tasks. The results in terms of accuracy are presented in Figure 3, where we observe that (1) disease knowledge infusion takes only three epochs to achieve the optimal performance on BioBERT over the CHQ answering task. (2) cross-entropy loss used by disease knowledge infusion can be enhanced by adding the term of maximizing the raw logits (Equation 4).

5 Conclusions

In this paper, we propose a new disease infusion training procedure to augment BERT-like pre-trained language models with disease knowledge. We conduct this training procedure on a suite of BERT models and evaluate them over disease-related tasks. Experimental results show that these models can be enhanced by this disease infusion method in nearly all cases.

References

Asma Ben Abacha, Eugene Agichtein, Yuval Pinter, and Dina Demner-Fushman. 2017. Overview of the medical question answering task at trec 2017 liveqa. In *TREC*.

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC bioinformatics*, 20(1):511.

Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. 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.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Wang, Chris Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. 2018. Construction of the literature graph in semantic scholar. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 84–91, New Orleans - Louisiana. Association for Computational Linguistics.

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)*, pages 3606–3611.

Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. 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, Florence, Italy. Association for Computational Linguistics.

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

John S Brownstein, Clark C Freifeld, and Lawrence C Madoff. 2009. Digital disease detection-harnessing the web for public health surveillance. *The New England journal of medicine*, 360(21):2153.

Tuhin Chakrabarty, Debanjan Ghosh, Smaranda Muresan, and Nanyun Peng. 2020. $\mathcal{R}^3$: Reverse, retrieve, and rank for sarcasm generation with commonsense knowledge. *arXiv preprint arXiv:2004.13248*.

Sherri De Coronado, Margaret W Haber, Nicholas Sioutos, Mark S Tuttle, Lawrence W Wright, et al. 2004. Nci thesaurus: using science-based terminology to integrate cancer research results. In *Medinfo*, pages 33–37.

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), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Kevin Donnelly. 2006. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279.

In Du Jeong, Moo In Park, Sung Eun Kim, Beom Jin Kim, Sang Wook Kim, Jie-Hyun Kim, Hye Young Sung, Tae-Hoon Oh, Yeon Soo Kim, The Korean Society of Neurogastroenterology, et al. 2017. The degree of disease knowledge in patients with gastroesophageal reflux disease: A multi-center prospective study in korea. *Journal of neurogastroenterology and motility*, 23(3):385.

Mor Geva, Ankit Gupta, and Jonathan Berant. 2020. Injecting numerical reasoning skills into language models. *arXiv preprint arXiv:2004.04487*.

Rezarta Islamaj Dogan, G Craig Murray, Aurélie Névéol, and Zhiyong Lu. 2009. Understanding pubmed® user search behavior through log analysis. *Database*, 2009.

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*, 3:160035.

Tassilo Klein and Moin Nabi. 2020. Contrastive self-supervised learning for commonsense reasoning. *arXiv preprint arXiv:2005.00669*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

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):1234–1240.

Yoav Levine, Barak Lenz, Or Dagan, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2019. Sensebert: Driving some sense into bert. *arXiv preprint arXiv:1908.05646*.

Carolyn E Lipscomb. 2000. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2019a. K-bert: Enabling language representation with knowledge graph. *arXiv preprint arXiv:1909.07606*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Oxford-English-Dictionary. 2020. Definition of disease in englis.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. 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, Florence, Italy. Association for Computational Linguistics.

Matthew E Peters, Mark Neumann, IV Logan, L Robert, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. 2019. Knowledge enhanced contextual word representations. *arXiv preprint arXiv:1909.04164*.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *arXiv preprint arXiv:2003.08271*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *arXiv preprint arXiv:2002.12327*.

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*, pages 1586–1596, Brussels, Belgium. Association for Computational Linguistics.

Fahad Saleem, Mohamed Azmi Hassali, Asrul Akmal Shafie, Muhammad Atif, Noman ul Haq, and Hisham Aljadhey. 2012. Disease related knowledge and quality of life: a descriptive study focusing on hypertensive population in pakistan. *Southern med review*, 5(1):47.

Soumya Sharma, Bishal Santra, Abhik Jana, TYSS Santosh, Niloy Ganguly, and Pawan Goyal. 2019. Incorporating domain knowledge into medical nli using knowledge graphs. *arXiv preprint arXiv:1909.00160*.

Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and Feng Wu. 2020. Skep: Sentiment knowledge enhanced pre-training for sentiment analysis. *arXiv preprint arXiv:2005.05635*.

Jorgen Urnes, Hermod Petersen, and Per G Farup. 2008. Disease knowledge after an educational program in patients with gerd–a randomized controlled trial. *BMC health services research*, 8(1):236.

Chih-Hsuan Wei, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Jiao Li, Thomas C Wiegers, and Zhiyong Lu. 2016. Assessing the state of the art in biomedical relation extraction: overview of the biocreative v chemical-disease relation (cdr) task. *Database*, 2016.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Wenhan Xiong, Jingfei Du, William Yang Wang, and Veselin Stoyanov. 2019. Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model. *arXiv preprint arXiv:1912.09637*.

Frank F Xu, Zhengbao Jiang, Pengcheng Yin, Bogdan Vasilescu, and Graham Neubig. 2020. Incorporating external knowledge through pre-training for natural language to code generation. *arXiv preprint arXiv:2004.09015*.

Yichong Xu, Xiaodong Liu, Chunyuan Li, Hoifung Poon, and Jianfeng Gao. 2019. Doubletransfer at mediqa 2019: Multi-source transfer learning for natural language understanding in the medical domain. *arXiv preprint arXiv:1906.04382*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Da Yin, Tao Meng, and Kai-Wei Chang. 2020. Sentibert: A transferable transformer-based architecture for compositional sentiment semantics. *arXiv preprint arXiv:2005.04114*.

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*.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does nlp benefit legal system: A summary of legal artificial intelligence. *arXiv preprint arXiv:2004.12158*.