KEBLM: Knowledge-Enhanced Biomedical Language Models

Tuan Manh Lai, ChengXiang Zhai, Heng Ji

[a] *Computer Science Department, University of Illinois Urbana-Champaign, 201 N. Goodwin Ave, Urbana, 61801, Illinois, United States*

Abstract

Pretrained language models (PLMs) have demonstrated strong performance on many natural language processing (NLP) tasks. Despite their great success, these PLMs are typically pretrained only on unstructured free texts without leveraging existing structured knowledge bases that are readily available for many domains, especially scientific domains. As a result, these PLMs may not achieve satisfactory performance on knowledge-intensive tasks such as biomedical NLP. Comprehending a complex biomedical document without domain-specific knowledge is challenging, even for humans. Inspired by this observation, we propose a general framework for incorporating various types of domain knowledge from multiple sources into biomedical PLMs.

We encode domain knowledge using lightweight *adapter modules*, bottleneck feed-forward networks that are inserted into different locations of a backbone PLM. For each knowledge source of interest, we pretrain an adapter module to capture the knowledge in a self-supervised way. We design a wide range of self-supervised objectives to accommodate diverse types of knowledge, ranging from entity relations to description sentences.

Once a set of pretrained adapters is available, we employ fusion layers to combine the knowledge encoded within these adapters for downstream tasks. Each fusion layer is a parameterized mixer of the available trained adapters that can identify and activate the most useful adapters for a given input. Our method diverges from prior work by including a knowledge consolidation phase,

during which we teach the fusion layers to effectively combine knowledge from both the original PLM and newly-acquired external knowledge using a large collection of unannotated texts. After the consolidation phase, the complete knowledge-enhanced model can be fine-tuned for any downstream task of interest to achieve optimal performance.

Extensive experiments on many biomedical NLP datasets show that our proposed framework consistently improves the performance of the underlying PLMs on various downstream tasks such as natural language inference, question answering, and entity linking. These results demonstrate the benefits of using multiple sources of external knowledge to enhance PLMs and the effectiveness of the framework for incorporating knowledge into PLMs. While primarily focused on the biomedical domain in this work, our framework is highly adaptable and can be easily applied to other domains, such as the bioenergy sector.

Keywords:
Pre-trained language models, Knowledge bases, Domain knowledge

---

1. Introduction

Pretrained language models (PLMs) such as BERT [1] and T5 [2] have recently revolutionized the field of natural language processing (NLP). The main idea is to pretrain a model on a large-scale corpus of unannotated text using one or more self-supervised learning objectives, such as the popular masked language modeling (MLM) objective [1, 3, 4]. PLMs have been shown to effectively capture rich semantic and syntactic patterns from plain texts [5, 6]. As such, for a task of interest with some supervised data, a PLM can typically be fine-tuned to achieve very competitive performance on the target task [7, 8, 9].

While the majority of PLMs are pretrained on a general-domain corpus such as Wikipedia, more and more PLMs are being introduced for more specific domains, such as scientific domains [10, 11, 12]. For example, SciBERT [10]

is a language model trained on a multi-domain corpus of 1.14M scientific publications from Semantic Scholar [13]. Another example is BioBERT [11], a model pretrained on large amounts of PubMed abstracts and PMC full-text articles. By being pretrained on domain-specific texts, these domain-specific PLMs are generally more effective than generic PLMs for NLP tasks within the corresponding domain [14].

However, all these PLMs are trained using only unstructured text content, typically by optimizing a self-supervised training objective. They do not explicitly leverage external knowledge from high-quality structured knowledge bases (KBs) such as UMLS [15] and PubChem [16]. As a result, these PLMs may not achieve satisfactory performance on knowledge-intensive tasks such as biomedical NLP. Indeed, comprehending a complex biomedical document without domain-specific knowledge is quite challenging, even for humans. While the current PLMs may acquire some domain-specific knowledge implicitly from the unstructured literature articles, such domain-specific knowledge is implicitly stored in their model parameters. Due to the exponential growth of scientific publications and knowledge [17], models that do not go beyond their fixed set of parameters will likely fall behind [18, 19, 20, 21]. In fact, recent studies on probing biomedical PLMs suggest that these models possess a very limited amount of biomedical factual knowledge compared to a typical knowledge base (KB) [22, 23]. The main reason is that biomedical documents, either formal (e.g., scientific papers) or informal ones (e.g., clinical notes), are written for domain experts [20, 24]. As such, they contain many highly specialized terms, acronyms, and abbreviations of entities, whose definitions and properties are not presented in the local contextual sentences that are used to train the existing PLMs. For example, in the BioRelEx dataset [25], a biomedical information extraction dataset, we find that about 65% of the annotated entity mentions are abbreviations of biological entities, and an example is shown in Figure 1.

Due to the limited capability of many existing PLMs in learning domain-

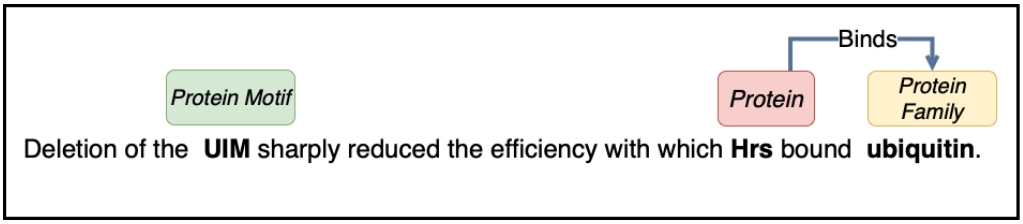


Figure 1: An example in the BioRelEx dataset [25]. **UIM** is an abbreviation of "Ubiquitin-Interacting Motif". In our preliminary experiments, we found that our baseline SciBERT model incorrectly predicts the mention as a "DNA" instead of a "Protein Motif", even though SciBERT was already pretrained on 1.14 million scientific papers.

specific knowledge from literature articles, recently, several methods have been proposed to enhance biomedical PLMs directly with external domain knowledge [20, 19, 26, 27]. For example, KECI [20], a biomedical information extraction framework, utilizes an entity linker as a bridge for transferring knowledge from UMLS [15] to neural models. Given that many high-quality domain knowledge bases already exist in scientific domains and human experts are also making an effort to maintain and grow such knowledge bases over time, such an approach of *knowledge-enhanced* PLMs is appealing. It can empower PLMs with more domain knowledge without requiring extra human effort and would also enable PLMs to scale up to incorporate more knowledge naturally over time.

Despite their effectiveness, many current methods for incorporating domain knowledge bases, such as KECI, can only leverage a single source of knowledge (e.g., UMLS). This limitation restricts the total amount of knowledge that can be utilized for downstream tasks. In the biomedical domain, there are many high-quality KBs that contain complementary knowledge [15, 28, 16]. It is thus important to incorporate all of them into PLMs to maximize the amount of encoded knowledge in the PLMs. However, this is technically challenging as the types of knowledge vary greatly, and each tends to require a different method of incorporation.

To address this challenge, we propose a novel general framework, called

KEBLM (Knowledge-Enhanced Biomedical Language Models), for incorporating various types of knowledge from multiple sources into biomedical PLMs. More concretely, KEBLM encodes knowledge using adapter modules [29, 30, 31], lightweight neural networks that are typically inserted into different layers of a backbone PLM. For each knowledge source of interest, we pretrain an adapter module to memorize the knowledge in it in a self-supervised way. We design a wide range of self-supervised objectives to accommodate diverse types of knowledge, ranging from entity-entity relations to description sentences. Given a set of pretrained adapters, we use fusion layers [30] to combine the knowledge encoded in the adapters for downstream tasks. Each fusion layer is a parameterized mixer of the available trained adapters that can identify and activate the most useful adapters for a given input.

Different from previous studies that also attempt to incorporate knowledge from multiple sources using adapters [21], our method explicitly includes a knowledge consolidation phase. During this phase, we teach the fusion layers to effectively combine knowledge from both the original PLM and newly-acquired external knowledge by using a large collection of unannotated texts. Following the consolidation phase, the complete knowledge-enhanced model can be fine-tuned for any downstream task of interest to achieve optimal performance. The knowledge consolidation phase is crucial, as different types of knowledge typically vary significantly. Consequently, the fusion layers may not learn to incorporate them effectively if relying solely on fine-tuning from a downstream task, especially in scientific domains where available task-specific datasets are relatively small in size.

We evaluate the effectiveness of KEBLM by instantiating it to incorporate three types of biomedical domain knowledge: (1) entity descriptions, (2) entity-entity relations, and (3) entity synonyms. We use multiple biomedical NLP datasets to study the impact of the incorporated knowledge on three representative downstream tasks, i.e., natural language inference (NLI), question answering (QA), and entity linking (EL). Our experiment results

Knowledge Type	Knowledge Base(s)	Example(s)
Entity descriptions	UMLS [15, 32] and PubChem [16]	Phenylephrine(1+) is an organic cation obtained by protonation of the secondary amino function of phenylephrine. It is an ammonium ion derivative and an organic cation. It is a conjugate acid of a phenylephrine.
Entity-entity relations	MSI (multiscale interactome) [33]	(SLBP, interacts with, CSTF3) (EPB41L2, interacts with, EFTUD2) (APOA2, has function, phosphatidylcholine biosynthetic process)
Entity synonyms	UMLS [15]	Synonym pairs: (Cancer, Malignant Neoplasms), (Influenza, Human Flu), (EGFR, Epidermal Growth Factor Receptor)

Table 1: Knowledge types we consider in this work.

show that KEBLM consistently outperforms the baseline PLMs on all the tasks in all measures. Furthermore, we observe that incorporating more domain knowledge generally leads to greater improvement. This showcases the effectiveness of KEBLM as a general framework for incorporating multiple sources of knowledge, as well as the overall benefits of including explicit domain-specific knowledge to enhance task performance. Note that while primarily focused on the biomedical domain in this work, our framework is highly adaptable and can be easily applied to other domains, such as the bioenergy sector.

In the following parts, we first describe our proposed framework for external knowledge incorporation in Section 2. We then discuss the conducted experiments and their results in Section 3. After that, Section 4 outlines previous related work. Finally, we conclude this work in Section 5.

2. Methods

Figure 2 shows an overview of KEBLM, our proposed framework. KEBLM encodes external domain knowledge using adapter modules [30, 31] (Section 2.1). We pretrain one adapter module for each knowledge source of interest (Section 2.2). As adapter modules are essentially lightweight neural networks, their pretraining process is typically less computationally demanding compared to the standard pretraining of full PLMs.

In this work, we consider three different types of knowledge: (1) entity descriptions, (2) entity-entity relations, and (3) entity synonyms. Table 1

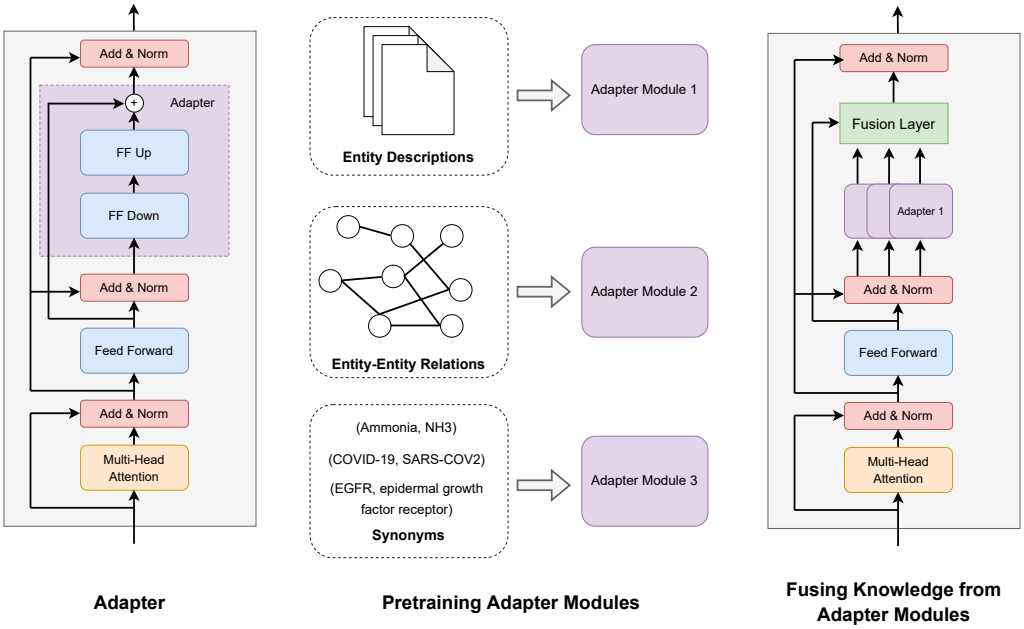


Figure 2: An overview of KEBLM. An adapter module is a group of adapters pretrained together to encode knowledge from a particular source of knowledge.

provides an overview of the knowledge types. As different adapter modules learn to encode different kinds of knowledge and features, we propose to use fusion layers to combine their knowledge for downstream tasks (Section 2.3). Each fusion layer is a parameterized mixer of the available trained adapters [30]. Basically, using fusion layers allows for the identification and activation of the most useful adapters for a given input, as the knowledge from some adapters may be more helpful than others for a specific task/input.

Compared to earlier studies that also integrate knowledge from various sources using adapters [21], our approach specifically includes a knowledge consolidation phase. In this phase, we teach the fusion layers to effectively combine knowledge from the original PLM and newly-acquired external knowledge by using a large collection of unannotated texts. This knowledge consolidation phase is discussed in more detail in Section 2.3.

2.1. Adapter Modules

A common transfer learning technique in NLP is *full fine-tuning*, which involves copying the weights of a PLM and tuning all of them on some downstream task of interest [34]. Despite its effectiveness [35, 1, 36], full fine-tuning can be computationally expensive as the entire PLM needs to be tuned. Adapter modules [29, 30] were introduced as an alternative method for more parameter-efficient adaptation of PLMs. Adapters are small neural networks added between layers of a PLM. During model tuning on a downstream task, only the parameters of the added adapters are updated while the weights of the original PLM are frozen. Therefore, adapter-based tuning adds only a small amount of parameters for each downstream task of interest.

In this work, we use adapter modules to encode external domain knowledge. For each knowledge source of interest, we pretrain an adapter module to memorize the knowledge in it. This approach enables a highly extensible integration of knowledge. When a new source of knowledge emerges, we need to pretrain a new adapter module; however, we do not need to update the parameters of any existing pretrained adapter modules. A related work, DAKI [21], also aims to incorporate domain knowledge from multiple sources using adapters. We compare and contrast our approach with DAKI in greater detail in Section 4.

We use a simple but effective bottleneck architecture for the adapters [30, 31], which is illustrated in the left part of Figure 2. Each layer of a typical Transformer-based PLM contains two primary sub-layers: a multi-head attention layer and a feed-forward layer [37]. In addition, a residual connection is employed around each of the two sub-layers, followed by layer normalization. We insert an adapter after each feed-forward sub-layer and its corresponding `Add & Norm` layer (see Figure 2).

Each adapter first projects the features it receives into a smaller dimension, applies a non-linearity (e.g., ReLU), and then projects the resulting vector back to the original dimension. There is also a skip-connection that connects

the output of the feed-forward layer to the output of the up-projection layer. If we denote the dimension of the hidden states in the backbone PLM as d and the bottleneck dimension as m, then the total number of new parameters in a single adapter is $2md + d + m$. In practice, $m \ll d$, therefore, the number of parameters per adapter is typically small.

Note that we use the term "adapter module" to refer to a group of adapters pretrained together to encode a particular type of knowledge. In other words, for a knowledge source of interest, we first add an adapter to every Transformer layer of the backbone PLM. After that, we pretrain them together on some self-supervised learning task to be discussed in Section 2.2. For example, suppose we want to enhance a PLM consisting of 12 Transformer layers with knowledge from three different sources. Then the number of adapter modules will be 3, and the total number of adapters will be $12 \times 3 = 36$.

In this work, we add adapters to all layers of the backbone PLM because we aim to adapt every single layer to capture the external knowledge. Despite the addition of adapters to all the layers, the computational cost remains low as the number of parameters in each adapter is minimal.

2.2. Adapters Pre-training

In this work, we explore three different types of knowledge: (1) entity descriptions, (2) entity-entity relations, and (3) entity synonyms. These knowledge types are popular and commonly available in various knowledge bases. Table 1 provides an overview of the knowledge types. For each knowledge type, we pretrain an adapter module using a self-supervised learning objective specifically designed for it. During the pretraining process, only the parameters of the adapter modules are updated while the weights of the backbone PLM are frozen. Note that our proposed `KEBLM` is highly extensible and not constrained to only the knowledge types discussed here.

In general, our approach to designing learning objectives is to encourage a model to accurately predict the information contained in a knowledge source of interest. However, the specific form of the objective depends on the

type of knowledge. For instance, since entity descriptions offer rich textual information, employing the masked language modeling objective [1] appears to be a natural choice. On the other hand, for more structured entity relations, a ranking objective would be more suitable, requiring the model to rank all the actual relations present in the knowledge source higher than negatively sampled relations.

Entity descriptions. Biomedical KBs typically have informative descriptions about many different entities. For example, at least 100 million pairs of concepts and corresponding definitions or descriptions can be constructed from UMLS [15]. Biomedical documents typically contain many highly specialized terms, acronyms, and abbreviations. Therefore, knowledge from the description sentences in external KBs can be extremely helpful when trying to comprehend biomedical documents. To this end, we propose to use the masked language modeling (MLM) objective [1] to incorporate knowledge from the description sentences.

The MLM objective is a form of denoising-autoencoding, where the task is to restore a corrupted input sequence. More specifically, given the textual description of some biomedical entity, we first mask some percentage of its tokens at random to produce a corrupted input sequence. The model then needs to predict the masked tokens.

Entity-entity relations. Several biomedical knowledge graphs (KGs), such as UMLS [15] and MSI [33], have a lot of information about the relations between different entities. Let $\mathcal{T} = \{(h, r, t) \mid h, t \in \mathcal{E},\ r \in \mathcal{R}\}$ be the collection of ordered triples in a KG of interest, where $\mathcal{E}$ and $\mathcal{R}$ are the sets of entities and relations (respectively). We aim to pretrain an adapter module to memorize all the relations stored in $\mathcal{T}$.

Concretely, the pre-training task is to train a model to assign high scores to correct positive triples in $\mathcal{T}$ and low scores to triples that are likely to be

incorrect [38]. We use the max-margin loss function:

$$\mathcal{L}(x) = \frac{1}{B}\sum_{i=1}^{B}\max(0, \lambda - f(x) + f(\tilde{x_i})) \tag{1}$$

where $f(.)$ takes a triple as input and returns a score indicating its plausibility. $x \in \mathcal{T}$ is a positive triple. $\tilde{x_i}$ is a negative triple constructed by swapping the head or tail entity of x with a random entity. B is the number of negative samples per positive triple. λ is a margin hyperparameter.

In order to compute $f(x)$, we first convert the triple $x = (h, r, t)$ to a textual sequence Text$[x]$ by concatenating the words in the names of its components. We then use the backbone PLM augmented with the adapter module to transform Text$[x]$ into a feature vector $\mathbf{x}$. Finally, we apply an additional linear layer to $\mathbf{x}$ to get the final plausibility score. This process is summarized as follows:

$$\begin{aligned} \mathbf{x} &= \text{reduce}\big(\phi_{\boldsymbol{\theta}_{\text{PLM}},\boldsymbol{\theta}_{\text{ER}}}\big(\text{Text}[x]\big)\big) \\ f(x) &= \text{FFNN}_s(\mathbf{x}) \end{aligned} \tag{2}$$

where $\phi_{\boldsymbol{\theta}_{\text{PLM}},\boldsymbol{\theta}_{\text{ER}}}(.)$ denotes the entire encoder stack consisting of the PLM and the adapter module. $\boldsymbol{\theta}_{\text{PLM}}$ denotes the parameters of the PLM, while $\boldsymbol{\theta}_{\text{ER}}$ denotes the parameters of the adapter module. reduce(.) is a function that returns the final hidden state of the encoder that corresponds to the first input token. FFNN_s is a feed-forward neural network with a single output dimension. Only $\boldsymbol{\theta}_{\text{ER}}$ and FFNN_s are updated during pretraining.

Entity synonyms. A biomedical KB such as UMLS [15] typically has a comprehensive collection of biomedical synonyms in various forms. For example, the 2020AA version of UMLS has 4M+ concepts and 10M+ synonyms that stem from over 150 controlled vocabularies such as MeSH, SNOMED CT, RxNorm, Gene Ontology, and OMIM [39]. Knowledge of biomedical synonyms can be useful for various downstream tasks such as entity linking [39, 40], information

extraction [41], and paraphrase detection [42]. To this end, we propose to use a contrastive training objective [39] to incorporate knowledge of biomedical synonyms.

Note that it is technically possible to utilize the same ranking objective presented for entity-entity relations for incorporating knowledge of entity synonyms. Nevertheless, we choose the contrastive training objective in this context, as previous research, such as CLIP [43], has demonstrated its efficacy in modeling the similarity between different objects. The contrastive loss function is highly effective for models that need to prioritize the measurement of similar objects over dissimilar ones.

Formally, let $\mathcal{E} = \{e_1, e_2, \ldots, e_K\}$ be the set of all entities in a KB of interest. We assume that each entity e is associated with a set of textual names $\mathcal{N}(e)$. For example, in UMLS, some of the names associated with the entity `C0004057` include "aspirin" and "2-(Acetyloxy)benzoic Acid". If two names are associated with the same entity, then we consider them as synonyms. Finally, let $\mathcal{N}$ denote the set of all names in the KB (i.e., $\mathcal{N} = \cup_{i=1}^{K} \mathcal{N}(e_i)$).

The objective is to learn a function $g : \mathcal{N} \rightarrow \mathbb{R}^d$ that maps each entity name in $\mathcal{N}$ to a feature vector. If n_i and n_j are synonyms, the similarity between $g(n_i)$ and $g(n_j)$ needs to be high (and vice versa). We model the function g as follows:

$$g(n) = \text{reduce}\big(\phi_{\boldsymbol{\theta}_{\text{PLM}}, \boldsymbol{\theta}_{\text{ES}}}\big(\text{Text}[n]\big)\big) \tag{3}$$

where $n \in \mathcal{N}$ is an entity name. Text$[n]$ consists of all the words in the name. $\phi_{\boldsymbol{\theta}_{\text{PLM}}, \boldsymbol{\theta}_{\text{ES}}}(.)$ is the encoder stack consisting of the PLM and the adapter module. $\boldsymbol{\theta}_{\text{ES}}$ denotes the parameters of the adapters.

In this work, we use the contrastive learning framework defined in [44] to train the function g. During pretraining, we freeze $\boldsymbol{\theta}_{\text{PLM}}$ and only update $\boldsymbol{\theta}_{\text{ES}}$. We use the cosine similarity function to evaluate the similarity of any two feature vectors.

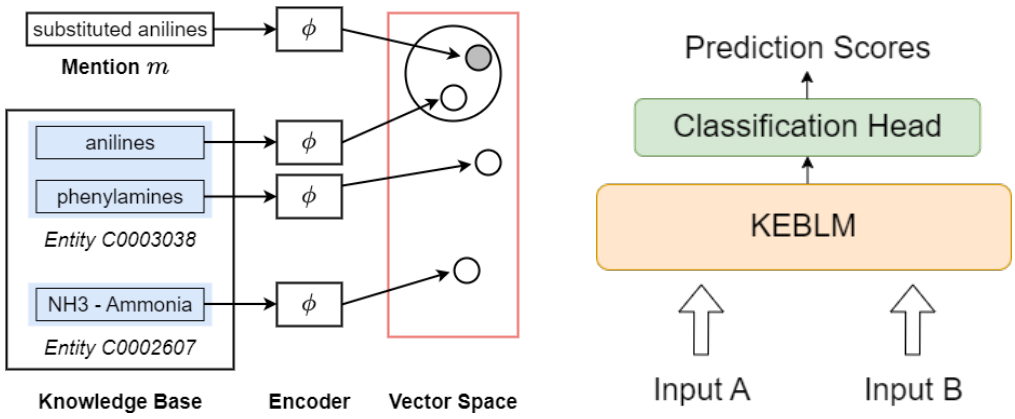
Lastly, it is worth noting that when discussing the incorporation of knowledge types such as entity-entity relations and entity synonyms, we introduce a function Text[x] that essentially transforms an object, like a relation triple or a mention, into a textual sequence. While the strategy for creating input text from such an object could impact the performance, we reserve this exploration for future research and employ only the simple strategies discussed here.

2.3. Knowledge Fusion

Once the adapter modules are pretrained, we use fusion layers to combine their knowledge for downstream tasks. We directly utilize the AdapterFusion mixture layers [30] that can learn to identify and activate the most useful adapters for a given input. We refer readers to [30] for a complete description of the mixture layers.

After incorporating randomly initialized fusion layers into the model stack (see the right part of Figure 2), it becomes possible to fine-tune the entire model for a specific downstream task of interest. However, since each adapter module encodes a distinct type of knowledge, the fusion layers may face difficulties in learning to effectively combine them, particularly if the fine-tuning dataset contains a limited number of examples. To address this issue, we propose conducting a *knowledge consolidation* phase to aid the fusion layers in its learning process.

More specifically, we first gather a large collection of biomedical texts (e.g., publication abstracts) that can be easily obtained from the internet. Next, we attach a masked language modeling (MLM) head to the model stack and train the entire system for the MLM task using the collected corpus. During this process, we freeze the parameters of the backbone PLM and the pretrained adapter modules, allowing only the fusion layers and the newly attached MLM head to be trained. The intuition behind this approach is that it forces the fusion layers to effectively synthesize knowledge from different adapter modules and the original PLM.



(a) Entity linking. (b) Text pair classification.

Figure 3: Our approaches for using KEBLM in downstream tasks such as entity linking (EL), natural language inference (NLI), and question answering (QA). The left part presents a high-level overview of our EL method, while the right part outlines our approaches for NLI and QA.

Following the knowledge consolidation phase, we simply remove the MLM head, rendering the entire stack ready for fine-tuning on any downstream task of interest.

3. Experiments and Results

3.1. Datasets and experimental settings

3.1.1. Downstream tasks

We evaluate our KEBLM on six datasets over three downstream tasks, including four entity linking (EL) datasets, one natural language inference (NLI) dataset, and one question answering (QA) dataset.

Entity Linking. Biomedical EL is the task of mapping entity mentions in a biomedical document to referent entities in a given KB [45, 46]. Our approach to EL is to train an encoder ϕ that encodes mentions and entity names into the same vector space [47, 46]. Before inference, we use ϕ to pre-compute embeddings for all the entity names in the KB. During inference, mentions

are also encoded by ϕ and entities are retrieved using the cosine similarity function. ϕ can be based on an existing PLM (e.g., SciBERT, BioBERT) or our newly proposed `KEBLM`. The left part of Figure 3 provides a high-level overview of our approach to EL.

We use four EL datasets: NCBI-d [48], BC5CDR-c and BC5CDR-d [49], and COMETA [50]. For each dataset, we follow the data split by [39]. We refer the readers to [39, 46] for more information about the problem formulation, the general EL approach, and the datasets.

Natural Language Inference. NLI is the task of determining whether a hypothesis is true (entailment), false (contradiction), or undetermined (neutral) based on a given premise. For example, consider the following premise and hypothesis:

- **Premise**: Watermelon stomach with gastric varices, without bleed in more than 2 years
- **Hypothesis**: Patient has hematemesis.

Informally speaking, "hematemesis" refers to the vomiting of blood. Therefore, the hypothesis contradicts the premise.

Compared to the general domain, there are relatively fewer studies on NLI in the biomedical domain [51, 52]. In this work, we simply formulate the task as a text pair classification problem (see the right part of Figure 3). Specifically, for each example, we concatenate the premise and hypothesis into a single input sequence and feed it into a Transformer-based model, such as `KEBLM`, with a classification head.

We train and evaluate NLI models on MedNLI [51], an NLI dataset consisting of sentence pairs extracted from MIMIC-III, a comprehensive clinical database. MedNLI has 11,232 premise-hypothesis pairs in the training set, 1,395 pairs in the development set, and 1,422 pairs in the test set.

Question Answering. We also evaluate `KEBLM` on the task of question answering (QA) using the PubMedQA dataset [53]. PubMedQA contains a

Dataset	Train Examples	Dev Examples	Test Examples	Estimated Max. Input Length (Words)
NCBI-d	5,134	787	960	12
BC5CDR-c	5,203	5,347	5,385	26
BC5CDR-d	4,182	4,244	4,424	17
COMETA	13,489	2,176	4,350	6
MedNLI	11,232	1,395	1,422	207
PubMedQA	450	50	500	487

Table 2: Statistics of the downstream task datasets. The table shows the number of train, dev, and test examples, as well as the estimated maximum input length (measured in terms of the number of words) for each dataset.

collection of research questions and corresponding reference texts taken from PubMed abstracts, each of which is labeled with whether the text contains the answer to the research question (yes/maybe/no). We use the original train/dev/test split, which consists of 450/50/500 questions, respectively, for our experiments. Similar to NLI, we also formulate QA as a text pair classification problem (see the right part of Figure 3). For each example, the input is the concatenation of the question and the reference text.

Table 2 presents a summary of the statistics for all the downstream task datasets. To ensure compatibility with the underlying PLM, inputs that exceed the maximum token length are truncated. Specifically, after preprocessing and tokenization, inputs are truncated to retain only the first N tokens, where N represents the maximum token limit allowed by the PLM.

3.1.2. Pretraining setup

In this work, we use BioBERT [11] and SciBERT [10] as our base PLMs. For each base PLM, we pretrain three different adapter modules to incorporate three different types of knowledge (see Table 1 for an overview): (1) entity descriptions, (2) entity-entity relations, and (3) entity synonyms.

We utilize UMLS [15] and PubChem [16] to gather entity descriptions. More specifically, following the procedure of [32], we first collect over 100 mil-

lion pairs of concepts and corresponding definitions or descriptions from UMLS [15]. In addition, we extract approximately 102,980 compound-description pairs from PubChem [54]. After combining the descriptions from the two sources and doing some filtering (e.g., removing PubChem descriptions that are too short), our final set consists of 130 million descriptions or definitions of a diverse range of entities from the chemistry and biomedical domains, including molecules, genes, and diseases.

We utilize a knowledge graph (KG) called **MSI** [33] to collect entity-entity relationships. This recent network encompasses diseases, proteins, genes, drug targets, and biological functions. Overall, MSI comprises 29,959 entities, 6 relation types, and allows for the extraction of 484,654 positive triples.

To collect information of biomedical synonyms from UMLS [15], we use the same procedure employed for pretraining SAPBERT [39].

Finally, we gather more than 30 million abstracts from PubMed for the knowledge consolidation phase. During the knowledge consolidation phase, the parameters of the base PLM are frozen, while only the parameters of the fusion layers and a new MLM head are updated. As a result, the knowledge consolidation process should be considerably less computationally demanding compared to the standard pretraining of full PLMs on all these abstracts.

To implement the adapters, we utilize the AdapterHub framework[1] [31]. Each adapter has a bottleneck architecture, as proposed by [55], which corresponds to the `PfeifferConfig` in AdapterHub. We initialize each adapter using the default parameters provided by `PfeifferConfig`, with the exception of the reduction factor, which we set to 4.

3.1.3. Hyperparameters

For fine-tuning on EL datasets, we utilize our existing codebase[2] that was previously used in a different study on biomedical EL [46]. We explore a

[1] `https://docs.adapterhub.ml/index.html`
[2] `https://github.com/laituan245/rescnn_bioel`

Models	Top-1 Accuracy (on test sets)			
	NCBI-d	BC5CDR-d	BC5CDR-c	COMETA
Previous SOTA Methods				
BioSyn [56]	91.1	93.2	96.6	71.3
SapBERT [39]	92.5	**93.8**	**96.8**	77.0
ResCNN (Max Pooling) [46]	92.4	93.1	**96.8**	80.1
BioBERT	92.0	93.3	96.2	80.6
KEBLM (BioBERT) - Ours	93.2$^{\uparrow}$	93.7$^{\uparrow}$	96.6$^{\uparrow}$	**80.8**$^{\uparrow}$
SciBERT	91.5	93.0	96.2	77.3
KEBLM (SciBERT) - Ours	**93.5**$^{\uparrow}$	93.3$^{\uparrow}$	96.5$^{\uparrow}$	77.8$^{\uparrow}$

Table 3: Overall test results on the four biomedical EL datasets. The best ones are highlighted in **bold**, while "↑" denotes that improvements are observed when comparing KEBLM with the corresponding baseline model. All observed improvements are statistically significant with a p-value < 0.05.

range of parameter values for the fine-tuning process, including lower learning rates of {1e-5, 5e-5}, upper learning rates of $\{0.001, 0.0001\}$, batch sizes of $\{64, 128\}$, and training epochs of $\{25, 50\}$. The lower learning rate is applied to update the backbone LM, adapter modules, and fusion layers (if applicable), while the upper learning rate is designated for updating other parameters in the entire model stack.

For fine-tuning on NLI and QA datasets, the optimal values are variant-specific. We experiment with the following range of possible values: a learning rate of $\{1e-5, 5e-5\}$, a batch size of $\{8, 16, 32\}$, the number of training epochs set to $\{10, 25, 50\}$, and a weight decay of $\{0, 0.01\}$. The maximum number of input tokens is set to 512. For each variant, we evaluate the test performance of the checkpoint that achieves the best score on the designated development set.

3.2. Performance on downstream tasks

Table 3 shows the performance of various entity linking (EL) models. We observe that KEBLM consistently improves the performance of both BioBERT and SciBERT on all datasets. Table 4 presents the overall results on the NLI and QA datasets. Similar to the EL results, KEBLM is also effective in

	MedNLI (Accuracy)	PubMedQA (Accuracy)
SciBERT - Ours	80.59	55.2
KEBLM (SciBERT) - Ours	**82.14**	**59.0**
SciBERT + MoP [19]	81.43	54.78
BioBERT - Ours	82.21	62.2
KEBLM (BioBERT) - Ours	**84.24**	**68.0**
BioBERT + MoP [19]	83.44	61.82
BioBERT + DAKI [21]	83.41	-

Table 4: Overall test results on the NLI and QA datasets. DAKI [21] did not use PubMedQA in their study.

Models	Top-1 Accuracy (on development sets)			
	NCBI-d	BC5CDR-d	BC5CDR-c	COMETA
BioBERT	94.3	93.5	98.2	80.3
KEBLM **(BioBERT) - Ours**	94.5	93.7	98.3	80.7
SciBERT	92.2	92.5	97.6	77.2
KEBLM **(SciBERT) - Ours**	94.2	92.6	98.2	77.3

Table 5: Overall results on the development sets of the four biomedical EL datasets.

enhancing the performance of the base PLMs. Furthermore, in Table 4, we compare KEBLM to MoP [19] and DAKI [21], which are previous methods that also incorporate external knowledge. KEBLM consistently outperforms these competing methods in terms of absolute performance scores. It is worth noting that, similar to our study, DAKI [21] aims to incorporate domain knowledge from multiple sources using adapters. However, a key distinction lies in our method's inclusion of an explicit knowledge consolidation phase.

For a more comprehensive comparison of the models, Table 5 and Table 6 also present the results on the development sets. Overall, these results align well with those from the test sets. For instance, based on the development set scores, it is evident that our proposed framework effectively integrates diverse external knowledge types, leading to improved performance in the target tasks.

	MedNLI (Accuracy)	PubMedQA (Accuracy)
SciBERT - Ours	82.94	54.0
KEBLM **(SciBERT) - Ours**	83.01	56.0
BioBERT - Ours	83.58	66.0
KEBLM **(BioBERT) - Ours**	84.66	70.0

Table 6: Overall results on the development sets of NLI and QA datasets.

3.3. Analysis

Ablation Study. We thoroughly examine the impact of the knowledge modules through an ablation study and present the findings in Table 7. Evidently, the knowledge consolidation phase plays a crucial role in enabling the fusion layers to effectively integrate knowledge from various adapter modules. A noticeable decline in performance is observed when the knowledge consolidation phase is omitted. Moreover, we find that, generally, incorporating more external knowledge leads to more improvement. When incorporating knowledge from two or more sources, we can achieve better performance on MedNLI than when we leverage only one single knowledge source. These results demonstrate the effectiveness of KEBLM as a general framework for incorporating multiple sources of knowledge.

Qualitative Analysis. We attempted to manually examine some predictions made by both our knowledge-enhanced models and the baseline models. Generally, it is not always possible to ascertain the exact reasons why a model made an error, given the inherent complexity of each model with hundreds of millions of parameters and the fact that interpretable machine learning remains an active area of research. As a result, it is not always straightforward to determine when knowledge proves beneficial. Nevertheless, to gain some insights, we will present instances where knowledge is evidently helpful.

First, we provide some qualitative analyses to demonstrate the strengths of our models over the baseline models in Table 8.

In the first example, which comes from MedNLI, the baseline model

	MedNLI (Accuracy)
KEBLM (BioBERT)	**84.24**
without KC	83.90
without KC and ED	83.68
without KC and ER	83.61
without KC and ES	83.26
without KC and [ED, ER]	82.63
without KC and [ED, ES]	82.42
without KC and [ER, ES]	83.19
BioBERT (without any external knowledge)	82.21

Table 7: Ablation analysis. Here, KC refers to the knowledge consolidation phase. In addition, ED, ER, and ES refer to the adapter modules that encode entity descriptions, entity-entity relations, and entity synonyms, respectively.

incorrectly predicts the relation between the given premise and hypothesis to be "neutral." This is likely because the baseline model does not understand the technical term "hematemesis," which refers to the vomiting of blood (informally speaking). However, the definition of this term is readily available on UMLS and is also part of our pretraining data. As such, it is likely the reason why KEBLM is able to correctly predict that the relation should be "contradiction."

In the second example related to QA, the reference text is a long abstract that does not provide an explicit yes or no answer to the given question. Instead, a large part of it discusses conducted analyses and numerical results (not shown in Table 8 due to space constraints), which can make it difficult for an automatic model to determine the correct answer. As a result, our baseline model, which does not incorporate external knowledge, incorrectly predicts the answer to be "no." However, by looking at the definition of "spasticity" in UMLS, we can see that it is a form of muscle disorder. With this knowledge, we can guess that reducing spasticity is likely to increase functional benefit, even without reading the abstract. This is likely the reason why KEBLM is

Input	Label	Model Predictions
Task: Natural Language Inference (MedNLI) **Premise**: Watermelon stomach with gastric varices, without bleed in more than 2 years. **Hypothesis**: Patient has hematemesis.	Contradiction	SciBERT: Neural KEBLM (SciBERT): Contradiction
Task: Question Answering (PubMedQA) **Question**: Does reducing spasticity translate into functional benefit? **Reference Text**: Spasticity and loss of function in an affected arm are common after stroke. Although botulinum toxin is used to reduce spasticity, its functional benefits are less easily demonstrated. This paper reports an exploratory meta-analysis to ...	Yes	BioBERT: No KEBLM (BioBERT): Yes
Task: Entity Linking (COMETA) **Mention and its context**: I was recomended the **5HTP** and as I said it initially worked but havn ' t been taking it since.	Oxitriptan (substance)	SciBERT: Azacitidine KEBLM (SciBERT): Oxitriptan

Table 8: Examples showing how external knowledge improves prediction accuracy.

able to return the correct answer of "yes," since one of the knowledge types we consider is definition sentences from UMLS.

The final example in Table 8, taken from the COMETA corpus for medical entities in social media, is challenging for our baseline model to handle. The context in this example, as is typical in tweets or Facebook posts, is relatively short. Additionally, the surface form of the target mention, "5HTP," differs from all the names of the correct entity stored in COMETA's KB. In contrast, by utilizing synonym information from UMLS, KEBLM can easily identify the correct entity and rank it at the top.

4. Related Work

In recent years, there have been many studies that explicitly aim to inject external knowledge into PLMs [57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67]. A promising direction is to utilize structured knowledge bases (KBs) to augment Transformer-based PLMs. Some notable studies include ERNIE [59] and

KnowBERT [58], where the entity information from KBs is explicitly linked with the input text during pre-training yielding entity-enhanced variants of BERT models. For ERNIE and KnowBERT to work, an entity linker is required to connect the input text to information in the external KBs. In contrast, `KEBLM`, our proposed framework, does not have such a requirement. This characteristic makes `KEBLM` applicable to new scientific domains that do not have any high-quality entity linkers.

Another line of work adopts the retrieve-and-read framework [57, 60, 61, 63, 67]. Typically, given an input of some NLP task, a retrieval component first retrieves potentially relevant text snippets (e.g., sentences or paragraphs) from a corpus (e.g., Wikipedia). After that, another model produces the final output conditioned on the original input and the retrieved information. While advancing the state-of-the-art of many knowledge-intensive tasks, most methods in this direction focus only on retrieving information from Wikipedia [57, 60, 67]. This is different from our `KEBLM`, which leverages knowledge from multiple sources.

Compared to the general domain, there have been fewer studies on incorporating external knowledge into biomedical models [68, 19, 20, 21, 69]. For instance, UmlsBERT [69] is a contextual embedding model that integrates clinical domain knowledge from the UMLS Metathesauru during the pre-training process. Another noteworthy study is that of SAPBERT [39], which is a pre-training scheme designed to learn information from a collection of biomedical synonyms from UMLS. In our research, we not only utilize knowledge from UMLS but also incorporate information from other knowledge bases, such as PubChem [16] and MSI [33]. Additionally, our study differs from SAPBERT in that we not only evaluate our proposed `KEBLM` on entity linking tasks but also on other downstream tasks, such as NLI and QA.

Mixture-of-Partitions (MoP) [19] is a novel approach for infusing knowledge by partitioning knowledge graphs into smaller sub-graphs. While MoP focuses only on knowledge triples of *(subject, relation, object)*, our `KEBLM` incorporates

a broader range of knowledge types. In this work, our specific instantiation of KEBLM also considers knowledge such as entity descriptions and synonyms in addition to entity relations. This makes KEBLM a more general and versatile approach to knowledge infusion. Additionally, KEBLM can be easily extended to apply the idea of graph partitioning of MoP when incorporating knowledge of entity relations.

A closely related work, DAKI [21], also aims to integrate domain knowledge from multiple sources using adapters. Unlike DAKI, our method explicitly includes a knowledge consolidation phase (see Section 2.3). During this phase, we train the fusion layers to effectively combine knowledge from both the original PLM and newly acquired external knowledge by utilizing a vast collection of unannotated texts. Furthermore, different from our work, DAKI does not explicitly incorporate synonym information from external KBs, which can be extremely useful for tasks such as entity linking. Our experimental results on MedNLI also suggest that KEBLM can be more effective than DAKI in incorporating external knowledge into biomedical PLMs (refer to Section 3).

5. Conclusions and Future Work

This work proposes KEBLM, a general framework for incorporating various types of domain knowledge from many sources into biomedical PLMs. Extensive experiments show that KEBLM is highly effective as it can consistently improve the performance of the underlying PLMs. In the future, we plan to extend KEBLM to incorporate other types of complex knowledge, such as molecule structures, and to explore its use for incorporating knowledge from general-domain KBs to tackle general-domain NLP tasks.

Acknowledgement

This research is based upon work supported by the Molecule Maker Lab Institute: An AI Research Institutes program supported by NSF under

Award No. 2019897, NSF No. 2034562, Agriculture and Food Research Initiative (AFRI) grant no. 2020-67021-32799/project accession no.1024178 from the USDA National Institute of Food and Agriculture, and NSF Award No. 2034562. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes, notwithstanding any copyright annotation therein.

This work was also funded by the DOE Center for Advanced Bioenergy and Bioproducts Innovation (U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under Award Number DE-SC0018420). Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the U.S. Department of Energy.

In addition, this research used the Delta advanced computing and data resource which is supported by the National Science Foundation (award OAC 2005572) and the State of Illinois. Delta is a joint effort of the University of Illinois Urbana-Champaign and its National Center for Supercomputing Applications.

References

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

[2] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou,

W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (140) (2020) 1–67.
URL http://jmlr.org/papers/v21/20-074.html

[3] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[4] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, Albert: A lite bert for self-supervised learning of language representations, in: International Conference on Learning Representations, 2020.
URL https://openreview.net/forum?id=H1eA7AEtvS

[5] K. Clark, U. Khandelwal, O. Levy, C. D. Manning, What does BERT look at? an analysis of BERT's attention, in: Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Association for Computational Linguistics, Florence, Italy, 2019, pp. 276–286. doi:10.18653/v1/W19-4828.
URL https://aclanthology.org/W19-4828

[6] A. Rogers, O. Kovaleva, A. Rumshisky, A Primer in BERTology: What We Know About How BERT Works, Transactions of the Association for Computational Linguistics 8 (2021) 842–866. arXiv:https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00349/1923281/tacl_a_00349.pdf, doi:10.1162/tacl_a_00349.
URL https://doi.org/10.1162/tacl_a_00349

[7] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing

Systems, Vol. 32, Curran Associates, Inc., 2019.
URL https://proceedings.neurips.cc/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf

[8] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, O. Levy, SpanBERT: Improving pre-training by representing and predicting spans, Transactions of the Association for Computational Linguistics 8 (2020) 64–77. doi:10.1162/tacl_a_00300.
URL https://aclanthology.org/2020.tacl-1.5

[9] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, Electra: Pre-training text encoders as discriminators rather than generators, ArXiv abs/2003.10555 (2020).

[10] I. Beltagy, K. Lo, A. Cohan, SciBERT: A pretrained language model for scientific text, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3615–3620. doi:10.18653/v1/D19-1371.
URL https://aclanthology.org/D19-1371

[11] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (4) (2019) 1234–1240. arXiv:https://academic.oup.com/bioinformatics/article-pdf/36/4/1234/32527770/btz682.pdf, doi:10.1093/bioinformatics/btz682.
URL https://doi.org/10.1093/bioinformatics/btz682

[12] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, ACM Trans. Comput. Healthcare 3 (1)

(oct 2021). doi:10.1145/3458754.
URL https://doi.org/10.1145/3458754

[13] W. Ammar, D. Groeneveld, C. Bhagavatula, I. Beltagy, M. Crawford, D. Downey, J. Dunkelberger, A. Elgohary, S. Feldman, V. Ha, R. Kinney, S. Kohlmeier, K. Lo, T. Murray, H.-H. Ooi, M. Peters, J. Power, S. Skjonsberg, L. Wang, C. Wilhelm, Z. Yuan, M. van Zuylen, O. Etzioni, 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), Association for Computational Linguistics, New Orleans - Louisiana, 2018, pp. 84–91. doi:10.18653/v1/N18-3011.
URL https://aclanthology.org/N18-3011

[14] P. Lewis, M. Ott, J. Du, V. Stoyanov, Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art, in: Proceedings of the 3rd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Online, 2020, pp. 146–157. doi:10.18653/v1/2020.clinicalnlp-1.17.
URL https://aclanthology.org/2020.clinicalnlp-1.17

[15] O. Bodenreider, The unified medical language system (umls): integrating biomedical terminology, Nucleic acids research 32 Database issue (2004) D267–70.

[16] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Y. Zaslavsky, J. Zhang, E. E. Bolton, Pubchem in 2021: new data content and improved web interfaces, Nucleic Acids Research 49 (2021) D1388 – D1395.

[17] L. Bornmann, R. Mutz, Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references, Journal of the Association for Information Science and Technology 66 (2015).

[18] M. E. Peters, M. Neumann, R. Logan, R. Schwartz, V. Joshi, S. Singh, N. A. Smith, Knowledge enhanced contextual word representations, 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, 2019, pp. 43–54. doi:10.18653/v1/D19-1005.
URL https://aclanthology.org/D19-1005

[19] Z. Meng, F. Liu, T. Clark, E. Shareghi, N. Collier, Mixture-of-partitions: Infusing large biomedical knowledge graphs into BERT, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 4672–4681. doi:10.18653/v1/2021.emnlp-main.383.
URL https://aclanthology.org/2021.emnlp-main.383

[20] T. Lai, H. Ji, C. Zhai, Q. H. Tran, Joint biomedical entity and relation extraction with knowledge-enhanced collective inference, 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, 2021, pp. 6248–6260. doi:10.18653/v1/2021.acl-long.488.
URL https://aclanthology.org/2021.acl-long.488

[21] Q. Lu, D. Dou, T. H. Nguyen, Parameter-efficient domain knowledge integration from multiple sources for biomedical pre-trained language models, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 3855–3865. doi:10.18653/v1/2021.findings-emnlp.325.
URL https://aclanthology.org/2021.findings-emnlp.325

[22] M. Sung, J. Lee, S. Yi, M. Jeon, S. Kim, J. Kang, Can language models be biomedical knowledge bases?, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 4723–4734. doi:10.18653/v1/2021.emnlp-main.388.
URL https://aclanthology.org/2021.emnlp-main.388

[23] Z. Yao, Y. Cao, Z. Yang, V. Deshpande, H. Yu, Extracting biomedical factual knowledge using pretrained language model and electronic health record context, arXiv preprint arXiv:2209.07859 (2022).

[24] Z. Zhang, N. Parulian, H. Ji, A. Elsayed, S. Myers, M. Palmer, Fine-grained information extraction from biomedical literature based on knowledge-enriched Abstract Meaning Representation, 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, 2021, pp. 6261–6270. doi:10.18653/v1/2021.acl-long.489.
URL https://aclanthology.org/2021.acl-long.489

[25] H. Khachatrian, L. Nersisyan, K. Hambardzumyan, T. Galstyan, A. Hakobyan, A. Arakelyan, A. Rzhetsky, A. Galstyan, BioRelEx 1.0: Biological relation extraction benchmark, in: Proceedings of the 18th BioNLP Workshop and Shared Task, Association for Computational Linguistics, Florence, Italy, 2019, pp. 176–190. doi:10.18653/v1/W19-5019.
URL https://aclanthology.org/W19-5019

[26] Z. Yuan, Y. Liu, C. Tan, S. Huang, F. Huang, Improving biomedical pretrained language models with knowledge, in: Proceedings of the 20th Workshop on Biomedical Language Processing, Association for Computational Linguistics, Online, 2021, pp. 180–190. doi:10.18653/

v1/2021.bionlp-1.20.
URL https://aclanthology.org/2021.bionlp-1.20

[27] Q. Xie, J. A. Bishop, P. Tiwari, S. Ananiadou, Pre-trained language models with domain knowledge for biomedical extractive summarization, Knowledge-Based Systems 252 (2022) 109460.

[28] C. J. Mattingly, M. C. Rosenstein, A. P. Davis, G. T. Colby, J. N. Forrest, J. L. Boyer, The comparative toxicogenomics database: a cross-species resource for building chemical-gene interaction networks., Toxicological sciences : an official journal of the Society of Toxicology 92 2 (2006) 587–95.

[29] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, S. Gelly, Parameter-efficient transfer learning for nlp, in: ICML, 2019.

[30] J. Pfeiffer, A. Kamath, A. Rücklé, K. Cho, I. Gurevych, Adapterfusion: Non-destructive task composition for transfer learning, arXiv preprint arXiv:2005.00247 (2020).

[31] J. Pfeiffer, A. Rücklé, C. Poth, A. Kamath, I. Vulić, S. Ruder, K. Cho, I. Gurevych, Adapterhub: A framework for adapting transformers, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 46–54.

[32] F. Remy, K. Demuynck, T. Demeester, BioLORD: Learning ontological representations from definitions for biomedical concepts and their textual descriptions, in: Findings of the Association for Computational Linguistics: EMNLP 2022, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 1454–1465.
URL https://aclanthology.org/2022.findings-emnlp.104

[33] C. Ruiz, M. Zitnik, J. Leskovec, Identification of disease treatment mechanisms through the multiscale interactome, Nature Communications 12 (2020).

[34] J. Howard, S. Ruder, Universal language model fine-tuning for text classification, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 328–339. doi:10.18653/v1/P18-1031.
URL https://aclanthology.org/P18-1031

[35] Q. Chen, Z. Zhuo, W. Wang, Bert for joint intent classification and slot filling, arXiv preprint arXiv:1902.10909 (2019).

[36] T. M. Lai, Q. H. Tran, T. Bui, D. Kihara, A simple but effective bert model for dialog state tracking on resource-limited systems, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 8034–8038.

[37] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, ArXiv abs/1706.03762 (2017).

[38] R. Nadkarni, D. Wadden, I. Beltagy, N. A. Smith, H. Hajishirzi, T. Hope, Scientific language models for biomedical knowledge base completion: An empirical study, ArXiv abs/2106.09700 (2021).

[39] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, N. Collier, Self-alignment pretraining for biomedical entity representations, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 4228–4238. doi:10.18653/v1/2021.naacl-main.334.
URL https://aclanthology.org/2021.naacl-main.334

[40] H. Yuan, Z. Yuan, S. Yu, Generative biomedical entity linking via knowledge base-guided pre-training and synonyms-aware fine-tuning, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 4038–4048. doi:10.18653/v1/2022.naacl-main.296.
URL https://aclanthology.org/2022.naacl-main.296

[41] K. Fundel, R. Küffner, R. Zimmer, RelEx—Relation extraction using dependency parse trees, Bioinformatics 23 (3) (2006) 365–371. arXiv:https://academic.oup.com/bioinformatics/article-pdf/23/3/365/647909/btl616.pdf, doi:10.1093/bioinformatics/btl616.
URL https://doi.org/10.1093/bioinformatics/btl616

[42] K. Dey, R. Shrivastava, S. Kaushik, A paraphrase and semantic similarity detection system for user generated short-text content on microblogs, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 2880–2890.
URL https://aclanthology.org/C16-1271

[43] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.

[44] X. Wang, X. Han, W. Huang, D. Dong, M. R. Scott, Multi-similarity loss with general pair weighting for deep metric learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5022–5030.

[45] J. Zheng, D. P. Howsmon, B. Zhang, J. Hahn, D. L. McGuinness, J. A.

Hendler, H. Ji, Entity linking for biomedical literature, BMC Medical Informatics and Decision Making 15 (2014) S4 – S4.

[46] T. Lai, H. Ji, C. Zhai, Bert might be overkill: A tiny but effective biomedical entity linker based on residual convolutional neural networks, in: Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 1631–1639.

[47] D. Gillick, S. Kulkarni, L. Lansing, A. Presta, J. Baldridge, E. Ie, D. Garcia-Olano, Learning dense representations for entity retrieval, in: Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 528–537. doi:10.18653/v1/K19-1049.
URL https://www.aclweb.org/anthology/K19-1049

[48] R. I. Dogan, R. Leaman, Z. Lu, Ncbi disease corpus: A resource for disease name recognition and concept normalization, Journal of biomedical informatics 47 (2014) 1–10.

[49] J. Li, Y. Sun, R. J. Johnson, D. Sciaky, C.-H. Wei, R. Leaman, A. P. Davis, C. Mattingly, T. C. Wiegers, Z. Lu, Biocreative v cdr task corpus: a resource for chemical disease relation extraction, Database: The Journal of Biological Databases and Curation 2016 (2016).

[50] M. Basaldella, F. Liu, E. Shareghi, N. Collier, COMETA: A corpus for medical entity linking in the social media, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 3122–3137. doi:10.18653/v1/2020.emnlp-main.253.
URL https://www.aclweb.org/anthology/2020.emnlp-main.253

[51] A. Romanov, C. Shivade, 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, 2018, pp. 1586–1596. doi:10.18653/v1/D18-1187.
URL https://www.aclweb.org/anthology/D18-1187

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

[53] Q. Jin, B. Dhingra, Z. Liu, W. Cohen, X. Lu, PubMedQA: A dataset for biomedical research question answering, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 2567–2577. doi:10.18653/v1/D19-1259.
URL https://aclanthology.org/D19-1259

[54] C. Edwards, C. Zhai, H. Ji, Text2Mol: Cross-modal molecule retrieval with natural language queries, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 595–607. doi:10.18653/v1/2021.emnlp-main.47.
URL https://aclanthology.org/2021.emnlp-main.47

[55] J. Pfeiffer, I. Vulić, I. Gurevych, S. Ruder, MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing

(EMNLP), Association for Computational Linguistics, Online, 2020, pp. 7654–7673. doi:10.18653/v1/2020.emnlp-main.617.
URL https://aclanthology.org/2020.emnlp-main.617

[56] M. Sung, H. Jeon, J. Lee, J. Kang, Biomedical entity representations with synonym marginalization, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 3641–3650. doi:10.18653/v1/2020.acl-main.335.
URL https://aclanthology.org/2020.acl-main.335

[57] E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli, J. Weston, Wizard of wikipedia: Knowledge-powered conversational agents, ArXiv abs/1811.01241 (2019).

[58] M. E. Peters, M. Neumann, R. Logan, R. Schwartz, V. Joshi, S. Singh, N. A. Smith, Knowledge enhanced contextual word representations, 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, 2019, pp. 43–54. doi:10.18653/v1/D19-1005.
URL https://aclanthology.org/D19-1005

[59] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, Q. Liu, ERNIE: Enhanced language representation with informative entities, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 1441–1451. doi:10.18653/v1/P19-1139.
URL https://aclanthology.org/P19-1139

[60] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W.-t. Yih, Dense passage retrieval for open-domain question answering, in:

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 6769–6781. doi:10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550

[61] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in Neural Information Processing Systems 33 (2020) 9459–9474.

[62] T. Févry, L. Baldini Soares, N. FitzGerald, E. Choi, T. Kwiatkowski, Entities as experts: Sparse memory access with entity supervision, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 4937–4951. doi:10.18653/v1/2020.emnlp-main.400. URL https://aclanthology.org/2020.emnlp-main.400

[63] G. Izacard, E. Grave, Leveraging passage retrieval with generative models for open domain question answering, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 874–880. doi:10.18653/v1/2021.eacl-main.74. URL https://aclanthology.org/2021.eacl-main.74

[64] F. Sun, F.-L. Li, R. Wang, Q. Chen, X. Cheng, J. Zhang, K-aid: Enhancing pre-trained language models with domain knowledge for question answering, Proceedings of the 30th ACM International Conference on Information & Knowledge Management (2021).

[65] O. Agarwal, H. Ge, S. Shakeri, R. Al-Rfou, Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training, in: Proceedings of the 2021 Conference of the North American

Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 3554–3565. doi:10.18653/v1/2021.naacl-main.278.
URL https://aclanthology.org/2021.naacl-main.278

[66] J. Kaur, S. Bhatia, M. Aggarwal, R. Bansal, B. Krishnamurthy, LM-CORE: Language models with contextually relevant external knowledge, in: Findings of the Association for Computational Linguistics: NAACL 2022, Association for Computational Linguistics, Seattle, United States, 2022, pp. 750–769. doi:10.18653/v1/2022.findings-naacl.57.
URL https://aclanthology.org/2022.findings-naacl.57

[67] F. Petroni, A. Piktus, A. Fan, P. Lewis, M. Yazdani, N. De Cao, J. Thorne, Y. Jernite, V. Karpukhin, J. Maillard, V. Plachouras, T. Rocktäschel, S. Riedel, KILT: a benchmark for knowledge intensive language tasks, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 2523–2544. doi:10.18653/v1/2021.naacl-main.200.
URL https://aclanthology.org/2021.naacl-main.200

[68] Y. He, Z. Zhu, Y. Zhang, Q. Chen, J. Caverlee, Infusing Disease Knowledge into BERT for Health Question Answering, Medical Inference and Disease Name Recognition, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 4604–4614. doi:10.18653/v1/2020.emnlp-main.372.
URL https://aclanthology.org/2020.emnlp-main.372

[69] G. Michalopoulos, Y. Wang, H. Kaka, H. Chen, A. Wong, UmlsBERT: Clinical domain knowledge augmentation of contextual embeddings using the Unified Medical Language System Metathesaurus, in: Proceedings of

the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 1744–1753. doi:10.18653/v1/2021.naacl-main.139.
URL https://aclanthology.org/2021.naacl-main.139