BERT-MK: Integrating Graph Contextualized Knowledge into Pre-trained Language Models

Bin He[1], **Di Zhou**[1], **Jinghui Xiao**[1], **Xin Jiang**[1], **Qun Liu**[1], **Nicholas Jing Yuan**[2], **Tong Xu**[3]
[1]Huawei Noah's Ark Lab
[2]Huawei Cloud & AI
[3]School of Computer Science, University of Science and Technology of China
{hebin.nlp, zhoudi7, xiaojinghui4, jiang.xin, qun.liu, nicholas.yuan}@huawei.com, tongxu@ustc.edu.cn

Abstract

Complex node interactions are common in knowledge graphs (KGs), and these interactions can be considered as contextualized knowledge exists in the topological structure of KGs. Traditional knowledge representation learning (KRL) methods usually treat a single triple as a training unit, neglecting the usage of graph contextualized knowledge. To utilize these unexploited graph-level knowledge, we propose an approach to model subgraphs in a medical KG. Then, the learned knowledge is integrated with a pre-trained language model to do the knowledge generalization. Experimental results demonstrate that our model achieves the state-of-the-art performance on several medical NLP tasks, and the improvement above MedERNIE indicates that graph contextualized knowledge is beneficial.

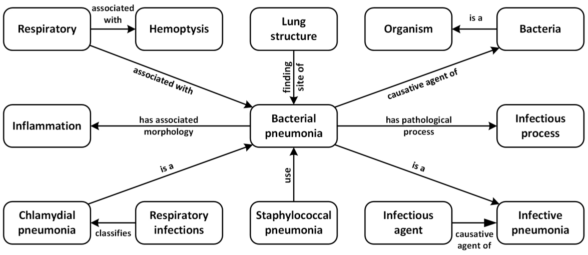


Figure 1: A subgraph extracted from a medical knowledge graph. The rectangles represent entities and directed arrows denote relations.

1 Introduction

In 1954, Harris (1954) proposed a distributional hypothesis that words occur in the same contexts tend to have similar meanings. Firth (1957) explained the context-dependent nature of meaning in linguistics by his famous quotation "you shall know a word by the company it keeps" . Although the above-mentioned distributional hypothesis is proposed for language models, if we look at the knowledge graph from the perspective of this hypothesis, we can find that similar hypothesis exists in knowledge graphs (KGs). We call it **KG distributional hypothesis**: *you shall know an entity by the relationships it involves.*

Given this hypothesis, contextualized information in language models can be mapped to knowledge graphs, which we call "**graph contextualized knowledge**". Figure 1 illustrates a knowledge subgraph that includes several medical entities. In this figure, four incoming and four outgoing neighboring nodes (hereinafter called "in-entity" and "out-entity") of node "*Bacterial pneumonia*" are linked by various relation paths. These linked nodes and correlations can be seen as "graph contextualized information" of entity node "*Bacterial pneumonia*". In this study, we will explore how to integrate graph contextualized knowledge into pre-trained language models.

Pre-trained language models learn contextualized word representations on large-scale text corpus through self-supervised learning methods, and obtain new state-of-the-art (SOTA) results on most downstream tasks (Peters et al., 2018; Radford et al., 2018; Devlin et al., 2019). This gradually becomes a new paradigm for natural language processing research. Recently, several knowledge-enhanced pre-trained language models have been proposed, such as ERNIE-Baidu (Sun et al., 2019), ERNIE-Tsinghua (Zhang et al., 2019a), WKLM (Xiong et al., 2019) and K-ADAPTER (Wang et al., 2020).

In this study, since we need to learn graph contextualized knowledge in a large-scale medical knowledge graph, ERNIE-Tsinghua (hereinafter called "ERNIE") is chosen as our backbone model. In ERNIE, entity embeddings are learned by TransE (Bordes et al., 2013), which is a popular transition-based method for knowledge representation learning (KRL). However, TransE cannot deal with the modeling of complex relations (Lin et al.,

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 2281–2290
November 16 - 20, 2020. ©2020 Association for Computational Linguistics

2018), such as 1-to-n, n-to-1 and n-to-n relations. This shortcoming will be amplified in the medical knowledge graph, in which many entities have a large number of related neighbors.

Inspired by previous work (Veličković et al., 2018; Nathani et al., 2019), we propose an approach to learn knowledge from subgraphs, and inject graph contextualized knowledge into the pre-trained language model. We call this model BERT-MK (a **BERT**-based language model integrated with **M**edical **K**nowledge), our contributions are as follows:

- We propose a novel knowledge-enhanced pre-trained language model BERT-MK for medical NLP tasks, which integrates graph contextualized knowledge learned from the medical KG.
- Experimental results show that BERT-MK achieves better performance than previous state-of-the-art biomedical pre-trained language models on entity typing and relation classification tasks.

2 Methodology

Our model consists of two modules: the knowledge learning module and the language model pre-training module. The first module is utilized to learn graph contextualized knowledge existing in KGs, and the second one integrates the learned knowledge into the language model for knowledge generalization. The details will be described in the following subsections.

2.1 Learning Graph Contextualized Knowledge

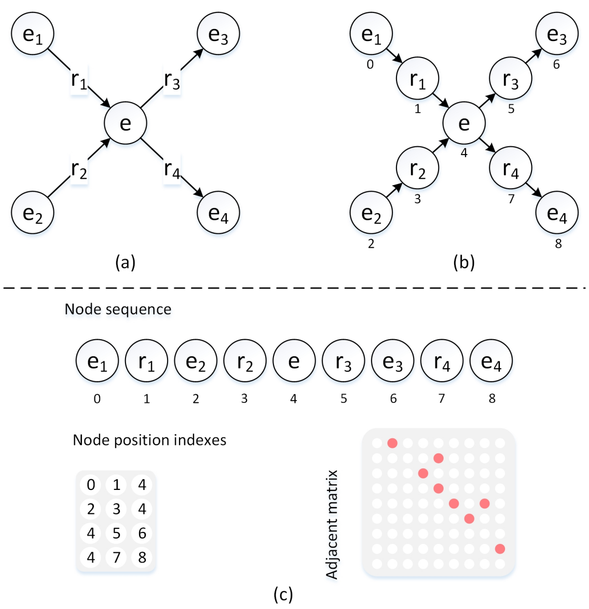
We denote a knowledge graph as $\mathcal{G} = (\mathcal{E}, \mathcal{R})$, where $\mathcal{E}$ represents the entity set and $\mathcal{R}$ is the set of relations between enity pairs. A triple in $\mathcal{G}$ is formalized as (e_s, r, e_o), where e_s is a subjective entity, e_o is an objective entity, and r is the relation between e_s and e_o. In Figure 1, two entities (rectangles) and a relation (arrow) between them constructs a knowledge triple, for example, (Bacterial pneumonia, *causative agent of*, Bacteria).

2.1.1 Subgraph Conversion

To enrich the contextualized information in knowledge representations, we extract subgraphs from the knowlege graph to be the modeling objectives, and the generation process is described in Algorithm 1. For a given entity, its two 1-hop in-entities and out-entities are sampled to generate a subgraph[1], and we repeat the generation process M times for each entity. Figure 2(a) shows an instance of the knowledge subgraph, which consists of four 1-hop and four 2-hop relations. In this study, we propose a Transformer-based (Vaswani et al., 2017) module to model subgraphs. Relations are learned

Algorithm 1: Subgraph generation.

```
Input: Knowledge graph $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T})$, duplicate number M
Output: Subgraph set $\mathcal{S}$
Initial $\mathcal{S} = []$;
foreach $e \in \mathcal{E}$ do
    $d_e^{\text{in}}$ = calculate_in_degree($\mathcal{G}, e$);
    $d_e^{\text{out}}$ = calculate_out_degree($\mathcal{G}, e$);
    $T_e^{\text{in}}$ = extract_in_triples($\mathcal{G}, e$);
    $T_e^{\text{out}}$ = extract_out_triples($\mathcal{G}, e$);
    $i = 0$;
    while $i < (d_e^{\text{in}} + d_e^{\text{out}}) * M/2$ do
        $T_i^{\text{in}}$ = random_sample($T_e^{\text{in}}, 2$);
        $T_i^{\text{out}}$ = random_sample($T_e^{\text{out}}, 2$);
        $subgraph = T_i^{\text{in}} + T_i^{\text{out}}$;
        $\mathcal{S} = \mathcal{S} + subgraph$;
        $i = i + 1$;
    end
end
return $\mathcal{S}$
```



0	1	4
2	3	4
4	5	6
4	7	8

Figure 2: Converting a subgraph extracted from the knowledge graph into the input of the model. (a) e refers to the entity, and r represents the relation. (b) Relations are transformed into sequence nodes, and all nodes are assigned a numeric index. (c) Each row in the matrix of node position indexes represents the index list of an triple in (b); the adjacent matrix indicates the connectivity (the red points equal to 1 and the white points are 0) between the nodes in (b).

[1] In this study, longer n-hop relations are not involved in the subgraph generation process, we leave more arbitrary subgraph to the future work.

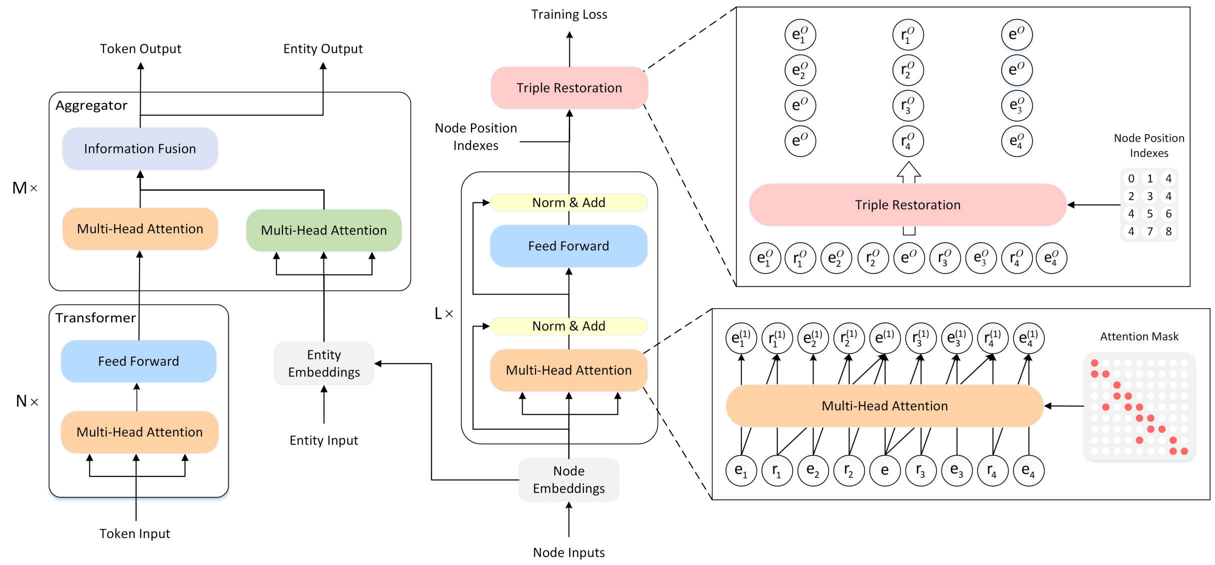


Figure 3: The model architecture of BERT-MK. The left part is the pre-trained language model, in which entity information learned from the knowledge graph is incorporated. The right part is GCKE module. The subgraph in Figure 2 is utilized to describe the learning process. e_1, $e_1^{(1)}$, e_1^O is the embedding of the input node, the updated node and the output node, respectively.

as nodes equivalent to entities in our model, and the relation conversion process is illustrated in Figure 2(b). Therefore, knowledge graph $\mathcal{G}$ can be redefined as $G = (V, E)$, where V represents the nodes in G, involving entities in $\mathcal{E}$ and relations in $\mathcal{R}$, and E denotes the directed edges among the nodes in V.

Then, subgraphs are converted into sequences of nodes. The conversion result of a subgraph is shown in Figure 2(c), including a node sequence, a node position index matrix and an adjacency matrix. Each row of the node position index matrix corresponds to a triple in the subgraph. For example, the triple (e_1, r_1, e) is represented as the first row $(0, 1, 4)$ in this matrix. In the adjacency matrix, the element $\mathbf{A}_{ij}$ equals 1 if the node i is connected to node j in Figure 2(b), and 0 otherwise.

2.1.2 GCKE

After the subgraph conversion preprocessing, the input samples to learn graph contextualized knowledge are generated. Formally, we denote the node sequence as $\{x_1, \ldots, x_N\}$, where N is the length of the input sequence. Besides, the node position index matrix and the adjacency matrix are defined as $\mathbf{P}$ and $\mathbf{A}$, respectively. Entity embeddings and relation embeddings are integrated in the same matrix $\mathbf{V}$, where $\mathbf{V} \in \mathbb{R}^{(n_e+n_r)\times d}$, n_e is the entity number in $\mathcal{E}$ and n_r is the relation type number in $\mathcal{R}$. The node embeddings $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ can be generated by looking up node sequence $\{x_1, \ldots, x_N\}$ in embedding matrix $\mathbf{V}$. $\mathbf{X}$, $\mathbf{P}$ and $\mathbf{A}$ constitute the input of the graph contextualized knowledge embedding learning module, called **GCKE**, as shown in Figure 3.

The inputs are fed into a Transformer-based model to encode the node information.

$$\mathbf{x}'_i = \bigoplus_{h=1}^{H} \sum_{j=1}^{N} \alpha_{ij}^h \cdot (\mathbf{x}_j \cdot \mathbf{W}_{\mathrm{v}}^h), \tag{1}$$

$$\alpha_{ij}^h = \frac{\exp(a_{ij}^h)}{\sqrt{d/H} \cdot \sum_{n=1}^{N} \exp(a_{in}^h)}, \tag{2}$$

$$a_{ij}^h = \mathtt{Masking}((\mathbf{x}_i \cdot \mathbf{W}_{\mathrm{q}}^h) \cdot (\mathbf{x}_j \cdot \mathbf{W}_{\mathrm{k}}^h)^{\mathrm{T}}), \mathbf{A}_{ji} + \mathbf{I}_{ij}), \tag{3}$$

where $\mathbf{x}'_i$ is the new embedding for node x_i. $\bigoplus$ denotes the concatenation of the H attention heads in this layer, α_{ij}^h and $\mathbf{W}_{\mathrm{v}}^h$ are the attention weight of node x_j and a linear transformation of node embedding $\mathbf{x}_j$ in the h^{th} attention head, respectively. The `Masking` function in Equation 3 restraints the contextualized dependency among the input nodes, only the *degree-in* nodes and the current node itself are involved to update the node embedding. The subfigure in the lower right corner of Figure 3 shows the contextualized dependencies. Similar to $\mathbf{W}_{\mathrm{v}}^h$, $\mathbf{W}_{\mathrm{q}}^h$ and $\mathbf{W}_{\mathrm{k}}^h$ are independent linear transformations of node embeddings. Then, the updated node representations are fed into the feed forward layer for further encoding. The aforementioned Transformer blocks are stacked by L times, and the output hidden states can be formalized as

$$\mathbf{X}^O = \{\mathbf{x}_1^O, \ldots, \mathbf{x}_N^O\}. \tag{4}$$

Then, the node position indexes $\mathbf{P}$ is utilized to restore triple representations:

$$\mathbf{T} = \texttt{TripleRestoration}(\mathbf{X}^O, \mathbf{P}), \tag{5}$$

where $\mathbf{P}_k = (e_s^k, r^k, e_o^k)$ is the position index of a valid knowledge triple, and $\mathbf{T}_k = (\mathbf{x}_{e_s^k}^O, \mathbf{x}_{r^k}^O, \mathbf{x}_{e_o^k}^O)$ is the representation of this triple. The subfigure in the upper right corner of Figure 3 shows the triple restoration process.

In this study, the translation-based scoring function (Han et al., 2018) is adopted to measure the energy of a knowledge triple. The node embeddings are learned by minimizing a margin-based loss function on the training data:

$$\mathcal{L} = \sum_{\mathbf{t}\in\mathbf{T}} \texttt{max}\{d(\mathbf{t}) - d(f(\mathbf{t})) + \gamma, 0\}, \tag{6}$$

where $\mathbf{t} = (\mathbf{t}_s, \mathbf{t}_r, \mathbf{t}_o)$, $d(\mathbf{t}) = |\mathbf{t}_s + \mathbf{t}_r - \mathbf{t}_o|$, $\gamma > 0$ is a margin hyperparameter, $f(\mathbf{t})$ is an entity replacement operation that the head entity or the tail entity in a triple is replaced and the replaced triple is an invalid triple in the KG.

2.2 Integrating Knowledge into the Language Model

Given a comprehensive medical knowledge graph, graph contextualized knowledge representations can be learned using the GCKE module. We follow the language model architecture proposed in (Zhang et al., 2019a), and utilize graph contextualized knowledge to enhance medical language representations. The pre-training process is shown in the left part of Figure 3. The Transformer block encodes word contextualized representation while the aggregator block implements the fusion of knowledge and language information.

According to the characteristics of medical NLP tasks, domain-specific finetuning procedure is designed. Similar to BioBERT (Lee et al., 2019), symbol “@” and “$” are used to mark the entity boundary, which indicate the entity positions in a sample and distinguish different relation samples sharing the same sentence. For example, the input sequence for the relation classification task can be modified into “[CLS] *pain control was initiated with morphine but was then changed to @ demerol* $, *which gave the patient better relief of @ his epigastric pain* $”. In the entity typing task, entity mention and its context are critical to predict the entity type, so more localized features of the entity mention will benefit this prediction process. In our experiments, the entity start symbol is selected to represent an entity typing sample.

Table 1: Statistics of UMLS.

# Entities	**# Relations**	**# Triples**
2,842,735	874	13,555,037
In-degree	**Out-degree**	**Median degree**
5.05	5.05	4

3 Experiments

3.1 Dataset

3.1.1 Medical Knowledge Graph

The Unified Medical Language System (UMLS) (Bodenreider, 2004) is a comprehensive knowledge base in the biomedical domain, which contains large-scale concept names and relations among them. The metathesaurus in UMLS involves various terminology systems and comprises about 14 million terms covering 25 different languages. In this study, a subset of this knowledge base is extracted to construct the medical knowledge graph. Non-English and long terms are filtered, and the final statistics is shown in Table 1.

3.1.2 Corpus for Pre-training

To ensure that sufficient medical knowledge can be integrated into the language model, PubMed abstracts[2] and PubMed Central full-text papers[3] are chosen as the pre-training corpus, which are open-access datasets for biomedical and life sciences journal literature. Since sentences in different paragraphs may not have good context coherence, paragraphs are selected as the document unit for next sentence prediction. The Natural Language Toolkit (NLTK)[4] is utilized to split the sentences within a paragraph, and sentences having less than 5 words are discarded. As a result, a large corpus containing 9.9B tokens is achieved for language model pre-training.

[2] https://www.ncbi.nlm.nih.gov/pubmed/.

[3] https://www.ncbi.nlm.nih.gov/pmc/.

[4] https://www.nltk.org/.

Table 2: Statistics of the datasets. Most of these datasets do not follow a standard train-valid-test set partition, and we adopt some traditional data partition ways to do model training and evaluation.

Task	Dataset	# Train	# Valid	# Test
Entity Typing	2010 i2b2/VA (Uzuner et al., 2011)	16,519	-	31,161
	JNLPBA (Kim et al., 2004)	51,301	-	8,653
	BC5CDR (Li et al., 2016)	9,385	9,593	9,809
Relation Classification	2010 i2b2/VA (Uzuner et al., 2011)	10,233	-	19,115
	GAD (Bravo et al., 2015)	5,339	-	-
	EU-ADR (Van Mulligen et al., 2012)	355	-	-

In our model, medical terms appearing in the corpus need to be aligned to the entities in the UMLS metathesaurus before pre-training. To make sure the coverage of identified entities in the metathesaurus, the forward maximum matching (FMM) algorithm is used to extract the term spans from the corpus aforementioned, and spans less than 5 characters are filtered. Then, BERT vocabulary is used to tokenize the input text into word pieces, and the medical entity is aligned with the first subword of the identified term.

3.1.3 Downstream Tasks

In this study, entity typing and relation classification tasks in the medical domain are used to evaluate the models.

Entity Typing Given a sentence with an entity mention tagged, this task is to identify the semantic type of this entity mention. For example, the type "*medical problem*" is used to label the entity mention "*asystole*" in the sentence "*he had a differential diagnosis of* $\langle \texttt{e} \rangle$ *asystole* $\langle / \texttt{e} \rangle$". To the best of our knowledge, there are no publicly available entity typing datasets in the medical domain. Therefore, three entity typing datasets are constructed from the corresponding medical named entity recognition datasets. Entity mentions and entity types are annotated in these datasets, in this study, entity mentions are considered as input while entity types are the output labels. Table 2 shows the statistics of the datasets for the entity typing task. Datasets can be download from here[5].

Relation Classification Given two entities within one sentence, this task aims to determine the relation type between the entities. For example, in sentence "*pain control was initiated with morphine but was then changed to* $\langle \texttt{e}_1 \rangle$ *demerol* $\langle / \texttt{e}_1 \rangle$, *which gave the patient better relief of* $\langle \texttt{e}_2 \rangle$ *his epigastric pain* $\langle / \texttt{e}_2 \rangle$", the relation type between two entities is *TrIP* (Treatment Improves medical Problem). In this study, three relation classification datasets are utilized to evaluate our models, and the statistics of these datasets are shown in Table 2. Datasets can be download from here[6].

3.2 Baselines

In addition to the state-of-the-art models on these datasets, we have also added the popular BERT-Base model and another two models pre-trained on biomedical literature for further comparison.

BERT-Base (Devlin et al., 2019) This is the original bidirectional pre-trained language model proposed by Google, which achieves state-of-the-art performance on a wide range of NLP tasks.

BioBERT (Lee et al., 2019) This model follows the same model architecture as the BERT-Base model, but with the PubMed abstracts and PubMed Central full-text articles (about 18B tokens) used to do model finetuning upon BERT-Base.

SCIBERT (Beltagy et al., 2019) In this model, a new wordpiece vocabulary is built based on a large scientific corpus (about 3.2B tokens). Then, a new BERT-based model is trained from scratch using this scientific vocabulary and the scientific corpus. Since a large portion of the scientific corpus consists of biomedical articles, this scientific vocabulary can also be regarded as a biomedical vocabulary, and helps improve the performance of downstream tasks in the biomedical domain.

3.3 Implementation Details

3.3.1 Graph Contextualized Knowledge

Firstly, UMLS triples are fed into the TransE model to achieve a basic knowledge representation. We

[5] https://drive.google.com/file/d/1OletxmPYNkz2ltOr9pyT0b0iBtUWxslh/view.

[6] https://drive.google.com/file/d/1-jDKGcXREb2X9xTFnuiJ36PvsqoyHWcw/view.

Table 3: Experimental results on the entity typing and relation classification tasks. Accuracy (Acc), Precision, Recall, and F1 scores are used to evaluate the model performance. The results reported in previous work are underlined. E-SVM is short for Ensemble SVM (Bhasuran and Natarajan, 2018), which achieves SOTA performance in the GAD dataset. CNN-M stands for CNN using multi-pooling (He et al., 2019), which is the SOTA model in the 2010 i2b2/VA dataset.

Task	Dataset	Metrics	E-SVM	CNN-M	BERT-Base	BioBERT	SCIBERT	BERT-MK
Entity Typing	2010 i2b2/VA	Acc	-	-	96.76	97.43	**97.74**	97.70
	JNLPBA	Acc	-	-	94.12	94.37	**94.60**	94.55
	BC5CDR	Acc	-	-	98.78	99.27	99.38	**99.54**
Relation Classification	2010 i2b2/VA	P	-	73.1	72.6	76.1	74.8	**77.6**
		R	-	66.7	65.7	71.3	71.6	**72.0**
		F	-	69.7	69.2	73.6	73.1	**74.7**
	GAD	P	79.21	-	74.28	76.43	77.47	**81.67**
		R	89.25	-	85.11	87.65	85.94	**92.79**
		F	83.93	-	79.33	81.66	81.45	**86.87**
	EU-ADR	P	-	-	75.45	81.05	78.42	**84.43**
		R	-	-	**96.55**	93.90	90.09	91.17
		F	-	-	84.71	87.00	85.51	**87.49**

use OpenKE toolkit (Han et al., 2018) to learn entity and relation embeddings. Knowledge embedding dimension is set to 100, while training epoch number is set to 10000.

Following the initialization method used in (Nguyen et al., 2018; Nathani et al., 2019), the embeddings produced by TransE are utilized to initialize knowledge representations of the GCKE module. We set the layer number to 4, and each layer contains 4 heads. Due to the median degree of entities in UMLS is 4 (shown in Table1), we set the count of in-entities and two out-entities to 4, so each subgraph contains four 1-hop and four 2-hop relations. The GCKE module runs 1200 epochs on a single NVIDIA Tesla V100 (32GB) GPU to learn graph contextualized knowledge. The batch size is set to 50000.

3.3.2 Pre-training

In this study, two pre-trained language models are trained. The first one is MedERNIE, a medical ERNIE model trained on the UMLS triples and the PubMed corpus, inheriting the same model hyper-parameters used in (Zhang et al., 2019a). Besides, the entity embeddings learned by GCKE module are integrated into the language model to train the BERT-MK model. In our work, we align the same pre-training epochs with BioBERT, which uses the same pre-training corpus as ours, and finetune the BERT-Base model on the PubMed corpus for one epoch.

3.3.3 Finetune

As shown in Table 2, there is no standard valid or test set in some datasets. For datasets containing a standard test set, if no standard valid set is provided, we divide the training set into new train/valid sets by 4:1. We preform each experiment 5 times under specific experimental settings with different random seeds. Besides, 10-fold cross-validation method is used to evaluate the model performance for the datasets without a standard test set. According to the maximum sequence length of the sentences in each dataset, the input sequence length for 2010 i2b2/VA (Uzuner et al., 2011), JNLPBA (Kim et al., 2004), BC5CDR (Li et al., 2016), GAD (Bravo et al., 2015) and EU-ADR (Van Mulligen et al., 2012) are set to 390, 280, 280, 130 and 220, respectively. The initial learning rate is set to 2e-5.

3.4 Results

3.4.1 Entity Typing

Table 3 presents the experimental results on the entity typing and relation classification tasks. For entity typing tasks, all these pre-trained language models achieve high accuracy, indicating that the type of a medical entity is not as ambiguous as that in the general domain. BERT-MK outperforms BERT-Base and BioBERT on three datasets, and is competitive with SCIBERT. Without using external knowledge in the pre-trained language model, SCIBERT achieves comparable results to BERT-MK, which proves that a domain-specific vocabulary is critical to the feature encoding of inputs. Long tokens are relatively common in the medical domain, and these tokens will be split into short pieces when a domain-independent vocabulary is used, which will cause an overgeneralization of lexical features. Therefore, a medical vocabulary generated by the PubMed corpus can be introduced into BERT-MK in the following work.

3.4.2 Relation Classification

On the relation classification tasks, BERT-Base does not perform as well as other models, which indicates that pre-trained language models require a domain adaptation process when used in restricted domains. Compared with BioBERT, which utilizes the same domain-specific corpus as ours for domain adaptation, BERT-MK improves the F score of 2010 i2b2/VA, GAD and EU-ADR by 1.1%, 5.21% and 0.49%, respectively, which demonstrates medical knowledge has indeed played a positive role in the identification of medical relations.

The following example provides a brief explanation of why medical knowledge improves the model performance of the relation classification tasks. "*On postoperative day number three , patient went into* $\langle \mathtt{e}_1 \rangle$ *atrial fibrillation* $\langle /\mathtt{e}_1 \rangle$ *, which was treated appropriately with* $\langle \mathtt{e}_2 \rangle$ *metoprolol* $\langle /\mathtt{e}_2 \rangle$ *and digoxin and converted back to sinus rhythm*" is a relation sample from the 2010 i2b2/VA dataset, and the relation label is *TrIP*. Meanwhile, the above entity pair can be aligned to a knowledge triple (*atrial fibrillation*, *may be treated by*, *metoprolol*) in the medical knowledge graph. Obviously, this knowledge information is advantageous to identify the relation type of the aforementioned example.

3.5 Discussion

3.5.1 TransE vs. GCKE

In order to explicitly analyze the improvement effect of the GCKE module on pre-trained language models, we compare MedERNIE (TransE-based) and BERT-MK (GCKE-based) on two relation classification datasets. Table 4 demonstrates the results of these two models. As we can see, integrating graph contextualized knowledge into the pre-trained language model, the performance increases F score by 0.9% and 0.64% on these two relation classification datasets, respectively.

In Figure 4, as the amount of pre-training data increases, BERT-MK always outperforms MedERNIE on the 2010 i2b2/VA relation dataset, and the performance gap has an increasing trend. However, on the GAD dataset, the performance of BERT-MK and MedERNIE are intertwined. We link the entities in each relation sample to the medical KG, and find that some entity pairs have a connected relationship in the KG. Statistical analysis on 2-hop neighbor relationships between these entity pairs shows that there are 136 cases in the 2010 i2b2/VA dataset, while only 1 in GAD. The second case shown in Table 5 gives an example of the observation described above. Triple (*CAD, member of, Other ischemic heart disease*) and (*Other ischemic heart disease, has member, Angina symptom*) are triples in the medical KG, which indicates entity pair *cad* and *angina symptoms* in the relation sample have a 2-hop neighbor relationship in the KG. GCKE learns these 2-hop neighbor relationships in 2010 i2b2/VA and produces an improvement for BERT-MK. However, due to the characteristics of

Table 4: TransE vs. GCKE on the 2010 i2b2/VA relation and GAD datasets.

Dataset	MedERNIE			BERT-MK		
	P	R	F	P	R	F
2010 i2b2/VA	76.6	71.1	73.8	77.6	72.0	**74.7**
GAD	81.28	91.86	86.23	81.67	92.79	**86.87**

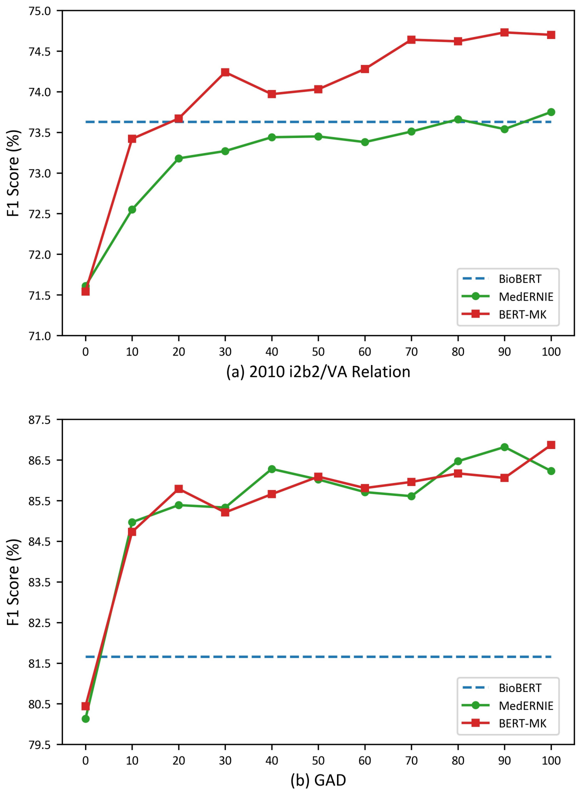


Figure 4: Model performance comparison with increasing amount of the pre-trained data. The x-axis represents the proportion of the medical data used for pre-training. 0 means no medical data is utilized, so the BERT-Base is used as an initialization parameter for the model finetuning. 100 indicates the model is pre-trained on the medical corpus for one epoch. BioBERT pre-trains on the PubMed corpus for one epoch, which is drawn with dashed lines in the figure as a comparable baseline.

Table 5: Case study on the 2010 i2b2/VA relation dataset. The bold text spans in two cases are entities. In the first case, the corresponding triple can help identify the relationship between the entity pair in this relation sample. NPP, no relation between two medical problems; PIP, medical problem indicates medical problem. MI, myocardial infarction; CAD, coronary artery disease.

	Cases	**The Corresponding Triples**	**BioBERT**	**MedERNIE**	**BERT-MK**	**Ground Truth**
1	... **coronary artery disease**, status post **mi** x0, cabg ...	(Coronary artery disease, associated with , MI)	NPP	PIP	PIP	PIP
2	0. **cad**: presented with **anginal symptoms** and ekg changes (stemi), with cardiac catheterization revealing lesions in lad, lcx, and plb.	(CAD, member of, Other ischemic heart disease); (Other ischemic heart disease, has member, Angina symptom)	NPP	NPP	PIP	PIP

the GAD dataset, the capability of GCKE is limited.

3.5.2 Effect of Different Corpus Sizes in Pre-training

Figure 4 shows the model performance comparison with different proportion of the pre-training corpus. From this figure, we observe that BERT-MK outperforms BioBERT by using only 10%-20% of the corpus, which indicates that medical knowledge has the capability to enhance pre-trained language models and save computational costs (Schwartz et al., 2019).

4 Related Work

Pre-trained language models represented by ELMO (Peters et al., 2018), GPT (Radford et al., 2018) and BERT (Devlin et al., 2019) have attracted great attention, and a large number of variant models have been proposed. Among these studies, some researchers devote their efforts to introducing knowledge into language models (Levine et al., 2019; Lauscher et al., 2019; Liu et al., 2019; Zhang et al., 2019b). ERNIE-Baidu (Sun et al., 2019) introduces new masking units such as phrases and entities to learn knowledge information in these masking units. As a reward, syntactic and semantic information from phrases and entities is implicitly integrated into the language model. Furthermore, a different knowledge information is explored in ERNIE-Tsinghua (Zhang et al., 2019a), which incorporates knowledge graph into BERT to learn lexical, syntactic and knowledge information simultaneously. Xiong et al. (2019) introduce entity replacement checking task into the pre-trained language model, and improve several entity-related downstream tasks, such as question answering and entity typing. Wang et al. (2020) propose a plug-in way to infuse knowledge into language models, and their method keeps different kinds of knowledge in different adapters. The knowledge information introduced by these methods does not pay much attention to the graph contextualized knowledge in the KG.

Recently, several KRL methods have attempted to introduce more contextualized information into knowledge representations. Relational Graph Convolutional Networks (R-GCNs) (Schlichtkrull et al., 2018) is proposed to learn entity embeddings from their incoming neighbors, which greatly enhances the information interaction between related triples. Nathani et al. (2019) further extend the information flow from 1-hop in-entities to n-hop during the learning process of entity representations, and achieves the SOTA performance on multiple relation prediction datasets, especially for the ones containing higher in-degree nodes. We believe that the information contained in knowledge graphs is far from being sufficiently exploited. In this study, we develop an approach to integrate more graph contextualized information, which models subgraphs as training samples. This module has the ability to model any information in the KG. In addition, this learned knowledge is integrated into the language model to obtain an enhanced version of the medical pre-trained language model.

5 Conclusion and Future Work

We propose a novel approach to learn more comprehensive knowledge, focusing on modeling subgraphs in the knowledge graph by a knowledge learning module. Additionally, the learned medical knowledge is integrated into the pre-trained language model, which outperforms BERT-Base and another two domain-specific pre-trained language models on several medical NLP tasks. Our work validates the intuition that medical knowledge is beneficial to some medical NLP tasks and provides a preliminary exploration for the application of medical knowledge.

In the follow-up work, some knowledge-guided tasks will be used to validate the effectiveness of the knowledge learning module GCKE. Moreover, we will explore some other knowledge injection ways to combine medical knowledge with language models, such as multi-task learning. More subgraph sampling strategies need to be explored, such as r-ego subgraph (Qiu et al., 2020) and degree-dependent subgraph.

Acknowledgment

The authors would like to thank all the anonymous reviewers for their insightful comments. Thank Yasheng Wang for his help in code implementation.

References

Iz Beltagy, Arman Cohan, and Kyle Lo. 2019. Scibert: Pretrained contextualized embeddings for scientific text. *arXiv preprint arXiv:1903.10676*.

Balu Bhasuran and Jeyakumar Natarajan. 2018. Automatic extraction of gene-disease associations from literature using joint ensemble learning. *PloS one*, 13(7):e0200699.

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

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*, pages 2787–2795.

Àlex Bravo, Janet Piñero, Núria Queralt-Rosinach, Michael Rautschka, and Laura I Furlong. 2015. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC bioinformatics*, 16(1):55.

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.

John R Firth. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*.

Xu Han, Shulin Cao, Xin Lv, Yankai Lin, Zhiyuan Liu, Maosong Sun, and Juanzi Li. 2018. Openke: An open toolkit for knowledge embedding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 139–144.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Bin He, Yi Guan, and Rui Dai. 2019. Classifying medical relations in clinical text via convolutional neural networks. *Artificial intelligence in medicine*, 93:43–49.

Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Nigel Collier. 2004. Introduction to the bio-entity recognition task at jnlpba. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications*, pages 70–75. Citeseer.

Anne Lauscher, Ivan Vulić, Edoardo Maria Ponti, Anna Korhonen, and Goran Glavaš. 2019. Informing unsupervised pretraining with external linguistic knowledge. *arXiv preprint arXiv:1909.02339*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*.

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*.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Yankai Lin, Xu Han, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. 2018. Knowledge representation learning: A quantitative review. *arXiv preprint arXiv:1812.10901*.

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

Deepak Nathani, Jatin Chauhan, Charu Sharma, and Manohar Kaul. 2019. Learning attention-based embeddings for relation prediction in knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4710–4723, Florence, Italy. Association for Computational Linguistics.

Dai Quoc Nguyen, Tu Dinh Nguyen, Dat Quoc Nguyen, and Dinh Phung. 2018. A novel embedding model for knowledge base completion based on convolutional neural network. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 327–333.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.

Jiezhong Qiu, Qibin Chen, Yuxiao Dong, Jing Zhang, Hongxia Yang, Ming Ding, Kuansan Wang, and Jie Tang. 2020. Gcc: Graph contrastive coding for graph neural network pre-training. *arXiv preprint arXiv:2006.09963*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, pages 593–607. Springer.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2019. Green ai.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Erik M Van Mulligen, Annie Fourrier-Reglat, David Gurwitz, Mariam Molokhia, Ainhoa Nieto, Gianluca Trifiro, Jan A Kors, and Laura I Furlong. 2012. The eu-adr corpus: annotated drugs, diseases, targets, and their relationships. *Journal of biomedical informatics*, 45(5):879–884.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. *International Conference on Learning Representations*.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Cuihong Cao, Daxin Jiang, Ming Zhou, et al. 2020. K-adapter: Infusing knowledge into pre-trained models with adapters. *arXiv preprint arXiv:2002.01808*.

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

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019a. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2019b. Semantics-aware bert for language understanding. *arXiv preprint arXiv:1909.02209*.

A Appendices

A.1 Comparison between MedERNIE and BERT-MK

As shown in Table 6, BERT-MK outperforms MedERNIE on all datasets except BC5CDR.

Table 6: MedERNIE vs. BERT-MK.

	Entity Typing (Acc)		
	2010 i2b2/VA	JNLPBA	BC5CDR
MedERNIE	97.37	94.46	**99.62**
BERT-MK	**97.70**	**94.55**	99.54
	Relation Classification (F)		
	2010 i2b2/VA	GAD	EU-ADR
MedERNIE	73.8	86.23	86.99
BERT-MK	**74.7**	**86.87**	**87.49**