Medical knowledge-enhanced prompt learning for diagnosis classification from clinical text

Yuxing Lu
Department of BigData and Biomedical AI
College of Future Technology
Peking University
Beijing, China
yxlu0613@gmail.com

Xukai Zhao
Department of Landscape Architecture
School of Architecture
South China University of Technology
Guangzhou, China
zhaoxukai0208@163.com

Jingzhuo Wang
Department of BigData and Biomedical AI
College of Future Technology
Peking University
Beijing, China
wangjinzhuo@pku.edu.cn

Abstract

Artificial intelligence based diagnosis systems have emerged as powerful tools to reform traditional medical care. Each clinician now wants to have his own intelligent diagnostic partner to expand the range of services he can provide. When reading a clinical note, experts make inferences with relevant knowledge. However, medical knowledge appears to be heterogeneous, including structured and unstructured knowledge. Existing approaches are incapable of uniforming them well. Besides, the descriptions of clinical findings in clinical notes, which are reasoned to diagnosis, vary a lot for different diseases or patients. To address these problems, we propose a Medical Knowledge-enhanced Prompt Learning (MedKPL) model for diagnosis classification. First, to overcome the heterogeneity of knowledge, given the knowledge relevant to diagnosis, MedKPL extracts and normalizes the relevant knowledge into a prompt sequence. Then, MedKPL integrates the knowledge prompt with the clinical note into a designed prompt for representation. Therefore, MedKPL can integrate medical knowledge into the models to enhance diagnosis and effectively transfer learned diagnosis capacity to unseen diseases using alternating relevant disease knowledge. The experimental results on two medical datasets show that our method can obtain better medical text classification results and can perform better in transfer and few-shot settings among datasets of different diseases.

1 Introduction

Clinical notes in Electronic Health Records (EHRs) are the medical texts written by a physician to address the patient's medical history, chief complaints and examinations during a patient's visit. Physicians can get the corresponding diagnosis through their expertise based on the patient's clinical notes. In the past decade, researchers have tried various methods for medical text classification tasks to assist doctors in their treatment.

Text classification models in the generic domain are developing most rapidly. Traditional machine learning methods, such as Naive Bayesian (NB) (Maron, 1961), K-Nearest Neighbor (KNN) (Cover and Hart, 1967), Support Vector Machine (SVM) (Joachims, 1998), and Random Forest (RF) (Breiman, 2001) are first introduced to solve text classification tasks. For deep learning models, TextCNN (Chen, 2015) is widely used, where Convolutional Neural Network (CNN) (Albawi et al., 2017) models are introduced to solving text classification problems. Whereafter, Pre-trained Language Models (PLMs), such as BERT (Devlin et al., 2018) achieve state-of-the-art results on several Natural Language Processing (NLP) tasks and thus has been widely used. However, these approaches are based on generic data and therefore ignore the high reliance on medical knowledge in medical text classification tasks. When applied directly to the medical field, these models often fail to achieve the same performance as in the generic field.

To address the knowledge-dependent medical text classification tasks, researchers have proposed a number of medical text classification models that incorporate knowledge. Garla and Brandt (2013) map medical text to corresponding medical concepts and is the first to conduct feature engineering. Yao et al. (2019a) use medical concept descriptions

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 278–288
July 14, 2023 ©2023 Association for Computational Linguistics

Raw Clinical Notes	Runny nose and coughing with phlegm started 4 days ago.
+ Prompt Template	Patient: Runny nose and coughing with phlegm started 4 days ago. Which disease does the patient get? [MASK]
+ Knowledge Enhanced Prompt Template	Cough with croup and recurrent infections is a symptom of bronchitis, dyspnoea is a symptom of bronchitis. Patient: Runny nose and coughing with phlegm started 4 days ago. Which disease does the patient get? [MASK]

Figure 1: Different template generation methods for clinical notes. Prompt learning method simply adds questions to the clinical notes, our **Med**ical **K**nowledge-enhanced **P**rompt **L**earning method incorporates heterogeneous medical knowledge in the template.

to improve distributed document representations. Gasmi (2022) use external terminology resources to expand and represent the text with a combination of different methods. Nevertheless, these models only learn the relationship between the text and the corresponding knowledge, without having a good generalization ability. Therefore they tend to be less effective when transferring to the medical domains beyond the training data.

In the medical field, there are rich sources of knowledge, such as expert knowledge (Flores et al., 2011), medical knowledge bases (Zuccon et al., 2013), medical knowledge graphs (Li et al., 2019), medical information on the web, etc. These knowledge present a heterogeneous structure(such as triples, SQLs and free texts, etc.) and cannot be well uniformed in the previous methods. Differences among knowledge sources prevent these models from learning by using knowledge prompt from all sources and thus may have bias when dealing with real-world data. Therefore, we hope to propose a model that is compatible with all the sources of medical knowledge.

To solve the above problems in medical text classification, we propose a **Med**ical **K**nowledge-enhanced **P**rompt **L**earing (MedKPL) model that can uniform different knowledge sources. The contribution of this paper can be summarized as follows: 1) We design the MedKPL model to uniform heterogeneous knowledge by transforming knowledge from different sources into free texts. Experiments prove that structured and unstructured texts can be uniform in our model, and both yield good results. 2) We use the MedKPL model to conduct medical text classification tasks on two Chinese EHR datasets and obtain state-of-the-art classification results through knowledge incorporation. 3) We evaluate the MedKPL model for few-shot learning among departments. The results show that our method can obtain good results in both zero-shot and few-shot scenarios, and can effectively transfer between departments that have low text similarity in a robust way.

2 Related Work

2.1 Knowledge Enhancement for PLMs

PLMs has become text representation method in most NLP tasks. Generic PLMs are usually trained on unstructured text corpus without domain knowledge. For example, BERT (Devlin et al., 2018) is trained on BooksCorpus (Zhu et al., 2015) and English Wikipedia, GPT-2 (Radford et al., 2019) and GPT-3 (Brown et al., 2020) use Common Crawl (Raffel et al., 2020) and WebText as training corpus. Due to training on generic datasets, most of contextual information learned by these PLMs lack domain knowledge, resulting in their lack of expertise in dealing with domain-specific problems.

Continuous Knowledge-enhancement uses knowledge encoders to get the embedding of knowledge and incorporate them into the process of training contextual representations of text. Know-BERT (Peters et al., 2019) propose Knowledge Attention and Recontextualization (KAR) and entity linking to incorporate knowledge into PLMs. ERNIE-THU (Zhang et al., 2019) introduce an knowledge fusion module, injecting entity embeddings through knowledge encoders. KEPLER (Wang et al., 2021) jointly optimize the knowledge embedding and language modeling objectives within the same PLM. DKPLM (Zhang et al., 2022) use pseudo token representations to embed long-tail entities which relieve computation burdens of previous methods.

Discrete Knowledge-enhancement retrieves knowledge directly from the knowledge graph and add them to training texts. K-BERT (Liu et al., 2020) and CoLAKE (Sun et al., 2020) directly reorganize the triples in the knowledge graph into texts and insert them directly into the training corpus, without pre-training any extra models. We also apply the ideas behind these methods to our work.

2.2 Prompt Learning

Prompt learning refers to transforming the original text via templates to leverage the contextual pattern learned by the PLMs. Brown et al. (2020) first use

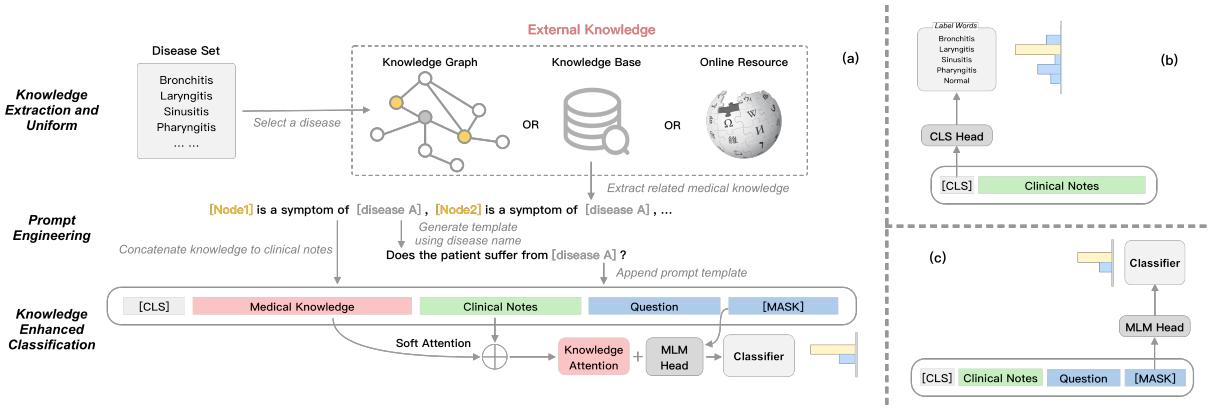

Figure 2: **The illustration of MedKPL and other methods.** (a) is the workflow of MedKPL, knowledge can be obtained from different knowledge sources and then incorporated into clinical notes through template construction, and the classifier (multi-classifier and binary classifier) can be further enhanced by using soft attention on knowledge prompt and clinical notes. (b) is the method for fine tuning at PLM to classify the embedding of the [CLS] token, and (c) is the method for regular prompt learning to predict the probability distribution of the [MASK] token.

the prompt learning method for text classification tasks and find it works well on few-shot learning scenarios. Schick and Schütze (2020) reformulate inputs as cloze questions for text classification. Schick et al. (2020) and Gao et al. (2020) extend previous methods by automatically generating label words and templates, respectively. Recently, some knowledge-related prompt learning methods have been proposed. Hu et al. (2021) incorporate external knowledge into the verbalizer with calibration. Chen et al. (2022) inject latent knowledge into learnable virtual type words and answer words.

Compared with these approaches, our approach can uniform heterogeneous knowledge to build prompt templates, which solves the differences brought by different knowledge formats sources. Our approach also provides a deep integration between clinical notes and knowledge prompts.

2.3 Medical Text Classification

How to apply external knowledge to medical text classification tasks is a topic that has been constantly explored by researchers. Garla and Brandt (2013) map clinical text to Unified Medical Language System (UMLS), and use those UMLS Concept Unique Identifiers (CUIs) as features to train classifiers on medical documents. Yao et al. (2019a) propose to distribute document representations with medical concept descriptions for the classification of traditional Chinese medicine clinical records. Yao et al. (2019b) combine rule-based features and knowledge-guided CNN for effective disease classification. Li and Yu (2020) use multi-filter Residual CNN to predict ICD codes. Chen et al. (2020) propose an attention-based bidirectional LSTM model for classifying outpatient categories according to textual content.

However, none of these works mentioned the model's transferability among departments and few-shot learning ability, which are issues that must be addressed to solve the medical long-tail problem and achieve truly trustworthy medical AI.

3 Method

The overall structure of our model is shown in Figure 2. Our model introduce disease $d \in D$ related medical knowledge prompt k_d into medical text classification tasks, where D is the disease set. The knowledge prompt can come from a variety of sources, e.g. expert knowledge, knowledge graphs, knowledge bases, online resources, etc. We use $p(y|x_i, k)$ to denote the probability of patient i getting disease y, where x_i is the clinical notes for patient i, and k is the set of knowledge prompts used for knowledge incorporation.

Specifically, we decompose the process of knowledge incorporation into three stages. 1) Extract medical knowledge of disease d from different knowledge sources and transform the knowledge into a uniform representation k_d. 2) Construct templates that incorporate knowledge prompts set k with clinical notes. We concatenate the collected medical knowledge prompts into natural text and generate the template based on the disease name d. 3) Predict labels using MLM on the [MASK] token in prompt template. It is also possible to integrate knowledge prompt and clinical notes at a

deep level by using PLM to represent knowledge prompt and clinical notes separately and aligning them using soft attention mechanisms to enhance the knowledge representation.

We will then go over our model's methodology and its three stages of knowledge incorporation.

3.1 Knowledge Extraction and Uniform

Unstructured knowledge is naturally available as part of the prompt template, while structured knowledge needs to be pre-processed. For structured medical knowledge, the most common organization form is the medical knowledge graph. Thus we take knowledge graph as our knowledge source and denote it as $\mathcal{G} = (\mathcal{E}, \mathcal{R})$ where $\mathcal{E}$ is the collection of all entities and $\mathcal{R}$ is the collection of all relations. In the knowledge graph $\mathcal{G}$, a relational knowledge triple is denoted as (e_h, r, e_t), where $e_h \in D$ is the head entity and e_t is the tail entity. r is the specific relation between e_h and e_t.

In a large-scale medical knowledge graph, a disease may have multiple relations, we denote the relation set of disease e_i as R_i. The distribution of triples related to disease e_i is very diverse and complex, and we need to find those triples $(e_i, r, e_j) \in \mathcal{G}, r \in R_i$ that are suitable for our medical knowledge-enhanced prompt learning method.

Specifically, we determine the refined relation set $R'_i = (r_1, r_2, \ldots, r_k), r_i \in \mathcal{R}_i$ based on the relationships commonly mentioned in the clinical notes for disease e_i. Then with the disease e_i and the refined relation set R'_i, we can retrieve all relevant triples $\mathcal{T}_i$ of disease e_i from the knowledge graph.

$$\mathcal{T}_i = \{(e_i, r_i, e_j) | r_i \in \mathcal{R}'_i, e_i, e_j \in \mathcal{E}\} \in \mathcal{G} \quad (1)$$

For those diseases lacking relevant medical knowledge, we consider using similar entities e_j for replacement, where $(e_i, r_{syn}, e_j) \in \mathcal{G}$ and r_{syn} is the relationship of synonym. For those diseases not in the entity set of the medical knowledge graph, we consider replacing them with other knowledge sources(e.g., online search engines).

Alternatively, we also consider using unstructured medical knowledge, such as knowledge bases and online search engines for replacement. Medical knowledge related to disease d can be represented as k_d. Since this unstructured knowledge is already in the form of text, we apply them directly to the subsequent processes.

3.2 Prompt Engineering

The core idea of the prompt learning method is to construct templates and use the contextual knowledge learned by the PLM during the pre-training process to make predictions on the masked words.

Different from the normal prompt approach, we want our templates to contain medical knowledge extracted from heterogeneous knowledge sources. Therefore, we propose a disease-adaptive template generation method. For a disease d, if the knowledge source is KG, we first extract all the required knowledge triples $\mathcal{T}_d$ from the KG and concatenate all the triples together into free texts. Given an example knowledge triple t =*(dyspnoea, a symptom of, bronchitis)*, the formed free-text knowledge would correspondingly be ***Dyspnoea is a symptom of bronchitis***. By concatenating all the triples, we can get the disease-related knowledge k_d in the text pattern.

The promoting function $f_{prompt}(k_d, x, d)$ contains medical knowledge and manual template engineering. We devise templates for binary classification tasks and multi-classification tasks seperately. These two tasks are different in practical medical application scenarios, where a multi-classification task can quickly determine which disease the patient is most likely to have, and a binary classification task can make predictions about the likelihood of a specific disease more precisely. For binary classification tasks, the prompt learning method will extend the input clinical notes x into

$$x' = [K_d][X] \text{ Does the patient suffer from } [D]? \ [MASK].$$

and for multi-classification tasks, the input clinical notes x will be turned into

$$x' = [K_d][X] \text{ Which disease the patient have? } [MASK].$$

where the slot $[K_d]$, $[X]$, $[D]$ are filled with k_d, x, d respectively. In this way, we convert the sequence classification task into a task of predicting the distribution of masked token [MASK].

By organizing all the heterogeneous knowledge into free texts, we can extend the knowledge sources of MedKPL to almost all types of medical knowledge.

3.3 Knowledge Enhanced Classification

By simply concatenating and adding knowledge to the template, we can use PLM to

learn the contextual association between clinical notes and knowledge prompt. However, this approach treats them in a sentence as a whole. To better explore the deeper connection between clinical notes and knowledge prompt, we integrate these texts in a deeper way.

Vector representation of the knowledge prompt $K = (k_1, k_2, \ldots, k_m)$ and clinical notes $C(c_1, c_2, \ldots, c_n)$ can be obtained by PLM, where m and n are the length of knowledge prompt and clinical notes respectively. We use the Soft Attention mechanism (Luong et al., 2015) to align clinical notes with knowledge prompt.

Specifically, we select the [CLS] token $k_1 \in K$ as the vector representation of the whole knowledge prompt and calculate the alignment vector a which is calculated by comparing the knowledge prompt representation k_1 with each clinical note word's hidden state $c_s \in C$:

$$a_s = align(k_1, c_s) = \frac{exp(score(k_1, c_s))}{\sum_{s'=0}^{n}(score(k_1, c_{s'}))} \tag{2}$$

where we use dot product function to compute scores.

$$score(k_1, c_s) = k_1^T c_s \tag{3}$$

Given alignment vector a as weight, the integrated vector i_t is computed as weighted average over all the words' representations in clinical notes. The integrated vector $i_t = \sum_{s=0}^{n} a_s c_s$ can enhance the most relevant part of the clinical notes with the knowledge prompt. For medical text classification, we sum the integrated vector i_t with the Masked Language Model (MLM) prediction x_{mlm} on [MASK] to get $xintegrate$ and compute the loss based on the classification tasks.

$$x_{mlm} = f_{MLM}(x', [MASK]) \tag{4}$$

$$x_{integrate} = W_x x_{mlm} + W_i i_t \tag{5}$$

where f_{MLM} is the masked language model of the PLM. For binary classification tasks, the loss function L_{binary} is computed directly between $x_{integrate}$ and the index of label words ("yes" or "no") in the PLM's vocabulary.

$$L_{binary} = CELoss(x_{integrate}, label) \tag{6}$$

where the CELoss is cross entropy loss. For multi-classification tasks, the loss function is computed by first map $x_{integrate}$ into the label space using a fully-connected layer and compute the cross entropy loss.

$$L_{multi} = CELoss(W_x x_{integrate} + b_x, label) \tag{7}$$

where W_x and b_x are learnable parameters in the model, and $label$ represents the categories in mult-classification tasks.

4 Experiments

4.1 Datasets

In this paper, we compare our results against many existing methods on two medical datasets. The first dataset is the Pediatric Patient EHR (PPE) used in (Liang et al., 2019), which contains 1,362,559 outpatient visits from 567,498 pediatric patients across 6 departments, each outpatient visit includes adverse event, chief complaint and history of present illness, some also have physical examination and image report. The second dataset is Adult-EMR, which contains 339,672 EHR records for 2556 diseases across 12 departments, each record includes chief complaint and history of present illness. We use the clinical notes of patients' history of present illnesses for training. In PPE, we use the clinical notes of the Hematology-Immunology department as normal control data and select six diseases from each of the other five departments (Respiratory (Resp.), Gastroenterology (Gast.), Psychiarty (Psy.), Neurology (Neuro.), Gynecology (Gyn.)) for experiments. In Adult-EMR, we use the clinical notes of the Respiratory Department as normal control data and select six diseases from Tumor and Cancer Department and Cardiology Department for experiments. The knowledge graph we use in our experiments is the DiseaseKG, which is an open-source Chinese medical knowledge graph from OpenKG.

4.2 Settings

In the multi-label classification task, we select 1000 samples from each of the k diseases $(k = 2, 4, 6)$ from a department and 1000 samples from normal control data for $k + 1$ classification task. The knowledge prompt is the concatenation of the truncated knowledge from

Table 1: Standard multi-classification accuracy on different departments. "+ERNIE" and "+DKPLM" means using knowledge-enhanced PLMs to replace BERT, "+Attn" means using the attention layer to enhance the classification performance. The results for each department are acquired by averaging the results for disease number k = 2, 4, 6.

	PPE						Adult EMR		
	Resp.	Gast.	Psy.	Neuro.	Gyn.	Overall	Tumor.	Cv.	Overall
LSTM	64.89	77.08	85.63	90.29	77.59	79.10	67.97	64.65	66.31
LSTM+Attn	65.49	78.89	85.58	87.58	79.59	79.42	62.87	74.35	68.61
CNN	69.31	81.98	85.76	91.41	81.38	81.97	71.25	73.05	72.15
Fine tuning	68.74	80.25	86.96	89.41	81.75	81.42	71.00	74.46	72.73
Prompt	71.05	82.51	89.06	91.47	82.17	83.25	71.43	76.45	73.94
Prompt+ERNIE	70.24	83.27	88.08	91.76	80.84	82.84	69.57	73.96	71.77
Prompt+DKPLM	73.94	84.77	88.76	91.69	81.82	84.20	72.14	76.62	74.38
MedKPL (Ours)	74.06	83.72	89.13	92.29	82.10	84.26	**73.14**	77.44	75.29
MedKPL+DKPLM (Ours)	**75.01**	**85.05**	**90.11**	**92.40**	**83.96**	**85.31**	72.71	**78.61**	**75.66**

Table 2: Standard binary classification accuracy on different departments. "+ERNIE" and "+DKPLM" means using knowledge-enhanced PLMs to replace BERT. The results for each department are acquired by averaging the results for disease number k = 2, 4, 6.

	PPE						Adult EMR		
	Resp.	Gast.	Psy.	Neuro.	Gyn.	Overall	Tumor.	Cv.	Overall
Prompt	88.44	87.77	97.28	92.31	93.56	91.87	86.67	91.13	88.90
Prompt+ERNIE	85.11	81.78	95.22	88.75	91.92	88.56	84.33	90.54	87.44
Prompt+DKPLM	89.92	89.17	92.14	95.58	97.31	92.82	89.17	95.10	92.14
MedKPL (Ours)	94.89	96.06	**98.69**	96.08	99.19	96.98	**96.33**	96.45	**96.39**
MedKPL+DKPLM (Ours)	**95.75**	**97.36**	**98.69**	**96.75**	**99.31**	**97.57**	94.33	**97.78**	96.06

all the diseases, the basic truncation length is 50 per disease. In the binary classification task, we select 500 samples from each of the k diseases $(k = 2, 4, 6)$ from a department and select $k * 500$ samples from the normal control data for binary classification task. The knowledge prompt used for normal control data is randomly selected from all extracted medical knowledge in binary classification tasks.

Standard Settings. For traditional NLP methods, we select LSTM (Liu et al., 2016), CNN (Chen, 2015) and LSTM (Chen et al., 2020) and LSTM with attention (Chen et al., 2020) for comparison.The word embedding for LSTM and CNN models is the 300-dimension skip-gram word embedding (Mikolov et al., 2013) pre-trained on Sogou News corpus (Li et al., 2018), and the word embedding for models applying PLM is BERT-base-chinese (Devlin et al., 2018) if not otherwise stated. For fine tuning, we take the classification token [CLS] and feed it into a fully connected layer for classification, as shown in Figure 2 (b). For prompt learning (Brown et al., 2020), we calculate the probability distribution of the [MASK] token and further predict the classification result, as shown in Figure 2 (c). In addition, we also try different knowledge-integrated PLMs for comparisons, such as ERNIE (Zhang et al., 2019) and medical version of DKPLM (Zhang et al., 2022). These models above are used as the baseline in our experiments. We use BERT and DKPLM as PLMs to conduct experiments on our method, the use of DKPLM on our method can be regarded as using medical knowledge in both the pre-training phase and the prompt learning phase.

Low-Resource Settings. In our experiments, we design a couple of different low-resource scenarios on binary classification tasks. The first is model transferring among departments. We compare the effect of 16-shot transfer learning among the five departments of the PPE dataset, comparing the results of fine tuning, prompt learning, and our method. In addition, we conduct 0-, 2-, 4-, 8-, and 16-shot

transfer learning experiments to compare the effectiveness of our method with other methods on few-shot learning tasks.

In all of our experiments, we use Adam (Kingma and Ba, 2014) as the optimizer with a learning rate of $1e-7$. The training epoch is 20, the batch size is 32, and the dropout rate is 0.5. Due to the average length of knowledge prompt is 48, we set the truncation lengths of a disease's knowledge prompt as 50, and we set the truncation length of the input clinical notes with prompt template as 128. We use the cross entropy loss as the loss function.

All experiments are conducted on a single NVIDIA Tesla V100. The evaluation metric is accuracy, which is widely used in text classification tasks (Lee and Dernoncourt, 2016).

4.3 Results

4.3.1 Standard Results

We first evaluate the performance on multi-classification tasks under standard text classification task settings. The results are shown in Table 1, where we compared to a range of baselines. The result shows that all MedKPL methods, consistently outperform traditional NLP methods, fine tuning and prompt learning baselines, indicating the effectiveness of our methods. Moreover, as a pre-trained PLM using medical knowledge, prompt learning method using DKPLM outperforms the standard prompt learning method by 0.95 percent in multi-classification performance on PPE dataset, showing the effectiveness of knowledge-enhanced PLMs. However, the knowledge-enhanced PLM ERNIE, which is trained on generic knowledge, is 0.41 percent weaker than the standard prompt learning method. This demonstrates that the incorporation of medical knowledge in the pre-training phase does benefit the medical downstream tasks. In addition, replacing BERT in our model with DKPLM can further yield better results.

We also conduct experiments on binary classification tasks and the results are shown in Table 2, where MedKPL outperforms other methods in a larger gap compared with multi-classification tasks. We conjecture this is because in the binary classification tasks, the knowledge prompt only contains knowledge of the selected disease, so the model can learn the relationship between the knowledge prompt and clinical notes in a more targeted way.

Table 3: The effect of different methods on transferring between departments, this table selects the results of transfer from Respiratory department (Resp.) to other 4 departments. We choose the sample size $shots = 16$ and the number of diseases $k = 6$ as the parameters in the transfer learning experiment.

Resp.→	Gast.	Psy.	Neuro.	Gyn.
Fine tuning	74.92	53.08	54.75	46.92
Prompt tuning	74.83	73.08	68.58	66.50
MedKPL	**85.83**(+11)	**86.83**(+13.75)	**84.42**(+15.84)	**80.83**(+14.33)

Table 4: The effect of transferring MedKPL from Respiratory department (Resp.) to Gastroenterology department (Gast.) with different sample sizes was tested with $shots = 0, 2, 4, 8, 16$ and number of diseases $k = 6$.

Resp.→ Gast.			
shots	Fine tuning	Prompt learning	MedKPL
0	53.97	72.75	**84.92**
2	60.85	86.24	**89.98**
4	68.25	87.57	**90.48**
8	71.43	89.98	**90.48**
16	71.16	89.42	**91.53**

For the analysis of each department of classification task on the PPE dataset, we observe that the model's performance in the psychiatry (Psy.) department and the gynecology (Gyn.) department are highest both on multi-classification and binary classification tasks. By looking at the clinical notes in these two departments, we conjecture that the model's good performance is due to the low noise contained in the texts of these two departments.

4.3.2 Low-Resource Results

We conduct experiments on transfer learning across departments in PPE dataset and select the results of transferring from the Respiratory department (Resp.) to other departments in Table 3. The results in Table 3 show that the transferability of our method among departments outperforms the fine tuning and prompt learning methods by a large margin.

According to the results, there is also an in-

Table 5: The impact of different knowledge sources on the effect of MedKPL model, where the **Structured** is obtained from the Knowledge Graph, the **Unstructured** is obtained from online resources such as Wikipedia, the **Plain Text** uses the phrase *The disease requires timely medical attention.* as the text that does not contain medical knowledge, and the **Random** refers to randomly selected knowledge for augmentation. The results for each department are acquired by averaging the multi-classification results for disease number $k = 2, 4, 6$.

	Resp.	Gast.	Psy.	Neuro.	Gyn.	Overall
Structured	**85.17**	84.30	**94.03**	92.24	**95.12**	**90.17**
Unstructured	85.13	**84.90**	90.59	84.93	94.17	87.94
Plain Text	72.21	71.64	86.42	**92.62**	84.03	81.39
Random	54.31	61.37	83.36	81.12	69.17	69.87

teresting phenomenon that departments with lower text similarity have a higher improvement on classification accuracy, we conjecture that this is because our knowledge incorporation approach allows our model to discover the association between knowledge prompt and clinical notes in a more direct way. Also, by calculating the variance for all the results, we get the variance of 104.17 for fine tuning method and 83.55 for prompt learning method, while the variance of our method is 56.15, which is much lower than that of fine tuning and prompt learning. Therefore we speculate that our method can achieve higher classification results while having good robustness at the same time.

Besides transferring to other departments, We have also tested our method under different transfer shots to further demonstrate our model's few-shot learning capability. The results of transferring from Respiratory department (Resp.) to Gastroenterology department (Gast.) with different shots are shown in Table 4. It can be observed that under the zero-shot scenario, our method is far superior to the fine tuning and prompt learning methods. As the sample size rises, all methods witnesses an increase in transfer effect, but our method is still the best among the three methods.

Overall, our MedKPL model is more capable of transferring among departments and can also be better adapted to few or zero-shot scenarios.

4.3.3 Comparison among Knowledge Sources

To demonstrate that our model can uniform heterogeneous knowledge as input, we test different knowledge sources and their corresponding classification effects, results are shown in Table 5.

We begin by contrasting the structured knowledge prompt, derived from the knowledge graph, with the unstructured knowledge prompt, sourced from online search. Our findings demonstrate that the structured knowledge prompt outperforms its unstructured counterpart in terms of classification accuracy. This suggests that there exists a trade-off between the quality and accessibility of knowledge. While the structured knowledge prompt is more refined and contains less noise and irrelevant information, it is also more challenging to access. Conversely, unstructured free-text knowledge prompts offer almost limitless accessibility. For cases involving plain text, we employ the sentence ***The disease necessitates expedient medical attention.*** as the knowledge prompt. However, we observed that this non-medical knowledge prompt yielded significantly lower classification performance than the previous two methods. Furthermore, we conducted an experiment to disrupt the knowledge prompt by augmenting clinical notes with a random, irrelevant piece of knowledge prompt. Our results indicate that this method is the least effective among the four knowledge sources, with some outcomes even lower than the fine-tuning method. These findings reinforce the notion that knowledge prompts can contribute to improved classification outcomes in our approach.

In general, our approach can handle heterogeneous medical knowledge in a uniform way. The structured knowledge pormpt works most effectively, but is relatively difficult to obtain, while the unstructured knowledge can be accessed more easily, but at the expense of some performance.

4.3.4 Ablation Study

To explore how much the knowledge prompt contributes to our model, we conduct some ablation experiments of the impact of two main components: length of knowledge and soft

Table 6: Ablation study on a) knowledge length and b) soft attention. We test the knowledge truncation length from 0 to full length and test methods with or without soft attention mechanism, experimental parameters are kept consistent and the number of diseases $k = 6$.

(a)

Knowledge length	Acc.
full-length	**95.67**
40	95.33
30	94.33
20	95.33
10	94.50
0	93.33

(b)

Attention	Acc.
w/o Attention	93.67
w/ Clinical notes	94.17
w/ Knowledge	93.83
w/ Soft Attention	**95.67**

attention mechanism. Results are shown in Table 6.

It is noteworthy that the average length of medical knowledge in the Respiratory department is 36. The experimental results presented in Table 6a reveal that the model performs optimally when the medical knowledge is not truncated. We hypothesize that this is because larger truncation lengths promote the seamless integration of medical knowledge. Additionally, we evaluated the knowledge-enhanced classification module depicted in Figure 2(a) by comparing the soft attention mechanism with only clinical notes embeddings or knowledge prompt embeddings. The results in Table 6b demonstrate that the soft attention mechanism is instrumental in directing the model's focus towards the knowledge-laden attributes of the clinical notes, thereby leading to superior classification outcomes.

5 Conclusion

In this paper, we propose a MedKPL model and achieve state-of-the-art classification results on two medical EHR datasets. With the advantage of knowledge extraction and uniform process, our model can eliminate the difference among different sources and organize all knowledge into one representation style. The knowledge incorporation and soft attention mechanism between knowledge prompt and clinical notes enable the model to be more robust and achieve appreciable improvement on medical text classification tasks. The introduction of knowledge and prompt learning method exploits better few-shot and zero-shot transferability among departments.

References

Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. 2017. Understanding of a convolutional neural network. In 2017 international conference on engineering and technology (ICET), pages 1–6. Ieee.

Leo Breiman. 2001. Random forests. Machine learning, 45(1):5–32.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901.

Che-Wen Chen, Shih-Pang Tseng, Ta-Wen Kuan, and Jhing-Fa Wang. 2020. Outpatient text classification using attention-based bidirectional lstm for robot-assisted servicing in hospital. Information, 11(2):106.

Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In Proceedings of the ACM Web Conference 2022, pages 2778–2788.

Yahui Chen. 2015. Convolutional neural network for sentence classification. Master's thesis, University of Waterloo.

Thomas Cover and Peter Hart. 1967. Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1):21–27.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

M Julia Flores, Ann E Nicholson, Andrew Brunskill, Kevin B Korb, and Steven Mascaro. 2011. Incorporating expert knowledge when learning bayesian network structure: a medical case study. Artificial intelligence in medicine, 53(3):181–204.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. arXiv preprint arXiv:2012.15723.

Vijay N Garla and Cynthia Brandt. 2013. Knowledge-based biomedical word sense disambiguation: an evaluation and application to clinical document classification. Journal of the American Medical Informatics Association, 20(5):882–886.

Karim Gasmi. 2022. Medical text classification based on an optimized machine learning and external semantic resource. Journal of Circuits, Systems and Computers.

Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Juanzi Li, and Maosong Sun. 2021. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. arXiv preprint arXiv:2108.02035.

Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In European conference on machine learning, pages 137–142. Springer.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Ji Young Lee and Franck Dernoncourt. 2016. Sequential short-text classification with recurrent and convolutional neural networks. arXiv preprint arXiv:1603.03827.

Fei Li and Hong Yu. 2020. Icd coding from clinical text using multi-filter residual convolutional neural network. Proceedings of the AAAI Conference on Artificial Intelligence, 34(05):8180–8187.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical reasoning on Chinese morphological and semantic relations. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 138–143, Melbourne, Australia. Association for Computational Linguistics.

Xuedong Li, Yue Wang, Dongwu Wang, Walter Yuan, Dezhong Peng, and Qiaozhu Mei. 2019. Improving rare disease classification using imperfect knowledge graph. BMC Medical Informatics and Decision Making, 19(5):1–10.

Huiying Liang, Brian Y Tsui, Hao Ni, Carolina Valentim, Sally L Baxter, Guangjian Liu, Wenjia Cai, Daniel S Kermany, Xin Sun, Jiancong Chen, et al. 2019. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. Nature medicine, 25(3):433–438.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. arXiv preprint arXiv:1605.05101.

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

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025.

Melvin Earl Maron. 1961. Automatic indexing: an experimental inquiry. Journal of the ACM (JACM), 8(3):404–417.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 26.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever,

et al. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res., 21(140):1–67.

Timo Schick, Helmut Schmid, and Hinrich Schütze. 2020. Automatically identifying words that can serve as labels for few-shot text classification. arXiv preprint arXiv:2010.13641.

Timo Schick and Hinrich Schütze. 2020. Exploiting cloze questions for few shot text classification and natural language inference. arXiv preprint arXiv:2001.07676.

Tianxiang Sun, Yunfan Shao, Xipeng Qiu, Qipeng Guo, Yaru Hu, Xuanjing Huang, and Zheng Zhang. 2020. Colake: Contextualized language and knowledge embedding. arXiv preprint arXiv:2010.00309.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. Kepler: A unified model for knowledge embedding and pre-trained language representation. Transactions of the Association for Computational Linguistics, 9:176–194.

Liang Yao, Zhe Jin, Chengsheng Mao, Yin Zhang, and Yuan Luo. 2019a. Traditional chinese medicine clinical records classification with bert and domain specific corpora. Journal of the American Medical Informatics Association, 26(12):1632–1636.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019b. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. BMC medical informatics and decision making, 19(3):31–39.

Taolin Zhang, Chengyu Wang, Nan Hu, Minghui Qiu, Chengguang Tang, Xiaofeng He, and Jun Huang. 2022. Dkplm: Decomposable knowledge-enhanced pre-trained language model for natural language understanding. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 11703–11711.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. arXiv preprint arXiv:1905.07129.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In Proceedings of the IEEE international conference on computer vision, pages 19–27.

Guido Zuccon, Amol S Wagholikar, Anthony N Nguyen, Luke Butt, Kevin Chu, Shane Martin, and Jaimi Greenslade. 2013. Automatic classification of free-text radiology reports to identify limb fractures using machine learning and the snomed ct ontology. AMIA Summits on Translational Science Proceedings, 2013:300.