Hindawi
Evidence-Based Complementary and Alternative Medicine
Volume 2023, Article ID 2088698, 11 pages
https://doi.org/10.1155/2023/2088698

Research Article

Combining the External Medical Knowledge Graph Embedding to Improve the Performance of Syndrome Differentiation Model

Qing Ye, Rui Yang, Chun-lei Cheng, Lin Peng, and Yong Lan

School of Computer, Jiangxi University of Chinese Medicine, Nanchang, Jiangxi, China

Correspondence should be addressed to Chun-lei Cheng; chunlei_cheng@163.com

Received 3 August 2022; Revised 10 October 2022; Accepted 24 November 2022; Published 1 February 2023

Academic Editor: Tiancai Wen

The electronic medical records (EMRs) of traditional Chinese medicine (TCM) include a wealth of TCM knowledge and syndrome diagnosis information, which is crucial for improving the quality of TCM auxiliary decision-making. In practical diagnosis, one disease corresponds to one syndrome, posing considerable hurdles for the informatization of TCM. The purpose of this work was to create an end-to-end TCM diagnostic model, and the knowledge graph (KG) created in this article is used to improve the model's information and realize auxiliary decision-making for TCM disorders. We approached auxiliary decision-making for syndrome differentiation in this article as a multilabel classification task and presented a knowledge-based decision support model for syndrome differentiation (KDSD). Specifically, we created a KG based on TCM features (TCMKG), supplementing the textual representation of medical data with embedded information. Finally, we proposed fusing medical text with KG entity representation (F-MT-KER) to get prediction results using a linear output layer. After obtaining the vector representation of the medical record text using the BERT model, the vector representation of various KG embedded models can provide additional hidden information to a certain extent. Experimental results show that our method improves by 1% (P@1) on the syndrome differentiation auxiliary decision task compared to the baseline model BERT. The usage of EMRs can aid TCM development more efficiently. With the help of entity level representation, character level representation, and model fusion, the multilabel classification method based on the pretraining model and KG can better simulate the TCM syndrome differentiation of the complex cases.

1. Introduction

With the fast advancement of the medical information technology in recent years, the number of electronic medical records (EMRs) accessible has increased. EMRs are mostly semistructured or unstructured text records of the diagnostic and treatment processes. EMRs serve as the central repository for big medical data, which contains a wealth of medical information and therapeutic concepts. The course record in medical records can be divided into four diagnostic information, physical examination, chief complaint, syndromes, treatment plan, and traditional Chinese medicine (TCM) prescription.

In general, there are few same syndromes in TCM diagnosis. Usually, a case corresponds to a syndrome, or a patient corresponds to a syndrome. The number of syndromes is far too many to classify. Therefore, we proposed the task of transforming the syndrome differentiation problem into identifying the nature and locations of the disease. This way, we can transform the TCM syndrome differentiation auxiliary decision-making task from the original multiclassification task into a multilabel classification task, where the disease's nature and location are regarded as labels. However, doctors give clinical prescriptions based on the diagnosed syndromes. As a result, extensive TCM clinical expertise and excellent TCM diagnostic understanding are required in the syndrome differentiation approach. In order to imitate the TCM dialectics process, we need to introduce the information implied by the symptoms and chief complaints in each medical record into the medical record text representation. So, we constructed a knowledge graph (KG) based on TCM features (TCMKG) to excavate hidden information in the medical records to solve this problem.

The language model of a deep neural network has developed into an efficient text representation method in recent years. BERT [1] has shown promise in some natural language processing (NLP) tasks. We adopted BERT for EMRs multilabel classification and extended the model with knowledge representation with hidden information provided in the shared task, such as four diagnostic information, chief complaint, and physical examination. Experiments on a test set of EMRs demonstrated the efficacy of our method.

The following are the contributions of this paper: (1) To get around the problem of not being able to use EMRs for information-based dialectics, we turned TCM's standard syndrome differentiation work into a multilabel categorization assignment. (2) In this paper, we proposed syndrome differentiation assistant decision-making based on knowledge (KDSD) to integrate implicit knowledge graph embedding (KGE) from TCMKG into the syndrome differentiation assistant task. (3) We designed a method to adopt the fusion of medical text and knowledge graph entity representation (F-MT-KER) to alleviate the apparent impact of the personalization of the TCM medical records on decision-making.

2. Related Works

Machine learning often accomplishes multilabel classification tasks by changing the current algorithm or employing a binary classification technique. Unlike traditional machine learning, the deep learning method uses various neural network structures to extract the semantic embedding of input text. XML-CNN [2] used a one-dimensional convolutional neural network along the sequence length and word embedding dimension to represent the text input. Slice considers supervised pretrained dense embeddings from XML-CNN model as the input of its hierarchical linear model. Recently, based on XML-CNN, AttentionXML [3] used BiLSTM, added attention to labels to design scoring functions, and used a hierarchical tag tree to warm up the model.

Many studies have been committed to pretraining language representation models to obtain language information from a text and then use this information for specific NLP tasks. The Elmo [4], GPT [5], and BERT [1] models have achieved better results in multiple NLP tasks. After fine-tuning, these pretraining models can be applied to various tasks. However, due to the lack of information linkages between vertical applications and open domains, these models cannot be easily translated to specialized domains. The model can be pretrained in specific fields to solve this problem, but it is not desirable in terms of computational time and cost. For example, RoBERTa [6], XLNET [7], SpanBERT [8], and ERNIE [9] are trained in the abovementioned way. Recent research has demonstrated that in addition to the input, adding other relevant knowledge to the model can improve the model's performance to variable degrees, such as reading comprehension [10], text classification [11], natural language inference [12], knowledge acquisition [13], and question answer [14]. Therefore, we argued that additional knowledge information could effectively benefit the existing pretraining model. Some studies have tried effectively controlling the external KGs through the joint representation learning of words and entities to achieve promising results [15–18]. Chen et al. [19] used BiLSTM to process text and introduced additional knowledge through specific attention mechanisms. Zhang et al. [20] tried to use external KG to enrich embedded information and improve language understanding.

In TCM data mining, Zhou [21] built a data model from multiple information entities and their interactions in outpatient data to fulfill large-scale clinical data integration and preprocessing activities, as well as to uncover useful clinical knowledge from the data. Liu et al. [22] approached a clinical data warehouse based on structured EMR data. Clinical terminology is collected to establish clinical hypotheses and aid in the discovery of clinical knowledge from large-scalereal-world TCM clinical data. The majority of TCM data mining involves constructing data into a database and then mining information from it to realize the application [23]. In recent years, the auxiliary decision-making systems based on expert systems [24] and the model-assisteddecision-making systems [25] based on deep learning have dominated TCM diagnosis. Joyce et al. [26] described the cognitive work of TCM tongue diagnosis and used it to find TCM diagnostic thinking patterns. An expert panel must assess the correctness of the diagnosis. Xie et al. [27] mined links between hidden items and inferred pathways from symptoms to syndromes using TCM knowledge graphs and reinforcement learning techniques. Yang et al. [28] utilized KGE to find triples connection pairings of syndromes and symptoms, and then created a score algorithm to find suggested syndromes. To apply clinically aided diagnostic and treatment models, Ruan et al. [29] stacks semantic-aware graph convolutions to learn efficient low-dimensional representations of nodes via metagraphs and self-attention, and predicts correct patterns via clustering and linking. Xie [30] employed multiclassifier integrated TCM syndrome element classification to help with diagnosis, mainly using the Nave Bayes, Weighted bipartite graph, SVM, and ProSVM algorithms. Since the number of syndromes in actual diagnosis is unclear, we split syndromes into labels of illness type and disease location, used multilabel classification to predict labels, and utilize a certain number of labels to replace ambiguity.

3. Methods

3.1. Overview. The KDSD model is divided into two components, as illustrated in Figure 1, the EMRs module and the KG module. We first develop the TCMKG using the EMRs and then represent all entities and relationships in KG through the KG embedding model. Input the EMRs into the model, and the EMRs module will extract the chief complaint and symptom information (four diagnostic information and physical examination). Meanwhile, the embedding representation of the knowledge entities in the

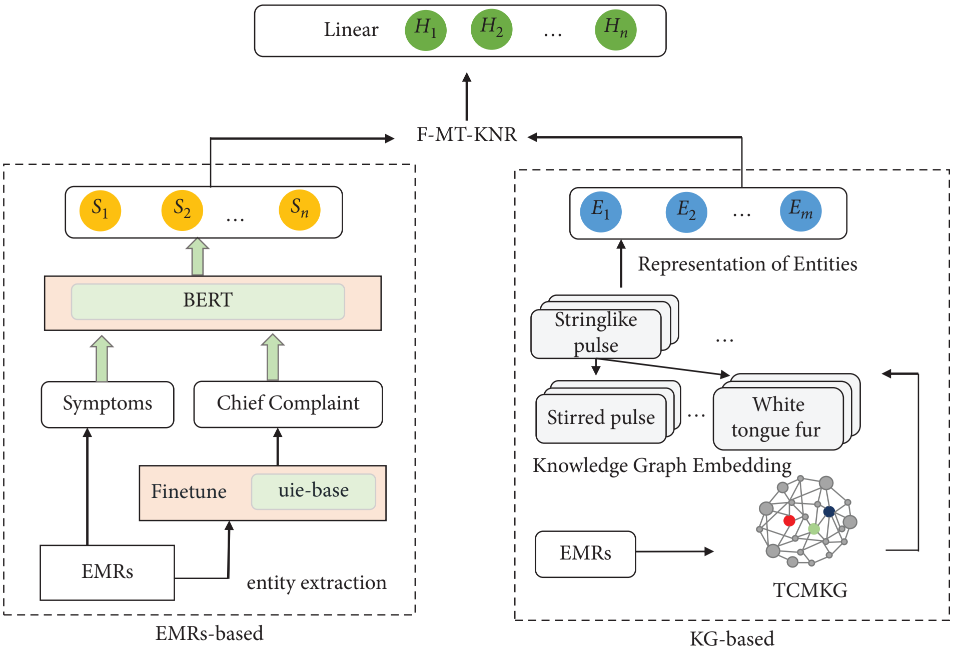

Figure 1: The architecture of the KDSD model.

KG module is obtained by extracting and linking the entities of the EMRs module. Finally, we fused text representations and KG entity embedding representations through the F-MT-KNR method designed in this paper to predict the nature and location of diseases through multilabel classification. The implementation details of the model are described below.

3.2. Construction of Knowledge Graph. Different doctors use different syndrome differentiation procedures, which results in them prescribing different syndromes for the same illness and even achieving different syndrome outcomes for the same patient. In TCM, syndromes include the relationship between evil and positive, internal and external pathogenic factors, the location of the disease, and the nature of the disease. Therefore, it is objective to use the nature and location of the disease to assist doctors in making clinical decisions.

In the data mining phase, we extracted the TCM symptom entities and chief complaint information from the desensitized EMRs. We then decomposed the information on the nature of the illness and its location in the syndrome. We designed the schema layer top-down [31], which is actually to classify these entities. As a result, the schema layer's categories and the data layer's medical entities establish "class-example" connections. The data layer is filled with entities simultaneously, and the TCMKG is completed. Figure 2 shows the partial TCMKG.

After desensitization, entity extraction, and relationship extraction from 12533 EMRs, we created TCMKG. The TCMKG database has 115012 triples, which include 18078 symptom entities, 2098 chief complaint entities, and 157 nature and location of the disease entities. In addition, Neo4j is employed as a data storage tool for TCMKG.

3.3. Representation of Knowledge Graph. In the EMRs of TCM, there is a very complex relationship between the four diagnostic information and syndrome or disease nature and location, which experts of TCM can only understand. By contrast, KG is a multirelational graph comprised of many different entities and their relationships [32]. In this paper, KG can accurately describe the relationship between symptoms and syndromes in TCM through information extraction, data mining, and knowledge reasoning. It can describe the evolution process and development law of knowledge. To provide accurate, traceable, interpretable, and inferential knowledge data for syndrome auxiliary decision-making. Besides, the representation learning of KG is the method of transforming these knowledge data with implicit information into a vector representation.

Given a knowledge graph K, which includes a collection of entities ε and relations $\mathcal{R}$, that is, $K \subseteq \varepsilon \times \mathcal{R} \times \varepsilon$. The triples are expressed as (h, r, t), where $h, r \in \varepsilon$ denote the head and tail entities, respectively, and $r \in \mathcal{R}$ denotes the relationship between them.

For each entity e and relationship r in K, the knowledge graph representation methods generate $e_v \in R^{d_e}$ and $r_v \in R^{d_r}$, where e_v and r_v are d_e and d_r dimensional vectors, respectively. Each embedding technique incorporates a scoring mechanism, $\mathcal{F}\colon \varepsilon \times \mathcal{R} \times \varepsilon \longrightarrow R$ will give some scores $\mathcal{F}(h, r, t)$ is assigned to a possible triple (h, r, t), $h, t \in \varepsilon$, , and $r \in \mathcal{R}$. The model is trained in such a way that for each correct triple $(h, r, t) \in K$ and wrong triple $(h', r', t') \notin K$, the model assigns a score such that $\mathcal{F}(h, r, t) > 0$ and $\mathcal{F}(h', r', t') < 0$. Typically, a scoring function is a function of (e_h, e_r, e_t).

We selected several representative KG embedding methods to conduct experiments on TCMKG (see the experimental part) and evaluated the indicators, among which TransE performed best.

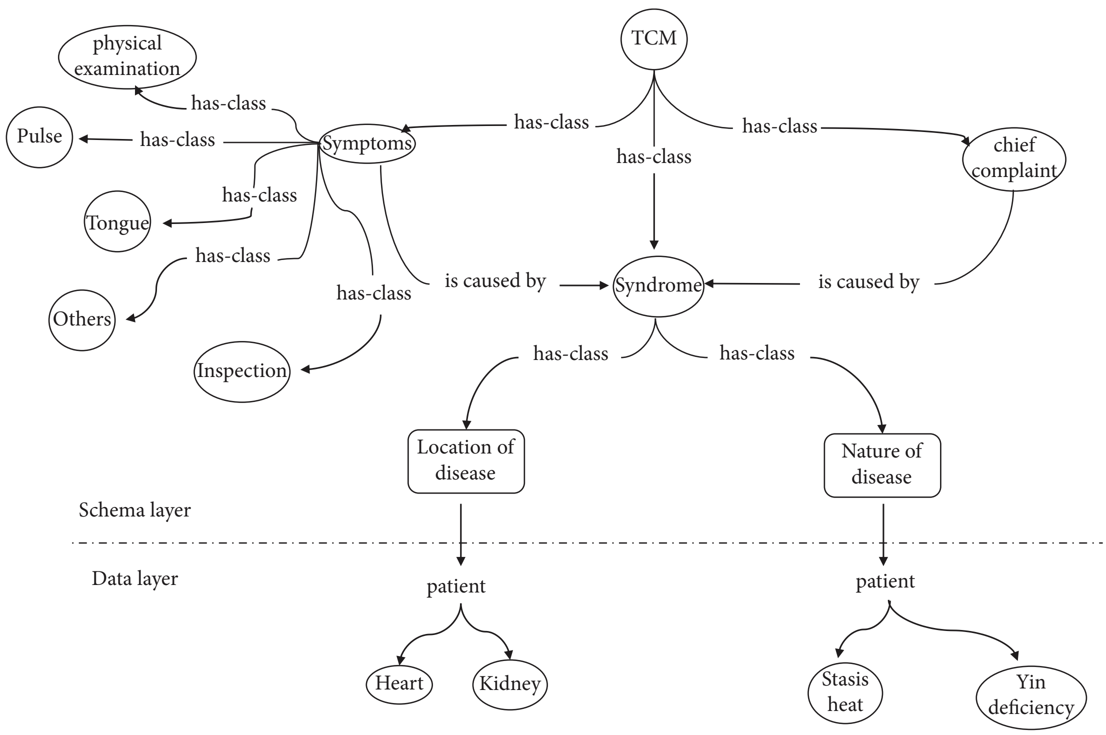

Figure 2: Ontology representation of TCMKG.

Bordes et al. [33] proposed TransE based on the assumption that the additional $h + t$ embedding should be near to that of t and that the scoring function should be defined as follows under L_1 or L_2 constraints:

$$f_r(h,t) = \|h + r - t\|_{L1/L2}. \tag{1}$$

As mentioned above, $h, t \in \varepsilon$, and $r \in \mathscr{R}$. After the KG embedding is completed, formula (1) makes $f(h,r,t) > 0$ for all true triples, and $f(h,r,t) < 0$ for all wrong triples.

3.4. EMR-Based Module. The BERT layer is used by the EMRs module to extract features from incoming text input. The embeddings of the KDSD model differ from those of the BERT model. We use outside information to complement the text's representation. We separate the information about the principal complaint from the information about the symptoms when we analyze the input data, and the format of the input sequence is specified by

$$[\text{CLS}]\text{symptoms}[\text{UNK}]\text{chief complaint}[\text{UNK}][\text{SEP}], \tag{2}$$

where [CLS] and [SEP] are the BERT model-specific special symbols, and [UNK] is the separator inside the text. Symptoms include pulse-taking, tongue examination, listening and smelling, inspection, and physical examination. Each symptom and chief complaint field is separated by [UNK] to facilitate the integration of subsequent KGE.

After entering the EMRs into the model, it is required to improve the text representation via the hidden layer. However, the symptom information and the chief complaint information are stated differently in EMRs, as is the length of the text. Since the maximum input length of BERT is generally 512 characters, considering that this paper needs to separate symptoms and chief complaints, and the description of private complaints is too complicated, the length of some medical records exceeds the limit of the model. In this paper, most of the information on the four diagnostic information is relatively standardized. To decrease the text length, we apply the fine-tuned Uie-base [34] model to extract the main complaint information from the text. Then, as stated in the following equations, we employed the multi-headself-attention [35] process to integrate the principal complaint information into the text representation of symptom information.

$$Q = K = V = W^S \text{Concat}\left([\text{C}]; S_{1\ldots M}, C_{1\ldots N}\right), \tag{3}$$

$$[C'] = \text{Concat}\left(\text{head}_1, \ldots, \text{head}_{\text{i}}\right)W^O, \tag{4}$$

$$\text{head}_{\text{i}} = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right), \tag{5}$$

where $[C]$ is the representation of [CLS] in the hidden layer state, $[C']$ is the representation of input text. $S_{1\ldots M}$ and $C_{1\ldots N}$ are symptom information embedding containing M values and chief complaint information embedding containing N values. W^s, W^O, W^Q, W^K, and W^V are trainable parameters.

Integration of additional medical knowledge.

The fusion information part aims to integrate the output of KGE into the output of BERT. BERT is to train the representation of each character in the unit of characters, while in the KGE module, we can only train the entities in KG into triples. Therefore, we proposed F-MT-KER to realize the fusion of characters and entity embeddings in KG, as shown in Figure 3.

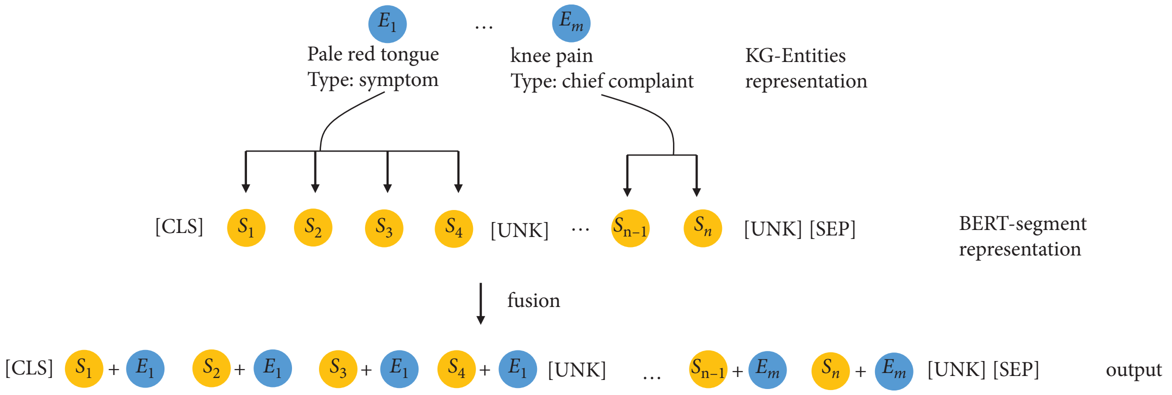
E_1
…
E_m
Pale red tongue
Type: symptom
knee pain
Type: chief complaint
KG-Entities representation
[CLS]
S_1
S_2
S_3
S_4
[UNK]
…
S_{n-1}
S_n
[UNK] [SEP]
BERT-segment representation
fusion
[CLS] S_1 + E_1 S_2 + E_1 S_3 + E_1 S_4 + E_1 [UNK] … S_{n-1} + E_m S_n + E_m [UNK] [SEP]
output

Figure 3: The process of fusion of knowledge graph entity information and character information of the BERT model.

$$S = \sum_{i=1}^{i=k}\left[\left(S_j, \ldots, S_{j+L\left(e_k\right)}\right) + e_k\right]. \qquad (6)$$

In equation (6), e_k is the embedding of an entity in KG, $S_j, \ldots, S_{j+L(e_k)}$ corresponds to symptom character set or chief complaint character set of e_k. L is the function of calculating the length of the entity and f is to select the corresponding e_k. The embedded representation of the same entity is added to the characters of each entity. Through the linear layer transformation, the fused vector dimension is changed from the dimension of the hidden layer to the dimension of the label quantity.

4. Results

4.1. Overview. The process of auxiliary decision-making for syndrome differentiation can be divided into the KG construction and embedding module, the text representation of the BERT model, and multilabel classification based on model fusion. For the medical records of the input model, we first take the medical record entity set through TCMKG and then link the entities in this entity set to the vector space embedded in TCMKG through entity matching. As a result, we can obtain vector representations of symptom entities and chief complaint entities. Ultimately, the nature and location of the disease recommended by the medical record can be obtained by inputting the embeddings containing hidden information into the model.

4.2. Dataset and Experimental Setup. We conducted experiments on the EMRs dataset and TCMKG. As seen in Figure 4, we removed sensitive information from the original EMRs and extract entities for pulse diagnosis, tongue diagnosis, and the other four diagnostic information based on the punctuation marks. After fine-tuning the medical record annotation data, the symptom information in the chief complaint is retrieved using the Uie-base model. For example, from the text information, the symptoms of “oral ulcer” and “unfavorable urination” may be extracted from “oral ulcer and unfavorable urination” that recur repeatedly for 2-3 years. It is worth mentioning that when the probability of the anticipated symptom item is less than 0.5, the entity is deleted.

The dataset includes 12533 EMRs from QIHUANG TCM. 10026 EMRs were used for training and 2507 were used for testing. Furthermore, our criterion for dividing the test set is that only the test data labels that appear in the training set can be divided into the test set. This part of the data accounts for a tiny proportion, so we used the model to generate their input randomly. There are 157 classifications for disease nature and disease location. Every text in the input model is a medical case, and each text has an average of 102 characters. The relationship, number of edges, and triples in the knowledge graph constructed in this paper are shown in Table 1.

We constructed the knowledge ontology based on Yu [31] and defined entity and relationship description as the core. It includes some knowledge in EMRs, such as pulse diagnosis, tongue diagnosis, smell diagnosis, observation diagnosis, chief complaint, and other knowledge.

This paper preprocessed EMRs through data desensitization, cleaning, structuring, filtering, and diagnostic label standardization. In data filtering, duplicate information, and information that has little impact on diagnosis (such as the date in chief complaint) will be deleted. On the one hand, it can fulfill the BERT model input’s format criteria. In TCMKG, on the other hand, it may also maintain concealed information. Our BERT model was bert-base-chinese, with the following primary settings: hidden size 768, maximum position embedding 512, number of epochs 50, number of attention heads 12, number of hidden layers 12, maximum input length 256, learning rate $2e-5$, and batch size 8. Our NVIDIA T4 GPU is the backbone of all our experiments (17.18G).

4.3. Evaluation Metrics. Table 2 displays the experimental outcomes on the EMRs dataset. P@k (Precision at k) was used as evaluation metrics.

$$P@k = \frac{1}{k}\sum_{l=1}^{k} y_{\mathrm{rank}(l)}. \qquad (7)$$

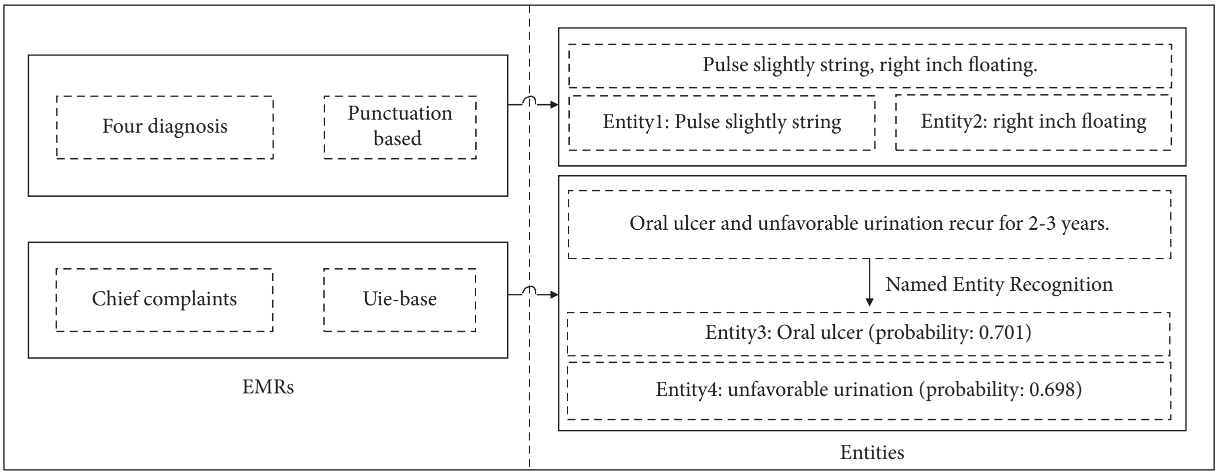
Four diagnosis
Punctuation based
Chief complaints
Uie-base
EMRs
Pulse slightly string, right inch floating.
Entity1: Pulse slightly string
Entity2: right inch floating
Oral ulcer and unfavorable urination recur for 2-3 years.
Named Entity Recognition
Entity3: Oral ulcer (probability: 0.701)
Entity4: unfavorable urination (probability: 0.698)
Entities

Figure 4: Entity extraction method for four diagnosis and chief complaint information.

The index of the lth highest prediction label is rank(l) and y is the true binary vector.

4.3.1. Hamming Loss. The Hamming loss is the proportion of wrongly predicted labels. The higher the performance, the lower the Hamming loss value. $\hat{y}_i$ is the predicted value for the jth label of a given sample, y_i is the corresponding true value, and n_{labels} is the number of classes or labels.

$$L_{\text{Hamming}}(y, \hat{y}) = \frac{1}{n_{\text{labels}}} \sum_{j=0}^{n_{\text{labels}}-1} 1\left(\hat{y}_i \neq y_i\right). \tag{8}$$

4.3.2. Average Precision. The weighted mean of precisions acquired at each threshold is used to summarize a precision-recall curve, with the increase in recall from the previous threshold used as the weight. Where P_n and R_n are the precision and recall at the nth threshold.

$$AP = \sum_n \left(R_n - R_{n-1}\right)P_n. \tag{9}$$

4.3.3. Label Ranking Loss. The label ranking loss function computes the ranking loss, which is weighted by the inverse of the number of ordered pairs of false and true labels.

$$\text{ranking}_{\text{loss}(y,\hat{f})} = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{sample}-1}} \frac{1}{\|y_i\|_0\left(n_{\text{labels}} - \|y_i\|_0\right)} \left|\left\{(k,l)\colon\ \widehat{f_{ik}} < \widehat{f_{il}}, y_{ik} = 1, y_{il} = 0\right\}\right|. \tag{10}$$

4.4. Performance on EMRs Dataset. Liu et al. [36] used a data improvement and model fusion technique to accomplish syndrome differentiation classification of lung cancer diagnosis using the neural network text classification models TextCNN [37], TextRNN [38], RCNN [39], and fastText [40]. Hu et al. [41] employed the CNN and fastText models to complete the task of yin and yang deficient syndrome text categorization.

In addition, this study employs AttentiveConvNet [42], DPCNN [43], Transformer [35], and AttentionXML as comparative models for the multilabel text classification problem. A variety of context information is added to local convolution processes by AttentiveConvNet. Long-distance text dependencies may be extracted using the DPCNN. AttentionXML is a structural model for tag trees based on the attention mechanism. Using a multilabel attention approach for text multilabel categorization, the most essential feature information is collected for each label, allowing the relationship information between the input text and each label to be properly comprehended. Moreover, the KG-based algorithm used the KG's information to forecast relationships that will deliver the desired results. This paper proposes the KDSD model. Bert-base-chinese is the pretraining model that is employed. By linking the linear layer after feature extraction using the model, a multilabel classification result is achieved.

4.5. The Results of Knowledge Graph Representation. In order to demonstrate the KGE model's efficacy, we used the mainstream KGE methods [33, 44–46] to evaluate the TCMKG. For each KGE approach, we produce the associated knowledge representation and add it to the KDSD model. Table 3 illustrates the KDSD model's performance when various KGE models are employed as knowledge supplements.

We employ the MRR (Mean Radical Ranking), MR (Mean Rank), and Hits@N to analyze the performance metrics of KGE [47]. The MRR is the average of the triples' reciprocal rankings. The value ranges from 0 to 1; the greater the number, the better the model. MR is the average of all the triples' rankings. The value ranges from 1 (ideal condition,

Table 1: Number of relation categories and examples of triples in TCMKG.

Relationship categories	Number of edges	*Triples*	
		Chinese	English
Chief complaint-syndrome	16277	<失眠, 主诉_证候, 胆热痰扰>	< Insomnia, chief complaint-syndrome, gallbladder heat, and phlegm disturbance>
Symptom-syndrome	62308	<两寸略弱, 主诉_证候, 胆热痰扰>	< Two inches slightly weaker, symptom-syndrome, gallbladder heat, and phlegm disturbance>
Syndrome-nature of disease	19327	<胆热痰扰,证候_病性, 热>	< Gallbladder heat and phlegm disturbance, syndrome-nature of disease, and heat>
Syndrome-location of disease	17100	<胆热痰扰,证候_病位, 胆>	< Gallbladder heat and phlegm disturbance, syndrome-location of disease, and gallbladder>

Table 2: The results on the EMRs dataset.

Models	P@1	P@3	P@5	Average precision	Hamming loss	Label ranking loss
AttentiveConvNet	0.416	0.404	0.384	0.263	0.081	0.136
DPCNN	0.690	0.63	0.574	0.528	0.045	0.079
Transformer	0.721	0.665	0.61	0.578	0.04	0.086
AttentionXML	0.749	0.702	0.645	0.537	0.043	0.078
TextCNN	0.753	0.686	0.636	0.603	0.039	0.069
TextRNN	0.769	0.719	0.668	0.646	0.034	0.068
FastText	0.809	0.748	0.686	0.676	0.034	0.046
BERT	0.89	0.876	0.845	0.857	0.026	0.016
KG-based	0.059	0.073	0.066	—	—	—
KDSD	0.9	0.879	0.852	0.866	0.029	0.014

Table 3: Performance of different embedding models on KDSD.

Methods	P@1	P@3	P@5
BERT	0.89	0.876	0.845
+TransE	0.9	0.879	0.852
+DistMult	0.878	0.856	0.829
+ComplEx	0.868	0.852	0.825
+ConvKB	0.886	0.862	0.833

all rankings are equal) to the number of corruptions. Hits@n is the proportion of calculated ranks that are higher than (or equal to) a rank of n. The value ranges from 0 to 1; the greater the number, the better the model.

5. Discussion

5.1. Principal Results. As seen in Table 2, based on pretraining models, BERT and KDSD models outperform fastText, TextCNN, Text RNN, Transformer, and other models. KDSD increases by 1%, 0.3%, and 0.52% in the metrics P@1, P@3, and P@5, respectively, as compared to the BERT model.

From the experimental results, the results of the pretraining model are better than the previous text-based multilabel classification method. Furthermore, including KGE information in the pretraining model can improve the model's performance. However, all KG-Based indicators were terrible, far lower than the KDSD results. This might be due to one of the two factors. One is that KG-based method only used the information embedded in the graph and obtains the labels through link prediction. Another possibility is that TCMKG is a KG constructed for TCM diagnosis. Although the diagnosis of each medical record is different, the reuse rate of the information of the four diagnoses and the chief complaint is high. Besides, various clinicians define diseases differently, the number of triples for some conditions may be inadequate for good prediction.

From Table 3, it can be seen that the average precision of the KDSD model is 0.9% higher than that of the BERT model. This is due to the fact that in actual text task processing, some training corpora are difficult to obtain; their overall number and the total number of words contained are very small, making them unsuitable for training models with embedding layers. However, these data are suitable for training models without embedding layers. In addition, it provides useful rules that may be extracted by the model. Using a pretrained model to encode the original text is an excellent choice in this instance because the pre-trained model is derived from a large corpus and can make the current text meaningful, despite the fact that these implications may not be domain-specific. However, these deficiencies can be remedied by utilizing fine-tuned models.

5.2. Results of KGE. The KG-based experiment is another experiment undertaken for this investigation. If only the knowledge graph and its embedded data are used, the link prediction is performed on multiple triples associated with the same medical record, followed by the application of the KG's embedded scoring function. The expected labels received were evaluated, and the label with the highest score was finally selected. The results are bad due to the high rate of recurrence of the four diagnoses in TCM, the excessive number of triple diagnoses that may be derived from a medical file, and the inability to establish a connection between them. When this experiment is compared to the KDSD model and the BERT model, it is clear that using the information of the KG alone will not provide excellent results; instead, a deep learning model must be used in conjunction with it to get superior outcomes.

Although the KG-based method has no advantages in various indicators, the results of KDSD are indeed better than BERT in some indicators. It shows that the fusion with the embedded information from TCMKG can improve the performance of syndrome decision-making model. Moreover, as shown in Table 3, KDSD has the best performance after using the TransE model, surpassing BERT in all indicators. However, the results of some models have even declined after being fused with BERT. Here we evaluated the performance of KGE. Table 4 shows the performance evaluation findings for several embedding strategies. As seen in the table, TransE outperforms other models in MRR, MR, Hits@N, and has the greatest performance in the KDSD model for link prediction.

We analyze the reasons for the following reasons. First, complicated embedding approaches, such as adding complex vector space and convolution methods, will not function when the repeat rate of the four diagnosis information and the principal complaint information is high. Second,

Table 4: Performance evaluation of knowledge graph embedding.

Methods	MRR	MR	Hit@1	Hit@10	Hit@100
TransE	0.827	171	0.764	0.912	0.946
DistMult	0.203	492	0.13	0.357	0.633
ComplEx	0.201	530	0.128	0.35	0.625
HolE	0.208	1103	0.133	0.363	0.539
ConvKB	0.157	1011	0.072	0.361	0.551

since the triples of TCMKG construction adopt "symptom-nature\location" and "chief complaint-nature\location," we think the TransE model's distance-based scoring function calculates the Euclidean distance between them, which has more extraordinary expression ability, utilizing this intuitive distance-based technique.

5.3. Limitations. In this section, we discuss some limitations of the KDSD model. First, the new medical record cannot ensure that the input symptoms and chief complaints can match the entities in the TCMKG to ensure the existing label. Therefore, the embedded knowledge integrated into the model has little information, affecting the model's performance. Second, due to the particularity of TCM dialectics, although there is enough information on the syndrome, some syndromes' descriptions are still too simple. For example, if the syndrome is damp-heat obstructing the meridian (湿热阻滞经脉), then, we can get that the location of the disease is the meridians (经脉), and the nature of the disease is obstruction (阻滞) and damp-heat (湿热). This syndrome contains little information about the nature and location of the disease, so such medical records will have an adverse impact on the model's prediction of multiple labels when tested.

6. Conclusion

The auxiliary syndrome differentiation work has been viewed as a multiclassification issue. This paper presents the KDSD model for converting multiclassification into a multilabel classification. The KDSD model improves the model's performance by integrating the textual information of the EMR and the entity information embedded in the KG. We constructed TCMKG by using EMRs. Then, we represented TCMKG by TransE, and connected the results to the input layer of BERT. Finally, the results are obtained through the full connection output of the model. The results show that although BERT has been an excellent multilabel classification model in recent years, adding domain-specific knowledge graph information can improve the model's performance.

In the future, we will further study the construction of KG in TCM and explore the critical technologies of model fusion to improve the accuracy of the syndrome differentiation decision-making system. We find that some symptom entities in EMRs are not included in TCMKG (we only consider the existence of labels when building the dataset, some symptom entities will generate vectors randomly when predicting) to introduce other TCM knowledge that contains more symptom, and syndrome entities is an effective feature for syndrome differentiation.

Abbreviations

EMRs:	Electronic medical records
TCM:	Traditional Chinese medicine
KG:	Knowledge graph
KDSD:	Knowledge-based decision support model for syndrome differentiation
TCMKG:	Knowledge graph based on TCM features
F-MT-KER:	Fusing medical text with KG entity representation
KGE:	Knowledge graph embedding.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Qing Ye and Rui Yang contributed equally to this work.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (no. 62141202), Jiangxi Provincial Department of Education Science and Technology Research Key Project (GJJ201204), Jiangxi Provincial Natural Science Foundation (no. 20224BAB206102), Jiangxi Postgraduate Innovation Special Fund Project (YC2021-S498), and Jiang-xi University of Chinese Medicine Science and Technology Innovation Team Development Program (no. CXTD22015).

References

[1] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the naacL-HLT*, pp. 4171–4186, Association for Computational Linguistics, Minneapolis, Minnesota, June 2019.

[2] J. Z. Liu, W. C. Chang, Y. X. Wu, and Y. M. Yang, "Deep learning for extreme multi-label text classification," in *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pp. 115–124, ACM, Shinjuku, Tokyo, Japan, August 2017.

[3] R. H. You, Z. H. Zhang, Z. Y. Wang, and S. Y Dai, "Attentionxml: label tree-basedattention-aware deep model for high-performance extreme multi-label text classification," in *Proceedings of the 33rd Conference on Neural Information*

Processing Systems (NeurIPS 2019), ACM, Vancouver, Canada, 2019.

[4] J. Sarzynska-Wawer, A. Wawer, A. Pawlak et al., "Detecting formal thought disorder by deep contextualized word representations," *Psychiatry Research*, vol. 304, no. 10, pp. 114135–135, 2021.

[5] T. Brown, B. Mann, N. Ryder, and M. Subbuah, "Language models are few-Shot Learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[6] Y. H Liu, M. Ott, N. Goyal, and J. F. Du, "Roberta: a robustly optimized bert pretraining approach," 2019, https://arxiv.org/abs/1907.11692.

[7] Z. Yang, Z. Dai, Y. Yang, and J. Carbonell, "Generalized autoregressive pretraining for language understanding," in *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, pp. 5753–5763, ACM, Vancouver, Canada, December 2019.

[8] M. Joshi, D. Q. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, "Spanbert: improving pre-training by representing and predicting spans," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 64–77, 2020.

[9] Y. Sun, S. H. Wang, Y. K. Li, and S. K. Feng, "Ernie: enhanced representation through knowledge integration," 2019, https://arxiv.org/pdf/1904.09223.pdf.

[10] T. Mihaylov and A. Frank, "Knowledgeable reader: Enhancing cloze-style reading comprehension with external commonsense knowledge," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 821–832, Association for Computational Linguistics, Melbourne, Australia, July 2018.

[11] K. L. Zhang, X. Zhao, L. Zhuang, and Q. Xie, "Knowledge-enabled diagnosis assistant based on Obstetric EMRs and knowledge graph," in *Proceedings of the 19th Chinese National Conference on Computational Linguistics*, pp. 1155–1165, Chinese National Conference on Computational Linguistics, Haikou, China, October 2020.

[12] Q. Chen, X. D. Zhu, Z. H. Ling, and D. Inkpen, "Neural natural language inference models enhanced with external knowledge," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pp. 2406–2417, Association for Computational Linguistics, Melbourne, Australia, July 2017.

[13] X. Han, Z. Y. Liu, and M. S. Sun, "Neural knowledge acquisition via mutual attention between knowledge graph and text," in *Proceedings of the Thirty-second AAAI conference on artificial intelligence*, pp. 4832–4839, AAAI, April 2018.

[14] A. Saxena, A. Tripathi, and P. Talukdar, "Improving multi-hop question answering over knowledge graphs using knowledge base embeddings," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4498–4507, Association for Computational Linguistics, July 2020.

[15] Z. Wang, J. W. Zhang, J. L. Feng, and Z. Chen, "Knowledge graph and text jointly embedding," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1591–1601, Association for Computational Linguistics, Doha, Qatar, October 2014.

[16] K. Toutanova, D. Q. Chen, P. Pantel, and H. Poon, "Representing text for joint embedding of text and knowledge bases," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 1499–1509, Association for Computational Linguistics, Lisbon, Portugal, September 2015.

[17] X. Han, Z. Y. Liu, and M. S. Sun, "Joint representation learning of text and knowledge for knowledge graph completion," 2016, https://arxiv.org/pdf/1611.04125.pdf.

[18] Y. X. Cao, L. Hou, J. Z. Li, and Z. Y. Liu, "Joint representation learning of cross-lingual words and entities via attentive distant supervision," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 227–237, Brussels, Belgium, November 2018.

[19] J. D. Chen, Y. Z. Hu, J. P. Liu, and Y. H. Xiao, "Deep short text classification with knowledge powered attention," in *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, pp. 6252–6259, AAAI, January 2019.

[20] Z. Y. Zhang, X. Han, Z. Y. Liu, and X. Jiang, "ERNIE: enhanced language representation with informative entities," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1441–1451, Association for Computational Linguistics, Florence, Italy, July 2019.

[21] X. Z. Zhou, S. B. Chen, B. Y. Liu et al., "Development of traditional Chinese medicine clinical data warehouse for medical knowledge discovery and decision support," *Artificial Intelligence in Medicine*, vol. 48, no. 2-3, pp. 139–152, 2010.

[22] B. Y. Liu, X. Z. Zhou, Y. H. Wang et al., "Data processing and analysis in real-world traditional Chinese medicine clinical data: challenges and approaches," *Statistics in Medicine*, vol. 31, no. 7, pp. 653–660, 2012.

[23] S. S. Fang, L. Dong, L. Liu et al., "HERB: a high-throughputexperiment-andreference-guided database of traditional Chinese medicine," *Nucleic Acids Research*, vol. 49, no. D1, pp. 1197–1206, 2021.

[24] M. Casal-Guisande, A. Comesaña-Campos, A. Pereira, J. B. Bouza-Rodríguez, and J. Cerqueiro-Pequeño, "A decision-making methodology based on expert systems applied to machining tools condition Monitoring," *Mathematics*, vol. 10, no. 3, p. 520, 2022.

[25] Y. Xie, "A tcm question and answer system based on medical records knowledge graph," in *Proceedings of the 2020 International Conference on Computing and Data Science (CDS)*, pp. 373–376, IEEE, Stanford, CA, USA, August 2020.

[26] J. K. Anastasi, L. M. Currie, and G. H. Kim, "Understanding diagnostic reasoning in TCM practice: tongue diagnosis," *Alternative Therapies in Health & Medicine*, vol. 15, no. 3, pp. 18–28, 2009.

[27] Y. H. Xie, L. Y. Hu, X. X. Chen, J. Feng, and D. Zhang, "Auxiliary diagnosis based on the knowledge graph of tcm syndrome," *Computers, Materials & Continua*, vol. 65, no. 1, pp. 481–494, 2020.

[28] R. Yang, Q. Ye, C. L. Cheng, S. H. Zhang, Y. Lan, and J. Zou, "Decision-making system for the diagnosis of syndrome based on traditional Chinese medicine knowledge graph," *Evidence-based Complementary and Alternative Medicine*, vol. 20229 pages, Article ID 8693937, 2022.

[29] C. Y. Ruan, Y. Wu, Y. Yang, and G. S. Luo, "Semantic-aware graph convolutional networks for clinical auxiliary diagnosis and treatment of traditional Chinese medicine," *IEEE Access*, vol. 9, no. 99, pp. 8797–8807, 2021.

[30] Y. H. Xie, Y. Y. Yan, J. Y. Li, and D. Zhang, "Research on assistant diagnostic method of TCM based on multi classifier integration," *Advances in Intelligent Systems Research*, vol. 153, pp. 371–376, 2017.

[31] T. Yu, J. H. Li, Q. Yu et al., "Knowledge graph for TCM health preservation: design, construction, and applications," *Artificial Intelligence in Medicine*, vol. 77, pp. 48–52, 2017.

[32] S. X. Ji, S. R. Pan, E. Cambria, P. Marttinen, and P. S. Yu, "A survey on knowledge graphs: representation, acquisition, and

applications," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 2, pp. 494–514, 2022.

[33] A. Bordes, N. Usunier, A. G. Duran, and J. Weston, "Translating embeddings for modeling multi-relational data," *Advances in Neural Information Processing Systems*, vol. 2, pp. 2787–2795, 2013.

[34] Y. J. Lu, Q. Liu, D. Dai, and X. Y. Xiao, "Unified structure generation for Universal information extraction," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 5755–5772, Association for Computational Linguistics, Dublin, Ireland, May 2022.

[35] A. Vaswani, N. Shazeer, N. Parmar, and J. Uszkoreit, "Attention is all you need," in *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*, NIPS, Long Beach, CA, USA, 2017.

[36] Z. Q. Liu, H. Y. He, S. X. Yan, Y. Wang, T. Yang, and G. Z. Li, "End-to-End models to imitate traditional Chinese medicine syndrome differentiation in lung cancer diagnosis: model development and Validation," *JMIR Medical Informatics*, vol. 8, no. 6, Article ID e17821, 2020.

[37] Y. Kim, "Convolutional neural network for sentence classification," 2014, https://arxiv.org/pdf/1408.5882.pdf.

[38] P. F. Liu, X. P. Qiu, and X. J. Huang, "Recurrent neural network for text classification with multi-task learning," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pp. 2873–2879, ACM, New York, USA, July 2016.

[39] R. Girshick, J. Donahue, and T. Darrell, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, IEEE, Columbus, OH, USA, September 2014.

[40] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of Tricks for efficient text classification," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 427–431, Association for Computational Linguistics, Valencia, Spain, April 2017.

[41] Q. A. Hu, T. Yu, J. H. Li, Q. Yu, L. Zhu, and Y. Gu, "End-to-end syndrome differentiation of Yin deficiency and Yang deficiency in traditional Chinese medicine," *Computer Methods and Programs in Biomedicine*, vol. 174, pp. 9–15, 2019.

[42] W. P. Yin and H. Schütze, "Attentive convolution: Equipping cnns with rnn-style attention mechanisms," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 687–702, 2018.

[43] R. Johnson and T. Zhang, "Deep Pyramid convolutional neural networks for text categorization," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 562–570, Association for Computational Linguistics, Vancouver, Canada, July 2017.

[44] T. Trouillon, J. Welbl, S. Riedel, and E. Gaussier, "Complex embeddings for simple link prediction," in *Proceedings of the 33 rd International Conference on Machine Learning*, pp. 2071–2080, ACM, New York, NY, USA, June 2016.

[45] B. Yang, W. Yih, X. He, J. F. Gao, and L. Deng, "Embedding entities and relations for learning and inference in knowledge bases," 2014, https://arxiv.org/abs/1412.6575.

[46] D. Q. Nguyen, T. D. Nguyen, D. Q. Nguyen, and D. Phung, "A novel embedding model for knowledge base completion based on convolutional neural network," in *Proceedings of the NAACL-HLT*, pp. 327–333, Association for Computational Linguistics, New Orleans, Louisiana, June 2017.

[47] Y. Q. Zhang, Q. M. Yao, T. X. Shao, and L. Chen, "NSCaching: simple and efficient negative sampling for knowledge graph embedding," in *Proceedings of the 2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pp. 614–625, IEEE, Macao, China, April 2019.