Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 19(10): 10533–10549.
DOI: 10.3934/mbe.2022492
Received: 31 May 2022
Revised: 28 June 2022
Accepted: 04 July 2022
Published: 25 July 2022

Research article

Graph-based structural knowledge-aware network for diagnosis assistant

Kunli Zhang[1,2], **Bin Hu**[1], **Feijie Zhou**[3], **Yu Song**[1,*], **Xu Zhao**[1] **and Xiyang Huang**[1]

[1] School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou, China
[2] Pengcheng Laboratory, Shenzhen, China
[3] The Third Affiliated Hospital of Zhengzhou University, Zhengzhou 450052, China

* **Correspondence:** Email: ieysong@zzu.edu.cn.

Abstract: Diagnosis assistant is an effective way to reduce the workloads of professional doctors. The rich professional knowledge plays a crucial role in diagnosis. Therefore, it is important to introduce the relevant medical knowledge into diagnosis assistant. In this paper, diagnosis assistant is treated as a classification task, and a Graph-based Structural Knowledge-aware Network (GSKN) model is proposed to fuse Electronic Medical Records (EMRs) and medical knowledge graph. Considering that different information in EMRs affects the diagnosis results differently, the information in EMRs is categorized into general information, key information and numerical information, and is introduced to GSKN by adding an enhancement layer to the Bidirectional Encoder Representation from Transformers (BERT) model. The entities in EMRs are recognized, and Graph Convolutional Neural Networks (GCN) is employed to learn deep-level graph structure information and dynamic representation of these entities in the subgraphs. An interactive attention mechanism is utilized to fuse the enhanced textual representation and the deep representation of these subgraphs. Experimental results on Chinese Obstetric Electronic Medical Records (COEMRs) and open dataset C-EMRs demonstrate the effectiveness of our model.

Keywords: electronic medical records; diagnosis assistant; knowledge graph; multi-label classification

1. Introduction

Diagnosis assistant aims to use patients' Electronic Medical Records (EMRs) to aid physicians in diagnosis. Applying artificial intelligence to diagnosis assistant is one of the most effective ways to alleviate the problem of insufficient medical resources [1]. Based on the Basic Specification of Electronic Medical Records (trial) [2], various medical institutions have accumulated massive EMRs resources. A large number of EMRs provide data support for diagnosis assistant. Diagnosis assistant takes the patients' EMRs as input and the diagnosis as output, which is often regarded as a classification task.

Research on diagnosis assistant can be traced back to expert systems in the 1950s. Specifically, the knowledge base was constructed by using the experience and knowledge of experts in the relevant fields, then the computer summarized and studied knowledge of the field through certain rules, and gave the corresponding diagnosis. Ledley et al. [3] applied the data model to clinical medicine for the first time, laying a foundation for subsequent diagnosis assistant. ShortLiffe et al. [4] adopted flexible knowledge representation to develop an antibacterial treatment information system which could carry out interactive consultation with doctors. Mekruksavanich et al. [5] constructed a web-based diabetes classification system for early diagnosis of diabetes by using nine subcategories of diabetes. Expert system, as an early research of artificial intelligence, requires little data and has a strong interpretability, which provides a basis for the subsequent development. However, the performance of expert system depends largely on the quality of the dataset and the standards established by experts.

Since the 2010s, Baati et al. [6] used Naive Bayes as a classifier to design a diagnostic system for lymphatic duct disease based on the UCI dataset. Calisir et al. [7] proposed a system for hepatitis diagnosis using a combination of feature extraction and classifier. The system first extracted features from the hepatitis database and performed approximate simplification, then used a classifier for diagnosis. Otoom et al. [8] used wearable sensors to monitor the real-time status of cardiac patients, then predicted and alerted on heart disease by classifiers. The application of machine learning has greatly promoted the application of artificial intelligence in diagnosis assistant. However, these methods cannot mine the deep information in EMRs and are difficult to deal with multi-label tasks.

With the popularity of deep learning, Liang et al. [9] used semantic information extracted from EMRs by Convolutional Neural Networks (CNN) to diagnose liver cancer. Kim et al. [10] developed a two-stage hierarchical prediction model based on Recurrent Neural Network (RNN) for continuous prediction of acute kidney injury by combining information from patients' EMRs. Du et al. [11] utilized an end-to-end Long Short-Term Memory (LSTM) using a multigraph structure to predict foodborne diseases, which can integrate multiple external factors and make accurate predictions considering the spatio-temporal characteristics of foodborne disease data. Sedik et al. [12] proposed a COVID-19 detection system based on CNN and Convolutional Long Short-Term Memory (ConvLSTM) , which diagnosed COVID-19 based on the patient's X-ray and CT images. In addition to EMRs, professional knowledge is also essential to diagnosis. The above methods do not introduce external medical knowledge into diagnosis assistant.

Knowledge Graph (KG) has developed rapidly in recent years and has been widely used in various fields of artificial intelligence, such as Question and Answer(QA), knowledge response, etc. Compared with traditional knowledge representation, KG has a broader coverage which can represent diverse semantic information and replicate the domain knowledge and clinical experience of medical experts quickly. Integrating KG into diagnosis assistant may significantly improve the accuracy of diagnosis. Zhang et al. [13] proposed a Knowledge-Enabled Diagnosis Assistant (KEDA) model for integrating Bidirectional Encoder Representation from Transformers(BERT) [14] model into external medical KG. They integrated embedding of KG into BERT through TransE [15] model, making KEDA to learn medical domain knowledge. However, KEDA only introduces the triples information of KG, without considering the graph structure information of KG. And the graph embedding obtained by TransE model was static, which cannot deal with 1-to-n, n-to-1 and n-to-n relationships.

In order to solve the above problems, this paper proposes Graph-Based Structural Knowledge-aware Network (GSKN) for diagnostic assistant. GSKN identifies entities in EMRs by a similarity-based approach and extracts these entity-related triples from the knowledge graph to construct subgraphs, then initializes the subgraph nodes using TransE. In order to address the shortcomings of TransE and introduce the graph structure information of KG, GSKN uses Graph Convolutional Neural Network (GCN) [16] to obtain the deep graph structure information and the dynamic representation of the nodes in these subgraphs. Unlike common classification tasks, the information in EMRs is extremely complex and the importance of different information to the diagnostic outcome varies. Considering the multiple presentations of EMRs and the degree of influence on diagnosis, the information in EMRs is divided into general information, key information and numerical information. For general information, BERT [14] model is employed to obtain its textual representation. Numerical information in EMRs also plays a crucial role in diagnosis. However, BERT model performs poorly in learning numerical information, so the numerical information is embedded in the textual representation of EMRs through a multi-head self-attentive mechanism [17]. Key information refers to the most salient pain or obvious symptoms of the patients, which is also the main basis for diagnosis. Inspired by Qu et al. [18], Key-Info embedding containing the key information is added to provide information enhancement for the textual representation of EMRs. A sentence that introduces too much outside knowledge may deviate from its original meaning, which is named knowledge noise (KN) problem. GSKN uses interactive attention to deeply integrate the deep representation of subgraphs with the enhanced text representation to solve the KN problem. Specifically, GSKN uses Knowledge-Knowledge Attention to evaluate the relative importance of knowledge components, then uses Knowledge-Text Attention to measure the semantic relevance between knowledge and text representation to solve the KN problem. Experimental results on Chinese Obstetric Electronic Medical Records (COEMRs) and open dataset C-EMRs [19] demonstrate the validity of our model. The main contributions of this paper are as follows:

1) Given that traditional sequence learning models have difficulty in handling graph structure data, we design a GSKN model to effectively integrate graph structure information.

2) Given that BERT model cannot distinguish different information in EMRs and cannot learn numerical information effectively, we add an enhancement layer for BERT to pay attention to the different effects of various information on the final results.

3) Given the problem of knowledge noise caused by the introduction of external knowledge, we propose an interactive attention mechanism to integrate text and knowledge. Subsequent experiments have demonstrated the effectiveness of the interactive attention mechanism.

2. Methods

2.1. Model architecture

GSKN can be divided into three parts: EMRs Modeling, Knowledge Modeling, and Attention Layer. The GSKN model diagram is shown in Figure 1. The left part of Figure 1 is EMR Modeling, which obtains the enhanced text representation of EMRs. The right part of Figure 1 is Knowledge Modeling, which acquires deep representation of subgraphs and dynamic representation of nodes. GSKN fuses the knowledge representation and text representation through Attention Layer to get the final output. The input of GSKN is all information in EMRs, and the output is the probability distribution of labels.

2.2. *EMRs modeling*

An example of COEMRs is shown in Figure 2. We take the original description of EMRs as general information, and extract key information and numerical information from it. Numerical information refers to some checks or indications expressed by numerical form, such as “P:84/min, R:21/min” in Figure 2. Key information refers to the main pain or the most obvious symptoms and (or) signs experienced by the patient, as well as the main reason for his/her treatment and the duration of the disease, such as “menopause for more than 8 months, vaginal bleeding for 14 hours” in Figure 2. Key information in EMRs is usually in a particularly refined language (generally about 20 words) to explain the core symptoms of admission, which is the main basis for diagnosis. For general information, BERT model is used as encoder to generate text representation, and the input text sequence is shown in Figure 3, where [CLS] is a specific classifier token and [SEP] is a sentence separator which is defined in BERT.

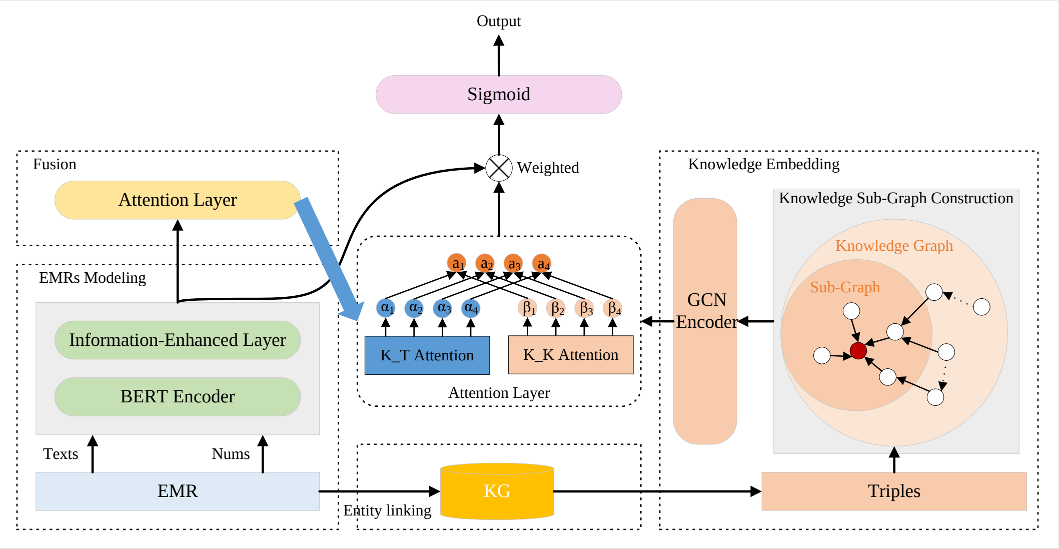


Figure 1. The architecture of the GSKN model.

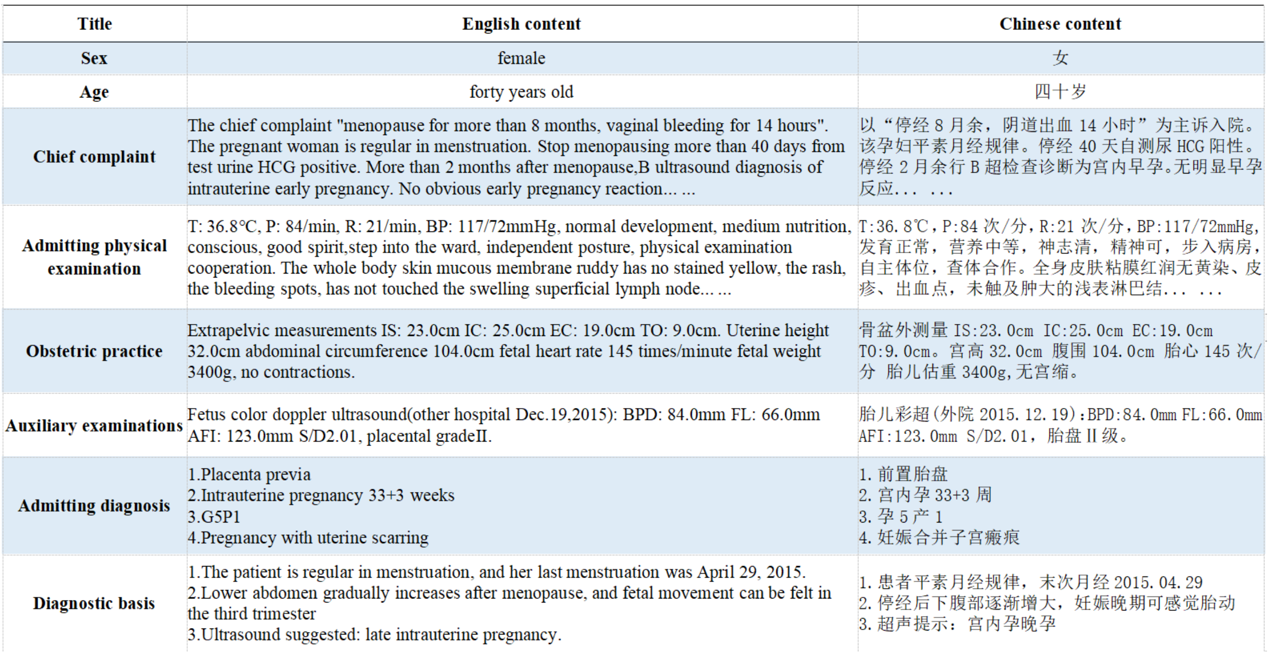
2.2.1. Numerical information

Numerical information in EMRs has a significant effect on diagnosis. For example, the older a pregnant woman is, the more risk factors she is exposed to and the more specific treatment should be available. At the same time, extracting the numerical information separately can meet BERT’s length limit of the input text sequence. GSKN uses a rule-based approach to extract 18 types of important numerical information such as blood pressure, age, body temperature, heart rate and so on. Numerical information needs to be standardized and normalized before it can be used as input because of the inconsistency of units and the problems of missing and error. The min-max normalization and zero-mean normalization are used in this step, as shown in Eqs (2.1) and (2.2):

$$y_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{2.1}$$

$$y_i' = \frac{x_i - \bar{x}}{S} \tag{2.2}$$

where x is the unprocessed numerical information, y_i is the intermediate result, y_i' is the final result of the completed processing, and S is the standard deviation.



Title	English content	Chinese content
Sex	female	女
Age	forty years old	四十岁
Chief complaint	The chief complaint "menopause for more than 8 months, vaginal bleeding for 14 hours". The pregnant woman is regular in menstruation. Stop menopausing more than 40 days from test urine HCG positive. More than 2 months after menopause,B ultrasound diagnosis of intrauterine early pregnancy. No obvious early pregnancy reaction... ...	以“停经 8 月余，阴道出血 14 小时”为主诉入院。该孕妇平素月经规律。停经 40 天自测尿 HCG 阳性。停经 2 月余行 B 超检查诊断为宫内早孕。无明显早孕反应... ...
Admitting physical examination	T: 36.8°C, P: 84/min, R: 21/min, BP: 117/72mmHg, normal development, medium nutrition, conscious, good spirit,step into the ward, independent posture, physical examination cooperation. The whole body skin mucous membrane ruddy has no stained yellow, the rash, the bleeding spots, has not touched the swelling superficial lymph node... ...	T:36.8℃，P:84 次/分，R:21 次/分，BP:117/72mmHg，发育正常，营养中等，神志清，精神可，步入病房，自主体位，查体合作。全身皮肤粘膜红润无黄染、皮疹、出血点，未触及肿大的浅表淋巴结... ...
Obstetric practice	Extrapelvic measurements IS: 23.0cm IC: 25.0cm EC: 19.0cm TO: 9.0cm. Uterine height 32.0cm abdominal circumference 104.0cm fetal heart rate 145 times/minute fetal weight 3400g, no contractions.	骨盆外测量 IS:23.0cm IC:25.0cm EC:19.0cm TO:9.0cm。宫高 32.0cm 腹围 104.0cm 胎心 145 次/分 胎儿估重 3400g，无宫缩。
Auxiliary examinations	Fetus color doppler ultrasound(other hospital Dec.19,2015): BPD: 84.0mm FL: 66.0mm AFI: 123.0mm S/D2.01, placental gradeII.	胎儿彩超（外院 2015.12.19）:BPD:84.0mm FL:66.0mm AFI:123.0mm S/D2.01，胎盘 II 级。
Admitting diagnosis	1.Placenta previa 2.Intrauterine pregnancy 33+3 weeks 3.G5P1 4.Pregnancy with uterine scarring	1. 前置胎盘 2. 宫内孕 33+3 周 3. 孕 5 产 1 4. 妊娠合并子宫瘢痕
Diagnostic basis	1.The patient is regular in menstruation, and her last menstruation was April 29, 2015. 2.Lower abdomen gradually increases after menopause, and fetal movement can be felt in the third trimester 3.Ultrasound suggested: late intrauterine pregnancy.	1. 患者平素月经规律，末次月经 2015.04.29 2. 停经后下腹部逐渐增大，妊娠晚期可感觉胎动 3. 超声提示：宫内孕晚孕

Figure 2. An example of COEMRs.

For the missing numerical information, the mean value of similar information is used as supplement. For the error data that obviously exceed the index value range, delete the error data first, then use the mean value to complete the error data. Numerical information processed by the above process can be input into the model as numerical features, and this process can be regarded as establishing the dependency relationship between numerical information and textual information. In GSKN, the multi-head self-attention is used to fuse them. The specific calculation processes are shown in Eqs (2.3) –(2.6):

$$Q = K = V = W^S Concat([C']; Num_{1...M}) \tag{2.3}$$

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.4}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{2.5}$$

$$[C'] = Concat(head_1, ..., head_h)W^O \tag{2.6}$$

where $[C]$ is the hidden layer state representation of [CLS], $[C']$ is the text representation after fusing numerical information. $Num_{1...M}$ is the Nums embedding containing M values, which is obtained by standardizing and normalizing the numerical information in EMRs. W^S, W^Q, W^K, W^V and W^O are trainable parameters.

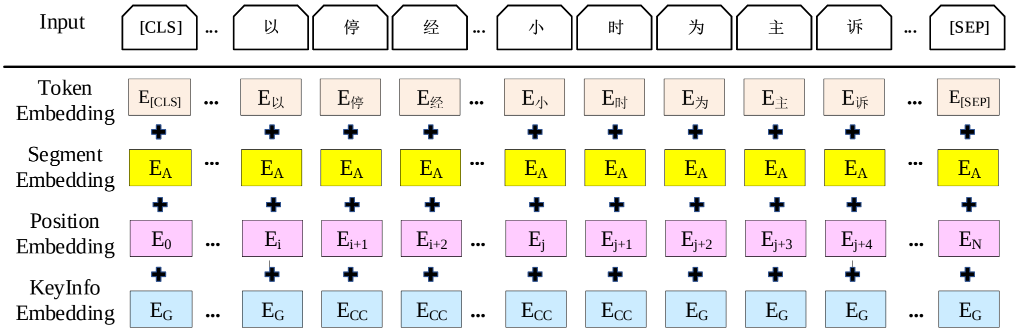


Figure 3. The input embedding of GSKN.

2.2.2. Key information

Key information refers to the main pain or the most obvious symptoms and (or) signs experienced by the patient, as well as the duration of the disease. For example, the key information in one EMRs is that "It is found that S/D value is high for 2 hours after menses stops for more than 7 months". The corresponding diagnosis is "fetal distress ...". The S/D value represents the resistance to the umbilical cord blood flow, and high resistance is a signal of fetal distress. Key information is more critical for the diagnosis than other texts in EMRs. We use a rules-based approach to extract key information from EMRs. Inspired by Qu et al. [18], KeyInfo embedding is added into the input of BERT. As shown in Figure 3, characters corresponding to key information are marked as E_{CC} in the KeyInfo embedding, and the remaining characters are indicated by E_G.

2.3. Knowledge modeling

Knowledge modeling consists of two parts: Sub-graphs Construction and Sub-graphs Encoding. In this module, the Chinese Obstetric Knowledge Graph (COKG) [23] and Chinese Medical Knowledge Graph(CMeKG) [24] are regarded as the source of external knowledge.

2.3.1. Sub-graphs construction

We use ICD-10 (International Classification of Diseases) [*], CSKB (Chinese Symptom Knowledge Base) [25] and knowledge graph to construct the corresponding disease and symptom dictionary. A rules-based approach is used to identify diseases and symptoms in EMRs. Then a similarity-based approach is used to link entities to triples in the knowledge graph. To be specific, the similarity of each

[*] https://icd.who.int/browse10/2019/en

entity in EMRs and all entities in dictionary is calculated by Levenshtein Distance, Jaccard Similarity Coefficient and Longest Common Substring. The mean value is taken as the final score, and the triple with the highest score is taken as the link result. Levenshtein Distance, also known as Edit Distance, is the minimum number of editing operations required to convert from one to the other between two strings. The permitted editing operations are substitution, insertion, and deletion. And the calculation formula is shown in Eq (2.7). Jaccard Similarity Coefficient, also known as Jaccard Index, is used to compare the similarity and difference between finite sample sets and is shown in Eq (2.8). Longest Common Substring refers to the longest common subsequence between two strings and is shown in Eq (2.9). The three scores are averaged to get the final score:

$$score_L = 1 - \frac{Levenshtein(e_i, k_j)}{max(len(e_i), len(k_j))} \tag{2.7}$$

$$score_J = jaccard(bigram(e_i), bigram(k_j)) = \frac{\left|bigram(e_i) \cap bigram(k_j)\right|}{\left|bigram(e_i) \cup bigram(k_j)\right|} \tag{2.8}$$

$$score_l = \frac{\left|longestcommonsubstring(e_i, k_j)\right|}{max(len(e_i), len(k_j))} \tag{2.9}$$

$$score_{final} = Average(score_L, score_J, score_l) \tag{2.10}$$

where e_i is the i-th possible entity in the EMRs and k_j is the j-th entity in the knowledge graph K.

The triples associated with these disease and symptom entities obtained through entity linking are extracted from the knowledge graph to construct subgraphs. Finally, these triples are rebuilt as subgraphs by joining the same entities and keeping the relationships as edges.

2.3.2. Sub-graphs encoding

Due to its simplicity and efficiency, TransE [15] is used to initialize the nodes of the subgraph. TransE is based on the translation-invariant phenomenon in the word vector space, where the vector corresponding to the relation r is viewed as a translation between the entity h and the entity t vector, which is $l_h + l_r = l_t$. The embedding obtained by TransE is static, which cannot deal with 1-to-n, n-to-1 and n-to-n relationships. GSKN obtain information from neighboring nodes and performs convolutional operations on them to update the embedding of the specified nodes (the disease or symptom entities in EMRs) by GCN. Let D_i and S_j denote the embedding representation of disease d_i and symptom s_j. The updated rules are shown in Eqs (2.11) and (2.12):

$$\hat{D}_i = ReLU(W_1 D_i + \sum_{u \epsilon N_{p_i}} \frac{W_2 D_u}{\left|N_{p_i}\right|} + \sum_{v \epsilon N_{c_i}} \frac{W_3 D_v}{\left|N_{c_i}\right|} + b_1) \tag{2.11}$$

$$\hat{S}_j = ReLU(W_4 S_j + \sum_{m \epsilon N_{p_j}} \frac{W_5 S_m}{\left|N_{p_j}\right|} + \sum_{k \epsilon N_{c_j}} \frac{W_6 S_k}{\left|N_{c_j}\right|} + b_2) \tag{2.12}$$

where W_1, W_2, W_3, W_4, W_5, W_6, b_1 and b_2 are trainable parameters. N_{p_i} is the set of parent and child nodes of node d_i, and N_{p_j} is the set of parent and child nodes of node s_j.

The embedding of disease d_i and symptom s_i are updated in training based on the representations of their parent nodes and child nodes through the above rules. Finally, these final representations obtained by two layers of GCN are read out using mean_node to obtain the final knowledge representation.

2.4. Attention layer

To make more effective utilization of knowledge representation, we propose an interactive attention-based mechanism. The interactive attention includes Knowledge-Text Attention (K-T), Knowledge-Knowledge Attention (K-K) and gating mechanism.

2.4.1. Knowledge-text attention

In order to reduce the ambiguity or noise caused by some unnecessary entities in knowledge graph, Knowledge-Text Attention is proposed to measure the semantic relevance between knowledge and text representation. The calculation process is shown in Eq (2.13):

$$\alpha_i = softmax(W_1^T tanh(W_2([C^{'}]; K_i) + b_1)) \tag{2.13}$$

where $[C^{'}]$ is the text representation, and K_i is the deep knowledge representation obtained by Knowledge Modeling. W_1^T and W_2 are trainable parameters, and b_1 is the offset. The weight α_i is positively correlated with the semantic similarity between corresponding knowledge and text.

2.4.2. Knowledge-knowledge attention

Knowledge is different, and the importance of each part of knowledge should be different. In order to evaluate the relative importance of each part of the knowledge representation, Knowledge-Knowledge Attention based on source2token self-attention is proposed to effectively measure its relative importance. The calculation process is shown in Eq (2.14):

$$\beta_i = softmax(W_3^T tanh(W_4 + b_2)) \tag{2.14}$$

where β_i represents the attention weight of the i-th knowledge relative to all knowledge in the knowledge set. W_3^T, W_4 and b_2 are trainable parameters.

Knowledge-Knowledge Attention is similar to feature selection, which is equivalent to a soft feature selector and can give greater weight to more important knowledge. For other unimportant knowledge, its weight tends to zero.

2.4.3. Gating mechanism

Gating Mechanism is employed to fuse the results of the above two-step calculations. For diagnosis assistant, external knowledge is not always necessary for the text, and too much knowledge may make noise. The specific calculation processes are shown in Eqs (2.15) – (2.17):

$$w_i = \sigma(W_5(\alpha_i; \beta_i) + b_3) \tag{2.15}$$

$$a_i = w_i \odot \alpha_i + (1 - w_i) \odot \beta_i \tag{2.16}$$

$$[C''] = \sum_{i=1}^{N} a_i [C']_i \tag{2.17}$$

where σ is a sigmoid function, W_5 is a trainable parameter, b_3 is an offset, and $[C'']$ is the final representation after fusion of knowledge. w_i is used to balance the relative importance between α_i and β_i. When w_i approaches 1, the introduced knowledge is the least, and when it approaches 0, the introduced knowledge will increase accordingly.

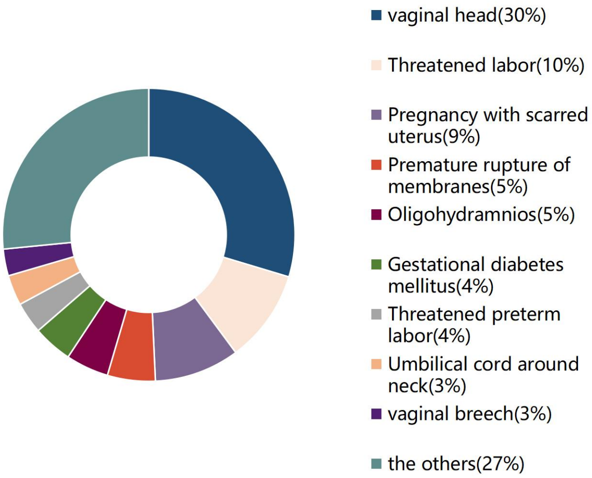


Figure 4. The labels distribution of Chinese Obsteric EMRs.

3. Experiments

3.1. Datasets

We used Chinese Obstetric Knowledge Graph (COKG) to conduct experiments on the COEMRs and Chinese Medical Knowledge Graph (CMeKG) to conduct experiments on C-EMRs.

3.1.1. Chinese obstetric electronic medical records

COEMRs includes 24,339 EMRs from multiple hospitals in China. They are pre-processed through the steps of anonymization, data cleaning, structuring, and diagnostic label standardization. Of these, 21,905 are used for training and 2434 are used for testing. The final number of diagnostic labels used is 180. The labels distribution is shown in Figure 4.

3.1.2. C-EMRs

C-EMRs comes from Yang et al. [19], which contains 18,590 medical records. Of these, 16,731 are used as the training set and 1859 are used as the test set. The diagnosis results are in a single-label format with 10 different categories, which are different from the COEMRs. The specific labels distribution is shown in Figure 5.

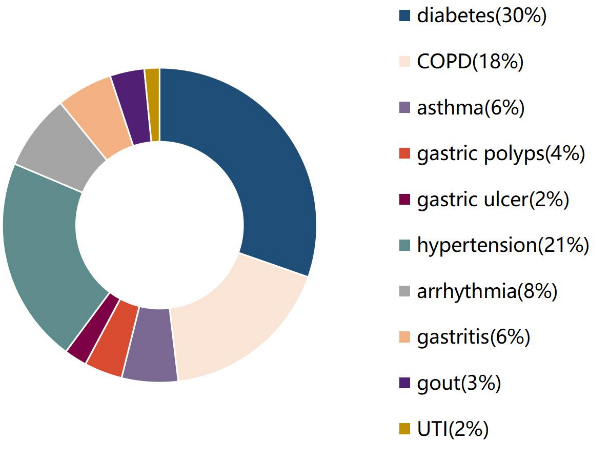


Figure 5. The labels distribution of C-EMRs.

3.1.3. Chinese obstetric knowledge graph

COKG uses the MeSH-like framework as the knowledge ontology to define the entity and relationship description system with obstetric diseases as the core. It contains knowledge from various sources such as the professional thesaurus, obstetrics textbooks, clinical guidelines, network resources, and other multi-source knowledge. COKG includes a total of 15,249 types of relations and 10,674 types of entities.

3.1.4. Chinese medical knowledge graph

CMeKG refers to a variety of international medical standards and clinical guidelines such as ICD, ATC and MeSH, as well as a variety of medical texts such as medical encyclopedia. It includes 11,076 types of diseases, 18,471 types of drugs, 14,794 types of symptoms and 1,566,494 types of relations.

3.2. Experimental setup

In this paper, EMRs are preprocessed by means of privacy removal, data cleaning, structuring, data filtering and standardization of diagnostic labels. During the data filtering process, duplicate or barely diagnostic information is removed. On the one hand, it can satisfy the limit of input length of BERT model. On the other hand, it can also retain useful information. The BERT model version is Bert-BASE-Chinese, and the main parameters are as follows: hidden size 768, max position embedding 512, num attention heads 12, num hidden layers 12, maximum input length 512, learning rate 5e-5, batch size 6, training epoch 20. All of our experiments were run on RTX2080ti GPU(12G).

3.3. Baselines

The experimental comparison model is as follows:

- TextCNN [20]: The classical text classification model adopts the CNN based on pre-trained word embedding.

- TextRCNN [21]: RNN in this model is used to obtain the text context information, and CNN is used to obtain the main components.
- TextRNN+Att [22]: Attention mechanism is introduced into TextRNN model to obtain important information in sentences.
- Transformer [17]: Transformer abandons the traditional CNN and RNN, and the network structure is completely composed of attention mechanism.
- BERT [14]: BERT is a pre-trained language representation model based on Transformer's bidirectional encoder representation. It used the Masked Language model (MLM) to generate a deep bidirectional language representation.

3.4. Results

3.4.1. Results on Chinese obstetric EMRs

Experimental results on COEMRs are shown in Table 1. The BERT model, which is pre-trained on a large scale of corpus, shows its strong adaptability, and only requires fine-tuning on the target corpus to achieve better effect. Although the traditional models based on CNN and RNN are inferior to BERT model in effect, they are more lightweight and have fewer parameters, and are suitable for fast training iterative tasks.

Table 1. The results on COEMRs.

Model	F1 (%)	P (%)	R (%)	HL (%)
TextCNN	76.88	87.95	68.28	0.0058
TextRNN	77.16	83.81	71.49	0.0059
TextRNN+Att	74.51	79.48	70.12	0.0243
Transformer	76.72	80.65	73.17	0.0198
BERT	79.74	80.63	78.87	0.0051
KEDA	81.11	83.22	79.12	0.0056
GSKN	81.42	83.56	79.35	0.0051

From the perspective of a series of BERT-based models, the KEDA model with external knowledge is better than the simple BERT model. After introducing the graph structure information contained in EMRs, GSKN improves by 1.68% compared with BERT model and 0.31% compared with KEDA model. Compared with the traditional sequence learning model, GSKN is more sensitive and effective in learning graph structure information. From the dimension of knowledge, the information of graph structure type is deeper than the triplet information introduced by KEDA model, and the information in the graph is also constantly improved in the process of training.

According to the occurrence frequency of each label in COEMRs, we divided the total labels into 8 groups, as shown in Figure 6, based on the number of occurrences of each label in the data. As can be seen from Figure 6, GSKN achieves higher performance than BERT and KEDA in all groups except 100–300 group, which also proves the feasibility of introducing deep graph structure information of knowledge graph. In 100–300 group, GSKN is slightly lower than KEDA, but higher than the BERT model without the introduction of external knowledge, demonstrating the feasibility of introducing medical knowledge for diagnosis assistant.

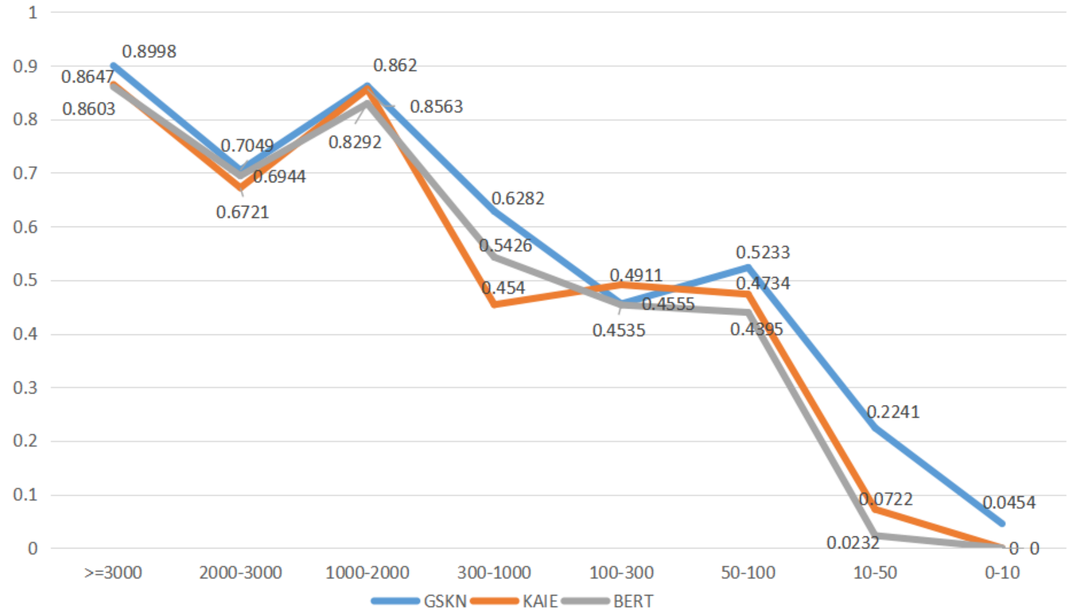


Figure 6. F1 values for each group of labels in GSKN, KEDA and BERT.

3.4.2. Results on C-EMRs

To further verify the validity of GSKN, a comparative experiment was carried out on the public dataset C-EMRs. Due to a large number of missing numerical information in this dataset, numerical information was not introduced into the model. Since the disease data in C-EMRs are different from COEMRs, it is not appropriate to use COKG as the source of knowledge, which focuses on obstetric diseases and symptoms. Therefore, CMeKG, which is more versatile, is further used as the source of external knowledge in this experiment. The experimental results are shown in Table 2. It can be seen from the data in the table that KEDA and GSKN integrating external knowledge graph information have better results on public dataset C-EMRs than traditional deep learning model and single BERT model. And GSKN with graph structure information has the best effect, reaching 82.01%. It further verifies the effectiveness of the corresponding fusion knowledge graph method proposed in this paper and proves that it is feasible to introduce external knowledge into diagnosis assistant.

Table 2. The results on C-EMRs.

Model	F1 (%)	P (%)	R (%)	HL (%)
TextCNN	87.75	92.31	83.32	0.0215
TextRNN	90.48	91.80	89.13	0.0171
TextRNN+Att	89.05	89.55	88.54	0.0198
Transformer	78.92	83.02	75.20	0.0365
BERT	91.57	92.12	91.03	0.0056
KEDA	91.95	92.04	91.87	0.0056
GSKN	92.02	91.88	92.16	0.0051

3.5. Ablation experiments

In order to verify the effectiveness of the interactive attention and information enhancement proposed in this paper, the following comparative experiments were carried out. Table 3 shows the results of removing information enhancement. "-KeyInfo" means to remove key information embedding, "-Num" means to remove numerical information. Key information can enhance the influence of chief complaint in diagnosis, and the model cannot distinguish chief complaint information well after removing key information embedding. Numerical information is a feature that cannot be accurately captured in BERT. Numerical information has a close influence in diagnosis.

Table 3. Ablation experiment of information enhancement layer.

model	F1 (%)	P (%)	R (%)	HL (%)
GSKN	81.42	83.56	79.35	0.0051
-KeyInfo	80.92	85.02	77.20	0.0051
-Num	80.32	84.47	76.56	0.0051
-KeyInfo&Num	80.06	84.51	76.06	0.0051

Table 4 shows the results of removing interactive attention, where "-K-T" means removing the Knowledge-Text Attention, "-K-K" means removing the Knowledge-Knowledge Attention and "-K-T&-K-K" means removing interactive attention. As can be seen from the experimental results in the table, the overall effect of the model is lower than that of GSKN, whether the interactive attention is removed completely or partially. In terms of the impact of removal on the model, it is obvious that complete removal has the greatest impact on the model performance. The model that completely removes the interactive attention is difficult to effectively screen the acquired knowledge, and may introduce more noise information, which intuitively shows the decrease of model effect. As can be seen from the Table 4, when the attention mechanism is completely removed, the performance of GSKN is even worse than that of BERT model without introducing external knowledge. It also proves that the interactive attention mechanism proposed in this paper can effectively solve the problem of knowledge noise. And K-K attention is designed to screen more effective knowledge, so the result after removing this part is also decreased.

Table 4. Ablation experiment of interactive attention.

model	F1 (%)	P (%)	R (%)	HL (%)
GSKN	81.42	83.56	79.35	0.0051
-K-T	80.75	82.12	79.43	0.0056
-K-K	80.63	82.35	78.98	0.0056
-K-K&K-T	79.44	80.57	78.36	0.0128
BERT	79.74	80.63	78.87	0.0051

3.5.1. Overall analysis

We analyzed the reasons leading to the deterioration of the model effect, and the specific reasons are as follows:

Both COEMRs and C-EMRs data are extremely unbalanced. In C-EMRs, “diabetes”, “hypertension” and “chronic obstructive pulmonary disease” account for the largest proportion, accounting for 30.3, 21.2 and 17.8% respectively, while “urinary tract infection” accounts for only 1.6%. We conducted further experiments on COEMRs to verify the effect of unbalanced data on the model. According to the frequency of each label appearing in the data, we divided the total labels into four groups: 0–10, 10–50, 50–300 and 300+. The number of labels in each group is 84, 43, 30, and 23, respectively. The result is shown in Figure 7. In the group of 0–10, although its P value reaches 1.0, its R value is only 0.0232. From the table, it can be seen that the higher the frequency of label occurrence, the higher the F1 value predicted by the model, and the smaller the gap between P value and R value. The more data the label corresponds to, the better the model learns.

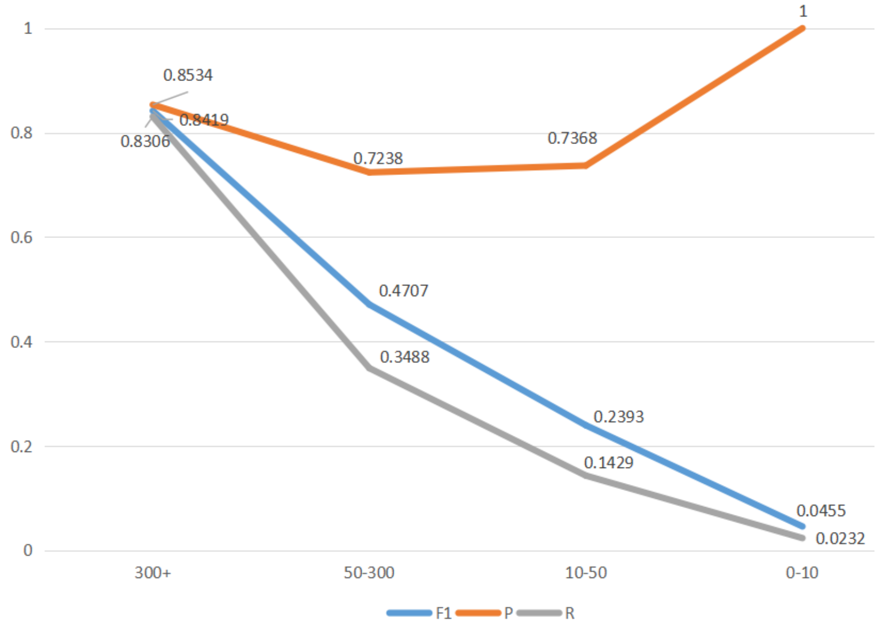


Figure 7. Results of each labels group.

There are different levels of description for different diseases in the knowledge graph. In COKG, there are 121 relations for group B streptococcal infection and only 1 relation for pregnancy with hypothyroidism. There is no doubt that the external knowledge introduced by the model for group B streptococcal infection is significantly better than that of pregnancy with hypothyroidism.

4. Conclusions

In this paper, diagnosis assistant is treated as a multi-label classification problem. We propose a GSKN model integrating external knowledge. GSKN utilizes the numerical information of EMRs, key information and sub-graphs constructed from external knowledge to enhance the textual representation of EMRs. We use the similarity method to obtain the entity set in EMRs and construct the sub-graph of knowledge graph. The deep representation of sub-graphs is obtained through Graph Convolutional Neural Network, and the deep fusion with text is carried out through interactive attention mechanism. Finally, the effectiveness of our method is verified with experiments on Chinese Obstetric EMRs and C-EMRs.

Acknowledgments

We thank the anonymous reviewers for their constructive comments, and gratefully acknowledge the support of Major Science and Technology Project of Yunnan Province (202102AA100021), Zhengzhou City Collaborative Innovation Major Projects (20XTZX11020), National Key Research and Development Program (2017YFB1002101), National Natural Science Foundation of China (62006211),Henan Science and Technology Research Project (192102210260), Henan Medicine Science and Technology Research Plan: Provincial and Ministry Co-construction Project (SB201901021), Henan Provincial Key Scientific Research Project of Colleges and Universities (19A520003, 20A520038), The MOE Layout Foundation of Humanities and Social Sciences (Grant No. 20YJA740033), Henan Social Science Planning Project (Grant No. 2019BYY016).

Conflict of interest

The authors declare there is no conflict of interest.

References

1. National Health and Family Planning Commission of the P. R. C., Guiding opinions of the General Office of the State Council on promoting the construction and development of medical consortium, *Bulletin of The State Council of the People's Republic of China*, 2017.
2. China's Ministry of Health, Basic specification of electronic medical records (trial), *Chin. Med. Rec.*, **11** (2010), 64–65.
3. R. S. Ledley, L. B. Lusted, Reasoning foundations of medical diagnosis: Symbolic logic, probability, and value theory aid our understanding of how physicians reason, *Science*, **130** (1959), 9–21. https://doi.org/10.1126/science.130.3366.9
4. E. H. Shortliffe, S. G. Axline, B. G. Buchanan, T. C. Merigan, S. N. Cohen, An artificial intelligence program to advise physicians regarding antimicrobial therapy, *Comput. Biomed. Res.*, **6** (1973), 544–560. https://doi.org/10.1016/0010-4809(73)90029-3
5. S. Mekruksavanich, Medical expert system based ontology for diabetes disease diagnosis, in *2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, (2016), 383–389. https://doi.org/10.1109/ICSESS.2016.7883091
6. K. Baati, T. M. Hamdani, A. M. Alimi, Diagnosis of lymphatic diseases using a naïve bayes style possibilistic classifier, in *2013 IEEE International Conference on Systems, Man, and Cybernetics*, (2013), 4539–4542. https://doi.org/10.1109/SMC.2013.772

7. D. Çalişir, E. Dogantekin, A new intelligent hepatitis diagnosis system: PCA-LSSVM, *Expert Syst. Appl.*, **38** (2011), 10705–10708. https://doi.org/10.1016/j.eswa.2011.01.014

8. A. F. Otoom, E. E. Abdallah, Y. Kilani, A. Kefaye, M. Ashour, Effective diagnosis and monitoring of heart disease, *Int. J. Software Eng. Appl.*, **9** (2015), 143–156. https://doi.org/10.14257/ijseia.2015.9.1.12

9. C. W. Liang, H. C. Yang, M. M. Islam, P. A. A. Nguyen, Y. T. Feng, Z. Y. Hou, et al., Predicting Hepatocellular Carcinoma with minimal features from electronic health records: Development of a deep learning model, *JMIR Cancer*, **7** (2021), e19812. https://doi.org/10.2196/19812

10. K. Kim, H. Yang, J. Yi, H. E. Son, J. Y. Ryu, Y. C. Kim, et al., Real-time clinical decision support based on recurrent neural networks for in-hospital acute kidney injury: External validation and model interpretation, *J. Med. Int. Res.*, **23** (2021), e24120. https://doi.org/10.2196/24120

11. Y. Du, H. Wang, W. Cui, H. Zhu, Y. Guo, F. A. Dharejo, et al., Foodborne disease risk prediction using multigraph structural long short-term memory networks: Algorithm design and validation study, *JMIR Med. Inf.*, **9** (2021), e29433. https://doi.org/10.2196/29433

12. A. Sedik, M. Hammad, A. El-Samie, E. Fathi, B. B. Gupta, A. El-Latif, et al., Efficient deep learning approach for augmented detection of Coronavirus disease, *Neural Comput. Appl.*, (2021), 1–18. https://doi.org/10.1007/s00521-020-05410-8 .

13. K. Zhang, X. Zhao, L. Zhuang, Q. Xie, H. Zan, Knowledge-enabled diagnosis assistant based on obstetric EMRs and knowledge graph, in *China National Conference on Chinese Computational Linguistics*, Springer, **23** (2020), 444–457. https://doi.org/10.1007/978-3-030-63031-7_32

14. J. Devlin, M. W. Chang, K. Lee, K. Toutanova, BERT: Pretraining of deep bidirectional transformers for language understanding, preprint, arXiv:1810.04805.

15. A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, in *Advances in Neural Information Processing Systems*, **26** (2013), 1–9. https://dl.acm.org/doi/10.5555/2999792.2999923

16. T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, preprint, arXiv:1609.02907.

17. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., Attention is all you need, in *Advances in Neural Information Processing Systems*, **30** (2017), 1–11. https://dl.acm.org/doi/10.5555/3295222.3295349

18. C. Qu, L. Yang, M. Qiu, W. B. Croft, Y. Zhang, M. Iyyer, BERT with history answer embedding for conversational question answering, in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, (2019), 1133–1136. https://doi.org/10.1145/3331184.3331341

19. Z. Yang, Y. Huang, Y. Jiang, Y. Sun, Y. J. Zhang, P. Luo, Clinical assistant diagnosis for electronic medical record based on convolutional neural network, *Sci. Rep.*, **8** (2018), 1–9. https://doi.org/10.1038/s41598-018-24389-w

20. Y. Chen, Convolutional neural network for sentence classification, preprint, arXiv:1408.5882.

21. S. Lai, L. Xu, K. Liu, J. Zhao, Recurrent convolutional neural networks for text classification, in *Twenty-ninth AAAI Conference on Artificial Intelligence*, (2015), 2267–2273. https://dl.acm.org/doi/abs/10.5555/2886521.2886636

22. P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, et al., Attention-based bidirectional long short-term memory networks for relation classification, in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, **2** (2016), 207–212. https://doi.org/10.18653/v1/P16-2034

23. K. Zhang, C. Hu, Y. Song, H. Zan, Y. Zhao, W. Chu, Construction of Chinese obstetrics knowledge graph based on the multiple sources data, in *Workshop on Chinese Lexical Semantics*, (2022), 399–410. https://doi.org/10.1007/978-3-031-06547-7_31

24. O. Byambasuren, Y. Yang, Z. Sui, D. Dai, B. Chang, S. Li, et al., Preliminary study on the construction of Chinese medical knowledge graph, *J. Chin. Inf. Process*, **10** (2019), 1–9.

25. H. Zan, Y. Han, Y. Fan, C. Niu, K. Zhang, Z. Sui, Construction and analysis of symptom knowledge base in Chinese, *J. Chin. Inf. Process*, **34** (2020), 30–37.

AIMS Press