Mathematical Biosciences and Engineering

MBE, 21(1): 1489–1507.
DOI: 10.3934/mbe.2024064
Received: 06 November 2023
Revised: 15 December 2023
Accepted: 19 December 2023
Published: 28 December 2023

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

Research article

A study on pharmaceutical text relationship extraction based on heterogeneous graph neural networks

Shuilong Zou[1], Zhaoyang Liu[2], Kaiqi Wang[2], Jun Cao[2], Shixiong Liu[1,*], Wangping Xiong[2,*] and Shaoyi Li[1]

[1] Nanchang Institute of science & Technology, Nanchang 330004, China
[2] School of Computer, Jiangxi University of Chinese Medicine, Nanchang 330004, China

* **Correspondence:** Email: mrsix408@163.com, 20030730@jxutcm.edu.cn.

Abstract: Effective information extraction of pharmaceutical texts is of great significance for clinical research. The ancient Chinese medicine text has streamlined sentences and complex semantic relationships, and the textual relationships may exist between heterogeneous entities. The current mainstream relationship extraction model does not take into account the associations between entities and relationships when extracting, resulting in insufficient semantic information to form an effective structured representation. In this paper, we propose a heterogeneous graph neural network relationship extraction model adapted to traditional Chinese medicine (TCM) text. First, the given sentence and predefined relationships are embedded by bidirectional encoder representation from transformers (BERT fine-tuned) word embedding as model input. Second, a heterogeneous graph network is constructed to associate words, phrases, and relationship nodes to obtain the hidden layer representation. Then, in the decoding stage, two-stage subject-object entity identification method is adopted, and the identifier adopts a binary classifier to locate the start and end positions of the TCM entities, identifying all the subject-object entities in the sentence, and finally forming the TCM entity relationship group. Through the experiments on the TCM relationship extraction dataset, the results show that the precision value of the heterogeneous graph neural network embedded with BERT is 86.99% and the F1 value reaches 87.40%, which is improved by 8.83% and 10.21% compared with the relationship extraction models CNN, Bert-CNN, and Graph LSTM.

Keywords: medical text; relation extraction; BERT; heterogeneous graph neural networks

1. Introduction

As the main carrier of medical knowledge, extracting effective information from texts is of great significance for modern clinical research [1]. Chinese medicine has been practiced for thousands of years, forming a rich text of books, medical cases, etc. Unlike ordinary texts, Chinese medicine texts have many technical terms, diverse language expressions, streamlined structures, and complicated semantic relationships [2,3], such as licorice and dried ginger decoction,... , divided into temperature and then served, contains both the temperature of medication and the frequency of medication, resulting in entity boundary identification difficulty, knowledge extraction of which should be upgraded from the language level to the content level [4]. Meanwhile, relationship extraction in the field of Chinese medicine is intricate and complex, and the relationship may exist between heterogeneous entities, which have both sibling relationship and inclusion relationship, etc. [5,6], for example, the compositional relationship between traditional Chinese medicines and formulas, the utility relationship between medicines and certificates, the correlation relationship between certificates and illnesses, the therapeutic relationship between formulas and certificates, and so on. Also, the mainstream extraction methods seldom take into account the correlation that exists between the entities and the relationships when extracting, resulting in insufficient semantic information. This leads to insufficient semantic information and has certain limitations in this scenario [7,8].

Relational extraction technique is an important part of information extraction and a prelude subtask of knowledge base building, which can effectively extract associations between words from massive text and convert native text into structured information [9,10]. 1490ith traditional entity-relationship extraction algorithms, neural network relationship extraction methods have stronger generalization ability to effectively capture textual information and apply it to a wider range of data domains, which is the backbone of current relationship extraction models [11,12]. In order to implement the relationship extraction subtask used for chemical pathogenic correlation, Zhou et al [13] proposed the integration of a neural network-based, a feature-based, and a kernel function based three-layer relation extraction framework to achieve the relation extraction task in related fields. Convolutional neural networks are also used in relationship extraction tasks. Zhang et al [14] proposed a hybrid model combining recurrent neural networks (RNN) and convolutional neural networks (CNN) for biomedical relationship extraction in a relationship extraction task in the supervised learning domain to learn latent features from a sequence of sentences. Quirk et al [15] proposed a graph network model for cross-sentence relationship extraction, which combines sentence sequences with features. In cross-sentence relationship extraction, where words in a sentence are used as nodes of a graph, edges are determined by neighboring words, dependencies, and semantic relationships, and reachable paths between entities are computed by the graph, thus enabling relationship extraction. Shi et al [16] proposed a graph neural network based distant supervision relation extraction (DSRE), which explores by breadth, for those direct neighbors that contribute more to the relationship prediction by assigning higher weights and deep exploration to determine the correlation between the relationship and higher order neighbors. Liang [17,18] et al. proposed a heterogeneous graph-based encoder to represent conversational content, using emotional personality-aware decoders to form appropriately related responses that can effectively perceive contextually relevant connections.

In the field of Chinese medicine, graph neural networks (GNN) have been proposed for the study of Chinese medicine. Zhao et al [19] constructed a Chinese medicine diagnosis and treatment algorithm model based on GNN and MLP to realize the process of intelligent recommendation from Chinese

medicine symptom data to clinical medication. Research on traditional Chinese medicine (TCM) relationship extraction of entity associations generally involves drinking tablets, prescription evidence, and symptoms, and are based on rule-based solutions. Zhou et al [20] proposed a combined calculation method based on bootstrapping and relationship weights, which integrates TCM literature data and some biological data to derive the relationship between TCM evidence and genes. Wan et al [21] gave a new heterogeneous entity factorial graph model, which integrates the TCM literature of the last few years and labels it as a discourse dataset, thus effectively extracting the relationships to the dataset. Thus, the relationships between physical objects such as Chinese medicines, formulas, and diseases of the dataset are extracted efficiently. The proposed methods for relational extraction in the field of TCM are different due to the specific content and application variability, and it is especially important to propose a relational extraction method suitable for a specific problem [22].

In summary, the relationship extraction technology has been rapidly developed, but the current mainstream relationship extraction model does not consider the association between entities and relationships during extraction, resulting in insufficient semantic information to form an effective structured representation. Therefore, this paper proposes a relationship extraction method based on heterogeneous graph attention network (HGAN) to solve the problem of feature generation in the process of relationship extraction for Chinese medicine antiquities. First, the adopted data are derived from TCM ancient books and the data format is standardized through text annotation and preprocessing operations to obtain a TCM ancient book corpus. Then using the HGAN model, the words and relations in the text are encoded into model input node vectors, the information from the word nodes and relation nodes are fused through multi-head attention and heterogeneous graph layers, and the attributes of the Chinese medicine prescription entities corresponding to the method of taking medicine are determined by the relation extraction layer. The method proposed in this paper has a higher accuracy in the relationship extraction of Chinese medicine ancient books and can obtain ternary groups more accurately.

2. Materials and methods

In this study, we propose a textual relation extraction model, and Figure 1 shows the general design of the whole study. First, the corpus is constructed and text crawling technique is used to obtain some ancient Chinese medicine texts to form a corpus of ancient Chinese medicine serving [23]. The Bidirectional Encoder Representation from Transformers (BERT) is later fine-tuned using a corpus of ancient Chinese medicine texts. Second, node vector embeddings are obtained, and the words as well as relations in the sentences are formed into node vectors by pre-training the language model [24,25] as inputs to the relation extraction model, given the sentences and the predefined relation categories. Then, the obtained word node and relation node vectors are utilized as inputs to the heterogeneous graph neural network layer [26,27] to fuse the word node and relation node representations. Finally, using the relationship extraction model based on the heterogeneous graph neural network architecture, the feature representations obtained from the heterogeneous graph neural network layer are input to the relationship extraction layer, and entities and attributes are obtained through the entity-attribute tagger [28], respectively, to determine the entity-relationship pairs in the whole sentence.

The text S is a complete sentence from an ancient text containing a classical prescription formula and a method of taking medicine, S contains a classical prescription entity, e_1, and multiple attributes of taking medicine, a_i, and the relational label, R, is the relationship between e_1 and a_i, e.g.,

Pueraria Mirifica Tang is e_1, warm drink is a_i, and temperature of taking medicine is R. That is, given the set of sentences S and relational labels R, the goal is to learn a function F that predicts possible relations between classical prescription entities and the attributes of taking medicine in a sentence of the relationship category.

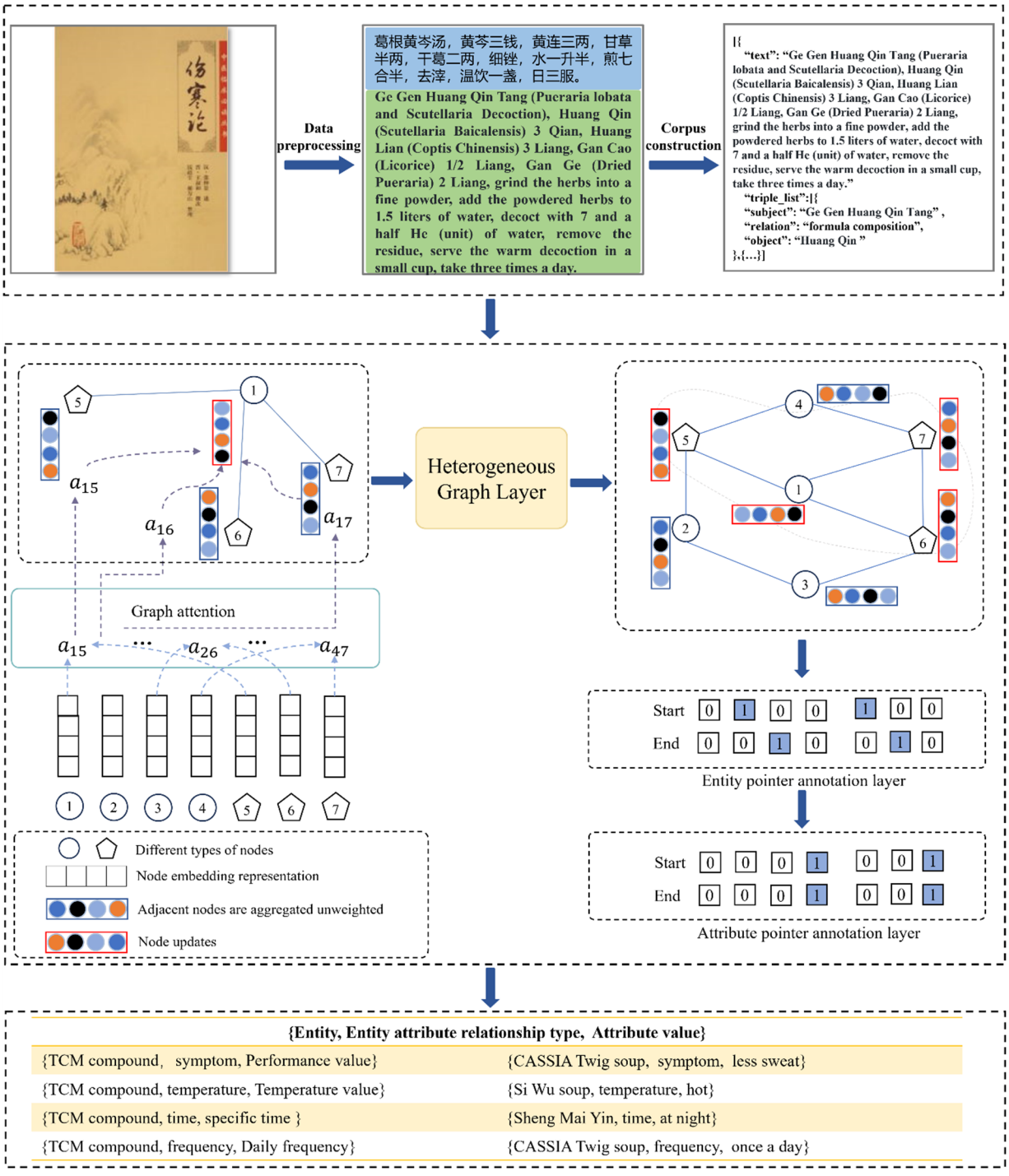

Figure 1. The overall flow of the research.

2.1. Corpus construction

Text crawler technology was used to gather a portion of ancient Chinese medicine text, from which we manually extracted sentences containing effective Chinese medicine prescriptions and medication methods. The collated ancient Chinese medicine text was then annotated and organized using the Doccano automated text annotation tool, including sequence annotation [29]. The annotated text included identified entities and attributes such as classical prescription, temperature, dosage, frequency, and nine other types. Additionally, we analyzed the relationships between the eight temperature groups of medication, dosage, frequency, time, solution type, and other factors. For the first time, a corpus of TCM ancient medication is formed, which is used for the training and testing of the relationship extraction model in this study, and it also supports the construction of a medication knowledge graph.

2.2. Word embedding layer

Model training starts with embedding operation, and this method proposes word node embedding as well as relation node embedding together to form the model input [30]. Among them, word nodes represent the linguistic content of a sentence and capture the linguistic context and semantics of the words in the sentence, while relation nodes represent the predefined relations between entities in a heterogeneous graph and are specialized for capturing semantic information related to the predefined relations between the entities in the graph, and they play different roles in comprehensively presenting the linguistic semantics and structural relations in a heterogeneous graph. Since ancient Chinese medical texts contain many symptoms, syndromes, and herbs specific to the field of Chinese medicine, the clinical records are written in archaic Chinese, in which the Chinese characters can have different meanings and orders in the sentences. In order to fill the gap between the generic Chinese BERT model and TCM ancient texts, we perform domain language modeling fine-tuning on the pre-trained BERT-Base-Chinese language model for TCM ancient texts. Assuming that the input ancient text is X, whose text word set is $S(x_1, x_2, x_3, \dots, x_n)$, n is the length of the text, where the ith word is denoted as x_i, and all the characters of the text token are inputted into BERT to get the word node representation as:

$$[h_1, h_2, \dots, h_n] = \text{BERT}([x_1, x_2, \dots, x_n]) \tag{1}$$

where h_i is the encoded hidden layer representation, the initial input of the node.

Also, the predefined relational labels are embedded as high dimensional vectors $[r_1, r_2, \dots, r_m]$ through the linear mapping layer.

$$[r_1, r_2, \dots, r_m] = \sigma([l_1, l_2, \dots, l_m]) \tag{2}$$

where r_j is the representation of the relationship node after encoding, which is used as an input to the model at the same time as h_i, and σ is the representation of the linear mapping transformation.

2.3. Graph attention layer

To incorporate more semantic information representation, we convert the original adjacency matrix into multiple attention-guided adjacency matrices by using multiple heads of attention to capture interactions between any two positions in a single sequence, with the matrices serving as inputs

to later heterogeneous graph layer computations. Each matrix corresponds to a fully connected graph, and each entry matrix is the weight of a node-to-node edge. This allows the models to jointly focus on information from different representation subspaces. The computation involves a query and a set of key-value pairs. The output is computed as a weighted sum of values, where the weights are computed by the query function with the corresponding keys.

Heterogeneous graph layers consider nodes of one class as neighbors of nodes in another class of graphs and update the node information using a method similar to graph attention networks [31]. The main idea comes from graph attention network, nodes in the layer can pay attention to the characteristics of their neighboring nodes, assigning different weights to different nodes in the neighborhood. Figure 2 shows the schematic diagram of node information updating.

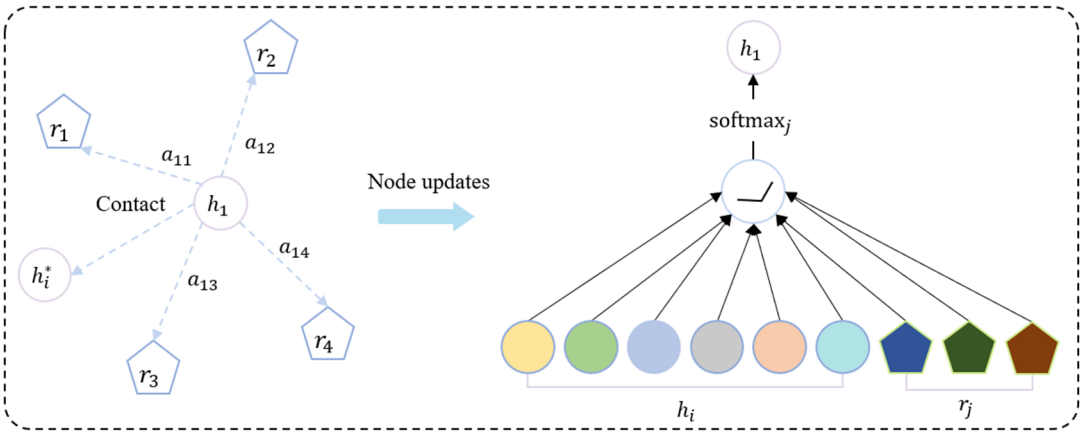

Figure 2. Node representation update.

The α_{ij} in the graph represents the probability of the attention weight of the ith node of the word to the jth node of the relation representation, which is calculated as follows:

$$w_{ij} = w_1[h_i, r_j] + b \tag{3}$$

$$\alpha_{ij} = \frac{\exp(w_{ij})}{\sum_{i=1}^{N} \exp(w_{ij})} \tag{4}$$

$$h_i^* = h_i + \sum_{j=1}^{m} \alpha_{ij} w_2 r_j \tag{5}$$

where w_1 and w_2 are trainable weight matrices, N is the sentence length, m is the number of predefined relations, h_i and r_j are word node and relation node representations, and h_i^* is the representation of word node information fused with relation information.

In order to make the neural network have enough ability to catch more complex features, a gate mechanism is used after the activation function to keep the nonlinear ability, which is calculated as follows:

$$e_i = \text{sigmoid}(w_3[h_i, h_i^*] + b) \tag{6}$$

$$\tilde{h_i} = e_i\, h_i^* + (1 - e_i) h_i \tag{7}$$

where w_3 is the trainable parameter weights and $h_i^{\sim}$ is the output of the final heterogeneous graph layer. Similarly, the relation nodes are also operated by the above node update principle to get the relation node representation that incorporates the word information. In addition, in order to avoid the possible gradient disappearance during the training process, the residual computation connection is added after the above steps. The specific calculation is as follows:

$$h = h_i^{\sim} + h_i \tag{8}$$

$$r = r_j^{\sim} + r_j \tag{9}$$

The word nodes combine all the relationship information to get the representation, and the relationship nodes depend on the updated word node information update, which is well semantic fusion, such a graph node information update strategy is more suitable for text with certain characteristics of the relationship extraction.

2.4. Heterogeneous graphical network coding layer

The heterogeneous graph layer (HGL) part of the heterogeneous graph neural network model is shown in Figure 3, where two types of semantic node representations are obtained from the word embedding operation described above, and then the set of word node representations h_i and the set of relation nodes r_j are spliced $[h_i, r_j]$.

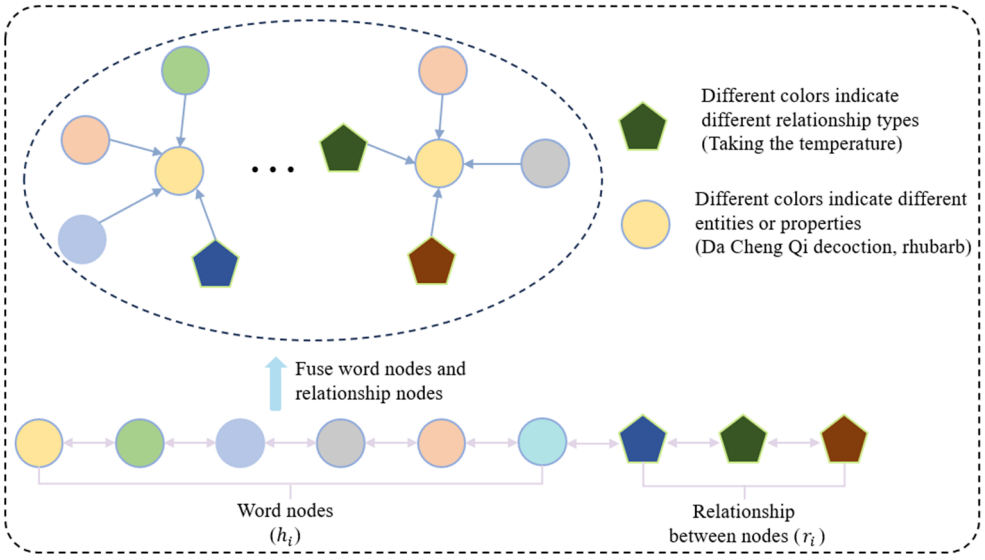

Figure 3. Heterogeneous graph network.

2.5. Two-stage decoding extraction layer

The conventional practice of relationship extraction is to first extract entities and attributes, and then classify the relationships between entity attributes to form entity-attribute relationship pairs. In this paper, we adopt a two-stage entity attribute identifier method, where the identifier uses a binary classifier to recognize the start and end positions of the entities. The advantage of this approach over

the mainstream entity sequence decoding is that it can deal with entity repetition well and reduces meaningless entity-relationship pairs to a certain extent.

After obtaining the final representations of word nodes and relation nodes, the entity attribute identifier recognizes all possible entity attributes in the word nodes respectively, thus obtaining the entity attribute relation pairs in the sentence. Specifically, the binary classifier identifies the start and end position of the entity, and uses [0, 1] to identify each token of the sentence, "1" in the start layer represents the start position of the entity, and "0" in the end layer means the position is the end position of the entity, the combination of the two layers determines the entity relationship pairs. The combination of the two layers identifies the entity. s represents the start position and e represents the end position. The calculation is as follows:

$$p_i^s = \text{sigmoid}(w_s h^l + b_s) \tag{10}$$

$$p_i^e = \text{sigmoid}(w_e h^l + b_e) \tag{11}$$

where h^l is the last heterogeneous layer output of the word node, w_s, w_e are the weights, b_s, b_e are the biases, and p_i^s, p_i^e are the probabilities that the ith word node is represented as the start position and the end position, respectively [32]. The entity identifier is optimally trained to determine the entity x span in sentence S by the following likelihood function.

$$p_\theta(x|S) = \prod_{\lambda\in(s,e)} \prod_i^n \left(p_i^\lambda\right)^{f\left\{y_i^\lambda=1\right\}} \left(1-p_i^\lambda\right)^{f\left\{y_i^\lambda=0\right\}} \tag{12}$$

where n is the sentence token length, $f\{\phi\} = 1$ if ϕ is true and 0 otherwise, and y_i^s and y_i^e denote the binary representations of the start and end positions of the ith token in the sentence, respectively. After identifying the entity, the word node, relationship node, and candidate entity information are further fused as inputs to the attribute identification network, and the possible probabilities of the start and end positions are obtained in the same way, which are calculated as follows:

$$h_o = w_4[h^l, r^l, \delta_k] + b_4 \tag{13}$$

$$p_i^{s_o} = \text{sigmoid}(w_s^o h_o + b_s^o) \tag{14}$$

$$p_i^{e_o} = \text{sigmoid}(w_e^o h_0 + b_e^o) \tag{15}$$

where r^l is the last layer output of the relation node; δ_k is denoted as the kth candidate head entity representation; w_4, w_s^o, and w_e^o are the parameter weights; b_4, b_s^o and b_e^o are the biases; and $p_i^{s\_o}$ *and* $p_i^{e\_o}$ are the probabilities of the ith word node representation as the start position and the end position of the tail entity.

The same entity is optimally trained with the maximum likelihood function, and given the sentence S, the relationship r between entities in the sentence, and the identified entity x, the likelihood function of the attribute o to be extracted is expressed as the following equation.

$$p_\theta(o|x,S,r) = \prod_{\lambda\in(s,e)} \prod_i^n \left(p_i^\lambda\right)^{f\left\{y_i^\lambda=1\right\}} \left(1-p_i^\lambda\right)^{f\left\{y_i^\lambda=0\right\}} \tag{16}$$

where θ is the attribute identifier parameter, the same as having such a parameter in the entity identifier. Based on the likelihood function of the entity attributes, the loss function of the model can

be identified as shown in the following equation.

$$L = \sum_{x \in D_i} \log p_\theta(x|S) + \sum_{r \in D_{i|x}} \log p_\theta(o|x,S,r), (x,r,o) \in D \quad (17)$$

The specific algorithmic pseudo-code flow for heterogeneous graph neural network relationship extraction is shown in Table 1.

The pseudo-code for the model training process is shown in Table 1, where S is the training statement, e is the representation of entity information in the statement, and r is the relationship between entities in the training statement. The output of the model is the combination of entity attributes and relationships between them. h_i is the ith word node representation after BERT encoding, r_j is the relationship node representation after linear mapping, and I_{list} is the fusion of word node representations and relationship node vectors as the model inputs. h and r denote the hidden layer representations of the words and relationships after the heterogeneous graph network layer, respectively, E_{sub} is the entity representation obtained by entity E_{sub} is the entity representation obtained by the entity recognizer, Ent_{list} refers to the spliced representation of entity, word and relation, and then Ent_{list} is inputted to the attribute identifier to get the attribute E_{obj}, and finally form the entity attribute-relationship pairs.

Algorithm 1. Pseudo code of heterogeneous graph neural network method.

Input：Training set $(S, e, r) \in D$
Output：Model test results *Out*

1：Parameter Initialization
2：**for** i in S
3： **for** j in r **do**
4： $h_i = \text{BERT}(i)$
5： $r_j = \text{linear(j)}$
6： $\text{I}_{\text{list}} = \{\text{h}_\text{i}\}+\{\text{r}_\text{j}\}$
7：**End for**
8：**for** batch in *D* **do**
9： $h, r = \text{Heterogeneous layer\_encoder}(Ilist)$
10： $E_{sub} = \text{SubjectEntity Tagger}(h)$
11： $\text{E}nt_{list} = \text{Contact}(h, r, E_{sub})$
12： $E_{obj} = \text{ObjectEntity Tagger}(\text{E}nt_{list})$
13： $Out = \{E_{sub}, E_{obj}, r\}$
14：**End for**
15：**return** ***Out***

3. Results

3.1. Experimental data

In order to validate the performance of the relationship extraction model proposed in this paper, and also to construct the knowledge graph of Chinese medicine taking medicine, relevant experiments are conducted on the text dataset based on the ancient books of Chinese medicine to validate it. The

dataset is derived from the classical Chinese medicine books "Treatise on Exogenous Febrile Disease", "The Essentials of the Golden Chamber", "The Essentials of Thousand Gold", etc. More than 5000 texts containing methods of taking medicines by prescription are obtained by using text crawler technology combined with manual screening. We removed spaces and non-Chinese characters from the data first, and then annotated with the doccano annotation tool to determine the candidate relationships by labeling the entities of the classical prescriptions and the attributes of the medication, to obtain the corpus of TCM ancient medication for the first time.

The precursor database contains a total of 20,763 pairs of relationship pairs between eight categories of classical prescription entities and serving attributes, and it is also divided into training, testing and validation sets according to a certain proportion. The original annotation information of the TCM ancient text dataset is shown in Table 1, which includes eight different relationship annotation styles, each of which has a clear category, while the entities and attributes correspond to specific instances, respectively.

Table 1. Partial labeling information.

Relationship category	Serial number	Classical prescription entity	Properties of medication
Medication schedule	18	Qin Pi Da Huang Tang	post-diet
Post-medication manifestations	455	Licorice and epimedium soup	sweating
Medication temperature	3086	Pueraria lobata	room temperature
Frequency of medication	1408	Bai Hu Tang	three doses per day
Medication solution	2033	Fu Yuan Tong Qi San	wine
Formulas	2	Pueraria Mirifica Scutellaria Soup	Scutellaria
Dose of medication	2	Pueraria Mirifica Scutellaria Soup	seventy-five percent
Classical prescription for diseases	40	Chai Hu Gui Zhi Tang	abdominal pain

The original labeled data is shown in Figure 4, in order to make the data conform to the model input, it is subjected to the preprocessing operation of removing the noisy data and format conversion, as shown in Figure 5, where text is the original textual information, and in the triple_list tuple, the first column is the entity of the scrip, the third column is the attribute of the method of taking the medicine, and the second column is the relationship between the two.

3.2. Experimental setup and evaluation indicators

The hyperparameters of the heterogeneous graph-based attention network model in this paper are shown in Table 2. By adjusting the learning rate lr and some other important parameters, the output of the model tends to be stable and optimal.

[{
"text" : "Ge Gen Huang Qin Tang (Pueraria lobata and Scutellaria Decoction) ,Huang Qin (Scutellaria Baicalensis) 3 Qian, Huang Lian (Coptis Chinensis) 3 Liang, Gan Cao (Licorice) 1/2 Liang, Gan Ge (Dried Pueraria) 2 Liang, grind the herbs into a fine powder, add the powdered herbs to 1.5 liters of water, decoct with 7 and a half He (unit) of water, remove the residue, serve the warm decoction in a small cup, take three times a day",
"relation list":[{
"subject": "Ge Gen Huang Qin Tang" ,
"relation" : "Composition",
"object": "Huang Qin "
},{
"subject": "Ge Gen Huang Qin Tang" ,
"relation" : "Composition",
"object": "Huang Qin "

},{
"subject": "Ge Gen Huang Qin Tang" ,
"relation" : "Composition",
"object": "Huang Qin "

},{
"subject": "Ge Gen Huang Qin Tang" ,
"relation" : "Composition",
"object": "Huang Qin "

},{
"subject": "Ge Gen Huang Qin Tang" ,
"relation" : "Composition",
"object": "Huang Qin "
},{

Figure 4. Original annotation data.

[{
"text" : "Ge Gen Huang Qin Tang (Pueraria lobata and Scutellaria Decoction) ,Huang Qin (Scutellaria Baicalensis) 3 Qian, Huang Lian (Coptis Chinensis) 3 Liang, Gan Cao (Licorice) 1/2 Liang, Gan Ge (Dried Pueraria) 2 Liang, grind the herbs into a fine powder, add the powdered herbs to 1.5 liters of water, decoct with 7 and a half He (unit) of water, remove the residue, serve the warm decoction in a small cup, take three times a day",
"triple list":[
["Ge Gen Huang Qin Tang ", "Composition", "Huang Qin "],
["Ge Gen Huang Qin Tang ", "Composition", "Huang Lian "],
["Ge Gen Huang Qin Tang ", "Composition", "Gan Cao "],
["Ge Gen Huang Qin Tang ", "Composition", "Gan Ge "],
["Ge Gen Huang Qin Tang ", "Dosage", "7 and a half He "],
["Ge Gen Huang Qin Tang ", "Temperature", "serve the warm decoction"],
["Ge Gen Huang Qin Tang ", "Frequency", "take three times a day"],
] },{…}]

Figure 5. Processed data.

Table 2. Model parameter.

Parameter name	Parameter value
epoch	100
max_length	256
batch_size	16
lr	1×10^{-1}

In the experiments in this section, precision, recall, and F1 values are used as metrics to evaluate the relational extraction model effectiveness. The specific formulas are as follows:

$$\text{precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{18}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{19}$$

$$\text{F1} = \frac{2*\text{precision}*\text{recall}}{\text{precision+recall}} \tag{20}$$

where TP is the number of correctly extracted relation pairs, FP is the number of misreported relation pairs, and FN is the number of missed relation pairs.

3.3. Analysis of the results of the relational extraction experiment

Model training allows us to compare the eight categories of pill-taking methods and evaluate their performance using test set and validation set data. The outputs of the evaluation metrics for each type of pill-taking method are shown in Table 3.

According to the analysis of the above results, among the eight categories of relationships, the F1 values of the three categories of relationships, namely, classical prescription composition, medication temperature, and medication time, exceed 90%, which is better than the other categories. One of the reasons may be that these three categories of relationship categories account for a larger proportion of the dataset, the greater the number of corpus will lead to stronger feature learning of the model, so that the accuracy rate will be improved. Another may be that these three categories of medication attributes are more differentiated in the textual features, and the more different the textual features are, the lower the error rate of the model in learning this part of the features. Therefore, these three categories of modeling effect are better than other relationship categories.

Table 3. Different relationships on experimental results.

Relationship category	Precision(P) %	Recall(R) %	F1 %
HP（Post-medication manifestations）	81.25	76.47	78.79
HT（Medication temperature）	90.63	94.16	92.36
HS（Medication solution）	85.71	80.49	83.02
HC（Formulas）	92.38	91.91	92.15
HT（Medication schedule）	91.26	92.27	91.76
HF（Frequency of medication）	89.54	82.53	85.89
HJ（Dose of medication）	81.94	84.69	83.29
HB（Diseases）	77.27	76.84	77.05

In the parameter settings of this experiment, the learning rate lr affects the convergence state of the model, in order to explore the lr value of the model, through a number of test experiments, replacing different lr values to observe the convergence state of its F1 value (take the similar three groups for comparison), the results are shown in Figure 6, the x-axis is the epoch and the y-axis is the F1 value. It is finally determined that the lr value of this experiment is taken as 1×10^{-1}.

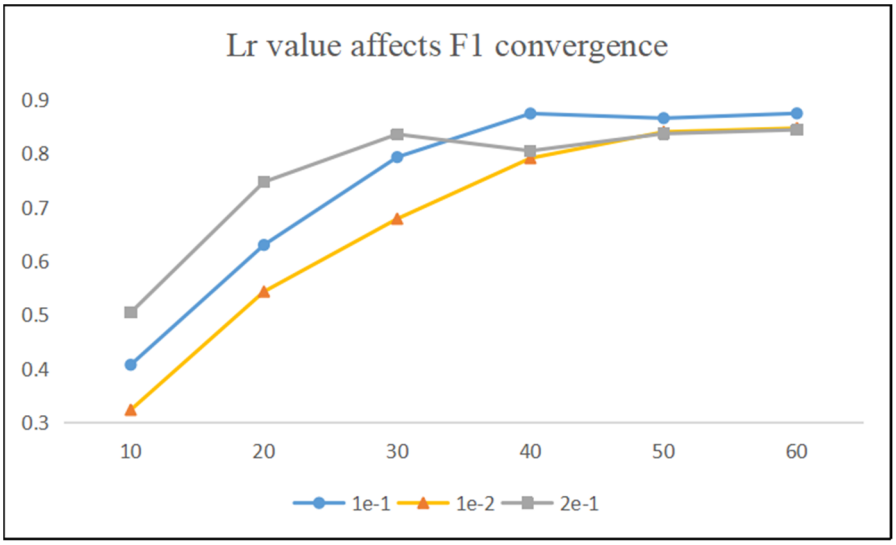

Figure 6. Influence of different lr values on model results.

In order to verify the applicability of this experimental model to TCM ancient texts, the TCMRel dataset (abstracts of TCM literature published on cnki.net in the past ten years) and non-CM text datasets are put into the model for comparison experiments, and the experimental results of each type of data are outputted, see Tables 4. The results show that the P-value, R-value, and F1-value of the TCM ancient texts corpus run for this experiment are better than those of other datasets, and that the present research model is more suitable for the relational extraction of TCM ancient texts.

Table 4. Different data test results.

Data sources	Precision(P) %	Recall(R) %	F1 %
Non-Chinese Medicine Corpus	78.98	75.89	77.40
TCMRel	83.67	77.74	80.59
Chinese Medicine Ancient Texts Corpus	86.99	87.81	87.40

In order to verify the effectiveness of the proposed model in this experiment, three traditional methods, conventional CNN, Bert_CNN and graph LSTM, are introduced as comparison models. The results of each experimental model are shown in Table 5.

Table 5. Different model test results.

Model	Precision(P) %	Recall(R) %	F1 %
CNN	68.33	68.21	68.26
Bert_CNN	72.31	70.14	71.21
Graph LSTM	78.16	76.24	77.19
GNN	80.10	81.31	81.13
HGAN	86.99	87.81	87.40

According to the results in Tables 5, the methods mentioned in this section show good results in the ancient Chinese medical text relationship extraction corpus with higher values of various

evaluation indices than the control approach. With the simple CNN model, it can automatically obtain the textual features and encode them into an end-to-end network approach to realize the distribution of relationship categories between the convolutional layer and the fully-connected layer Softmax, whereas in this experiment, the F1 value outputted due to the unimproved model is only 68.26%. With the Bert_CNN model, the feature sequences of the entity information in the article are first obtained, and then the possible associated entity pairs are predicted. Due to the lack of a priori knowledge for support, there are more relationship pairs between entities, which affects the experimental results, and the F1 value is 68.87%. Graph LSTM relationship extraction method does an embedding operation on the words, and the LSTM method is used in graph neural network for feature learning, outputs contextual entity representations, and then carries out the relationship extraction, which is applicable to the N-element relationship extraction and is not practical for this corpus. The utility is not strong and the F1 value is 75.62%. Ancient Chinese medicine texts usually contain complex and large graph structures, in which the relationships between entities (herbs, diseases) may present a complex structure with multiple levels and jumps, and it is difficult for the GNN model to adapt well to this highly complex graph structure, thus affecting the experimental results, with an F1 value of 81.13.

The method proposed in this chapter combines the predefined relational a priori knowledge in the sentence in the encoding stage to obtain a feature sequence that incorporates relational information, Then, it utilizes a two-stage approach in the decoding stage to first identify the scrip entities in the sentence, and then uses the identified scrip entity information and the encoded identifiers to identify the attribute representations of medication methods in the relational group. For the reasons mentioned above, the model achieves better results on this dataset, with an integrated F1 value reaching 87.40%.

To better test the stability of the model, we set the number of different types of iterations on this dataset and gradually increase it until the maximum number of iterations in the parameter settings is reached. Set all models to the same parameters, output the F1 value, and observe the iteration results, as shown in Figure 7, the x-axis is the epoch and the y-axis is the F1 value. Also, output the comprehensive evaluation results of each modeling, including P-value, R-value, and F1-value, and the results are shown in Figure 8, the x-axis is the epoch, and the y-axis is the F1 value.

According to the comparison results in Figure 7, the F1 value of each model increases with the number of iterations, showing an initial increase and then tends to be smooth, and the convergence speed of the CNN model is slower than that of the other three models; although the Bert-CNN has a faster convergence speed, but its F1 value is lower; the graph LSTM model has a faster convergence speed and higher F1 value, and it is less stable when its convergence is followed by a more substantial oscillation.

From the comparison of Figures 8, it is concluded that the evaluation indexes based on the HGAN model present the best results when the number of iterations reaches 40 times, and with the increase of the number of iterations there is a smooth fluctuation within the controllable range. Therefore, both in terms of convergence speed and stability, the research model proposed in this paper is optimal and suitable for the relational extraction of TCM ancient texts.

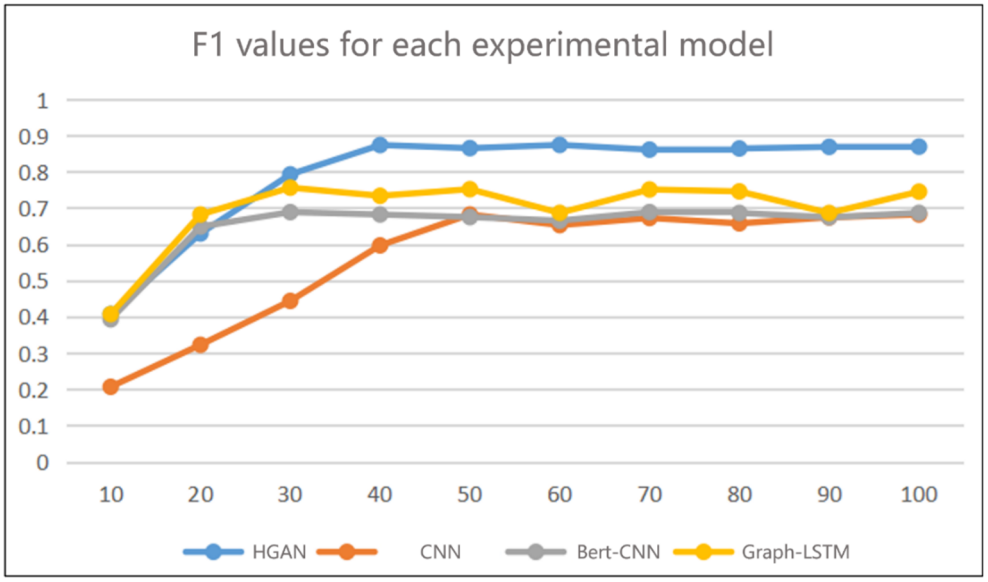

Figure 7. Iterative results of F1 values of each model.

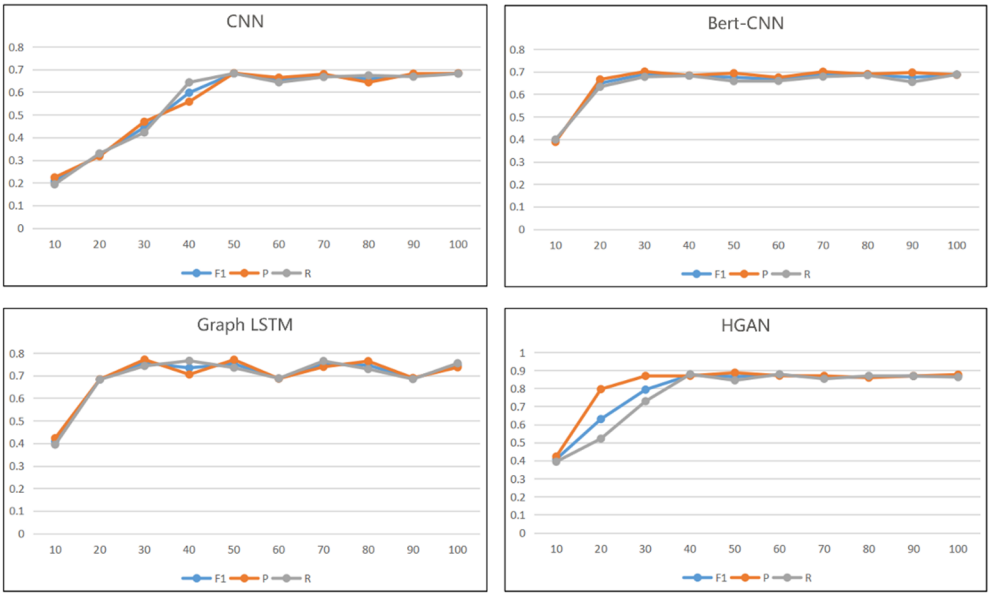

Figure 8. Model iteration results.

4. Discussion

In this study, we propose and validate a heterogeneous graph neural network relationship extraction model adapted to Chinese medicine texts, which achieves significant performance improvement on the Chinese medicine relationship extraction task. The semantic associations and complex relationships between heterogeneous entities in TCM texts are successfully captured by BERT fine-tuned word embedding and combined with graph attention mechanism. The experimental results show that the heterogeneous graph neural network embedded with BERT reaches 86.99% and 87.40% in terms of PRECISION and F1 values, which are at least 8.83% and 10.21% higher,

respectively, compared to other relation extraction models.

First, our model employs BERT fine-tuning, which utilizes a pre-trained language representation model, enabling the model to better understand complex structural and semantic information when processing ancient Chinese medicine utterances, and helps to locate entities and identify relationships more accurately. In contrast, traditional relation extraction models such as CNN, Bert-CNN, and graph LSTM may fail to capture the special linguistic features of TCM texts well and thus have some limitations in processing TCM texts.

Second, we introduce heterogeneous graph networks and graph attention mechanisms to effectively model the associations between words, phrases, and relation nodes. Semantic relationships in TCM texts are often multi-layered and multi-dimensional, and traditional models may fail to handle these heterogeneous relationships well, resulting in relatively low performance. Our model improves the accuracy of relationship extraction by more comprehensively understanding the structural information of the text, especially the sensitivity to heterogeneous relationships.

Despite the performance gains achieved by our model on the TCM relationship extraction task, there are still some limitations. For example, the generalization ability of the model may be limited by the training data, and more diverse and rich data are needed for validation. In addition, the adaptability of the model for pharmaceutical texts from other domains requires further research and validation.

Overall, our results emphasize the importance of using pre-trained BERT models and introducing heterogeneous graph neural networks in TCM texts. These factors make our model more adapted to deal with the linguistic features and complex relationships of TCM texts, and it shows superior performance compared to traditional relational extraction models.

5. Conclusions

This paper proposes a relationship extraction method based on heterogeneous graph neural network, using heterogeneous graph structure to achieve model construction, word node and relationship node information fusion to update the node information, and a two-stage entity extraction method in the decoding stage to obtain entity attributes in the sentence and the formula entity-medication method attribute relationship group. Through training and validation on the corpus of ancient Chinese medicine texts, it is proved that the method proposed in this paper has certain advantages over other models and is applicable to the relationship extraction of ancient Chinese medicine texts.

However, in this paper, the relationship extraction model in Chinese medicine is found to rely heavily on a large amount of accurately labeled training data. Moreover, the medical rationality of the data also requires verification and validation by field experts. In future research, it would be beneficial to explore Few-shot Learning methods or even Zero-shot Learning, which can effectively learn from a smaller number of samples and possess the capability to learn quickly and efficiently from examples. By enhancing the model's ability to learn from examples, its dependency on labeled data can be reduced.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (82274680, 82160955) and the university-level research team of Jiangxi University of Traditional Chinese Medicine for the innovation team of Chinese medicine preparation technology and equipment (CXTD22006).

Conflict of interest

The authors declare there is no conflict of interest.

References

1. C. Yang, D. Xiao, Y. Luo, B. Li, X. Zhao, H. Zhang, A hybrid method based on semi-supervised learning for relation extraction in Chinese EMRs, *BMC Med. Inf. Decis. Mak.*, **22** (2022), 169–181. https://doi.org/10.1186/s12911-022-01908-4
2. Q. Hu, T. Yu, J. Li, Q. Yu, L. Zhu, Y. Gu, End-to-End syndrome differentiation of Yin deficiency and Yang deficiency in traditional Chinese medicine, *Comput. Methods Programs Biomed.*, **174** (2019), 9–15. https://doi.org/10.1016/j.cmpb.2018.10.011
3. L. Gong, J. Jiang, S. Chen, M. Qi, A syndrome differentiation model of TCM based on multi-label deep forest using biomedical text mining, *Front. Genet.*, **14** (2023). ttps://doi.org/10.3389/fgene.2023.1272016
4. T. Qi, S. Qiu, X. Shen, H. Chen, S. Yang, H. Wen, et al., KeMRE: Knowledge-enhanced medical relation extraction for Chinese medicine instructions, *J. Biomed. Inf.*, **120** (2021), 103834. https://doi.org/10.1016/j.jbi.2021.103834
5. H. Wan, M. F. Moens, W. Luyten, X. Zhou, Q. Mei, L. Liu, et al., Extracting relations from traditional Chinese medicine literature via heterogeneous entity networks, *J. Am. Med. Inf. Assoc.*, **23** (2016), 356–365. https://doi.org/10.1093/jamia/ocv092
6. X. Chen, C. Ruan, Y. Zhang, H. Chen, Heterogeneous information network based clustering for precision traditional Chinese medicine, *BMC Med. Inf. Decis. Making*, **19** (2019). https://doi.org/10.1186/s12911-019-0963-0
7. X. Liu, Y. Liu, H. Wu, Q. Guan, A tag based joint extraction model for Chinese medical text, *Comput. Biol. Chem.*, **93** (2021). https://doi.org/10.1016/j.compbiolchem.2021.107508
8. H. Chang, H. Zan, T. Guan, K. Zhang, Z. Sui, Application of cascade binary pointer tagging in joint entity and relation extraction of Chinese medical text, *Math. Biosci. Eng.*, **19** (2022), 10656–10672. https://doi.org/10.3934/mbe.2022498
9. T. Savalia, A. Shukla, R. Bapi, A unified theoretical framework for cognitive sequencing, *Front. Psychol.*, **7** (2016). https://doi.org/10.3389/fpsyg.2016.01821
10. H. Le, D. Can, N. Collier, Exploiting document graphs for inter sentence relation extraction, *Biomed. Semantics*, **13** (2022), 15. https://doi.org/10.1186/s13326-022-00267-3
11. Y. Lin, S. Shen, Z. Liu, H. Luan, M. Sun, Neural relation extraction with selective attention over instances, *Ann. Meet. Assoc. Comput. Linguist.*, (2016), 2124–2133. https://doi.org/10.18653/v1/P16-1200

12. L. Luo, Z. Yang, M. Cao, L. Wang, Y. Zhang, H. Lin, A neural network-based joint learning approach for biomedical entity and relation extraction from biomedical literature, *J. Biomed. Inf.*, **103** (2020). https://doi.org/10.1016/j.jbi.2020.103384
13. H. Zhou, Deng H, Chen L, Yang Y, Jia C, Huang D, Exploiting syntactic and semantics information for chemical-disease relation extraction, *Database*, **2016** (2016), baw048. https://doi.org/10.1093/database/baw048
14. Y. Zhang, H. Lin, Z. Yang, J. Wang, S. Zhang, Y. Sun, et al., A hybrid model based on neural networks for biomedical relation extraction, *J. Biomed. Inf.*, **81** (2018), 83–92. https://doi.org/10.1016/j.jbi.2018.03.011
15. C. Quirk, H. Poon, Distant supervision for relation extraction beyond the sentence boundary, preprint, arXiv: 1609.04873.
16. Y. Shi, Y. Xiao, P. Quan, M. Lei, L. Niu, Distant supervision relation extraction via adaptive dependency-path and additional knowledge graph supervision, *Neural Netw.*, **134** (2021), 42–53. https://doi.org/10.1016/j.neunet.2020.10.012
17. Y. Liang, F. Meng, Y. Zhang, Y. Chen, J. Xu, J. Zhou, Infusing multi-source knowledge with heterogeneous graph neural network for emotional conversation generation, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **35** (2021). https://doi.org/10.1609/aaai.v35i15.17575
18. Y. Liang, F. Meng, Y. Zhang, Y. Chen, J. Xu, J. Zhou. Emotional conversation generation with heterogeneous graph neural network, *Arti. Intell.*, **308** (2022). https://doi.org/10.1016/j.artint.2022.103714
19. X. Chu, B. Sun, Q. Huang, S. Peng, Y. Zhou, Y. Zhang, Quantitative knowledge presentation models of traditional Chinese medicine (TCM): A review, *Arti. Intell. Med.*, **103** (2020). https://doi.org/10.1016/j.artmed.2020.101810
20. X. Zhou, B. Liu, Z. Wu, Y. Feng, Integrative mining of traditional Chinese medicine literature and MEDLINE for functional gene networks, *Arti. Intell. Med.*, **41** (2007), 87–104. https://doi.org/10.1016/j.artmed.2007.07.007
21. T. Li, À. Bravo, L. Furlong, B. Good, A. Su, A crowdsourcing workflow for extracting chemical-induced disease relations from free text, *Database*, **2016** (2016). https://doi.org/10.1093/database/baw051
22. X. Yang, C. Wu, G. Nenadic, W. Wang, K. Lu, Mining a stroke knowledge graph from literature, *BMC Bioinf.*, **22** (2021). https://doi.org/10.1186/s12859-021-04502-z
23. G. Meng, Y. Huang, Q. Yu, Y. Ding, D. Wild, Y. Zhao, et al., Adopting text mining on rehabilitation therapy repositioning for stroke, *Front. Neuroinf.*, **13** (2019), 17. https://doi.org/10.3389/fninf.2019.00017
24. M. Ji, J. Zhou, N. Wei, AFR-BERT: Attention-based mechanism feature relevance fusion multimodal sentiment analysis model, *PLoS One*, **17** (2022). https://doi.org/10.1371/journal.pone.0273936
25. J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, et al., BioBERT: A pre-trained biomedical language representation model for biomedical text mining, *Bioinformatics*, **4** (2020), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682
26. H. Gong, X. You, M. Jin, Y. Meng, H. Zhang, S. Yang, et al., Graph neural network and multi-data heterogeneous networks for microbe-disease prediction, *Front. Microbiol.*, **13** (2022). https://doi.org/10.3389/fmicb.2022.1077111

27. Q. Liu, C. Long, J. Zhang, M. Xu, D. Tao, Aspect-aware graph attention network for heterogeneous information networks, *IEEE Trans. Neural Netw. Learn. Syst.*, (2022). https://doi.org/10.36227/techrxiv.19311104
28. Q. Zhao, D. Xu, J. Li, L. Zhao, F. A. Rajput, Knowledge guided distance supervision for biomedical relation extraction in Chinese electronic medical records, *Expert Syst. Appl.*, **204** (2022), 117606. https://doi.org/10.1016/j.eswa.2022.117606
29. J. Chen, W. Lin, S. Yang, M. F. Chiang, M. R. Hribar, Development of an open-source annotated glaucoma medication dataset from clinical notes in the electronic health record, *Transl. Vis. Sci. Techn.*, **11** (2022), 20. https://doi.org/10.1167/tvst.11.11.20
30. P. Kumar, B. Raman, A BERT based dual-channel explainable text emotion recognition system, *Neural Netw.*, **150 (**2022), 392–407. https://doi.org/10.1016/j.neunet.2022.03.017
31. G. Dai, X. Wang, X. Zou, C. Liu, S. Cen, MRGAT: Multi-relational graph attention network for knowledge graph completion, *Neural Netw.*, **154** (2022), 234–245. https://doi.org/10.1016/j.neunet.2022.07.014
32. T. Dai, J. Zhao, D. Li, S. Tian, X. Zhao, S. Pan, Heterogeneous deep graph convolutional network with citation relational BERT for COVID-19 inline citation recommendation, *Expert Syst. Appl.*, **213** (2023), 118841. https://doi.org/10.1016/j.eswa.2022.118841

AIMS Press