Entity Relation Extraction of Traditional Chinese Medicine Influenza Based on Bi-GRU+GBDT

Yanhua Zhao[1], Jianxun Zhang[1(✉)], and Yue Li[2]

[1] Tianjin University of Technology and Education, Tianjin, China
zhangjx@tute.edu.cn
[2] Tianjin University of Traditional Chinese Medicine, Tianjin, China

Abstract. In this paper, an algorithm based on Bi-directional Gated Recurrent Unit (Bi-GRU) and Gradient Boosting Decision Tree (GBDT) is proposed to extract the entity relationship of Traditional Chinese Medicine influenza. Firstly, the word vector is used as the input data set and the word vector is constructed by the word embedding model Word2Vec tool. Then the sentence feature is extracted by Bi-GRU and the attention mechanism is integrated to improve the accuracy of feature extraction. Finally, the feature vector is input into the GBDT algorithm for classification training and prediction to complete the Traditional Chinese Medicine influenza entity relationship extraction. In this paper, a variety of different entity relation extraction algorithms are compared with this algorithm to verify the effectiveness of the algorithm. This algorithm improves the stability of the model and effectively solves the problem of insufficient generalization ability of the model. Therefore, when studying the relationship extraction of Traditional Chinese Medicine texts, we can give priority to using Bi-GRU+GBDT model. Also through the experiment to adjust the model parameters and comparison to get the optimal parameters of Traditional Chinese Medicine influenza relationship extraction experiment.

Keywords: relationship extraction · deep learning · gated cyclic neural network · gradient lifting tree · knowledge graph

1 Introduction

Traditional Chinese Medicine is the wisdom crystallization of several generations. After continuous inheritance and improvement, it has a complete theoretical system, has a good guiding significance for clinical practice, and provides a unique method for the diagnosis and treatment of diseases. As a substitute of modern medicine, Traditional Chinese Medicine is getting more and more attention all over the world, and a large number of Traditional Chinese Medicine

Y. Cao and X. Shao (Eds.): MONAMI 2022, LNICST 474, pp. 138–153, 2023.
https://doi.org/10.1007/978-3-031-32443-7_10

research papers are published every year [1]. At the same time, large-scale analysis of a large number of Traditional Chinese Medicine literature has become an interesting research field in recent years, because such an analysis can excavate the collective knowledge of Traditional Chinese Medicine researchers and supplement the main body of medical knowledge. In addition, Traditional Chinese Medicine is a very complex medical system, involving a variety of entities, and there can be many types of complex relationships among these entities [2]. Therefore, there are many ideas, methods and clinical experiences that have not been discovered. Therefore, it is necessary to further collate and analyze medical records and literature to dig out more knowledge content to supplement and improve the existing theoretical system of Traditional Chinese Medicine. Information technology is undoubtedly the most favorable assistant to this work, so how to use modern emerging information technology to explore the knowledge of Traditional Chinese Medicine is an important direction worthy of scholars' attention. In the vertical field, the medical field is one of the most widely used knowledge graphs [3]. Dan Zhu et al. [4] constructed the knowledge graph related to fatty liver disease. In the process, they used the methods of common and individual experience analysis, analyzed and studied the clinical trials of many famous Traditional Chinese Medicine on fatty liver disease, and found out the relationship between syndrome and treatment. Tong Ruan [5] and others construct the medical knowledge graph by using the medical concepts obtained from the medical website and the relationship between them. In the study of cervical radiculopathy, Kang Li [6] and others obtained the common syndrome type and commonly used drugs for treatment of cervical radiculopathy through association rule analysis. Their research data came from China knowledge Network. Hong Wu [7] constructed the knowledge graph of "symptom-disease-prescription-Traditional Chinese Medicine" by using the entities, attributes and relationships extracted from "Compendium of Materia Medica", "Collection of typical cases in China" and Traditional Chinese Medicine diagnosis and treatment data. Taking the obstetrics and gynecology textbooks as the data source. Xuejiao Zhao [8] and others proposed to use related technologies such as knowledge extraction to construct the gynecology and obstetrics knowledge graph in order to share the common knowledge of gynecology and obstetrics medicine related to science popularization. Entity relation extraction is a key step in the process of information extraction in the construction of knowledge graph. Relationship extraction (RE) refers to the extraction of the relationship between entities, so that scattered entities can be linked through relationship extraction [9], and then knowledge storage is carried out to form a related semantic network [10]. Relationship extraction (RE) refers to the extraction of the relationship between entities from unstructured text [11], which determines the category of the relationship according to the characteristics of the entity. Because the data in the field of Traditional Chinese Medicine is of great potential value and significance to human beings, the practical value of building a knowledge graph for the field of Traditional Chinese Medicine is highlighted.

2 Related Work

According to the form of the extracted corpus, the relation extraction model includes relation extraction for sentences and relation extraction for paragraphs. The difference between the two lies in whether two related entities appear in a sentence or in a paragraph. According to Chinese grammatical habits, generally speaking, two related entities and the relationship between them can be expressed clearly in a sentence. Therefore, this paper uses the relation extraction of sentences. According to whether the relationship type is predefined or not, the relationship extraction model can be divided into schema-based relation extraction and open relation extraction. The former means that the relationship of the entity pair can only be selected from the predefined category, while the latter means that the entity has no limit to the relationship. In the construction of domain knowledge graph, schema is fixed, so the relationship type of entity pair is predefined. This paper aims at relation extraction based on schema. Knowledge graph is used to describe concepts and their related relationships in the real world. In the medical field, that is, the composition of two medical entities with semantic relations and their semantic relations, it is a good intuitive knowledge representation. Entities are the nodes in the knowledge graph network, and the relationship is the type of semantic relationship between the two entities, that is, the edge of the network connection node [12]. Relationship category is the relationship between entities, and entities are connected through relationships, thus forming a complete semantic knowledge network. In this paper, according to the text data of influenza in Traditional Chinese Medicine, five kinds of relations are defined, which are symptomatic, dialectical, treatment, use and contain. Among them, there is a symptomatic relationship between “patient” and “symptom”. The type of relationship is defined as “symptomatic” relationship, and the patient’s syndrome is distinguished according to the patient’s symptom. Therefore, there will be a dialectical relationship between “patient” and “syndrome”, and the relationship type is defined as “dialectical” relationship, and the relative treatment method is adopted according to the syndrome differentiation. There will be a therapeutic relationship between “syndrome” and “treatment”. The relationship type is defined as “treatment” relationship, and what prescription will be used to the patient after confirming the treatment method. Therefore, there will be a use relationship between “patient” and “prescription”. The relationship type is defined as “use” relationship, according to what kind of Traditional Chinese Medicine is needed according to the prescription used, therefore, there will be an inclusive relationship between “prescription” and “Traditional Chinese Medicine”. The relationship type is defined as a “contain” relationship. First of all, the entity pairs in the sentence are identified, and then the category of the relationship between the entity pairs is marked manually. A total of 1210 sample data are selected for entity relationship tagging, and the labeled data are divided into training set and test set according to the 80% and 20% standards. The distribution ratio accords with the common proportion of the data.

3 Design of Entity Relation Extraction Algorithm of Traditional Chinese Medicine Influenza Based on Bi-GRU+GBDT

There are two main methods used in entity relationship extraction, each of which has its own advantages and disadvantages. The first method is to use remote supervision to obtain training data, so that the labeling work can be reduced, but if the previous entity recognition errors will be passed on to the relationship extraction work. The second method is to combine the two tasks of entity recognition and relationship extraction between entities, so that the two can be more fully integrated, and the entity information can be fully utilized. However, the accuracy of this model is only 64%. When the accuracy of entity recognition is more than 80%, the first method is better than the second method. In the previous work, the entity recognition accuracy is much higher than 80%. The first method can be determined.

Entity relation extraction models can be divided into three categories [13]: pattern-based methods [14,15], statistical machine learning [16] and neural networks. The method based on pattern, that is, the traditional rule-based method, the rules need to be designed in advance, and the quality of the rule design determines the quality of the subsequent relationship extraction task. If the rule design is not good, it will not achieve the desired effect, and the waste of time has no effect. Statistical machine learning methods need to spend a lot of time and energy to extract relational features, which is also a very arduous task. Therefore, in comparison, in-depth learning in the field of neural network research can well complete the task of relationship extraction.

3.1 Framework of Bi-GRU+GBDT Entity Relationship Extraction Algorithm

Recurrent Neural Network (RNN) is a kind of artificial neural network and one of the representative algorithms of deep learning. It is a kind of neural network model with memory ability [17]. The advantage of cyclic neural network in relation extraction is that it can be used to extract long-distance dependent information in sentences, but it also has some disadvantages, such as easy to fall into gradient explosion [18]. The emergence of Gated Recurrent Neural Network solves this problem, and its principle is to control the flow of information through the door that can be learned [18]. The threshold Gated Recurrent Unit (GRU) used in this paper is one of the categories of gated cyclic neural networks, which is essentially similar to LSTM. Both belong to RNN.GRU, which is a variant of LSTM, and can be regarded as a LSTM without input gates, that is, from three gate functions of LSTM to two gate functions, including update gate and reset gate. Because there is no input gate, it writes all the contents of the memory unit to the overall network [18] at each time step, as shown in Fig. 1. Where r_t represents a reset door that determines the extent to which previous information is forgotten, and z_t represents an update door, which determines what information needs to be added and forgotten.

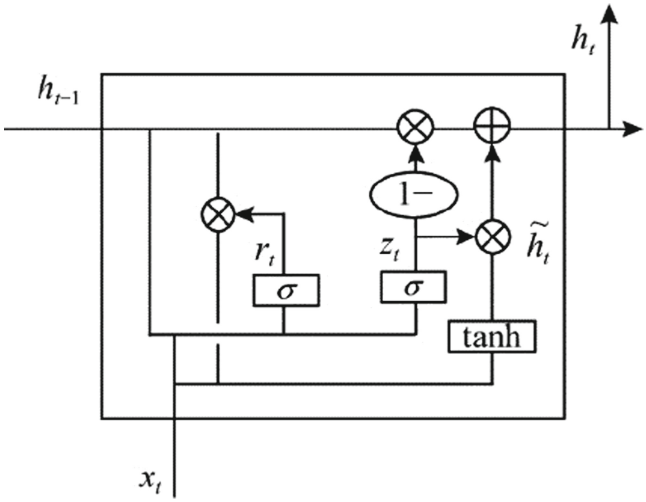

Fig. 1. GRU model structure.

$$\begin{cases} z_t = \sigma(W_z * [h_{t-1}, x_t]) \\ r_t = \sigma(W_r * [h_{t-1}, x_t]) \\ \widetilde{h_t} = tanh(W_{h_t} * [r_t * h_{t-1}, x_t]) \\ h_t = (1 - z_t) * h_{t-1} + z_t * \widetilde{h_t} \end{cases} \tag{1}$$

The problem of entity relationship extraction is usually regarded as multi-classification, and softmax is often used as the classifier in the deep learning model of relationship extraction, but the generalization ability of this classifier is insufficient, so the effect of relationship extraction is not very good. The Gradient Boosting Decision Tree (GBDT) is an integrated classifier, and the GBDT model can automatically find features and combine them effectively [19]. Gradient lifting decision tree is a lifting tree model based on CART regression tree model. Its core idea is to use negative gradient approximation to simulate residuals. In the process of generating each tree, the residual of the previous tree is calculated, and the next tree is fitted on the basis of the residual, so that the residual obtained on the next tree is reduced [20]. GBDT can combine and discretize features automatically. After the establishment of the decision tree, the path from the root node to each leaf node is a combination of different features, and each leaf node represents a unique feature combination. The lifting tree model can be expressed as an additive model of the decision tree:

$$f_M(x) = \sum_{m=1}^{M} T_m(x) \tag{2}$$

where: the decision T_m (x) represents the $m - th$ decision tree, and M represents the number of trees. The loss function is determined by negative gradient approximation, and the GBDT loss function is defined as L(y, f)

$$L(y, f) = \sum_{i=1}^{m} L(y_i, f(x_i)) \tag{3}$$

where: the loss function $L(y_i, f(x_i)$ represents the gradient to the tree $f(x_i)$. The process of gradient lifting tree algorithm (GBDT) is as follows: step1: During initialization, c is taken as the mean value of all the training samples, and the initial learner is obtained.

$$f_0(x) = c \tag{4}$$

step2: Iterative training m = 1, 2 ... N. Take the residual r_{mi} of the previous step as the new value of the sample, take the data as the training data of the next tree, get a new regression tree, its corresponding leaf node region is R_{mj}, and calculate the best fitting value is c_{mj}. j = 1, 2 ... J. The regression tree represented by J is the number of leaf nodes, and then update the learner:

$$r_{mi} = -[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}]_{f(x)=f_{m-1}(x)} \tag{5}$$

$$c_{mj} = arg \min_{C} \sum_{x \in R_{mj}} L(y_i, f_{m-1}(x_i) + c) \tag{6}$$

step3: Get the final learner GBDT:

$$\widehat{f}(x) = f_M(x) = f_M(x) + \sum_{m=1}^{M} \sum_{j=1}^{J} c_{mj} I(x \in R_{mj}) \tag{7}$$

where: I is the indicator function. If $x \in R_m j$, then I is 1, otherwise I is 0.

The Bi-GRU+GBDT entity relationship extraction model is mainly divided into the following parts, including input embedding layer, Bi-GRU layer, attention mechanism layer and output layer.

(1) Input embedding layer
The input embedding layer serves as the input layer of the subsequent Bi-GRU layer. It represents the word vector of each sentence in the corpus, and then trains it by embedding words into the Word2vec to form the required input matrix of the subsequent model for use by the subsequent model.

(2) The Bi-GRU layer
The GRU model selected in this paper has relatively few hyper-parameters, and the structure is simpler, so it is easier to train than other cyclic neural network models. In order to enable the neural network to learn both forward sequence information and reverse sequence information, the bi-directional GRU model is selected for training (Fig. 2).

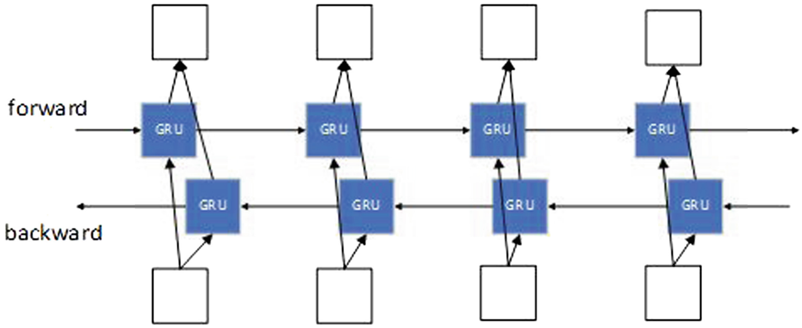

Fig. 2. Bi-directional GRU model.

(3) The attention mechanism layer The traditional RNN neural network model is not very effective in the overall sentence extraction, and the ability to learn context information is very poor, so in order to better learn the semantic information of the context, this paper adds the attention mechanism layer after the sentence feature extraction. It will focus on the characteristics of relational categories, and will consider the importance of key words or words to relational features. By assigning weights to different features, the forward and reverse sentence feature vectors transferred from the Bi-GRU layer can be calculated by stitching and weighting. The attention mechanism layer structure diagram is shown in Fig. 3.
a_{di}: The attention probability of the ith word to the dth word.

$$h_{di} = U_a * \tan(U_b * h_d + U_c * h_i + b) \tag{8}$$

$$a_{di} = \frac{exp(h_{di})}{\sum_{j=1}^{T} exp(h_{dj})} \tag{9}$$

Among them: U_a, U_b, U_c represent the weight matrix, h_d represents the forward output of the Bi-GRU layer, h_i represents the backward output, b represents the bias vector, and T represents the length of the sentence. After passing through the attention mechanism layer, the new output feature vector is H_t. As shown in formula (10).

$$H_t = \sum_{i=1}^{m} a_{di} * h_m \tag{10}$$

(4) The output layer inputs The feature vectors of sentences passing through the attention mechanism layer into the GBDT algorithm. This paper uses gradient lifting to train and predict classification, iterates to build a decision tree, and finally obtains the relationship categories between entities and entities contained in each sentence. The model architecture is shown in Fig. 4.

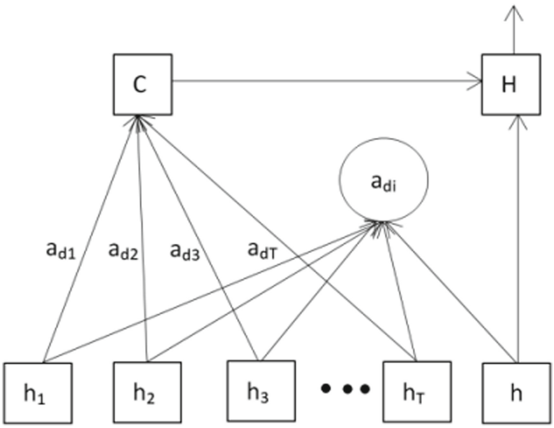

Fig. 3. Attention mechanism model.

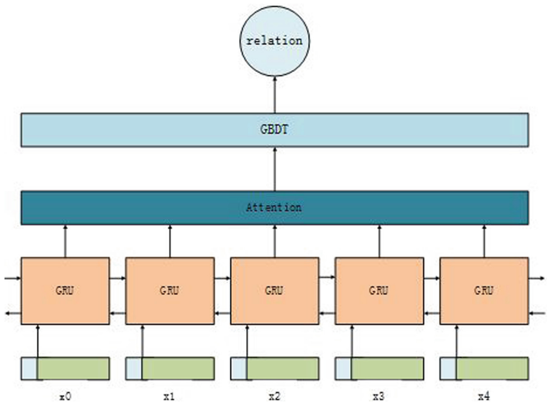

Fig. 4. Bi-GRU+GBDT model structure diagram.

3.2 Bi-GRU+GBDT Entity Relationship Extraction Algorithm Flow

Traditional Chinese Medicine influenza entity relationship extraction algorithm is mainly divided into six steps: acquisition of corpus data, pre-processing of corpus data, generation of word vector training model, training and learning relationship extraction model, test model, relationship category extraction. Using Word2vec to embed words into the training set samples to form a vector matrix as the input of the model, after the Bi-GRU model has forward output and backward output, the forward and reverse sentence feature vectors transferred from the Bi-GRU layer are spliced and weighted through the attention mechanism layer, and the sentence feature vectors are calculated. Finally, the final relation category is obtained by using the GBDT algorithm in the output layer. Among them, after obtaining the corpus data, a pair of related entities are extracted from the sentence, and the category of the relationship between the two entities in the sentence is tagged. There are 1210 sample data, and five relation categories are set in advance. For the specific category description, see the definition of the relationship category in the above work. The tagging format of each sample is (entity 1, entity 2, relationship, the sentence). In addition, the labeled data is divided into training set and test set according to the criteria of 80% and 20%, and the distribution proportion accords with the common proportion of the data. The flow of the algorithm is shown in Fig. 5.

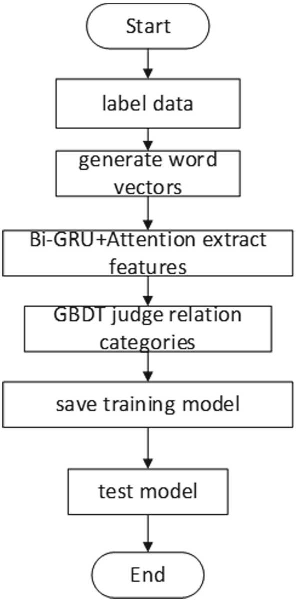

Fig. 5. Algorithm flow chart.

4 Experiment and Result Analysis

In order to verify the effectiveness of this model, we use a variety of relational extraction models as comparative experiments, that is, the results of other models are compared with the experimental results of entity recognition of this model. Among them, the model comparison experiments include Bi-GRU model, GBDT model, Bi-GRU+GBDT model, and the m value of gradient lifting tree GBDT, that is, the optimization of the number of trees and the corresponding performance of the model.

4.1 The Source of Experimental Data

The data comes from Traditional Chinese Medicine ancient books such as "A hundred Clinical Books of Traditional Chinese Medicine in the past 100 years in China". Because these ancient books only contain PDF files and are scanned into picture format, these documents need to be processed regularly, and the text data about Traditional Chinese Medicine influenza can be extracted by OCR technology. According to the definition of the relationship between entities in the Traditional Chinese Medicine Language System (TCMLS), this paper defines the relationship among symptoms, dialectics and treatment of cases in ancient books of Traditional Chinese Medicine, and forms the SPO triple form of "entity 1, relationship, entity 2" [21]. At present, the triple in the sentences in the open domain data set is an one-to-one relationship. According to the writing format of medical records in the field of Traditional Chinese Medicine in real life, a sentence generally includes multiple entities, as well as multiple relationships, that is an one-to-many triple relationship. From this point of view, the corpus of the relationship extraction experiment in this paper is not limited to the one-to-one relationship of a sentence, but contains multiple relationships between multiple entities in a sentence [21]. In this experiment, a total of 1210 sentences are prepared, and then the entity pairs are identified and marked, and the labeled data are divided into training set and test set, accounting for 80% and 20% respectively.

4.2 The Experimental Instructions and Parameter Settings Preprocess the Data

After obtaining the corpus, then use the word embedding model Word2vec to construct the word vector and train it, using the word vector as the input of the bidirectional GRU [22], then extract the features through the bidirectional GRU and merge the coding sequences of the two directions, and then get the complete features of the sentence, and then pay attention to the features of the relation category through the attention mechanism layer. Assign weights to different features, calculate the output features and relationship category labels, and finally build a gradient lifting tree according to the output of the attention mechanism layer, and judge the final relationship category through model training and iteration.

The experiment is carried out in the Windows10 operating system based on Pytorch framework, and the experimental environment is shown in Table 1. The experimental steps and the hyper-parameter related settings of the model are as follows: first, the first word embedding layer trains the word vector through the Word2vec tool, and reads out the corpus that needs to be analyzed from the file. The dimension of the word vector is set to 128D. The value of this dimension is generally related to the size of the corpus. If it is a larger corpus, you can increase the dimension as needed. The default value of window is usually 5. In practical use, the size of the window can be adjusted dynamically according to the actual demand. The parameter min_count refers to the minimum word

Table 1. Configuration table of experimental environment.

Environment	configuration
memory	64G
CPU	Intel(R)Core(TM) i7-6700
operating system	Windows 10 64-bit
programming language	python 3.7
Deep Learning Framework	Torch 1.10.0
Compiler	PyCharm Community Edition

frequency of the word vector that needs to be calculated. This value can remove some obscure low-frequency words [23]. Because it contains more professional words, it can be set to 1, and the value can be adjusted according to the size of the corpus. Then, the word vector is read out from the pre-trained word vector model, which is used as the input data of the subsequent network model, and the data is divided into two parts: the training set and the verification set. Each word of the sentence in the sample of the training set has a labeled relational category tag, the words are converted into word vectors, and the sentences are input into the model in batches for training. The size of the batch_size can be adjusted according to the size of the corpus. Here, batch_size is set to 64, which means that 64 sentences are entered into the network model at a time, and the hidden layer in the network model maps the sentence features of the data to the high-dimensional space, and the next parameter neuron is used to further divide the features in the high-dimensional space, using linear division, so the more neurons are needed to achieve a high-precision model. With the structure of the network model will be more complex, and easily lead to over-fitting, to sum up, set the number of neurons to 128, before the model starts training, set a number of iterations, that is, the number of traversing samples. After each iteration, re-traverse the sample and continue to iterate until the specified number of iterations is reached. After the training, the model is tested, the test set is input into the model, and the accuracy of the model is tested [9]. The specific hyper-parameter settings are shown in Table 2.

4.3 Experimental Evaluation Criteria

In this paper, three models of Bi-GRU, GBDT and Bi-GRU+GBDT are added as the control group, and the three models are constructed by Word2vec [24]. In order to show the experimental results comprehensively and truly, when evaluating the model, this chapter mainly uses the following three factors as the evaluation index of the model: accuracy, recall and F value [25]. The accuracy is expressed by P, which refers to Precision, and the recall rate is expressed by R, which refers to the Recall. Accuracy refers to all the predicted results, the number of real samples in the predicted positive samples, that is, the proportion of the total number of results consistent with the actual results [26]. In popular

Table 2. Parameter related settings in the model.

Parameter name	Parameter setting
batch_size	64
learn_rate	0.003
epoches	60
min_count	1
window	5
emb_size	128
hidden_size	128
optimizer	Adam

terms, it is find the right. Recall rate refers to the number of predicted correct samples in the positive sample, that is, the proportion of the number of entities identified to all entities [27]. In reality, the accuracy and recall rate may contradict each other, at this time, the F value is needed to balance the two indicators. The F value is a combination of accuracy and recall rate to evaluate the model as a whole [24]. The relevant formulas for accuracy, recall and F value are shown in formula (3). The TP in the formula indicates that the identified entity is also the desired entity, that is, the number of identified entities, that is, the number of identified entities that are not relevant, that is, the number of useless entities, that is, the number of wrong entities identified, and TP+FP represents the total number of entities identified, whether useful or not. FN represents the number of unrecognized entities but in fact they are the number of entities needed, that is, the number of unrecognized but related entities [28], and TP+FN represents the total number of entities needed for manual labeling.

$$\begin{cases} P = \frac{TP}{TP+FP} * 100\% \\ R = \frac{TP}{TP+FN} * 100\% \\ F = \frac{P*R*2}{P+R} * 100\% \end{cases} \tag{11}$$

4.4 Experimental Results and Analysis

Two model methods are designed and compared with this method, and the model training is carried out according to the parameters set in Sect. 3.2, and then in order to verify the effectiveness of the method used in Traditional Chinese Medicine influenza entity relationship extraction, the model is tested with the test set, and the parameter values of these groups are consistent. Table 3 and Fig. 6 are the final comprehensive comparison results of different entity recognition model experiments.

Table 3 and Fig. 6 show the final experimental results of the three relationship extraction models Bi-GRU, GBDT and Bi-GRU in Traditional Chinese Medicine influenza relationship extraction, and the evaluation indicators are the three elements mentioned above. First of all, comparing the recognition results of GBDT

Table 3. Comparison of final results of different models.

Model	Precision	Recall	F value
GBDT	80.47%	80.95%	80.71%
Bi-GRU	84.26%	84.03%	84.14%
Bi-GRU+GBDT	88.98%	88.59%	88.77%

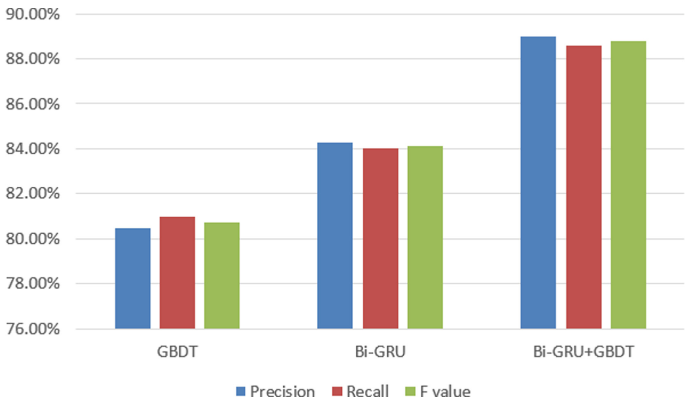

Fig. 6. Comparison of final results of different models.

and Bi-GRU, the result of Bi-GRU is better than that of GBDT model, because the ability of Bi-GRU model to automatically extract the deep relational features of sentences is stronger, so we can get better experimental results. Compared with Bi-GRU and GBDT, the effect of this model Bi-GRU+GBDT is improved, and GBDT is added as the classifier of the model on the basis of Bi-GRU. Because GBDT can improve the generalization ability of entity relationship recognition, thus improving the accuracy of relationship extraction. The experimental results of precision recognition of the model used in this paper in relation extraction are shown in Table 4.

Table 4. Bi-GRU+GBDT relation extraction experiment result table.

Model	Precision	Recall	F value
Symptomatic	89.36%	85.96%	87.63%
Dialectical	82.24%	85.48%	83.83%
Treatment	88.56%	89.33%	88.94%
Use	92.57%	90.61%	91.58%
Contain	92.16%	91.59%	91.87%
Synthesis	88.98%	88.59%	88.77%

From the test results, it can be seen that the better recognition effects of relation categories in relation extraction are “contain” and “use”, while those with poor recognition results are “dialectics” and “therapy”. The reason for this phenomenon is that there are fewer sentences related to these two kinds of relations in the data set for model training and learning, so the model recognition effect of this kind of relationship is worse than that of other categories. There is another part of the reason, when labeling data, there may be incomplete labeling, and some errors will lead to poor results.

In order to explore the effect of the number of decision trees m on the model, this paper sets the initial value to 10 when adjusting the number of gradient lifting trees m, and increases continuously with 10 as a unit, which is divided into 8 tests to test the effect of the model. The test results are shown in Table 5 and Fig. 7.

As can be seen from Table 5 and Fig. 7, m increases continuously from the initial value 10, and with the continuous increase of m, the corresponding F

Table 5. Comparison of experimental results of value adjustment.

m	Precision	Recall	F value
10	62.39%	62.73%	62.56%
20	73.66%	73.69%	73.67%
30	79.34%	78.92%	79.13%
40	83.79%	83.52%	83.65%
50	85.45%	85.39%	85.42%
60	88.98%	88.59%	88.77%
70	87.36%	87.51%	87.43%
80	87.34%	86.81%	87.07%

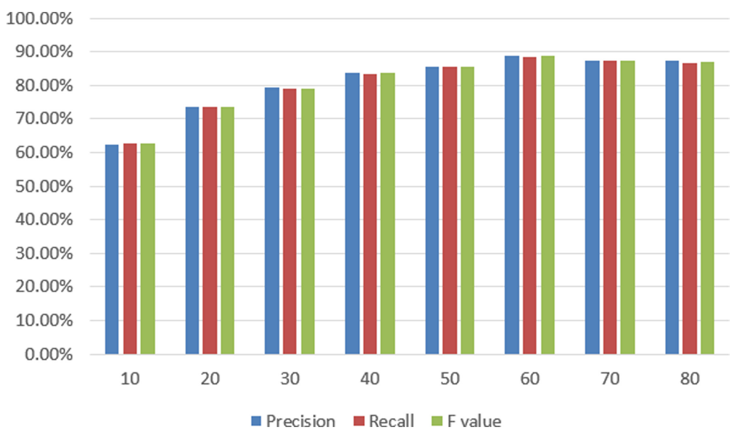

Fig. 7. Comparison of final results of different models.

value of the model increases gradually. When the m value reaches 60, the F value reaches the maximum, and the model effect is the best. When the m value continues to increase, the F value shows a small downward trend, but tends to be stable, so the model effect is the best when the number of gradient lifting trees m is 60.

5 Concluding Remarks

This paper mainly introduces the entity relation extraction of Traditional Chinese Medicine influenza based on Bi-GRU+GBDT algorithm, in which the Bi-GRU model has strong ability to automatically extract the deep relational features of sentences, and the traditional deep learning model is prone to lack of generalization ability in relation extraction, and GBDT just makes up for this deficiency, so adding GBDT on the basis of Bi-GRU model can improve the accuracy of the model. The three models of Bi-GRU, GBDT and Bi-GRU+GBDT are used for training, and then the effect of the model is evaluated by three evaluation indexes. Finally, the relationship recognition effects of the three models are compared, which verifies the effectiveness of this model in relation extraction compared with other models. Therefore, in the study of relationship extraction of Traditional Chinese Medicine texts, Bi-GRU+GBDT model can be given priority. In addition, through a large number of experiments to find the best performance of the model, the m value of the gradient lifting tree GBDT, that is, the number of trees, to provide experience for the future text relationship extraction of Traditional Chinese Medicine.

References

1. Wan, H.Y., Moens, M.F., Luyten, W., et al.: Extracting relations from Traditional Chinese Medicine literature via heterogeneous entity networks. J. Am. Med. Inf. Assoc. **23**(2), 356–365 (2015)
2. Liu, L.: Research on intelligent diagnosis of disease based on medical knowledge graph. Hunan University, Hunan (2018)
3. Xiu, X.: Research on the construction of tumor knowledge map based on Chinese electronic medical records. Peking Union Medical College (2019)
4. Zhu, D.: Study on the law of syndrome and treatment of fatty liver disease treated by famous Traditional Chinese Medicine and the construction of knowledge graph. Chinese Academy of Traditional Chinese Medicine (2019)
5. Ruan, T., Wang, H.: Construction of medical and health semantic knowledge base based on ontology. China Inf. Soc. (e-Health) (06), 50–51 (2014)
6. Li, K., Cui, K., Ao, F., Tang, X., Wang, Y., Li, Y.: Study on clinical application of Traditional Chinese Medicine in cervical radiculopathy. Pract. Comb. Tradit. Chin. Western Med. **19**(09), 7–11 (2019)
7. Wu, H.: Study on the Construction and Application of knowledge Graph of Traditional Chinese Medicine based on Compendium of Materia Medica. Zhengzhou University (2020)
8. Zhao, X.: Research and implementation of knowledge graph construction in obstetrics and gynecology. Chin. Digit. Med. **14**(01), 3–5 (2019)

9. Zhang, S.: Design and implementation of Intelligent question answering system for Coal Mine Industry. Hebei Engineering and Technology (2020)
10. Yuji, Y., Bin, X., Hu, J., Tong, M., Zhang, P., Zheng, L.: An accurate and efficient method for constructing domain knowledge graph. J. Softw. **29**(10), 2931–2947 (2018)
11. Di, Y.: Literature information extraction based on deep learning and its application in brain connection research. Huazhong University of Science and Technology (2020)
12. Qu, Q.: Research on Treatise on febrile Diseases based on Natural language processing. Anhui University of Traditional Chinese Medicine (2021)
13. Xu, H., et al.: More data, more relations, more context and more openness: a review and outlook for relation extraction. arXiv (2020)
14. Meng, J., et al.: MetaPAD: meta pattern discovery from massive text corpora. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 877–886 (2017)
15. Zheng, S., Yu, P., Chen, L., Huang, L., Xu, W.: DIAG-NRE: a deep pattern diagnosis framework for distant supervision neural relation extraction, pp. 1419–1429 (2018)
16. Pawar, S., Palshikar, G.K., Bhattacharyya, P.: Relation extraction: a survey. arXiv (2017)
17. Du, S., Yu, H., Zhang, H.: Research progress of text classification based on deep learning. J. Netw. Inf. Secur. **6**(04), 1–13 (2020)
18. Song, Z., Yan, R.: Chinese text emotion classification model based on CNN-BIGRU. Comput. Technol. Dev. **30**(02), 166–170 (2020)
19. Gong, J.: Research and application of click rate prediction algorithm based on federated learning. Beijing Jiaotong University (2021)
20. Zhong, J.: Research on credit card fraud recognition based on feature combination of GBDT. Lanzhou University (2020)
21. Xie, X.: Study on the construction technology of disease knowledge graph for orthopedic consultation of Traditional Chinese Medicine. Kunming University of Science and Technology (2019)
22. Li, P.: Research on emotion Classification and Application of Stock investors based on Multi-core convolution Neural Network. Hangzhou University of Electronic Science and Technology (2020)
23. Lv, R.: Research and implementation of multi-tag text classification based on deep learning. Southeast University (2018)
24. Qu, Q., Kan, H.: Named entity recognition of Traditional Chinese Medicine text based on BERT-Bi-LSTM+CRF. Electron. Des. Process **29**(19), 40–43, 48 (2021)
25. Tang, L.: Information extraction method based on domain knowledge graph and its application in medical text. Southwest Jiaotong University (2020)
26. Meng, C., Dong, Y.: Text emotion analysis based on decomposed convolution neural network. Comput. Digit. Eng. **47**(8), 1970–1973, 2101 (2019)
27. Li, L.: Automatic bug allocation based on active learning. Dalian University of Technology (2013)
28. He, Y., Liu, S., Qian, L., Zhou, G.: Disease name recognition based on syntactic and semantic features. Chin. Sci. Inf. Sci. **48**(11), 1546–1557 (2018)