Neural Computing and Applications (2022) 34:2739–2748
https://doi.org/10.1007/s00521-021-05897-9

S.I: COGNITIVE-INSPIRED COMPUTING AND APPLICATIONS

Traditional Chinese medicine entity relation extraction based on CNN with segment attention

Tian Bai[1,3] · Haotian Guan[1,3] · Shang Wang[2,3] · Ye Wang[1,3] · Lan Huang[1,3]

Received: 1 December 2020 / Accepted: 3 March 2021 / Published online: 20 March 2021

Abstract
Extracting medical entity relations from Traditional Chinese Medicine (TCM) related article is crucial to connect domain knowledge between TCM with modern medicine. Herb accounts for the majority of Traditional Chinese Medicine, so our work mainly focuses on herb. The problem would be effectively solved by extracting herb-related entity relations from PubMed literature. In order to realize the entity relation mining, we propose a novel deep-learning model with improved layers without manual feature engineering. We design a new segment attention mechanism based on Convolutional Neural Network, which enables extracting local semantic features through word embedding. Then we classify the relations by connecting different embedding features. We first test this method on the Chemical-Induced Disease task and the experiment show better result comparing to other state-of-the-art deep learning methods. Further, we apply this method to a herbal-related data set (Herbal-Disease and Herbal Chemistry, HD-HC) constructed from PubMed to explore entity relation classification. The experiment shows superior results than other baseline methods.

Keywords Traditional Chinese medicine · Relation extraction · Convolutional neural network · Segment attention mechanism

1 Introduction

TCM (Traditional Chinese Medicine) has attracted worldwide attention as an alternative to modern medicine [1]. Herb is a specific term in the field of Traditional Chinese Medicine, which is mainly composed of herbal medicine (root, stem, leaf, fruit), animal medicine (viscera, skin, bone, organs, etc.) and mineral medicine. And herbal medicine accounts for the majority of Traditional Chinese Medicine, so our work mainly focuses on herb. Entity relation mining in this domain has become an important research topic [2]. Wu et al. [3] were the first to apply relation extraction to TCM for the purpose of connecting TCM with modern life sciences [4]. There are already many effective methods for extracting the relation of biomedical entities, but related works on herb relation extraction from the PubMed literature are scarce. Therefore, this paper studies the problem of herb-related entity relations mining from PubMed literature and tries to propose an effective solution. It aims to solve the knowledge isolation problem in TCM knowledge bases [5] by

✉ Lan Huang
huanglan@jlu.edu.cn

Tian Bai
baitian@jlu.edu.cn

Haotian Guan
guanht20@mails.jlu.edu.cn

Shang Wang
wangshang17@mails.jlu.edu.cn

Ye Wang
wangye_15@mails.jlu.edu.cn

[1] College of Computer Science and Technology, Jilin University, Changchun 130012, China

[2] College of Software, Jilin University, Changchun 130012, China

[3] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China

extracting herbal-related entity relationships through the methods we proposed.

Entity relation extraction from text has been widely studied in biomedicine [6, 7]. The entity relations include relations such as disease-specific, drug-protein and chemical-protein. The implementation method has progressed from traditional rule-based or co-occurrence-based approaches to machine learning methods such as Support Vector Machine (SVM) and logistic regression. To avoid fussy feature engineering, deep learning technology has been applied to the relation extraction task with superior results. But relevant works are seldom adapted to TCM area.

Currently, the limitation of studies concerning entity relations extraction in TCM related can be put into two categories. First, language gap exists between TCM practices and researches. It is widely seen that most studies are based on Chinese corpora. But in recent years the scientific research literature related to TCM has been published in English. Such corpus limitations typically cause TCM knowledge to be isolated from modern medical knowledge. Second, accuracy improvement is labile among various implementation approaches. Relation mining in TCM was initially based on rules and co-occurrence methods. Machine learning methods are subsequently introduced into this field and achieve good results. However, the classical machine learning algorithms improve the accuracy of relation extraction primarily through complicated manual feature engineering. It, therefore, constrains the potential of accuracy improvement for classical machine learning. In contrast, deep learning methods can eliminate the need for engineered features, but few studies have yet applied this method to extract entity relation of herb-related from PubMed literature.

In order to extract Chinese medicine-specific entity relations more reliably and efficiently from PubMed literature, we propose a novel architecture with an improved layer for entity relation extraction. The methods we proposed consist of two parts: the first part implements a Convolutional Neural Network with a SEGment ATTention Mechanism (SEGATT-CNN) to extract word-level features represented by word2vec. The second part connects the different embedding features to achieve the final relation classification using a machine learning classifier.

To test the performance of our method, two data sets are used for evaluation. First, we test relevant deep learning methods on Chemical-Induced Disease (CID) data set [8], and then we test our propose methods along with state-of-the-art model. Our method obtains a better result. Second, we further apply this method to solve the problem of herbal-related relation extraction. We constructed HD-HC data set from PubMed abstracts to test the performance of the method in this paper. The results show that our model has better performance than the baseline method. Especially, it well adapts to the unbalanced positive and negative samples.

Therefore, the main contribution of this paper can be summarized as follow. First, we combine the CNN model to propose a novel deep learning method with segment attention mechanism (SEGATT-CNN), one of our contributions is a new input representation especially designed combined with a word-level attention mechanism. Second, we use the high-order feature tensor output by the SEGATT-CNN model and the word embedding features encoded by TF-IDF to act on the relation classification layer to improve the generalization of overall model. Third, we test the proposed method in Chemical-Induced Disease data set and evaluate our model along with other current models on this data set. The methods we proposed achieve reliable and effective performance. Fourth, based on the obtained PubMed corpus, two types of "coarse-grained" entity relations of herb-disease and herb-chemical are defined. This corpus was constructed with the assistance of domain experts. We solve the problem of extracting herbal-related entity relations and comprehensively verify the effectiveness of the proposed method.

2 Related work

Relation extraction, which is an important aspect of information extraction [8], has attracted extensive attention in Natural Language Processing (NLP). The extracted relationships are often applied to deeper research subjects [9]. There are many developed technologies for biomedical relation extraction, for example, considerable work has been constructed to solve the Chemical-Induced Disease (CID) problem. However, due to the lack of a publicly annotated corpus for TCM, the extraction of specific types of relations requires further exploration through deep learning techniques.

In recent research regarding biomedical entity relation extraction, Gu et al. [10] addressed the CID task based on intra-sentence levels by constructing a large number of engineered features and using a traditional machine learning model (Maximum Entropy, ME), which achieved a good classification result. In a later paper, Gu et al. [11] applied a CNN to solve the CID task and also achieved good results. Based on the CNN model, Zhou et al. [12] conducted an exploratory study by integrating word POS, head and SDP-seq features to solve this problem. Li et al. [13] used a CNN with an attention mechanism to extract CID relations. Later, Li et al. [14] proposed a new recurrent Piecewise Convolutional Neural Network (RPCNN) model to solve the above problems by composite RNN and CNN that achieved state-of-the-art results. Deep learning

technology, which achieves superior performance with only employing word embedding, can effectively replace feature engineering [15–18].

In TCM, relevant works concerning entity relation extraction from the literature are scarce. The extraction methods primarily include rule-based methods, co-occurrence methods, and traditional machine learning algorithms. In a recent study, Wan et al. [19] used a factor graph model to explore the relationships among herbs, prescriptions, symptoms and diseases in a Chinese corpus and compared the results with a baseline SVM. Wang et al. [20] discussed the relation extraction problem of effect relations and conditional effect relations in TCM publications and compared the effects of applying rule-based and feature-based relation classification models from this aspect. Yang et al. [21] used an SVM classifier combined with syntactic features to extract the relations between prescriptions and diseases, and this model achieved good results.

In general, current studies usually use self-constructed Chinese corpora to conduct entity relation mining in the field of TCM. They ignore the fact that more and more herb-related researches are published in English, and PubMed actually has rich resources on both TCM and modern western medicine. However, few studies have extracted herb-related entity relation by using deep learning method from PubMed. And text classification based on convolutional neural networks (CNN) has got more attention recently [22, 23]. This article aims to fill research gaps in this field by studying entity relationship extraction issues related to Chinese herbal (such as herbs-diseases and herbs-chemicals). We propose a novel deep learning method for relation extraction based on the dataset constructed by PubMed. Further, we compare the proposed method with the traditional machine learning and neural network algorithms to test whether it is reliable in solving this problem.

3 Method

In this section, we describe the model with the improved layer in detail. Sect. 3.1 describes a new input method based on entity relation classification, as shown in Fig. 1a-a1. Sect. 3.2 describes the SEGATT-CNN model, as shown in Fig. 1b-b1. The method for concatenating the text features of different embedding for relation classification is described in Sect. 3.3, as shown in Fig. 1c. Figure 1 shows a flow chart of our model and its different layers.

3.1 Input layer

One of our contributions is design of a new input for entity relation classification. The contexts are split into five disjoint regions based on the two relation arguments: the left context, the left entity1, the middle context, the right entity2 and the right context. The middle context often contains the most relevant relation information, but the information carried by the left and right contexts cannot be ignored completely. Therefore, we want to focus attention on the semantic information associated with the two related entities. Inspired by [24], we designed three context information combinations: (1) a combination of the left context, entity_1, the middle context and the entity_2 as x_1; (2) a combination of entity_1, the middle context and entity_2 as x_2; (3) a combination of entity_1, the middle context, entity_2 and the right context as x_3. Each part includes two entities as input to capture additional semantic characteristics that can be to distinguish the entity relation type. The extended middle context can force the attention mechanism and the CNN model to pay special attention to it. For specific operations in this input layer, 'padding' is used to supplement the text length after segmentation. Figure 2 shows how the inputs are combined.

Figure 2 describes this procedure. It shows an example sentence: 'in Uighur traditional medicine, < HB > euphorbia humifusa wild < HE > is used to treat < DB > fungal diseases < DE > , and recent studies suggest that it is the EA content which is responsible for its therapeutic effect.' It is easy to see that the entities ' < HB > Euphorbia Wild < HE > ' and ' < DB > Mycosis < DE > ' divide the text into five separate parts. A combination of the left context, entity1, the middle context and the entity2 as the left input. Similarly, the $\times$ 2 and $\times$ 3 described above are the middle input and the right input. The middle context directly describes the relation of the two entities, while the right context is a detailed interpretation of that relation. Taking the whole context as input, the key words in the middle part can be given increased attention through the attention mechanism. However, it is easy to lose important words during feature extraction because of the max-pooling layer of the convolution neural network. Therefore, the context is extended using segmented input to ensure that the important feature is not lost.

In the experiment, we use abstracts most of which come from PubMed as training set and use word2vec to train word embedding. Using the above methods as input, each input can be expressed as: $x_i \in T \times k$ (i = 1,2,3) which expresses the ith segment by its length (T, the number of words) and the k dimensions of the word embedding.

3.2 SEGATT-CNN for feature extraction

In this section, we propose a new model for local feature extraction called SEGATT-CNN model. Usually, adding the attention mechanism [25] in the input layer measures the importance of each word based on the whole context as

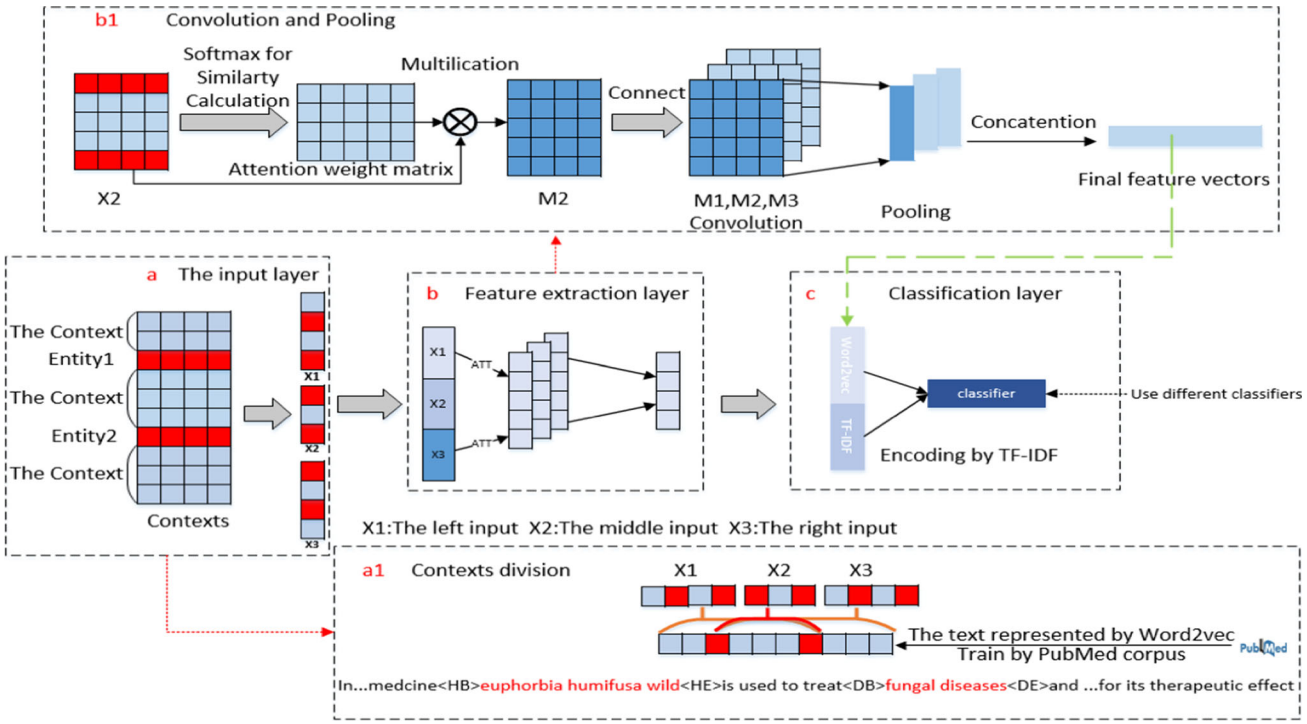

Fig. 1 The architecture of our model with different layers: **a** and **a1** describe the input layer construction process, **b** and **b1** show the details of processing with SEGATT-CNN, and **c** shows the feature connections used for classification

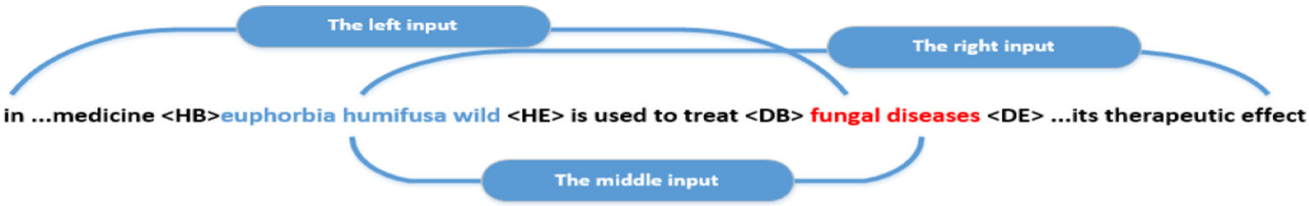

Fig. 2 Examples of three contexts combining information

input. This approach assigns less weight to words with less sensitivity for relation classification. However, we change this method and used the attention mechanism to focus on more segmented domain local information features. Next, we will introduce the operations of the various layers. Figure 3 shows the details of the process in each layer of the SEGATT-CNN model.

3.2.1 SEGATT layer

The entity divides the text into different segments, and each context may contain semantic information important for distinguishing the entity relation. Adding an attention mechanism to each segment helps to focus on the sensitive words within the segment domain for relation classification. In this layer, the vector matrix $M_i \in T \times k$ ($i = 1,2,3$) is obtained by the attention mechanism for each independent input and then connected in the outputs of the three parts. The output matrix M is as follows:

$$\alpha_w = \text{soft}\max(V^T * X_i) \tag{1}$$

$$M_i = X_i \alpha_w^T \tag{2}$$

$$M = M_1 \oplus M_2 \oplus M_3 \tag{3}$$

We let V denote the attention vector, which represents the weight of each row of X_i, and M is the feature matrix after concatenation. We use this word attention mechanism to capture the specific significant words and jointly trained with the other CNN components.

3.2.2 Convolutional and pooling layer

Furthermore, to obtain valuable features, we use a CNN to capture context information. The convolutional layer is

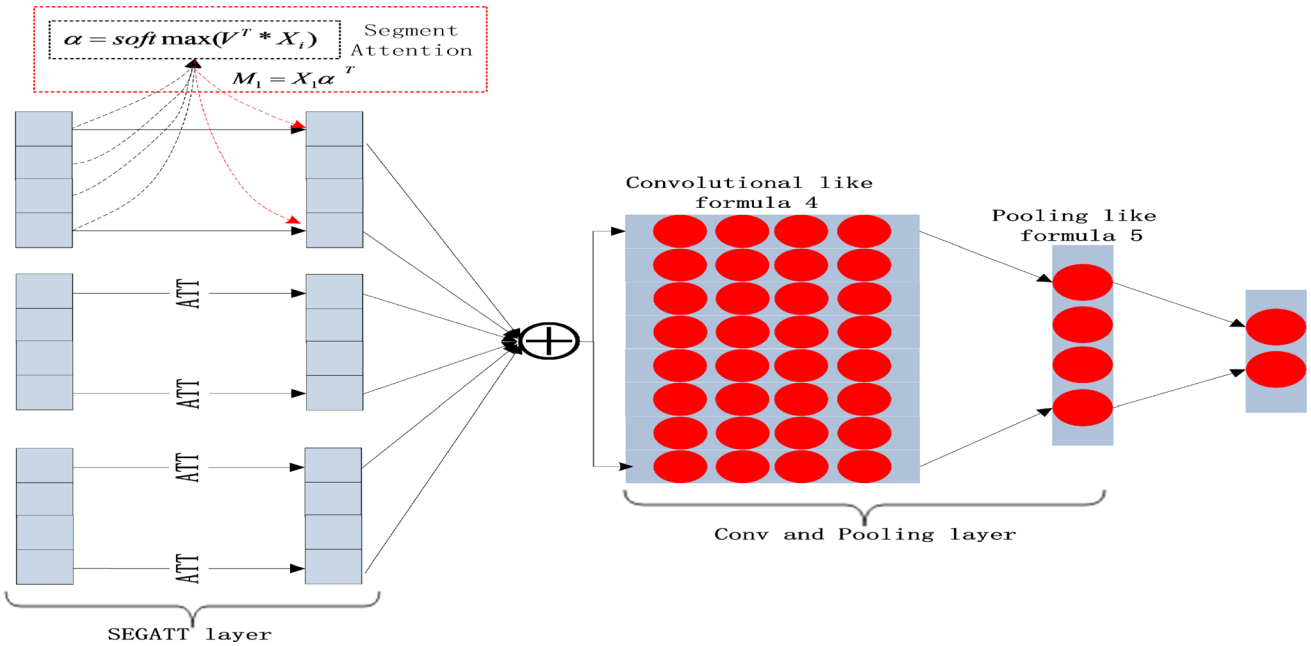

Fig. 3 Overview of the SEGATT-CNN layer architecture

intended to capture the local sentence semantics and compress this valuable information into feature maps. When the input to the convolution layer is the feature matrix M, M is the output of the SEGATT layer. The convolution operates as follows:

$$y_i = f(W_j^{h\times k} * X_{i:i+h-1} + b) \tag{4}$$

where $b \in \mathrm{R}$ is the bias, f is a nonlinear activation function such as a hyperbolic tangent function, and $W^{h\times k}$ is a filter applied to a window of h words to produce a new feature. A feature y_i is generated from a window of words $X_{i:\ i+h\ -\ 1}$.

A pooling operation is then utilized to further abstract the features generated from the convolution operation. In this layer, we use Max-pooling and Average-pooling to extract more effective features from feature maps to represent more accurate semantic information. Therefore, we concatenate these two pooling features to yield the output:

$$\mathrm{c}_{i\mathrm{ave}} = \mathrm{ave}(y_i) \tag{5}$$

$$\mathrm{c}_{i\max} = \max(y_i) \tag{6}$$

$$\mathrm{c}_i = c_{i\mathrm{ave}} \oplus c_{i\max} \tag{7}$$

where y_j is the new feature map for each convolution operation and c_i is the output of the pooling layer.

3.3 Connection and classification

Finally, we concatenate the two types of embedding features for relation classification. In this paper, we employed two methods for feature vector construction. One uses the SEGATT-CNN model to extract word2vec-encoded word embedding features. The other is constructed by introducing the Term Frequency-Inverse Document Frequency (TF-IDF) method to encode word features [26, 27]. The TF-IDF algorithm reflects the importance of a word in the text that has strong discrimination ability. TF-IDF encoded word features do not contain semantic features; this approach typically regards every word as a separate entity—which ignores connections between entities. Therefore, it is necessary to combine it with the word2vec representation. The formula for combining the two methods is as follows:

$$P = \{\ \mathrm{c}_1, c_2 ... c_m\} \tag{8}$$

$$Q = \{q_1, q_2 ... q_n\} \tag{9}$$

$$\mathrm{y} = F(C_1 P \oplus C_2 Q) \tag{10}$$

where P is the fully connected output after pooling, Q is the feature matrix from the TF-IDF encoding, which is a feature-connection operation and F represents a machine learning classifier such as SVM. C_1 and C_2 represent two characteristic forms connected by different weights.

4 Experiments

The proposed method is evaluated with two experiments. First, we test relevant deep learning models in CID task and compare with our proposed method to evaluate the performance. Second, we apply our method to solve the herbal relation extraction problem and then tested the effectiveness of the method by comparing it with some

baseline models. In this study, we used the word2vec toolkit to train embeddings on part of abstracts from PubMed. Figure 4 shows the work flow of our experiments.

4.1 Evaluate the proposed method on CID task

In this section, we compare our proposed method with other deep learning models on CID task, and the results show that our proposed method is more reliable than others.

4.1.1 Data set

The Chemical-Disease Relation (CDR) extraction data set [8] was created for extracting chemical and disease relations from the biomedical literature. The main purpose of the CID relation extraction task is to promote the in-depth study of entity relation extractions for biomedical applications. The CID relation extraction task based on intra-sentence levels is a decomposition of the relation extraction task of CID. In this paper, we only focus on the relation extraction task at the intra-sentence level.

We adopt precision (P), recall (R), and the F-score (F) to evaluate the models' performances. The formulas are as follows:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{11}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{12}$$

$$F\text{ - score} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{13}$$

where TP, FP, and FN represent the number of true positives, false positives and false negatives returned by the model, respectively. The F-score is the weighted average of precision and recall. We further introduce Area under the Curve (AUC) and accuracy in another experiment.

We explore the influence of the relationship classification method based on mixed characteristics on the experimental results. For this method, we pretrain two feature output modules, respectively, to make each model reach the optimal output feature tensor. Firstly, the influence of feature vectors connected by different weights on the classification results was tested. We took F value as the measurement standard and drew the contour curve of F value according to the results. It can be seen from Fig. 5 that the value range of C_2 in (10) is [0.4–0.6], so set $C_2 = 0.5$.

According to the experiment analysis, using different datasets and determining the optimal results, $C_2 = 0.5$ is obtained; that is, a representation in which the two output vectors are connected with equal weights after standardization.

4.1.2 Result and analysis

In order to test the performance of our proposed method, we applied the designed SEGATT-CNN model and the method of two-feature hybrid use SEGATT-CNN model combined with SVM classifier (SEGATT-CNN_SVM). Compared our methods with relevant deep learning models on this data set. The results are shown in Table 1. Compared with the model by Xu et al. [28] developed for the CDR task of the BioCreative V challenge, the LSTM and CNN method by Zhou et al. and the CNN method by Gu et al., RelSCAN [29] achieved the highest precision. But for the other two indicators, Li et al.'s RPCNN model achieved best results. The F value is an important indicator in classification tasks. Li et al.'s RPCNN model achieved significant advantages, which is more reliable than others. The RPCNN model is currently the best method for the overall performance of this task. This method combines the CNN and RNN models and introduces the attention mechanism. However, our SEGATT-CNN model only use word embedding features achieved good results, although its F value is not the highest. But compared to the all recent methods, our SEGATT-CNN_SVM method achieved best F-scores.

In summary, comparing the performance of all models, we obtained good results in terms of overall performance by only using word2vec embedding and TF-IDF embedding features without manual feature engineering. Our model's F-score is 2.74% higher than the RPCNN model.

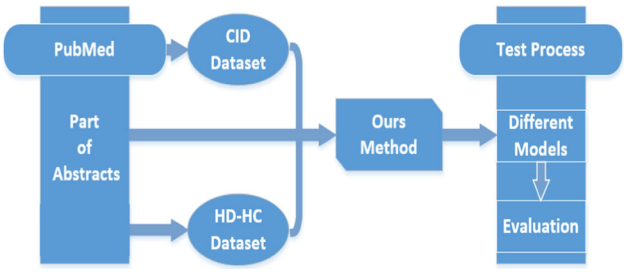

Fig. 4 The work flow of our experiments

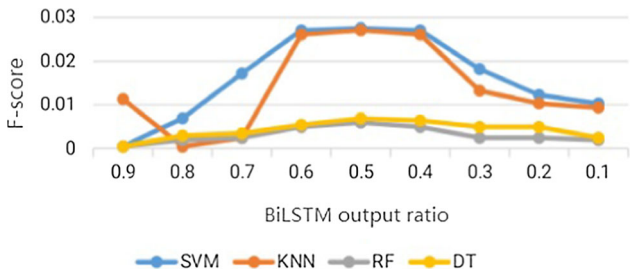

Fig. 5 The F-scores of different BiLSTM output radio

Table 1 Experiment results on the CID task

methods	P	R	F
Gu et al.'s CNN	0.597	0.550	0.572
Zhou et al.'s LSTM	0.5491	0.5141	0.531
Zhou et al.'s CNN	0.411	0.553	0.472
Li et al.'s CNN	0.5780	0.5420	0.5594
Li et al.'s RPCNN	0.5517	0.6363	0.5910
Xu et al.'s	0.5960	0.440	0.5073
RelSCAN	**0.637**	0.413	0.501
SEGATT-CNN	0.5874	0.5728	0.5775
SEGATT-CNN_SVM	0.5867	**0.6539**	**0.6184**

Bold indicates the best value among 9 methods

Thus, compared with existing deep learning methods, our proposed method is more effective in solving this problem. After looking up the wrongly classified data, we found that when the text expression is complex (when the trigger words that dominate the semantic relationship such as induced and cause are not included), the SEGATT-CNN_SVM model can still more effectively focus on the overall semantic information of the text, thereby improving the classification accuracy.

4.2 Applying the proposed method to herbal relation extraction tasks

We applied the proposed method to the two data sets we constructed and compared it with other baseline methods. The experiment results show that our method can effectively extract herbal-related entity relations.

4.2.1 Relation definition and problem definition

By summarizing and analyzing the related research in the field of biomedicine and TCM, we can draw an entity relation structure diagram such as the one shown in Fig. 6. It is easy to observe that an herbal entity in the TCM field can be connected with most noun entities. There are many

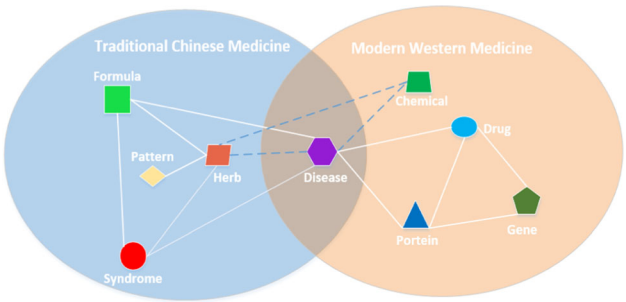

Fig. 6 Entity relation and extraction method between TCM and Western medicine

open data sets and related studies about entity relations in biomedicine [30]. Therefore, we defined the entity relations between herbs and diseases and between herbs and chemicals as the research objectives of this paper.

4.2.2 Data set construction

To explore the extraction of entity relations in the field of TCM, we built a corpus based on PubMed of biomedical literature. Figure 7 depicts the construction process.

4.2.2.1 Data Collection We collected and preprocessed the data set using the following steps:

1. We collected 2863 English herb names (including the Latin herb name) from various TCM resource websites. Then, we retrieved abstracts of related articles (22,600 articles) from PubMed.
2. We obtained and merged common diseases and chemical entity names 24,471 and 42,636, respectively, from Medical Subject Headings (MESH) and existing entity relation data sets.
3. The Stanford CoreNLP processing tool was used to preform sentence segmentation of the texts. Then, a dictionary-based method was adopted to mark the entities in the sentences.

To help us accurately select data samples, we defined the following simple rules:

1. When only two entities exist in the sentence (herbs and diseases or herbs and chemicals), we retained the sentence and classified it.
2. When the sentence contained multiple entities such as herbs and chemicals (diseases), the nearest herbal and chemical (disease) was selected as the included entity of the sentence and then classified.
3. Any notes appearing in parentheses (except for herb aliases and disease or chemical abbreviations) were ignored.

4.2.2.2 Data Annotation To identify a positive and effective correlation between TCM and biomedical entities, we developed relation-labeling guidelines for the two types of entity relations and used them to guide the annotation of the two different datasets. Figure 8 shows a sample relation between herbs and chemicals:

Herbal-chemical relation: If the chemical can be extracted from the herb or if the herb contains the chemical, we mark it as a positive relationship. If the herb merely cooccurs with the chemical or is otherwise not part of the extraction or composition relations, we label it as a negative sample.

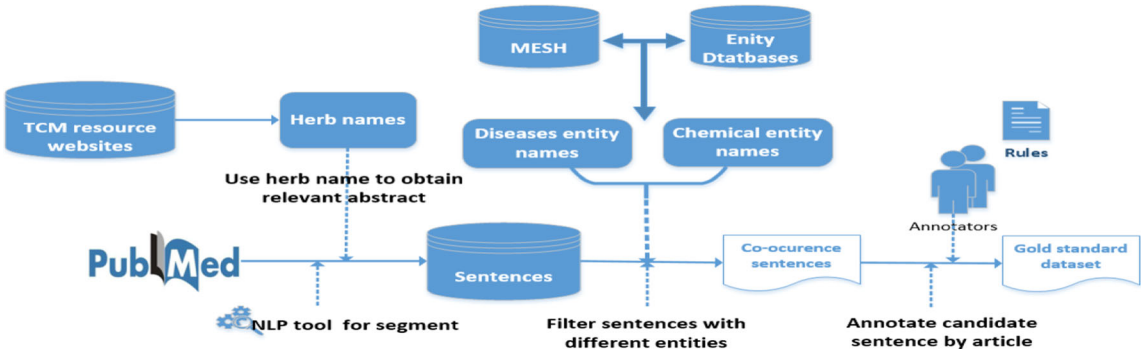

Fig. 7 The workflow for constructing our corpus

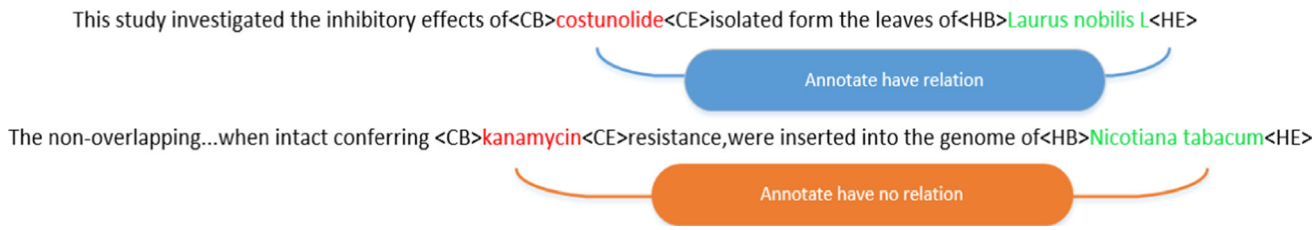

Fig. 8 A labeled sample of a relationship between herbal and chemical

Herbal-disease relation: If an herb can cure the disease or has a significant effect on the disease, we mark it as a positive sample; otherwise, we mark it as a negative sample.

4.2.2.3 Dataset Description Finally, we labeled 4732 samples of herbs-diseases and 4666 samples of herbs-chemicals. The dataset statistics are reported in Table 2.

4.2.3 Results and analysis

This part of the experiment was mainly intended to verify that the proposed method in this paper is effective when applied to the task of extracting herbal-related entity relations. To demonstrate the advantages of the proposed method, we compare it with the following representative baseline methods: 1. several machine learning classifiers (such as: KNN, SVM); 2. traditional CNN model for text classification (CNN); 3. the SEGATT-CNN model with a segment input mechanism (SEGATT-CNN); 4. our method combined with different classifiers (such as: KNN Combined SEGATT-CNN).

Table 2 The statistics of the labeled dataset

Number of Candidate Relations		
Relation Type	Labeled	
	Related	Unrelated
Herb-disease	1766	2966
Herb- chemical	2271	2395

In this experiment, we use a pre-trained word embedding with 100 dimensions. The hyperparameters of the SEGATT-CNN model are as follows: The convolutional window sizes are 3 or 4, the number of filters is 100, the optimizer is Adam, the dropout rate is 0.25, and the batch size is 200. To fairly evaluate the performances of the different methods, we randomly selected 35% of the entire data set as a test set and others as the training set. The experiment was repeated 100 times, and we adopted the average value as the final evaluation result.

The results reported in Table 3 are those for the task of classifying the true relations between herbs and diseases. The SVM, K-Nearest Neighbor (KNN), Decision Tree (DT) and Random Forest (RF) classifiers combined with our feature fusion approach achieved better results than did the machine learning methods and the standard CNN and SEGATT-CNN models. Compared with the machine learning method, our classification approach increased the F-score by 20.85%, 19.19%, 23.04%, and 8.22% when combined with SVM, KNN, DT, and RF, respectively. But for binary classification tasks, AUC is also an important measure. It shows that the model can correctly identify the positive relationship of Herb-Disease. The two methods proposed in this paper have achieved obvious advantages in AUC. Although our approach combined with different classifiers results in different performances, the SVM classifier achieves the best classification effect.

Table 3 Experiment results for herb and disease tasks

Methods	P	R	F	AUC	Accuracy
Standard CNN	0.9085	0.8670	0.8861	0.9067	0.9169
SEGATT-CNN	0.9209	0.8856	0.9023	0.9198	0.9285
KNN	0.7240	0.7063	0.7149	0.7729	0.7897
KNN combined SEGATT-CNN	0.9156	0.8983	0.9068	0.9244	0.9310
Decision tree	0.7098	0.6095	0.6535	0.7291	0.7593
Decision tree combined SEGATT-CNN	0.9075	0.8621	0.8839	0.9047	0.9288
Random forest	0.9023	0.7882	0.8412	0.8686	0.8890
Random forest combined SEGATT-CNN	0.9394	0.9081	0.9234	0.9366	0.9438
SVM	0.7977	0.6784	0.7330	0.7879	0.8156
SVM combined SEGATT-CNN	0.9543	0.9292	0.9415	0.9513	0.9569

The results on the task to classify the true relations between herbs and chemicals are shown in Table 4. Our method still achieves good overall performance results, and the experiment results are similar to those in Table 3. Compared with the machine learning methods, the combinations with the SVM, KNN and decision tree all have large improvements (F-score improvements of 16.48%, 14.2%, 15.59%, respectively), and the combination with the random forest achieves a small improvement (an F-score improvements of 5.1%). And from the AUC results, it can be seen that after the feature is mixed, the accuracy of the relationship recognition can be improved.

The results in Tables 3 and 4 show that in the several methods we tested: The CNN model can classify relations more effectively than traditional machine learning algorithms. Moreover, the SEGATT-CNN model is obviously better than the traditional CNN model, which also demonstrates that combining the new context representation with the word-level attention mechanism can improve the local feature extraction ability of convolutional neural networks. The classification method based on feature fusion also improves the SEGATT-CNN results, which means that using this architecture; we can adopt different classifiers to achieve more effective solution. In general, its F value is about 3%-5% higher than using the deep-learning model alone. Of course, the corresponding word vector representation method could also be replaced to accommodate different data sets. Therefore, the SVM classifier combined with the method of this paper can be more effectively applied to the issue of herbal-related entity relation extraction.

5 Conclusions

Based on the PubMed corpus, this paper studies the entity relationship extraction problem related to Chinese herb. And propose a novel deep-learning model with improved layers. We tested our novel method with two experiments. First, we tested relevant deep learning models in CID task and compared with our proposed method to evaluate the performance. The results showed that our method more reliable in solving CID relation extraction task. Second, we applied propose method to solve the problem of herbal-related relation extraction considered in this paper. Compared with several baseline methods, our method has an absolute advantage and is proved to be effective in solving this problem. In the future, we will study the application of

Table 4 Experiment results for herb and chemicals tasks

Methods	P	R	F	AUC	Accuracy
Standard CNN	0.8955	0.8771	0.8856	0.9037	0.9041
SEGATT-CNN	0.9205	0.8866	0.9027	0.9185	0.9191
KNN	0.8045	0.7594	0.7812	0.7922	0.7931
KNN combined SEGATT-CNN	0.9314	0.9153	0.9232	0.9257	0.9260
Decision tree	0.7645	0.7175	0.7388	0.7528	0.7538
Decision tree combined SEGATT-CNN	0.9104	0.8802	0.8947	0.8988	0.8993
Random forest	0.8669	0.8797	0.8731	0.8757	0.8756
Random forest combined SEGATT-CNN	0.9387	0.9092	0.9236	0.9264	0.9268
SVM	0.8274	0.7647	0.7947	0.8066	0.8078
SVM combined SEGATT-CNN	0.9670	0.9522	0.9595	0.9607	0.9609

unsupervised methods in this field and consider building a Traditional Chinese Medicine knowledge grap.

Funding This work is supported by the Development Project of Jilin Province of China (Nos.20200801033GH, YDZJ202101ZYTS128), Jilin Provincial Key Laboratory of Big Data Intelligent Computing (No.20180622002JC), The Fundamental Research Funds for the Central University, JLU.

Declarations

Conflict of interest The author(s) declared no potential conflicts of interest with respect to the research, author- ship, and/or publication of this article.

References

1. Liu J, Chen Z (2011) Traditional Chinese medicine in the new century. Front Med 5(2):111–114
2. Chai H, Hai LU, Liu QC (2015) Overview of research methods for natural language processing in traditional Chinese medicine. J Med Inf 36(10):58–63
3. Wu Z, Zhou X, Liu B, Chen J (2004) 'Text mining for finding functional community of related genes using TCM knowledge'. In European Conference on principles of data mining & knowledge discovery, pp.459–470
4. Fang YC, Huang HC, Chen HH, Juan HF (2008) TCMGeneDIT: a database for associated traditional Chinese medicine, gene and disease information using text mining. Bmc Complement Altern Med 8:58
5. Yu T, Li J, Yu Q, Tian Y, Shun X, Xu L, Zhu L, Gao H (2017) Knowledge graph for TCM health preservation: design, construction, and applications. Artif Intell Med 77:48–52
6. Golshan PN, Dashti HAR, Azizi S, Safari L (2018) 'A study of recent contributions on information extraction'. The 4th national conference on distributed computing and big data processing
7. Haihong HE, Zhang WJ, Xiao SQ, Cheng R, Hu YX, Zhou XS, Niu PQ (2019) A survey of entity relationship extraction based on deep learning. J Softw 30(6):1793–1818
8. Li J, Sun Y, Johnson RJ, Sciaky D, Wei CH, Leaman R, Davis AP, Mattingly CJ, Wiegers TC, Lu Z (2016) 'BioCreative V CDR task corpus: a resource for chemical disease relation extraction', Database the J Biol Databases Curation, vol. 2016, Article Number: baw068
9. Bai T, Gong L, Wang Y et al (2016) A method for exploring implicit concept relatedness in biomedical knowledge network. BMC Bioinf 17(9):53–56
10. Gu J, Qian L, Zhou G (2016)'Chemical-Induced disease relation extraction with various linguistic features', Database, vol. 2016, Article Number: baw042
11. Gu J, Sun F, Qian L, Zhou G (2017) Chemical-Induced disease relation extraction via convolutional neural network. Database J Biol Databases Curation 1:2017
12. Zhou H, Deng H, Chen L, Yang Y, Chen J, Huang D (2016) 'Exploiting syntactic and semantics information for chemical–disease relation extraction', Database J Biol Databases Curation, vol. 2016, Article Number: w48
13. Li H, Chen Q, Tang B, Wang X (2017) 'Chemical-Induced disease extraction via convolutional neural networks with attention', 2017 IEEE international conference on bioinformatics and biomedicine (BIBM), Kansas City, MO,USA, pp. 1276–1279
14. Li H, Ming Y, Chen Q, Tang B, Wang X, Yan J (2018) Chemical-Induced disease extraction via recurrent piecewise convolutional neural networks. BMC Med Inform Decis Mak 18(S2):45–51
15. Li Y, Jin R, Luo Y (2018) Classifying relations in clinical narratives using segment graph convolutional and recurrent neural networks (Seg-GCRNs). J Am Med Inform Assoc 26(3):262–268
16. Wang D, Su J, Yu H (2020) Feature extraction and analysis of natural language processing for deep learning English language. IEEE Access 8:46335–46345
17. Luo Y, Cheng Y, Uzuner Ã, Szolovits P, Starren J (2017) Segment convolutional neural networks (Seg-CNNs) for classifying relations in clinical notes. J Am Med Inform Assoc 25(1):93–98
18. Bai T, Wang C et al (2020) A novel deep learning method for extracting unspecific biomedical relation. Concurr Comput Pract Exp 32(1):e5005
19. Wan H, Moens MF, Luyten W, Zhou X, Mei Q, Liu L, Tang J (2016) Extracting relations from traditional chinese medicine literature via heterogeneous entity networks. J Am Med Inf Asmsociation Jaia 23(2):356–365
20. Wang J, Poon J (2017) 'Relation extraction from traditional Chinese medicine journal publication'. In IEEE international conference on bioinformatics & biomedicine, pp.15–18
21. Yang XH, Shan YH, Xie D, Li XD (2017) Relation extraction of traditional Chinese medicine prescription and disease based on literature abstracts data. Mod Tradit Chin Med Mater Medica-World Sci Technol 19(7):1167–1172
22. Han H, Liu J, Liu G (2018) Attention-based memory network for text sentiment classification. IEEE Access 6:68302–68310
23. Xiang Y, Xu Y, Yu Z et al (2019) CNN-based text multi-classifier using filters initialised by N-gram vector. Int J Inf Commun Technol 15(4):419
24. Vu NT, Adel H, Gupta P, Schütze H (2016) 'Combining recurrent and convolutional neural networks for relation classification'. In: proceedings of NAACL-HLT, pp. 534–539
25. Luong MT, Pham H, Manning CD (2015) 'Effective approaches to attention-based neural machine translation'. In: proceedings of the 2015 conference on empirical methods in natural language processing. Association for Computational Linguistics, Lisbon, Portugal, pp. 1412–1421
26. Ye W, Zhi Z, Shan J, Liu J, Mi L (2017)'Comparisons and selections of features and classifiers for short text classification.' In, IOP conference series-materials science and engineering (Iop Publishing Ltd) Vol. 261
27. Amin S, Uddin MI, Hassan S et al (2020) Recurrent neural networks with TF-IDF embedding technique for detection and classification in tweets of dengue disease. IEEE Access 8:131522–131533
28. Xu J, Wu Y, Zhang Y, Wang J, Lee HJ, Xu H (2016) 'CD-REST: a system for extracting chemical-induced disease relation in literature', Database, vol. 2016 Article Number:baw036
29. Chika Onye S, Akkeleş A, Dimililer N (2018) RelSCAN—A system for extracting chemical-induced disease relation from biomedical literature. J Biomed Inf 87(2018):79–87
30. Bai T, Ge Y et al (2019) BERST: an engine and tool for exploring biomedical entities and relationships. Chin J Electron 28(4):797–804