International Journal of Machine Learning and Cybernetics (2021) 12:355–363
https://doi.org/10.1007/s13042-020-01176-6

ORIGINAL ARTICLE

Chinese medical relation extraction based on multi-hop self-attention mechanism

Tongxuan Zhang[1] · Hongfei Lin[1] · Michael M. Tadesse[1] · Yuqi Ren[1] · Xiaodong Duan[2] · Bo Xu[1,3]

Received: 30 December 2019 / Accepted: 1 August 2020 / Published online: 14 August 2020

Abstract
The medical literature is the most important way to demonstrate academic achievements and academic exchanges. Massive medical literature has become a huge treasure trove of knowledge. It is necessary to automatically extract implicit medical knowledge from the medical literature. Medical relation extraction aims to automatically extract medical relations from the medical text for various medical researches. However, there are a few kinds of research in Chinese medical literature. Currently, the popular methods are based on neural networks, which focus on semantic information on one aspect of the sentence. However, complex semantic information in the sentence determines the relation between entities, the semantic information cannot be represented by one sentence vector. In this paper, we propose an attention-based model to extract the multi-aspect semantic information for the Chinese medical relation extraction by multi-hop attention mechanism. The model could generate multiple weight vectors for the sentence through each attention step, therefore, we can generate the different semantic representation of a sentence, respectively. Our model is evaluated by using Chinese medical literature from China National Knowledge Infrastructure (CNKI). It achieves an F1 score of 93.19% for therapeutic relation tasks and 73.47% for causal relation tasks.

Keywords Chinese medical literature · Multi-hop self-attention mechanism · Relation extraction · Natural language processing (NLP)

Abbreviations

NLP	Natural language processing
CNKI	China National Knowledge Infrastructure
PPIs	Protein–protein interactions
DDIs	Drug–drug interactions
CPIs	Chemical–protein interactions
CDIs	Chemical–disease interactions
SAE	Stacked autoencoder
CNNs	Convolutional neural networks
RNN	Recurrent neural network
LSTM	Long short term memory network
Bi-LSTM	Bidirectional long short term memory network
SVM	Support vector machine
McDepCNN	Multi-channel dependency-based convolutional neural network

✉ Hongfei Lin
hongfeilin@dlut.edu.cn

[1] Dalian University of Technology, Dalian, China

[2] Dalian Minzu University, Dalian, China

[3] State Key Laboratory of Cognitive Intelligence, iFLYTEK, Hefei, People's Republic of China

1 Background

The medical literature contains a large quantity of valuable inter-entity relations. Currently, the common medical relation extraction tasks include: protein–protein interactions (PPIs) [1], drug–drug interactions (DDIs) [2], chemical–protein interactions (CPIs) [3] and chemical–disease interactions (CDIs) [4]. In promoting the development of medical fields, this information has important research value and plays an important role. For datasets extracted from medical relations, there are some publicly available datasets, such as Aimed [5], DDIExtraction 2013 [6], BioCreative V-CDR [4] and so on. All of this data comes from the MEDLINE database, the largest medical literature database in the world. Compared with these English datasets, there is a lack of relevant public datasets in China, so that the development of Chinese medical text mining is relatively late.

With the rapid development of the Chinese medical field, Chinese medical literature has grown at an explosive rate. Most of the Chinese medical literature is included in China National Knowledge Infrastructure (CNKI) that providing a wealth of literature resources for the development of medical science in China. According to the literature statistics of CNKI, the number of medical literature published is large every year. A wealth of medical knowledge has accumulated in the medical literature. However, few researchers conducted research on relation extraction in Chinese literature. For the research of relation extraction in Chinese medical literature, it is still an issue worthy of attention.

The early medical relation extraction task mainly relies on the template matching that needs to analyze the text features and manually summarize the grammar rules, and then match the relations in the new text. This method required the annotator to have high requirements for linguistics and medical knowledge. Blaschke [7] et al. built more than ten templates by syntactic information, contextual information, and other information, to identify genes and protein entities from medical literature and judging the relation between entities. Corney et al. [8] used a custom template to replace the original template in GATE[1] to implement a rule-based information extraction system. Due to the different syntactic information and different language expression of datasets, it would lead to a lower recall of relation extraction, so that the generalization ability of the template matching method is very limited.

As the medical relation extraction task become a focus of research, many institutions have conducted relevant evaluations and prompted many machine learning methods to be applied to this task. Early machine learning focused on eigenvector methods to map text features into a high-dimensional vector as the final feature vectors in classifiers. Alam et al. [9] manually extracted varied linguistic features, such as vocabulary information, phrases information, and dependency syntax information, to extract the relation on chemical–disease relation data. Kim et al. [10] also extracted rich features and input into a linear support vector machine model (SVM) for the drug–drug interaction extraction task. Machine learning methods generally have better generalization and portability than template matching methods. However, the quality of feature selection often determines the performance of relation extraction and needs to use external resources or tools to extract features, such as part-of-speech tagger, syntax-dependent approach. Errors generated by these external tools can have a cascading effect on the performance of the relation extraction.

However, machine learning methods rely on manual construction of features, which is laborious and time-consuming. With the advance of deep learning and word representation learning, lots of researchers apply deep learning methods to the medical relation extraction. Peng et al. [11] used a multi-channel dependency-based convolutional neural network (McDepCNN), which has the best performance on the PPI dataset in 2017. Zhang et al. [12] integrated the sentence sequence and shortest dependency path by the hierarchical recurrent neural networks (RNNs)-based method for the DDI extraction task. Zhang et al. [13] used CNN and RNN for biomedical relation extraction, it is called a hybrid deep neural model. With strong representations of data, better performances, and less feature engineering, deep learning methods became the most popular methods of relation extraction.

Bidirectional long short-term memory (Bi-LSTM) and CNN are the most widely employed neural network structures in medical relation extraction. CNN [14] reduces model training parameters by them sharing and local connections. It is more suitable for short text classification tasks. For dealing with long-distance features and obtain the information, The Bi-LSTM [15] is designed. However, different parts of the sentence have different effects on the classification, focusing on more important information will help to improve the classification results. Thus, some researches have added an attention mechanism to select more important information in the sentence [16]. However, these methods only learned a sentence vector representation that could not capture more complex information in a sentence.

Therefore, in this paper, we design an attention-based model to obtain additional vector representation of the sentence, which considers the previous self-attention memory information. In conclusion, the main contributions are:

- We first evaluate the multi-hop self-attention model on the task of Chinese medical relation extraction. By each step of attention, different words weight could be generated for the sentence.
- We could obtain the multi-aspect semantic information from Chinese medical literature by the model. The final predication combines the results from different sentence representations.
- We conduct our experiment on Chinese medical literature datasets. Extensive evaluations show that our method establishes a new state-of-the-art on Chinese literature.

2 Methods

We apply the multi-hop attention to the Chinese medical relation extraction task. Our model learns complex semantic information based on multiple sentence vector representations, and comprehensively considers different information to achieve the final classification result. The model is divided into four parts: (1) embedding layer, (2) Bi-LSTM layer, (3)

[1] https://gate.ac.uk/.

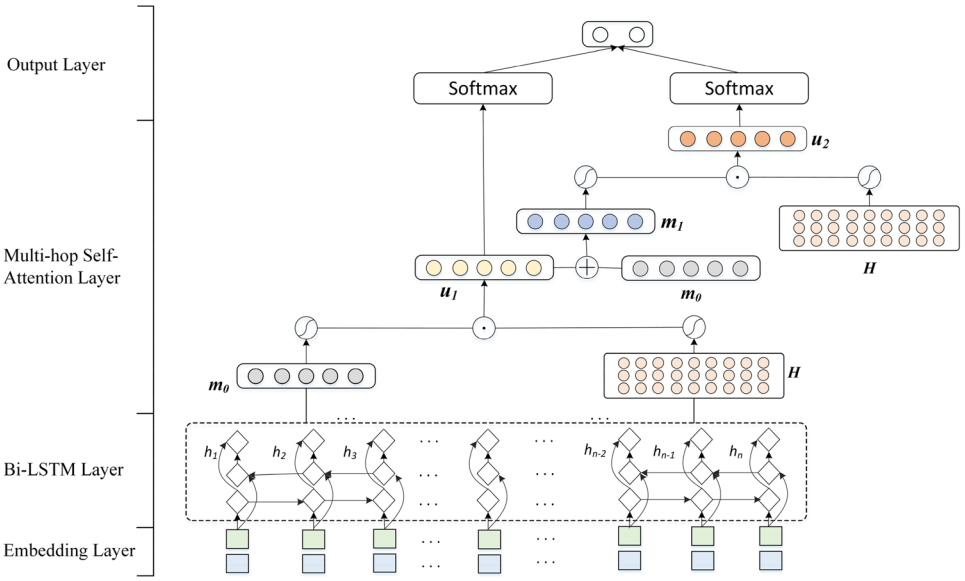

Fig. 1 Architecture of our proposed model

multi-hop self-attention layer, (4) output layer. The input of the model is sentence representation, which is obtained by concatenation of the word embedding and the position embedding. Next, the Bi-LSTM layer learns the deep semantic information of the text. Then, the multi-hop self-attention mechanism extracts complex semantic information. Through multiple iterations, the model can obtain a different representation for different information of the sentence. Finally, these vector representations are input the fully connected layer, respectively. By the softmax function, we calculate the classification probability, and all the classification probabilities are averaged as the final classification result. The model structure diagram is shown in Fig. 1:

2.1 Text preprocessing

After the data preprocessing is completed, the next step is to segment each token of the sentence. However, the original tool is not good for the segmentation of medical proprietary words. To solve this issue, we expand the medical entity dictionary we built, the keywords of the medical literature, and the entity words identified in the word segmentation dictionary. The experimental results show that the new word segmentation dictionary could improve the performance.

2.2 Feature design

In this paper, we extracted two features for model training. In each sentence, the word representation x_i is got by concatenating the word embedding and position embedding as the input of the model. For the word x_i in the sentence, each word x_i is represented by a vector $x_i = [E_{wordi}; E_{posi}]$.

Word embedding. As the input of the model, the word embedding plays a crucial role in model training. The Word2Vec [17] is a popular word embedding method, and it is an unsupervised learning method. During training, Word2Vec learns low-dimensional vector representations for words. Meanwhile, it solves the memory overflow issue caused by one-hot encoding. In recent years, Word2Vec has been used in text classification, sentiment analysis, and other natural language processing (NLP) tasks. In our model, we use the skip-gram model of Word2Vec to pre-train word embeddings. The training data are from the abstracts of the CNKI medical literatures by using web crawlers. The final training data size is about 2G.

Position embedding. The closer the words are to the entities, the more important they are. Based on this assumption, we extract the relative distances between each entity and other words as an auxiliary feature in the relation extraction task. Zeng et al. [18] used the position embedding as one part of the input in the relation extraction task, and the performance is better than without the position embedding. The position feature refers to the relative distance of other words to each entity. For example, the position features calculated afterword segmentation are shown in Fig. 2:

For the word x_i in the sentence, there are two position features pos_1 and pos_2, which are concatenated as the final position feature $pos = [pos_1; pos_2]$. The position feature vector is obtained by random initialization, and the vector representation is updated during the relation extraction model training process.

2.3 Bi-LSTM layer

During the training, the traditional RNN import the output of the previous moment to learn the sequence information of the sentence. However, when the sentence is too long, the previous information will be lost. Therefore, LSTM [19] is

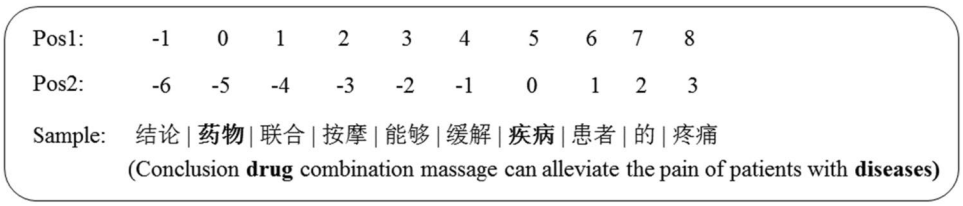

Fig. 2 A sample of position feature

proposed. It could not only retain the previous information but also solve the vanishing gradient problem.

For an input sequence is $X = (x_1, x_2, \ldots, x_n)$, and the output through the Bi-LSTM layer is $H = (h_1, h_2, \ldots, h_n)$. Also, there is a cell state C_t at each moment t, which is used to preserve long-term information. The cell C_{t-1} is the state at last moment *t-1*. There are only a few linear interactions during training. The structure of an LSTM unit incorporates three gates which are the input, forget and output gate. The formulas are:

$$f_t = \sigma(W_f \bullet [h_{t-1}, x_t] + b_f) \tag{1}$$

$$i_t = \sigma(W_i \bullet [h_{t-1}, x_t] + b_i) \tag{2}$$

$$\tilde{C}_t = tanh(W_C \bullet [h_{t-1}, x_t] + b_C) \tag{3}$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{4}$$

$$o_t = \sigma(W_o \bullet [h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_t = o_t * tanh(C_t) \tag{6}$$

where W are parameters to be trained, x_t is the input token at the moment t, h_{t-1} is the hidden state of the *t-1* cell, h_t is the hidden state of the t cell, b are the bias vectors, σ indicates the *sigmoid* function and *tanh* is an activation function.

A forward LSTM mechanism can only learn forward information in a text sequence, but the effect of the text classification model often depends on the contextual information of the text. Therefore, in this paper, we use the forward and backward LSTM as the feature extraction layer of relation extraction task. Because the Bi-LSTM model can obtain richer contextual information, it is better than the uni-directional LSTM. Finally, we concatenate the two hidden states as the final output $h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$ at time step t.

2.4 Self-attention mechanism

In NLP, the attention mechanism has developed rapidly. The main reasons are divided into three aspects. Firstly, the neural network with the attention mechanism outperforms the best model in multiple NLP tasks. Secondly, although the deep learning could obtain good performance, the training process is usually difficult to explain. The weighting mechanism assigns the weight to each word in the sentence, to increase the interpretability of neural networks. Finally, the attention mechanism can learn the contextual information of the text and solve the issue of long-distance dependency. With the development of attention mechanism research, many variant models have been produced, and the self-attention model is the most widely used in text classification [20, 21] and relation extraction [22, 23]. The self-attention could automatically learn the weight for each word. It assigns different weights to each word so that to distinguish the importance of the word in the sentence.

The hidden representation of the sentence is $H = (h_1, h_2, \ldots, h_n)$, the formula to calculate the word weight using the self-attention mechanism is as follows:

$$\beta = softmax(w^T tanh(WH + b)) \tag{7}$$

$$u = \sum_t \beta^T h_t \tag{8}$$

where W indicates the trainable parameter matrix, w is the trainable parameter vector, b is bias vector. β is the weight for the each word. t is the t-th word in the sentence. Finally, we obtain the vector representation u of the sentence.

2.5 Multi-hop self-attention mechanism

The earliest multi-hop attention mechanism was proposed in the question answering task [24]. Multiple attention could focus on different parts of the sentence to obtain multiple information in the sentence. The multi-hop attention mechanism translates the different information into multiple vector representations and considers each important information to select the best answer. Currently, the common relation extraction model has a problem with single semantic information. For example, the self-attention mechanism concentrates text information on the most relevant part of the task but ignores other information. But we could use multiple self-attention mechanisms to assign different weights to the words based on the previous memory weights. So that we could obtain complex semantic information in sentences and obtain multiple sentence vector representations. Inspired by the above ideas, we propose a multi-hop self-attention model to improve the

effectiveness of medical relation extraction. We use multiple attention iterations to obtain the complex relationship between the entities.

After the Bi-LSTM layer, we obtain the sentence hidden vectors H, the length of the sentence is n. We need to define an array of parameters M to save the word weight for memory after self-attention. The m^k denotes a memory vector, it is saved in M and could guide the next attention step. When calculating the word weight of the k^{th} self-attention, the formula is as follows:

$$S^k = \tanh(W_h^k H) \odot \tanh(W_m^k m^k) \tag{9}$$

$$\beta^k = \text{softmax}(W_s^k S^k) \tag{10}$$

where W indicate the attentive weight matrices. The initialization vector of m is obtained by averaging the sentence hidden vector after the Bi-LSTM. The m^k parameter is recursively updated by:

$$\begin{cases} m^0 = \frac{1}{N}\sum_t h_t \\ m^k = m^{k-1} + u^k \end{cases} \tag{11}$$

where u^k represents the sentence vector representation obtained after the weighted summation of the word vectors. The formula is as follows:

$$u^k = \sum_t \beta^k h_t \tag{12}$$

We could obtain the different u^k after each step of the self-attention so that k vector representation of the sentence is calculated. Then by using u^k, we calculate the classification probability. The finally multiple classification results are averaged as a basis of the classification. The formula is as follows:

$$R^k = softmax(u^k) \tag{13}$$

$$R = \frac{1}{k}\sum_k R^k \tag{14}$$

In this paper, we set the number of self-attention steps is 2, which the result is the best. When the k is 2, the multi-hop self-attention mechanism structure is shown in Fig. 3. It can be seen from the figure that each step of the self-attention updates m^k until k is 2, and interactives with the deep sentence vector representation to obtain a new sentence representation u^k.

2.6 Output and training

The loss function is the cross-entropy function of the model. The formula is as follows:

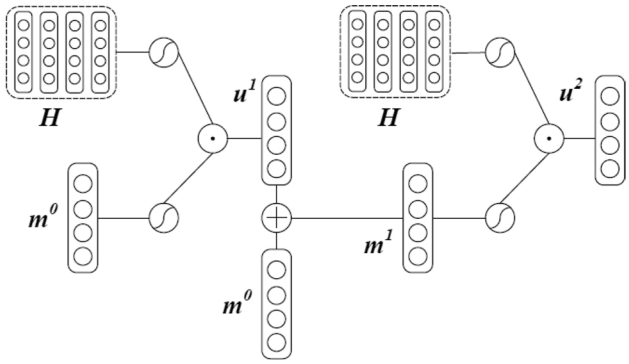

Fig. 3 The multi-hop self-attention mechanism structure

$$C = -\sum_i y_i ln R_i \tag{15}$$

The parameters are trained by minimizing the loss function. y_i is the real classification result and R_i is prediction result.

3 Results

3.1 Datasets

In this paper, the dataset is extracted from 4000 Chinese medical literature abstracts. Chinese medical literature is all from the Chinese core journal of PKU. We use crawler technology to obtain literature abstracts from CNKI. The sentence splitting was performed after the complete annotation of the abstracts. As a result of annotation, the 4000 Chinese medical literature abstracts generated 5490 sentences as training samples. Each semantic relation in the dataset consists of a specific type of entity pair. Then, the entity pairs with the same type are divided into the same group.

Before the relation extraction, the entities in the sentence have all been annotated. For sentences of multiple entities, we combine the different types of entities and then generate new instances for different entity pairs. In each instance, we focus on only two entities. In this paper, we only retain the sentences that contain therapeutic relation and causal relation. Therefore, we focus on the therapeutic relation extraction and causal relation extraction task. There is a therapeutic relation sample in Fig. 4:

There are more than two entities in an instance. To clearly distinguish the entity pairs with judgment in each instance, we replace drug name and disease name with “drug” and “diseases”. During the experiment, each dataset was randomly selected 10% as the test set, and the sentence length of each dataset was counted. Here we present the details of two datasets in Table 1.

Fig. 4 The therapeutic relation sample of data processing

Table 1 The details of different types of dataset

Relation type	Length of sentence	Positive number	Negative number	All number
Therapeutic relation	120	2032	2164	4196
Causal relation	150	530	764	1294

Table 2 Training details of medical relation extraction model based on multi-hop self-attention mechanism

Layer	Hyper-parameter	Value
Embedding	Word embedding dimension	200
	Position embedding dimension	15*2
Bi-LSTM	The output of Bi-LSTM	50*2
Attention	k	2
Training parameters	Dropout	0.5
	Optimizer	Adam
	Learning rate	0.001
	Batch size	8

3.2 Evaluation metrics and training details

The input of the model is the concatenation of word embeddings and position embeddings. Then the deep semantic information is obtained through the Bi-LSTM layer and multi-hop self-attention mechanism. Finally, the results are classified by the softmax function. The learning rate of the gradient descent method is Adam optimizer. As a classification task, we calculate the precision (P), recall (R) and F-score (F1) to evaluate the performance of our model. When determining the dimension of position embedding, 10, 15, and 20 were tried in the experiment, and finally the result with 15-dimensional of position embedding is the best. In the multi-hop self-attention mechanism, the number of self-attention steps k is 1, 2, and 3, respectively, and according to the performance of the model, we select k is 2. The parameter settings are shown in Table 2.

3.3 Experimental results

In this paper, to verify that the Bi-LSTM model based on multi-hop self-attention mechanism has better performance in relation extraction tasks. We select several mainstream relation extraction models as comparative experiments. The experiment results are shown in Table 3.

Table 3 Comparison of performance with other methods

Methods	Therapeutic relation			Causal relation		
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
CNN	87.01	91.89	89.38	60.0	68.95	64.16
Bi-LSTM	85.27	98.48	91.40	59.09	68.42	63.41
Bi-LSTM + self-attention	85.96	97.94	91.56	60.0	71.43	65.22
Bi-LSTM + multi-head attention	88.71	96.34	92.14	65.91	76.28	70.72
Our model	88.65	98.23	93.19	69.23	78.26	73.47

Because the relation extraction task could be viewed as a classification task, the CNN and Bi-LSTM methods are commonly used as the basic model. In this paper, our model also compares with these two models. The self-attention mechanism is widely used in various NLP tasks because it

Table 4 Comparison of step of self-attention in therapeutic relation extraction

k	P (%)	R (%)	F1 (%)
1	85.96	97.94	91.56
2	88.65	98.23	93.19
3	87.56	97.69	92.34

Table 5 Comparison of step of self-attention in causal relation extraction

k	P (%)	R (%)	F1 (%)
1	68.42	72.22	70.27
2	69.23	78.26	73.47
3	69.69	76.67	73.02

sentence. Meanwhile, adding multi-head attention can assign weights to multiple parts of the Bi-LSTM output vector to capture different parts of semantic information. Therefore, the performance is better than the self-attention mechanism. The Bi-LSTM model based on the multi-hop self-attention mechanism proposed in this paper can obtain more complex semantic relations in sentences. Through multiple iterations of self-attention, it can capture different important information in each iteration. Compared with other methods, our model achieved significant results. The F1-score was obtained on our methods on the therapeutic relation task (93.19%) and causal relation task (73.47%).

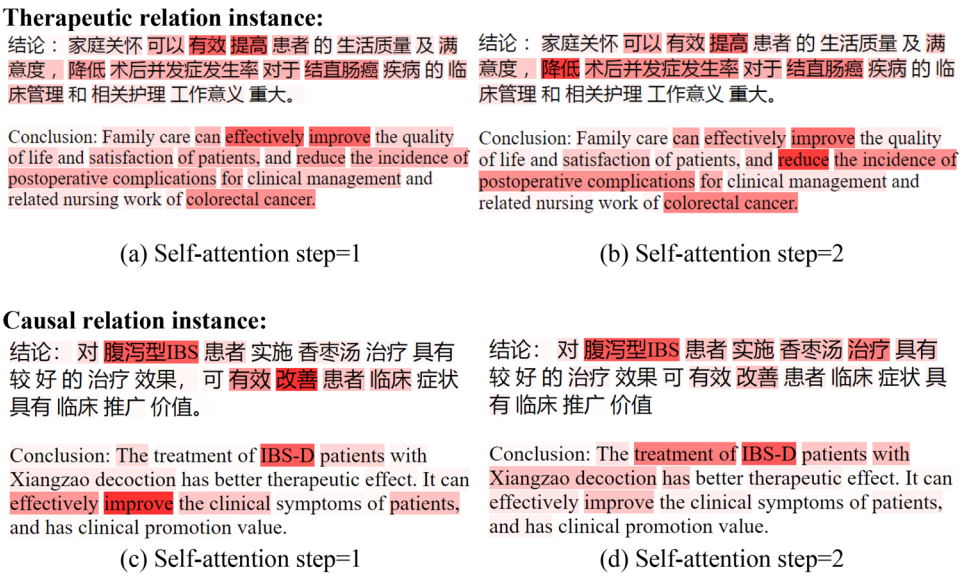

Fig. 5 The weight of the example on different self-attention steps

can learn important information in the text by assigning different weights to the words, and the performance is better in the classification task. Therefore, a lot of researchers applied it to medical relation extraction [25]. Also, the multi-head attention [26] mechanism also obtained high attention in NLP. This mechanism divided the sentence hidden vector representation into multiple parts, and did the attention operation for each part, and then connect them to obtain the final output result. The performance of multi-head attention is better than the single attention mechanism in the classification effect. Therefore, we also use the multi-head attention in medical relation extraction. The input of all models is the sentence vector representations by concatenation of the word embedding and position embedding.

As shown in Table 3, the multi-hop self-attention mechanism outperforms other methods in the Chinese medical relation extraction task. In the first two lines, the Bi-LSTM achieved higher performance than the CNN model, because the contextual information extracted by Bi-LSTM is important although CNN can extract the local features of a

3.4 Effect of the number of self-attention step

To verify the influence of the number of self-attention step k on the result, Table 4 shows the results with the different numbers of steps on the therapeutic relation task, and Table 5 shows the results with the different numbers of steps on the causal relation task. From Tables 4 and 5, we see that when the number of multi-hop self-attention mechanism iteration steps is 2, the performance of the model is the best. Meanwhile, the optimal number of iterations may be related to sentence length. Therefore, multiple self-attention steps may be effective for long sentence information detection.

3.5 Case study

We visualize the heat map of some sentences generated by our multi-hop self-attention model in Fig. 5. Each word in the sentence, the stronger the color, the larger the attention weight. This shows that the word plays an important role in

classification. As shown in Fig. 5, we provide the heat map with a different number of multi-hop self-attention steps. The sentences in Fig. 5a and b are from the therapeutic relation dataset, and the sentences in Fig. 5c and d are from the therapeutic relation dataset.

Figure 5a and b show the sentence weight generated through the first-hop self-attention and the second-hop self-attention, respectively. In this example, we observe that the second step could focus on information in the sentence that is different from the first step. The entities in the sentence are 'family care' and 'colorectal cancer' respectively. The word 'improve' could indicate the positive of the view between the two entities. However, the word 'reduce' also could play an important role in the classification of relation extraction. Finally, through the combination of the two aspects of the sentence, it can be better classified. In Fig. 5c and d, the word 'improve' and 'treatment' could play an important role in the classification of relation extraction.

4 Conclusions

Although deep learning solved the drawbacks of machine learning relying on manual extraction features, most neural network models are limited to a single vector representation. However, the semantic relationship between two entities in a sentence is often complicated, especially for long sentences. Therefore, we employ the multi-hop self-attention mechanism to general multiple sentence representations. In the Chinese medical relation extraction task, our model focus on different semantic information between two entities by multiple iterations of attention. Experimental results on two Chinese medical datasets show that our model is highly effective among existing experiments, achieving state-of-the-art results with little feature engineering.

Funding This work has been supported by the Natural Science Foundation of China (No. 61632011, 61572102). The Postdoctoral Science Foundation of China (No. 2018M641691). The Foundation of State Key Laboratory of Cognitive Intelligence, iFLYTEK, P.R. China (COGOS-20190001). The funding bodies did not play any role in the design of the study, data collection and analysis, or preparation of the manuscript.

References

1. Wang Y, You ZH, Yang S et al (2019) A high efficient biological language model for predicting protein–protein interactions. Cells 8(2):122
2. Ryu JY, Kim HU, Lee SY (2018) Deep learning improves prediction of drug–drug and drug–food interactions. Proc Natl Acad Sci 115(18):E4304–E4311
3. Kringelum J, Kjaerulff S K, Brunak S et al (2016) ChemProt-3.0: a global chemical biology diseases mapping. Database 2016
4. Wei C H, Peng Y, Leaman R et al (2016) Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task. Database 2016
5. Bunescu R, Ge R, Kate RJ et al (2005) Comparative experiments on learning information extractors for proteins and their interactions. Artif Intell Med 33(2):139–155
6. Segura Bedmar I, Martínez P, Herrero Zazo M (2013) Semeval-2013 task 9: extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). Association for Computational Linguistics, pp 341–350
7. Blaschke C, Valencia A (2002) The frame-based module of the SUISEKI information extraction system. IEEE Intell Syst 17(2):14–20
8. Corney DPA, Buxton BF, Langdon WB et al (2004) BioRAT: extracting biological information from full-length papers. Bioinformatics 20(17):3206–3213
9. Alam F, Corazza A, Lavelli A et al (2016) A knowledge-poor approach to chemical-disease relation extraction. Database 071
10. Kim S, Liu H, Yeganova L et al (2015) Extracting drug-drug interactions from literature using a rich feature-based linear kernel approach. J Biomed Inform 55:23–30
11. Peng Y, Lu Z (2017) Deep learning for extracting protein-protein interactions from biomedical literature. arXiv preprint arXiv:1706.01556
12. Zhang Y, Zheng W, Lin H et al (2017) Drug–drug interaction extraction via hierarchical RNNs on sequence and shortest dependency paths. Bioinformatics 34(5):828–835
13. Zhang Y, Lin H, Yang Z et al (2018) A hybrid model based on neural networks for biomedical relation extraction. J Biomed Inform 81:83
14. Lee K, Qadir A, Hasan SA, Datla V, Prakash A, Liu J, Farri O (2017) Adverse drug event detection in tweets with semi-supervised convolutional neural networks. In: Proceedings of the international conference on World Wide Web, pp 705–714
15. Li F, Zhang M, Fu G, Ji D (2017) A neural joint model for entity and relation extraction from biomedical text. BMC Bioinform 18(1):198
16. Alimova I, Solovyev V (2018) Interactive attention network for adverse drug reaction classification. In: Conference on artificial intelligence and natural language. Springer, pp 185–196
17. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781
18. Zeng D, Liu K, Lai S et al (2014) Relation classification via convolutional deep neural network. In: Proceedings of the 25th international conference on computational linguistics (COLING), pp 2335–2344
19. Hochreiter S (1998) The vanishing gradient problem during learning recurrent neural nets and problem solutions. Int J Uncertain Fuzziness Knowl Based Syst 6:107–116
20. Dong Y, Liu P, Zhu Z et al (2019) A fusion model-based label embedding and self-interaction attention for text classification. IEEE Access 8:30548–30559
21. Wu X, Cai Y, Li Q et al (2018) Combining contextual information by self-attention mechanism in convolutional neural networks for text classification. In: International conference on web information systems engineering, Springer, Cham, pp 453–467
22. Du J, Han J, Way A et al (2018) Multi-level structured self-attentions for distantly supervised relation extraction. arXiv preprint arXiv:1809.00699
23. Huang Y, Du J (2019) Self-attention enhanced CNNs and collaborative curriculum learning for distantly supervised relation extraction. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), pp 389–398

24. Tran NK, Niedereée C (2018) Multihop attention networks for question answer matching. In: The 41st international ACM SIGIR conference on research & development in information retrieval, ACM, pp 325–334
25. Zhou P, Shi W, Tian J et al (2016) Attention-based bidirectional long short-term memory networks for relation classification. In: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: short papers), pp 207–212
26. Vaswani A, Shazeer N, Parmar N et al (2017) Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008