Document-level Biomedical Relation Extraction Using Graph Convolutional Network and Multi-head Attention

Jian Wang, Xiaoyu Chen, Yu Zhang, Yijia Zhang, Jiabin Wen, Hongfei Lin, Zhihao Yang, Xin Wang

Submitted to: JMIR Medical Informatics
on: December 30, 2019

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.
Document-level Biomedical Relation Extraction Using Graph Convolutional Network and Multi-head Attention

Jian Wang, Xiaoyu Chen, Yu Zhang, Yijia Zhang, Jiabin Wen, Hongfei Lin, Zhihao Yang, Xin Wang

Corresponding Author:
Yijia Zhang
Phone: +860411-84708498
Email: zhyj@dlut.edu.cn

Abstract

Background: Automatically extracting relations between chemicals and diseases plays an important role in biomedical text mining. Chemical-disease relation (CDR) extraction aims at extracting complex semantic relationships between entities in documents, which contain intra- and inter-sentence relations. Most previous methods do not consider dependency syntactic information across the sentences, which are very valuable for the relations extraction task, in particular for extracting the inter-sentence relations accurately.

Methods: In this paper, we propose a novel end-to-end neural network based on the graph convolutional network (GCN) and multi-head attention. To improve the performance of inter-sentence relation extraction, we construct the document-level dependency graph to capture the dependency syntactic information across sentences. GCN is applied to capture the feature representation of the document-level dependency graph. The multi-head attention mechanism is employed to learn the relative important context features from different semantic subspaces. To enhance the input representation, the deep context representation (ELMo) is used in our model instead of traditional word embedding.

Results: The experimental results show that our method achieves an F-score of 63.5% which is superior to other state-of-the-art methods. The GCN model can effectively exploit the across sentence dependency information to improve the performance of inter-sentence CDR extraction. Both the ELMo and multi-head attention are helpful in CDR extraction task.

(jmIR Preprints 30/12/2019:17638)
DOI: https://doi.org/10.2196/preprints.17638

Preprint Settings
1) Would you like to publish your submitted manuscript as preprint?
✔ Please make my preprint PDF available to anyone at any time (recommended).
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?
✔ Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http://pubmednow.jmir.org">PubMed Now! service</a> my accepted manuscript PDF will automatically be made openly available.
Original Paper

Document-level Biomedical Relation Extraction Using Graph Convolutional Network and Multi-head Attention

Jian Wang¹, Xiaoyu Chen¹, Yijia Zhang¹, Yu Zhang¹, Jiabin Wen², Hongfei Lin¹, Zhihao Yang¹, Xin Wang¹

¹School of Computer Science and Technology, Dalian University of Technology, Dalian, 116024, China
²Department of VIP, the Second Hospital of Dalian Medical University, Dalian, 116024, China

§Corresponding author

Email addresses:

Jian Wang: wangjian@dlut.edu.cn
Xiaoyu Chen: xiaoyu1127@mail.dlut.edu.cn
Yijia Zhang: zhyj@dlut.edu.cn
Yu Zhang: zhangxiaoyu041024@gmail.com
Jiabin Wen: wenjiabin2005@sina.com
Hongfei Lin: hflin@dlut.edu.cn

https://preprints.jmir.org/preprint/17638 [unpublished, non-peer-reviewed preprint]
Document-level Biomedical Relation Extraction Using Graph Convolutional Network and Multi-head Attention

Abstract

**Background:** Automatically extracting relations between chemicals and diseases plays an important role in biomedical text mining. Chemical-disease relation (CDR) extraction aims at extracting complex semantic relationships between entities in documents, which contain intra- and inter-sentence relations. Most previous methods do not consider dependency syntactic information across the sentences, which are very valuable for the relations extraction task, in particular for extracting the inter-sentence relations accurately.

**Methods:** In this paper, we propose a novel end-to-end neural network based on the graph convolutional network (GCN) and multi-head attention. To improve the performance of inter-sentence relation extraction, we construct the document-level dependency graph to capture the dependency syntactic information across sentences. GCN is applied to capture the feature representation of the document-level dependency graph. The multi-head attention mechanism is employed to learn the relative important context features from different semantic subspaces. To enhance the input representation, the deep context representation is used in our model instead of traditional word embedding.

**Results:** The experimental results show that our method achieves an F-score of 63.5% which is superior to other state-of-the-art methods. The GCN model can effectively exploit the across sentence dependency information to improve the performance of inter-sentence CDR extraction. Both the deep context representation and multi-head attention are helpful in CDR extraction task.

**Keywords:** Biomedical relation extraction, Dependency Graph, Multi-head attention, Graph convolution network
Introduction

The valuable biomedical information and knowledge are still hidden in the exponentially growing biomedical literature, such as chemical-disease relation (CDR). Extracting the relation between chemicals and diseases is an important task in biomedical text mining, which plays an important role in various biomedical researches, such as clinical medical treatment, drug development and biomedical knowledge discovery [1], [2], [3]. However, extracting CDR from the biomedical literature manually is time-consuming and difficult to keep up-to-date. Thus, the BioCreative V community [4] proposes a task of extracting CDR in the biomedical literature automatically to promote the research on the CDR extraction.

So far, a lot of methods have been proposed for automatic relation extraction between chemicals and diseases, which can be divided into three categories: rule-based methods [5], feature-based methods [6], [7], [8], [9] and deep neural network-based methods [10], [11], [12], [13]. Rule-based methods aim to formulate the heuristic rules for CDR extraction. Lowe et al. [5] developed a pattern-based system with some heuristic rules to extract chemical-induced disease (CID) relations within the same sentence. The heuristic rules are used to extract the most likely CID relations, when no patterns match a document. Generally speaking, rule-based methods are simple and effective. But the rule-based methods are difficult to be applied to a new task or dataset.

Feature-based methods aim at designing rich features including semantic and syntactic information. Xu et al. [6] utilized the text features including context information and entity information incorporating with domain knowledge to extract CID relations. Since the syntactic information carried in the dependency graph of the sentence is crucial to CDR extraction, some studies also developed the syntactic features. Gu et al. [7] utilized various linguistic features to extract CID relations with maximum entropy model. They leveraged lexical features for both intra- and inter-sentence level relation extraction and developed the dependency features only for intra-sentence level relation extraction. Zhou et al. [8] utilized the shortest dependency path (SDP) between chemical and disease entities to extract structured syntactic features. Feature-based methods achieve better performance than rule-based methods. However, traditional feature-based methods only use the dependency trees to extract local syntactic dependencies for the intra-sentence level relation extraction, without considering the syntactic dependencies across sentences for the document-level relation extraction. Besides, designing the rich features is a time-consuming and laborious work.

Recently, the deep neural network has been widely used in various natural language processing (NLP) tasks. Some studies have developed deep neural network-based methods for the biomedical relation extraction. Long short-term memory (LSTM) models and convolutional neural network (CNN) models are two major neural networks. Zhou et al. [10] applied LSTM and CNN models based on traditional word embedding to capture context features for CDR extraction and achieve a good performance. Gu et al. [11] proposed a CNN-based model to capture context and dependency features for intra-sentence level relation extraction. Nguyen et al. [13] investigated character-based word embedding into the CNN-based relation extraction model. Traditional word embedding, such as word2vec, cannot vary according to linguistic contexts effectively. Peters et al. [14] proposed deep contextualized word representations called ELMo based on a deep bidirectional language model. ELMo can generate a more comprehensive representation for each word based on the sentence context. Therefore, integrating ELMo with a deep neural network may improve the performance of CDR extraction.
In both CNN-based and LSTM-based models, it is hard to distinguish the relevant and irrelevant context features for the relation extraction. Recent work [15] suggested attention mechanism can capture the most important semantic information for the relation extraction. Vaswani et al. [16] introduced a multi-head attention mechanism that applied the self-attention mechanism multiple times to capture the relatively important features from different representation subspaces. Thus, multi-head attention mechanism can be used to improve the performance of the CDR extraction.

Dependency trees are often used to extract local dependencies for intra-sentence level CDR extraction. However, existing studies ignored the non-local dependency across sentences which is crucial to inter-sentence level CDR extraction. Quirk et al. [17] introduced a document graph which can derive features within and across sentences. Thus, we also construct a document-level dependency graph that can extract dependencies for intra- and inter-sentence level CDR extraction simultaneously. Recently the graph convolution network (GCN) [18] has been effectively used for encoding document graph information. Thus, GCN can operate directly on the document-level dependency graph to capture long-range syntactic information which is useful for CDR extraction.

In this paper, we evaluate the effectiveness of the deep contextualized word representations, multi-head attention mechanism and GCN in the CDR extraction task. To improve the performance of inter-sentence relation extraction, we construct the document-level dependency graph to capture the dependency syntactic information across sentences. Based on the document-level dependency graph, we propose a novel end-to-end model to extract CID relations from the biomedical literature. Firstly, we use ELMo, POS embedding and position embedding to construct the input representation, and employ the multi-head attention with Bi-LSTM to capture the relatively important context features. Then, we employ the GCN to capture the long-range dependency features based on the document-level dependency graph. Next, we combine the context features and long-range dependency features as the final feature representation, and apply a Softmax function to implement relation classification. Finally, we evaluate our model on CDR corpus.

Methods

Chemical-disease relation extraction

CDR extraction task is a challenging task which is proposed by the BioCreative V community. It aims to extract CDR from biomedical literature automatically and accurately. It is composed of two subtasks: (1) Disease named entity recognition and normalization (DNER); (2) Chemical-induced diseases (CID) relation extraction. In this paper, we focus on CID relation extraction task.

CDR extraction task is a document-level biomedical relation extraction problem, which is different from traditional biomedical relation extraction task. Traditional biomedical relation extraction only considers relation within a single sentence, such as protein-protein interaction (PPI) [19], drug-drug interaction (DDI) [20]. However, the CID relation is not only expressed within a single sentence, but also expressed across several sentences. Figure 1 shows an illustration of CDR extraction. It is extracted from the CDR corpus whose PMID is 6203632. Among these sentences, the texts in bold are mentions of chemical and disease entities. In Figure 1, we mark the corresponding entity type and Medical Subject Headings concept identifiers (MeSH ID) [21] after the entity mention in the sentence.
As we can see in Figure 1, the chemical D007545 has two intra-sentence level co-occurrences with disease D006332 in the sentence 1 and the sentence 2, while it has an inter-sentence level co-occurrence with disease D006965. However, not all occurrences of chemicals and diseases are considered as a CID relation. For example, the chemical D007545 doesn’t have a CID relation with the disease D006984 in the sentence 4 because the concept of the disease D006984 is too general to reflect a CID relation.

![Figure 1 The illustrative examples of CID relation.](https://preprints.jmir.org/preprint/17638)

**Relation instance construction**

Firstly, we should construct relation instances for both training and testing stages. All the instances generated from the disease and chemical mentions in the document are pooled into two groups at intra- and inter-sentence level, respectively. The former means a chemical-disease mention pair is in the same sentence. The latter means a mention pair is in a different sentence. If the relation between the chemical and disease entity of the mentioned pair is annotated as a CID relation in the document, then this mention pair is constructed as a positive instance; otherwise, this mention pair is constructed as a negative instance. We apply several effective heuristic rules for both intra- and inter-sentence level instances. The details are as follows:

1) Relation Instance Construction for Intra-Sentence Level

   (1) All chemical-disease entity mention pairs that appear in the same sentence are constructed as intra-sentence level instances.

   (2) If multiple mentions refer to the same entity in a sentence, the mentions in the nearest distance should be constructed as an instance.

   For instance, chemical D007545 and disease D006332 in sentence 1 form an intra-sentence level positive instance, while chemical D007545 and disease D006984 in sentence 4 form an intra-sentence level negative instance.

2) Relation Instance Construction for Inter-Sentence Level

   (1) Only the chemical-disease entity pairs which are not involved in any intra-sentence level are considered as inter-sentence level instances.

   (2) If multiple mentions refer to the same entity, the chemical and disease mentions in the nearest distance are chosen.

   According to our heuristic rules, chemical D007545 in sentence 4 and disease D006965 in sentence 5 are regarded as an inter-sentence level instance because there are no mentions of them in the same sentence. Chemical D007545 in sentence 1 and disease D006965 in sentence 5 will be omitted because the distance of them is not the nearest.

   Also, chemical D007545 in sentence 4 and disease D006984 in sentence 5 are not
regarded as an inter-sentence level instance because chemical D007545 already has intra-sentence level co-occurrence with disease D006984 in sentence 4.

**Document-level dependency graph**

To generate features for entity pairs within and across sentences, we introduce a document-level dependency graph with nodes representing words and edges representing intra- and inter-sentence dependency relations. Figure 2 shows an example of document-level dependency graph for two sentences. In this paper, we use the following three types of intra- and inter-sentence dependency edges.

- **Syntactic dependency edge:** The syntactic structure is crucial to biomedical relation extraction. Hence, we use syntactic dependency edges derived from Stanford dependency syntactic parser as intra-sentential edges. For instance, “conj” denotes the syntactic relation between the word “stopped” and “followed” in the same sentence.

- **Adjacent sentence edge:** Dependencies between sentences are useful for document-level relation extraction. Thus, we consider the sentence as a node in a type of discourse dependency tree. And we add an edge between the dependency roots of adjacent sentences as an inter-sentential edge, which is a simple but effective approach. For instance, “next” denotes the syntactic relation between two sentences.

- **Self-node edge:** we add self-node edges to all the nodes of the graph in order to enable GCN to not only learn information based on neighbor nodes but also learn the node information itself.

![Diagram of document-level dependency graph](https://preprints.jmir.org/preprint/17638)

**Figure 2** An example document-level dependency graph for two sentences expressing a CID relation. The chemical and disease entity mention is highlighted in bold. For simplicity, we omit self-node edges.

**Model architecture**

The schematic overview of our model is shown in Figure 3. In short, our model mainly consists of four parts: the input representation layer, the Bi-LSTM layer, the multi-head attention layer and the GCN layer. The inputs of our model are text sequences. The input layer will generate a deep contextualized word representation for each word. Recent studies \[22\], \[23\] have suggested that the part of speech (POS) and the position of each word are useful for biomedical relation extraction. Hence, we concatenate the deep contextualized word representation and POS and position embedding as the whole word representation.
representation. The Bi-LSTM layer will get contextual features from the word representation. The multi-head attention layer will apply the self-attention mechanism multiple times to capture the relative semantic features from different representation subspaces. The GCN layer will operate over the document-level dependency graph to capture long-range syntactic features. We employ max pooling over the outputs of the multi-head attention layer and GCN layer and then concatenate these two vectors as the final representation. Finally, we employ a fully connected layer and the Softmax function to identify the CID relation. Our model will be described in detail in the following section.

Figure 3 Overview of our model. The input representation consists of ELMo, POS embedding and position embedding. In Multi-head self-attention layer, we only show the detailed self-attention computation for the word “administration”. In GCN layer, we only show the detailed graph convolution computation for the word “administration”.

**Input representation**

We use ELMo instead of traditional word representation in our model. Traditional word representation generates a fixed representation vector for the same word. However, ELMo is the function of the entire input sentence based on a bidirectional language model, so that it can generate different representation vectors for the same word according to the different sentence context.

Given a sequence \( \{t_1, t_2, \ldots, t_N\} \) denotes the word tokens in a sentence \( S \). Given a token \( t_i \), the forward language model calculates the probability of the token \( t_i \) based on the previous tokens \( \{t_1, t_2, \ldots, t_{k-1}\} \) of \( t_i \) in the sentence \( S \) as follows:

\[
p_{\text{forward}}(t_1, t_2, \ldots, t_N) = \prod_{k=1}^{N} p(t_k | t_1, t_2, \ldots, t_{k-1}) \quad (1)
\]

Similarly, the backward language model calculates the probability of the token \( t_i \) based on the back tokens \( \{t_1, t_2, \ldots, t_{k-1}\} \) of \( t_i \) in the sentence \( S \) as follows:

\[
p_{\text{backward}}(t_1, t_2, \ldots, t_N) = \prod_{k=1}^{N} p(t_k | t_{k+1}, t_{k+2}, \ldots, t_N) \quad (2)
\]
Combining the forward and the backward language models as a bidirectional language model that maximizes the log-likelihood as follows:

\[
\sum_{k=1}^{N} \left( \log p(t_1|t_1, t_2, \ldots, t_{k-1}) + \log p(t_N|t_{k+1}, t_{k+2}, \ldots, t_N) \right)
\]

ELMo can represent the semantic and syntactic information of the word. In our model, we use a linear combination of the hidden state in each layer of the bidirectional language model to generate a deep contextualized representation for word.

The POS and the position information of a word are crucial to biomedical relation extraction. Therefore, we also utilize POS embedding and position embedding to enhance the representation ability of the input. The POS embedding represents the POS feature of a word, and the position embedding reflects the relative distance between the word and the target entity. Given a word at position \(i\), we obtain its POS embedding \(w_{p,i}\) and position embedding \(w_{d,i}\) based on mapping matrices \(M_p\) and \(M_d\), respectively.

Finally, the whole word representations concatenate deep contextualized word representations, POS embedding and position embedding as follows:

\[
w_i = [w_{c,i}; w_{p,i}; w_{d,i}]
\]

**Bi-LSTM**

The LSTM model is a variant of RNN models that has been used in many NLP tasks successfully. The LSTM model overcomes the vanishing gradient problem by introducing a gating mechanism [24]. Therefore, it is suitable to capture the long-term dependency feature. The LSTM unit consists of three components: the input gate \(i_t\), the forget gate \(f_t\), and the output gate \(o_t\). At the time step \(t\), the LSTM unit utilizes the input word \(x_t\), the previous hidden state \(h_{t-1}\) and the previous cell state \(c_{t-1}\) to calculate the current hidden state \(h_t\) and cell state \(c_t\). The equations are as follows:

\[
f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{5}
\]

\[
o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{6}
\]

\[
g_t = \tanh(W_g x_t + U_g h_{t-1} + b_g) \tag{7}
\]

\[
i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{8}
\]

\[
c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{9}
\]

\[
h_t = o_t \odot \tanh(c_t) \tag{10}
\]

where \(W, U\) and \(b\) is weight and bias parameters, and \(\odot\) denotes element-wise multiplication.

In this paper, we use Bi-LSTM model that can capture the forward and backward context features simultaneously. The Bi-LSTM model combines a forward LSTM and a backward LSTM. Given the hidden state of the forward LSTM \(h_t^f\) and the hidden state of the backward LSTM \(h_t^b\), the final hidden state is concatenated as \(h_t = [h_t^f; h_t^b]\).

**Multi-head attention**

The Bi-LSTM model learns the context features from the input sequences automatically and effectively. However, these features make different contributions to the biomedical relation extraction. In our model, we capture the relatively important features by introducing multi-head attention mechanism. The essence of multi-head attention is
applying self-attention mechanism multiple times so that it may let model learn the relatively important features from different representation subspaces. The self-attention mechanism generates the output based on a query and a set of key-value pairs. The output is the weighted sum of the values, where the weight assigned to each value is computed by applying attention function to the query with the corresponding key. In our study, we deal with the output of the Bi-LSTM model by multi-head self-attention. And we use dot-product attention function instead of the standard additive attention function [25] as follows:

\[ \text{Attention}(Q, K, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d}} \right) V \]  

where \( Q, K, V \in \mathbb{R}^{n \times d} \) represent query, key and value matrixes, respectively. \( \sqrt{d} \) is the scaling factor, \( d \) is the dimension of the output of the Bi-LSTM model.

The main idea of the multi-head attention is applying the self-attention mechanism multiple times. If the multi-head attention contains \( h \) heads, the \( i \)-th attention head can be calculated as \( \text{head}_i = \text{Attention}(Q, K, V) \). Thus, the final multi-head attention is the concatenation of \( \text{head}_1, \text{head}_2, \ldots, \text{head}_h \) as \( \text{MultiHead}(Q, K, V) = \text{Concat} \{ \text{head}_1, \text{head}_2, \ldots, \text{head}_h \} W^D \). The output of the multi-head attention layer is a matrix of \( \mathbb{R}^{n \times d} \).

**Graph convolutional network**

Graph convolutional network (GCN) is an adaptation of convolutional neural network [26], which operates on graphs. Given a graph with \( n \) nodes, the graph structure can be represented as an \( n \times n \) adjacency matrix \( A \). In this paper, we convert document-level dependency graph into its corresponding adjacency matrix \( A \), where \( A_{ij} = 1 \) if there is a dependency edge going from token \( i \) to token \( j \), otherwise \( A_{ij} = 0 \). The dependency graph can be calculated as an undirected graph [27], which means \( A_{ij} = A_{ji} \). And we add a self-node edge to all the nodes in the graph, which means \( A_{ii} = 1 \). Since the degree of a node in the dependency graph varies a lot, this may bias the output representation towards favoring high-degree nodes regardless of the information carried in the node. To solve this issue, we normalize the activations in the graph convolution before feeding it through the nonlinearity. Finally, the graph convolution operation for node \( i \) at the \( l \)-th layer where \( h_i^{(l-1)} \) and \( h_i^{(l)} \) denotes the input representation and the output representation of node \( i \) can be defined as follows:

\[ h_i^{(l)} = \rho \left( \sum_{j=1}^{n} A_{ij} W^{(l)} h_j^{(l-1)}/d_i + b^{(l)} \right) \]

where \( W^{(l)} \) is the weight matrix, \( b^{(l)} \) is the bias vector, \( d_i = \sum_{j=1}^{n} A_{ij} \) is the degree of node \( i \) in the dependency graph, and \( \rho \) is an activation function (e.g., ReLU).

The GCN model takes the output of the Bi-LSTM model as the input word representation \( h_0^{(0)}, \ldots, h_n^{(0)} \), then we stack the graph convolution operation over \( L \) layers and get \( h_0^{(L)}, \ldots, h_n^{(L)} \) as the output word representations of the GCN model.

Note that the GCN model presented above uses the same parameters for all edges in the dependency graph.

**Relation classification**

To make use of the output word representation of the GCN model for relation extraction,
we generate the sentence representation as follows:

\[ h_{\text{sent}} = f(h^{L}) = f(\cdot) \]

where \( h^{L} \) denotes the output representations at the last layer \( L \) of the GCN model, and \( f: \mathbb{R}^{n \times d} \rightarrow \mathbb{R}^{d} \) is a max pooling function that maps \( n \) output vectors to the sentence vector.

Inspired by recent work [28], [29], entity information is central to relation classification. Therefore, we also obtain the chemical entity representation \( h_c \) as

\[ h_c = f(h_{\text{chemical}}^{L}) \]

Similarly, we can obtain the disease entity representation \( h_d \). The feature representation of the whole GCN model is \( h_{\text{GCN}} = [h_{\text{sent}}; h_c; h_d] \).

We also obtain the feature representation \( h_{\text{att}} \) from the output of the multi-head attention layer by applying max pooling to the multi-head attention matrix. Finally, we concatenate \( h_{\text{GCN}} \) and \( h_{\text{att}} \) to form the final representation \( h_{\text{final}} = [h_{\text{GCN}}; h_{\text{att}}] \) for relation classification. Then the final representation is fed into a 2-layer perceptron as follows:

\[ h_1 = \text{ReLU}(W_1 h_{\text{final}} + b_1) \] (14)

\[ h_2 = \text{ReLU}(W_2 h_1 + b_2) \] (15)

where \( W_1, W_2 \) are weight matrices, \( b_1, b_2 \) are bias vectors.

Finally, the hidden representation \( h_2 \) is fed to a Softmax function to calculate the confidence of the CID relation:

\[ o = \text{softmax}(W_o h_2 + b_0) \] (20)

where \( o \) is the output, \( W_o \) is the weight matrix and \( b_0 \) is the bias vector.

**Results and discussion**

**Dataset**

We evaluate our model on the CDR corpus which is released by BioCreative V task. The CDR dataset is the benchmark dataset for the CID relation extraction task which consists of 1500 PubMed abstracts: 500 each for training, development and test set. Table 1 shows the details of the dataset.

| Task dataset | No. of Abstracts | No. of CID relations |
|--------------|------------------|----------------------|
| Training     | 500              | 1038                 |
| Development  | 500              | 1012                 |
| Test         | 500              | 1066                 |
| Total        | 1500             | 3116                 |

In this paper, the gold entity annotations provided by BioCreative V are used to evaluate our model. All the comparison methods reported in this paper are evaluated with gold entity annotations. Therefore, it is fair and comparable. And we measure the CID relation extraction performance with Precision(P), Recall(R), and F-score(F).

**Experimental settings**

The dimensions of POS embedding and position embedding are both 100. The dimension of ELMo is 1024. The dimensions of the LSTM hidden layer and the GCN layer are 500 with the dropout proportion \( p=0.5 \). The dimensions of 2-layer perceptron
are also 500 with the dropout proportion $p=0.5$. Our model is trained by Adam [30] with a learning rate 0.001 and a mini-batch size of 32.

In addition, our model is implemented based on an open-source deep learning library PyTorch\(^1\). We use StanfordNLP\(^2\) to obtain the POS of the word and the dependency tree. And we use AllenNLP\(^3\) pre-trained ELMo representations for the deep contextualized word representations.

### Experimental results

#### Effect of input representation

We evaluate the effectiveness of the input representation of our model. We use the same model that we propose and change the input representations. The comparison performance of different input representations is presented in Table 2.

| Input representation                  | P(%) | R(%) | F(%) |
|--------------------------------------|------|------|------|
| Word                                 | 47.3 | 71.7 | 57.0 |
| Word+position                        | 49.1 | 71.4 | 58.2 |
| Word+position+POS                    | 51.6 | 71.8 | **60.1** |
| ELMo                                 | 57.0 | 67.4 | 61.8 |
| ELMo+position                        | 54.2 | 74.9 | 62.9 |
| ELMo+position+POS                    | 56.3 | 72.7 | **63.5** |
| BioBERT+position+POS                 | 57.9 | 70.1 | 63.4 |

‘Word’: the input representation of the model is the word embedding which is pre-trained by word2vec.

‘Word+position’: the input representation of the model is the concatenation of the word embedding and position embedding.

‘Word+position+POS’: the input representation of the model is the concatenation of the word embedding, position embedding and POS embedding.

‘ELMo’: the input representation of the model is the deep contextualized word representation.

‘ELMo+position’: the input representation of the model is the deep contextualized word representation and position embedding.

‘ELMo+position+POS’: the input representation of the model is the deep contextualized word representation, position embedding and POS embedding.

‘BioBERT+position+POS’: the word representation is generated from the last hidden layer of the BioBERT [31] in a feature-based approach which means the parameters of the BioBERT are not fine-tuned. The input representation of the model is the BioBERT word representation, position embedding and POS embedding.

In Table 2, we can observe that the model achieves an F-score of 57.0% when we only use the pre-trained word embedding as the input representation. When we concatenate the pre-trained word embedding and position embedding, the F-score is improved from 57.0% to 58.2% which yields a 1.2% improvement. When we concatenate the pre-trained word embedding, position embedding and POS embedding as the input representations, we yield another 1.9% improvement compared with only using the pre-trained word embedding and position embedding. The result indicates that

---

\(^1\) https://pytorch.org/
\(^2\) https://nlp.stanford.edu/software/
\(^3\) https://allennlp.org/elmo
both POS and position features are effective for the CID relation extraction. The deep contextualized word representation ELMo significantly outperforms the pre-trained word embedding and yield a 4.8% improvement of F-score. The result indicates that ELMo can generate a more comprehensive representation for the word according to the sentence context, which results in a better CDR performance. Similarly, combining the position and POS embedding with the deep contextualized word representation can further improve the performance. When we concatenate the deep contextualized word representation, position embedding and POS embedding as the input representation, we achieve the best F-score of 63.5%. We also use the word representations generated from the BioBERT in a feature-based approach and achieve an F-score of 63.4% which is similar with using ELMo.

**Effect of the attention mechanism**

We evaluate the effectiveness of multi-head self-attention mechanism. We use the same model architecture that we propose, but deal with the output of Bi-LSTM by different attention mechanism. The attention mechanism is divided into two categories: single-head attention mechanism and multi-head attention mechanism. In single-head attention mechanism, we use three kinds of attention function: additive attention, general attention and scaled dot-product attention as follows:

\[
\begin{align*}
    h_{\text{att}} &= \sum_{i=1}^{N} \alpha_i h_i \quad (21) \\
    \alpha_i &= \frac{\exp(e_i)}{\sum_i \exp(e_i)} \quad (22) \\
\end{align*}
\]

\[
\begin{align*}
    e_i &= v^T \tanh(W_1 h_i + W_2 s) \quad \text{additive attention} \\
    e_i &= s^T W h_i \quad \text{general attention} \\
    e_i &= \frac{s^T h_i}{\sqrt{d}} \quad \text{scaled dot product attention} \quad (23)
\end{align*}
\]

where \( h \) is the output of the Bi-LSTM, \( W_1, W_2, s, v \) are parameter matrices, \( \sqrt{d} \) is the scaling factor, \( d \) is the dimension of the output of the Bi-LSTM model. The formula of the multi-head attention is described in formula (11). The comparison performance of different attention mechanism is presented in Table 3.

| Attention mechanism     | P(%) | R(%) | F(%) |
|-------------------------|------|------|------|
| w/o attention           | 55.1 | 71.3 | 62.2 |
| additive attention      | 55.9 | 70.3 | 62.3 |
| general attention       | 55.3 | 71.8 | 62.5 |
| scaled dot-product      | 54.9 | 73.3 | 62.8 |
In Table 3, we can see that using the attention mechanism can improve the performance of the CID relation extraction. The multi-head attention mechanism is more helpful than other single-head attention mechanisms. This suggests that the multi-head attention mechanism can capture more valuable features from different representation subspaces.

Effect of the attention heads

We evaluate the effectiveness of the number of heads of the multi-head attention mechanism. In this comparative experiment, we use the deep contextualized word representation, position embedding and POS embedding as the input representation and the dimensions of query Q, key K and value V are the same. As is shown in Table 4, we only vary the number of heads of the multi-head attention.

Table 4 The effect of the attention heads on performance

| Heads | P(%)  | R(%)  | F(%)  |
|-------|-------|-------|-------|
| 2     | 57.2  | 68.2  | 62.2  |
| 4     | 56.9  | 70.6  | 63.0  |
| 5     | 56.3  | 72.7  | **63.5** |
| 8     | 57.0  | 70.2  | 62.9  |
| 10    | 54.4  | 75.4  | 63.2  |

In Table 4, we can see that the multi-head attention mechanism can effectively improve the performance of the CID relation extraction. We can observe that the F-score ranges from 62.2% to 63.5% when setting a different number of heads. When the number of heads is too little or too large, the performance will drop off. In short, we achieve the best F-score of 63.5% when we set the number of heads as 5.

Ablation study

To examine the contributions of two main components, namely, multi-head attention layer and GCN layer, we run an ablation study. The experimental results are shown in Table 5. The results contain intra-sentence level, inter-sentence level and relation merging which means merging the intra- and inter-sentence level results as the final document-level result.

Table 5 An ablation study for our model

| Model          | Intra-sentence level | Inter-sentence level | Relation merging |
|----------------|----------------------|----------------------|------------------|
|                | P(%) | R(%) | F(%) | P(%) | R(%) | F(%) | P(%) | R(%) | F(%) |
| w/o multi-head | 58.2 | 82.9 | 68.4 | 44.7 | 44.3 | 44.5 | 55.1 | 71.3 | 62.2 |
| w/o GCN        | 62.6 | 74.1 | 67.9 | 43.6 | 48.4 | 45.9 | 57.1 | 66.4 | 61.4 |
| Ours           | 59.1 | 81.5 | **68.5** | 47.8 | 52.2 | **49.9** | 56.3 | 72.7 | **63.5** |

We can observe that removing either multi-head attention layer or GCN layer reduces
the performance of the model. This suggests that both layers can learn effective features. When we remove the multi-head attention layer and the GCN layer, the F-score drops by 1.3% and 2.1%, respectively.

Especially, we can observe that adding either multi-head attention layer or GCN layer improves the performance in the inter-sentence level relation extraction by a large margin. When we remove the multi-head attention layer and GCN layer, the inter-sentence level F-score drops by 5.4% and 4.0%, respectively. This suggests that the multi-head attention layer can capture the relatively important features from different representation subspaces and the GCN layer can capture long-range syntactic features for inter-sentence level relation extraction.

**Comparison with related work**

We compare our model with several state-of-the-art methods of the CID relation extraction. These methods are divided into two categories: methods without additional resources (without KBs) and methods using additional resources (with KBs). The results are presented in Table 6.

Pattern rule-based: Lowe et al. [5] developed a pattern-based system with some heuristic rules to extract CID relations within the same sentence, and achieved an F-score of 60.8%.

ME: Gu et al. [7] developed a machine learning-based system that utilized simple but effective manual linguistic features with maximum entropy model. They built rich manual features for intra-sentence level and inter-sentence level instances, respectively. And they achieved an F-score of 58.3%.

LSTM+SVM: Zhou et al. [10] developed a hybrid system which consists of a feature-based model that utilized flat features and structure features with SVM and a neural network model based on LSTM. Their model achieved an F-score of 56.0. After using additional post-processing heuristic rules, they achieved a 5.3% improvement of F-score.

CNN+ME: Gu et al. [11] proposed a maximum entropy model for inter-sentence level relation extraction and a convolutional neural network model for intra-sentence level relation extraction. They achieved an F-score of 60.2%. And they also used additional post-processing heuristic rules to improve performance that makes F-score raise to 61.3%.

BRAN: Verga et al. [12] proposed the Biaffine Relation Attention Network based on multi-head self-attention model which can predict relationships between all mention pairs in the document. The model achieved an F-score of 62.1%.

GCNN: Sahu et al. [18] proposed a labelled edge graph convolutional neural network model on a document-level graph. The model achieved an F-score of 58.6%.

SVM_Xu: Xu et al. [6] explored four different knowledge bases to extract the knowledge features and achieve an F-score of 67.2%.

SVM_Pons: Pons et al. [9] extracted three sets of features which are prior knowledge, statistical and linguistic information from the document. They achieve an F-score of 70.2%.

KCN: Zhou et al. [32] proposed a convolutional neural network which integrated both relation representations and entity representations learned from KBs. The model achieved an F-score of 71.3%.

| Category       | System       | Method                        | P(%) | R(%) | F(%) |
|---------------|--------------|-------------------------------|------|------|------|
|               |              |                               |      |      |      |

Table 6 Comparison with related work
In Table 6, the deep neural network-based methods achieved competitive performance in the CID relation extraction task. For example, Sahu et al. [18] use graph convolutional network to capture dependency information and achieve an F-score of 58.6%. Compared with other deep neural network-based methods, we not only employ the multi-head attention to capture the relatively important semantic features but also use the graph convolutional network to capture the valuable syntactic features from the document-level dependency graph automatically and effectively. We also observe that some studies [7], [10], [11] design and extract rich semantic and syntactic features for the relation extraction task and use additional post-processing heuristic rules to improve performance. Our method is an end-to-end neural network based model and achieves a high F-score of 63.5% without using post-processing heuristic rules.

As we can see in Table 6, the methods with KBs outperform the methods without KBs significantly. This suggests that prior knowledge is much useful for CID relation extraction. In this paper, we focus on the effectiveness of GCN and multi-head attention mechanism rather than the prior knowledge. We will attempt to integrate the biomedical knowledge to further improve the performance of our method in our future work.

**Visualization of Multi-head Attention Mechanisms**

To understand our multi-head self-attention mechanism clearly, we visualize the attention weights of an example sequence in Figure 4. Different colors represent different heads. The darker the color is, the higher the attention weight is. In Figure 4, the word pays different levels of attention to different words in different heads. For the word ‘Cardiac’, the word ‘Pilsicainide’ has the higher weight score in the second head, however, the words ‘Torsades’ and ‘Pointes’ have the higher weight score in the last head. For the word ‘Pilsicainide’, the words ‘Cardiac’ and ‘Death’ have the higher weight score in the fourth head. Thus, the multi-head self-attention mechanism can make model capture the relatively important features from different representation subspaces.
Error analysis

To understand our model better, we perform an error analysis on the output of our final results. There are two main types of errors: false positive (FP) errors and false negative (FN) errors. We list some examples to analyze the errors. In FP errors, some instances are non-relations but are mistaken as CID relations. For the sentences “Carbamazepine(Chemical: D002220)-induced cardiac dysfunction(Disease: D006331),” and “A patient with sinus bradycardia and atrioventricular block(Disease: D054537), induced by carbamazepine(Chemical: D002220),” the disease D006331 is the hypernym of the disease D054537. According to the labeling rules of the CDR corpus, we need to extract the most specific relations. Thus, the first sentence doesn’t express a CID relation and the second sentence expresses a CID relation. However, our model extracts a CID relation between the chemical D002220 and the disease D006331 in the first sentence incorrectly because the first sentence is the common sentence pattern that expresses a CID relation. In FN errors, several CID relations are not recognized. One of the main reasons is some inter-sentence level instances are removed by the heuristic rules in the relation instance construction stage because the sentence distance is more than 3. In the future, we will consider preferable pre-processing and post-processing techniques to solve the above problems.

Conclusions

In this paper, we propose a novel end-to-end neural network based on GCN and multi-head attention. The document-level dependency graph is constructed to capture the dependency syntactic information across sentences. We apply GCN to capture the long-range dependency syntactic features, which can improve the performance of inter-sentence level relation extraction. And we employ the multi-head attention mechanism to capture the relative important context features from different semantic subspaces. ELMo is used in our model to enhance the input representation. We evaluate the effectiveness of ELMo, multi-head attention mechanism and GCN on the BioCreative V CDR dataset. Experimental results show that ELMo, multi-head attention and GCN can significantly improve the performance of the CDR extraction. Our method achieves an F-score of 63.5%, which is superior to other state-of-the-art methods.

There are many large-scale knowledge bases, such as Comparative Toxicogenomics Database (CTD), UMLS, MeSH, UniProt and the commercial system Euretos.
Knowledge Platform, etc. These knowledge bases contain a great amount of structured data as the form of triples (entity, relation, entity), where relation represents the relationship between two entities. Some works suggest that integrating the structured information from the knowledge bases may improve the performance of the CDR extraction. In future work, we will integrate the biomedical knowledge to further improve the performance of our method.

Acknowledgements

Authors’ contributions: Jian Wang and Yijia Zhang leaded the method application, experiment conduction and the result analysis. Xiaoyu Chen, Yu Zhang and Jiabin Wen participated in the data extraction and preprocessing. Yu Zhang and Xin Wang participated in the manuscript revision. Hongfei Lin and Zhihao Yang provided theoretical guidance and the revision of this paper.

Funding: The work is supported by grants from National Natural Science Foundation of China (No. 61572098 and 61572102).

We would like to thank the Natural Science Foundation of China. We also would like to thank all anonymous reviewers for valuable suggestions and constructive comments.

Conflicts of Interest

None declared

Abbreviations

CDR: Chemical-disease relation; CID: Chemical-induced disease; CTD: Comparative Toxicogenomics Database; DNER: Disease named entity recognition and normalization; DDI: Drug-drug interaction; PPI: Protein-protein interaction; MeSH ID: Medical Subject Headings concept identifiers; SDP: Shortest dependency path; POS: Part of speech; NLP: Natural language processing; CNN: Convolutional neural network; LSTM: Long short-term memory; ELMo: Deep context representation; GCN: Graph convolutional network; P: Precision; R: Recall; F: F-score; FP: False positive; FN: False negative

References

1. Islamaj Dogan R, Murray G C, Névéol A. Understanding PubMed® User Search Behavior Through Log Analysis. Database. 2009; 2009: bap018.
2. Hurle MR, Yang L, Xie Q, Rajpal DK, Sanseau P, et al. Computational drug repositioning: from data to therapeutics. Clinical Pharmacology & Therapeutics. 2013; 93(4): 335-341.
3. Qu JF, Ouyang D, Hua W, Ye YX, Li XM. Distant supervision for neural relation extraction integrated with word attention and property features. Neural Networks. 2018; 100: 59-69.
4. Wei CH, Peng YF, Leaman R. Assessing the state of the art in biomedical relation extraction: Overview of the BioCreative V Chemical Disease Relation (CDR) Task. Database. 2016; 2016: baw032.
5. Lowe DM, O’Boyle, Noel M, Sayle RA. Efficient chemical-disease identification and relationship extraction using Wikipedia to improve recall. Database. 2016; 2016: baw039.
6. Xu J, Wu YH, Zhang YY, et al. CD-REST: a system for extracting chemical-induced disease relation in literature. Database. 2016; 2016: baw036.
7. Gu JH, Qian LH, Zhou GD. Chemical-induced disease relation extraction with various linguistic features. Database. 2016; 2016: baw042
8. Zhou HW, Deng HJ, He J. Chemical-disease relations extraction based on the shortest dependency path tree. In: Proc. the fifth BioCreative challenge evaluation workshop, 2015, pp. 240-246.
9. Pons E, Becker BFH, Akhondi SA, Afzal Z, Van Mulligen EM, et al. Extraction of chemical-induced diseases using prior knowledge and textual information. Database. 2016; 2016: baw046.
10. Zhou HW, Deng HJ, Chen L, Yang YL, et al. Exploiting syntactic and semantics information for chemical-disease relation extraction. Database. 2016; 2016: baw048.
11. Gu JH, Sun FQ, Qian LH, Zhou GD. Chemical-induced disease relation extraction via convolutional neural network. Database. 2017; 2017: bax024.
12. Verga P, Strubell E, McCallum A. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. In: Proceedings of NAACL-HLT 2018, New Orleans, USA, 2018.
13. Nguyen DQ, Verspoor K. Convolutional neural networks for chemical-disease relation extraction are improved with character-based word embeddings. In: Proceedings of the BioNLP 2018 workshop, 2018.
14. Peters ME, Neumann M, Iyyer M, et al. Deep contextualized word representations. In: Proceedings of NAACL-HLT 2018, New Orleans, USA, pp. 2227–2237, 2018.
15. Li LF, Nie YP, Han WH, Huang JM. A Multi-attention-Based Directional Long Short-Term Memory Network for Relation Extraction. In: Liu D., Xie S., Li Y., Zhao D., El-Alfy ES. (eds) Neural Information Processing. ICONIP 2017. Lecture Notes in Computer Science, vol 10638. Springer, Cham
16. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, USA, 2017, pp. 5998–6008.
17. Quirk C, Poon H. Distant supervision for relation extraction beyond the sentence boundary. arXiv preprint arXiv:1609.04873, 2016.
18. Sahu SK, Christopoulou F, Miwa M, et al. Inter-sentence Relation Extraction with Document-level Graph Convolutional Neural Network. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, pp. 4309–4316, 2019.
19. Krallinger M, Vazquez M, Leitner F, et al. The Protein-Protein Interaction Tasks of BioCreative III: Classification/Ranking of Articles and Linking Bio-Ontology Concepts to Full Text. BMC Bioinformatics. 2011; 12 Suppl 8(Suppl 8): S3.
20. Segura-Bedmar I, Martínez P, María H. Lessons learnt from the ddiextraction-2013 shared task. Journal of Biomedical Informatics. 2014; 51: 152-164.
21. Coletti MH, Bleich HL. Medical subject headings used to search the biomedical literature. Journal of the American Medical Informatics Association. 2001; 8(4): 317-323.
22. Zhao ZH, Yang ZH, Luo L, et al. Drug drug interaction extraction from biomedical literature using syntax convolutional neural network. Bioinformatics. 2016; 32: 3444–3453.
23. Zhang YJ, Zheng W, Lin HF, et al. Drug–drug interaction extraction via hierarchical RNNs on sequence and shortest dependency paths. Bioinformatics. 2018; 34: 828–835.
24. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation. 1997; 9(8): 1735-1780.
25. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. Computer Science. 2014.
26. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proceedings of the IEEE. 1998; 86(11): 2278–2324.
27. Zhang YH, Qi P, Manning CD. Graph Convolution over Pruned Dependency Trees Improves Relation Extraction. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, pp. 2205-2215, 2018.
28. Santoro A, Raposo D, Barrett D, Malinowski M, et al. A simple neural network module for relational reasoning. In Advances in neural information processing systems, 2017, pp. 4974-4983.
29. Lee K, He LH, Lewis M, Zettlemoyer L. End-to-end neural coreference resolution. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017.
30. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. Computer Science. 2014.
31. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020; 36(4): 1234–1240.
32. Zhou HW, Lang CK, Liu Z, et al. Knowledge-guided convolutional networks for chemical-disease relation extraction. BMC Bioinformatics. 2019; 20(1).
Supplementary Files
Other materials for editor/reviewers onlies