Bioinformatics, 34(8), 2018, 1381–1388
doi: 10.1093/bioinformatics/btx761
Advance Access Publication Date: 24 November 2017
Original Paper

Data and text mining

An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition

Ling Luo[1], Zhihao Yang[1,*], Pei Yang[1], Yin Zhang[2], Lei Wang[2,*], Hongfei Lin[1] and Jian Wang[1]

[1]College of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China and [2]Beijing Institute of Health Administration and Medical Information, Beijing 100850, China

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on July 19, 2017; revised on October 26, 2017; editorial decision on November 19, 2017; accepted on November 23, 2017

Abstract

Motivation: In biomedical research, chemical is an important class of entities, and chemical named entity recognition (NER) is an important task in the field of biomedical information extraction. However, most popular chemical NER methods are based on traditional machine learning and their performances are heavily dependent on the feature engineering. Moreover, these methods are sentence-level ones which have the tagging inconsistency problem.

Results: In this paper, we propose a neural network approach, i.e. attention-based bidirectional Long Short-Term Memory with a conditional random field layer (Att-BiLSTM-CRF), to document-level chemical NER. The approach leverages document-level global information obtained by attention mechanism to enforce tagging consistency across multiple instances of the same token in a document. It achieves better performances with little feature engineering than other state-of-the-art methods on the BioCreative IV chemical compound and drug name recognition (CHEMDNER) corpus and the BioCreative V chemical-disease relation (CDR) task corpus (the F-scores of 91.14 and 92.57%, respectively).

Availability and implementation: Data and code are available at https://github.com/lingluodlut/Att-ChemdNER.

Contact: yangzh@dlut.edu.cn or wangleibihami@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Chemical named entity recognition (NER) aims to automatically detect the chemical mentions in biomedical literature, which is a fundamental step for further biomedical text mining and has received much attention recently. It is particularly challenging due to the following reasons: various ways of naming chemical, ambiguities caused by the frequent occurrences of abbreviations and acronyms, new chemical constantly and rapidly reported in scientific publications, and a number of symbols mixed with common words (Liu *et al.*, 2015). To promote the performance of chemical NER systems, chemical NER has been the subtasks of several public challenges in the biomedical domain such as the chemical and drug named entity recognition (CHEMDNER) task in the BioCreative IV (Krallinger *et al.*, 2015) and the drug-drug interaction (DDIExtraction) challenge (Segura Bedmar *et al.*, 2013).

In the previous works, the state-of-the-art CRF-based chemical NER methods (Leaman *et al.*, 2015; Lu *et al.*, 2015; Rocktäschel *et al.*, 2012; Usié *et al.*, 2014) depend on effective feature engineering, i.e. the design of effective features using various NLP tools and knowledge resources, which is still a labor-intensive and skill-dependent task. Recently, deep learning has become prevalent in the machine learning research community. For the NER task in general domain

(such as news), several neural network architectures have been proposed (Collobert *et al.*, 2011; Huang *et al.*, 2015; Lample *et al.*, 2016; Ma and Hovy, 2016). Furthermore, they have also been used to identify the biomedical entities, including genes and proteins (Li *et al.*, 2015), diseases (Sahu and Anand, 2016) and chemicals (Chalapathy *et al.*, 2016). Among others, the model of bidirectional Long Short-Term Memory with a conditional random field layer (BiLSTM-CRF), exhibits promising results (Huang *et al.*, 2015; Lample *et al.*, 2016).

Both the above-mentioned standard traditional machine learning methods and deep learning methods in practice treat NER as a sentence-level task, i.e. they treat each sentence as a separate document, with multiple instances of the same token in different sentences of a document viewed multiple entirely independent tagging problems. However, these sentence-level NER methods lead to the tagging inconsistency problem (i.e. the same mentions in a document are tagged with different labels). As shown in the example of Table 1, the mentions in italic type were recognized by the sentence-level BiLSTM-CRF model. In the biomedical abstract, the same chemical mentions may appear several times in the different sentences (e.g. 'VK2', the abbreviation of the chemical entity 'Vitamin K2'). Reasonably, these mentions should be tagged with the same labels. However, the mentions 'VK2' in bold type were not recognized by the model. This is so-called the tagging inconsistency problem. To alleviate the problem, the rule-based post-processing step to enforce tagging consistency is often employed in NER methods. In addition, there have been some efforts on enforcing tag consistency with the use of non-local information to improve supervised sequential models (Finkel *et al.*, 2005; Ratinov and Roth, 2009). But the consensus seems to be that the non-local information has a relatively modest effect on performance.

Our work focuses on the above-mentioned two issues: the performance dependency on the feature engineering for neural network architecture methods and tagging inconsistency of the sentence-level NER. And, in this paper we propose a novel attention-based bidirectional Long Short-Term Memory with a conditional random field layer (Att-BiLSTM-CRF) approach for document-level chemical NER to mitigate the issues. The main contributions of our work can be summarized as follows:

- Our neural network architecture relies on a novel attention mechanism to capture similar entity attention at the document-level. This allows it to view the related tokens in different sentences of a document as a dependent tagging problem. Moreover, we consider four different alternatives to compute the score of attention matrix. The experimental results show that our method can significantly improve the tagging consistency.
- Domain features used in traditional NER methods [such as part of speech (POS), chunking and dictionary features] with neural network architectures (including BiLSTM-CRF and our Att-BiLSTM-CRF) for chemical NER are investigated. The experimental results show that our method can achieve the state-of-the-art performance with little feature engineering.

Owing to the above contributions, our method achieves the state-of-the-art performances for chemical NER on the BioCreative IV CHEMDNER corpus and the BioCreative V chemical-disease relation (CDR) corpus (the F-scores of 91.14 and 92.57%, respectively).

Table 1. An example of the tagging inconsistency problem

Vitamin K2 covalently binds to Bak and induces Bak-mediated apoptosis. *Vitamin K2* (*VK2*, *menaquinone*) is known to have anticancer activity in vitro and in vivo. Although its effect is thought to be mediated, at least in part, by the induction of apoptosis, the underlying molecular mechanism remains elusive. Here, we identified Bcl-2 antagonist killer 1 (Bak) as a molecular target of **VK2**-induced apoptosis. **VK2** directly interacts with Bak and induces mitochondrial-mediated apoptosis. Although Bak and Bcl-2-associated X protein (Bax), another member of the Bcl-2 family, are generally thought to be functionally redundant, only Bak is necessary and sufficient for **VK2**-induced cytochrome c (cyt c) release and cell death. Moreover, *VK2-2, 3 epoxide*, an intracellular metabolite of **VK2**, was shown to covalently bind to the *cysteine-166* residue of Bak. Several lines of evidence suggested that the covalent attachment of **VK2** is critical for apoptosis induction. Thus this study reveals a specific role for Bak in mitochondria-mediated apoptosis. This study also provides insight into the anticancer effects of **VK2** and suggests that Bak may be a potential target of cancer therapy.

2 Materials and methods

In this section, firstly, embedded features used in our neural network model are described. Secondly, the basic BiLSTM-CRF model is introduced. At last, our Att-BiLSTM-CRF model is presented.

2.1 Features

Recently distributed feature representation is widely used in the field of NLP, especially for the deep learning methods. Our method uses word and character embeddings as basic features. In addition, to investigate the effects of traditional features (such as POS, chunking and dictionary features) for deep learning methods, these features are added into the models as additional features. Details of the features are presented as follows.

2.1.1 Word embedding

Word embedding, also known as distributed word representation, can capture both the semantic and syntactic information of words from a large unlabeled corpus and has attracted considerable attention from many researchers (Lai *et al.*, 2016). Compared with the bag-of-words (BOW) representation, word embedding is low-dimensional and dense. In recent years, several tools, such as word2-vec (Mikolov *et al.*, 2013) and GloVe (Pennington *et al.*, 2014) have been widely used in the field of NLP. To achieve a high-quality word embedding, we downloaded a total of 1 918 662 MEDLINE abstracts from the PubMed website with the query string 'chemical'. Then these abstracts and CHEMDNER corpus were used to train word embedding by the word2vec tool using the skip-gram model as pre-trained word embedding.

2.1.2 Character embedding

In addition to the word-level features, character-level features in an entity name contain rich structure information of the entity. These features (such as character n-grams, prefixed and suffixes) are commonly employed in the current chemical NER methods (Liu *et al.*, 2015). Unlike the previous traditional methods in which character features are based on hand-engineering, character embedding can be learned while training. They can not only learn interior representations of the entity names, but also alleviate the out-of-vocabulary problem (Rei *et al.*, 2016). First, a character lookup table containing a character embedding for every character was initialized randomly. Then the character embedding corresponding to every character in a word was given in both direct and reverse orders to a bidirectional Long Short-Term Memory (BiLSTM). At last, the concatenation of the forward and backward representations from the BiLSTM was used as the character-level feature of the word.

2.1.3 Additional features

In our experiments, the effects of two kinds of traditional manually designed features (i.e. linguistic feature and domain resource feature) for our neural network model are also explored.

Due to the complexity of the natural language, some linguistic features are often employed in the chemical NER systems based on traditional machine learning methods (Eltyeb and Salim, 2014). We also explored the effect of linguistic features (such as POS and chunking) for our model. The POS information and chunking information of each word were generated by the GENIA tagger (http://www.nactem.ac.uk/GENIA/tagger/). Then two lookup tables were used to output 25-dimensional POS embedding and 10-dimensional chunking embedding, respectively.

To optimize the chemical NER systems, chemical dictionaries as a form of domain knowledge are often added to the set of features. In fact, there are a large number of available chemical lexical resources currently, such as Jochem (Hettne *et al.*, 2009), ChEBI (Degtyarenko *et al.*, 2008) and CTD (Davis *et al.*, 2009). We used Jochem dictionary to generate our dictionary feature. Firstly, longest possible matches between the normalized token sequences and dictionary entries were captured. Then for each token in the match, the feature was encoded in BIO (Begin, Inside, Outside) tagging scheme. At last, a lookup table was used to output 5-dimensional dictionary embedding.

2.2 BiLSTM-CRF model

The architectures of basic BiLSTM-CRF model and our Att-BiLSTM-CRF model are illustrated in Figure 1A and B, respectively. The former is similar to the ones presented by Huang *et al.* (2015), Lample *et al.* (2016) and Ma and Hovy (2016).

Given a sentence, the model predicts a label corresponding to each of the input tokens in the sentence. Firstly, through the embedding layer, the sentence is represented as a sequence of vectors $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_n)$ where n is the length of the sentence. Next, the embeddings are given as input to a BiLSTM layer. In the BiLSTM layer, a forward LSTM computes a representation $\overrightarrow{\mathbf{h}_t}$ of the sequence from left to right at every word t, and another backward LSTM computes a representation $\overleftarrow{\mathbf{h}_t}$ of the same sequence in reverse. These two distinct networks use different parameters, and then the representation of a word $\mathbf{h}_t = [\overrightarrow{\mathbf{h}_t}; \overleftarrow{\mathbf{h}_t}]$ is obtained by concatenating its left and right context representations. LSTM memory cell is implemented as Lample *et al.* (2016) did.

Then a tanh layer on top of the BiLSTM is used to predict confidence scores for the word having each of the possible labels as the output scores of the network.

$$\mathbf{e}_t = \tanh(\mathbf{W}_e \mathbf{h}_t) \tag{1}$$

where the weight matrix $\mathbf{W}_e$ is the parameter of the model to be learned in training.

Finally, instead of modeling tagging decisions independently, the CRF layer is added to decode the best tag path in all possible tag paths. We consider $\mathbf{P}$ to be the matrix of scores output by the network. The t^{th} column is the vector $\mathbf{e}_t$ obtained by the Equation (1). The element $P_{i,\,j}$ of the matrix is the score of the j^{th} tag of the i^{th} word in the sentence. We introduce a tagging transition matrix $\mathbf{T}$, where $T_{i,\,j}$ represents the score of transition from tag i to tag j in successive words and $T_{0,\,j}$ as the initial score for starting from tag j. This transition matrix will be trained as the parameter of model. The score of the sentence $\mathbf{X}$ along with a sequence of predictions $\mathbf{y} = (y_1, \ldots, y_t, \ldots, y_n)$ is then given by the sum of transition scores and network scores:

$$s(\mathbf{X}, \mathbf{y}) = \sum_{i=1}^{n} (T_{y_{i-1}, y_i} + P_{i, y_i}) \tag{2}$$

Then a softmax function is used to yield the conditional probability of the path $\mathbf{y}$ by normalizing the above score over all possible tag paths$\tilde{\mathbf{y}}$:

$$p(\mathbf{y}|\mathbf{X}) = \frac{e^{s(\mathbf{X}, \mathbf{y})}}{\sum_{\tilde{\mathbf{y}}} e^{s(\mathbf{X}, \tilde{\mathbf{y}})}} \tag{3}$$

During the training phase, the objective of the model is to maximize the log-probability of the correct tag sequence. At inference time,

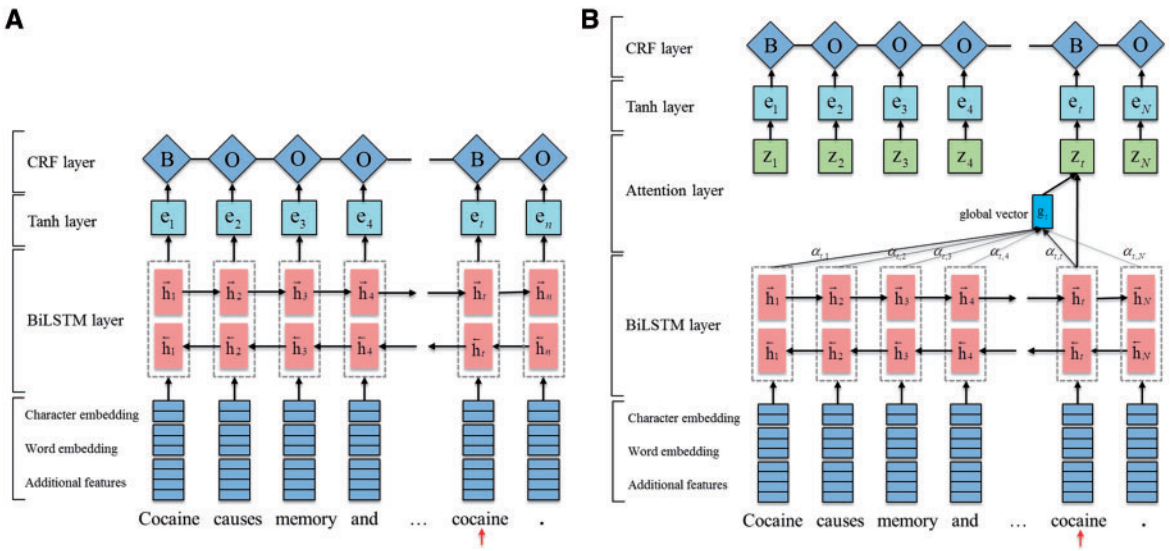

Fig. 1. The architectures of BiLSTM-CRF model and our Att-BiLSTM-CRF model. (**A**) The basic BiLSTM-CRF model. (**B**) Our Att-BiLSTM-CRF model. In the models, the BIO (Begin, Inside, Outside) tagging scheme are used. In the figure (**B**), only the attention weights of the target word are illustrated for clarity. The attention weight is larger when the word is more related to the target word, and the color of attention line is darker

we predict the best tag path that obtains the maximum score given by:

$$\arg\max_{\tilde{\mathbf{y}}} s(\mathbf{X}, \tilde{\mathbf{y}}) \tag{4}$$

This can be computed using dynamic programming, and the Viterbi algorithm (Viterbi, 1967) is chosen for this inference.

2.3 Att-BiLSTM-CRF model

Similar to most traditional machine learning NER methods, the above-mentioned BiLSTM-CRF method is also a sentence-level NER method, suffering from the tagging inconsistency problem. To solve the problem, previous works often employ rule-based post-processing to enforce tagging consistency. For example, if a mention was tagged by the NER model at least twice within a document, any of the mention in the document that the model had not identified was also tagged (Leaman *et al.*, 2015). However, these post-processing methods do not necessarily improve the performance of the model since if the entity is mistakenly tagged by the model, the post-processing will introduce more noise. Therefore, we design an Att-BiLSTM-CRF model so as to automatically ensure tagging consistency in a document. The architecture of our model is illustrated in Figure 1B.

Attention mechanism has gained popularity recently in image, speech and NLP fields (Bahdanau *et al.*, 2015; Mnih *et al.*, 2014). For NER task, Bharadwaj *et al.* (2016) and Rei *et al.* (2016) introduced the attention mechanism to enhance their model performances. Though both methods focus on the character-lever representations by the attention mechanism, they are still the sentence-level NER methods. More recently, Pandey *et al.* (2017) presented a model similar to our Att-BiLSTM-CRF to extract knowledge of Adverse Drug Reactions (ADRs). However, their attention mechanism focuses on which encoded elements contribute to the generation of the current unit or the prediction of an ADR in sentence-level. Different from them, we apply the attention mechanism to focus on the related tokens in the different sentences of a document to address the tagging inconsistency problem.

For an input document $\mathbf{D} = (\mathbf{X}_1, \ldots, \mathbf{X}_t, \ldots, \mathbf{X}_m)$ consisting of m sentences, each sentence is expressed as $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_n)$ where n is the the number of the words in the sentence. We define N as the number of the words in the document. Like BiLSTM-CRF model, the embeddings described in Section 2.1 are firstly given as input to a BiLSTM layer. Then a new attention layer on top of the BiLSTM layer is used to capture similar word attention at the document-level. In the attention layer, we introduce an attention matrix $\mathbf{A}$ to calculate the similarity between the current target word and all words in the document. The attention weight value $\alpha_{t,j}$ in the attention matrix is derived by comparing the t^{th} current target word representation $\mathbf{x}_t$ with the j^{th} word representation $\mathbf{x}_j$ in the document:

$$\alpha_{t,j} = \frac{\exp\left(score(\mathbf{x}_t, \mathbf{x}_j)\right)}{\sum_k \exp\left(score(\mathbf{x}_t, \mathbf{x}_k)\right)} \tag{5}$$

Here, the score is referred as an alignment function for which we define the following four alternatives (manhattan distance, euclidean distance, cosine distance and perceptron):

$$\text{score}(\mathbf{x}_t, \mathbf{x}_j) = \begin{cases} \mathbf{W}_a \left| \mathbf{x}_t - \mathbf{x}_j \right| \\ \mathbf{W}_a (\mathbf{x}_t - \mathbf{x}_j)^T (\mathbf{x}_t - \mathbf{x}_j) \\ \dfrac{\mathbf{W}_a (\mathbf{x}_t \bullet \mathbf{x}_j)}{|\mathbf{x}_t|\,|\mathbf{x}_j|} \\ \tanh(\mathbf{W}_a [\mathbf{x}_t; \mathbf{x}_j]) \end{cases} \tag{6}$$

where the weight matrix $\mathbf{W}_a$ is a parameter of the model, and $\bullet$ is the element-wise product. For the score of alignment function, the score values of cosine distance and perceptron are larger when the two vectors $\mathbf{x}_t$ and $\mathbf{x}_j$ are more similar. On the contrary, the score values of manhattan distance and euclidean distance are smaller when the two vectors are more similar. Therefore, the final scores of manhattan distance and euclidean distance are calculated by the maximum scores (the maximum score of the current target word representation with all word representations in the document) minus the scores to make their final scores larger when the vectors are more similar.

Then a document-level global vector $\mathbf{g}_t$ is computed as a weighted sum of each BiLSTM output $\mathbf{h}_j$:

$$\mathbf{g}_t = \sum_{j=1}^{N} \alpha_{t,j} \mathbf{h}_j \tag{7}$$

Next, the document-level global vector and the BiLSTM output of the target word are concatenated as a vector $[\mathbf{g}_t; \mathbf{h}_t]$ to be fed to a tanh function to produce the output of attention layer.

$$\mathbf{z}_t = \tanh(\mathbf{W}_g [\mathbf{g}_t; \mathbf{h}_t]) \tag{8}$$

Then a tanh layer on top of the attention layer is used to predict confidence scores for the word having each of the possible labels as the output score of the network:

$$\mathbf{e}_t = \tanh(\mathbf{W}_e \mathbf{z}_t) \tag{9}$$

At last, similar to BiLSTM-CRF model, the CRF layer is added to decode the best tag path in all possible tag paths. For an input document $\mathbf{D}$, the score of the document along with a tag path $\mathbf{y}$ is then given by the sum of transition scores and network scores:

$$s(\mathbf{D}, \mathbf{y}) = \sum_m \sum_{i=1}^{n} (T_{y_{i-1}, y_i} + P_{i, y_i}) \tag{10}$$

Compared with the Equation (2), the Equation (10) yields the score at the document-level instead of sentence-level.

Next, like BiLSMT-CRF model, a softmax function is used to yield the conditional probability of the path. During the training phase, the objective of the model is to maximize the log-probability of the correct tag sequence. Viterbi algorithm is used to compute optimal tag sequences for inference.

3 Results

3.1 Experimental datasets and settings

In our experiments, two corpora (http://www.biocreative.org/resources/) released by the BioCreative challenge were used: the CHEMDNER corpus (Krallinger *et al.*, 2015) and the CDR task corpus (Li *et al.*, 2016). Overall statistics for each dataset are provided in Supplementary Material: *BioCreative Corpora*. Like many teams in the challenge, the original training set and development set were used as the training set. Then we randomly selected 10% of the training set as the validation set to tune the hyper-parameters. The chemical NER performance was measured with an F-score which attributes equal importance to precision and recall (F1 score) on the test set.

The parameters of our model in the word embedding are initialized with 50-dimensional pre-trained word embeddings (The performances of the higher dimensional word embeddings were also tested, but no better performances were achieved. And the results are provided in Supplementary Material: *Performance of Word*

Embeddings.) and other parameters are initialized at random from a uniform distribution. Then all parameters are optimized using stochastic gradient descent (SGD) (Bottou, 1991) to maximize the log-probability of the correct tag sequence. In addition, we tuned the hyper-parameters on the validation set by random search (Bergstra and Bengio, 2012). The main hyper-parameters of our model are shown in Supplementary Material: *Hyper-parameter Settings*. The number of epochs is chosen by early stopping strategy (Prechelt, 1998) on the validation set. Our models are implemented using open-source deep learning library Theano (http://deeplearning.net/software/theano/).

3.2 The effect of alignment functions for the Att-BiLSTM-CRF model

As described in Section 2.3, four alignment functions (manhattan distance, euclildean distance, cosine distance and two-layer multilayer perceptron) are designed for our Att-BiLSTM-CRF model. To select the best alignment function, we investigated the effect of these functions on performance. In the experiments, word and character embeddings are used as inputs of the model and alignment function. The performances of the attention-based model with different alignment functions on the CHEMDNER corpus are shown in Table 2.

The results show that the model using the euclidean distance achieves the highest F-score of 90.84%. However, there is no significant difference among the F-scores of these alignment functions except the two-layer multilayer perceptron. The reason may be that euclildean, manhattan and cosine distances are common similarity measure methods that are simple and effective in machine learning. Compared with them, the two-layer multilayer perceptron has more complex structure so that it is difficult to optimize.

3.3 Performance comparison of document-level and sentence-level methods

In previous works, most NER methods treat NER as a sentence-level task. This is often accompanied by the tagging inconsistency problem as discussed in Section 1. In contrast, our Att-BiLSTM-CRF model is designed to alleviate this problem by the document-level attention. In our experiments, the two document-level and two sentence-level methods were compared (all models only used word and character embeddings as inputs).

Table 2. Performances of our attention-based model with different alignment functions on the CHEMDNER corpus

Method	Precision	Recall	F-score
Manhattan	91.23	89.92	90.57
Euclidean	**91.65**	**90.04**	**90.84**
Cosine	91.59	89.67	90.62
Perceptron	91.13	88.86	89.98

The bold values denote the highest values.

Table 3. Performance comparison of the document-level and sentence-level methods on the CHEMDNER corpus

Method	Precision	Recall	F-score
BiLSTM-CRF(sent)	91.31	87.73	89.48
BiLSTM-CRF(sent)+Post	90.38	89.81	90.10
BiLSTM-CRF(doc)	90.12	88.46	89.28
Att-BiLSTM-CRF (sent)	91.63	88.63	90.10
Att-BiLSTM-CRF (doc)	**91.65**	**90.04**	**90.84**

The bold values denote the highest values.

The performance comparisons of various models on the CHEMDNER corpus are shown in Table 3. Compared with the original sentence-level BiLSTM-CRF model described in Section 2.2 [BiLSTM-CRF(sent)], our document-level Att-BiLSTM-CRF model described in Section 2.3 [Att-BiLSTM-CRF(doc)] obtains better performance (improvements of 0.34, 2.31 and 1.36% in precision, recall and F-score, respectively). The result shows that the recall is improved significantly without loss of precision. We analyzed the results and found that the main reason of recall improvement is the alleviation of the tagging inconsistency problem. We performed an error analysis to help illustrate the reason of performance improvement of our model. We randomly selected 20% of the test set (600 abstracts) and reviewed the output results of the models. Overall, these abstracts contain 5155 chemical entity mentions corresponding to 2602 unique chemical entities in document-level (totally 966 entities appear more than once in a document). In the result of BiLSTM-CRF(sent), 95 entities recognized have the tagging inconsistency problem. Of these entities, 65 were tagged with the consistent labels by our Att-BiLSTM-CRF (doc) model, in which 48 were tagged correctly according to the gold standard. Some results of two models are provided in Supplementary Material: *Examples of NER Results* due to space limitation.

In addition, the result of the BiLSTM-CRF(sent) with a consistency post processing method (i.e. tagging any mention already tagged twice) is provided in Table 3 [BiLSTM-CRF(sent)+Post]. The results show our Att-BiLSTM-CRF(doc) method can achieve better performance than BiLSTM-CRF(sent) with the consistency post processing method. The method improves the recall of the entity by enforcing tagging consistency while, for some incorrectly tagged entities by the model, it will introduce noise (i.e. incorrectly tagging any mention already incorrectly tagged twice) leading to the precision decline.

For the BiLSTM-CRF model, we implemented a document-level version, BiLSTM-CRF(doc) (i.e. documents are directly used as inputs of the model instead of sentences). However, the document-level version does not achieve higher F-score than the sentence-level vision (89.28% versus 89.48%). The main reason is that LSTM model is a biased model. Although LSTM can solve hard long time lag problems with the gating mechanism (Hochreiter and Schmidhuber, 1997), the fact that later words are more dominant than earlier words leads to recognition difficulty on long sentences (Lai *et al.*, 2015). Therefore, it is not effective to capture global document-level information by simply extending sentences to documents as inputs of the model.

For our Att-BiLSTM-CRF model, we also investigated the effect of a sentence-level version, Att-BiLSTM-CRF(sent) (i.e. sentences are directly fed into the model instead of documents). It can achieve a higher F-score than the sentence-level BiLSTM-CRF model, but a lower one than the document-level Att-BiLSTM-CRF model (90.10% versus 89.48% and 90.84% in F-score). It demonstrates our attention mechanism can learn the richer context information even at the sentence-level and, nevertheless, document-level model can achieve a better performance with the alleviation of the tagging inconsistency problem.

3.4 The effect of additional features on performance

We also investigated the effect of three additional features (POS, chunking and dictionary embeddings mentioned in Section 2.1.3) on the performance of our model and Table 4 shows the results of

different combinations of these features on the CHEMDNER corpus. The baselines only use word and character embeddings as inputs of the model and alignment function while the additional features are introduced into BiLSTM-CRF model and Att-BiLSTM-CRF model. For the former, the concatenation of the word embedding, character embedding and additional features as input is fed into the BiLSTM layer. For the latter, three styles are designed for adding additional features. **Style-I**: the additional features are only added into the BiLSTM layer and the attention weight values are computed with the original word and character embeddings. **Style-II**: the additional features are only added into the attention layer (i.e. attention weight values are computed with the original word and character embeddings and additional features) and the original word and character embeddings as inputs are fed into the BiLSTM layer. **Style-III**: the additional features both are added into both BiLSTM and attention layers.

For the BiLSTM-CRF model, when only the chunking embedding is added, higher F-score (an improvement of 0.49% in F-score over the baseline) is achieved. The main reason is that some entity boundary errors can be revised by the chunking information. When only the POS embedding is added, the model only achieves a small improvement (an improvement of 0.15% in F-score). Among others, the dictionary embedding contributes most to the BiLSTM-CRF model (an improvement of 0.58% in F-score), which demonstrates that the information of prior chemical entities provided by JoChem dictionary can help boost the performance. When all additional features are added, the best performance (an improvement of 0.83% in F-score) is achieved.

For our Att-BiLSTM-CRF model, it achieves the best performance (91.14% in F-score over the baseline, an improvement of 0.30%) when only dictionary feature is added with the Style-II. Compared with the BiLSTM-CRF model, our model achieves the smaller improvement (0.83% versus 0.30% in F-score) when the additional features are added. It demonstrates that our model is more robust and it is less affected by the additional features. The plausible reason is that our model itself has learned sufficient effective features automatically from the word and character embeddings with our attention mechanism.

3.5 Performance comparison with other existing methods

To further demonstrate the effectiveness of our method, the performance comparison between our Att-BiLSTM-CRF method and other state-of-the-art methods are performed on the BioCreative IV CHEMDNER corpus and the BioCreative V CDR corpus as shown

Table 4. The effect of additional features on performance on the CHEMDNER corpus

	BiLSTM-CRF			Att-BiLSTM-CRF			
Feature	Precision	Recall	F-score	style	Precision	Recall	F-score
Baseline	91.31	87.73	89.48	Baseline	91.65	90.04	90.84
+chunk	90.62	89.33	89.97	Style-I	91.89	89.47	90.66
				Style-II	91.40	90.52	90.96
				Style-III	92.01	89.67	90.82
+POS	91.84	87.51	89.63	Style-I	92.43	89.26	90.82
				Style-II	91.98	89.64	90.80
				Style-III	91.95	89.31	90.61
+dic	91.40	88.76	90.06	Style-I	92.48	88.98	90.70
				Style-II	92.29	90.01	**91.14**
				Style-III	92.10	89.88	90.98
+dic	91.64	88.90	90.25	Style-I	92.38	89.29	90.81
+chunk				Style-II	92.18	89.64	90.89
				Style-III	92.14	89.65	90.88
+dic	91.49	89.17	**90.31**	Style-I	92.18	89.20	90.66
+chunk				Style-II	91.87	90.01	90.93
+POS				Style-III	92.44	89.60	91.00

Note: Chunk denotes the chunking embedding, POS denotes the POS embedding and dic denotes the dictionary embedding. The bold values denote the highest values.

Table 5. Performance comparison with other existing methods on the CHEMDNER and the CDR corpora

	BioCreative IV CHEMDNER				BioCreative V CDR			
Method	Precision	Recall	F-score	Δ	Precision	Recall	F-score	Δ
tmChem (2015)	89.09	85.75	87.39	3.75	–	–	–	–
Lu et al. (2015)	88.73	87.41	88.06	3.08	–	–	–	–
TaggerOne (2016)	–	–	–	–	94.20	88.80	91.40	1.17
RNNA-CRF	91.14	88.27	89.68	1.46	92.40	89.69	91.03	1.54
BiLSTM-CRF	91.31	87.73	89.48	1.66	92.82	88.52	90.62	1.95
BiLSTM-CRF*	91.49	89.17	90.31	0.83	92.85	90.44	91.63	0.94
Att-BiLSTM-CRF	91.65	90.04	90.84	0.30	92.88	91.07	91.96	0.61
Att-BiLSTM-CRF*	92.29	90.01	**91.14**		93.49	91.68	**92.57**	

Note: Δ denotes the F-score improvement of our model and * denotes the best version of adding additional features. The bold values denote the highest values.

in Table 5 (in our experiments, all models are trained with the same data so that the results are comparable).

Among other existing methods, tmChem (Leaman *et al.*, 2015) and the method of Lu *et al.*'s (2015) are both CRF-based methods, in which much feature engineering is employed to improve the performance. tmChem extracts rich hand-crafted features including general linguistic features, prefixes and suffixes, character features, Roman numerals and Greek letters, semantic features, chemical elements, case pattern features and contextual features. Lu et al.'s method uses word-level and character-level features, and multi-scale word clustering based on a skip-gram model is used to further improve the performance. TaggerOne (Leaman and Lu, 2016) is a semi-markov model for joint NER and normalization, which achieves an F-score of 91.40% on the CDR corpus. More recently, Pandey *et al.* (2017) presented a model similar to our Att-BiLSTM-CRF to extract knowledge of Adverse Drug Reactions (ADRs). And we rebuilt their model (RNNA-CRF) to chemical NER. To make the comparison fair, the same 50-dimensional word embedding was also used for RNNA-CRF and its hyper-parameters were tuned on the validation set likes our model. The results show that our method obtains the state-of-the-art results with much less feature engineering where word and character embeddings are used and tuned automatically during the training process. Moreover, owing to document-level attention mechanism, our Att-BiLSTM-CRF model without additional features achieves better performance than other sentence-level neural network-based models and our Att-BiLSTM-CRF model with additional features achieves the best performances so far on the BioCreative CHEMDNER and CDR corpora (91.14% and 92.57% in F-score, respectively).

4 Conclusions

In this paper, we present a novel neural network approach to document-level chemical NER by introducing a document-level attention mechanism, which allows the model to focus on tagging consistency across multiple instances of the same token in a document. In addition, we explored the effect of additional domain features for the neural network models on the chemical NER task. The experimental results show that (i) our attention mechanism that is introduced to capture the document-level correlation information between words has been proved to be effective to alleviate the tagging inconsistency problem; and (ii) our Att-BiLSTM-CRF model is more robust and it is less affected by the removal of manually designed features as discussed in Section 3.4. Owing to these two advantages, it can still achieve the state-of-the-art performance on the CHEMDNER and CDR corpora with only word and character embeddings (90.84 and 91.96% in F-score, respectively).

Our Att-BiLSTM-CRF approach exhibits promising results for chemical NER in the biomedical literature. In addition, the approach can be easily adapted to other domain. However, it may have shortcoming on the identification of some other entity types where consistency is not desirable. For example, articles mentioning genetic diseases sometimes use the same abbreviation to refer to both the gene and the disease. In the future work, we will improve our model to solve the problem. Moreover, the NER approach need to capture longer distance dependency information when applied to the full text articles than the abstracts, and our approach will be applied to full text articles in our future works.

Funding

This work was supported by the grants from the National Key Research and Development Program of China (No. 2016YFC0901902, funding body: Ministry of Science and Technology of China), Natural Science Foundation of China (No. 61272373, 61572102 and 61572098, funding body: National Natural Science Foundation of China), and Trans-Century Training Program Foundation for the Talents by the Ministry of Education of China (NCET-13-0084, funding body: Ministry of Education of China).

Conflict of Interest: none declared.

References

Bahdanau,D. *et al.* (2015) Neural machine translation by jointly learning to align and translate. In: *International Conference on Learning Representations*.

Bergstra,J. and Bengio,Y. (2012) Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, **13**, 281–305.

Bharadwaj,A. *et al.* (2016) Phonologically aware neural model for named entity recognition in low resource transfer settings. *EMNLP*, **2016**, 1462–1472.

Bottou,L. (1991) Stochastic gradient learning in neural networks. In: *Proceedings of Neuro-Nimes*, p. 91.

Chalapathy,R. *et al.* (2016) An investigation of recurrent neural architectures for drug name recognition. In: *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pp. 1–5.

Collobert,R. *et al.* (2011) Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, **12**, 2493–2537.

Davis,A.P. *et al.* (2009) Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical–gene–disease networks. *Nucleic Acids Res.*, **37**, D786–D792.

Degtyarenko,K. *et al.* (2008) ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.*, **36**, D344–D350.

Eltyeb,S. and Salim,N. (2014) Chemical named entities recognition: a review on approaches and applications. *J. Cheminf.*, **6**, 17.

Finkel,J.R. *et al.* (2005) Incorporating non-local information into information extraction systems by gibbs sampling. In: *Proceedings of the 43rd annual meeting on association for computational linguistics.*, pp. 363–370.

Hettne,K.M. *et al.* (2009) A dictionary to identify small molecules and drugs in free text. *Bioinformatics*, **25**, 2983–2991.

Hochreiter,S. and Schmidhuber,J. (1997) Long short-term memory. *Neural Comput.*, **9**, 1735–1780.

Huang,Z. *et al.* (2015) Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv: 1508.01991.

Krallinger,M. *et al.* (2015) CHEMDNER: The drugs and chemical names extraction challenge. *J. Cheminform.*, **7**, S1.

Lai,S. *et al.* (2016) How to generate a good word embedding. *IEEE Intell. Syst.*, **31**, 5–14.

Lai,S. *et al.* (2015) Recurrent convolutional neural networks for text classification. *AAAI*, 2267–2273.

Lample,G. *et al.* (2016) Neural architectures for named entity recognition. In: *Proceedings of NAACL-HLT.* pp. 260–270.

Leaman,R. and Lu,Z. (2016) TaggerOne: joint named entity recognition and normalization with semi-Markov Models. *Bioinformatics*, **32**, 2839–2846.

Leaman,R. *et al.* (2015) tmChem: a high performance approach for chemical named entity recognition and normalization. *Journal of Cheminformatics*, **7**,.

Li,L. *et al.* (2015) Biomedical named entity recognition based on extended recurrent neural networks. In: *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, pp. 649-652.

Li,J. *et al.* (2016) BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, **2016**, 1–10.

Liu,S. *et al.* (2015) Drug name recognition: approaches and resources. *Information*, **6**, 790–810.

Lu,Y. *et al.* (2015) CHEMDNER system with mixed conditional random fields and multi-scale word clustering. *J. Cheminf.*, **7**, S4.

Ma,X. and Hovy,E. (2016) End-to-end sequence labeling via bi-directional lstm-cnns-crf. In: Proceedings of ACL, pp. 1064–1074.

Mikolov,T. *et al*. (2013) Distributed representations of words and phrases and their compositionality. *Adv. Neural Inf. Process. Syst*., **2**, 3111–3119.

Mnih,V. *et al*. (2014) Recurrent models of visual attention. *Adv. Neural Inf. Process. Syst*., **2**, 2204–2212.

Pandey,C. *et al*. (2017) Improving RNN with attention and embedding for adverse drug reactions. In: *Proceedings of the 2017 International Conference on Digital Health*. ACM, pp. 67–71.

Pennington,J. *et al*. (2014) Glove: Global vectors for word representation. In: *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, vol. 12, pp. 1532–1543.

Prechelt,L. (1998) Automatic early stopping using cross validation: quantifying the criteria. *Neural Netw*., **11**, 761–767.

Ratinov,L. and Roth,D. (2009) Design challenges and misconceptions in named entity recognition. In: *Proceedings of CoNLL*. Association for Computational Linguistics, pp. 147–155.

Rei,M. *et al*. (2016) Attending to characters in neural sequence labeling models. arXiv preprint arXiv: 1611.04361.

Rocktäschel,T. *et al*. (2012) ChemSpot: a hybrid system for chemical named entity recognition. *Bioinformatics*, **28**, 1633–1640.

Sahu,S.K. and Anand,A. (2016) Recurrent neural network models for disease name recognition using domain invariant features. In: *Proceedings of ACL*, pp.2216–2225

Segura Bedmar,I. *et al*. (2013) *Semeval-2013 Task 9: Extraction of Drug-Drug Interactions from Biomedical Texts (Ddiextraction 2013)*. In: Proceedings of the 7th International Workshop on Semantic Evaluation, Atlanta, Association for Computational Linguistics, pp. 341–350.

Usié,A. *et al*. (2014) CheNER: chemical named entity recognizer. *Bioinformatics*, **30**, 1039–1040.

Viterbi,A. (1967) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inf. Theory*, **13**, 260–269.