IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 10, 2019, accepted May 26, 2019, date of publication June 4, 2019, date of current version June 18, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2920734

Named Entity Recognition From Biomedical Texts Using a Fusion Attention-Based BiLSTM-CRF

HAO WEI, MINGYUAN GAO, AI ZHOU, FEI CHEN, WEN QU, CHUNLI WANG, AND MINGYU LU
Information Science and Technology College, Dalian Maritime University, Dalian 116026, China

Corresponding author: Mingyu Lu (lumingyu@dlmu.edu.cn)

This work was supported by the Natural Science Foundation of China under Grant 61272369.

ABSTRACT Biomedical named entity recognition (BNER) is the basis of biomedical text mining and one of the core sub-tasks of information extraction. Previous BNER models based on conventional machine learning rely on time-consuming feature engineering. Though most neural network methods improve the problems with automatic learning, they cannot pay attention to the significant areas when capturing features. In this paper, we propose an attention-based BiLSTM-CRF model. First, this model adopts a bidirectional long short-term memory network (BiLSTM) to obtain more complete context information. At the same time, the attention mechanism is proposed to improve the vector representations in BiLSTM. We design different attention weight redistribution methods and fuse them. It effectively prevents the significant information loss when extracting features. Finally, combining BiLSTM with conditional random field (CRF) layer effectively solves the problems of the inability to handle the strong dependence of tags in the sequence. With the simple architecture, our model achieves a reasonable performance on the JNLPBA corpus. It obtains an F1-score of 73.50. Our model can enhance the ability of the neural network to extract significant information and does not rely on any feature engineering, with only general pre-training word vectors. It makes our model have high portability and extendibility.

INDEX TERMS Biomedical text, named entity recognition, attention mechanism, long short-term memory, conditional random field.

I. INTRODUCTION

Named entity recognition (NER) is the basic task of information extraction and data mining. Its main research content is to identify entities (persons, organizations and so on) in massive texts. In recent years, with the exponential increase of biomedical texts, how to extract and mine structured valuable information from massive data has received more and more attention. The first and essential step in obtaining these information is identifying the entities in biomedical texts. Therefore, biomedical named entity recognition (BNER) has become one of the most important research points.

The biomedical entities include genes, proteins, diseases, drugs, chemicals and so on. But in biomedical texts, the entities' naming rules are not clear and a large number of phrases and abbreviations have difficulty to be identified. The above problems also make it more difficult to identify biomedical entities than in general field. Previous works for BNER usually employed conventional machine learning methods, such as conditional random field (CRF), hidden markov model (HMM), support vector machine (SVM), maximum entropy markov model (MEMM), etc. Li *et al.* [1] proposed a feature coupling generalization framework to generate lower dimensional features based on co-occurrence information and term frequency in unlabeled data. Then they combined them with CRF to obtain an F1-score of 89.05 on the BC2GM corpus. Torii *et al.* [2] developed a BNER system called BioTagger-GM, which integrated four machine learning models including MEMM, CRF and so on, and they confirmed the applicability in multiple corpora. The Skip-Chain CRF model was constructed by Liao and Wu [3] for BNER, which could fully obtain the dependence of the context information with a long sentence. It obtained F1-scores of 73.20 on the JNLPBA corpus. Tang *et al.* [4] evaluated three different types of word representation features (clustering-based representation, distributional representation and word embeddings). Then they sent these features into CRF. It obtained F1-scores of 71.39 and 80.96 on JNLPBA

The associate editor coordinating the review of this manuscript and approving it for publication was Yin Zhang.

and BC2GM corpora respectively. CRF even became the best choice for BNER because it had shown good performance on different kinds of sequence labeling tasks [5]. However, these studies must rely on the feature engineering. They had to design much complicate features manually which not only required linguistic and domain insight but also was time-consuming.

Driven by artificial intelligence and cognitive computing, deep learning has found an increasingly wide utilization in all fields [6]–[13]. Therefore, more and more NER studies adopt the neural networks to automatically learn the feature representations from text sequences. Because it can avoid the costly feature engineering. The commonly used neural network models include convolutional neural network (CNN), recurrent neural network (RNN), long short-term memory network (LSTM), etc. In general field, Huang *et al.* [14] proposed the combination of bidirectional LSTM (BiLSTM) and CRF model, which could use both backward and forward input features to achieve better results. Ma and Hovy [15] used a combination of CNNs and LSTM-CRF to identify entities. Chiu and Nichols [16] adopted BiLSTM-CNN model to obtain reasonable results on CoNLL-2003 corpus. Lample *et al.* [17] used BiLSTM-CRF model with character-level representation. These models have reached or surpassed results of the conventional methods based on feature engineering. In BNER, Yao *et al.* [18] first used neural networks to generate word embeddings on unlabeled biological texts, and then established a multi-layer neural network. Their system achieved F1-score of 71.01 on JNLPBA corpus and closed to the state-of-the-art performance. Li *et al.* [19] combined the twin word embeddings with sentence vector and adopted the BiLSTM to identify biomedical entities. Their model could achieve an F1-score of 88.61 on the BC2GM corpus. Zeng *et al.* [20] constructed the BiLSTM-CRF model to identify drug entities. They adopted CNN to obtain the character-level feature representations and combined them with word-level representations as the inputs. The BiLSTM-CRF model achieved F1-scores of 92.04 and 79.26 on DDI2011 and DDI2013 corpora respectively. Zhao *et al.* [21] proposed a multi-label CNN (MCNN) model to identify disease entities and chemical entities. They treated NER as a classification task and adopted the multi-label mechanism to obtain adjacent output tags. The MCNN obtained F1-scores of 85.17 and 87.83 on NCBI-Disease and CDR corpora respectively.

However, there are a large number of function words in the biomedical literature that can lead to information redundancy. They will obstruct the acquisition of significant information when the neural network models are trained to capture features. It may result in information loss if the BNER models cannot focus on the significant areas. Therefore, how to make the neural network models pay more attention to significant areas becomes an important factor to improve performance. Regarding the issue above, we adopt the attention mechanism to improve them. Bahdanau *et al.* [22] first proposed the attention mechanism in the field of machine translation. By adding the attention structure to the decoder model, this model can focus on the significant parts of the original sentences when decoding and it can reduce the loss of important information. Now the attention mechanism has been widely used in NLP. Many researchers have applied their continuous improvement to machine translation [23], image description [24], syntactic analysis [25], and reading comprehension [26], etc. In BNER, Pandey *et al.* [27] proposed an Encoder-Decoder model based on bidirectional RNN with attention mechanism and used to extract adverse drug reactions from a large number of electronic health records. Luo *et al.*. [28] adopted an attention-based BiLSTM-CRF model for document-level BNER. They used the attention mechanisms between different sentences to optimize the tagging inconsistency problem. It achieved the best performances on CHEMDNER (F1-score:91.14) and CDR (F1-score:92.57) corpora.

This paper aims to improve the accuracy of identifying biomedical entities by enhancing the model's ability to capture significant information. We propose an attention-based BiLSTM-CRF model. First, it adopts a bidirectional LSTM network, which not only solves the problem of distance dependence in longer sentences, but also considers the bidirectional context information. Then, the attention mechanism is used to redistribute weights of the vector representations in hidden state. It makes our model pay more attention to significant areas of the text sequence when capturing features. In order to better calculate the attention weight distribution, we propose different methods of attention weight redistribution and effectively fuse them. Finally, the CRF layer we used can effectively improve the problem that the BiLSTM cannot handle the strong dependence of the tags in sequence labeling. Our model can effectively improve the accuracy of identifying biomedical entities without relying on any feature engineering.

The remainder of the paper is organized as follows. In Section 2, we review our method. Section 3 reports the experimental settings. Section 4 analyzes and discusses the experimental results in detail. Section 5 draws conclusions.

II. METHODS

The attention-based BiLSTM-CRF framework is shown in Figure 1. First, the original data is searched by the pre-trained vectors dictionary and converted into word embeddings. Then the BiLSTM layer is used to capture features in the word embedding sequence. Next, we use the fusion-based attention mechanism to redistribute the weight for the output of BiLSTM. Finally, we adopt the CRF layer to parse the tags corresponding to the text sequence. Our model's framework includes four parts: Input (Feature and Embedding), BiLSTM, Attention and CRF.

A. INPUT LAYER

In this paper, we represent the text sequence with only word vectors. So we input each word as the discrete feature to the embedding layer. Word embeddings are the expression that maps words to real low-dimensional vectors. It is used to

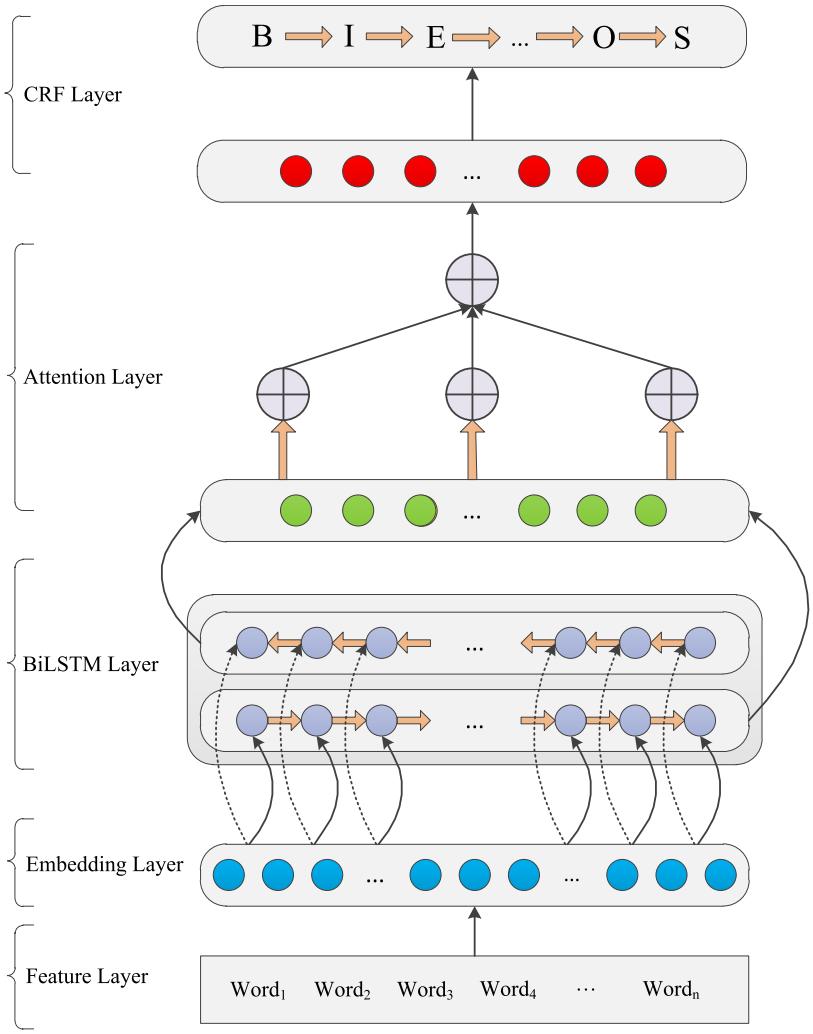


FIGURE 1. The attention-based BiLSTM-CRF model overall framework. It includes five parts: Input (Feature and Embedding), BiLSTM, Attention and CRF.

express the vocabulary instead of the conventional one-hot method, because it can solve the one-hot method of dimensional disaster and the inability to express vocabulary relationships. Recently, with the development of deep learning in the NLP, word embeddings have also been more widely used. The use of word embeddings as direct input or extra feature expressions has significantly improved the performance of NLP related models. At present, the commonly used word embeddings training tools are Word2vec [29], Glove [30], etc. We use Stanford's publicly available Glove 100-dimensional embeddings (https://nlp.stanford.edu/projects/glove/) which are trained on 6 billion words from Wikipedia and web texts as the pre-trained embeddings for our model.

B. BILSTM LAYER

LSTM [31], [32] is an improving architecture based on RNN. It introduces a memory unit and gate mechanism to realize the use of longer distance information in sentences and solves the problem of gradient disappearance or gradient explosion in RNN. The designed door structure can selectively save

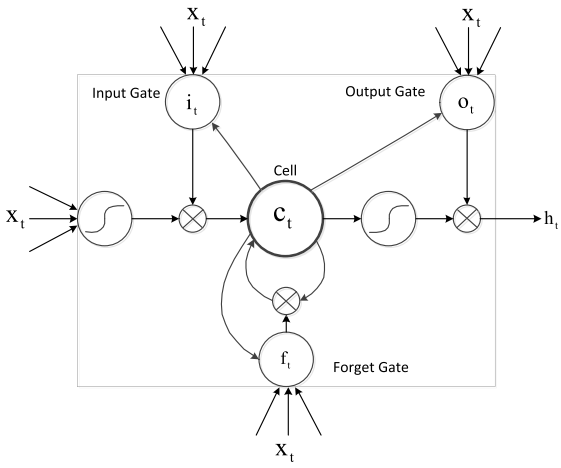


FIGURE 2. The memory unit of LSTM.

context information, so it is more suitable for sequence labeling problems such as NER. The memory unit of LSTM is shown in Figure 2. The mathematical expressions of the

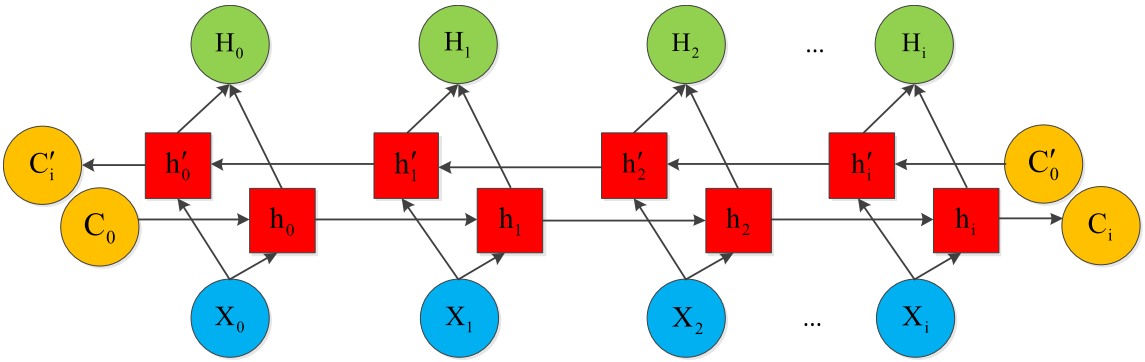


FIGURE 3. The BiLSTM model framework.

LSTM model are as follows:

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (1)$$
$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad (2)$$
$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (3)$$
$$\widetilde{c_t} = tanh(W_c[h_{t-1}, x_t] + b_c) \quad (4)$$
$$c_t = i_t * \widetilde{c_t} + f_t * c_{t-1} \quad (5)$$
$$h_t = o_t * tanh(c_t) \quad (6)$$

where σ represents the *sigmod* activation function. *tanh* represents the hyperbolic tangent function. x_t represents the unit input. i_t, f_t, o_t represent the input gate, the forget gate, and the output gate at time t. W and b represent the weights and bias of the input gate, the forget gate, and the output gate respectively. $\widetilde{c_t}$ denotes the current state of the input. c_t denotes the update state at time t. h_t denotes the output at time t.

However, LSTM only considers the forward information of the sentence and cannot obtain the backward information. In the task of BNER, the backward information also contains important features. In response to the above problems, we adopt the bidirectional LSTM (BiLSTM) proposed by Graves and Schmidhuber [33] in BNER, which is shown in Figure 3. The BiLSTM model captures two different feature representations of the forward and backward for each sentence. Then it merges them to obtain the complete hidden layer feature representation. BiLSTM can effectively obtain the bidirectional context information and excavate more hidden features.

C. ATTENTION LAYER

This section includes two parts: attention mechanism and attention fusion. It mainly describes the attention weight calculation and fusion method proposed in this paper.

1) ATTENTION MECHANISM

Attention is the mechanism by which the brain's attention is focused on significant areas when simulating human observation of things. The core idea is to imitate the brain when humans are observing things. It focuses on a critical area of the observed things, while ignoring other non-critical areas. The attention mechanism in the neural network model can be understood as a weight redistribution mechanism. Combining the attention mechanism with our model can effectively highlight the role of keywords in text data. When the attention mechanism is not added, the model only obtains the global context information and cannot focus on the significant information acquisition. After the attention mechanism is adopted, the significant areas can be highlighted by increasing weights. The acquisition of significant information can effectively improve the performance of the BNER models. In this paper, the relevant attention mechanism calculation formulas are as follow:

$$H_i = [\overrightarrow{h_i} \oplus \overleftarrow{h_i}] \quad (7)$$
$$\alpha_i = \frac{exp(e_i)}{\sum_{k=1}^{n} exp(e_k)} \quad (8)$$
$$e_i = tanh(W^T H_i + b) \quad (9)$$

where the weight calculation formula of the redistribution of each word is obtained by (8). Where α_i is the normalized processing result of e_i, and e_i represents the energy function which is used to quantify the importance of every word in the sentence and is obtained by formula (9). Where W represents the weight parameter of the hidden state H_i. In this paper, we propose different attention weight redistribution methods as follows.

In formula(10), this is a attention weight redistribution method based on scoring mechanism. We regard the attention weight of each word's hidden state as the importance score, and then concatenate the score with the hidden state before outputting. The symbol $\oplus$ denotes concatenate operation.

$$H_i' = H_i \oplus \alpha_i \quad (10)$$

In formula(11), this is a attention weight redistribution method based on merge operation. Similar to formula (10), we still aim to guide hidden state by the attention weight. The final hidden state H_i' is obtained by merging hidden state H_i and the attention weight α.

$$H_i' = H_i + \alpha_i \quad (11)$$

In formula (12), this is a attention weight redistribution method based on dynamic scaling. We multiply the hidden state of each word by the attention weight score. It represents the importance of each word in the task. The attention weight calculation method changes the probability matrix of the BiLSTM output, which may improve the sequence labeling results of the CRF layer. Besides, the symbol $*$ denotes element-wise multiplication.

$$H_i' = H_i * \alpha_i \tag{12}$$

2) ATTENTION FUSION

The final state H_i' denotes the output of the BiLSTM, where the size of the H_i' in different attention calculation methods are $n * (d_u + d_\alpha)$, $n * d_u$ and $n * d_u$ respectively. n represents the number of words, and d_u represents the number of the hidden units. Corresponding to the different attention output, we can get different probability matrices P by linear mapping. In order to maintain the stability of attention weight calculation and avoid multiple attention matrix to cause information redundancy, we calculate the similarity of two probability matrices separately in the stage of attention fusion. Then we merge the two matrices with the most similarity, and finally input them into the CRF layer. The relevant attention fusion formulas are as follow:

$$dist(P_i, P_j) = \sqrt{\sum_{i=1,j=1,i\neq j}^{n} (P_i - P_j)^2} \tag{13}$$

$$(P_i, P_j)' = (P_i, P_j)_{min(dist(P_i,P_j))} \tag{14}$$

$$P' = avg(P_i + P_j) \tag{15}$$

where *dist* represents the similarity of any two probability matrices, and P represents the probability matrix obtained by different attention weight redistribution methods. We perform the addition operation on the two most similar matrices to obtain the output of BiLSTM and serve them as the input to the CRF layer.

D. CRF LAYER

After the BiLSTM outputs the feature sequence, the sequence labeling result L can be obtained by the classification decision functions. However, in the task of BNER, the output sequence tags have strong restrictions and dependencies. For example, in the entity tagging scheme, ''B-protein'' cannot appear after ''I-protein''. The classification decision functions is not sufficient to label the biomedical sequences effectively.

CRF [20] is an undirected graph model. It can obtain the global optimal labeling sequence by considering the relationship between adjacent tags. In this paper, the CRF layer is added after the output layer of BiLSTM, so that not only can the context information be combined, but also the dependencies between the output tags can be effectively considered. For sentences $S = \{s_1, s_2, \cdots, s_n\}$ is fed into the training network, the probabilistic matrix P is the output after the fusion-based attention operation in BiLSTM, where the size of P is $n * m$. n represents the number of words, and m represents the number of class labels. p_{ij} is the prediction that the *i-th* word is the *j-th* label probability in a sentence. For a prediction sequence $L = \{l_1, l_2, \cdots, l_n\}$, its probability can be expressed as:

$$P(S, L) = \sum_{i=0}^{n} A_{l_i, l_{i+1}} + \sum_{i=1}^{n} P_{i, l_i} \tag{16}$$

A is a transfer matrix of size $m+2$. For example, the transition probability from the labels i to j can be expressed as A_{ij}, where l_1 and l_n represent labels for the start and end of the predicted sentence. Therefore, the probability of generating the all possible tags sequence L under the *softmax* is:

$$P(L \mid S) = \frac{exp^{P(S,L)}}{\sum_{\widetilde{L} \in L_S} exp^{P(S,\widetilde{L})}} \tag{17}$$

$\widetilde{L}$ represents the truth of tags sequence.

The likelihood function of training the tags sequence is:

$$log(P(L \mid S)) = P(S, L) - log\left(\sum_{\widetilde{L} \in L_S} exp^{P(S,\widetilde{L})}\right) \tag{18}$$

L_S represents all the valid tags sequence that can be obtained by the likelihood function formula. In the prediction stage, the formula of the output tags sequence with the largest overall probability is:

$$L^* = argmax_{\widetilde{L} \in L_S} P(S, \widetilde{L}) \tag{19}$$

III. EXPERIMENTAL SETTINGS

This section mainly describes the experimental settings of the model training. It includes optimizer, regularization, tagging scheme, hyperparameter, corpus and evaluation measures.

A. OPTIMIZER

We use the Adam [34] (Adaptive Moment Estimation) algorithm in training. Adam is different from the conventional gradient descent methods. The conventional gradient descent methods maintain a fixed learning rate to update all parameters, and the learning rate do not change during training. Adam obtains independent adaptive learning rates for different parameters by calculating the first moment estimation and second moment estimation of the gradient. It makes the parameters update more stable.

B. REGULARIZATION

Dropout was proposed by Srivastava *et al.* [35] in 2012 to prevent overfitting of neural network models during training. The main idea of dropout is to randomly disable some neurons with a certain probability, which can make the model have better generalization because it does not rely on some local features.

C. IOBES TAGGING SCHEME

The most common format labeling method in NER is BIO method. B (Beginning) indicates that the word is a named entity start, I (Inside) indicates that the word is in a named

entity, and O (Outside) indicates that the word is outside the named entity. Because the naming rules of the biomedical entities are unclear and usually contain multiple words in a entity, we use the more detailed format tagging method IOBES to clearly code the vocabulary information in each entity. Adding the E (End) and S (Single) tagging methods respectively denotes the end of the entity and a single entity. Compared to BIO, IOBES can represent the more fine-grained sequence structure, and previous studies have shown that IOBES performs better than BIO [36]–[38]. By the above method, an example is as follows:

- **SENTENCE: IL-2 gene** expression and **NF-kappa B** activation through **CD28** requires reactive oxygen production by **5-lipoxygenase**.
- **BIO: B-DNA I-DNA** O O **B-protein I-protein** O O **B-protein** O O O O O **B-protein** O
- **IOBES: B-DNA E-DNA** O O **B-protein E-protein** O O **S-protein** O O O O O **S-protein** O

D. HYPER-PARAMETERS

Table 1 shows the settings of various hyper-parameter values. The dimension based on the word embeddings is set to 100. The Adam is used as the optimizer and the dropout rate is set by 0.5. The learning rate is 0.001 and the gradient clipping value is 5. We set the batch size value is 20.

TABLE 1. Experimental hyper-parameters.

Hyper-parameter	Value
word dim	100
LSTM dim	100
Dropout rate	0.5
initial learning rate	0.001
gradient clipping	5
optimizer	Adam
batch size	20
tag schema	IOBES

E. CORPUS

We use the JNLPBA corpus as our experimental data set. This corpus contains five entity types include DNA, RNA, Cell Type, Cell Line and Protein. It consists 2000 MEDLINE abstracts are used as training sets and 404 MEDLINE abstracts are used as test sets. The details of the corpus are shown in Table 2 and Table 3.

TABLE 2. Counts of basic statistics in JNLPBA corpus.

JNLPBA	Basic statistics		
	Abstracts	Sentences	Words
Training	2000	20546	472006
Test	404	4260	96780

F. EVALUATION MEASURES

We adopt three performance evaluation measures to show our model's performance, include Precision(P), Recall(R) and F-score(F1). In this paper, we use the entity-level evaluation measures with exact matches. The calculation formulas are:

$$P = \frac{TP}{TP + FP} \tag{20}$$

$$R = \frac{TP}{TP + FN} \tag{21}$$

$$F1 = \frac{2 * P * R}{P + R} \tag{22}$$

TP (True Positives) indicates the correct number of samples in the positive examples, *FP* (False Positives) indicates the number of incorrect samples in the negative examples, and *FN* (False Negatives) indicates the number of incorrect samples in the positive examples.

IV. RESULTS AND DISCUSSIONS

All experimental results are shown in the following tables. Table 4 shows the performance comparison between our model and the baseline models. Table 5 shows the performance comparison of the different word embedding dimensions including 50, 100, 200 and 300. Table 6 shows the performance comparison of the different LSTM units dimensions including 50, 100, 150 and 200. Table 7 shows the performance comparison of the different optimizer including SGD, Adam and AdaGrad. Table 8 shows the performance of using Dropout. Table 9 shows the performance comparison between our model and some previous models. Finally, we analyze the error cases in the test set.

A. COMPARISON WITH BASELINE MODELS

In the first set of experiment, we compare the performance of our model with baseline models. The experimental results are arranged according to the F1-score as shown in Table 4. We adopt the BiLSTM-CRF [17] which is currently the most widely used in sequence labeling tasks as the baseline model. As can be seen from the table below, our single attention methods have different degrees of performance improvement comparing to the baseline model. However, our fusion-based attention method maximizes the performance of the model, with an F1-score of 73.50. It proves the effectiveness of our method. In summary, we compare our model with the single-attention BiLSTM-CRF models and the baseline model (BiLSTM-CRF). Our model have obtained obvious performance improvements.

B. COMPARISON OF EMBEDDING DIMENSIONS

The dimensions of word embedding also affect the performance of the model. We set four different dimensions of pre-training word embedding provided by Glove, including 50, 100, 200 and 300. From the results in table 5, When the word embedding dimension is equal to 100, the highest F1-score is 73.50. So we set the dimension of the pre-trained word embedding to 100.

C. COMPARISON OF LSTM UNITS DIMENSIONS

The dimensions of the hidden units in the LSTM can change the training parameters of the model, thereby affecting the

TABLE 3. Counts of different types of entities in JNLPBA corpus.

JNLPBA	Counts of different types of entities					
	DNA	RNA	Cell Type	Cell Line	Protein	Total
Training	9534	951	6718	3830	30269	51301
Test	1056	118	1921	500	5067	8662

TABLE 4. Performance comparison of baseline models on the JNLPBA.

Methods	DNA	RNA	Cell Type	Cell Line	Protein	Overall	Δ
Ours	71.67	73.47	75.38	60.33	74.41	73.50	-
ATT1-BiLSTM-CRF	70.68	76.54	74.60	58.49	74.05	72.82	**-0.68**
ATT2-BiLSTM-CRF	71.03	74.59	75.22	58.54	73.96	72.91	**-0.59**
ATT3-BiLSTM-CRF	70.50	73.11	75.75	59.45	74.13	73.14	**-0.36**
BiLSTM-CRF	69.81	69.46	75.17	61.15	73.56	72.62	**-0.88**

*ATT1, ATT2, and ATT3 respectively correspond to formulas (10), (11) and (12).

TABLE 5. Performance comparison of the different word embedding dimensions.

	Dimensions	Precision	Recall	F1-Score
Embedding	50	69.02	76.00	72.34
	100	**71.57**	**75.55**	**73.50**
	200	70.82	75.26	72.97
	300	70.36	75.87	73.01

TABLE 6. Performance comparison of the different LSTM units dimensions.

	Dimensions	Precision	Recall	F1-Score
LSTM	50	70.83	75.12	72.91
	100	**71.57**	**75.55**	**73.50**
	150	69.77	77.36	73.37
	200	70.47	74.47	72.42

overall performance and computational complexity. In order to get the best hyperparameters, we set up hidden units of different dimensions, including 50, 100, 150 and 200. From table 6, when the hidden unit dimension is 100, the model obtains the highest F1-Score of 73.50. That the hidden units is too few will result in insufficient ability of capturing feature. With the increase of the dimension, the increase of the training parameters leads to an increase of computational complexity. Therefore, we set the hidden unit dimension of LSTM to 100.

D. COMPARISON OF OPTIMIZATION METHODS

This paper adopts the Adam optimization method during model training. In order to prove the effectiveness of adam, we set up a comparison experiment including sgd and adagrad. The experimental results are shown in Table 7. SGD is widely used in the optimization of neural networks and the main principle is to randomly draw a batch of fixed volumes from training samples. For example, the gradients and errors are calculated iteratively and the parameters are updated. The main principle of AdaGrad is to randomly sample a batch of fixed-capacity samples from a training sample, calculate the gradient and error, then update the learning rate and gradient parameters. The experimental results show that Adam converges faster when using the same parameters. SGD and AdaGrad do not achieve satisfactory results compared with Adam in our methods.

TABLE 7. Performance comparison of the different optimization methods.

Method	Precision	Recall	F1-Score
SGD	64.03	68.04	65.98
AdaGrad	60.17	65.65	62.79
Adam	**71.57**	**75.55**	**73.50**

TABLE 8. Performance comparison of using dropout.

Dropout	Precision	Recall	F1-Score
No	69.26	73.42	71.28
Yes	71.57	75.55	73.50
Δ	**2.31**	**2.13**	**2.22**

E. COMPARISON OF USING DROPOUT

In order to verify the effectiveness of the dropout, another set of comparative experiment is set up in this paper. The experimental results are shown in table 8.

This experiment is set to compare the performance differences before and after using dropout. The experimental results show that our model overall performance improves significantly after using dropout. It indicates that the regularization method we use is effective.

F. COMPARISON WITH EXISTING MODELS

Finally, we compare the performance of our model with some previous models. The experimental results are arranged according to the F1-score as shown in Table 9.

In these studies, POSBioTM-NER [39] incorporated various morphological information, part-of-speech and noun phrase as the input features. The SVM and CRF were merged to identify biomedical entities. Finkel *et al.* [40] made full use of local features and external resources to recognize the biomedical named entities. ABNER [41] and Gimli [44] were all the open source BNER systems based on CRF. Tsuruoka *et al.* [42] developed a biomedical sequence labeling system called GENIA Tagger based on the maximum entropy model and the tagging algorithm. It performed on

TABLE 9. Performance comparison with previous models on the JNLPBA.

Method	F1-score on each JNLPBA type					Overall performance		
	DNA	RNA	Cell Type	Cell Line	Protein	Precision	Recall	F1-Score
Song et al.[39]	60.08	64.07	64.48	57.33	69.07	67.82	64.80	66.28
Finkel et al.[40]	67.86	68.83	69.06	52.40	72.67	71.62	68.56	70.06
Settles[41]	65.10	61.60	72.00	56.00	72.60	69.10	72.00	70.50
Yao et al.[18]	65.90	60.76	73.50	55.25	71.29	64.86	76.10	71.01
Tsuruoka et al.[42]	66.20	64.29	74.31	57.81	72.79	67.45	75.78	71.37
Tang et al.[4]	-	-	-	-	-	70.78	72.00	71.39
Chang et al.[43]	-	-	-	-	-	-	-	71.85
Campos et al.[44]	69.27	67.24	70.49	58.64	74.68	72.85	71.62	72.23
Zhou et al.[45]	69.83	64.10	75.13	59.23	73.77	69.42	75.99	72.55
Zhu et al.[46]	68.64	66.95	73.58	59.44	74.37	-	-	72.57
Li et al.[19]	-	-	-	-	-	**74.77**	70.85	72.76
Tsai-H. et al.[47]	70.00	72.65	72.77	57.39	75.12	72.01	73.98	72.98
Liao et al.[3]	69.80	68.50	73.70	**66.60**	**76.50**	72.80	73.60	73.20
Wang et al.[48]	-	-	-	-	-	70.91	76.34	73.52
Lyu et al.[49]	-	-	-	-	-	71.24	76.53	73.79
Lee et al.[50]	-	-	-	-	-	72.68	**83.21**	**77.59**
Ours	**71.67**	**73.47**	**75.38**	60.33	74.41	71.57	75.55	73.50

the JNLPBA corpus with an F1-score of 71.37. In the study by Chang *et al.* [43], the biomedical word embeddings were induced into the CRF model as extra feature inputs. After applying them, the performance of the model had been significantly improved. PowerBioNE is a BNER system developed by Zhou *et al.* [45] based on HMM, which adopted various evidential features and resolved the data sparseness problem by K-Nearest Neighbor (KNN) algorithm. NERBio [47] is the best system which is implemented as a rule-based method on JNLPBA corpus and it obtains an F1-score of 72.98. The skip-chain CRF proposed by Liao and Wu [3] is the best system which is implemented as a conventional machine learning method on the same corpus and it obtains an F1-score of 73.20. However, the F1-score of our model is 0.52 better than NERBio and 0.30 better than skip-chain CRF model. Our model without any feature engineering achieves better performances than these rule-based and conventional machine learning methods.

Compared with recent deep learning methods, Zhu *et al.* [46] first adopted the CNN structures to BNER, and it achieved the F1-score of 72.57 on the JNLPBA corpus. The performance of our model is similar to that of Wang *et al.* [48] and they proposed a multi-task method with higher model complexity. Lyu *et al.* [49] proposed a model of BiLSTM-RNN. The F1-score is 0.29 higher than our model, but the precision of the model is 0.33 lower than ours. Moreover, the method requires pre-training the biomedical texts to obtain the word embeddings. We have achieved similar performance on the pre-trained word embeddings provided in the general field. Lee *et al.* [50] proposed the BioBERT language model to train biomedical corpora on eight V100 GPUs for 20 days. It greatly enhances the best performance of biomedical NLP tasks. But the training process is time-consuming and complex, what's more, it requires high experimental environment and operating platform. If BioBERT adopts the general pre-training vectors like our model, it's Precision, Recall, and F1-Score are 69.57, 81.20, and 74.94, respectively. Although the F1-Score is 1.44 higher than our model, the Precision is 2 lower than ours. In summary, our method achieves reasonable performance under the premise of using public general pre-training word vectors and simple architecture. The model has high portability and extendibility without relying on domain knowledge, high-performance devices and large-scale computing.

G. ERROR ANALYSIS

In this section we analyze the error cases on the JNLPBA corpus test set and classify them into the following categories.

1) The boundary is not clear. For example, the conjunctions ''and'' and ''or'' are included in ''T- and B-lymphocyte'' and ''B or HeLa cells'', which is often difficult to determine whether the conjunction is part of an entity.

2) The corpus annotation inconsistency. For example, ''beta-2-microglobulin'' is labeled as ''O'' in ''Tumor and serum beta-2-microglobulin expression in women with breast cancer.'', but in ''To investigate whether the tumor expression of beta-2-microglobulin (beta 2-M) could serve as a marker of tumor biologic behavior.'' is labeled as ''S-protein''.

3) Entity nesting. For example, both ''tumor beta 2-M'' and ''beta 2-M'' can be considered as ''protein'', which also increases the difficulty of model learning features.

4) Entity contains multiple words. For example, ''IL-2 receptor alpha (IL-2R alpha) gene'' is a ''DNA'' that contains six words and brackets. In this case, it is difficult to judge the boundary of the entity.

The above analysis shows that the main reason for the error in entity labeling is the complexity and annotation inconsistency of the corpus itself. For this problem we should adopt dynamic word vectors or continue to enhance the ability of our model to capture features.

V. CONCLUSION

In this paper, we propose a fusion attention-based BiLSTM-CRF model to replace the early methods (dictionary-based

and rule-based) and the conventional machine learning models in BNER. First, this paper adopts BiLSTM to obtain the bidirectional context information. Then, we propose different attention weight redistribution methods and fuse them. It improves the ability of BiLSTM to pay attention to more significant areas when capturing features. Finally, considering the strong dependencies between tags, the probability matrix from BiLSTM is input to the CRF layer to parse sequence tags. Our model has better flexibility and it does not rely on any complicate feature engineering. Besides, the attention mechanism can prevent the effective information loss to capture significant features more accurately. Our model produces reasonable performance in the JNLPBA corpus, with only general pre-training word vectors and simple architecture. In future work, we plan to adopt more abundant feature representations such as knowledge-based features, character-level features, part-of-speech features and so on. And we also plan to improve the attention mechanism calculation methods.

REFERENCES

[1] Y. Li, H. Lin, and Z. Yang, "Incorporating rich background knowledge for gene named entity classification and recognition," *BMC Bioinf.*, vol. 10, no. 1, p. 223, 2009.

[2] M. Torii, Z. Hu, C. H. Wu, and H. Liu, "BioTagger-GM: A gene/protein name recognition system," *J. Amer. Med. Informat. Assoc.*, vol. 16, no. 2, pp. 247–255, 2009.

[3] Z. Liao and H. Wu, "Biomedical named entity recognition based on skip-chain Crfs," in *Proc. Int. Conf. Ind. Control Electron. Eng.*, Aug. 2012, pp. 1495–1498.

[4] B. Tang, H. Cao, X. Wang, Q. Chen, and H. Xu, "Evaluating word representation features in biomedical named entity recognition tasks," *BioMed Res. Int.*, vol. 2014, Mar. 2014, Art. no. 240403.

[5] K. Li, W. Ai, F. Zhang, L. Jiang, K. Li, and K. Hwang, "Hadoop recognition of biomedical named entity using conditional random fields," *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 11, pp. 3040–3051, Nov. 2015.

[6] M. Chen, F. Herrera, and K. Hwang, "Cognitive computing: Architecture, technologies and intelligent applications," *IEEE Access*, vol. 6, pp. 19774–19783, 2018.

[7] K. Hwang and M. Chen, *Big–Data Analytics for Cloud, IoT and Cognitive Computing*. Hoboken, NJ, USA: Wiley, 2017.

[8] M. Chen, W. Li, G. Fortino, Y. Hao, L. Hu, and I. Humar, "A dynamic service-migration mechanism in edge cognitive computing," Aug. 2018, *arXiv:1808.07198*. [Online]. Available: https://arxiv.org/abs/1808.07198

[9] M. Chen, Y. Miao, X. Jian, X. Wang, and I. Humar, "Cognitive-LPWAN: Towards intelligent wireless services in hybrid low power wide area networks," *IEEE Trans. Green Commun. Netw.*, vol. 3, no. 2, pp. 409–417, Jun. 2019.

[10] M. Chen, X. Shi, Y. Zhang, D. Wu, and M. Guizani, "Deep features learning for medical image analysis with convolutional autoencoder neural network," *IEEE Trans. Big Data*, to be published.

[11] M. Chen, J. Zhou, G. Tao, J. Yang, and L. Hu, "Wearable affective robot," *IEEE Access*, vol. 6, pp. 64766–64776, 2018.

[12] M. Chen, Y. Hao, K. Lin, L. Hu, and Z. Yuan, "Label-less learning for traffic control in an edge network," *IEEE Netw.*, vol. 32, no. 6, pp. 8–14, Nov./Dec. 2018.

[13] Y. Ma, Y. Hao, M. Chen, J. Chen, P. Lu, and A. Košir, "Audio-visual emotion fusion (AVEF): A deep efficient weighted approach," *Inf. Fusion*, vol. 46, pp. 184–192, Mar. 2019.

[14] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," Aug. 2015, *arXiv:1508.01991*. [Online]. Available: https://arxiv.org/abs/1508.01991

[15] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional LSTM-CNNS-CRF," Mar. 2016, *arXiv:1603.01354*. [Online]. Available: https://arxiv.org/abs/1603.01354

[16] J. P. C. Chiu and E. Nichols, "Named entity recognition with bidirectional LSTM-CNNs," Nov. 2015, *arXiv:1511.08308*. [Online]. Available: https://arxiv.org/abs/1511.08308

[17] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," Mar. 2016, *arXiv:1603.01360*. [Online]. Available: https://arxiv.org/abs/1603.01360

[18] L. Yao, H. Liu, Y. Liu, X. Li, and M. W. Anwar, "Biomedical named entity recognition based on deep neutral network," *Int. J. Hybrid Inf. Technol*, vol. 8, no. 8, pp. 279–288, 2015.

[19] L. Li, L. Jin, Y. Jiang, and D. Huang, "Recognizing biomedical named entities based on the sentence vector/twin word embeddings conditioned bidirectional LSTM," in *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*. Berlin, Germany: Springer-Verlag, 2016, pp. 165–176.

[20] D. Zeng, C. Sun, L. Lin, and B. Liu, "LSTM-CRF for drug-named entity recognition," *Entropy*, vol. 19, no. 6, p. 283, Jun. 2017.

[21] Z. Zhao, Z. Yang, L. Luo, L. W. author, Y. Zhang, H. Lin, and J. Wang, "Disease named entity recognition from biomedical literature using a novel convolutional neural network," *BMC Med. Genomics*, vol. 10, no. 5, p. 73, Dec. 2017.

[22] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," Sep. 2014, *arXiv:1409.0473*. [Online]. Available: https://arxiv.org/abs/1409.0473

[23] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," Aug. 2015, *arXiv:1508.04025*. [Online]. Available: https://arxiv.org/abs/1508.04025

[24] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[25] O. Vinyals, L. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. Hinton, "Grammar as a foreign language," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2773–2781.

[26] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, "Teaching machines to read and comprehend," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1693–1701.

[27] C. Pandey, Z. Ibrahim, H. Wu, E. Iqbal, and R. Dobson, "Improving RNN with attention and embedding for adverse drug reactions," in *Proc. Int. Conf. Digit. Health*, 2017, pp. 67–71.

[28] L. Luo, Z. Yang, P. Yang, Y. Zhang, L. Wang, H. Lin, and J. Wang, "An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition," *Bioinformatics*, vol. 34, no. 8, pp. 1381–1388, 2017.

[29] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," Jan. 2013, *arXiv:1301.3781*. [Online]. Available: https://arxiv.org/abs/1301.3781

[30] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[31] R. Jozefowicz, W. Zaremba, and I. Sutskever, "An empirical exploration of recurrent network architectures," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2342–2350.

[32] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[33] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Netw.*, vol. 18, no. 5, pp. 602–610, 2005.

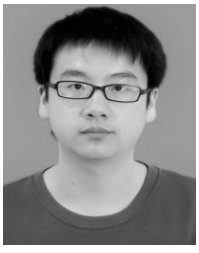
[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Dec. 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980

[35] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[36] H.-J. Dai, P.-T. Lai, Y.-C. Chang, and R. T.-H. Tsai, "Enhancing of chemical compound and drug name recognition using representative tag scheme and fine-grained tokenization," *J. Cheminform.*, vol. 7, no. S1, p. S14, 2015.

[37] L. Ratinov and D. Roth, "Design challenges and misconceptions in named entity recognition," in *Proc. 13th Conf. Comput. Natural Lang. Learn.* Stroudsburg, PA, USA: Assoc. Comput. Linguistics, 2009, pp. 147–155.

[38] B. Tang, H. Cao, Y. Wu, M. Jiang, and H. Xu, "Recognizing clinical entities in hospital discharge summaries using structural support vector machines with word representation features," *BMC Med. Informat. Decis. Making*, vol. 13, p. S1, Apr. 2013.
[39] Y. Song, E. Kim, G. G. Lee, and B.-K. Yi, "POSBIOTM-NER in the shared task of BioNLP/NLPBA 2004," in *Proc. Int. Joint Workshop Natural Lang. Process. Biomed. Appl.* Stroudsburg, PA, USA: Assoc. Comput. Linguistics, 2004, pp. 100–103.
[40] J. Finkel, S. Dingare, H. Nguyen, M. Nissim, C. Manning, and G. Sinclair, "Exploiting context for biomedical entity recognition: From syntax to the web," in *Proc. Int. Joint Workshop Natural Lang. Process. Biomed. Appl.* Stroudsburg, PA, USA: Assoc. Comput. Linguistics, 2004, pp. 88–91.
[41] B. Settles, "ABNER: An open source tool for automatically tagging genes, proteins and other entity names in text," *Bioinformatics*, vol. 21, no. 14, pp. 3191–3192, 2005.
[42] Y. Tsuruoka, Y. Tateishi, J.-D. Kim, T. Ohta, J. McNaught, S. Ananiadou, and J. Tsujii, "Developing a robust part-of-speech tagger for biomedical text," in *Proc. Panhellenic Conf. Inform.* Volos, Greece: Springer, 2005, pp. 382–392.
[43] F. X. Chang, J. Guo, W. R. Xu, and S. R. Chung, "Application of word embeddings in biomedical named entity recognition tasks," *J. Digit. Inf. Manage.*, vol. 13, no. 5, pp. 321–327, 2015.
[44] D. Campos, S. Matos, and J. L. Oliveira, "Gimli: Open source and high-performance biomedical name recognition," *BMC Bioinf.*, vol. 14, no. 1, p. 54, 2013.
[45] G. Zhou, J. Zhang, J. Su, D. Shen, and C. Tan, "Recognizing names in biomedical texts: A machine learning approach," *Bioinformatics*, vol. 20, no. 7, pp. 1178–1190, 2004.
[46] Q. Zhu, X. Li, A. Conesa, and C. Pereira, "GRAM-CNN: A deep learning approach with local context for named entity recognition in biomedical text," *Bioinformatics*, vol. 34, pp. 1547–1554, May 2018.
[47] R. T.-H. Tsai, C.-L. Sung, H.-J. Dai, H.-C. Hung, T.-Y. Sung, and W.-L. Hsu, "NERBio: Using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition," *BMC Bioinf.*, vol. 7, p. S11, Dec. 2006.
[48] X. Wang, Y. Zhang, X. Ren, Y. Zhang, M. Zitnik, J. Shang, C. Langlotz, and J. Han, "Cross-type biomedical named entity recognition with deep multi-task learning," Jan. 2018, *arXiv:1801.09851*. [Online]. Available: https://arxiv.org/abs/1801.09851
[49] C. Lyu, B. Chen, Y. Ren, and D. Ji, "Long short-term memory RNN for biomedical named entity recognition," *BMC Bioinf.*, vol. 18, no. 1, p. 462, 2017.
[50] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," Jan. 2019, *arXiv:1901.08746*. [Online]. Available: https://arxiv.org/abs/1901.08746

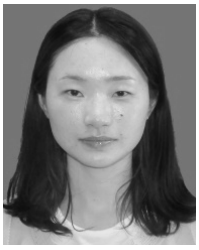


HAO WEI received the M.Sc. degree from Yunnan Normal University, in 2017. He is currently pursuing the Ph.D. degree with the Information Science and Technology College, Dalian Maritime University. His research interests include natural language processing, text mining, and information extraction for biomedical literatures.

MINGYUAN GAO received the bachelor's degree from Dalian Jiaotong University, in 2017. He is currently pursuing the master's degree in science with the School of Information Science and Technology, Dalian Maritime University. His main research interest includes natural language processing.

AI ZHOU received the master's degree from Hong Kong Polytechnic University, in 2017. She is currently pursuing the Ph.D. degree with the School of Information Science and Technology, Dalian Maritime University. Her research interests include natural language processing and stylometry.

FEI CHEN received the Ph.D. degree in management science and engineering from Dalian Maritime University, China, in 2010, where he is currently a Lecturer with the Information Science and Technology College. His research interests include database, machine learning, and natural language processing.



WEN QU received the Ph.D. degree in computer software and theory from Northeastern University, China. She is currently a Lecturer with the School of Information Science and Technology, Dalian Maritime University, Dalian. Her research interests include data mining and machine learning.

CHUNLI WANG received the Ph.D. degree in computer science from the Dalian University of Technology, in 2003. She is currently a Professor with the School of Information Science and Technology, Dalian Maritime University, China. Her research interests include pattern recognition and machine learning.

MINGYU LU was born in 1963. He received the Ph.D. degree from Tsinghua University, in 2002. He is currently a Profession and a Doctoral Supervisor with Dalian Maritime University. His research interests include data mining, pattern recognition, machine learning, and natural language processing.