ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20230324000477
Original scientific paper

Chinese Named Entity Recognition Method for Domain-Specific Text

He LIU, Yuekun MA*, Chang GAO, Jia QI, Dezheng ZHANG*

Abstract: The Chinese named entity recognition (NER) is a critical task in natural language processing, aiming at identifying and classifying named entities in text. However, the specificity of domain texts and the lack of large-scale labelled datasets have led to the poor performance of NER methods trained on public domain corpora on domain texts. In this paper, a named entity recognition method incorporating sentence semantic information is proposed, mainly by adaptively incorporating sentence semantic information into character semantic information through an attention mechanism and a gating mechanism to enhance entity feature representation while attenuating the noise generated by irrelevant character information. In addition, to address the lack of large-scale labelled samples, we used data self-augmentation methods to expand the training samples. Furthermore, we introduced a Weighted Strategy considering that the low-quality samples generated by the data self-augmentation process can have a negative impact on the model. Experiments on the TCM prescriptions corpus showed that the F1 values of our method outperformed the comparison methods.

Keywords: attention mechanism; data augmentation; domain text; meta-learning; named entity recognition

1 INTRODUCTION

The named entity recognition task which aims to identify useful knowledge from unstructured Chinese text is the basis of natural language processing tasks. However, NER methods obtained using public corpus training do not perform well on domain-specific texts, which have their own unique linguistic characteristics. For example, TCM prescriptions texts have a large number of TCM terms that may not be recognized by general NER models trained in news articles or social media texts.

At present, Chinese entity extraction research objects are mainly unstructured public domain text data [1-3], while domain knowledge is mostly recorded in professional terms with more concise language, which intensifies the reliance of entity semantic information on the semantic information of the context. The cost of good labelling guidelines, as well as personnel training [4], makes manual labelling expensive and leads to a lack of large-scale domain labelled datasets. Consequently, the Chinese named entity recognition task for specific domains is more challenging. To address the above problems, Ma et al. [5] started from the study of domain text features and designed a multi-granularity text feature extractor to extract contextual semantic information of text from different granularities. In addition, conditional generative adversarial networks are used to generate pseudo-training samples to address the effect of data size on the model, and excellent results have been achieved on several TCM datasets. The adversarial generative network has its own drawback of long computation time, although it can solve the adverse effects brought by data scarcity. Jia et al. [6] used a distant supervision named entity recognition method, which used domain dictionaries and unlabeled text to generate training samples, and proposed a method for spanwise detection of entities based on this. Ma et al [7] used dictionary information to enhance character features to improve the accuracy of the task. These two methods outperform other methods in the TCM datasets. However, domain dictionaries are always incomplete, and do not contain all entities in practice, as well as the cost of constructing a high-quality dictionary is high. Yang used cross-resource word vectors to bridge low and high resource domains for knowledge transfer [8]. Although transfer learning-based method can improve the accuracy of domain named entity recognition task, the existence of unused entity types in different domains is still an unavoidable problem, especially for the NER task of text data in TCM domain.

In the language domain, researchers hold that the meaning of a word is determined by the context in which it occurs, i.e., the distributional hypothesis, and it has been well verified in the task of learning word vector representations in natural language processing. In Chinese, sentences are composed of characters and word groups (or phrases), and the characters and word groups together constitute the complete sentence semantics, so making full use of the semantic information of the sentences is able to enhance the semantic information of the characters. In the work of Jia et al. [6], the sentence semantic feature vector was spliced as an additional feature in order to enhance the semantic information representation by directly following each character fragment vector representation that was to be detected, while the method did not take into account the low correlation between the unsubstantiated characters or words in the sentence and the semantics of the whole sentence. Therefore, each character in a sentence must be considered for its relevance to the overall sentence semantics in practical applications. Moreover, a small amount of domain labelled data can also have an impact on the model performance.

Data self-augmentation is a method that can effectively alleviate data scarcity. Dai and Adel [9] first applied the data self-augmentation method to the NER task and achieved impressive results. However, the aspect of semantic correctness will inevitably produce low-quality augmented data with latent noise [10], which will inevitably affect the overall performance of the model. Therefore, reducing the impact of low-quality augmented data on the domain NER task is also a challenge.

In this paper, we propose a Chinese named entity recognition method for domain text. Specifically, our method involves two sub-methods: (1) a named entity recognition method combining semantic information of sentences, and (2) a data augmentation method incorporating entity replacement and meta-weighting learning. For the named entity recognition method incorporating global sentence semantic information, the

dynamic vector embedding representation is first generated by encoding the sentence sequence using a pre-training model; meanwhile the pre-training model is able to generate a sentence semantic vector representation. Secondly, the method takes the overall semantic information of the sentence as the backbone feature and uses the attention mechanism and gating mechanism to control the adaptive incorporation into the character information, which is used to enhance the information representation of the characters themselves in the sentence. Then, the temporal information is further captured using the bidirectional long short-Term memory (BiLSTM) network to form the complete contextual semantic features for the sequence labelling task, and finally the classification prediction is performed using conditional random fields (CRF). The data augmentation method that incorporates entity replacement and meta-weighting first uses entity replacement to form new training samples for the entities of the sentences, while the purpose of meta-weighting is to weaken the effect of low-quality augmentation data on the model to better improve the model performance. Experimental results on the datasets of TCM prescriptions domain show that our method outperforms other comparative methods.

2 RELATED WORKS

2.1 Research Status of TCM Prescriptions and NER

The domain NER technology is dependent on the NER's technology development. Early domain NER approaches used lexicon and rule-based pattern matching methods, e.g., some research scholars constructed a Uyghur personal name data dictionary for Uyghur NER [11], and if there are entities in the text that are not included in the dictionary, they are manually entered into the dictionary for the next recognition. Based on this, more accurate extraction of entities is achieved by constructing relevant rules, such as in the work of Li et al. [12], where chemical substance name extraction is performed by constructing the rule of chemical + preposition + chemical prefix chemical symbol. Although pattern matching methods are able to obtain entities accurately, entity rule setting relies on domain experts and the migration effect is poor. Moreover, the field dictionaries also require long-term maintenance and the labor cost is higher.

With the statistical machine learning era that follows, the accuracy of domain NER tasks is greatly improved and the dependence on domain experts is reduced. The domain NER based on statistical machine learning is critical in feature extraction, which comes from a collection of features that reflect the characteristics of a certain class of entities. Yu et al. [13] used Cascading Hidden Markov (HMM) for Chinese names recognition, and Guo et al. [14] used CRF for NER in the tourism domain. Although statistical machine learning can significantly improve the accuracy of domain NER task, it is limited by a high-quality, large-scale labelled corpus.

Currently, deep learning-based methods are the dominant methods for domain NER, which enable machines to automatically identify latent features. Deng et al [15] introduced the BiLSTM-CRF model to the TCMNER task and significantly improved the accuracy. Li et al. [16] achieved good results in the medical NER task by integrating attention into the BiLSTM model, which significantly improved the information loss problem caused by distance through the attention mechanism. Methods based on deep learning have the same dependency on a high-quality corpus. Zhang et al [17] combined the pre-training model BERT and BiLSTMCRF to address the problem of fuzzy entity recognition and less labelled data in the TCM domain. Shi Yuan [18] used domain knowledge to improve the performance of named entity recognition models for less resourced domains. However, this method also required a high quality large-scale lexicon along with a high quality word vector embedding representation, which is more difficult in some domains. All the above methods use neural networks to mine implicit global semantic features in sentences to improve model performance, while ignoring the utilization of overall sentence semantic information. Therefore, a NER model incorporating semantic information of sentences is constructed in this paper.

2.2 Global Sentence Semantic Representation

Sentences are the basic unit of language usage. In natural language processing tasks, global sentence semantics was commonly used in text classification [19], text similarity matching [20] and sentiment classification tasks [21], etc. For these tasks, researchers have used a neural network such as CNN or BiLSTM to encode the input text to obtain contextual semantic information, and then form a global sentence semantic representation through relevant operations such as averaging pooling. With the widespread use of attention mechanisms in natural language tasks, some researchers have introduced attention mechanisms to enhance the contextual semantic information of sentences to form more accurate global sentence semantic representations. In a recent relational extraction task, Xu et al. [22] proposed a global gating mechanism (GGM) for transferring global semantic information to local information representations to enhance the expressiveness of the sentence information itself. In the named entity recognition task, Jia et al. [6] proposed a span-level named entity recognition model and used the [CLS] feature vector generated by the BERT model as a global sentence semantic representation linked after each character span fragment feature that was to be detected and used to enhance the features of each span to be detected. Inspired by previous work and based on the characteristics of Chinese language, we proposed a word-sentence adaptation unit to enhance character information feature representation by controlling global sentence semantic information using attention mechanism and gating mechanism.

2.3 Data Augmentation

Data augmentation technique is an effective strategy to alleviate data scarcity in deep learning, and has been widely used in natural language processing tasks. The mainstream techniques for data augmentation technologies are word replacement, mention replacement and paraphrasing. (1) Word replacement is a local replacement in a given sentence, such as Wei and Zhou et al. [23] who used synonym replacement of words in a sentence to

generate pseudo-training data for improving the accuracy of text classification tasks. Wang et al [24] expanded the training data by replacing words in the source and target sentences and thus generating additional parallel sentence pairs to improve the accuracy of the machine translation task under low resource conditions. (2) Mention replacement is entity replacement using entities of the same entity type in the training set, such as Zhao et al. [25] who exchanged all male entities and female entities in the datasets. (3) Paraphrasing is sentence-level rewriting and does not significantly change the semantic information of the sentence. For example, Zhang et al. [26] used a translation interface to translate the source data language into another language and later into the source data language for data augmentation. However, noise is inevitably introduced in the data augmentation process, which has a negative impact on the model. Therefore, this paper introduced a data augmentation method incorporating entity replacement and meta-weighting learning

3 MODEL

In this section, we will describe the proposed NER model in detail. Fig. 1 shows the overall architecture, which consists of four main components: (1) text encoding layer, (2) char-sentence adaptation layer, (3) temporal feature encoding layer and (4) feature decoding layer. The text encoding layer is composed of a pre-training BERT model, which aims to generate character embedding vectors containing global contextual semantic features and global sentence semantic feature embedding vectors for the input sentences.

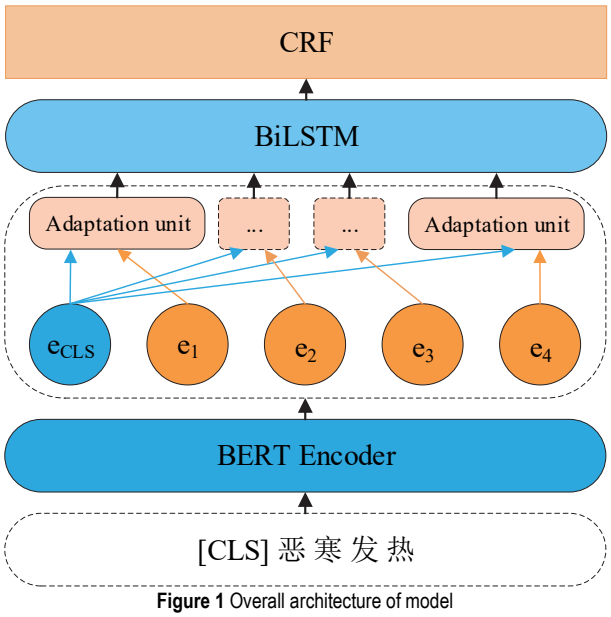

Figure 1 Overall architecture of model

The char-sentence adaptation layer consists of multiple identically structured char-sentence adaptation units, which is used to fuse global sentence semantic features with character semantic features to enhance the character semantic representation. The temporal feature encoding layer uses BiLSTM to further capture the temporal features of the input sentences, and can obtain more complete representation of the contextual semantic features. The feature decoding layer is composed of CRF, which aims to decode the contextual semantic features and predict the optimal label sequences.

3.1 Text Encoding Layer

To enhance the semantic representation of entity contextual information, the pre-training BERT model is used to encode sentences to form character embedding feature representations. The pre-training BERT model consists of a multi-layer bidirectional Transformer [27] encoder, which is trained and generates deep linguistic representations in a masked manner. The inputs to the BERT model are token embedding, location embedding and segmental embedding. Token embedding is usually a character vector, location embedding is used to mark the location information of each token, and segmental embedding is used to distinguish different sentences. The model is able to generate vector representations corresponding to the specific context of the sentence. In practice, the BERT model usually adds [CLS] flags for representing the overall information of a sentence and has been shown to be effective for natural language processing task [28]. In NER task, [CLS] flag vectors are usually discarded and are not involved in model training. In this paper, [CLS] will be used to enhance the information of each character in a sentence and will be described in detail in Section 3.2.

As in Fig. 1, given a sequence of input sentence characters $X = \{x_1, ..., x_n\}$ and n denoting the length of the input sentence sequence, we first form a new sequence of sentence characters $\tilde{X} = \{[cls], x_1, ..., x_n\}$ by adding the [CLS] flag to the sentence character sequence. Then the BERT model is used to generate the character embedding sequence $E = \{e_{CLS}, e_1, ..., e_n\}$, e_{CLS} represent the semantic information of the whole sentence sequence, and the process can be formalized as follows:

$$E = BERT(\tilde{X}) \tag{1}$$

3.2 Char-Sentence Adaptation Layer

In Chinese, sentences consist of characters and words (or phrases) which are capable of expressing the overall semantics of all characters. Linguistic studies have shown that the focus of sentence semantic understanding is to capture the semantic gravity of the sentence. The gravity of a sentence is the main part of the sentence, carried by the word, phrase or part of the sentence or the whole sentence, which covers all entity information and potential category information. From the characteristics of Chinese language and based on the goal of giving full play to the semantic information of sentences, we propose a char-sentence adaptation layer, which consists of multiple identical char-sentence adaptation units, and adaptively incorporates the sentence feature vectors into the character information to enhance the semantic information representation of characters.

To incorporate sentence semantic information into character information, inspired by recent research on BERT adapters [29], we propose a novel char-sentence

adaptation unit, as shown in Fig. 2, which adaptively assigns weights to sentence semantic feature vectors using an attention mechanism and achieves character semantic vector augmentation through a gating mechanism. The char-sentence adaptation unit accepts two inputs: a character feature vector and a sentence feature vector. For the character at position i in the sentence, the input is denoted as (e_i, e_{CLS}), e_i denotes the character feature vector and e_{CLS} is the sentence feature vector, and both vectors are the output of BERT. As shown in Fig. 1, the sentence semantic feature vector e_{CLS} is derived by summarizing multiple character semantic feature vectors. To more accurately use the sentence semantic features to enhance the character features, we introduce a character-sentence attention mechanism, which calculates the similarity between each character feature and sentence feature as the weight of each character receiving sentence semantic feature augmentation.

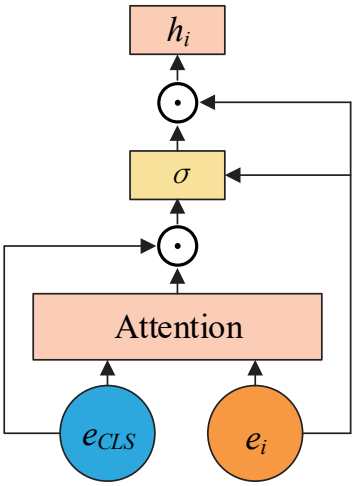

Figure 2 Char-sentence adaptation unit

Specifically, this paper first calculates the similarity between each character vector e_i and the sentence feature vector e_{CLS} as the weights of the semantic features of the sentences:

$$a_i = \text{soft}\max(e_i W_{\text{atten}} (e_{CLS})^T) \tag{2}$$

where W_{atten} is the weight matrix of attention. Thus, the weighted sentence feature vectors v_i corresponding to each character can be calculated as follows:

$$v_i = a_i e_{CLS} \tag{3}$$

Then the enhanced data are obtained by Eqs. (4) and (5):

$$g_i = \sigma(W_e e_i + W_v v_i + b) \tag{4}$$

$$h_i = e_i \odot g_i \tag{5}$$

where W_e and W_v are the learnable weight matrices, b is the bias matrix, $\sigma(\bullet)$ denotes the Sigmoid function, and $\odot$ denotes the vector multiplication. Among them, Eq. (4) is used to retain the global information related to the current character and forget the irrelevant global information, and Eq. (5) achieves the transfer of global information to the current character feature vector representation by performing the multiplication of gating information g_i and character vector e_i, so as to achieve the purpose of enhancing the semantic information of the character vector using the sentence semantic information.

3.3 Temporal Feature Encoding Layer

The character vector representation generated by the pre-training BERT model contains rich raw semantic information. To more accurately identify entities in sentences, this paper fuses the vector matrix E output from the feature encoding layer and the vector matrix $\tilde{E}$ output from the char-sentence adaptation layer to obtain a more complete character feature matrix as follows:

$$\tilde{E} = E \oplus H \tag{6}$$

where $\oplus$ indicates that the corresponding elements are summed.

Temporal features are crucial for the NER task. Although the fused character sequence feature matrix $\tilde{E}$ is rich in contextual semantic information, temporal features are still lacking in these features. To capture the temporal information of the character sequence feature matrix $\tilde{E}$, this paper follows previous work [30] and uses BiLSTM to encode the character feature matrix $\tilde{E}$, which is used to capture the temporal features of the character feature matrix $\tilde{E}$ to form a more complete contextual feature information, and the process can be formalized as follows:

$$\tilde{h}_1, ..., \tilde{h}_n = \text{BiLSTM}(h_1, ..., h_n) \tag{7}$$

3.4 Feature Decoding Layer

Considering the dependency between sequential labels, we use the CRF to decode the features. Given that the output of BiLSTM is $\tilde{H} = \{\tilde{h}_1, ..., \tilde{h}_n\}$, the label fraction O is first calculated by linear variation layer as follows:

$$O = W\tilde{H} + b \tag{8}$$

where W is the parameter of the linear transform layer and b is the bias parameter of the linear transform layer. For the label sequence $y = \{y_1, y_2, ..., y_{n-1}, y_n\}$, the probability is defined in this paper as follows:

$$p(y|x) = \frac{\exp\left(\sum_i \left(O_{i,y_i} + T_{y_{i-1},y_i}\right)\right)}{\sum_{\tilde{y}} \exp\left(\sum_i \left(O_{i,\tilde{y}_i} + T_{\tilde{y}_{i-1},\tilde{y}_i}\right)\right)} \tag{9}$$

where T is the transfer matrix and $\tilde{y}$ denotes the candidate label sequence.

3.5 Training Goals

In this paper, the model is trained by minimizing the negative log similarity loss at the sentence level, and given the training data $\{x_j, y_j\}, 1 \le j \le N$ calculates the loss as follows:

$$L = -\sum_{j} \log(p(y_j \left| x_j \right)) \tag{10}$$

Finally, the Viterbi algorithm [31] is used to calculate the best label sequence.

4 A DATA AUGMENTATION METHOD INCORPORATING ENTITY REPLACEMENT AND META-WEIGHTING LEARNING

4.1 Entity Replacement

Data augmentation is a method that can effectively address data scarcity, and in this paper, entity replacement is used to expand the training set. Entity replacement refers to replacing entities in sentences using synonyms to generate pseudo-data based on the raw token data. In this work, entities with the same category in the TCM prescriptions training set are considered as synonyms and a TCM prescriptions synonym dictionary is constructed. Specifically, for a given training sample (*X*, *Y*), the sampled entities in sample *X* are replaced using entities of the same category from the TCM synonym dictionary. For the presence of multiple entities in the training sample *X*, one of the entities is randomly selected as the sampled entity. Then a pseudo-training sample $(\overline{X}, \overline{Y})$ is obtained. For example, given a sample of "恶寒发热[B-symptom I-symptom I-symptom E-symptom]", a pseudo-sample of "恶寒无汗 [B-symptom I-symptom I-symptom E-symptom]" is generated after entity replacement.

4.2 Meta Weighted Strategy

We adopt the data augmentation method which only replaces the entities in the sentences, and do not consider whether the overall semantics of the sentences changes after the replacement, so there may be low-quality pseudo-samples in the training set. The terminology of traditional Chinese medicine prescriptions is simple and concise, and its semantic information is highly dependent on the semantic information of context, so it is very important to control the quality of pseudo-training data. Thus, we introduce the meta weighted mechanism proposed by Ren et al. [32] into the TCM prescriptions NER task, where we dynamically assign weights to each of the pseudo-training data based on the gradient direction. The main idea is to use a set of small, clean and unbiased samples as validation samples to guide the model training, and use the loss generated by this set of samples to weight the pseudo-training data in each batch, i.e., if the distribution and gradient descent direction of the pseudo-training samples are close to the distribution and gradient descent direction of the validation samples, then the samples need to increase the weights, and vice versa, the weights need to be reduced.

Specifically, given a set of training data $\{(X_i, Y_i), 0 < i < N\}$ a set of pseudo-training data $\{(\overline{X}_i, \overline{Y}_i), 0 < i < N\}$ is generated using the entity replacement method, while a small set of training data $\{(X_i, Y_i), 0 < i < M\}$ is randomly taken from the training data set as the validation set, and $M << N$. The expected loss of minimizing a small batch of the pseudo-training set in this paper:

$$\frac{1}{N}\sum_{i=1}^{N} f_i(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(f\left(\overline{X}_i; \theta\right), \overline{Y}_i\right) \tag{11}$$

where $f_i(\theta)$ denotes the loss of the *i*-th sample in the pseudo-training data, then the training goal of this paper is to minimize the weighted loss as follows:

$$\theta^*(w) = \arg\min_{\theta} \sum_{i=1}^{N} w_i f_i(\theta) \tag{12}$$

where $w_i \ge 0$ denotes the learning weight of the *i*-th sample in the pseudo-training data. The optimal parameters are further determined by the performance in the validation set w^*:

$$w^* = \arg\min \frac{1}{M}\sum_{i=1}^{M} f_i\left(\theta^*(w)\right) \tag{13}$$

To improve efficiency, an approximate online approach is used to update the weights, using stochastic gradient descent (SGD) to calculate the loss. At each step of training, a small batch of pseudo-training data $\{(\overline{x}_i \overline{y}_i), 0 < i < n\}$, with *n* as the minimum batch sample size, is used to adjust the parameters according to the direction of descent of the expected loss:

$$\theta_{t+1} = \theta_t - \alpha \nabla \left(\frac{1}{n}\sum_{i=1}^{n} f_i\left(\theta_t\right) \right) \tag{14}$$

where α is the step size, and then add interference to each pseudo-sample using ε_i:

$$f_{i,\epsilon}(\theta) = \varepsilon_i f_i(\theta) \tag{15}$$

$$\hat{\theta}_{t+1}(\varepsilon) = \theta_t - \alpha \nabla \sum_{i=1}^{n} f_{i,\varepsilon}(\theta)\Big|_{\theta=\theta_t} \tag{16}$$

The loss is calculated based on the updated parameter θ for the smallest batch of pseudo-training data as follows:

$$Loss(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\left(f\left(\overline{x}_i; \hat{\theta}_{t+1}\right), \overline{y}_i\right) \tag{17}$$

To generalize parameter $\hat{\theta}$ to the pseudo-training data set, ε is used as a meta-loss to generate sample weights, which are normalized along small batches as follows:

$$\hat{w}_i = \sigma\left(-\nabla_{\varepsilon_i} L(\hat{\theta})\Big|_{\varepsilon_i=0}\right) \tag{18}$$

$$w_i = \frac{\hat{w}_i}{\sum_j \hat{w}_j + \delta} \tag{19}$$

where $\sigma(\cdot)$ is the Sigmoid function and δ is a minimal value used to prevent the divisor from being 0. From this, the weight of each pseudo-sample can be calculated. Finally, the weights are used to weight the losses generated by the pseudo-training data to obtain the final losses.

In this work, the raw clean training samples are used as validation set samples, and Fig. 3 shows how the pseudo-samples are formed and how they participate in model training.

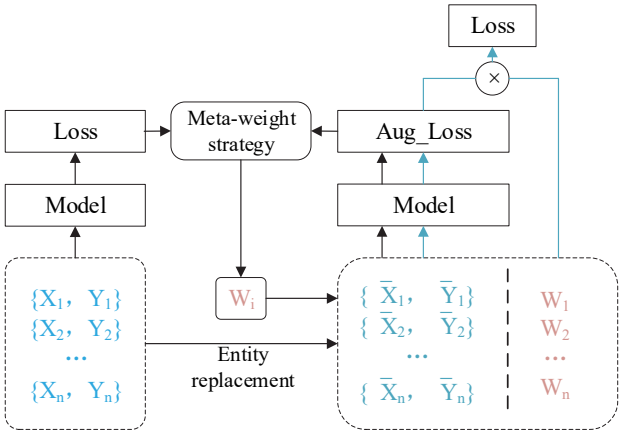

Figure 3 Meta weight strategy framework

5 EXPERIMENTS

5.1 Experimental Settings

1) Data set

In this paper, we use BIOES labeling scheme with six entity types, namely: "prescription", "dosage", "pulse", "tongue", "symptom", and "herbal medicine". In this work, the ratios of training set, test set and validation set are 8:1:1, and the detailed statistics are shown in Tab: 1.

Table 1 Datasets statistics

Data Set	Number of sentences	Number of characters	Number of entities
Training Set	3648	82490	5768
Validation Set	456	10337	1236
Test Set	456	9571	1154

2) Experimental details

In this paper, the BERT-base (Chinese version) model is used as an encoder and its internal parameters are fine-tuned with a learning rate of $3e^{-5}$ and $1e^{-4}$ for other parameters in the model, and the model parameters are updated to using the Adam optimizer. The learning rate of SGD is $1e^{-3}$. Gradient cropping is used to avoid gradient explosion with a parametric number of 5. The number of pseudo-samples involved in training is the same as the number of training samples when data augmentation methods are involved in this paper. All experiments are performed on a NIVIDIA GeForce RTX 3090 (24G) GPU.

3) Evaluation metrics

In this paper, the performance of the models is evaluated using $F1$ value, precision rate (P) and recall rate (R). To fairly evaluate the model performance, these experiments were carried out under standard resource and small sample conditions, and the average value of multiple tests for each model: three experiments were carried out for each model under standard resource conditions, and six experiments were carried out for each model under small sample conditions.

4) Baseline model

To comprehensively evaluate the effectiveness of our approach, a comparison would be made with the following methods.

BERT-CRF model: based on the training model BERT, many studies showed good results in several domain NER task by combining BERT and CRF models, such as in agriculture [33].

Zhang et al. [34] proposed to generate character embedding vectors for text using a pre-training model BERT for the problem of data scarcity in TCM domain, and then trained the model using BiLSTM-CRF framework. The effectiveness of the approach was demonstrated experimentally on a text datasets in the TCM diagnosis domain.

Dai and Adel [9] were the first to study data augmentation methods for NER task, expanding the training data using word replacement, mention replacement and paraphrase data augmentation methods, respectively, and obtaining optimal experimental results for word replacement methods on biomedical domain datasets.

Ma et al. [35] used a two-tower model to solve the problem of low-resource NER. They used two BERT Encoders to encode token and label separately, and then calculated the similarity between each token and all labels in the text to be predicted, and the label with the highest similarity was the final prediction. The optimal results were obtained in multiple NER datasets.

5.2 Experimental Results

The experimental results as shown in Tabs. 2 and 3 verify the effectiveness of the proposed method under low and high resource conditions, respectively.

1) Small samples setting

In this paper, 100, 200 and 300 data are randomly selected in the raw training set of TCM prescriptions data for experimental validation. Tab. 2 shows the experimental results of this paper's model and other baseline models in the small-sample scenario, and it can be observed: (1) the performance of both the proposed model and other baseline models improves significantly with the increase of training data, which indicates that the supervised NER model is more dependent on the data scale. (2) The use of semantic information of labels is based on the assumption that the model is able to generalize its meaning from the data. Compared with the generic domain entity labels, the entity labels in TCM domain are more abstract, so the performance of the method proposed by Ma et al. is lower in the small samples TCM prescriptions domain. (3) The training data generated by the data augmentation method can improve the performance of the model. (4) The method

proposed in this paper outperforms other models with an average improvement of $\left(\frac{1.99+3+1.53}{3}=2.17\%\right)$ in $F1$ values compared to the method proposed by Zhang et al. and an average improvement of $\left(\frac{0.84+0.13-0.83}{3}=0.04\%\right)$ compared to the method proposed by Dai and Adel et al. It demonstrates that our proposed method of augmenting character semantic features using sentence semantic features is effective in the small samples TCM prescriptions named task, especially at 100 data, with significantly better results than other methods.

2) Standard resource setting

Tab. 3 shows the experimental results of the proposed model and other baseline models for the TCM prescriptions datasets with the standard resources: (1) With sufficient data, our model is still able to achieve comparable results with the comparison model, which proves that the char-sentence adaptation unit we designed can control the information flow well and avoid the possible negative effects caused by information redundancy. (2) Under the condition of standard resources, Dai and Adel's method performed slightly worse, which results from the fact that every training data (including pseudo-training data) is treated equally, even if it is noisy. Also, the phenomenon demonstrates the necessity of introducing the meta weighted strategy low-quality training data negative impact.

Table 2 Performance of model under small samples conditions

Data volume (piece)	Model	*F*1 / %	*P* / %	*R* / %
100	BERT-CRF	71.90	67.74	76.61
	Zhang et al.	75.35	72.16	78.86
	Dai and Adel	76.50	71.54	82.21
	Ma et al.	72.98	69.76	76.54
	Ours	77.34	74.11	80.87
200	BERT-CRF	78.49	74.48	82.97
	Zhang et al.	79.90	78.17	81.72
	Dai and Adel	82.77	79.26	86.60
	Ma et al.	77.23	74.27	80.46
	Ours	82.90	78.76	87.51
300	BERT-CRF	83.23	80.20	86.49
	Zhang et al.	84.09	83.67	84.51
	Dai and Adel	86.45	84.11	88.93
	Ma et al.	84.00	81.57	86.58
	Ours	85.62	83.16	88.23

Table 3 Performance of each model under standard resource conditions

Model	*F*1 / %	*P* / %	*R* / %
BERT-CRF	90.51	89.72	91.79
Zhang et al.	91.00	89.62	92.41
Daiand Adel	90.81	89.49	92.16
Ma et al.	90.54	89.01	92.13
Ours	91.35	90.18	92.74

Table 4 Performance of data augmentation and meta weight strategy in the model

Model	*F*1 / %	*P* / %	*R* / %
100 data			
Ours	77.34	74.11	80.87
Ours + DA	78.85	76.34	81.56
Dai and Adel +MWS	79.12	77.26	81.81
Our + DA + MWS	80.33	77.56	83.97
200 data			
Ours	82.90	78.76	87.51
Ours + DA	83.45	79.28	88.09
Dai and Adel +MWS	84.77	81.26	88.60
Our + DA + MWS	85.00	82.41	87.77
300 data			
Ours	85.62	83.16	88.23
Ours + DA	86.71	84.55	89.06
Dai and Adel+MWS	86.85	84.62	89.13
Our + DA + MWS	86.92	84.73	89.22
Standard resources			
Ours	91.35	90.18	92.74
Ours + DA	91.37	90.27	92.69
Dai and Adel +MWS	91.26	90.15	92.66
Our + DA + MWS	91.59	90.36	92.85

5.3 Effectiveness of Data Augmentation Strategy

In this section, we evaluate the effectiveness of the data augmentation method and meta weighted strategy based on entity replacement in an integrated manner through an ablation study. The contents involved are as follows: (1) using the data augmentation (DA) method proposed in the model; (2) using the data augmentation (DA) method and the meta weighted strategy (MWS) proposed in the model; and (3) using the meta weighted

strategy (MWS) based on the method proposed by Dai and Adel. The experimental results in Tab. 4 show that (1) the data augmentation method can improve the accuracy of the named entity task of TCM prescriptions, especially the improvement of the model performance in the scenario with small samples is significant; (2) the meta weighted strategy can control the influence of training samples on the model and improve the model performance, which proves that the meta weighted strategy is also effective in the NER task; (3) the experiments also prove from the side that the model proposed is effective in the small samples scenario.

5.4 Discussion

In this section, the role of word-sentence adaptation unit and data augmentation method in small sample scenarios is further discussed.

1) Role of char-sentence adaptation unit

To fully verify the effectiveness of the char-sentence adaptation unit designed in the loss-resource TCM prescriptions NER task, the experimental analysis is performed in a loss-resource scenario. We set three conditions, namely "model with unused sentence semantics", "spliced sentence semantics vectors to character vectors" and "using char-sentence adaptation units", and the experimental results are shown in Fig. 4. The char-sentence adaptation unit proposed can dynamically allocate sentence semantic features for each character, improving the performance of the model in low-resource scenarios.

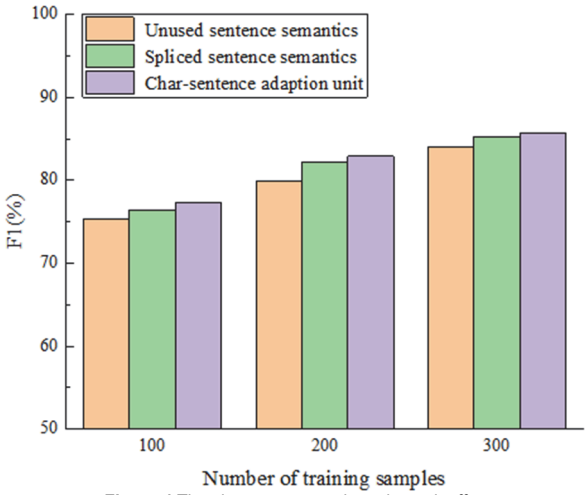

Figure 4 The char-sentence adaptation unit effect

2) The impact of data augmentation scale on model performance

The scale of training samples is an important factor affecting the model. To verify the effect of the number of pseudo-training samples on the model performance, experimental validation is performed in this paper under the conditions of 100 and 200 training samples. As shown in Fig. 5, more pseudo-training samples obtain better performance within a certain range. However, as the number of pseudo-training samples increases, the rising trend of the model performance slows down, and when the number of pseudo-data samples reaches a certain value, the improvement of the model performance tends to level off. It is also observed that the model performance stabilized after 400 pseudo-training samples under the condition of 100 training samples (Fig. 5, left) and stabilized after 300 pseudo-training samples under the condition of 200 training samples (Fig. 5, right), which shows that the model needs less and less pseudo-training samples as the number of training samples increases.

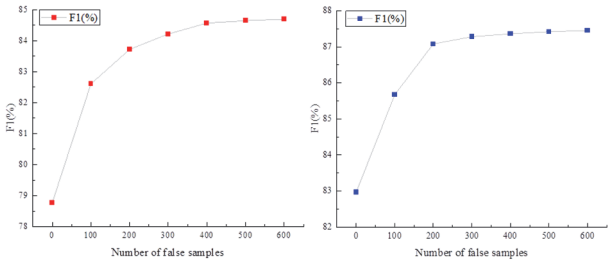

Figure 5 Effect of the number of pseudo-training samples on model performance

3) The effect of data augmentation on model performance improvement under different scale of sample data

Data augmentation is a technique to increase the number of training sets by algorithms, which can effectively alleviate the problem of insufficient data. To verify the effectiveness of data augmentation methods on improving model performance under different scale sample data, experimental analysis is performed in this subsection. We set the ratio of the number of pseudo-training samples to the number of raw training samples to 1:1. Fig. 6 shows the effect of data augmentation techniques on different scales of TCM prescriptions datasets, and it can be observed that: in the TCM prescriptions NER model proposed in this paper, there is no significant improvement in the performance of the model by data self-augmentation technique after the training samples reach 3000.

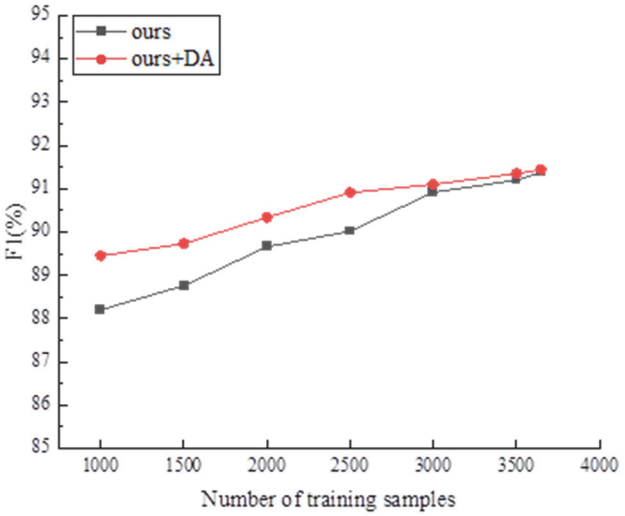

Figure 6 Experimental evaluation of data augmentation in data sets of different Sizes

The reasons for the analysis are as follows: (1) the pseudo-training samples and real training samples generated by data augmentation are treated simultaneously, among which the low-quality pseudo-training samples affect the overall performance of the model; (2) in the TCM

prescriptions text, the entities are relatively dense and the syntax is relatively simple which is more conducive to model learning, and when the data reach a certain level, the model can obtain relatively stable performance.

6 CONCLUSION

In this paper, by analyzing the Chinese language features and combining the knowledge of semantic understanding in linguistics, we designed a named entity recognition method combining sentence semantic information, which can efficiently utilize the sentence semantic information to enhance the character semantic features. Furthermore, we designed a data augmentation method incorporating entity replacement and meta-weighting to expand the labelled data and reduce the negative impact of the generated samples on the model, which improves the accuracy of the domain named entity recognition task. Experiments conducted on the TCM prescriptions datasets showed that our method is effective in a low-resource environment and remains effective under the condition of sufficient training data.

Acknowledgements

This work has been supported by Hebei Province 333 Talent Funding Project "Brain-like Intelligent Knowledge Discovery Technology Research" (Project-number: A201803082).

7 REFERENCES

[1] Cao, P., Chen, Y., Liu, K., Zhao, J., & Liu, S. (2018). Adversarial transfer learning for Chinese NER with self-attention mechanism. *Proceedings of the 2018 conference on empirical methods in natural language processing*, 182-192. https://doi.org/10.18653/v1/D18-1017

[2] Wu, F., Liu, J., Wu, C., Huang, Y., & Xie, X. (2019). Neural Chinese NER via CNN-LSTM-CRF and joint training with word segmentation. *The World Wide Web Conference*, 3342-3348. https://doi.org/10.1145/3308558.3313743

[3] Girsang, A. S., Siswanto, A. V., & Noveta, B. K. (2022). Flood mapping based on online news using named entity recognition. *Journal of System and Management Sciences*, *12*(5), 213-229. https://doi.org/10.33168/JSMS.2022.0513

[4] Huang, J., Li, C., Subudhi, K., Jose, D., Balakrishnan, S., Chen, W., & Han, J. (2020). Few-shot NER: A comprehensive study. arXiv preprint arXiv:2012.14978. https://doi.org/10.48550/arXiv.2012.14978

[5] Ma, Y., Liu, Y., Zhang, D., Zhang, J., Liu, H., & Xie, Y. (2022). A Multigranularity Text Driven NERCGAN Model for Traditional Chinese Medicine Literatures. *Computational Intelligence and Neuroscience*, *2022*. https://doi.org/10.1155/2022/1495841

[6] Jia, Q., Zhang, D., Xu, H., & Xie, Y. (2021). Extraction of traditional Chinese medicine entity: design of a novel span-level NER method with distant supervision. *JMIR Medical Informatics*, *9*(6), e28219. https://doi.org/10.2196/28219

[7] Ma, Y., Liu, H., Zhang, D., Gao, C., & Liu, Y. (2023). A Named Entity Recognition Method Enhanced with Lexicon Information and Text Local Feature. *Tehnički vjesnik*, *30*(3), 899-906. https://doi.org/10.17559/TV-20230121000257

[8] Yang, Z., Salakhutdinov, R., & Cohen, W. (2017). Transfer learning for sequence tagging with hierarchical recurrent networks. arXiv preprint arXiv:1703.06345. https://doi.org/10.48550/arXiv.1703.06345

[9] Dai, X. & Adel, H. (2020). An Analysis of Simple Data Augmentation for Named Entity Recognition. *Proceedings of the 28th International Conference on Computational Linguistics*, 3861-3867. https://doi.org/10.18653/v1/2020.coling-main.343

[10] Wu, L., Xie, P., Zhou, J., Zhang, M., Chunping, M., Xu, G., & Zhang, M. (2022). Robust self-augmentation for NER with meta reweighting. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4049-4060. https://doi.org/10.18653/v1/2022.naacl-main.297

[11] Anwar, A., Li, X., Yang, Y., Dong, R., & Turghun, O. (2020). Constructing Uyghur Name Entity Recognition System using Neural Machine Translation Tag Projection. *Proceedings of the 19th Chinese National Conference on Computational Linguistics*, 1006-1016. https://doi.org/10.1007/978-3-030-63031-7-18

[12] Nan L., Yanting, Z., & Jiuming, J. (2010).Research on Chinese chemical name recognition based on heuristic rules. *New Technology of Library and Information Service*, *26*(5), 13-17. https://doi.org/10.11925/infotech.1003-3513.2010.05.03

[13] Hongkui, Y., Huaping, Z., & Qun L. (2006).Chinese named entity identification using cascaded hidden Markov model. *Journal of Communications*, *27*(2), 87-94. https://doi.org/10.3321/j.issn:1000-436X.2006.02.013

[14] Jianyi, G. & Dagang, T. (2009). Named entity recognition for the tourism domain based on cascaded conditional random fields. *Journal of Chinese Information Processing*, *23*(5), 47-52. https://doi.org/10.3969/j.issn.1003-0077.2009.05.007

[15] Deng, N., Fu, H., & Chen, X. (2021). NER of traditional Chinese medicine patents based on BiLSTM-CRF. *Wireless Communications and Mobile Computing*, *2021*, 1-12. https://doi.org/10.1155/2021/6696205

[16] Li, L., Zhao, J., Hou, L., Zhai, Y., Shi, J., & Cui, F. (2019). An attention-based deep learning model for clinical NER of Chinese electronic medical records. *BMC Medical Informatics and Decision Making*, *5*(235), 1-11. https://doi.org/10.1186/s12911-019-0933-6

[17] Zhang, M., Yang, Z., Liu, C., & Fang, L. (2020).Traditional Chinese medicine knowledge service based on semi-supervised BERT-BiLSTM-CRF model. *2020 International Conference on Service Science (ICSS)*, 64-69. https://doi.org/10.1109/ICSS50103.2020.00018

[18] Shi, Y. (2022). Using Domain Knowledge for Low Resource Named Entity Recognition. arXiv preprint arXiv:2203.14738. https://doi.org/10.48550/arXiv.2203.14738

[19] Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2022). Deep learning--based text classification: a comprehensive review. *ACM computing surveys (CSUR)*, *54*(3), 1-40. https://doi.org/10.1145/3439726

[20] Qurashi, A., Holmes, V., & Johnson, A. (2020). Document processing: Methods for semantic text similarity analysis. *2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, 1-6. https://doi.org/10.1109/INISTA49547.2020.9194665

[21] Ke, Z., Liu, B., Xu, H., & Shu, L. (2021). Classic: Continual and contrastive learning of aspect sentiment classification tasks. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6817-6883. https://doi.org/10.18653/v1/2021.emnlp-main.550

[22] Xu, S., Sun, S., Zhang, Z., Xu, F., & Liu, J. (2022). BERT gated multi-window attention network for relation extraction. *Neurocomputing*, *492*, 516-529. https://doi.org/10.1016/j.neucom.2021.12.044

[23] Wei, J. & Zou, K. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th*

International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 6382-6388. https://doi.org /10.18653/v1/D19-1670

[24] Wang, X., Pham, H., Dai, Z., & Neubig, G. (2018). Switch Out: an efficient data augmentation algorithm for neural machine translation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 856-861. https://doi.org/10.18653/v1/D18-1100

[25] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *2*, 15-20. https://doi.org/10.18653/v1/N18-2003

[26] Zhang, Y., Ge, T., & Sun, X. (2020). Parallel data augmentation for formality style transfer. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3221-3228. https://doi.org/10.18653/v1/2020.acl-main.294

[27] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Proceedings of the 31st International Conference on Neural*, 600-6010. https://doi.org10.5555/3295222.3295349

[28] Reimers, N. & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamesebert-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982-3992. https://doi.org/10.18653/v1/D19-1410

[29] Liu, W., Fu, X., Zhang, Y., & Xiao, W. (2021). Lexicon enhanced Chinese sequence labeling using BERT adapter. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, *1*, 5847-5858. https://doi.org/10.18653/v1/2021.acl-long.454

[30] Lai, P., Ye, F., Zhang, L., Chen, Z., Fu, Y., Wu, Y., & Wang, Y. (2022).PCBERT: Parent and Child BERT for Chinese Few-shot NER. *Proceedings of the 29th International Conference on Computational Linguistics*, 2199-2209. https://aclanthology.org/2022.coling-1.192

[31] Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, *13*(2), 260-269. https://aclanthology.org/10.1109/TIT.1967.1054010

[32] Ren, M., Zeng, W., Yang, B., & Urtasun, R. (2018). Learning to reweight examples for robust deep learning. *International conference on machine learning*, *80*, 4334-4343. https://doi.org/10.48550/arXiv.1803.09050

[33] Zhang, S. & Zhao, M. (2020). Chinese agricultural diseases NER based on BERT-CRF. *5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*, 1148-1151. https://doi.org/10.1109/ICMCCE51767.2020.00252

[34] Zhang, M., Yang, Z., Liu, C., & Fang, L. (2021). Traditional Chinese medicine knowledge service based on semi-supervised BERT-BiLSTM-CRF model. *International Conference on Service Science (ICSS)*, 64-69. https://doi.org/10.1109/ICSS50103.2020.00018

[35] Ma, J., Ballesteros, M., Doss, S., Anubhai, R., Mallya, S., Al Onaizan, Y., & Roth, D. (2022). Label semantics for few shot Named Entity Recognition. *Findings of the Association for Computational Linguistics: ACL 2022*, 1956-1971. https://doi.org/10.18653/v1/2022.findings-acl.155

Contact information:

He LIU
College for Artificial Intelligence,
North China University of Science and Technology,
Tangshan 063210, China
E-mail: liuh@stu.ncst.edu.cn

Yuekun MA
(Corresponding author)
1) College for Artificial Intelligence,
North China University of Science and Technology,
Tangshan 063210, China
2) School of Computer and Communication Engineering,
University of Science and Technology Beijing, Beijing 100083, China
3) Hebei Key Laboratory of Industrial Intelligent Perception,
Tangshan 063210, China
E-mail: mayuekun@ncst.edu.cn

Chang GAO
College for Artificial Intelligence,
North China University of Science and Technology,
Tangshan 063210, China
E-mail: 775480184@qq.com

Qi JIA
Inspur Electronic Information Industry Co., Ltd.
E-mail: jiaqi01@inspur.com

Dezheng ZHANG
(Corresponding author)
School of Computer and Communication Engineering,
University of Science and Technology Beijing, Beijing 100083, China
E-mail: zdzchina@ustb.edu.cn