A Supervised Named Entity Recognition Method Based on Pattern Matching and Semantic Verification

Nan Gao[1], Zhenyang Zhu[1], Zhengqiu Weng[2,1], Guolang Chen[2,3], Min Zhang[1]

[1] College of Computer Science & Technology, Zhejiang University of Technology, China
[2] Department of Information Technology, Wenzhou Polytechnic, China
[3] School of Management, Zhejiang University, China
gaonan@zjut.edu.cn, sivan.zhu@qq.com, derisweng@163.com, 2019f098@zju.edu.cn, larainezm@qq.com

Abstract

Named entity recognition is a basic task in the field of natural language processing and plays a pivotal role in tasks such as information extraction, machine translation, and knowledge graph construction. It has also received widespread attention in financial, biological and pharmaceutical industries. This paper proposes a method of weakly supervised learning to recognize the complex named entities (commonly composed of multiple small entity sequences, hereinafter referred to as CNEs) in the corpus, which makes it difficult to determine the boundaries of such entities. To improve the recognition accuracy, our method Masked-BiLSTM-CRF is proposed to separate the context semantic relationship determination from the entity boundary confirmation. This method is based on two aspects to solve the above problems: (1) Semantic model based on CNEs mask processing. Before training, the CNEs in the corpus will be masked, and then use the masked corpus training the semantic model through BiLSTM-CRF, which can verify whether the context semantics of the corresponding location entities are correct. (2) A weakly supervised CNEs boundary confirmation model based on sequential patterns. In the small sample data set, the target CNE candidate set is found by sliding window combined with sequence pattern matching, and then it is effectively screened and judged by the semantic understanding model obtained in (1). The experimental results show that compared with the named entity recognition method based directly on BiLSTM-CRF on the weakly-supervised named entity recognition in financial field, our proposed method improves F1-Score in the small data training sample set by nearly 9%, and it has some generalization ability.

Keywords: Named entity recognition, Weakly supervised learning, Deep learning, Pattern matching

1 Introduction

The research on Name Entity Recognition (NER) first appeared in a paper on extracting and identifying company names published by Rau [1] at the 6th IEEE Artificial Intelligence Application Conference held in 1991, which Laid the beginning of NER research. Over time, NER has continued to innovate in technology, from initial rule-based methods to statistical-based machine learning methods, such as hidden Markov models [2], support vector machines [3], conditional random fields (CRF) [4], etc. With the development of deep learning [5-6], NER technology has also ushered in a new wave of development: the addition of word vector technology and neural network technology [7-8], NER technology has been greatly improved in the level of semantic recognition, such as the currently widely used BiLSTM-CRF model [9], which has a huge improvement in processing sequence annotation compared with the previous technology.

With the development of big data [10], block chain [11] and the Internet of Things [12-14], data is becoming more and more accessible, but labeled data is still not easy to obtain. The training of NER model requires a lot of labeled data, which is often difficult to meet in real production: the labeling of large sample data according to quality and quantity requires a lot of manpower, money and time, and even experts are required to participate in some fields. Another difficulty in data collection is that most of the previous research on natural language has focused on data-rich professional fields such as medicine, news, and biology. In recent years, with the birth of more and more pre-trained models such as BERT [15], XLNet [16], and the disclosure of large word vectors such as Google word2vec [17], Tencent AI-Lab Chinese word embedding [18]. Therefore, research on weakly-supervised or even unsupervised NER based on a small amount of labeled corpus in an industry is of great significance at this stage.

Complex named Entity (CNE) in Chinese training corpus for small labeled data, which is generally a composite entity composed of multiple small entity sequences, leads to the problem that the NER model training is difficult to converge and the recognition

*Corresponding Author: Guolang Chen; E-mail: 2019f098@zju.edu.cn
DOI: 10.3966/160792642020122107006

accuracy is reduced. This paper proposes a scheme for named entity recognition based on pattern matching and semantic verification.

The main contributions of our works include the following aspects.

(1) A method suitable for CNEs recognition (Masked- BiLSTM-CRF) is proposed. Compared with the current mainstream entity recognition scheme, our method can effectively determine the entity boundary according to the context semantics when identifying the CNE in the text, and improve the recognition accuracy.

(2) A weakly-supervised iterative training method that optimizes the sample selection process is proposed, named Optimized-Bootstrapping, which uses sample similarity metrics to select more reliable training samples augmentation to effectively control the model jitter problem during the iterative training process.

(3) Various experimental performance tests on the whole proposed approaches for model optimization are individually evaluated. This paper takes CNE such as project name in the financial field data as the research object, and compares the scheme in this paper with the current common NER schemes (CRF and BiLSTM-CRF) in this field. The experimental results show that our proposed scheme has greatly improved the accuracy of CNE recognition, and the proposed method is theoretically generalization ability, which lays the foundation for subsequent research.

The rest of this paper is listed as follows. Section 2 introduces the related works on Chinese named entity recognition, Financial domain named entity recognition, and the weak supervised named entity recognition. Section 3 describes our proposed Masked-BiLSTM-CRF and weakly supervised learning method. Section 4 shows the performance evaluation results from various experimental analysis. Finally, Section 5 concludes this paper.

2 Related Work

2.1 Chinese Named Entity Recognition

In the recognition of named entities in the Chinese domain, the input of a Chinese corpus based on deep learning models can be divided into word vectors and character vectors. If the sentence is converted into word vector, the effect of word segmentation will directly affect the accuracy of the NER model. However, if the word is replaced with character vector, although the word segmentation error is avoided, intrinsic information between words in semantics is lost. For the above problems, Zhang et al. [19] proposed a method of combining the LSTM model based on words and characters, adding a layer of word-based prediction to the output of the characters as the final output of Lattice-LSTM. For noisy data, Yang et al. [20] were inspired by adversarial network learning, and used the BiLSTM model to learn standard labeled data and noisy labeled data respectively, then optimized the learning quality of the module through adversarial learning, which make it converge on real data. Similarly, [21] uses BiLSTM-Attention as the generator model of the generative adversarial network, and CNN as the discriminator model to integrate positive sample labeled data from the crowdsourced labeled data set which is consistent with the distribution of expert labeled data, to solve the problem of lack of labeled data in the field. [22] proposed a chemical drug named entity recognition method based on attention mechanism. This method first learns word vectors from massive biological texts, then uses BiLSTM model to learn character vectors, and then uses word vectors and character vectors as another input of BiLSTM model and combines the attention mechanism to obtain the context representation of words in the full text, so as to improve the consistency of entity recognition.

2.2 Financial Domain Named Entity Recognition

Compared with news, medicine and other fields, the application of named entity recognition in the financial field has attracted much attention, but there are few published researches at present. The reason is that the text data disclosed in the financial field is relatively less, the labeled data is rare, and the process of labeled data is time-consuming. Wang et al. [23] proposed a new method for identifying named entities in financial news texts. This method first combines domain dictionaries and conditional random fields to identify financial entities (such as stock names, etc.), and then uses mutual information, boundary entropy, and context features to identify abbreviated financial entities. It achieves 91% recognition accuracy on Chinese financial data sets. In [24], the training process was deployed on Hadoop framework by combining conditional random fields with collaborative training, which reduced the training time while satisfying the recognition accuracy. [25] proposed a method of combining entity relationship extraction and entity recognition. By introducing rules and grammatical feature extraction to obtain the entities relationship, and using the obtained relationship to assist the recognition of financial entities to improve the recognition accuracy. From the above researches, it can be found that the entities identified in the financial field still concentrate on some simple entities.

2.3 Weak Supervised Named Entity Recognition

Weakly supervised learning [26] is a commonly used machine learning method in the case of insufficient labeled data. First, the labeled data is learned to make the model obtain a certain discrimination ability. On this basis, the unlabeled data is analyzed, utilized, and then performed iterative learning reinforcement.

Riloff and Jones [27] proposed a mutual bootstrap algorithm, which starts with a given type of artificial seed entities to find out the features and patterns in the large corpus where these entities are located. After that, find new entities through the relationship characteristics and patterns of these contexts. Cucchiarelli and Velardi [28] proposed a method of named entity recognition based on grammatical analysis, which discovers more and more accurate entity context relationships through grammatical structure. Tsuboi et al. [29] proposed a method of training conditional random field model by using incomplete labeled data. This method imitates the way of modeling a correct path, modeling all possible paths in incomplete labeled data, so as to achieve CRF model with incomplete labeled data. Duan Chaoqun [30] obtained word vectors in specific fields through RNNLM neural network training in NER research for lack of labeled data, using conditional random fields with real-valued features. He used iterative training for corpus with only a small amount of labeled data under the bootstrapping learning framework. Compared with the use of public word2vec word vectors, it has a significant improvement in human name recognition. Zafarian et al. [31] used an unlabeled bilingual corpus to extract useful features, and used these features and a small amount of training data to convert NER supervised models. Then they used the graph-based weakly supervised learning method to train a CRF-based supervised classifier, and expanded the training set by filtering out high-confidence prediction data. Aryoyudanta et al. [32] used labeled and unlabeled data for the Indonesian language and used a collaborative training method to obtain a named entity classifier based on support vector machines. The classifier was used to annotate new data and the results showed that F1 values reached 76.5%. Xu et al. [33] proposed that a sufficient corpus and a large number of unlabeled texts should be combined in the formal domain to improve recognition performance in the study of Chinese social media NER. To build a Mandarin NER module, [34] used data augmentation to expand the data set, and proposed a hybrid NER method that combines established rules and sentence patterns to successfully identify strange words and perform well in disaster management systems.

3 Methodology

The corpus data in this paper is a contract announcement document in the financial field. The project entity included is the CNE mentioned above, which consists of several small entities, for example:

Text sentence: “甲方和乙方经过三个月的协商，终于签署了关于成都天府国际空港新城公共服务配套总承包项目的合同书，此项目签约金额为 100000 元人民币.”

CNE: “成都天府国际空港新城公共服务配套总承包项目.”

The analysis of the text corpus shows that the CNEs include the following distribution features: (1) Although the CNEs in the text vary widely in text composition, its contextual semantic relationships are rare, basically focusing on project signing agreements and project winning information, etc.; (2) Although the CNEs are different, their composition patterns are rule-based. For example, the composition model of the CNE above is “region + industry + type”. Although character sequences are quite different, there are some similarities between part-of-speech sequence and word vector sequence.

Combining the above features, this paper uses the method of separating context semantic analysis and CNEs boundary determination to identify CNE in the text, that is, first using pattern matching to find the target CNE candidate set in the text, and then masking CNE to named entity, using the context semantic understanding model to determine whether the context semantic relationship of the entity is correct. The entire process is shown in Figure 1.

3.1 Text Pre-processing and Context Semantic Understanding Model Training

3.1.1 Word Segmentation

In this paper, we use Jieba word segmentation tool and select its search engine word segmentation mode to segment the text, and then counts the results of all text segmentation to obtain the frequency weight of each word. For example, the frequency of “目的” appears in all texts is 134. The frequency of the word “项目” is 1789. Then we use the word frequency weight table as the word segmentation weight to segment the words accurately. For example, in the sentence “保证该项目的质量”, because the word frequency weight of “项目” is higher than that of “目的”, the “项目” will be preferentially divided into one word, which can reduce the segmentation errors of common words.

3.1.2 Word Vector Representation

After effective word segmentation, how to vectorize the word is a problem that ensues. Word vectors add some contextual semantic information to the original discrete phrases, and similar words are also similar in vector space. At present, the commonly used word vector training methods mainly include two neural network language models, skip-gram [35] and CBOW [36] provided by Word2vec. The final training results of these two models are more seriously affected by the training corpus. In the case of insufficient or poor corpus, the originally semantically similar phrases may be far apart in the vector space, and it is impossible to

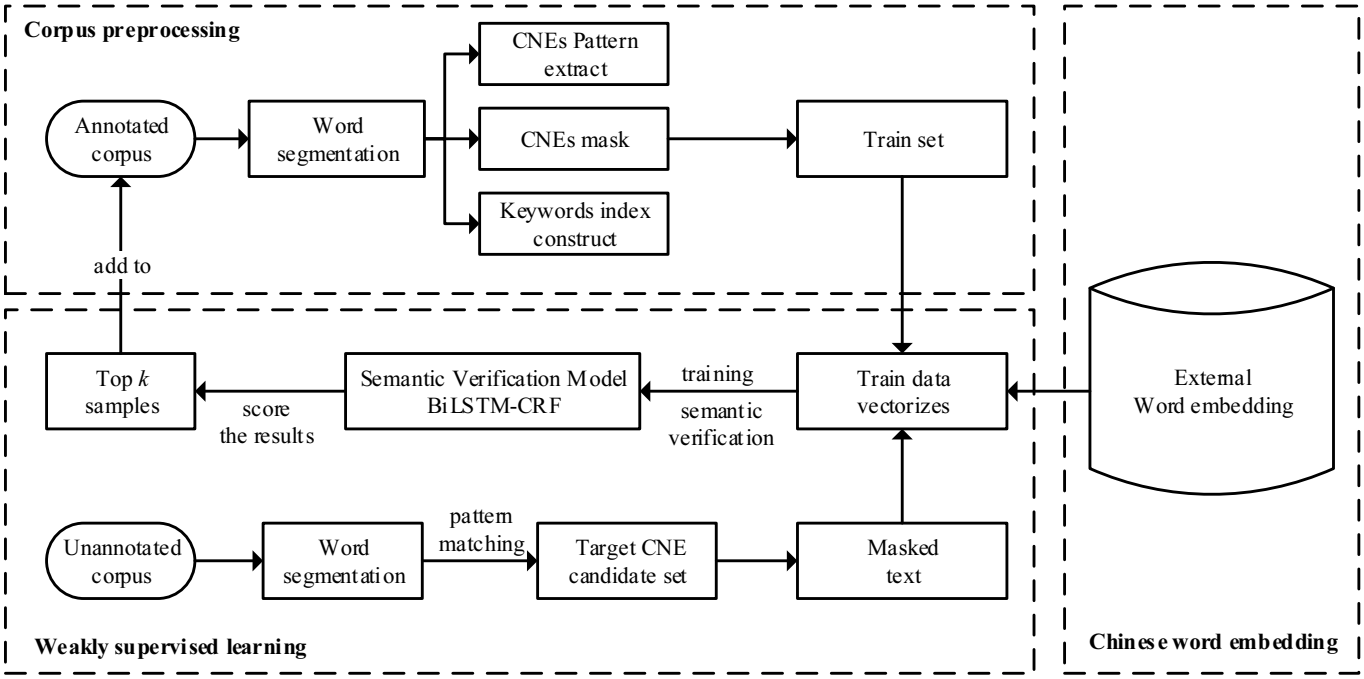


Figure 1. Weakly supervised complex NER framework

generate a new word vector representation separately. Considering the limitation of the amount and quality of text data, this paper does not adopt the method of training word vectors, but chooses the large Chinese word vector pre-training model (Chinese word embedding) provided by AI-Lab as the word vector representation of word segmentation. It provides more than 8 million words vectors, covering various fields, and the distribution of word vectors is relatively reasonable. Each word consists of a 200-dimensional vector, which can be expressed as: V= $\{D_1, D_2, \ldots, D_{200}\}$, where $D_i(i=1, 2,\ldots 200)\cdot(-1, 1)$.

3.1.3 CNEs Mask and Patterns Extract

The existence of CNEs makes the sentence structure more complicated, which will lead to the introduction of some interference features that affect the accuracy of the model when extracting text sequence features, and reduce the generalization ability of the model.

Based on the above considerations, this paper adopts the method of masking CNE as named entities to simplify the text structure (this process is hereinafter referred to as mask). The simplified text can make the model focus on the extraction of the semantic features of the entity context during training, thereby improving the semantic understanding of the model. In this paper, when training the context semantic understanding model, the CNE in the text is uniformly masked as the word "项目", and the "项目" there is marked as an entity (B-project). Because the term "项目" originally existed in the text in large amounts, these non-entity "项目" solved the problem of data imbalance in model training, that is, the model would extract features based on the contextual semantic relationship of "项目", and it is not recognized that "项目" is an entity.

The pattern extraction process of CNE accompanied by the above masking process, that is, when each CNE is masked in the labeled data, its pattern sequence features are extracted. The pattern sequence mainly includes two parts: noun sequence and part-of-speech sequence. Noun sequences can reflect the focus of entity, such as the "region + industry + type" mentioned above, while part-of-speech sequences can reflect their compositional characteristics from the level of syntactic dependence. The structure mode has many forms according to the actual situation. The entity pattern extraction process is shown in Figure 2.

3.1.4 Train Context Semantic Understanding Model

After obtaining simplified text data based on the above steps, we use the popular BiLSTM-CRF model to train a context semantic understanding model in this step. Its structure is mainly divided into two layers, a BiLSTM network layer and a CRF layer.

The structure of the entire network is shown in Figure 3. The CRF layer is mainly responsible for transferring features, it will consider the order between the output label sequences, such as to avoid outputting label sequences such as {I, I, B} that violate the objective sequence of the label sequence. For the input sequence X = $\{x_1, x_2, \ldots, x_n\}$, it can be predicted that the output label sequence Y=$\{y_1, y_2, \ldots, y_n\}$ scores are as follows.

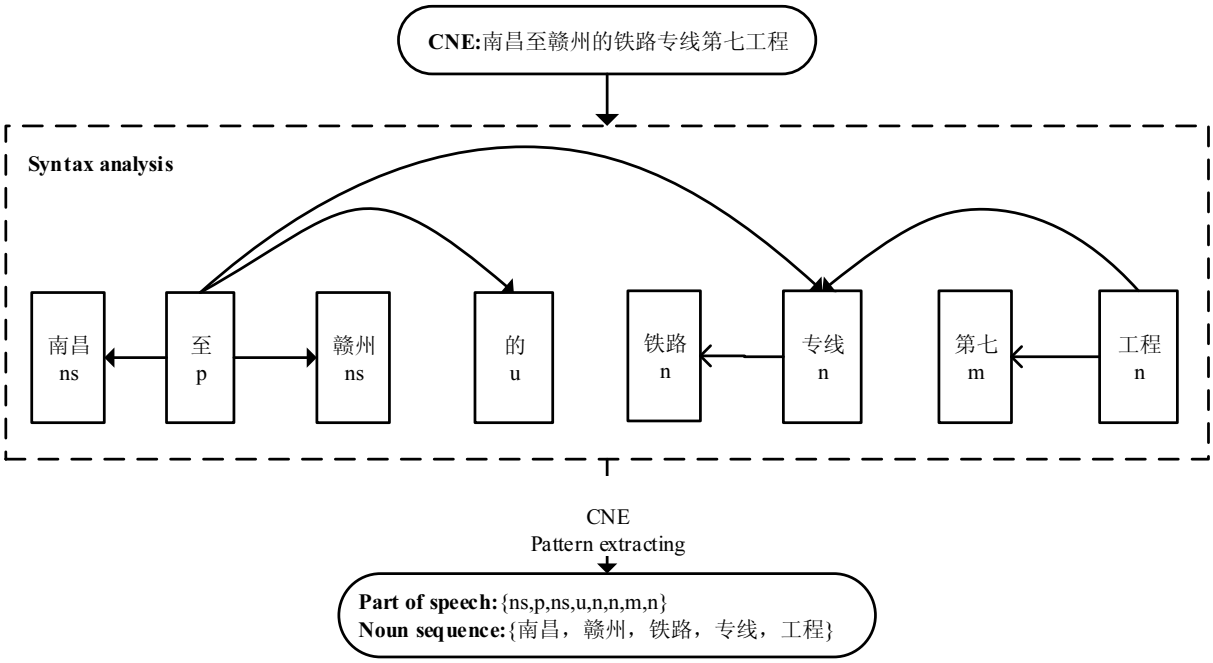


Figure 2. The diagram of complex entity pattern extract

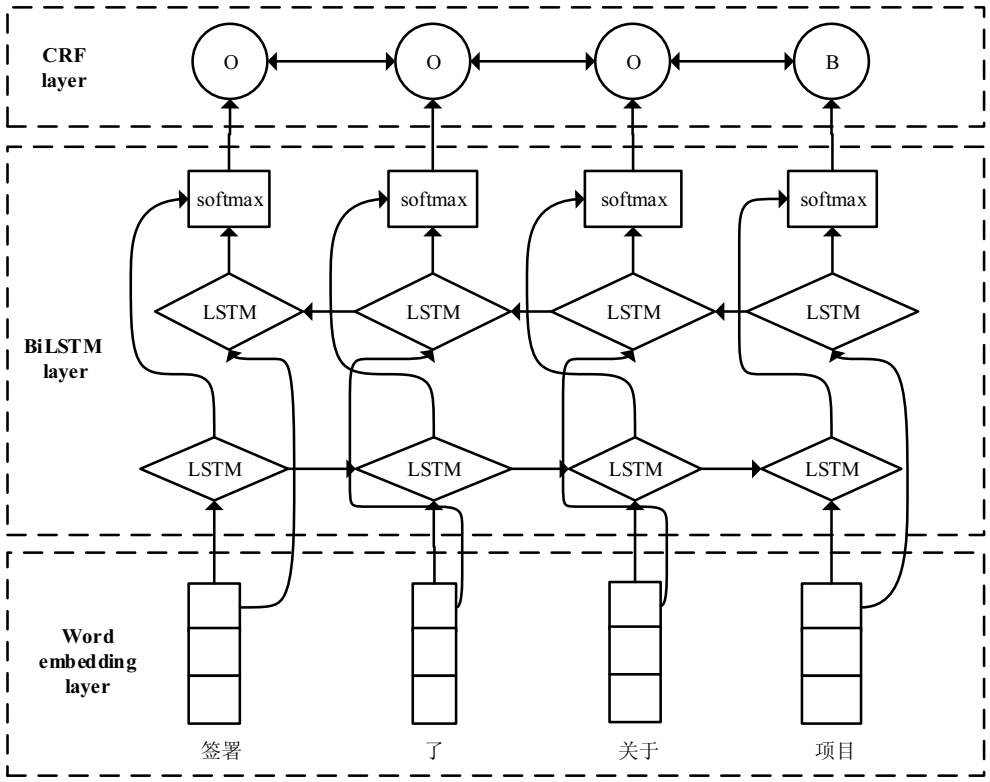


Figure 3. The structure of BiLSTM-CRF

$$s(X,Y)=\sum_{i=0}^{n} A_{y_i,y_{i+1}} + \sum_{i=1}^{n} P_{i,y_i} \quad (1)$$

In formula (1), P_{i,y_i} is the probability that the softmax output at the ith position is y_i, and the transition probability from y_i to y_{i+1}.

3.2 CNE Recognition and Weakly Supervised Learning

Through the above processing steps, the CNE pattern set EP_{set} and a context semantic understanding model M can be obtained. In the identification of CNEs, the effective determination of entity boundaries is a difficult problem. In this paper, a method of

combining sliding windows and pattern matching is used to preliminarily select the target CNE candidate set through EP_{set}, and then use model *M* to verify whether the context semantics of the target CNE is correct. The target CNE recognition process is shown in Figure 4.

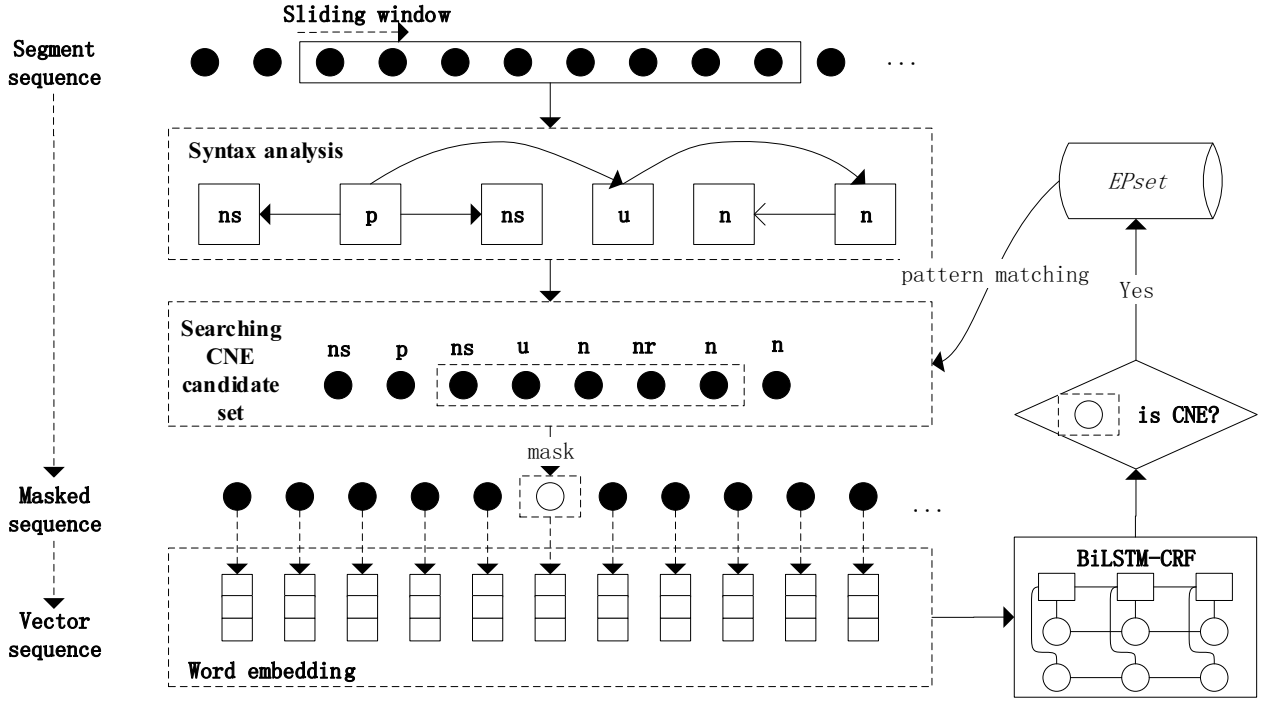


Figure 4. The diagram of target CNE recognition

3.2.1 Efficient Search Keyword Index Construction

In order to improve the efficiency of pattern matching, this paper adopts a high-dimensional spatial index technology to establish a keyword vector index for each word in EP_{set}, through which we can quickly find out the top N words closest to the input word vector in time complexity of O(logn). And then locate the corresponding pattern accord to these words. When recognizing CNEs in unlabeled data, we need segment the corpus at first to get a sequence $S=\{(X_1, T_1), (X_2, T_2)\ldots(X_n, T_n)\}$, where X_i(i=1, 2,…n) represents the word segmentation and T_i(i=1, 2,…n) represents the part of speech. After that, set a window of length L, $W = \{(X_{m+1}, T_{m+1}), (X_{m+2}, T_{m+2}), \ldots(X_{m+L}, T_{m+L})\}$ on this sequence, and then search the target CNEs by executing the pattern matching algorithm in the window. The index structure is shown in Figure 5.

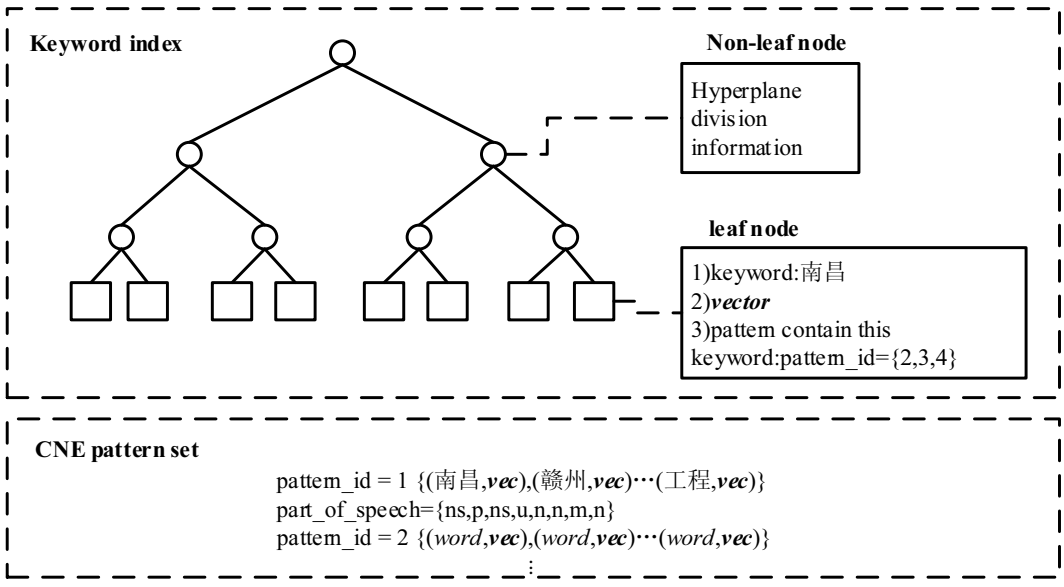


Figure 5. The structure of keyword Index diagram

The distance D between the vectors is calculated using the normalized Euclidean distance:

$$D_{u,v} = \sqrt{2 - 2 \times \cos(u, v)} \quad (2)$$

3.2.2 Target CNEs Searching

In this stage, the pattern matching algorithm will be executed in the sliding window to search for the target CNEs combining with the above index structure, and finally the target CNE candidate set is obtained. As the CNEs pattern extraction stage records their noun sequence feature, the algorithm also focuses on the noun sequence feature during the matching process. The matching process begins by traversing all the words in the sliding window. For each noun, find the N most similar words and their corresponding CNE pattern sequences from the keyword index, and then traverse these pattern sequences, if the head and tail are successfully matched in a pattern, this part is marked as the target CNE. The specific process is shown in Algorithm 1.

Algorithm 1. Pattern matching search target CNEs algorithm.

Input: sliding window word sequence W = {(X_{m+1}, T_{m+1}), (X_{m+2}, T_{m+2}),…(X_{m+L}, T_{m+L})};
output: target CNE candidate set TE_{set}.

```
TE_set ={}
For X_i in W [0 to L] & T_i=noun:
  <keys, patterns>=findTopKeyInIndex(X_i) /* find top N similar words and patterns contain them */
    For each key in keys:
      If similarity (key, X_i) >α:
        Entity_begin = X_i
      If Entity_begin exists:
        For each word in reverse (pattern):
          For X_j in W [L to i] & T_j =noun:
            If similarity (word, X_j) >α:
              Entity_end = X_j
      If Entity_begin & Entity_end all exist:
        Put W [i, j] to TE_set
Return TE_set
```

In the above algorithm, Similarity (X, Y) is the calculation function for words similarity measure. X、Y are n-dimensional word vectors. X={x_1, x_2, …, x_n}, Y={y_1, y_2, …, y_n}, and the function is expressed as follows:

$$Similarity(X, Y) = \frac{\sum_{i=0}^{n}(x_i \times y_i)}{\sqrt{\sum_{i=0}^{n}(x_i)^2} \times \sqrt{\sum_{i=0}^{n}(y_i)^2}} \quad (3)$$

The algorithm determines whether two nouns are related based on the similarity threshold α, above it indicates a successful match. When there is no target CNE in the sliding window, just move it right to the next noun.

3.2.3 Verify Target CNEs

The algorithm determines whether two nouns are related based on the similarity threshold α, above it indicates a successful match. When there is no target CNE in the sliding window, just move it right to the next noun.

In the above steps, the target CNE candidate set is obtained, but the final determination of the CNEs need to be verified by the context semantic understanding model. In this step, we traverse the TE_{set} to mask the target CNEs into short entities, and then use the context semantic model M (we trained before) to confirm. According to the above pattern matching algorithm, relatively longer target CNEs stored in TE_{set} can be ensured first, so in the process of traversal verification, the longer target CNEs will also be verified first. The entire verification algorithm is shown in Algorithm 2.

Algorithm 2. Verify target CNEs algorithm.

Input: word segmentation sequence S = {X_1, X_2, …, X_n}, Model *M*, target CNE candidate TE_{set};
output: verified CNEs.

```
Masked_S_set = {S}
For TE_i in TE_set [0 to n]:
  C={}
  For TE_j in TE_set [0 to i]:
    If TE_i ∩ TE_j != Ø: /* Mutually exclusive CNEs*/
      Put TE_j to C /* record these CNEs*/
  For each ms in Masked_S_set:
    New_MS_set={}
    If ms.maksed_entities ∩ C = Ø:
      new_ms=ms.mask (TE_i) /* Mask the current CNE to form a new statement */
      Put new_ms to New_MSset
  Masked_Sset = Masked_S_set ∪ New_MS_set
For ms in Masked_S_set:
  Input_S=ms.trans2embedding (Chinese word embedding) /* Convert to word vector */
  Label_S=M.predict (Input_S) /* Using model M to predict the label sequence */
  If all Label_S [mask_index] ==‘B-project’: /
    Return ms. maksed_entities
Return null /* there is no CNE*/
```

3.2.4 Weakly Supervised Iterative Learning Based on Scoring Function

Another part of the work in this paper is weakly supervised iterative optimization. This process introduces the scoring function to optimize the

selection of incremental samples during the iterative process based on bootstrapping. Given a small amount of initial labeled corpus set L and a large number of unlabeled corpus set U, the optimized bootstrapping process is as follows:

(a) Extract the CNE patterns in the labeled data set L according to the Section 3.1.3;

(b) Mask the CNEs in the set L into short entities to simplify the corpus, and then use the simplified corpus to train the context semantic understanding model M;

(c) Identify the CNEs in the unlabeled data set U according to the Section 3.2.1 and 3.2.2;

(d) Score all recognition results according to the scoring function and sort the corpus according to the score;

(e) The top K corpus after sorting are relatively reliable incremental samples in this iteration identification;

(f) Label this part of the incremental samples and add it to the set L, and update the TE_{set}, and the keyword high-dimension index in the Section 3.2.1;

(g) Repeat (b) ~ (f) until U is empty.

The flow chart of this process is shown in Figure 6.

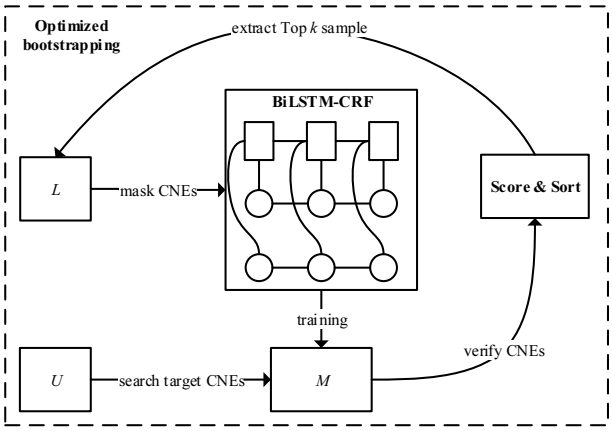


Figure 6. The weakly supervised NER diagram

The misjudgment of incremental samples will lead to the introduction of wrong information in the model training process, and the error will even gradually enlarge with the iterative process. Therefore, in the process of weakly supervised iterative training, minimizing the introduction of erroneous samples is the key to improving the accuracy of the final model. Due to the particularity of CNE, this paper adopts a scoring function (formula 5) based on the similarity measure between samples in weakly supervised learning to select incremental samples with higher reliability. Because the CNEs determined in the incremental samples are obtained from the labeled corpus through pattern matching. Therefore, if the incremental sample t1 determines that there is a CNE e1, and the original matched labeled corpus is t2 and contain e2, then the judgment reliability of the sample can be measured from two aspects: (1) the similarity φ_1 of the sequence composition of CNE e1 and e2; (2) the similarity φ_2 of the sequence composition of text t1 and text t2.

The higher the value φ_1 of φ_2 and, the more similar the sub-sample is to the original labeled text in terms of text structure, and the more reliable the decision. The text sequence is calculated using Levenshtein string editing distance:

$$Lev_{a,b}(i,j)=\begin{cases}\max(i,j) & if\ \min(i,j)=0\\ \min\begin{cases}Lev_{a,b}(i-1,j)+1\\ Lev_{a,b}(i,j-1)+1\\ Lev_{a,b}(i-1,j-1)+1\ \ (a_i\neq b_i)\end{cases} & otherwise\end{cases} \quad \textbf{(4)}$$

In formula (4), a and b represent the sequence part-of-speech symbol array, i and j represent the array subscript. After calculating φ_1 and φ_2 , they are weighted and normalized, and the final scoring function expression is:

$$Score(\varphi_1,\varphi_2)=\frac{1}{\omega_1\times e^{-\varphi_1}+\ \omega_2\times e^{-\varphi_2}} \quad \textbf{(5)}$$

In formula (5), ω_1 and ω_2 are weights, and the default value is 1. The higher the Score value, the more reliable the sub-sample decision.

4 Experimental Results

The data we used in this paper comes from the contract documents in the financial field, which manually marked the entities such as Party A and Party B, the name of the signed project, the name of the contract, etc. The labels are O, B-jiafang, I-jiafang, B-yifang, I- yifang, B-xiangmu, I-xiangmu, B-hetong, I-hetong, where the project entities are the CNEs. There are a total of 3,000 documents in the corpus data set. After data extraction, a total of 11,718 long sentences are collected. During the experiment, it is randomly divided into two parts, one of which is trained and the other is test data. The distribution of data is mainly shown in Table 1.

Table 1. Data set

Data	Number of sentences	Number of entities
Train set	10000	4789
Test set	1718	662

4.1 Experimental Comparison Settings

The experimental part of this paper mainly compares the Masked-BiLSTM-CRF proposed in this paper with the currently popular CRF model based on word annotation, and the BiLSTM-CRF model based on character and word annotation (hereinafter referred to as Character-BiLSTM- CRF and Word-BiLSTM-CRF), BiLSTM-CRF model based on the word Bert pre-

training vector (hereinafter referred to as Bert-BiLSTM-CRF). The comparison experiment is mainly divided into three parts: (1) the recognition effect of the CNEs in the fully supervised mode; (2) the initial marked data amount is 3000, and the weak entity recognition effect of the final model is obtained through iterative training of weak supervision; (3) The influence of the key parameter setting in the Masked-BiLSTM-CRF on the final recognition effect. We use the BIO annotation format to label the text, and the accuracy, Precision, Recall and F1-Score values are used as the evaluation criteria of the model's performance.

4.2 Experimental Parameter Setting

The corpus data used in the experiment is segmented by periods, and the truncation length of the sentence length after word segmentation is 50. If it is less than 50, it is filled with leading zeros when converted into word vectors. The deep learning model framework uses BiLSTM in keras. The parameter settings are shown in Table 2.

Table 2. Parameter settings

Parameter	Value
Number of word	21937
Dimension of word vector	200
Number of LSTM unit	100
Optimizer	Adam
Loss function	CRF_loss_function
Dropout	0.2
Batch_size	16
Epoch	10
Validation_split	0.15
α	0.6

In the above parameter setting, for the word similarity threshold α, after many tests, when the domain between words in the AI-Lab word vector is irrelevant, $\alpha < 0.6$, so 0.6 is used as the cut-off point in this experiment.

4.3 Comparative Experiment

(1) In the Chinese corpus, the input of BiLSTM-CRF can be divided into two types: character vector and word vector. In this paper, two training models, CRF and BiLSTM-CRF, are used to compare their effects under different inputs under fully supervised learning. The model convergence comparison is shown in Figure 7.

It can be found from Figure 7 that Masked-BiLSTM- CRF can converge faster, which shows that it can better extract the text features after the mask.

(2) In the fully supervised learning mode, the method proposed in this paper compares the experimental results with the more popular methods as shown in Table 3 below.

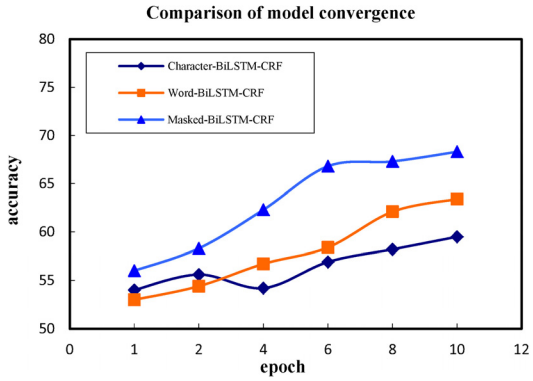


Figure 7. Comparison of model convergence diagram

Table 3. Fully supervised learning experiment results

Model	Precision	Recall	F1-Score
CRF	45.2	42.7	43.9
Character-BiLSTM-CRF	47.9	65.3	55.2
Bert-BiLSTM-CRF	44.3	58.6	50.4
Word-BiLSTM-CRF	55.6	64.4	59.7
Masked-BiLSTM-CRF	**64.6**	**67.2**	**65.8**

From the results in Table 3, it can be seen that Masked-BiLSTM-CRF has greatly improved the accuracy of entity recognition compared to other methods, mainly reflected in the determination of the boundary of the entity.

(3) In the weakly supervised learning mode, the experimental results of this method are compared with the mainstream methods. The initial data is set to 3,000 marked data as shown in Table 4.

Table 4. Weakly supervised learning experiment results

Model	Precision	Recall	F1-Score
CRF	41.2	38.9	40.0
Character-BiLSTM-CRF	43.8	53.6	48.2
Word-BiLSTM-CRF	42.5	58.7	49.3
Masked-BiLSTM-CRF	**60.5**	**57.4**	**58.9**

(4) Masked-BiLSTM-CRF different parameter settings affect the final result. The effect of window length setting on the recognition effect is shown in Figure 8; the effect of weakly supervised initial data amount seed and incremental sample number k on the result is shown in Figure 9.

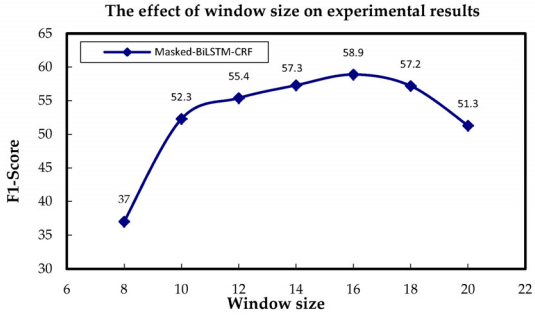


Figure 8. The effect of window length on experimental

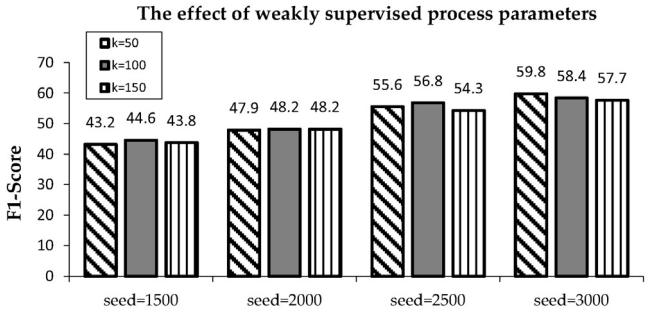


Figure 9. The effect of seed and k on experimental results

5 Conclusions

The difference between Masked-BiLSTM-CRF proposed in this paper and other named entity recognition methods such as CRF and BiLSTM is that the proposed method divides the recognition of named entities into two processes, the process of matching entity search based on pattern features and the verification based on context semantic relationship, so the proposed method retains the high generalization of the model while the features are controllable. Although we take the financial field data as the entry point of the research, the method proposed in this paper is not limited to this field. Any other field data that contains CNEs can theoretically use this method to identify them. If a researcher wants to transfer this method to other fields (such as biology, biological text corpus also contains a large number of CNEs), first need to provide a more reasonable distribution of word vectors in the corresponding field, and then mark a small amount of corpus and data analysis to determine the size of the sliding window can mark the entire data set in the form of weak supervision.

The proposed method has the following shortcomings: (1) the time complexity of the pattern matching process is high; (2) in the weakly supervised training process, the typical selection of CNE patterns in the seed data, that is, the effective selection of seed data; (3) The effect of the setting of the window size on the recognition effect, at this stage, it is not possible to recognize CNEs whose length exceeds the window. In view of the above problems, in future research, we can proceed from two aspects: (1) hierarchical pattern categories by clustering entity patterns to reduce matching time and extend the coverage of CNEs patterns of seed data; (2) sliding window Introduce a dynamic adjustment mechanism in the setting of, that is, the size of the window should be dynamically adjusted as the text sequence changes, so as to reduce the search range of the entity while avoiding the problem that the entity length exceeds the window and cannot be identified.

Acknowledgments

This research was sponsored by Major Scientific Research Projects of Wenzhou Polytechnics (No. WZY2020001), funded by Wenzhou Scientific Research Projects for Underdeveloped Areas (WenRenSheFa [2020] 61 (No. 5)), sponsored by the First Batch of Teaching Reform Research Projects in the 13th Five Year Plan of Higher Education in Zhejiang Province (No. jg20180585), supported by Zhejiang Province "the 13th Five-Year Plan" for Collaborative Education Project between Industry and University.(The E-Commerce "1+X" Curriculum Design based on Univerisity and Industry Co-cultivation).

References

[1] L. F. Rau, Extracting Company Names from Text, *The Seventh IEEE Conference on Artificial Intelligence Application*, Miami Beach, FL, USA, 1991, pp. 29-32.

[2] L. R. Rabiner, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, *Proceedings of the IEEE*, Vol. 77, No. 2, pp. 257-286, February, 1989.

[3] T. Joachims, Making Large-scale Support Vector Machine Learning Practical, in: B. Scholkopf, C. J. C. Burges, A. J. Smola (Eds.), *Advances in Kernel Methods*, MIT Press, 1999, pp. 169-184.

[4] J. Lafferty, A. Mccallum, F. Pereira, Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, *18th International Conference on Machine Learning*, Williamstown, MA, USA, 2001, pp. 282-289.

[5] S. He, Z. Li, Y. Tang, Z. Liao, F. Li, S. Lim, Parameters Compressing in Deep Learning, *Computers, Materials and Continua*, Vol. 62, No. 1, pp. 321-336, 2020. DOI: 10.32604/cmc.2020.06130.

[6] J. Zhang, W. Wang, C. Lu, J. Wang, A. K. Sangaiah, Lightweight Deep Network for Traffic Sign Classification, *Annals of Telecommunications*, Vol. 75, No.3, July, 2019. DOI: 10.1007/s12243-019-00731-9.

[7] J. Zhang, X. Jin, J. Sun, J. Wang, A. K. Sangaiah, Spatial and Semantic Convolutional Features for Robust Visual Object Tracking, *Multimedia Tools and Applications*, Vol. 79, No. 21-22, pp. 15095-15115, June, 2020.

[8] J. Zhang, Z. Xie, J. Sun, X. Zou, J. Wang, A Cascaded R-CNN with Multiscale Attention and Imbalanced Samples for Traffic Sign Detection, *IEEE Access*, Vol. 8, pp. 29742-29754, February, 2020.

[9] W. Ma, H. Yu, K. Zhao, D. Zhao, J. Yang, J. Ma, Tibetan Location Name Recognition Based on BiLSTM-CRF Model, *International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, Guilin, China, 2019, pp. 412-416.

[10] J. Wang, Y. Yang, T. Wang, R. S. Sherratt, J. Zhang, Big Data Service Architecture: A Survey, *Journal of Internet Technology*, Vol. 21, No. 2, pp. 393-405, March, 2020.

[11] J. Zhang, S. Zhong, T. Wang, H. Chao, J. Wang, Blockchain-Based Systems and Applications: A Survey, *Journal of Internet Technology*, Vol. 21, No. 1, pp. 1-14, January, 2020.

[12] J. Wang, Y. Gao, C. Zhou, R. S. Sherratt, L. Wang, Optimal Coverage Multi-Path Scheduling Scheme with Multiple Mobile Sinks for WSNs, *Computers, Materials & Continua*, Vol. 62, No. 2, pp. 695-711, 2020.

[13] J. Wang, X. Gu, W. Liu, A. K. Sangaiah, H.-J. Kim, An Empower Hamilton Loop based Data Collection Algorithm with Mobile Agent for WSNs, *Human-centric Computing and Information Sciences*, Vol. 9, Article number: 18, May, 2019.

[14] J. Wang, Y. Tang, S. He, C. Zhao, P. K. Sharma, O. Alfarraj, A. Tolba, LogEvent2vec: LogEvent-to-Vector Based Anomaly Detection for Large-Scale Logs in Internet of Things, *Sensors*, Vol. 20, No. 9, 2451, May, 2020.

[15] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, Minneapolis, MN, USA, 2019, pp. 4171-4186.

[16] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, Q. V. Le, XLNet: Generalized Autoregressive Pretraining for Language Understanding, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems*, Vancouver, Canada, 2019, pp. 5754-5764.

[17] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space, *1st International Conference on Learning Representations*, Scottsdale, Arizona, USA, 2013, pp. 1-12.

[18] Y. Song, S. Shi, J. Li, H. Zhang, Directional Skip-Gram: Explicitly Distinguishing Left and Right Context for Word Embeddings, *2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers) (HLT-NAACL)*, New Orleans, Louisiana, USA, 2018, pp. 175-180.

[19] Y. Zhang, J. Yang, Chinese NER Using Lattice LSTM, *56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia, 2018, pp. 1554-1564.

[20] Y. S. Yang, M. Zhang, W. Chen, W. Zhang, H. Wang, M. Zhang, Adversarial Learning for Chinese NER from Crowd Annotations, *32th AAAI Conference on Artificial Intelligence*, New Orleans, Louisiana, USA, 2018, pp. 1627-1635.

[21] H. Zhang, Y. B. Guo, T. Li, Domain Named Entity Recognition Combining GAN and BiLSTM-Attention-CRF, *Journal of Computer Research and Development*, Vol. 56, No. 9, pp. 1851-1858, September, 2019.

[22] P. Yang, Z. H. Yang, L. Luo, H. Lin, J. Wang, An Attention-Based Approach for Chemical Compound and Drug Named Entity Recognition, *Journal of Computer Research and Development*, Vol. 55, No. 7, pp. 1548-1556, July, 2018.

[23] S. Wang, R. Xu, B. Liu, L. Gui, Y. Zhou, Financial Named Entity Recognition Based on Conditional Random Fields and Information Entropy, *Machine Learning and Cybernetics*, Lanzhou, China, 2014, pp. 838-843.

[24] Y. Wu, The Method Study and System Implementation of Named Entity Recognition in the Finicial Field, M.S. Thesis, Harbin Institute of Technology, Harbin, China, 2015.

[25] E. T. Khaing, M. M. Thein, M. M. Lwin, Stock Trend Extraction Using Rule-based and Syntactic Feature-based Relationships between Named Entities, *Advanced Information Technologies*, Yangon, Myanmar, 2019, pp. 78-83.

[26] Z. H. Zhou, A Brief Introduction to Weakly Supervised Learning, *National Science Review*, Vol. 5, No. 1, pp. 44-53, January, 2018.

[27] E. Riloff, R. Jones, Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping, *16th National Conference on Artificial Intelligence*, Orlando, Florida, USA, 1999, pp. 474-479.

[28] A. Cucchiarelli, P. Velardi, Unsupervised Named Entity Recognition Using Syntactic and Semantic Contextual Evidence, *Computational Linguistics*, Vol. 27, No. 1, pp. 123-131, March, 2001.

[29] Y. Tsuboi, H. Kashima, S. Mori, H. Oda, Y. Matsumoto, Training Conditional Random Fields Using Incomplete Annotations, *International Conference on Computational Linguistics*, Manchester, UK, 2008, pp. 897-904.

[30] C. Q. Duan, *Research on the Named Entity Recognition in the Domain of Lack of Annotated Data*, M.S. Thesis, Harbin Institute of Technology, Harbin, China, 2015.

[31] A. Zafarian, A. Rokni, S. Khadivi, S. Ghiasifard, Semi-supervised Learning for Named Entity Recognition Using Weakly Labeled Training Data, *International Symposium on Artificial Intelligence and Signal Processing*, Mashhad, Iran, 2015, pp. 129-135.

[32] B. Aryoyudanta, T. B. Adji, I. Hidayah, Semi-supervised Learning Approach for Indonesian Named Entity Recognition (NER) Using Co-training Algorithm, *International Seminar on Intelligent Technology and Its Applications*, Lombok, Indonesia, 2017, pp. 7-12.

[33] J. J. Xu, H. F. He, X. Sun, X. C. Ren, S. J. Li, Cross-Domain and SemiSupervised Named Entity Recognition in Chinese Social Media: A Unified Model, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 26, No. 11, pp. 2142-2152, November, 2018.

[34] H. Kung, C. Hsieh, C. Ho, Y. Tsai, H. Chan, M. Tsai, Data-Augmented Hybrid Named Entity Recognition for Disaster Management by Transfer Learning, *Applied Sciences*, Vol. 10, No. 12, pp. 1-17, June, 2020.

[35] T. Mikolov, W. Yih, G. Zweig, Linguistic Regularities in Continuous Space Word Representations, *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics*, Atlanta, Georgia, USA, 2013, pp. 746-751.

[36] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed Representations of Words and Phrases and their Compositionality, *27th Annual Conference on Neural Information Processing Systems*, Lake Tahoe, Nevada, USA, 2013, pp. 3111-3119.

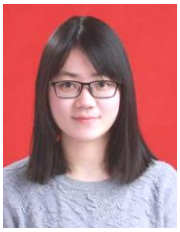
Biographies

Nan Gao received the M.S. degree in safety technology and engineering from Tsinghua University in 2008, and the Ph.D. degree in computer science from the University of South Carolina in 2014. She is currently an Associate Professor with the College of Computer Science and Technology, Zhejiang University of Technology. Her research interests include algorithm optimization, data mining, and artificial intelligence.

Zhenyang Zhu received the B.S. degree in computer science and technology from the Zhejiang University of Technology, Hangzhou China, in July 2017. He is currently pursuing the M.Eng. degree in Software Engineering with the Zhejiang University of Technology. His current research interest includes the NLP.

Zhengqiu Weng received the M.S. degree in software engineering from the Beijing Institute of Technology, Beijing, in 2005. She is currently pursuing the Ph.D. degree in computer science and technology with the Zhejiang University of Technology, and also an Associate Professor with Wenzhou Polytechnic, Zhejiang, China. Her current research interests include networks security and big data technologies.

Guolang Chen received the M.S degree in Computer technology from Hangzhou Dianzi University, Hangzhou, China, in 2007. He is an associate professor at Wenzhou Polytechnic, Wenzhou, China. He is also an Visiting scholar of Zhejiang University (Hangzhou, China). His current research interests include Network technology and E-commerce.



Min Zhang is currently pursuing the M. Eng. degree in computer science and technology with the Zhejiang University of Technology. Her current research interest includes the next generation wireless LAN and heterogeneous network.