**RESEARCH ARTICLE**

# A Research Toward Chinese Named Entity Recognition Based on Transfer Learning

Hui Kang[1,2] · Jingwu Xiao[3] · Yunpeng Zhang[3] · Lei Zhang[3] · Xu Zhao[3] · Tie Feng[1]

**Abstract**
To improve the performance of named entity recognition in the lack of well-annotated entity data, a transfer learning-based Chinese named entity recognition model is proposed in this paper. The specific tasks are as follows: (1) first/, a data transfer method based on entity features is proposed. By calculating the similarity of feature distribution between low resource data and high resource data, the most representative entity features are selected for feature transfer mapping, and the distance of entity distribution between the two domains is calculated to make up the gap between the data of the two domains then model is trained by high resource data. (2) Then, an entity boundary detection method is proposed. This method utilizes the BiLSTM+CRF as the main structure and integrates character boundary information to assist the attention network to improve the model's ability to recognize entity boundaries. (3) Finally, multiple named entity recognition methods are selected as baseline methods for comparison, and experiments are conducted on several datasets. The results show that the model proposed in this paper improves the accuracy of named entity recognition by 1%, the recall rate by 2%, and the $F1$ value by 2% on average in the field with low-resource.

**Keywords** Named entity recognition · Transfer learning · LSTM · CRF

## 1 Introduction

Named Entity Recognition (NER), as one of the tasks of natural language processing, is designed to detect Named entities from a given text and classify them into predefined categories [1].

At present, many scholars have made relevant researches on Chinese named entity recognition. Lin Guanghe et al. [2] proposed a joint entity recognition method based on the attention mechanism of character level word representation model and bi-LSTM-CRF model, which significantly improved the recall rate of the system. Song Ye-xuan et al. [3] used data with a small number of labels for modeling, and combined with the empirical distribution of labels to deal with the noise in the data set. The source domain data set used in this model was from the common data set in the public domain, and the target domain included entities such as the names of diseases, pests and plants. Some scholars took the characteristics of Chinese characters as auxiliary information and combined with the neural network model for research. Liu Yuhan et al. [4] designed an end-to-end model of BiLSTM+CRF structure combined with the characters of Chinese characters and trained the model on the independently constructed large-scale data set, which verified the good recognition effect of the model in the financial field. However, the disadvantages of this model are slow convergence and difficult to identify entities with other meanings. Dong et al. [5], taking the radicals of Chinese characters as auxiliary information, proposed a named entity recognition model based on neural network to improve the model recognition effect by enhancing the representation information of Chinese characters.

However, the existing Chinese named entity recognition still has the following two problems to be studied: (1) Limitations of domain named entity recognition: In different cultures, fields and backgrounds, the extension of named

✉ Tie Feng
  fengtie@jlu.edu.cn

[1] College of Computer Science and Technology, Jilin University, Changchun 130012, China

[2] Key Laboratory of Symbolic Computation and Knowledge Engineering, Ministry of Education, Jilin University, Changchun 130012, China

[3] College of Software, Jilin University, Changchun 130012, China

entities is different, which is the fundamental problem to be solved by named entity recognition technology [6]. As data in different fields usually have domain-specific characteristics, and data between domains cannot be directly applied to model training across domains, the current research on named entity recognition has only achieved good results in some fields and limited entity types [7]. The lack of data resources in some special fields leads to the absence of annotated data sets, which makes it difficult to directly carry out the training of models. It is necessary to collect and annotate data sets manually, or to turn to unsupervised learning. (2) Named entity boundary problem: At present, the number and types of named entities in different fields are very large, and it is difficult to achieve a comprehensive and different named entities constitution law is not the same, the nesting situation of named entities is complex, one entity nested in another entity is quite common, especially the organization name, for example, "吉林大学白求恩第一医院" is nested in the name of the organization, which is also the name of the organization, "吉林大学" and the name of "白求". Most of the existing named entity recognition models take character level vector representation as input, and it is difficult to accurately judge the boundary of named entities with uncertain length, which leads to the difficulty to improve the recall rate of the model.

This paper proposes a named entity recognition model based on transfer learning, aiming at the shortcomings of current named entity recognition technology in cross-domain recognition and entity boundary recognition. The specific work is as follows: First, a data migration method based on entity characteristics is proposed. The BERT model based on pre-training is applied to carry out text vector quantization. By calculating the feature distribution similarity between low resource data and high resource data, the entity feature with the most category representative is selected for feature transfer mapping, and the distance of entity distribution between the two domains is calculated, so as to make up the gap between the two domains data. The neural network model is trained with high resource data with perfect annotation. Then, an entity boundary detection method based on lexical information is proposed. BiLSTM+CRF is used as the main framework of the model, and character boundary information is used to assist attention network to improve the ability of the model to recognize entity boundaries, so as to further improve the entity recognition performance of the whole model. According to the comparative test results, compared with the ordinary model, the addition of the boundary detection method can improve the $F1$ value by 2%, and the accuracy and recall rate by 1% and 3%, respectively. Finally, a variety of named entity recognition methods based on transfer learning are selected as the baseline method for comparison, and experiments are carried out on data sets of different fields. The results show that the proposed model improves the accuracy of entity recognition by 1%, the recall rate by 2%, and the F1 value by 2% on average in the less labeled domain.

The main contributions of this paper are as follows:

1. A data transfer method based on named entity features is proposed. The data in source domain is adjusted by the features between source domain data and target domain data, and the highly complete data is used as the training set to train the deep neural network model to identify the entities in target domain.
2. Based on the lexical information of the text, a named entity boundary detection and recognition method was constructed, and CRF was built on the BiLSTM neural network model to classify labels, so as to improve the overall recognition performance of the model;
3. A Chinese named entity recognition model based on transfer learning—Neftl-BOUD model is proposed. Firstly, the named entities in the source domain data set are used to build a 300-dimensional word vector for the text to be processed by using pre-trained BERT model, and the similar features between the source domain and the target domain are screened and eliminated. The feature gap of the remaining part of the word vector is reduced by the vector distance formula to generate the data set conforming to the feature of the target domain. The deep neural network model combined with the boundary recognition method is used as the training set, and the parameters are continuously iterated to generate the final model to recognize the named entity.
4. Design comparative experiments on real data sets in multiple fields, and analyze the advantages, defects and deficiencies of the model through the experimental results.

In this article, Sect. 2 introduces the related research work; Sect. 3 introduces the model, including data migration layer and named entity recognition network layer; Sect. 4 introduces the data set processing, comparison model and experimental design; Sect. 5 concludes.

## 2 Related Work

Since the 18th century, professionals have been trying various methods to predict stocks, including the EMH hypothesis, statistics, economics, and artificial intelligence. The most common methods are: basic analysis and machine learning. Models, deep learning analysis methods, etc. These analysis methods will be briefly introduced below. Understanding the professional knowledge and related concepts of these stock markets plays a vital role in analyzing stock market, predicting stock prices, and formulating

corresponding trading strategies, and briefly introduces the attention mechanism and LSTM.

## 2.1 Named Entity Recognition Problem

### 2.1.1 Rule-Based and Statistics-Based Methods

The rule-based method relies on the manual formulation of basic semantic correlation rules, and weights are assigned to the rules respectively. Finally, named entities are recognized by the matching degree between the entities and the rules. Collins et al. [8] proposed a method combining unsupervised learning and artificial rules. In this method, a few sub-rule decision lists were added to obtain more rules, and the pre-established rules were finally applied for entity recognition. Huang Degen et al. [9] obtained statistical information from large-scale toponymy and real text corpus, summarized rules according to the features of place names, and carried out entity recognition by calculating the constitutive credibility of place names. The advantage of statistics-based method is that knowledge can be labeled by training large scale corpus, so that the labeling results have better coverage. Li Huilin et al. [10] combined the CONDITIONAL random field model with the neural network model and proposed the named entity deep layer model based on block representation.

### 2.1.2 Machine Learning-Based Methods

The machine learning-based method converts the named entity recognition task into sequence annotation task, and represents each training sample according to the annotated data in the form of feature engineering. Then the machine learning algorithm is used to train the model from the annotated data to improve the performance. In 1998, Bikel et al. [11] proposed a method based on Hidden Markov Models (HMMs). They used HMMs variants to recognize English place names and institution names, and discussed the influence of training data set size on the model in the experiment. Lafferty et al. [12] proposed the application of Conditional Random Field (CRF) to build the model and perform the segmentation and annotation of named entities. Conditional random fields can relax the strong independence assumption in the HIDDEN Markov model and avoid the limitations of the model.

### 2.1.3 Application of Deep Learning Algorithm

In recent years, with the rapid development of deep neural network technology, the named entity recognition method based on neural network has become the most important research direction because it can train the high-performance model almost without the application of complex feature engineering. Collobert et al. [13] proposed a named entity recognition model based on neural network for the first time. In this model, a large amount of unlabeled text data is used to train the model updating parameters, and the knowledge from named entity recognition and pos tagging is transferred to the blocks by joint training, but the association information between distant words is not taken into account. Lample et al. [14] proposed two new neural architectures for Chinese named entity recognition. One was composed of bidirectional Long short-term Memory (LSTM) and conditional random fields, and the other was marked segmentation based on transfer. Ma et al. [15] proposed an end-to-end model to automatically obtain information from the character-level representation by using the combination of bidirectional LSTM, CNN and CRF for named entity recognition. Chiu et al. [16] proposed a hybrid neural network architecture by combining BiLSTM with convolutional neural network (CNN). This model does not require feature engineering to automatically detect features at word and character levels, and the performance of $F1$ values exceeds that of systems using heavy feature engineering. Zhang et al. [17] proposed a lSTM-CRF model of lattice structure, which is an extension of the character-level model and can be completely independent of the word segmentation method. However, due to the characteristics of RNN, the model is prone to generate information loss. The purely data-driven neural network model has achieved good results in named entity recognition, and some scholars have introduced more information assisted model recognition on this basis to achieve better results. Liu et al. [18] proposed a character sensing NLL model, which used the word level knowledge contained in word embedding and the character level knowledge automatically extracted by the model to complete the task of named entity recognition. Liu et al. [19] combined semi-Markov and CONDITIONAL random fields and introduced external toponymy module to build neural named entity recognition model. Zhang Han et al. [20] trained the neural network model by using data related to the generative adversary-network generation domain, and created new feature representations of words to solve the problem of diversified names of named entities.

## 2.2 Transfer Learning

Transfer learning aims to improve the task learning performance in the target domain by using a large amount of labeled data and pre-training model in the source domain. With its advantages such as little dependence on data and labeling and relaxation of independent homogeneous distribution constraints [21], transfer learning has become a powerful tool to solve NER with scarce resources. Wang Hongbin et al. [22] proposed a case-based transfer learning algorithm, and the experiment proved that the algorithm

improved the experimental performance of Chinese named entity recognition task. The model-based transfer learning method does not require additional high resource language alignment information. It mainly utilizes the similarity and correlation between the source domain and the target domain to transfer some parameters or feature representations of the source model to the target model, and adaptively adjusts the target model. Ando et al. [23] proposed a transfer learning framework, which shared structural parameters among multiple tasks and improved the performance of multiple tasks including NER. Although model-based transfer learning has achieved good performance, the following problems still need to be solved: without considering the differences among resources, the model parameters and feature representation are forcibly shared between languages or domains; Resource data imbalance, high resource language or domain training set size is usually much larger than low resource training set size, ignoring these differences between domains, resulting in poor generalization ability. Cao et al. [24] applied adversarial transfer learning to NER task for the first time, and proposed a Chinese NER adversarial transfer learning model with self-attention mechanism by combining the information of NER task and Chinese word segmentation task. Zhou et al. [6] proposed a dual adversarial migration network model, which well solved the problems of presentation differences and data resource imbalance, and achieved significant improvements in cross-language and cross-domain NER migration.

## 3 Chinese Named Entity Recognition Model Based on Transfer Learning

This chapter combines the data migration model with the entity boundary detection and recognition model, and proposes a named entity recognition model based on transfer learning, which is mainly divided into two parts: data migration layer and named entity recognition network layer. The model framework is shown in Fig. 1. Mesh the model first BERT model was applied to the written judgment of the real data word vector training, to get word to scale, and then, using the method of data migration based on named entity feature building has perfect annotation data as the training set, finally the text word vectors as the input of neural network combining vocabulary information for physical boundary detection enhance the performance of named entity recognition.

### 3.1 Data Migration Layer Based on Named Entity Features

This section proposes a migration learning method based on named entity features, which can filter named entities

with similar features in the source domain and the target domain, and fill the gap between the data of the two domains according to the similar features, so as to achieve the effect of cross domain migration. It ensures that the target domain can have perfect and sufficient annotation data and provides sufficient data guarantee for subsequent experiments.

#### 3.1.1 Entity Feature Selection

1.  Vectorization of text based on Bert model

    Google proposes a word vector generation model [25] based on pre training. Literature [25] provides two sets of models with different parameters. This chapter uses the first model to generate word vectors with fixed dimensions. In the input layer of Bert model, the input text content is regarded as a sequence of words and represented by, where it represents the ith entity in the text sequence, and the generated word vector is expressed as. Because different texts have different length segments, the length of the input word sequence is different. However, the order of characters in the text sequence needs to be completely retained and cannot be changed, let alone recorded and stored in disorder. The subsequent neural network needs to carry out semantic coding according to the order of characters.

    As can be seen from Fig. 2, taking $F1$ value as the index, the recognition performance of the model gradually increases with the continuous increase of word vector dimension. However, when the word vector dimension exceeds 300, the performance of the model begins to decline slightly. At the same time, the training time of the computer also increases exponentially. Therefore, in the subsequent experiments, a 300-dimensional fixed length word vector was selected as the input of the neural network.

2.  Entity feature similarity calculation

    First, it is assumed that there are few labels in the target domain data set, and highly similar labels are selected by statistical method. Because the two data fields have labels that can be mapped to each other or completely different label sets, in order to migrate effective information before feature transfer, this paper combines the data of the source domain and the target domain, and then clusters the combined data set according to the features, Thus, the features with high correlation are gathered. Then calculate the similarity of each feature distribution between the source domain data and the target domain data as the sorting basis, and sort the features in each cluster in descending order. Finally, the highest-ranking feature from each cluster is selected as the feature that needs to be migrated eventually.
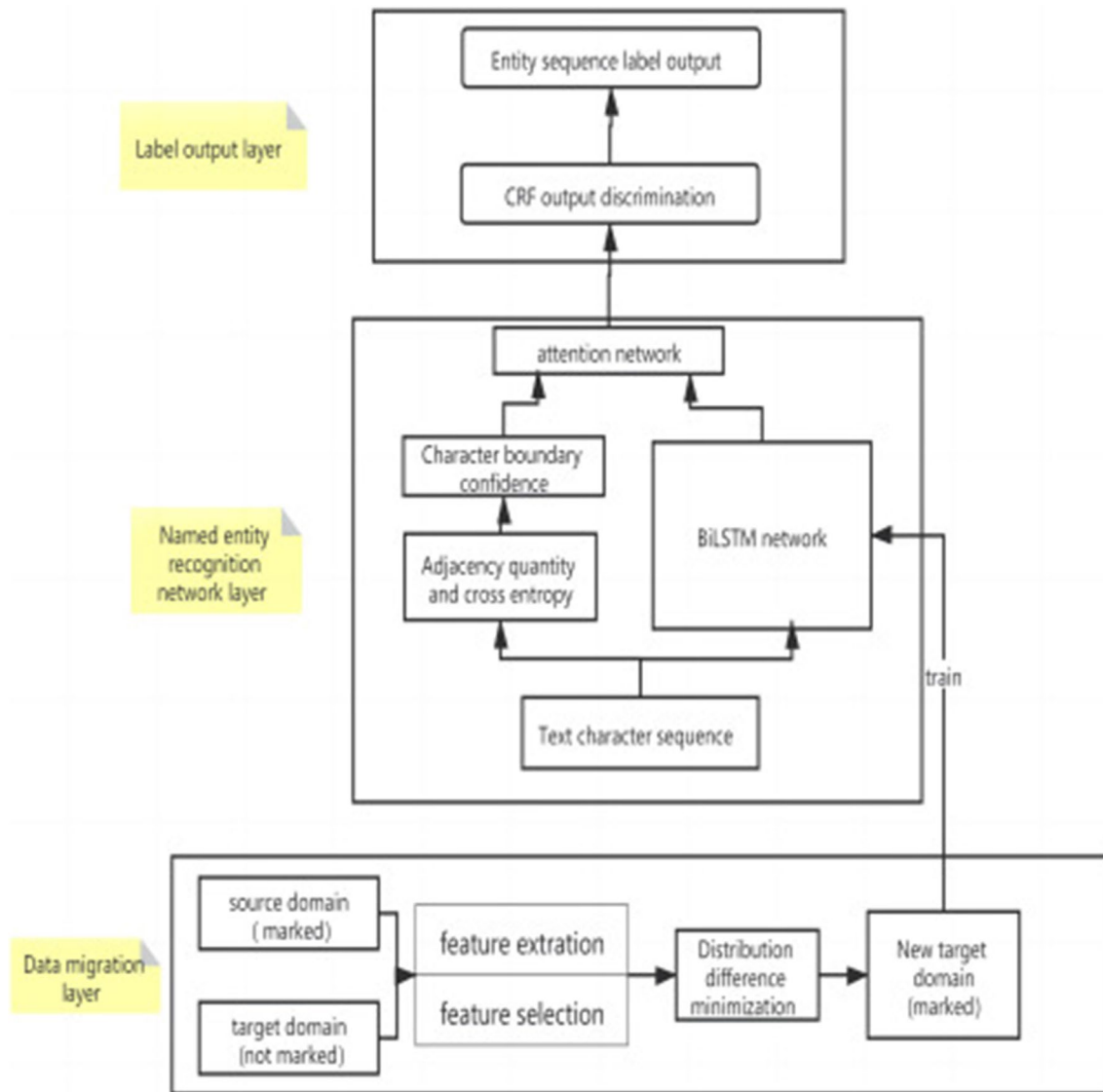
**Fig. 1** Framework of Chinese named entity recognition model based on Transfer Learning

In order to represent the semantic features of text, this paper introduces text classification features as an important basis for feature selection of named entities. This chapter proposes the following three definitions to quantify the relationship between text features and entity features in text data:

**Definition 1** (*Text classification features*) given a text data, there will always be one or more text classification features corresponding to it, i.e:

$$\mathbf{C}_{(\text{text})} = \{\mathbf{c_1}, \mathbf{c_2}, \dots, \mathbf{c_n}\} \tag{1}$$

**Definition 2** (*Entity similar features*) for the target domain T, K features with the highest similarity in the source domain s are called entity similar features, i.e

$$\mathbf{Fe}_{(\mathbf{S})} = \{\mathbf{f_1}, \mathbf{f_2}, \dots, \mathbf{f_k}\} \tag{2}$$

**Definition 3** (*Entity feature text category similarity*) for each entity similarity feature fi ($i = 1, 2, \dots, k$) and text classification feature, there is a certain similarity, and the similarity between them can be recorded as:

$$\mathbf{s_i} = \text{simi}(\mathbf{f_i}, \mathbf{C}_{(\text{text})}) \tag{3}$$

Sort the entity similar features according to the text classification features, arrange the obtained set in descending order, and take the highest $n$ features as the basis for subsequent data migration. The input sequence in this section is marked as $\text{Cont}_{(n)} = \{a_1, a_2, \dots, a_n\}$, where $a_i$ represents the $i$-th eigenvalue in the word vector sequence. $t(i, k)$ is used to
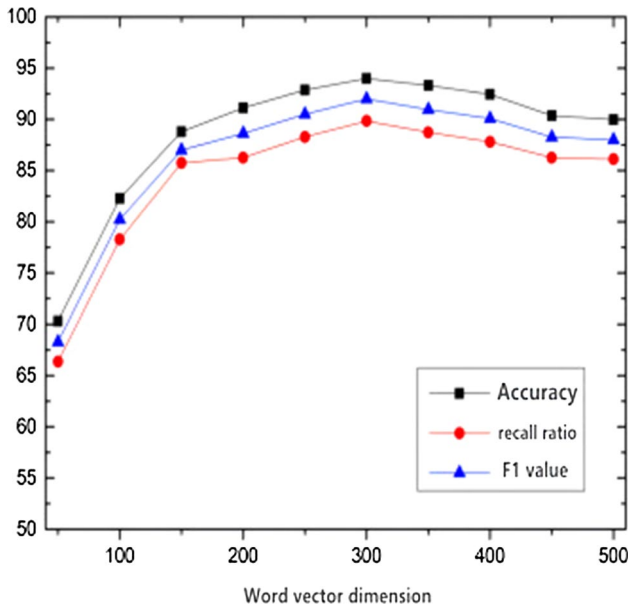
**Fig. 2** Effect of word vector dimension on model performance

represent the k-th chalracter of the *i*-th word in the sentence sequence, which can be regarded as an extension of character-based features. The input of the model is the character sequence $S = \{c_1, c_2, \ldots, c_n\}$ and all character subsequences matching the words in the dictionary, and the corresponding special weight $w_{b,e}^d$ represents the starting character and index of this sequence. For the character vector input by the model, its basic representation in the character level model is shown in formula 4:

$$\mathbf{x_j^c} = \mathbf{e^c}(\mathbf{c_j}) \tag{4}$$

The basic recursive structure of the model is composed of the character unit vector c and the hidden vector of each $c_j$, where $c_j$ is used to record the information flow from the beginning of the sentence to the sequence mark.

The named entity NES in the source domain is vectorized as $\{a_1, a_2, \ldots, a_n\}$, and the named entity NET in the target domain is vectorized as $\{\beta_1, \beta_2, \ldots, \beta_n\}$. According to the feature selection method, m classification basis features are selected as the main category features. In order to narrow the feature gap between the named entities in the source domain and the target domain, it is necessary to shorten the distance by minimizing the distribution difference.

### 3.1.2 Entity Distribution Difference Minimization

To find the distribution similarity between the source domain and the target domain, the next step is how to measure and narrow the distance between the two domain

entities. The source domain dataset $D_s = \{(x_{s1}, y_{s2}), \ldots, (x_{sn}, y_{sn})\}$, where $x_{si} \in R^m$ and $y_{si}$ are the labels corresponding to $x_{si}$, the target domain is represented as XA due to lack of labels, and the centers of the source domain and the target domain are defined as $\mu_s = \frac{1}{n_s} \sum_{i=1}^{n_s} x_{si}$ and $\mu_t = \frac{1}{n_t} \sum_{i=1}^{n_t} x_{n_{ti}}$. First, the distance between the clustering centers of the source domain and the target domain should be as close as possible, that is, $(w^s \mu S - \mu T)^2$ should be as small as possible. The core of measurement is to measure the difference between the data of two entities, and calculating the distance and similarity between two vectors is the basis of much machine learning. Calculating the distance and similarity between two vectors is the basis of many machine learning. The MMD square distance of two random variables is shown in formula 5.

$$\mathbf{M\,M\,D}^2(\mathbf{X}, \mathbf{Y}) = \left\| \sum_{i=1}^{\mathbf{n_1}} \phi(\mathbf{x}_i) - \sum_{j=1}^{\mathbf{n_2}} \phi(\mathbf{y}_j) \right\|_{\mathcal{H}}^2 \tag{5}$$

where $\phi(\cdot)$ is the mapping, which is used to map the original variable to the reproducing kernel Hilbert space.

It is assumed to migrate from the source domain *s* to the target domain T, where the training examples are XS and XT, respectively. WS and WT are used to represent the model parameter sets of source task and target task, respectively. Model parameters are divided into two groups, namely task specific parameters and shared parameters, as shown in formulas 6 and 7.

$$\mathbf{W_s} = \mathbf{W_{s, \, spec}} \cup \mathbf{W}_{shared} \tag{6}$$

$$\mathbf{W_t} = \mathbf{W_{t, \, spec}} \cup \mathbf{W}_{shared} \tag{7}$$

The shared parameter $W_{shared}$ is optimized by two tasks, and the task specific parameters $W_{s, \, spec}$ and $W_{t, spec}$ train each task, respectively. During each iteration, a task (source domain or target domain) is sampled from $\{W_s, W_t\}$ according to the distribution law. Then from a given task in the sampling of a batch of entity instance for subsequent model training, according to the given task of loss function gradient update back propagation, at the same time, the specific parameter updating Shared tasks, then repeat the iterative process, until all the batch data training completion or the change of the loss function is less than a certain threshold iteration to stop. In the iterative process, AdaGrad is used to dynamically calculate the learning rate of each iteration. Because the tasks in the source domain and the tasks in the target domain may have different convergence rates, the early stop method is used in the training process of the target task to prevent the problems of over fitting and negative migration.

### 3.1.3 Data Migration Algorithm Workflow

The steps of data migration algorithm based on entity features are as follows:

1. Input the source domain data and target domain data set, and preprocess the data set such as removing stop words.
2. According to the distribution characteristics of a given data set, K sample points are selected as the central points for clustering, and the entities subject to the same distribution and the central points of their corresponding distribution are found.
3. According to the Bert model, the entity text is vectorized, and the word vector dimension is set to 300 dimensions.
4. All principal axis features in the entity vector are filtered according to the feature selection method.
5. According to the Maximum Mean Discrepancy algorithm, the minimum value of the data distance between the source domain and the target domain and the distance from the center of the data domain are calculated.

6. The data migration weight matrix is calculated by minimizing the center distance and maximizing the data domain distribution.
7. The weight matrix is applied to the neural network layer to generate a new target domain data set.

The algorithm flow is shown in Fig. 3.

## 3.2 Entity Boundary Detection Method Based on Vocabulary Information

This section proposes an entity boundary detection method based on lexical information. This method guides the attention network by calculating the character boundary information in the text sequence, pre judges the entity boundary, improves the accuracy of named entity boundary recognition, and improves the recall rate of entity recognition.

### 3.2.1 Entity Boundary Detection Method

In order to alleviate the problem of unclear entity boundary recognition, this paper proposes a boundary enhanced neural

**Fig. 3** Data migration flow chart based on entity features
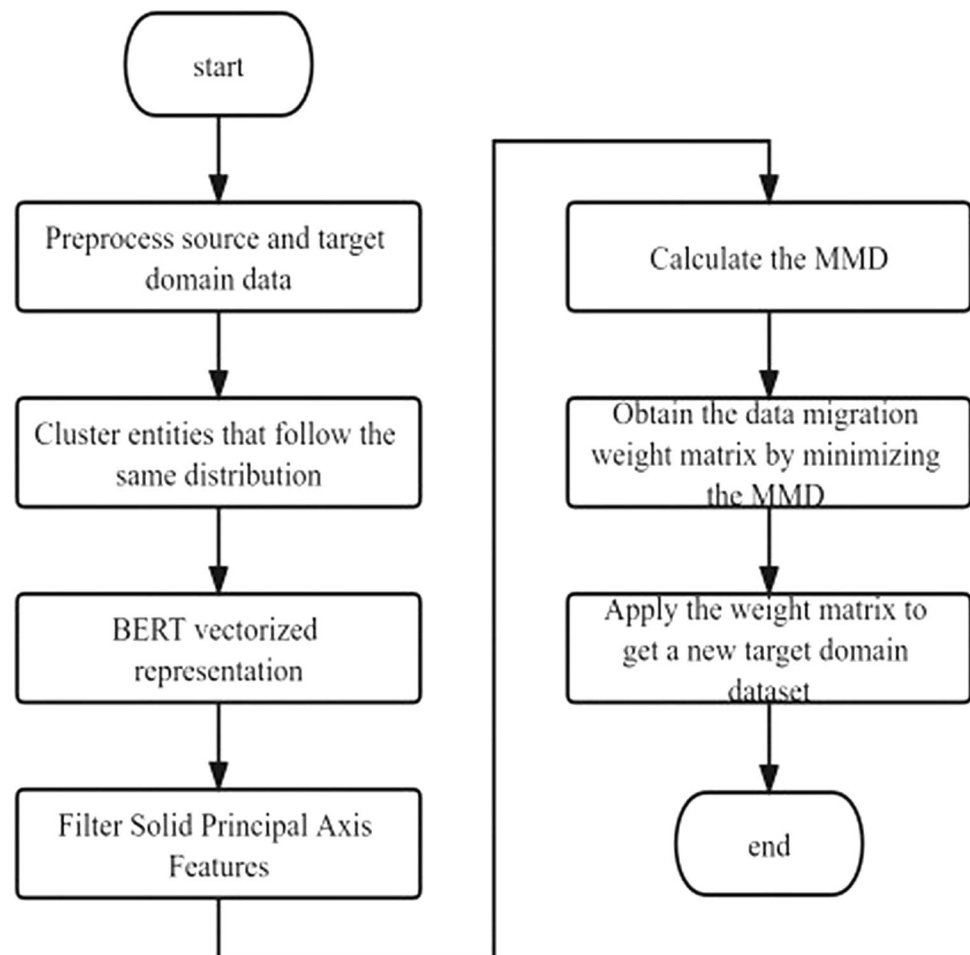
network model, as shown in Fig. 4. The accuracy and recall of named entity recognition are increased by introducing lexical adjacency information and adding boundary detection task.

Because the task of named entity recognition based on conditional random field is to find the optimal path in a tag space, the final recognition effect can be improved by interpreting the left and right boundaries of entity names in advance. The introduction of lexical information can often strengthen the entity boundary, especially for a large range of entity boundaries.

1. Character boundary information statistics

The number of these easily classified negative sample labels is very large, but it plays a far less important role in the training effect of the model than the positive sample. While avoiding the difficult to classify negative samples, it is also very necessary to make effective use of the easy to classify negative samples. Boundary detection is an important step in this paper. In order to effectively enhance the detection of entity boundary, this paper takes the word boundary information and branch entropy in text sequence as auxiliary information to guide boundary recognition.

*Hypothesis* there is a big difference between the forward information and the backward information of a character, and the character at this position may be a boundary. Definition of word boundary information of character *w*:

$$\mathbf{B\ I(w)} = \min \left\{ \mathbf{L_{BI}(w)}, \mathbf{R_{BI}(w)} \right\} \tag{8}$$

where $L_{B\ I}(w)$ and $R_{B\ I}(w)$ represent the number of characters in different categories to the left and right of character w, respectively.

Taking the term "故意伤害罪" as an example, the following are examples from the real data set provided by the website: China judgments Online [20]:

1. Accused the defendant of intentional injury in the original trial.
2. His behavior has constituted the crime of intentional injury.
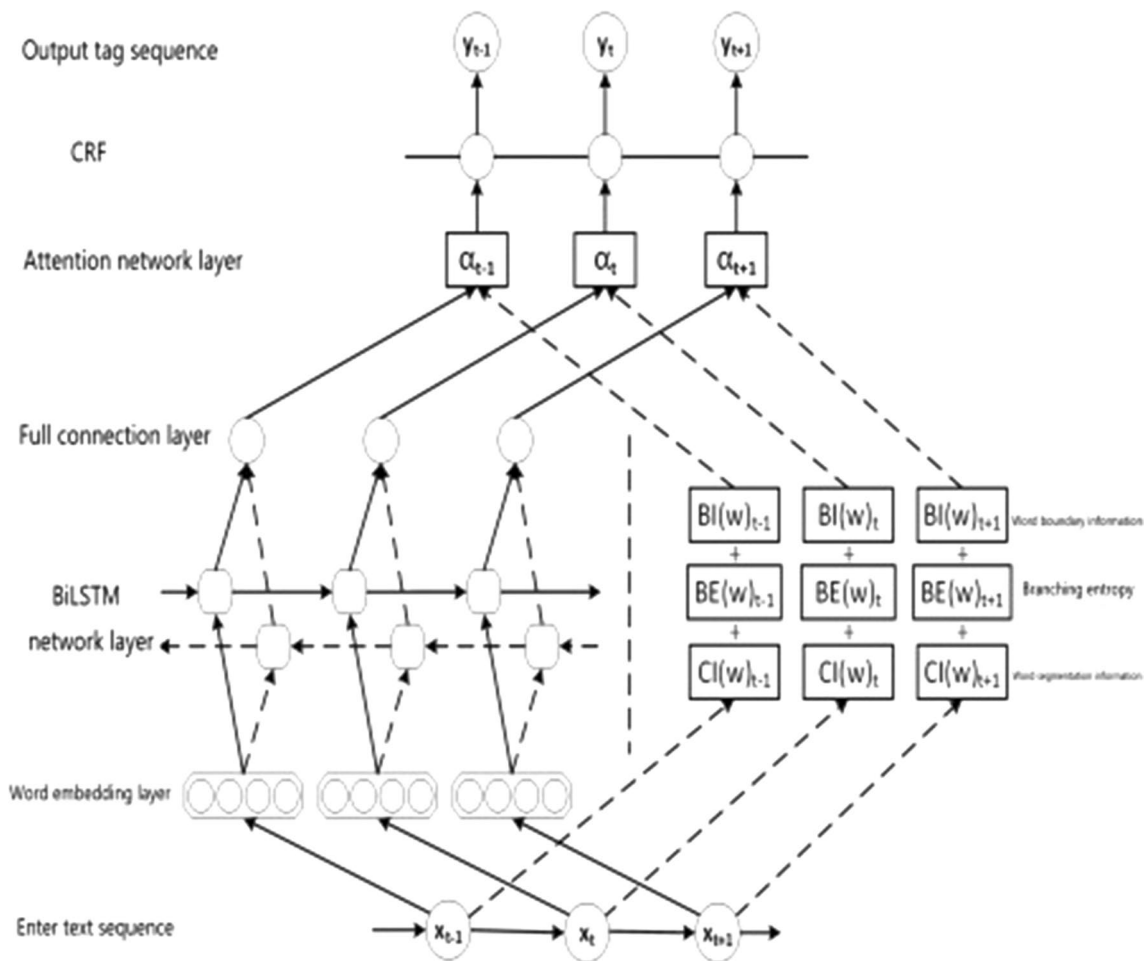


**Fig. 4** Schematic diagram of boundary enhanced neural network model

3. The defendant commits the crime of intentional injury and is sentenced to fixed-term imprisonment of two years.
4. He was criminally detained on mm/DD/yyyy for the crime of intentional injury involving the suspect.
5. The defendant shall be investigated for criminal responsibility for the crime of intentional injury according to law.

According to the above corpus, we can get$_{BI}(w) = 3, R_{BI}(w) = 5$, then BI $(w) = 3$. Branch entropy definition of string s:

$$BE(w) = -\sum_{x \in X} P(x \mid \mathbf{w}) \log P(\mathbf{x} \mid \mathbf{w}) \tag{9}$$

where $x$ is the set of all characters adjacent to character W.

The above information has a wide range of significance for a large number of words in the real corpus, but many of the actually detected entities do not appear in the corpus, so it is impossible to correctly calculate the effective word boundary information. To solve this problem. This chapter introduces the method of word segmentation to supplement the information, which aims to pre-judge the entity boundary through the method of word segmentation. Specifically, given a text sequence $C = \{x_1, x_2, \ldots, x_n\}$, perform word segmentation with two granularities (maximum granularity and minimum granularity) on it:

$$\mathbf{C}\ \mathbf{I}(\mathbf{w}) \in \{\mathbf{0}, \mathbf{1}, \mathbf{2}\} \tag{10}$$

For a character $x_i$ only one of the following three situations will occur: (1) both times are determined as boundaries; (2) once judged as boundary; (3) not determined as a boundary. The number of times characters are judged as boundaries is used as auxiliary information to further make up for the problem that the entity to be recognized does not appear in the corpus.

Combine the auxiliary information related to the above three character boundaries to obtain the reliability of word boundary information, as shown in formula 11,

$$Po = \alpha \times BI(\mathbf{w}) + \beta \times BE(\mathbf{w}) + \gamma CI(\mathbf{w}) \tag{11}$$

Among them, $a, \beta$ and $\gamma$, respectively, represent the weight values corresponding to word boundary information, cross entropy and word segmentation information. According to fuzzy theory, the three are randomly assigned in the process of model training. The same value means that the proportion of statistical information is the same and has the same impact on the boundary.

2. Sequence semantic coding based on BiLSTM

In this section, the LSTM + CRF structure is used as the main network structure of the model. The input sequence of the LSTM network is marked as $S = \{c_1, c_2, \ldots, c_n\}$, where $c_i$ represents the ith character in the sequence. Similarly, $S = \{w_1, w_2, \ldots, w_m\}$ can be further regarded as a word sequence, that is, where $w_j$ represents the $j$-th word in the word. Use $t(i, k)$ to represent the $k$-th character of the $i$-th word in the sentence sequence.Reference [26] proposed a Bert pre training language model. The model uses a bidirectional transformer encoder to integrate the context semantic information on both sides of the word vector.

### 3.2.2 Boundary Detection Method Framework and Structure

1. Algorithm flow

This paper proposes an entity boundary enhancement network model, which assists and guides the neural network to recognize the character boundary through three statistical information: character boundary information, branch entropy and word segmentation information. The pseudo code of the algorithm is shown in Table 1.

The algorithm flow is as follows:

1. Create the BiLSTM neural network and initialize the corresponding parameters $W^T$;

**Table 1** Entity boundary detection algorithm

| Entity boundary detection algorithm based on lexical information |
|---|
| Input: Entity text word vector:$S = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}\}$ |
| Output: Entity boundary confidence: Po |
| Begin: |
| 1) for each xi in $S$ : |
| 2)   $L_{BI}(w) \leftarrow \text{sum}(\text{X}\_left)$ |
| 3)   $R_{BI}(w) \leftarrow \text{sum}(\text{X}right)$ |
| 4)   $BI(w) \leftarrow \min\{L_{BI}(w), R_{BI}(\text{w})\}$ |
| 5)   $BE(w) \leftarrow -\text{sum}(P(x \mid w) * \log(P(x \mid w))$ |
| 6)   $CI(w) \leftarrow \text{cut\_word\_method(content)}$ |
| 7) end for |
| 8) if $L_{BI}(w) == 0$ and $R_{BI}(w) == 0$ : |
| 9)   pre\_NER(S) |
| 10)else: |
| 11)   $P_o \leftarrow W^T * BI(w) + (1 - W^T) * BE(w) + CI(w)$ |
| End |

2. Based on the real corpus, the boundary information $BI(w)_i$, branch entropy $BE(w)_i$ and word segmentation information $CI(w)_i$ of each character in the text sequence to be recognized are calculated;
3. Embedding the word vector into the input text sequence;
4. The word vector is used as the input of neural network and encoded by BiLSTM network;
5. The boundary information $BI(w)_i$, branch entropy $BE(w)_i$ and word segmentation information $CI(w)_i$ are used to guide the attention neural network to pool the corresponding character coding;
6. Input the pooled character code into CRF model for sequence annotation;
7. Boundary discrimination is performed on the labeling results of the sequence. If it is a boundary, the boundary confidence is improved, otherwise, the boundary confidence is reduced.

2. Code implementation

The weights and super parameters in LSTM module are adjusted by random gradient descent method in each iteration, and the two stages are carried out synchronously. The first stage is the input text based on word segmentation, and the second stage is the input text based on NER. Since the corpus of word segmentation is much larger than that of NER, the corpus of the former is randomly sampled in each training stage. Each sample contains 13,500 sentences for training stage 1 and 1350 sentences for training stage 2. The model is trained until the $F1$ value of NER model converges on the validation data set, or up to 30 epochs are trained.

### 3.3 BiLSTM+ CRF Module

The goal of BiLSTM model is to train the model with named entity recognition ability according to the well labeled data set. It is generated by initializing model parameters and iterative training. Therefore, when training BiLSTM network, the experiment aims to minimize the mean square deviation of real entity annotation and hidden state sequence, uses Adam algorithm for small batch gradient descent training, and constantly updates the model parameters until the algorithm converges. Compared with HMM, CRF can calculate the state sequence distribution according to the whole sequence rather than the sequence distribution of the state at the next time. Therefore, CRF layer is stacked on the BiLSTM network layer to judge the final output sequence. Adding a layer of CRF after LSTM can learn the characteristics of state sequence and save complex feature engineering when solving the sequence annotation task.

**Table 2** Experimental results of solid boundary detection method

| Model | Prec. (%) | Rec. (%) | $F1$ (%) | |
|---|---|---|---|---|
| NEFTL | 94.21 | 93.56 | 93.88 | Name |
| NEFTL-Boud | 95.03 | 96.11 | 95.57 | |
| NEFTL | 93.24 | 91.39 | 92.31 | Place name |
| NEFTL-Boud | 94.13 | 94.2 | 94.16 | |
| NEFTL | 89.82 | 88.35 | 89.08 | Organization |
| NEFTL-Boud | 91.73 | 91.86 | 91.79 | Name |

## 4 Experimental Analysis

### 4.1 Verification Experiment of Boundary Detection Method

#### 4.1.1 Data Set

The data set used in this section is the open source data of Microsoft Research Asia (MSRA). The data set contains a total of 46,365 texts. The data set is divided according to the ratio of 3:1:1 to obtain 27,819 texts in the training set, 9273 texts in the verification set and 9237 texts in the test set. There are three entity types in the data set, namely person name, place names and organization names (labeled per, LOC and org, respectively). The source domain data selected in the subsequent experimental stage of this paper is also based on this data set.

#### 4.1.2 Evaluation Index

At present, the most commonly used evaluation criteria for NER are precision, recall and $F1$ score. Their calculation methods are shown in formulas 12, 13 and 14, respectively:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{12}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{13}$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{14}$$

where TP (true positive) represents the number of correct entities recognized; FP (false positive) indicates the number of wrong entities identified; FN represents the number of unrecognized entities.

### 4.1.3 Experimental Results and Analysis

The performance of the model NEFTL-Boud proposed in this paper and its accuracy, recall and $F1$ score on named entities are shown in Table 2. The performance of the model is achieved at a dropout ratio of 0.3 and a learning rate of 0.001. As a comparative experiment, the NEFTL model in the table does not include the method of boundary recognition. In order to compare the performance of boundary detection methods, this section selects MsrA Microsoft Research Asia open source data set for experimental comparison. The main objects identified are three common entities in the text, namely person name, place name and organization name. It can be seen from the experimental results that the overall accuracy, recall and $F1$ value of the model are improved. Among them, $F1$ value is improved by 2% and recall is improved by 3% on average. In conclusion, the entity boundary detection method proposed in this paper can significantly improve the recall rate of entity recognition.

## 4.2 Named Entity Recognition Model Verification Experiment

In this paper, the discriminant named entity annotation method of word sequence annotation is used. The data set is annotated with BIOES annotation mode (beginning, inside, outside, end, single), and the annotation set is BIOES, as shown in Table 3.

### 4.2.1 Dataset

This paper uses the following three data sets: ACL paper data set, China referee network desensitization data set and ccks2018 data set:

1. ACL paper data set: this data set is the resume data collected in the paper ACL 2018 Chinese NER using Lattice LSTM [27], and the format of the data is shown in Table 4. Each line is composed of a word and its corresponding labels, with tabs as separators, and sentences are separated by a blank line. The number of labels is 320,000. The statistical information of the data set is

**Table 3** Comparison of entity labels and their meanings

| Entity tagging | Meaning |
| --- | --- |
| B | Named entity prefix |
| I | Named entity word |
| E | Named entity suffix |
| O | Unnamed entity |
| S | Named entity of a word |

**Table 4** Example table of dimension set

| Character | Remark | Character | Remark |
| --- | --- | --- | --- |
| 李 | B-NAME | 现 | O |
| 某 | I-NAME | 任 | O |
| 某 | E-NAME | 甘 | B-ORG |
| ， | O | 肃 | I-ORG |
| 中 | B-COUN | 友 | I-ORG |
| 国 | E-COUN | 谊 | I-ORG |
| 国 | O | 公 | I-ORG |
| 籍 | O | 司 | E-ORG |
| ， | O | 主 | B-TITLE |
| 汉 | B-RACE | 管 | I-TITLE |
| 族 | E-RACE | 会 | I-TITLE |
| ， | O | 计 | E-TITLE |

shown in Table 5. The target domain data set uses the real judgment document data provided by China judgment document network, including civil, criminal, administrative and enforcement.

2. Desensitization data set of China referee network: Legal referee documents usually contain a large number of person names, place names and organization names, which are very similar to general data sets, and referee documents have strong structure. According to these characteristics, In the experimental stage, only some data need to be manually labeled as test set and verification set to complete the whole training and testing of the model.

3. Ccks2018 dataset: structured storage of electronic medical record information is an important step in building the knowledge map in the medical field. Ccks2018 data set provides 600 well-marked electronic medical record texts, including body parts, symptoms, drug and disease names, surgery and other entity categories.

### 4.2.2 Model Comparison Method

In order to evaluate the overall performance of NEFTL-Boud model and the effectiveness of other network structures, four control models are established, respectively, according to the main model in machine learning method and neural network method, as follows:

**Table 5** Statistics of data set information

| Category | Number of sentences | Number of entities |
| --- | --- | --- |
| Training set | 62,000 | 108,822 |
| Test set | 23,300 | 32,450 |
| Validation set | 23,100 | 32,010 |

1. NEFTL+HMM: This method uses HMM model as the main framework. Other parts are exactly the same as the NEFTL-Boud model.
2. NEFTL+CRF: This method uses CRF as the main framework to distinguish the sequence annotation in the final output layer. other parts are exactly the same as the NEFTL-Boud model.
3. NEFTL+BiLSTM: This method uses BiLSTM model as the main framework. Other parts are exactly the same as the NEFTL-Boud model.
4. NEFTL+BiLSTM+CRF: This method uses BiLSTM-CRF model as the main framework. Other parts are exactly the same as the NEFTL-Boud model.

### 4.2.3 Baseline Comparison Method

In order to evaluate the performance of NEFTL-Boud model in migrating data to complete the task of named entity recognition, experiments will be compared with the following four Chinese NER methods or models based on migration learning: (1) POISE [27]. (2) NER-CWS [24]. (3) TL-NER [28]. (4) Trans-NER [29].

### 4.2.4 Experimental Environment

The named entity recognition model based on transfer learning proposed in this paper is trained and tested in the experimental environment shown in Table 6.

**Table 6** Experimental environment and parameters

| Experimental environment | Parameter |
| --- | --- |
| Operating system | Windows 10 |
| CPU | Intel Core i7-8700 3.19GHz |
| Memory | 8GB |
| Programing language | Python 3.6 |
| IDE | Pycharm |
| Deep learning framework | PyTorch |

### 4.2.5 Result Analysis

The hyperparameters of BiLSTM network model are shown in Table 7. The fitting effect during model training is shown in Fig. 5. In order to prevent over fitting caused by too many iterations of model training and reduce model performance, the experiment adopts early stop method for training. The comparative experimental results are shown in Fig. 6.

The following conclusions can be drawn by comparing the experimental results: For reasons of privacy protection, people's names usually end with "*" or "some" in the judgment documents, resulting in low recall rate and poor performance of HMM, CRF and BiLSTM named entity recognition classification models in people's name recognition; NEFTL-Boud introduces lexical information for entity boundary detection, which significantly improves the recall rate of entity recognition.

Through the comparative experimental results of different named entity recognition methods based on transfer learning, the following conclusions can be drawn:

1. Compared with trans NER model, the entity recognition performance of NEFTL-Boud is greatly improved,
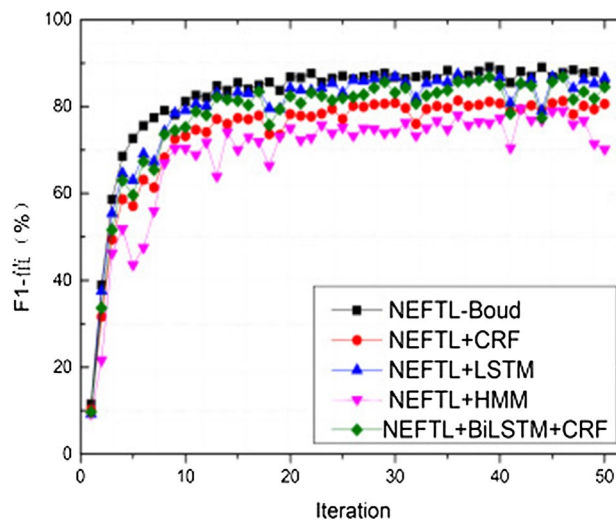


**Fig. 5** Model training fitting effect

**Table 7** BiLSTM-based network model hyperparameters

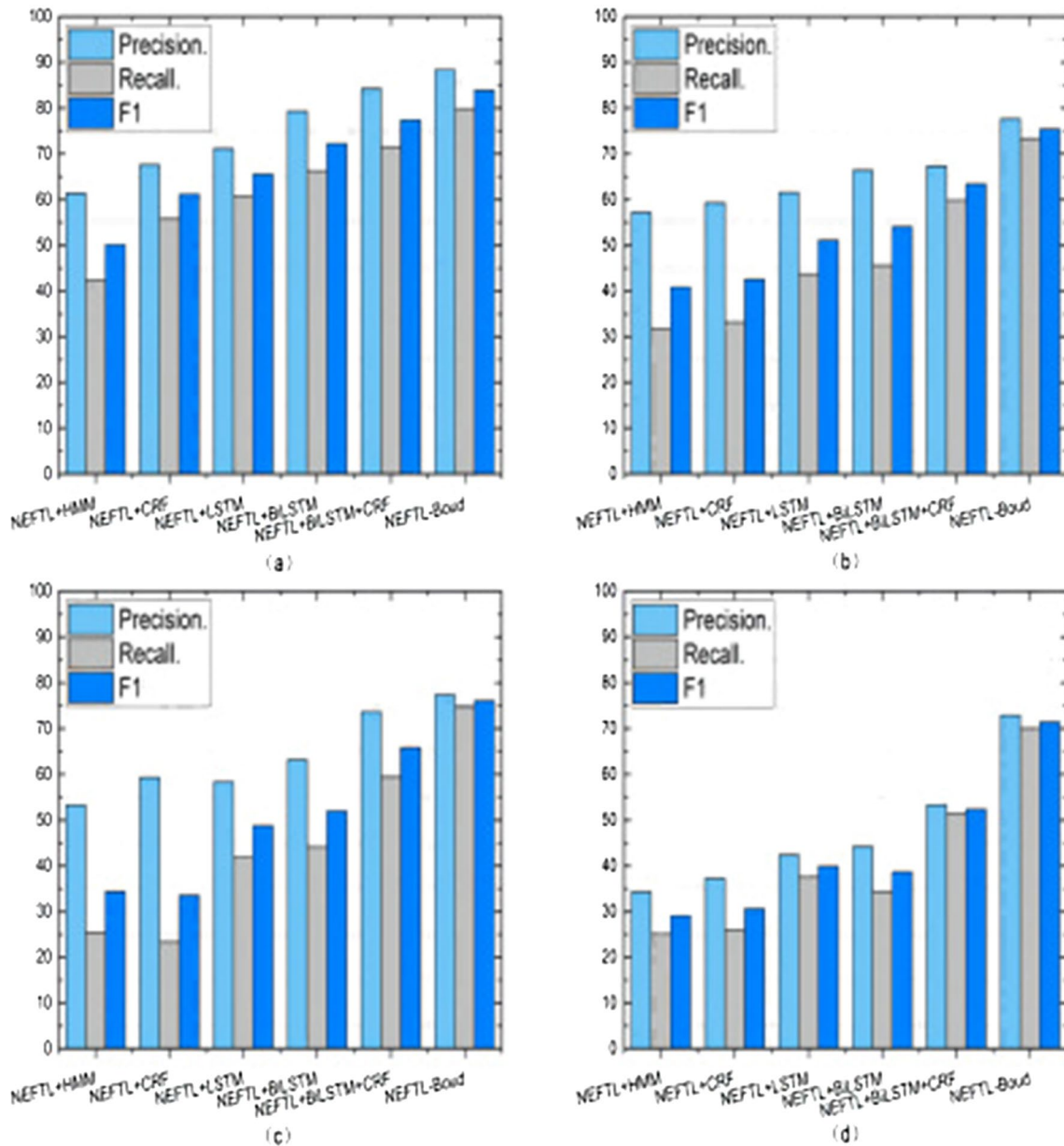| Super parameter name | Super parameter meaning | Optimal value |
| --- | --- | --- |
| $w$ | Dimension of word vector | 300 |
| $H$ | Hidden layer | 3 |
| Dropout rate | Abandonment percentage | 0.5 |
| epoch | Number of iterations | 30 |
| Batch size | Batch size | 64 |
| learning rate | Learning rate | 0.025, decreasing to 0.001 |
| BiLSTM L2 | L2 regularization | 0.0001 |

**Fig. 6** Histogram of control experiment results

which shows that BERT model has great representation ability and can accurately construct text word vector for data migration; Boundary detection methods have a significant impact on the recall rate of entity recognition.

2.  NER + CWS model also uses the word boundary information of the text to assist named entity recognition. The difference is that this method uses anti transfer learning to train the multi structure model. Although it maintains a good recognition effect, it also increases the burden of calculation.

3.  NEFTL-Boud introduces lexical information for entity boundary detection, which significantly improves the recall rate of entity recognition. At the same time, the boundary information and branch entropy in the boundary detection method take the non-entity data around the entities in the data set as the statistical information, which alleviates the problem of data imbalance to a certain extent.

## 5 Summary

In order to improve the ability of named entity recognition in professional segmentation field, this paper proposes a named entity recognition method based on transfer learning. The work completed includes the following:

1. Analyze the basic principles, advantages and existing key problems of traditional and deep neural network-based named entity recognition methods, and investigate the application of transfer learning technology in the field of named entity recognition.
2. In order to improve the cross-domain recognition performance of named entity recognition, a data migration method based on named entity characteristics is proposed. This method minimizes the difference between the source domain and the target domain by matching the similar features between the high resource data and the low resource data, so that the source domain data with high quality tags can be used to train the model, and the task of identifying named entities in the target domain can be solved.
3. In order to improve the recall rate of model identification entities, a method of entity boundary detection was proposed to solve the problem of fuzzy entity boundary. BiLSTM network is used to overlay CRF as the main framework of the model, and the statistical information such as character boundary information is used to guide attention network analysis to distinguish the possibility that characters are boundaries, so as to improve the accuracy of entity boundary detection, and finally improve the overall performance of the model.
4. The identification method proposed in this paper is evaluated using real data sets, and several experiments are conducted on the latest named entity recognition method in different fields and data sets of different sizes. The experimental results show that the proposed method improves the accuracy, recall rate and $F1$ value of named entity recognition.
5. Named entity recognition based on transfer learning can rapidly develop model building and training in the field of less annotation by relying on the general data set with perfect annotation. Extracting effective information from unstructured text and storing it in a structured way is also a very important part of natural language processing tasks. The method proposed in this paper can be effectively applied to the late review of legal documents and the retrieval of medical information and other special scenarios.

**Availability of Data and Material** The datasets generated during and/or analysed during the current study are available in the [A-Research-toward-Chinese-Named-Entity-Recognition-Based-on-Transfer-Learning] repository, https://github.com/xjw-JLU/A-Research-toward-Chinese-Named-Entity-Recognition-Based-on-Transfer-Learning.git.

## Declarations

**Conflict of interest** We declare no conflicts of interest regarding the publication for the paper.

**Ethics approval and consent to participate** Not applicable.

**Consent for publication** We all agree to publish this paper.

## References

1. Yadav, V., Bethard, S.: A survey on recent advances in named entity recognition from deep learning models (2019). arXiv preprint arXiv:1910.11470
2. Guanghe, L., Shaowu, Z., Hongfei, L.: Named entity identification based on fine-grained word representation. J. Chin. Inf. Process. **32**(11), 62–7178 (2018)
3. Yexuan, S., Zhao, C., Gang, W.: Named entity recognition based on partially labelled data and empirical distribution. J. Chin. Inf. Process. **35**(4), 51–57 (2021)
4. Yuhan, L., Changjian, L., Ruifeng, X., Wangda, L., et al.: Utilizing glyph feature and iterative learning for named entity recognition in finance text. J. Chin. Inf. Process. **34**(11), 74–83 (2020)
5. Dong, C., Zhang, J., Zong, C., Hattori, M., Di, H.: Character-based lstm-crf with radical-level features for chinese named entity recognition. In: Natural Language Understanding and Intelligent Applications: 5th CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2016, and 24th International Conference on Computer Processing of Oriental Languages,

ICCPOL 2016, Kunming, China, December 2–6, 2016, Proceedings 24, pp. 239–250. Springer (2016)

6. Zhou, J.T., Zhang, H., Jin, D., Zhu, H., Fang, M., Goh, R.S.M., Kwok, K.: Dual adversarial neural transfer for low-resource named entity recognition. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3461–3471

7. Chen, S., Ouyang, X.: Overview of named entity recognition technology. Radio Commun. Technol. **46**(3), 251–260 (2020)

8. Collins, M., Singer, Y.: Unsupervised models for named entity classification. In: 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora

9. DeGen, H., Guangling, Y.: Identification of Chinese place names based on statistics. J. Chin. Inf. Process. **17**(2), 37–42 (2003)

10. Huilin, L., Yumei, C., Muzhen, S.: Deep network model for text named entity recognition. J. Chin. Comput. Syst. **40**(1), 50–57 (2019)

11. Bikel, D.M., Miller, S., Schwartz, R., Weischedel, R.: Nymble: a high-performance learning name-finder (1998). arXiv preprint arXiv:cmp-lg/9803003

12. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001)

13. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. J. Mach. Learn. Res. **12**, 2493–2537 (2011)

14. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition (2016). arXiv preprint arXiv:1603.01360

15. Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional lstm-cnns-crf (2016). arXiv preprint arXiv:1603.01354

16. Chiu, J.P., Nichols, E.: Named entity recognition with bidirectional lstm-cnns. Trans. Assoc. Comput. Linguist. **4**, 357–370 (2016)

17. Zhang, Y., Yang, J.: Chinese ner using lattice lstm (2018). arXiv preprint arXiv:1805.02023

18. Liu, L., Shang, J., Ren, X., Xu, F., Gui, H., Peng, J., Han, J.: Empower sequence labeling with task-aware neural language model. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32

19. Liu, T., Yao, J.-G., Lin, C.-Y.: Towards improving neural named entity recognition with gazetteers. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 5301–5307

20. Zhang, H., Guo, Y., Li, T.: Domain named entity recognition combining gan and bilstm-attention-crf. J. Comput. Res. Dev. **56**(9), 1851–1858 (2019)

21. Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q.: A comprehensive survey on transfer learning. Proc. IEEE **109**(1), 43–76 (2020)

22. Wang, H., Shen, Q., Xian, Y.: Research on Chinese named entity recognition fusing transfer learning. J. Chin. Mini-Micro Comput. Syst **38**, 346–351 (2017)

23. Ando, R.K., Zhang, T., Bartlett, P.: A framework for learning predictive structures from multiple tasks and unlabeled data. J. Mach. Learn. Res. **6**(11) (2005)

24. Cao, P., Chen, Y., Liu, K., Zhao, J., Liu, S.: Adversarial transfer learning for chinese named entity recognition with self-attention mechanism. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 182–192

25. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding (2018). arXiv preprint arXiv:1810.04805

26. Francis, S., Van Landeghem, J., Moens, M.-F.: Transfer learning for named entity recognition in financial and biomedical documents. Information **10**(8), 248 (2019)

27. Sheng, J., Wumaier, A., Li, Z.: Poise: efficient cross-domain Chinese named entity recognization via transfer learning. Symmetry **12**(10), 1673 (2020)

28. Peng, D., Wang, Y., Liu, C., Chen, Z.: Tl-ner: a transfer learning model for Chinese named entity recognition. Inf. Syst. Front. **22**, 1291–1304 (2020)

29. Wang, Y.-r., Peng, D.-l., Chen, Z., Liu, C.: Trans-ner: a Chinese named entity recognition model supported by transfer learning. J. Chin. Comput. Syst.