IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 August 2022, accepted 4 October 2022, date of publication 12 October 2022, date of current version 18 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3213976

RESEARCH ARTICLE

Chinese Address Recognition Method Based on Multi-Feature Fusion

YANSONG WANG[1], MENG WANG[1], CHAOLING DING[1], XINGHUA YANG[1], AND JIAN CHEN[2]
[1]Chery HuiYin Motor Finance Service Company Ltd., Wuhu 241000, China
[2]Yangtze River Delta Information Intelligence Innovation Research Institute, Wuhu 241000, China

Corresponding author: Jian Chen (chenj@ustc.win)

This work was supported in part by the Key Research and Development Plan of Anhui Province under Grant 202104a05020071.

ABSTRACT A place name is a textual identification of a specific spatial location by people and is an important carrier of geographical information. The recognition of Chinese place names is of great importance in information retrieval and event extraction. The traditional approach is to transform the recognition of Chinese place names into a sequential annotation problem, with commonly used classification models such as support vector machines and conditional random fields. In this paper, Chinese address recognition is converted into a sequential annotation task, and a multi-feature fusion approach to Chinese address recognition is proposed. A deep learning network architecture model based on the fusion of character, word, and address features is constructed to convert characters, words, and their features into vector representations; finally, the sequential annotation of sentences is performed by CRF to achieve the recognition and extraction of address information. On the autonomously constructed dataset, the proposed method MFBL (Multi-Feature-BiLSTM) improves in accuracy by 4 to 10 percentage points compared to other methods, demonstrating that the MFBL model has better performance in the address recognition task.

INDEX TERMS Address recognition, named entity recognition, deep learning, conditional random fields.

I. INTRODUCTION

A place name is a textual identification of a specific spatial location by people and is an essential carrier of geographic information. Identifying Chinese place names is significant in information retrieval and event extraction. However, Chinese addresses have the characteristics of diverse sources and different descriptions. The critical factor for subsequent tasks is how to accurately parse Chinese addresses and identify address entities. The current research on Chinese address recognition mainly includes the following three methods: one is the rule-based address recognition method, which extracts Chinese addresses according to the feature words of address elements and then realizes the recognition of address elements. However, it is difficult to parse and extract non-standard addresses or complex addresses effectively, and thus lacks adaptability. Another method, a statistical model of place name recognition through a large-scale prediction database, is based on statistics and machine learning. The model combines the lexical information of place-name phrases and the context information of sentences, which can solve the problem of semantic ambiguity to a certain extent. The last is a method based on deep learning, which realizes the purpose of address recognition by mining the potential regular features in the data. A method that has been used more recently is the combination of BiLSTM and conditional random field CRF to build an address recognition model.

The methods based on rules and statistics have certain limitations. They rely on the construction of the standard address library and are ineffective in dealing with disordered and missing addresses. Moreover, these methods lack the understanding of address semantics and cannot extract address semantic information effectively. However, the methods based on deep learning still have much room for improvement. In response to this problem, this paper proposes a BiLSTM address semantic recognition model based on multi-feature fusion. First, the Trie syntax tree structure is used to

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

construct a standard address tree, and then a deep learning network architecture model based on character, word, and address feature fusion is used to convert characters, word tags, and their features into vectors, and then input the obtained sentence representation vectors into bidirectional recurrent network layers to obtain sentence semantic information, and finally use CRF to perform sequence labeling of sentences to realize the information identification and extraction of address.

The contributions of this paper are as follows:

1) We propose an address feature recognition method based on deep learning combined with syntax tree rules. Based on the characteristics of Chinese addresses and the unique properties of address feature words, we extract address feature words as the key factors of address semantic representation, and then construct an address syntax tree for Chinese addresses to realize the matching and recognition of addresses.

2) This paper proposes a BiLSTM address semantic recognition model based on multi-feature fusion, an address semantic representation model based on the fusion of character, word, and address features. The first important component of this architecture is responsible for converting word and word tokens and their features into vector representations, then the obtained sentence representation vectors are input to two bidirectional recurrent network layers to obtain address semantic information, and finally the sequence annotation of addresses is performed by CRF to realize the recognition and extraction of address information.

II. RELATED WORK

The current difficulties in identifying Chinese addresses are mainly reflected in the following points: Chinese address descriptions are diverse, and there is no unified specification and coding; Chinese address descriptions are more arbitrary, no unified rules can be followed, and there are redundancy and lack of address descriptions. The Chinese address naming method is relatively updated, and it is difficult to form a dictionary database with comprehensive coverage and high accuracy, so it is impossible to identify Chinese addresses by matching.

Chinese addresses have specific rules and characteristics. Based on the analysis of address rules, each element of the address can be identified. Relevant scholars analyzed and summarized the characteristic words and parsing rules of addresses and realized address recognition based on specific terms and regulations. However, the problem of ambiguity in addresses cannot be solved. For this reason, Tan proposed a mechanism that combines rule tree and ambiguity storage, which partially solves the problem of incompleteness and ambiguity in Chinese addresses [1]. In general, although address recognition based on address rules has a specific effect, the natural complexity of addresses makes it difficult to collect and cover all regulations. The address matching algorithm based on rules and dictionaries proposed by Zhang et al. [2] uses address element feature words and address feature dictionary to extract the most effective matching elements from non-standard addresses to realize the identification of address elements. However, it relies on the completeness of address elements, making it challenging to formulate rules. Relevant scholars use the Chinese address tree model for research [3], [4], [5]. The Chinese address extraction method based on the address tree model proposed by Kang et al. uses the topological relationship as a spatial constraint to extract standard addresses from non-standard addresses. However, it cannot list all non-standard address and place name element sets. To overcome this situation, Wu et al. [6] proposed a multi-strategy address matching algorithm, which combines character similarity and addresses element extraction strategy, and uses the method based on a feature word dictionary and sequence annotation to perform the matching algorithm with standard address database.

The statistical-based word segmentation method considers that the more adjacent characters appear in the same order, the more likely it is to form a word. Based on this idea, related scholars take the address string as the observation sequence and the address element type as the label sequence [7], obtain the address parsing model by training the labeled dataset, and use the parsing model to mark the address to identify the address element type. Statistics-based method models include Hidden Markov Model, Conditional Random Field Model, Support Vector Machine Model, etc. [8], [9], [10]. Zhu et al. [11] summarized the characteristics of Chinese addresses and trained a Chinese address parsing model based on conditional random fields. Based on the conditional random field model, Xu et al. [12] used the fuzzy matching method of local information of Chinese addresses to standardize and parse Chinese addresses. Yuan [13] and others combined statistical methods and rules to design a crime based on statistics and regulations, which explicitly affects disambiguation and improves the efficiency of Chinese address recognition. Related scholars have also proposed similar methods [14], [15], [16], [17].

Zang et al. [18] proposed a construction method based on an address semantic model, which avoids the ambiguity of address semantics by introducing SVM and effectively improves the accuracy of address recognition. Song et al. [10] proposed a Chinese address matching algorithm to identify unstructured addresses by establishing a spatial relationship address model and an address library logic model. However, it only covered 1,000 residential addresses, which were too small in the experiments. Luo and Huang [19] proposed a standardization method based on a finite state machine and an address hierarchy model, which is simple in principle and has strong operability. However, its shortcoming is that when it is aimed at specific practical problems, the model's generalization ability is not strong and has significant limitations. Duan et al. [20] proposed an algorithm for extracting administrative divisions of Chinese addresses based on conditional random fields. An expression model of administrative divisions was obtained by constructing a feature corpus template. Li et al. [21] proposed a rule-based method for the

extraction of administrative divisions to obtain the complete administrative divisions corresponding to address elements by establishing a set of address element rules. Liu et al. [22] proposed a method based on distinct characters, which recognizes Chinese addresses by segmenting the address field. It is simple in principle and easy to use, but it also faces the problem of low accuracy. Other scholars have also conducted in-depth research to address semantics [23], [24], [25], [26].

However, the methods mentioned earlier based on rules and statistics do not consider semantic information, so these methods have poor performance in the case of the irregular distribution of the center of gravity of some non-standard address information. To address such problems, many studies have begun to apply neural networks to address semantic representation tasks by using CNN [27], [28], [29], Recurrent Neural Networks (RNN) [30], [31], [32], Long Short-term Memory Networks (Long Short Term Memory, BiLSTM) [33], [34] or some fusion methods [23], [35], [36], etc. Specifically, there are many studies on the semantic parsing and recognition of Chinese addresses. Zhang et al. [37] used BiLSTM and a four-lemma tagging method to tag the address dataset, which improved the Chinese word segmentation effect. Cheng et al. [38] proposed to use BiLSTM combined with CRF to build a Chinese address parsing model, which improved the accuracy of Chinese address recognition. Wang et al. [39] realized the parsing and recognition of Chinese addresses on a small amount of annotated datasets by combining the improved Transformer and CRF. Some scholars have also done other CRF-based Chinese entity recognition research.

In conclusion, the current identification methods of Chinese addresses all rely on the integrity of the standard address library or use standard methods to identify Chinese addresses, which are not from the point of view of understanding address semantics. The neural network-based method can effectively solve the problem of the lack of semantic information in address matching and the poor effect of traditional methods on differences between address elements. However, for such models, how to effectively integrate the contextual information of global and local contexts is an important issue. By analyzing the characteristics of Chinese address structure, this paper proposes a Chinese address recognition method based on the fusion of characters, words, and their semantics. This method does not rely on the address feature library and realizes accurate identification of Chinese addresses from the perspective of address semantic understanding.

III. METHOD

A. PREREQUISITES

Chinese addresses contain three features: address elements, part of speech, and syntax. Chinese geographical names are usually composed of multiple elements, and each address element belongs to an independent part of the geographical name entity. The address elements are composed of ordinary characters and characteristic words, among which the characteristic words that can reflect the actual semantics and location information of the address can better reflect the essential difference between the address elements. Therefore, the characteristic words are a mark for distinguishing address elements and dividing address levels. In this section, we propose a deep learning network architecture model based on the fusion of character, word, and address features according to the characteristics of the Chinese address structure. The first critical component of the architecture is responsible for converting characters, word tokens, and their features into vectors and then inputting the resulting sentence representation vectors into two bidirectional recurrent network layers to obtain address semantic information. Finally, the sequence labeling of the address is carried out through CRF to realize the identification and extraction of the address.

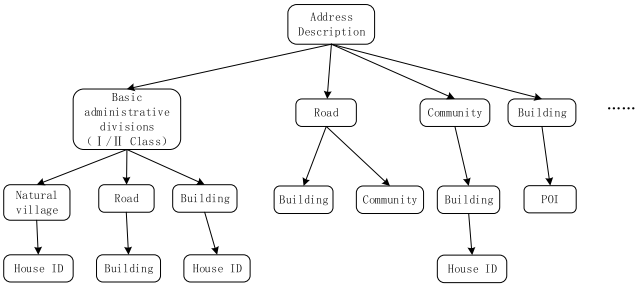


FIGURE 1. Trie grammar tree constructed for Address element extraction model.

A Chinese address is composed of multiple address elements. A valid address element should include the name of the administrative division, the name of the street, the name of the community, the door address, the name of the landmark, and the name of the point of interest. The description of Chinese addresses usually covers the following types: one is to describe the administrative area from wide to narrow, such as "No. 88 Ruixiang Road, Jiujiang District, Wuhu City, Anhui Province", which is a standardized address. Using address element feature words to distinguish address elements, such as "province", "city", "district", "road", "number", etc. An address tree model can be constructed through these characteristic words. The other is to use the building address as the abbreviation of the address, such as "Wuhu City Wanjiang Fortune Center Building", or use relative addresses to describe, such as "Next to Zhongjiang Avenue opposite Wuhu Municipal Government", such non-normalized addresses need to be normalized before identification. Chinese address elements include multiple levels, such as provinces and municipalities directly under the Central Government as the first level, provincial capitals and prefecture-level cities as the second level, districts and counties as the third level, and streets and towns as the fourth level. Taking streets and townships as an example, the feature sets that may include corresponding address elements are towns, townships, offices, neighborhood committees, communities, and streets. Based on the unique

attributes of the address feature word, this paper will extract the address feature word as the key factor of address semantic representation and construct the address syntax tree shown in Figure 1, where the nodes correspond to the address elements in the address. In this paper, the syntax tree is used to match Chinese addresses, and corresponding to the standard addresses that conform to the syntax tree, the address semantic model described below will be used for entity recognition.

B. CHINESE ADDRESS RECOGNITION MODEL

In this paper, we propose a novel Chinese address recognition model, Multi-Feature-BiLSTM (MFBL). According to the characteristics of Chinese addresses, the proposed Chinese address semantic recognition model is established by fusing the characters and word attributes of Chinese addresses. Specifically, the address semantic recognition is divided into three stages: Firstly, extracting character-level information by converting the address character information into vector representation by incorporating the local and global features of characters in address text. Secondly extracting word-level information by obtaining the forward and backward context dependencies of words in the address text. The word-based address semantic information can be extracted in this stage, and the address semantic representation is synthesized by leveraging the association between address indicator words appearing in the address text. Finally, the CRF module is used to synthesize the context feature vector of the previous structure output to obtain the category probability value of each character, and the annotation sequence with the maximum probability is selected as the recognition and annotation result of address elements. The main algorithm is described as follows:

Algorithm 1 Address Semantic Recognition Algorithm Based on MFBL

Input: *address text A*

Output: sequences of entity tags for each address element in address A

Step1. Initialize the resulting sequence *sepResult* as an empty sequence.

Step2. Construct syntax tree *divisionTree*.

divisionTree ← ***BuildTree***(A)

Step3. Traverse the syntax tree nodes to match address indicators.

for *ele* **in** $[a_i, a_j]$ **do**

for *node* **in** *divisionTree* **do**

if *headof* (*ele*, *len*(*node*)) == *node*:

sepList ← *node*

ele.delete(*node*)

if *node* == ***LastNode***(*A*):

sepList ← *ele*

sepResult ← *sepList*

Step4. Identify address elements in A based on MFBL model.

CRF ← ***MFBL*** (*sepResult*)

The address semantic recognition model takes the address text as input, generates the semantic fusion representation vector of the address text based on the character level and word level respectively, and then uses BiLSTM to obtain the address semantics representation. In the end, CRF is used to complete the recognition of address elements. The overall construction of the model is shown in Figure 2. The MFBL model is composed of an embedding module, a semantic feature fusion module, and a CRF module. The specific details of each module are explained in the following.

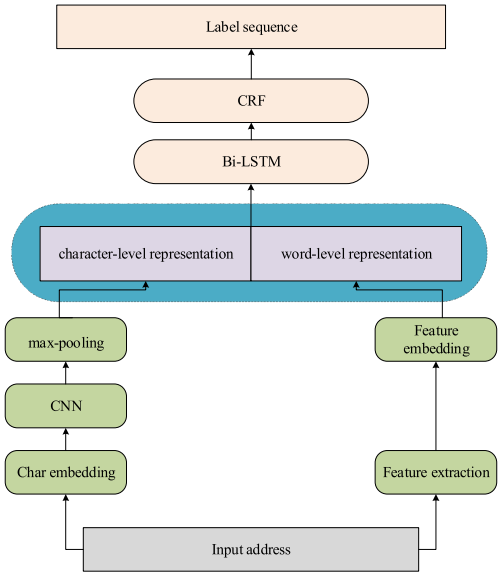


FIGURE 2. The overall framework of the MFBL model.

C. FUSION REPRESENTATION OF CHINESE ADDRESS SEMANTICS

In this paper, we use the address semantic representation that fuses character and word features. In detail, firstly, the character-level embedding vector representation is obtained from the input address text, and the local semantic information is extracted through the convolutional network. Then the generated character-based token sequence representation is passed to the input layer of BiLSTM. Secondly, segment the address text and use a pre-trained language model to encode the segmented sequence into a representation vector and feed it to the word embedding layer. The overall structure is shown in Figure 3.

1) CHARACTER-LEVEL REPRESENTATION

In order to learn more latent semantic connections in the address text, we take the Chinese character features as an input dimension and encode them globally and locally to learn richer semantic information in them. Specifically, firstly, the BiLSTM structure is used to encode the characters in the address text in a bidirectional way, and then the self-attention mechanism is used to effectively obtain the correlation between the characters, which is used as the global semantic information at the character level. Then, we use a convolutional neural network to extract the local features of the characters, and the max-pooling layer is used to remove the redundant local semantic information, and finally, the character-level local features are obtained. The overall

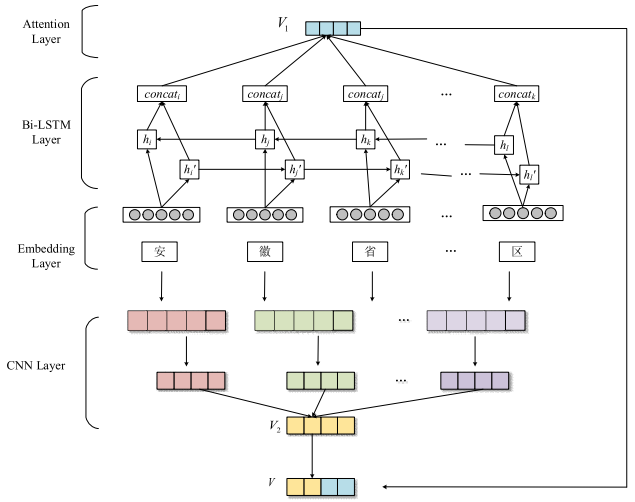


FIGURE 3. The overall structure of the character-level embedding model.

structure of the semantic encoder based on the character level is shown in Figure 4:

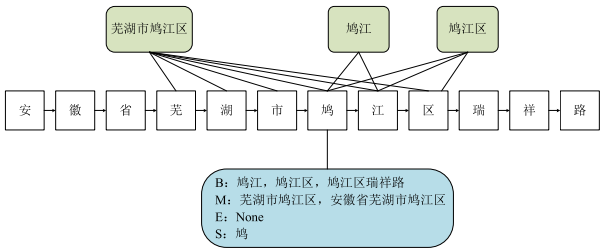


FIGURE 4. Character-based local and global feature embedding.

For the character w_t' in address text at position t, the pre-trained language model Bert is leveraged to convert it into an embedding vector:

$$w_t = bert(w_t') \tag{1}$$

The embedding vectors of the characters are sent into the BiLSTM network, and the semantic representation output of BiLSTM is obtained as $h_t = [\overrightarrow{h}_t; \overleftarrow{h}_t]$, where $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$ denote the forward and backward output of BiLSTM network. Based on h_t, the self-attention mechanism captures the relationship weight between every two characters in the address text, and the relevant calculation formulas are shown as follows:

$$h_t^A = \tanh(W_c[c_t; h_t]) \tag{2}$$

$$c_t = \sum_s \alpha_t h_s \tag{3}$$

$$\alpha_t = soft\max(\chi \tanh(w_a h_s + w_b h_t)) \tag{4}$$

where c_t is the context vector, w_a, w_b, w_c are the weight matrices, and χ is the randomly initialized parameter vector.

The convolutional neural network is used to extract the local features of characters, and then manipulate the output results with the max-pooling layer to retain the most important features from learned features. The relevant formula for local feature extraction using CNN is as follows:

$$CNN(x_t) = Mask(x_t')^* K \tag{5}$$

where x_t' is the embedded input representation of the character, K is the convolution kernel size, and *mask* means padding the input sequence with zero to unify the input dimension. After that, the output results are compressed by a max-pooling layer to retain the most relevant information for subsequent predictions. At time t, the obtained character-level local features are:

$$h_t^C = Max(CNN(x_t)) \tag{6}$$

2) WORD-LEVEL REPRESENTATION

In this paper, we not only use character-level features but also use word-level features. By introducing the character-level and word-level features, we can make full use of the boundary and semantic information of the input text.

Specifically, this method assigns four labels to each character: B, M, E, and S, where B denotes the latent word set beginning with the current character, M denotes the latent word set containing the current character, E denotes the latent word set at the end of the current character, and S denotes the current character.

For example, if the character "鸠" in the text "安徽省芜湖市鸠江区瑞祥路" (Ruixiang Road, Jiujiang District, Wuhu City, Anhui Province) is S, then the set of B, M, E and S is:

{

B: 鸠江, 鸠江区, 鸠江区瑞祥路(Jiujiang, Jiujiang District, Ruixiang Road, Jiujiang District,)

M: 芜湖市鸠江区, 安徽省芜湖市鸠江区(Jiujiang District, Wuhu City, Jiujiang District, Wuhu City, Anhui Province)

E: None

S: 鸠(Jiu)

}

For the B, M, E, and S sets corresponding to each word, the word frequency is used to calculate the corresponding weight, and then get multiple word vectors through the weighting summation method. Finally, the vector representation of B, M, E, and S sets is concatenated as the joint representation of the character and word information, which is used as the input of the model,

$$Emb(B, M, E, S) = concat[Emb(B), Emb(M), Emb(E), Emb(S)] \tag{7}$$

where $Emb(x)$ is a function that maps a word or word set to the representation vector. Then using word frequencies to calculate the corresponding weights,

$$Emb(B) = \alpha_1 * Embedding(B) \tag{8}$$

$$Emb(M) = \alpha_2 * Embedding(M) \tag{9}$$

$$Emb(E) = \alpha_3 * Embedding(E) \tag{10}$$

$$Emb(S) = \alpha_4 * Embedding(S) \tag{11}$$

where $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ denote the proportions of word frequency and $Embedding(x)$ is the word embedding function.

D. FEATURE FUSION

A multi-feature fusion strategy is used to represent the character-level features, containing both global and local features. Multi-feature fusion is a robust and efficient strategy that makes full use of the most significant features to achieve better results. Character-level-based feature fusion can combine multiple related features into a global information representation of the original input sequence. In the feature fusion stage, an adaptive connectivity strategy is used to fuse global and local features. The multi-feature fusion is expressed as follows:

$$h_t = u_1 h_t^A + (1 - u_1) h_t^C \tag{12}$$

where h_t^A and h_t^C are the features obtained from Section 3.2.1, u_1 is the parameter used to adjust the importance degree of these two features.

Finally, the fused character-level representation vector h_t and the enhanced representation vector $Emb(B, M, E, S)$ are concatenated to obtain the representation vector I of the final input layer, and then the concatenated vector was fed into the BiLSTM network.

E. CRF

The fusion representation vector of multi-level features is fed into the BiLSTM network to fully mine richer semantic information. It is assumed that the output sequence of the BiLSTM network is $X = (x_1, x_2, \ldots, x_n)$, and the corresponding label sequence is $Y = (y_1, y_2, \ldots, y_n)$. Conditional Random Field (CRF) is a discriminative probabilistic model, which combines the advantages of Hidden Markov Model (HMM) and Maximum Entropy Model (HEM), and is suitable for the sequence labeling task. When the labeling sequence of X is Y, its probability is calculated by the following formula,

$$p(Y|X) = \frac{e^{s(X,y)}}{\sum_{\widetilde{y} \in Y_x} e^{s(X,\widetilde{y})}} \tag{13}$$

where $e^{s(X,y)}$ is the score of the true path and $\sum_{\widetilde{y} \in Y_x} e^{s(X,\widetilde{y})}$ is the sum of the scores in all possible paths. To maximize the label probability, we use the maximum likelihood function,

$$\begin{aligned} \log(p(y|X)) &= \log\left(\frac{e^{s(X,y)}}{\sum_{\widetilde{y} \in Y_x} e^{s(X,\widetilde{y})}}\right) \\ &= S(X, y) - \log\left(\sum_{\widetilde{y} \in Y_x} e^{s(X,\widetilde{y})}\right) \end{aligned} \tag{14}$$

where $S(X, y) = \sum_i (u_{x_i, y_i} + P[y_i, y_{i-1}])$, then the solution of the maximum likelihood function is transformed into the following formula,

$$y^* = \arg\max S(X, \widetilde{y}) \tag{15}$$

where, u_{x_i, y_i} denotes the probability of element x_i to be labeled as y_i, and P denotes the label transition matrix.

IV. EXPERIMENT

A. SETTING

In this paper, the deep learning framework Keras 2.3.0 based on CUDA 10.0 was used to build the network model. The experiment was carried out on Ubuntu 18.04 LTS system with memory DDR4 32G, 3.6GHz i7-7700 Intel(R) Core(TM) CPU, NVIDIA GeForce GTX 1080 Ti.

B. DATASETS

In order to evaluate the stability of the model proposed in this paper, we used the standard address library to construct datasets containing 268973 Wuhu city address information. Then we selected 90% of the datasets, about 242076 data as the training set, and the remaining 26897 data as the test set. In addition, the ratio of positive and negative samples in the training set and test set is about 3:1. Before model training, the address data needs to be marked. In this paper,"BIO" annotation system is used to annotate address data. First of all, the Chinese processing tool jieba was used to perform segmentation of address data before annotation. Considering that address, as a kind of short text with a special structure, may contain a large number of specific words of place names, we used a self-defined stop word set to perform segmentation. Then we annotated each word according to the rules in the "BIOE" annotation system. As shown in Table 1, B-begin, I-inside, O-outside, and E-end were directly annotated at the end of each address element, and then automatically converted to BIOE format.

TABLE 1. Examples of Chinese address corpus annotation based on the "BIOE".

Data type	Examples
Raw data	Wanjiang Wealth Plaza Building, No.88 Ruixiang Road, Jiujiang District, Wuhu City, Anhui Province (安徽省芜湖市鸠江区瑞祥路88号皖江财富广场大楼)
Annotation data	安 B-Province，徽I-Province，省E-Province，芜B-City，湖I-City，市E-City，鸠B-County，江I-County，区E-County，瑞B-Street，祥I-Street，路 E-Street，88 B-Street_Number，号 E-Street_Number，皖 B-Local_area，江 I-Local_area，财I-Local_area，富I-Local_area，广 I-Local_area，场 I-Local_area，大 I-Local_area，楼 E-Local_area

C. EXPERIMENTAL SETUP

In this paper, the dimension of character features of Chinese characters is set as 20 dimensions, and the word2vec model is used to encode vectorization for each Chinese character. The address data with less than 20 dimensions is encoded with

0 to complement the 20-dimensional coding, and then each word in the address data is represented as the corresponding word vector, which is fused as the vector representation of the whole address data. In terms of the setting of hyper-parameters, the output dimension of each word is set as 768 dimensions according to the possible length of address data. And the semantic representation dimension of the output address data after representation is 100 dimensions. After the semantic representation is completed, the two semantic vectors are respectively input into the network structure of the next layer.

In the training process, the batch size is set to 1024, and the two-layer BiLSTM network is used to obtain the global context information. Combined with the CNN method to obtain local context information, we set dropout to 0.5. Then, the output results of BILSTM and CNN are concatenated and then fed into the self-attention network as a feature matrix. Finally, a 100-dimensional representation vector is output as the semantic representation of address data. For this model, Adam optimizer with epoch 25, learning rate 0.01, dropout 0.5, beta1 0.9, beta2 0.999, and decay rate 0.1 are used as the optimization method of the model. The specific parameters of the model are shown in Table 2.

TABLE 2. Hyper parameter setting of SGAM model.

Parameter name	Parameter value
epoch	25
batch_size	1024
optimizer	Adam
learning_rate	0.01
dropout	0.5

D. EXPERIMENTAL RESULTS AND ANALYSIS

In terms of evaluation metrics, in order to effectively evaluate the prediction results, we select some reference metrics to measure the final results, including accuracy, precious, recall, and F1-score. The higher the accuracy is, the more accurate the model is for the sequence annotation of address data. The higher the F1 score, the better the overall performance of the model.

1) ABLATION EXPERIMENTAL ANALYSIS

In order to verify the effectiveness of the MFBL model, the ablation experiments were performed:

a. The first group only obtains global features at the character level, and uses BiLSTM and attention mechanism to extract global features from the input character representation;

b. The second group only obtains character-level local features through the CNN model and maximum pooling operation;

c. The third group only obtains the features encoded by the word and then obtains the corresponding word set for each character. Finally, it is concatenated with a single character representation;

d. The fourth group is the global and local feature representation at the fusion character level;

e. The fifth group is the multi-feature fusion representation method based on character and word coding proposed in this paper.

TABLE 3. Experiment Results on comparison experiment.

Model	F1	Acc	Recall	Pre
①	0.7245	0.7367	0.7562	0.7782
②	0.7471	0.7682	0.7821	0.7932
③	0.8234	0.8342	0.8431	0.8342
④	0.8362	0.8472	0.8732	0.8261
⑤	**0.8924**	**0.9022**	**0.8943**	**0.8820**

As shown in Table 3, it can be found that the experiment based on a single feature makes the performance of the model deteriorate. The overall accuracy, recall, and F1-score are all poor whether only character-based global features or local features are selected or simply word-level features are used as the input of neural networks. The reason is that the input information of the model is less and the model is not fully trained, which leads to the poor performance of the model on the test corpus. For the fourth group of experiments, the global features and local features at the character level are used as the common input. It can be seen that the performance of entity recognition is improved, and the results are slightly better than those of the third group of experiments with word-level features as the input. The reason is that the extraction of global and local features of characters fully mines the potential information contained in the characters, which is helpful for the model to fully train the corpus. From the fifth group of experiments, it can be seen that the proposed model tends to be more effective with the increase of fused features, which shows that the fused features are helpful to improve the performance of entity recognition from a multi-level perspective.

2) COMPARISON WITH BASELINE MODEL EXPERIMENT

In order to verify the effectiveness of the MFBL model proposed in this paper, the model proposed in this paper is compared with the classical model. This paper sets up the following groups of comparison model experiments:

a. The first group only uses the BILSTM-CRF model for address sequence annotation;

b. The second group uses the BILSTM-CRF model and adds the attention mechanism to the experiment;

c. The third group combines with CNN network to obtain local context information and runs the experiments with BILSTM-CNN-CRF model;

d. The fourth group is the MFBL model proposed in this paper. The attention mechanism is added to BILSTM-CRF and the CNN network is combined for joint training.

TABLE 4. Experiment Results on comparison experiment.

Model	F1	Acc	Recall	Pre
BiLSTM-CRF	0.7645	0.7723	0.7843	0.7563
BiLSTM-Attention-CRF	0.8516	0.8691	0.8565	0.8531
BiLSTM-CNN-CRF	0.8438	0.8535	0.8491	0.8457
MFBL	**0.8924**	**0.9022**	**0.8943**	**0.8820**

The results of the comparison experiment are shown in Table 4, from which it can be concluded that the MFBL model proposed in this paper has achieved the best results in terms of accuracy, recall, and F1-score. The result indicates the effectiveness of the proposed method in Chinese address recognition. From the table, the second group adopts the entity recognition method combined with the attention mechanism, which improves the overall effect of the model. It shows that adding the attention mechanism can learn effective features from a global perspective. From the third group of experimental results, it is found that using CNN to obtain local effective features can also improve the performance of the model. Meanwhile, by comparing the experimental results of the fourth group, the second group, and the third group, it can be seen that the model proposed in this paper improves the performance of other models by about 4-5 percentage points in terms of F1-score. It is directly proved that the model cannot effectively capture some key information in the address when only considering the attention mechanism or the local information obtained by CNN. At the same time, the F1-score proves that the accuracy improvement of MFBL model is not affected by the proportion of positive and negative samples in the dataset. However, the overall learning ability of the model is indeed enhanced compared with other ablation models.

3) EXPERIMENTAL SUMMARY

For the problem that the redundant information of Chinese addresses cannot be recognized, a Chinese address recognition model based on multi-feature fusion is proposed in this paper. The actual experimental results show that this method has a good effect on many metrics. By comparing the simple BILSTM-CRF model, CNN network and attention mechanism are used to extract local features and global features of address data, which can effectively improve the performance of the model. In this paper, the particularity of address is considered in Chinese address recognition, and the semantic features of address are extracted from multiple dimensions. However, the association between an address and the geographical entity is not studied. In the next step, we will consider and try to introduce information such as a geographic information graph to enhance the accuracy of recognition, and the generalization ability of unknown address datasets also needs to be further studied.

V. CONCLUSION

This paper proposes a BiLSTM address semantic recognition model based on multi-feature fusion, using a grammar tree structure to construct a standard address tree, followed by a deep learning network architecture model based on the fusion of word, word and address features, converting word and word tokens and their features into vector representations to obtain sentence semantic information, and finally performing sequence annotation of sentences through CRF to achieve address information recognition and extraction. The experiments show that the MFBL model proposed in this paper has better performance in the address recognition task. However, with the development of urbanization, the description of Chinese addresses has great differences and volatility. In the face of the complexity and irregularity of Chinese addresses, our method has some limitations. To solve this problem, we need to collect a wider range of data sets, optimize the data with appropriate methods, and further improve the experimental methods, so that we can more accurately understand the semantic information of Chinese addresses and complete the semantic identification of addresses.

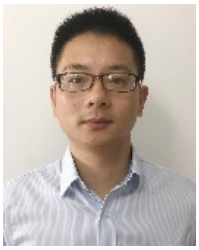
REFERENCES

[1] K. Tan, "Rule-based Chinese address segmentation and matching methods," Shandong Univ. Sci. Technol., Shandong, China, Tech. Rep., 2011, doi: 10.7666/d.D300758.
[2] X. Zhang, G. Lv, and B. Li, "Rule-based approach to semantic resolution of Chinese adresses," *J. Geo-Inf. Sci.*, vol. 12, no. 1, pp. 9–16, 2010.
[3] M. Kang, Q. Du, and M. Wang, "Chinese address extraction based on address tree model," *Acta Geodaetica et Cartographica Sinica*, vol. 44, no. 1, pp. 99–107, 2015.
[4] C. Xu, F. Zhang, Z. Du, Y. Zhang, M. Chen, and R. Liu, "A multi-level address-matching algorithm based on Hash function and double-array trie-tree," *J. Zhejiang Univ.*, vol. 41, no. 2, pp. 217–222, 2014.
[5] B. Li, "Research on key technologies of Chinese address coding," Nanjing Normal Univ., Nanjing, China, Tech. Rep., 2009, doi: 10.7666/d.y1726610.
[6] R. Wu, H. Long, X. Xiong, and Y. Peng, "A multi-strategy combined address matching algorithm," *J. Henan Polytech. Univ.*, vol. 38, no. 5, pp. 124–129, 2019.
[7] R. Jian, "Establishment of address standardization model based on statistical method," Yunnan Univ., Yunnan, China, Tech. Rep., 2015.
[8] Y. Quan, "New progress of Chinese word segmentation technology in China," *Sci. Res.*, no. 47, p. 70, 2015.
[9] X. Zhang, T. Wang, and H. Chen, "Research on named entity recognition," *Comput. Sci.*, vol. 32, no. 4, pp. 44–48, 2005.
[10] Z. Song, "Chinese address matching algorithm for natural language understanding," *J. Remote Sens.*, vol. 17, no. 4, pp. 788–801, 2013.
[11] F. Zhu, T. Zhao, Y. Liu, and Y. Zhao, "Research on Chinese address resolution model based on conditional random field," *J. Phys., Conf.*, vol. 1087, Sep. 2018, Art. no. 052040.
[12] Y. Xu, B. Shen, and X. Xu, "Non normalized Chinese address resolution method based on conditional random field," *Geography Geographic Inf. Sci.*, vol. 35, no. 2, pp. 12–18, 2019, doi: 10.3969/j.issn.1672-0504.2019.02.003.
[13] X. Yuan, "Design and implementation of Chinese address word segmentation system based on statistics and rules," Southeast Univ., Nanjing, China, Tech. Rep. GB/T 7714-2015, 2018.
[14] H. Zhang, F. Ren, H. Li, R. Yang, S. Zhang, and Q. Du, "Recognition method of new address elements in Chinese address matching based on deep learning," *ISPRS Int. J. Geo-Inf.*, vol. 9, no. 12, p. 745, Dec. 2020.

[15] X. Yao and Y. Lu, "A post-processing approach to Chinese address recognition," in *Proc. 8th Int. Conf. Fuzzy Syst. Knowl. Discovery (FSKD)*, Jul. 2011, pp. 1906–1910.
[16] C. Long, X. Zhu, K. Huang, J. Sun, Y. Hotta, and S. Naoi, "An efficient post-processing approach for off-line handwritten Chinese address recognition," in *Proc. 8th Int. Conf. Signal Process.*, 2006, pp. 1–4.
[17] Y. Ling, J. Yang, and L. He, "Chinese organization name recognition based on multiple features," in *Proc. Pacific-Asia Workshop Intell. Secur. Inform.* Berlin, Germany: Springer, 2012, pp. 136–144.
[18] Y. Zang, B. Wang, and X. Qu, "Discussion on the construction method of Chongqing Chinese semantic address model," *Geospatial Inf.*, vol. 13, no. 3, pp. 122–125, 2015.
[19] M. Luo and H. Huang, "A Chinese address standardization method based on finite state machine," *Comput. Appl. Res.*, vol. 33, no. 12, pp. 3691–3695, 2016.
[20] Y. Duan, X. Li, and S. Huang, "Extraction method of Chinese address administrative division based on conditional random fields," *J. Wuhan Univ. Eng.*, vol. 37, no. 11, pp. 47–51, 2015.
[21] X. Li, S. Huang, and T. Lu, "Algorithm for extracting administrative divisions of non standardized Chinese addresses," *Comput. Appl.*, vol. 37, no. 3, pp. 876–882, 2017.
[22] Z. Liu, X. Xia, and F. Zhou, "Word segmentation based on Chinese address information," *J. Shenyang Univ. Aeronaut. Astronaut.*, vol. 25, no. 4, pp. 63–66, 2008.
[23] G. Wu, G. Tang, Z. Wang, Z. Zhang, and Z. Wang, "An attention-based BiLSTM-CRF model for Chinese clinic named entity recognition," *IEEE Access*, vol. 7, pp. 113942–113949, 2019.
[24] H. Sui, Y. Jianping, Z. Hongxian, and Z. Wei, "Sentiment analysis of Chinese micro-blog using semantic sentiment space model," in *Proc. 2nd Int. Conf. Comput. Sci. Netw. Technol.*, Dec. 2012, pp. 1443–1447.
[25] Z. Xie, Z. Sun, L. Jin, H. Ni, and T. Lyons, "Learning spatial-semantic context with fully convolutional recurrent network for online handwritten Chinese text recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 8, pp. 1903–1917, Aug. 2018.
[26] D. Sui, Y. Chen, K. Liu, J. Zhao, and S. Liu, "Leverage lexical knowledge for Chinese named entity recognition via collaborative graph network," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 3830–3840.
[27] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 213–228.
[28] Y. Kim, S. Kim, T. Kim, and C. Kim, "CNN-based semantic segmentation using level set loss," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1752–1760.
[29] T. Gui, R. Ma, Q. Zhang, L. Zhao, Y.-G. Jiang, and X. Huang, "CNN-based Chinese NER with lexicon rethinking," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 4982–4988.
[30] M. Qi, J. Qin, A. Li, Y. Wang, J. Luo, and L. Van Gool, "StagNet: An attentive semantic RNN for group activity recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 101–117.
[31] J. Guo, J. Cheng, and J. Cleland-Huang, "Semantically enhanced software traceability using deep learning techniques," in *Proc. IEEE/ACM 39th Int. Conf. Softw. Eng. (ICSE)*, May 2017, pp. 3–14.
[32] J. Cheng, Y. Sun, and M. Q.-H. Meng, "A dense semantic mapping system based on CRF-RNN network," in *Proc. 18th Int. Conf. Adv. Robot. (ICAR)*, Jul. 2017, pp. 589–594.
[33] Y. Lin, M. Kang, Y. Wu, Q. Du, and T. Liu, "A deep learning architecture for semantic address matching," *Int. J. Geograph. Inf. Sci.*, vol. 34, no. 3, pp. 559–576, Mar. 2020.
[34] J. Chen, J. Chen, X. She, J. Mao, and G. Chen, "Deep contrast learning approach for address semantic matching," *Appl. Sci.*, vol. 11, no. 16, p. 7608, Aug. 2021.
[35] H. Llorens, E. Saquete, and B. Navarro, "TimeML events recognition and classification: Learning CRF models with semantic roles," in *Proc. 23rd Int. Conf. Comput. Linguistics*, 2010, pp. 725–733.
[36] B. Y. Lin, F. Xu, Z. Luo, and K. Zhu, "Multi-channel BiLSTM-CRF model for emerging named entity recognition in social media," in *Proc. 3rd Workshop Noisy User-Generated Text*, 2017, pp. 160–165.
[37] W. Zhang, "Design and optimization of Chinese address matching system," Wuhan Res. Inst. Posts Telecommun., Wuhan, China, Tech. Rep., 2018.
[38] B. Cheng, W. Li, and H. Tong, "Chinese level address segmentation based on BiLSTM-CRF," *J. Earth Inf. Sci.*, vol. 21, no. 8, pp. 1143–1151, 2019.
[39] F. Wang, "Research on Chinese address word segmentation method for a small amount of tagged data," Zhejiang Univ., Zhejiang, China, Tech. Rep., 2020, doi: 10.27461/d.cnki.gzjdx.2020.000564.

YANSONG WANG was born in 1979. He received the master's degree from the University of Science and Technology of China, in 2006. He has been working as a Vice Manager with Chery HuiYin Motor Finance Service Company Ltd., since 2017. His research interests include artificial intelligence and big data.

MENG WANG was born in 1993. He received the master's degree from Anhui Normal University, in 2020. He has been working as an Engineer with Chery HuiYin Motor Finance Service Company, Ltd., since 2021. His research interest includes machine learning.



CHAOLING DING was born in 1988. He received the master's degree from Huangshan University, in 2011. He has been working as an Engineer with Chery HuiYin Motor Finance Service Company Ltd., since 2020. His research interest includes data mining.

XINGHUA YANG was born in 1996. He received the master's degree from the Anhui University of Finance and Economics, in 2022. He has been working as an Engineer with Chery HuiYin Motor Finance Service Company Ltd., since 2022. His research interest includes NLP.

JIAN CHEN was born in 1989. He received the master's degree from Sichuan University, China, in 2014. He has been an Assistant Engineer with the Big Data Laboratories, Yangtze River Delta Information Intelligence Innovation Research Institute, since 2020. His research interests include natural language processing and data mining.