Research on Extraction Method of Financial Knowledge Based on How Net

Chaoyang Geng
School of Computer Science and Engineering
Xi'an Technological University
Xi 'an, 710021, China
E-mail:541211200@qq.com

Peng Liu
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
E-mail: 643973010@qq.com

JieJie Zhao
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
E-mail:13734100641@163.com

Dan Yang
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
E-mail: 1330311378@qq.com

***Abstract*—In order to obtain the knowledge information of financial texts more efficiently and make the extracted information such as entity relation attribute more accurate, this paper studies the grammatical features of financial news texts and the semantic features of How Net, and puts forward the scheme of financial information extraction based on How Net. First, the phrase matching is carried out in the dictionary. Then the neural network is used for weighting, BiLSTM is used for character vector feature enhancement training, and then conditional random field (CRF) is used to complete named entity recognition, and then the relationship extraction of entity pairs from the dependency syntax is carried out to complete the research on the construction method of knowledge extraction of text in the financial field. The experimental results show that this model is superior to the other three models in entity recognition, and the overall performance is improved by about 1.2%. In relation extraction, the accuracy and recall rate of the model algorithm adopted in this paper are improved by 5% and 1.5% respectively, which shows that the improvement of the algorithm is effective.**

Keywords-Named Entity Identification in the Financial Sector; Relation Extraction; How Net

I. INTRODUCTION

Information in the financial field is directly or indirectly related to all fields of social life, and it has the characteristics of huge data, various types and rich text formats. At present, with the rapid development of artificial intelligence, the fintech industry has gradually turned from manual to intelligent. News texts in the financial field are mostly unstructured text information, which contains a large amount of knowledge. Knowledge extraction can gather the entity relationship attributes and other information in financial news texts together in the form of triples to obtain efficient information for financial practitioners.

The entity recognition of financial texts plays a vital role in completing the construction of knowledge extraction in the financial field. Therefore, the named entity recognition of financial information texts is one of the main research contents of this paper. At present, there are two new challenges in naming and recognition in the financial field: the first is to identify the name of the organization: the name of the organization in the financial text often contains some personal names, place names, different abbreviations or unknown words, and its naming rules are not restricted by a unified format; The second is the recognition of terms in the financial field: entities in terms in the financial field are characterized by complex types and irregular updates. The boundary of professional terms in the financial field cannot be properly divided by common word segmentation tools, and colloquial

expressions of financial terms cause ambiguity identification, such as track and leverage.

Entity relationship extraction technology extracts structured relational triples from unstructured or semi-structured financial texts, which can not only promote the construction of knowledge base, but also provide support and help for intelligent retrieval and semantic analysis of financial knowledge. At present, the key problem in entity relation extraction is that overlapping triples seriously affect the efficiency of triplet recognition.

By studying the grammatical features and semantic features of financial news text, this paper designs the entity relation extraction algorithm of financial news text integrating How Net, and completes the knowledge extraction in the financial field. This paper proposes a scheme of financial entity identification based on How Net and BiLSTM-CRF, then elaborates on feature selection and named entity annotation, and then completes the relationship extraction between entity pairs through a relationship extractor including dependency syntax module. On this basis, the entity relationship extraction flow chart of financial news text is built. Finally, the effectiveness of the scheme is verified by experiments.

II. INFORMATION EXTRACTION RELATED WORK

As one of the important basic technologies of information extraction, entity relation extraction is mainly aimed at identifying the required entities and the semantic relationships between entities from unstructured natural language texts. This technology can provide important support for the construction of knowledge base and knowledge graph and the realization of intelligent search and intelligent question and answer.

A. Named entity recognition

The main task of named entity recognition in the financial field is to use natural language processing and artificial intelligence technology to extract financial entities in professional fields other than the names of people, places and institutions, such as the company name, company name, financial product name, project name, etc.

In the field of financial data, people first began to study entity recognition through a rule-based and dictionary-based approach, and the entities studied were only the abbreviation of the name of the institution. Wang Ning et al. [1] summarized the structural features and context information of the company name, then constructed six knowledge bases for identifying the company name manually, and completed entity identification through two traversal search and identification. The experimental results showed that in the closed test, the accuracy of the recognition of the organization name could reach 97.3%, and the recall rate reached 89.3%. But in addition to being time consuming, this approach requires a strong background in finance, so the researchers focused on statistical machine learning methods.

Statistical machine learning is a method based on feature engineering. This method is to manually select and design relevant features, then vectorize these features, and then use mathematical function models to make predictions. Although the statistics-based machine learning method has achieved good results in the data sets of the general domain, it is still difficult to identify financial named entities in the field of finance, especially in the field of finance. Therefore, Wang et al. [2] proposed a financial named entity recognition method based on conditional random field and mutual information and information entropy. Since stock names often appear in financial texts, the author adopted domain dictionary to identify stock names. They integrated the rule features of company name suffix, place name prefix, title and subsequent words into a linear CRF classifier to identify the complete financial naming entity. The experimental results showed that the accuracy of the model increased by about 6% after adding mutual information and information entropy.

Because of its powerful learning ability, deep learning has been applied to the field of natural language processing in recent years. Li Qianwen [3] proposed the structure of Word-BiLSTM-Attention-CRF cyclic neural network model, which was based on the BiLSTM-CRF model, and then took the vector combining words as the input

vector, and then added the attention mechanism into the network structure. This method effectively improved the effect of named entity recognition in the financial field. Wang Guo-ming [4] proposed a Bi-LSTM and CRF model, namely BLaC model (Bert-bi-Lstm-attention-CRF), which takes BERT pre-trained word vector as input and integrates the Attention mechanism. Experimental results show that this model is more effective than other models in named entity recognition. He Yunqi et al. [5] improved the performance of entity recognition by combining CRF models with a series of syntactic and semantic features, and achieved good results.

B. Relationship Extraction

Relationship extraction is to retrieve the existing relationships among entities on the basis of extracting a large number of entities, which is one of the important tasks of information extraction. Currently, the mainstream entity relationship extraction methods can be divided into three categories: supervised method, statistical machine learning method and deep learning method.

For relation extraction of unstructured text, the earliest method adopted by researchers is artificial construction rules for extraction. For example, Yu Li et al. [6] adopted Bootstrapping method for relation extraction in geographical domain, which can automatically mine descriptors representing entity relations by considering various text features such as parts of speech and location. Make it suitable for large-scale relationship extraction. However, rule-based methods often require a strong background of domain knowledge, as well as in-depth observation and analysis of data, which requires a high labor cost.

The method based on statistical machine learning defines all possible categories in advance, then trains the classifier according to the annotated corpus, and uses the classifier to predict entity relationships. For example, Kambhatla [7] et al. used dependency parsing to obtain various syntactic and semantic features of the corpus, and then used the maximum entropy model to classify them, providing ideas for subsequent researchers.

With the development of deep learning, people gradually pay attention to the field of deep learning. Liu et al. [8] proposed a relationship extraction model based on dynamic long - and short-term networks, in which entity information, entity location information and semantic features were introduced. The accuracy rate reached 72.9%, the recall rate reached 70.8%, and the F1 value reached 67.9%. By integrating the attention mechanism with the improved convolutional neural network model, Zhao Pengwu [9] dynamically extracted entity relationships by focusing on contextual semantic information, and automatically extracted features from training data by using the improved convolutional neural network model. Then, through comparison and verification with multi-dimensional experimental results of different models and entity relationship extraction efficiency of different vector training sets, it is proved that the prediction accuracy and global performance of CNN+Attention model are superior to SVM, LR, LSTM, BiLSTM and CNN model.

III. INFORMATION EXTRACTION BASED ON HOW NET

A. How Net

How Net holds the idea of reductionism, that the meaning of all words can be composed of "meaning", that is, the most basic semantic unit that should not be divided. And How Net believes that meanings and words in real world scenarios can be represented with limited meanings. In How Net, semantic primitives are unique and can accurately represent the semantic information of words by marking some complex semantic relationships between them, such as subordination and upper and lower relation. Therefore, this paper adds semagrams to word representation learning, aiming to improve the quality of word representation learning through semagrams embedding.

How Net uses various relationships to express the meanings of many meanings, such as definitions and modifiers, and forms a complex hierarchy. In addition, the knowledge base mainly describes concepts, including more than 2000 semantic primitives. Knowledge System Description Language (KDML) is its special language, which can describe the relationship

between concepts and meanings in the knowledge base. Therefore, this paper proposes a knowledge extraction method based on How Net. Firstly, entity extraction is carried out by combining How Net and bidirectional long and short time memory network. Then, dependency parsing module is added to the model and the relationship between entity pairs is extracted by using it to form an information extraction framework based on How Net.

B. BiLSTM-CRF Model

The network structure of BilSTM-CRF model is shown in Figure 1, which is composed of word embedding layer, BiLSTM layer and CRF layer:

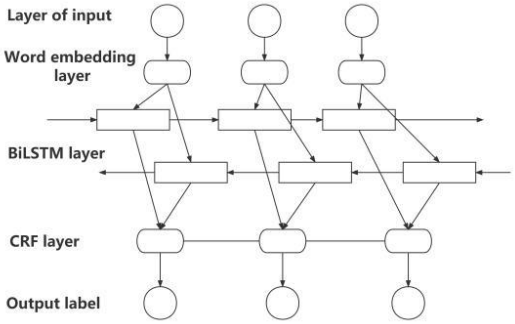

Figure 1. Network structure of BiLSTM-CRF model

1) Word embedding layer: This layer turns preprocessed words into low-dimensional dense vectors that can be processed by the model. The resulting vectors are distributed representations of words, which can represent semantic distance between words through the relationship between words and context. The process of this layer is to first train the word model with unmarked data, and then build a word vector matrix according to the word list, where v represents the size of the word list, d represents the dimension of the word vector, and then convert the text information into the id corresponding to the word list z, and map the id to the word vector matrix of the text, and s represents the number of words contained in the text.

2) BiLSTM network layer: BiLSTM is a bidirectional short and long time memory network, which is composed of forward LSTM and backward LSTM. Bidirectional text information can be extracted through BiLSTM. In this layer, the word vector matrix obtained by the word embedding layer is input into the forward LSTM and backward LSTM respectively, and then the forward LSTM output and backward LSTM output are combined according to the position to obtain the output of BiLSTM, which represents the score of each word corresponding to each category. For example, code the sentence "domestic crude steel production record high" and input "domestic", "crude", "steel", "output", "new high" forward to get five vectors $\{h_{L0}, h_{L1}, h_{L2}, h_{L3}, h_{L4}\}$. Input "New high", "output", "steel", "crude" and "domestic" successively in the back to get five vectors $\{h_{R0}, h_{R1}, h_{R2}, h_{R3}, h_{R4}\}$. And then to implicit vector before and after to the implicit vector according to the position of the splicing get $\{[h_{L0}, h_{R4}], [h_{L1}, h_{R3}], [h_{L2}, h_{R2}], [h_{L3}, h_{R1}], [h_{L4}, h_{R0}]\}$, $\{h_0, h_1, h_2, h_3, h_4\}$.

3) CRF layer: This layer is a conditional random field layer, which can "judge" the probability of the current state according to the state of the surrounding nodes. Such as X, Y is random variable, P (Y | X) is the probability distribution of under the condition of X Y, if random variable Y consisting of undirected graph G (V, E) is a markov random field, says the probability of undirected graph is a conditional random field (CRF). This layer can combine the output of BiLSTM layer as its input and take the state transition matrix as one of its parameters, and then learn and predict the label information to get the best label sequence of the text. Its calculation is shown in Equation (1).

$$score(X, y) = \sum_{i=1}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i} \qquad (1)$$

Where, A is the transfer matrix, representing the probability of transferring from the label to the label, the probability of the first word being labeled as the label, and the probability fraction of the input sentence sequence being labeled as the label sequence of is. The maximum value can be obtained to obtain the maximum label sequence of the current sample sentence.

C. Feature Selection

The words in a sentence are the basis of named entity recognition, so this paper takes the grammar and semantics of the words as the basic features.

Grammatical features are composed of words, word structures, parts of speech and morphemes, which refer to the surface features of words and sentences. Semantic features are in-depth word mining to find the correlation between named entities. This paper uses How Net as semantic features to improve the effect of named entity recognition in financial texts.

Since many words are often polysemous or have multiple words with one meaning, researchers have combined the knowledge of linguistics, cognitive science and artificial intelligence to build a semantically based dictionary to solve the non-unique semantic relationship between words. Such as the 1980 s by Princeton university George Miller team to build a large dictionary WordNet, its nouns, verbs, adjectives and adverbs in the form of a synonym set (synsets) stored in the database, the main relationship between words are synonyms. Domestic dictionaries include How Net, a large language knowledge base marked by Dong Zhendong and Dong Qiang for decades. After years of development and expansion, this knowledge base contains almost all the words used in daily life, and has extremely important application value in terms of lexical similarity calculation, text classification and information retrieval.

How Net believes that words and meanings can be described by the original meaning, and the original meaning is unique, and the semantic relationship between the meanings is also marked with complex. By adding How Net into the entity recognition model as a semantic feature, this paper can deeply explore the relationship between entities. In How Net, the record of each entry is composed of four parts: word, example, part of speech and concept, as shown in Figure 2(a). Take "net profit" as an example. Its composition in How Net is shown in Figure 2(b)

W_X = WORD

E_X = Examples of Words

G_X = Part of speech

DEF = Definition of Concepts

(a)

NO.=094557
W_C=net profit
G_C=noun [2] [jing4 li4]
E_C=
W_E=net profit
G_E=noun
E_E=
DEF={wealth|money:domain={finance | financial}, quantity = {net net |}, {earn earn | : possession = {~}}

(b)

Figure 2. How Net composition

In figure 2 (b), the "net profit" by DEF is described, in DEF "wealth | money" is the first meaning of the original "net profit", "{finance | financial}, quantity = {net net |}, {earn earn | : possession = {~}}" is the first original righteousness "wealth | money" domain limit; How Net takes the first semantic as the main meaning to express the semantic concept of words, and the semantic is to reflect the essence of the concept. Therefore, the first semanthems of entity classes and event classes in How Net are added into the named entity recognition stage of this paper as semantic features to improve the entity recognition effect.

D. Dependency syntax module

One of the factors causing noise propagation in relation extraction is that the text is long and the words describing the relation in the text are obscure, so the effective information in the sentence cannot be better learned by the relation classification model. The dominant and dominated relation between the words forming a sentence is called dependency relation, which is represented by arcs and dominated by core verbs in dependency parsing. Therefore, the syntactic structure of a sentence can be represented by a structural graph with words as nodes and relations as edges.

One way to express the interdependence between words and syntactic structure information in a sentence is to use syntactic dependency tree, which takes key verbs in a sentence as root nodes and relations as edges to form a tree-like structure, which can clearly show the logical relationship between words in a sentence. In the syntactic dependency tree, no matter how far the physical distance is, as long as there is a mutual modifying relationship between words, the distance in the tree will be very close, which can solve the problem of irregular expression.

The syntactic dependency tree is usually trained as an input feature of text, and then the corresponding weight is obtained, which can play the role of predicting relations. In this study, dependency parsing is used to extract important information from the text. In this part, only dependency subtrees containing entity pairs in the sentence are used instead of the dependency tree structure of the whole sentence, and the dependency subtrees containing entity pairs are encoded into local feature vectors, which is convenient to reduce noise and strengthen the direct relationship features of entity pairs. This study uses syntactic dependency tree for analysis, constructs a graph representation of dependency tree, and then uses the graph representation to calculate the attention factor. It uses the bidirectional long and short memory network to train the semantic features of the text, and finally adds the semantic features and attention weights to realize the extraction of inter-entity relations.

For the purpose of logical relation between words in sentences and relation extraction, this paper firstly uses tools to obtain the expression of dependency tree of text. The dependency tree takes core words as the root, and then describes the relation between words and adds them to the tree according to grammatical rules. After adding the relation between all words of a sentence to the tree, the syntactic dependency tree of the sentence is formed. Syntactic dependency can well describe the relationship between words. Traditional machine learning algorithms usually convert the syntactic dependency into a vector and then combine it with the semantic vector of the text as the input of machine learning, but this method cannot achieve effective integration with the semantic vector. Therefore, this paper uses syntactic dependency tree to generate graph representation. Then the attention is calculated by using the graph representation. Finally, the attention vector is multiplied by the semantic vector to realize the real fusion of syntactic vector and semantic vector. The specific steps are divided into the following 4 steps:

1) Generate syntax dependency tree. The syntactic analysis tree of "Synchronized development of Shanghai Pudong and legal system construction" is obtained through word segmentation and syntactic analysis, as shown in Figure 3.

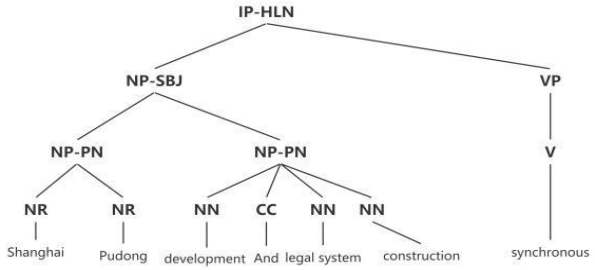

Figure 3. Syntactic analysis tree of "Shanghai Pudong Development and legal construction Synchronization"

2) According to the dependency tree results in FIG.3, the structure of directed graph is generated. In the directed graph, the specific dependency relationship in the graph is ignored, and only whether there is a relationship between words is concerned.

$$S_{ij} = \begin{cases} 1, \text{There is a dependency} \\ 0, \text{There is no dependency} \end{cases} \tag{2}$$

3) Use the two-dimensional table generated in step 2 to calculate the sentence attention distribution, as shown in Equation (3):

$$a_{t,i} = \frac{\exp(S_{t,i})}{\sum_{i=1}^{T} \exp(S_{t,i})} \tag{3}$$

4) The sentence's attention distribution is obtained, and the attention weight and the hidden vector trained by the bidirectional gated neural network are fused to calculate the semantic value, as shown in Equation (4) :

$$C_t = \sum_{i=1}^{T} a_{t,i} h_i \tag{4}$$

IV. EXAMPLE SIMULATION

The financial relationship data set in this pap er is 200,000 financial and economic news materi als crawled from the People's Internet as the corpu s, and then the BIO annotation system is adopted t o label the data set. The data set obtained in this st

udy is randomly divided into two parts, 80% of w hich are taken as the training set and 20% as the te st set.

A. *Experimental results of entity recognition*

In the feature extraction experiment, this stud y first extracts the Chinese word W_C from How Net and defines DEF (the first meaning of DEF), t hen takes W_C as the key and DEF as the corresp onding value to build the semantic dictionary requ ired by this paper, and finally uses the semantic di ctionary to conduct semantic annotation for the te xt in the financial field. In this study, HanLP was first used as a basic word segmentation tool, and t hen BiLSTM-CRF model integrated with How Ne t semantic features was used for analysis to obtain the experimental results of named entity recogniti on in this study. As shown in Table 1, feature extr action and BIO annotation results of "Internet Fin ance Association of China is registered and establi shed in Shanghai with the approval of Ministry of Civil Affairs" are obtained by using HanLp word segmentation tool and combined with How Net se mantics.

TABLE I. "CHINA... "IS VALID

Participles	The part of speech	How Net Semantics	BIO
China	ns	Local, national	B-ORG
Internet	n	The Internet	I-ORG
Finance	n	Business, money	I-ORG
Association	n	group	I-ORG
With	p	Function words	O
Ministry of Civil Affairs	nt	institutions	B-ORG
approval	v	Give one's consent	O
Shanghai	ns	place	B-ADDR
registered	v	record	O
establishe	v	To establish	O

In the experiment, Precision (P), Recall (R) a nd F1 (F-score) were used as evaluation indexes f or named entity recognition performance. As can be seen from Figure 4, this model is higher than th e other three models in terms of accuracy, recall r ate and F1 value. In the comparison experiment be tween BiLSTM model and BILSTM-CRF model, because BiLSTM model does not consider the infl uence of context and semantic environment on the label classification of words and words, the predi ction labels of each element are independent of ea ch other without dependence, and the CRF layer c an add some constraints to the final prediction lab els. These constraints can be learned automaticall y by the CRF layer from the training data set duri ng training. Therefore, the overall level of BiLST M model is lower than that of BILSTM-CRF mod el.

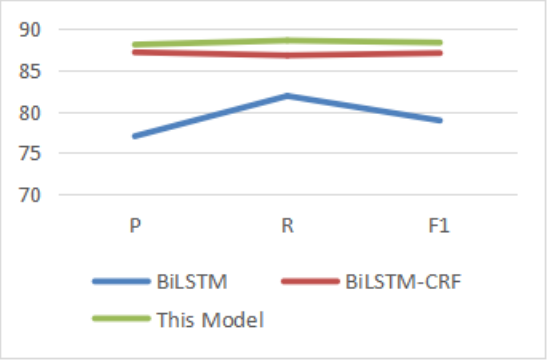

Figure 4. Comparison of P,R and F1 values

On the basis of the above advantages of the model, after adding How Net semantic features to the BiLSTM-CRF model, the overall performance is improved by about 1.2%, that is, the model can obtain more information than the comparison mo del, alleviating the imbalance between samples. T herefore, BiLSTM-CRF based on word embeddin g and How Net has achieved the best entity recog nition effect in each evaluation index.

B. *Relationship extraction results*

In this study, LTP of Harbin Institute of Tech nology was adopted as the support to build a synta ctic tree, and the machine learning training was co mpleted through Python programming. The param eters of the machine learning part were set as follo ws: Word vector dimension of training was set to 200, because this dimension was the best compreh ensive index of training time and effect. dropout v alue is 5.0; The learning rate was 0.01; Number of iterations: 200, hidden layer dimension is 200.

In this paper, Precision, Recall and F1 were u sed to evaluate the performance of the model algo rithm.

The algorithm in this paper is compared with the most classic LSTM algorithm, Bi-LSTM algo rithm and Bi-LSTM+self-ATT algorithm. Among them, Bi-LSTM+self-ATT algorithm is a bidirecti onal gated neural network integrated with self-atte ntion, which uses self-attention and attention is no t well integrated into the semantic features. The ac curacy and recall rate of the model algorithm used in this paper are improved by 5% and 1.5% respe ctively, which shows that the improvement of the algorithm is effective.

The experimental results are shown in Table 2.

TABLE II. COMPARISON OF EXPERIMENTAL RESULTS

Model	Precision	Recall	F1
LSTM	35.8	42.4	38.8
Bi-LSTM	57.5	55.4	56.4
Bi-LSTM+self-ATT	65.2	56.8	60.7
This algorithm	70.2	58.3	63.7

V. CONCLUSIONS

In this paper, a large number of entities are obtained by adding the semantic elements in How Net as features into entity recognition. Then, the syntactic dependency tree is used to add them into the relation extractor to extract the relationship between entities, and the knowledge extraction of financial texts is completed. In the aspect of entity recognition, the phrase is matched in the dictionary, then weighted by the neural network, the character vector is trained by BiLSTM, and then the conditional random field (CRF) is used to complete named entity recognition. By comparing with BiLSTM model and BilSTM-CRF model, it is proved that the algorithm improves the overall performance of named entity recognition by about 1.2%. In relation extraction, dependency syntax method and How Net semantic features are integrated to carry out relation extraction for the obtained entity pairs. The algorithm in this paper is compared with the most classical LSTM algorithm, Bi-LSTM algorithm and Bi-LSTM+self-ATT algorithm. It is proved that the accuracy and recall rate of this algorithm are improved by 5% and 1.5% respectively.

REFERENCES

[1] Wang Ning, GE Ruifang, YUAN Chunfa, HUANG Jinhui, LI Wenjie. Recognition of company names in Chinese Financial News., 2002(02):1-6.

[2] WANG S, XU R,LIU B,et al.Financial named entity recognition based on conditional random fields and information entropy[C]//2014 International Conference on Machine Learning and Cybernetics. Lanzhou: IEEE, 2014: 838-843.

[3] Qian-yu li. Financial entity oriented knowledge map construction research [D]. Yunnan university of finance and economics, 2020, DOI: 10.27455 /, dc nki. Gycmc. 2020.000492.

[4] Wang Guoming. Based on the deep study of the financial sector knowledge map construction research [D]. The northeast normal university, 2021. The DOI: 10.27011 /, dc nki. Gdbsu. 2021.000237.

[5] He Yunqi, Liu Suwen, Qian Longhua, et al. Disease Name Recognition Based on Syntactic and Semantic features [J]. Science China Information Science, 2018, 48 (11): 1546-1557.

[6] Yu L, Lu F, Liu X. A bootstrapping based approach for open geo-entity relation extraction [J].Acta Geodaetica et Cartographica Sinica, 2016, 45(5): 616-622.

[7] Kambhatla N. Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction[C]//Proceedings of the ACL Interactive Poster and Demonstration Sessions. 2004: 178-181.

[8] Liu J, Ren H, Wu M, et al. Multiple relations extraction among multiple entities in unstructured text [J]. Soft Computing, 2018, 22(13): 4295-4305.

[9] Zhao Pengwu, Li Zhiyi, Lin Xiaoqi. Chinese Character Relation Extraction and Recognition Based on Attention Mechanism and Convolutional Neural Network [J/OL]. Data analysis and knowledge discovery: 1-16 [2022-05-26]. HTTP: / / http://kns.cnki.net/kcms/detail/10.1478.G2.20220511.1654.008.html