2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD) | 978-1-6654-0527-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/CSCWD54268.2022.9776155

An Automatic Topic-oriented Structured Text Extraction Method based on CRF and Deep Learning

Hai-xia LANG[1], Yu-ying LI[1],Yan WANG[1,3*], Haibo WANG[2,3], Jianmin DONG[2,3]
[1]College of Computer Science, Inner Mongolia University,Hohhot, 010021, China
[2]Inner Mongolia Discipline Inspection Information Center, Hohhot, 010021, China
[3]Inner Mongolia Big Data Laboratory for Discipline Inspection and Supervision, Hohhot, 010021, China

***Abstract*—Automatic extraction of text information plays an important role in machine translation, knowledge mapping and other fields. In recent years, with the rapid development of computer technology and the popularization of Internet application, the resources acquired by people through the Internet show explosive growth. Facing the massive information resources, how to extract the required information quickly and effectively and convert it into structured data has become a hot topic of current research. Based on this, this paper proposes a automatic topic-oriented text extraction method combining BiLSTM and CRF models. This method firstly establishes the text extraction topic, then carries on the automatic entity recognition to the data related topic, finally forms the standard structured data, thereby realizes the unstructured text data to the specific structure of the information block, which lays the foundation for knowledge mining. Taking the data collected in ACL 2018 Chinese NER as the test data, the precision of our algorithm is 95.74%. Compared with the traditional neural network method, our text information extraction method can effectively identify more entity information in the text data, and improve its effect in practical application.**

Keywords—Information extraction, knowledge graph, Neural network, BilSTM-CRF

I. INTRODUCTION

With the continuous development and progress of computer technology, resources acquired by people through the Internet show explosive growth. How to extract the required information quickly and accurately from the ever-increasing massive data and transform it into structured data has become a hot topic in current research. Information extraction is produced and developed under this background. Information extraction, namely, extracting specific event or fact information from natural language text, helps us automatically classify, extract and reconstruct massive content, and provides data support for subsequent data mining, information retrieval and knowledge discovery services. This information usually includes entities, relationships, events, and so on. For example, times, places, and key people are extracted from news stories, or product names, development dates, and performance metrics are extracted from technical documents.

Information extraction technology is dedicated to extracting structured information from natural language text, and has been successfully applied in many fields such as information retrieval, automatic summary and text classification. It has effectively solved the problem of extracting key information quickly from massive knowledge and extracting structured data from free documents or semi-structured documents. Automatic text information extraction is an important part of text information extraction [1]. This technology automatically extracts and filters irrelevant information and integrates the valuable information contained in the text into an appropriate form. Therefore, it is of great practical significance to study how to use the existing computer technology to automatically extract important information from text for structured storage.

Based on this, this paper proposes an information extraction method based on BiLSTM and CRF model, which takes Chinese resume data as the experimental object to realize the identification of name, position, educational background and other information, providing basic applications for subsequent database construction, information retrieval, data mining and so on.

II. RELATED WORKS

Valuable structured data in text is output through information extraction techniques [2]. The target information is extracted from various texts and stored in a unified format. At present, the text information extraction methods [3] are mainly divided into three categories: information extraction method based on rules, information extraction method based on statistical learning and information extraction method based on deep learning.

A. Information extraction method based on rules

The approach based on rules relies on the study of domain knowledge, text syntax and grammar, and builds relevant rule to extract information. In order to solve the problem of automatic extraction and structured storage of bamboo germplasm data, Li et al. [4] proposed a structured method of bamboo species data based on regular extraction model. The method takes the attributes of bamboo species database as the extraction mode and uses the regularization expression to construct extraction rules. Experimental results show that the proposed model has high accuracy and can effectively extract the corresponding bamboo species information. Zhang et al. [5] proposed an information extraction method for urban rail

* Corresponding author

transit safety events, which realized semi-automatic information extraction of safety events and structured representation of their results, providing efficient data support for emergency decision-making of emergencies. Yu et al. [6] proposed a rule-based maritime free text information extraction method by constructing a custom maritime thesaurus and compiling extraction rules, which provided efficient support for safety assessment and effectiveness verification of risk measures.

However, this kind of method requires a lot of professional knowledge, and it cannot identify and extract effective information outside the dictionary or rules. The effectiveness of the method depends entirely on the richness of the dictionary, which has low recall rate and poor portability, and requires a lot of time and manpower.

B. Information extraction method based on statistical learning

In comparison with the rule-based information extraction method, the accuracy of this method is improved. The method has better scalability, and reduces the limitations of building dictionaries or rules manually. The methods based on statistical learning mainly use machine learning algorithms to extract text data and divide it into specific categories, such as names, place names, and organizations. These methods mainly include hidden Markov model (HMM), conditional random field (CRF), maximum entropy model (ME) and so on. Liang et al. [7] proposed an improved second-order HMM model for fine extraction of text information, which reasonably considered the correlation between probability and model state and effectively improved the performance of information extraction. Zhou et al. [8] used conditional random field to establish a text information extraction model, and the extraction accuracy of this model reached more than 90%, much higher than that of HMM model.

However, such methods require the selection and pre-processing of artificial features on data sets, namely feature engineering, and the quality of feature engineering will determine the effect of the model [9]. In many tasks, the collection and processing of data sets spend a lot of manpower and time, and training the model needs to provide a lot of feature engineering. Therefore, this method is inefficient and requires a large number of features to be designed manually, and its recognition performance largely depends on the accuracy of the designed features.

C. Information extraction method based on deep learning

In recent years, deep learning emerged and achieved rapid development. The method based on the deep learning can automatically select and extract data features, so it can also realize network modeling of text language, which provides a new method for natural language processing tasks, and greatly improves the performance of various tasks in language processing. Hou et al. [10] proposed a medical event extraction model based on BiLSTM, which not only avoids the disadvantage of insufficient generality of traditional machine learning methods, but also avoids the problem of information loss caused by separate classification of multi-attribute problems. Cao et al. [11] utilized the method combining CNN and CRF to extract text features from electronic medical records, which saved the complicated process of design of artificial feature and extraction, and effectively improved the accuracy rate and the recall rate. Zhang et al. [12] combined CRF and BiLSTM model to extract the entity information of tenderer and bidding agent in the text sequence of business, and the F1 score can be as high as 87.86%. Aiming at the problem of low accuracy and relatively rough identification of time and space information from micro-blogs by existing methods, Wu et al. [13] proposed a method for fine identification of time and space of emergencies in micro-blogs based on BiLSTM and CRF. This method can extract the time and space information of emergencies in micro-blogs more accurately, and provide technical support for rapid perception and accurate application of emergencies. The F1 score is 91.2%.

In view of the excellent performance of neural network models in information extraction tasks in recent years, automatic mining of hidden features can effectively solve the problem of discovering new words, while reducing the problem of artificial definition of features and over-dependence on domain knowledge. From the perspective of deep learning, this paper proposes a text information extraction model based on BiLSTM-CRF. Taking the resume text data as the experimental object, the model firstly transforms the resume text into the feature vector by embedding layer. Then it is sent to BiLSTM model for training, to dig the semantic information of entities and other words in the resume text at a deeper level, and better capture the implied information in the context. Finally, it is combined with CRF layer to solve the dependency between output tags and obtain the global optimal entity tag sequence.

III. MODEL AND ALOGIRTHM

With the development of neural network model, many studies have pointed out that applying deep-learning model to text information extraction task can achieve good results and improve the accuracy of subsequent tasks. The method in this paper fully considers the learning ability of the model and the ability to understand the text of a specific topic. By combining deep learning model-BiLSTM and random field model-CRF, it realizes the structured extraction of specific topic information from massive text information, which is called Topic-driven Information Extraction Method (T-BILSTM-CRF).

A. Overall framework of the model

The overall structure of the text information extraction model is shown in Fig.1. The first layer is the embedding layer. Using sentences from the text as units, it regards a sentence containing n characters as X, $X=(X_1,X_2,...,X_n)$. X_i represents the dictionary ID of the i-th character of a sentence. Therefore, a vector representation of each character is

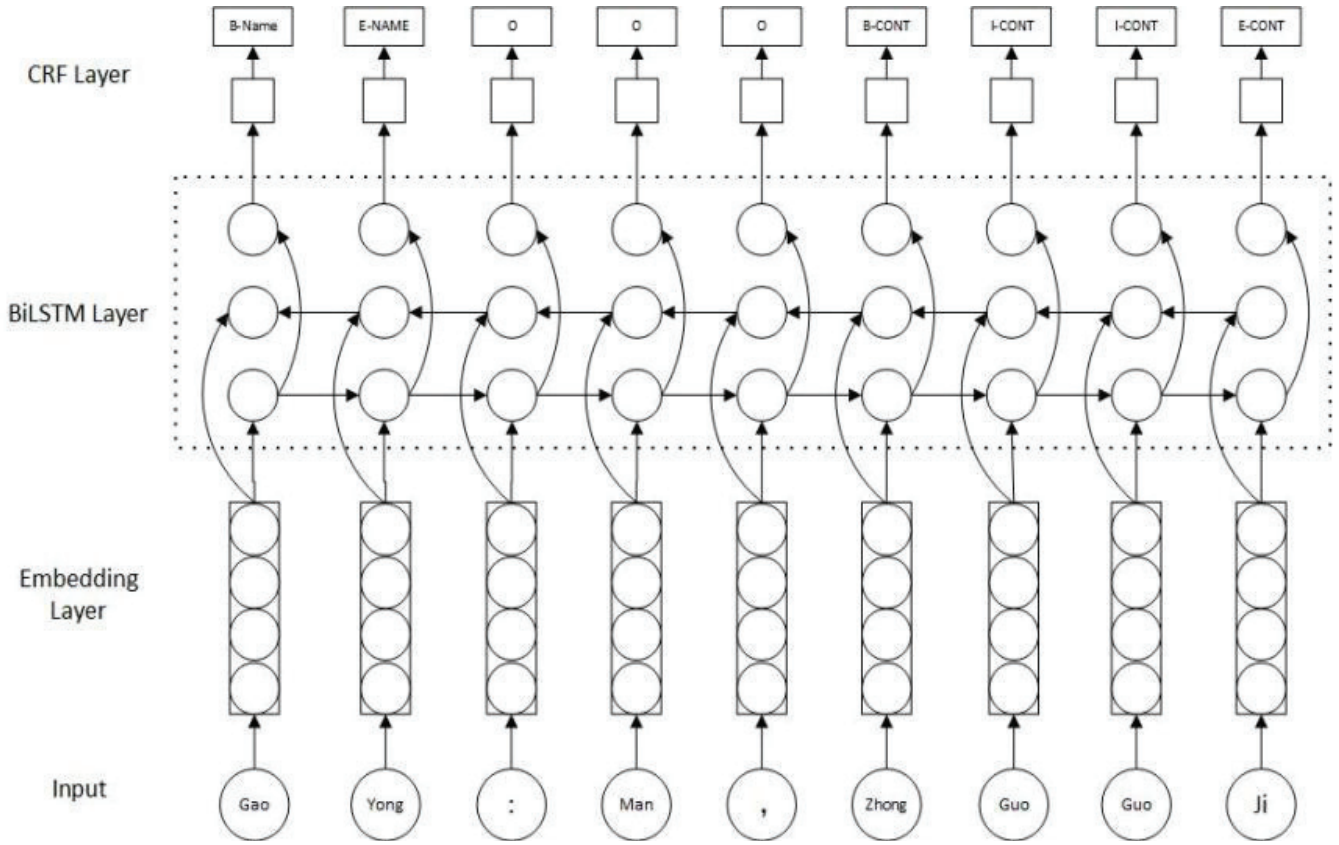

Fig.1. BiLSTM-CRF text information extraction model

obtained. The dimension is the size of the dictionary. The randomly initialized embedding matrix is used to map each character X_i to a word vector C_i with low density, $C_i \in R^d$, d is the dimension of the word vector. Then BiLSTM layer encodes the word vector sequences of the previous layer to automatically extract sentence features. It comprehensively considers forward feature extraction and reverse feature extraction, and constructs two hidden layers with opposite directions. In this way, it can better capture bidirectional semantic dependence and achieve better semantic expression effect. Finally, CRF layer is used to add constraints on labels, decode and output the prediction label sequence with the highest probability. Then the annotation type of each character can be obtained. It extracts and classifies the entities in the sequence and finally realizes the extraction of text information. The main difference between this model and other deep learning text information extraction models is the use of bidirectional long and short-term memory network. Because it has stronger contextual long distance semantic learning ability, so it can better solve the problem of polysemy, dig deeper features of text information, and provide richer semantic information for downstream tasks.

B. BiLSTM Layer

LSTM is an improved RNN, which effectively solves the problem of gradient disappearance existing in traditional RNN and realizes the effective utilization of long distance information. BiLSTM model is a combination of forward LSTM model and backward LSTM model. Since one-way LSTM can only encode data from one direction, that is, it cannot encode information from back to front, leading to understand sentence inadequately, forward LSTM and backward LSTM are combined to form BiLSTM model to fully learn semantic information between sentence contexts. The calculation process of LSTM model is shown in Fig.2.

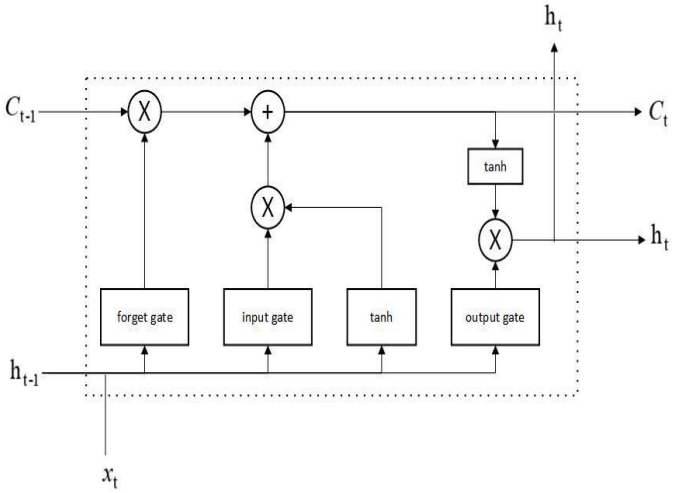

Fig.2. LSTM Unit Structure

Firstly, the LSTM model uses a forget gate to determine what information needs to be discarded in the previous cell. It receives the output of the last moment and the input of the present moment. A weight from 0 to 1 can be calculated by formula 1, indicating the change from complete rejection to complete retention.

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad (1)$$

The input gate controls the information to be added in this unit, and the calculation formula is shown in formula 2, 3 and 4.

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (2)$$

$$\tilde{c}_t = \tanh(W_c[h_{t-1}, x_t] + b_c) \quad (3)$$

$$C_t = i_t \times \tilde{C}_t + f_t \times C_{t-1} \quad (4)$$

The output gate is used to control which information is output for the current period of time, and the calculation process is shown in formula 5 and 6.

$$O_t = (W_0[h_{t-1}, x_t] + b_0) \quad (5)$$

$$h_t = O_t \times \tanh(c_t) \quad (6)$$

In the formula from 1 to 6, σ and tanh is the activation function. W and b respectively represent the weight matrix and bias vector of the three gates. x_t is the current unit input. c_t represents the state of a memory cell. $\tilde{c}_t$ represents the state at time T, which is the intermediate state obtained from the current input. Its main function is to update the state at the current time. h_t is the output at time T.

In the BiLSTM model, the input layer receives the feature vectors generated by the embedded layer, and takes the positive order and reverse order of the feature vectors as the input of the forward LSTM and backward LSTM respectively. Forward LSTM and backward LSTM are calculated according to formulas 1 to 6 respectively. Then the forward hidden state vector $\overrightarrow{h_t} = \{\overrightarrow{h_1}, \overrightarrow{h_2}, \ldots, \overrightarrow{h_t}\}$ and backward hidden state vector $\overleftarrow{h_t} = \{\overleftarrow{h_1}, \overleftarrow{h_2}, \ldots, \overleftarrow{h_t}\}$ are obtained. The final hidden state vector $h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}] \epsilon R^m$ is obtained by position splicing. Finally, the hidden state vector of the whole sentence $(h_1, h_2, h_3, \ldots, h_t) \epsilon R^{t \times m}$ is output. The hidden state vector is mapped to the *k*-dimensional space, where *k* is the size of the label set, and finally the feature matrix *P* of the whole sequence is generated, $P = (p_1, p_2, p_3, \ldots, p_n) \epsilon R^{n*k}$.

C. CRF Layer

The dependency between output prediction tags is also an important aspect of information extraction. For example, a tag that starts with "I-name" is illegal. A word can only start with B or O. B-per I-per is valid and B-per I-org is invalid. This kind of illegal situation can be avoided by using conditional random field. By adding some constraints to the predicted tags, the dependency between tags is captured by probability transfer matrix and illegal terms are excluded, then we can obtain an optimal prediction sequence.

For any given input sequence *X*, $X = (x_1, x_2, \ldots, x_n)$, the CRF comprehensive evaluation function of the corresponding label sequence Y, $Y = (y_1, y_2, \ldots, y_n)$, can be expressed by formula 7.

$$score(X, Y) = \sum_{i=0}^{n} A_{y_{i-1}, y_i} + \sum_{i=1}^{n} P_{i, y_i} \quad (7)$$

In Formula 7, A represents the transfer fraction matrix. A_{y_{i-1}, y_i} represents the score transferred from y_{i-1} to y_i. P_{i, y_i} is the score of BiLSTM layer output characteristic matrix, representing the non-normalized probability of mapping the i-th word x_i to label y_i.

The probability P(Y|X) of predicted sequence can be calculated by the following softmax function.

$$P(Y|X) = \frac{\exp(score(x,y))}{\sum_{y' \in Y_x} \exp(score(x,y'))} \quad (8)$$

The likelihood function of the prediction sequence is obtained by taking logarithms of both sides of Formula 8.

$$\ln(p(Y|X)) = score(x, y) - \ln(\sum_{y' \in Y_x} score(x, y')) \quad (9)$$

When decoding with Viterbi algorithm, y' represents the real annotation sequence, Y_x represents all possible annotation sequences, and the maximum score of the output prediction label sequence. Y* is obtained through the idea of dynamic programming, which is the final sequence annotation result of CRF layer.

$$Y^* = \underset{y' \in Y_x}{\operatorname{argmax}}(x, y') \quad (10)$$

IV. Experiments

A. Data

The resume data set collected in ACL 2018 Chinese NER Using Lattice LSTM [14] was used in the experiment. There are eight categories of entities: NAME, CONT, TITLE, ORG, LOC, RACE, EDU and PRO. During the experiment, the training set contained 127,919 characters. The test set contains 15,576 words. The validation set contains a total of 14,352 words.

B. Annotation strategy and evaluation metrics

Common labeling strategies include BIO, BIOES etc. All sentences were tagged by the BIOES in the experiment, in which B tag refers to the begining of a name, I tag refers to the continuation of a name, E tag refers to the ending of a name, O tag refers to a word not under tagging vocabulary(otherwise known as other tag), S tag refers to a single token.

Precision (P), Recall (R) and F1 scores are used as evaluation metrics in the experiment, and the calculation formula is as follows.

$$P = \frac{\text{The correct number of entities identified}}{\text{The number of all entities identified}} \times 100\% \quad (11)$$

$$R = \frac{\text{The correct number of entities identified}}{\text{The number of entities labeled}} \times 100\% \quad (12)$$

$$F1 = \frac{2PR}{P+R} \times 100\% \quad (13)$$

C. Experimental environment and parameter configuration

The environment configuration in the experiment is shown in Table 1.

TABLE 1. ENVIRONMENT CONFIGURATION

Configuration	Specification
OS	Windows10
CPU	Intel(R) Core(TM) i5-6200U CPU @ 2.30GHz 2.40GHz
Memory	8GB
Python	3.7.3
Pytorch	1.5.0

Adam optimizer is used during the training, and the specific parameter configuration is shown in Table 2.

TABLE 2. PARAMETER SETTING

Parameter	Value
batch_size	64
epoch	30
learn_rate	0.001
optimizer	Adam
Embedding_size	128
Hidden_size	128

D. Experimental results and comparative analysis

The results of T-BilSTM-CRF model on the Chinese resume data set are shown in Table 3. It can be seen from Table 3, the method proposed in the paper has high precision in the all types of entities, and the model achieves good results in combination with contextual semantic information. The F1 score of nationality identification is the highest. The model has low precision, recall rate and F1 score for organization, professional recognition. The reason may be that there are a lot of nesting among entities. For example, "Peking University" is marked as the name of organizational structure, and the model identifies "Beijing" as a place name. The nesting is the main factor affecting the accuracy of the model.

In order to verify the validity of T-BilSTM-CRF model, a comparative experiment was set up in this paper. The comparison between the experimental results of the model and those of other models is shown in Table 4.

TABLE 3. EXPERIMENTAL RESULTS

Entity Class		P（%）	R（%）	F1（%）
NAME	B-NAME	99.06%	93.75%	96.33%
	I-NAME	94.05%	96.34%	95.18%
	E-NAME	100%	99.11%	99.55%
CONT	B-CONT	100%	100%	100%
	I-CONT	100%	100%	100%
	E-CONT	100%	100%	100%
TITLE	B-TITLE	94.16%	91.97%	93.05%
	I-TITLE	95.30%	87.51%	91.24%
	E-TITLE	98.96%	98.32%	98.64%
ORG	B-ORG	95.50%	96.02%	95.76%
	I-ORG	95.32%	96.12%	95.72%
	E-ORG	91.56%	90.24%	90.89%
LOC	B-LOC	100%	83.33%	90.91%
	I-LOC	100%	80.95%	89.47%
	E-LOC	100%	83.33%	90.91%
RACE	B-RACE	100%	92.86%	96.30%
	E-RACE	100%	100%	100%
EDU	B-EDU	97.32%	97.32%	97.32%
	I-EDU	95.98%	93.30%	94.62%
	E-EDU	99.08%	96.43%	97.74%
PRO	B-PRO	83.33%	90.91%	86.96%
	I-PRO	83.56%	89.71%	86.52%
	E-PRO	91.18%	93.94%	92.54%

TABLE 4. COMPARISON RESULTS OF DIFFERENT MODELS

Model	P（%）	R（%）	F1（%）
HMM	91.49%	91.22%	91.30%
CRF	95.43%	95.43%	95.42%
BiLSTM	95.59%	95.58%	95.55%
T-BiLSTM-CRF	95.74%	95.72%	95.70%

It can be seen from Table 4 that the F1 score of the model proposed in this paper is 4.4% higher than that of the HMM model and 0.28% higher than that of the CRF model. Compared with all the models in Table 4, T-BilSTM-CRF model has the highest F1 score.

V. CONCLUSION

To sum up, this paper proposes an automatic text information extraction method based on deep learning and CRF for text information extraction task. By comparing the experimental results, it can be concluded that this method can effectively improve the precision of structured extraction of text information for a specific topic. The effect of extraction is

better than the traditional neural network model, and more entity information can be identified, which verifies the effectiveness of the method. The next focus of this paper will be to increase the size of the training data set, train a larger text information extraction model, and further improve the generalization ability of the model, so as to build a richer text information corpus.

ACKNOWLEDGMENTS

This work was supported in part by Natural Science Foundation of China under Grants 62162047, Natural Science Foundation of Inner Mongolia under Grants 2019ZD15, 2019MS06029, Inner Mongolia Science and Technology Plan Project under Grant 2019GG372, 2021GG0155, the Open-topic of Inner Mongolia Big Data Laboratory for Discipline Inspection and Supervision (IMDBD2020012, IMDBD2021014), the Self-topic of Engineering Research Center of Ecological Big Data, Ministry of Education,Inner Mongolia Engineering Laboratory for Cloud Computing and Service Software, Inner Mongolia Key Laboratory of Social Computing and Data Processing, Inner Mongolia Engineering Lab of Big Data Analysis Technology.

REFERENCES

[1] Pedro Domingos, Geoff Hulten. Mining high-speed data streams[P]. Knowledge discovery and data mining, 2000.

[2] Sachi Nandan Mohanty, E. Laxmi Lydia, Mohamed Elhoseny, Majid M. Gethami Al Otaibi, K. Shankar. Deep learning with LSTM based distributed data mining model for energy efficient wireless sensor networks[J]. Physical Communication,2020,40:

[3] Li Xiangyang, MIAO Zhuang. Techniques of Information Extraction from Free Text[J]. Information Science,2004(07):815-821+829.

[4] Li Xin, Li Shaowen, Xu Gaojian, Lin Jianbin. Research on bamboo species data structure method based on regular extraction[J]. Computer technology and development,2018,28(06):147-150+155.

[5] Zhang Meng, Chen Jiahui, Sun Ranran, Li Xiaolu, Zhu Guangyu. Rule-based Urban Rail Traffic Safety Event Information Extraction and Knowledge Element Representation[J]. Science Technology and Engineering, 201,21(15):6435-6440.

[6] Yu Chen, MAO Zhe, Gao Song. Research on maritime free text information extraction method based on rules[J]. Traffic information and safety,2017,35(02):40-47.

[7] Liang Jiguang, TIAN Junhua, XIONG Ling. Research on information extraction based on second-order HMM[J]. Journal of information science,2011,30(07):169-171+141.

[8] Zhou Jing, Wu Junhua, Chen Jia, Chen Shenyan. Text Information Extraction Based on CONDITIONAL Random Field CRF Model[J]. Computer Engineering and Design,2008(23):6094-6097.

[9] Zeng Yong. A new approach to Chinese named entity recognition based on BilstM-CRF model[D],2020.

[10] Hou Weitao, JI Donghong. Medical event recognition based on bi-lstm [J]. Application research of computers,2018,35(07):1974-1977.

[11] Cao Yiyi, Zhou Yinghua, SHEN Fahai, Li Zhixing. Research on named entity recognition of chinese electronic medical record based on CNN-CRF[J]. Journal of chongqing university of posts and telecommunications (natural science edition),2019,31(06):869-875.

[12] Zhang Yingcheng, Yang Yang, Jiang Rui, Quan Bing, Zhang Lijun, Ren Xiaolei. Business entity recognition model based on bilstm-crf[J]. Computer engineering,2019,45(05):308-314.

[13] Wu Jianhua, Hu Lieyun, Zhao Yu, Dai Peng, Xiong Jiaqi. Bilstm-crf and Classification and hierarchical annotation based on the temporal and spatial information identification of emergencies in microblog[J]. Geography and Geo-Information Science, 201,37(03):1-8.

[14] Yue Z , Jie Y . Chinese NER Using Lattice LSTM[C]// The 56th Annual Meeting of the Association for Computational Linguistics (ACL). 2018.