7-8-2023

# A Deep Learning Entity Extraction Model for Chinese Government Documents

Hao Ding
*Nanjing University*, 415058975@qq.com

Xuwen Li
*Nanjing University*, 1325730950@qq.com

jialin du
*Nanjing University*, jialin_du@smail.nju.edu.cn

Guangwei Hu
*Nanjing University*, hugw@nju.edu.cn

# A Deep Learning Entity Extraction Model for Chinese Government Documents

*Completed Research Paper*

**Hao Ding**
School of Information Management
Nanjing University
Nanjing, China
dinghao@smail.nju.edu.cn

**Xuwen Li**
School of Information Management
Nanjing University
Nanjing, China
1325730950@qq.com

**Jialin Du**
School of Information Management
Nanjing University
Nanjing, China
jialin_du@smail.nju.edu.cn

**Guangwei Hu**
School of Information Management
Nanjing University
Nanjing, China
hugw@nju.edu.cn
(Corresponding author)

## Abstract

*In this paper, we propose a combined Whole-Word-Masking based Robustly Optimized BERT pretraining approach with dictionary embedding entities recognition model for Chinese documents. By using multiple feature vectors generated by such as Roberta and domain dictionaries as embedding layers, the contextual semantic information of the text is fully considered. Meanwhile, Bi-directional Long Short-Term Memory(BiLSTM) and a multi-head attention mechanism are used to learn the information of long-distance dependency of the text. We use conditional random field(CRF) to obtain the global optimal annotation sequence, which is expected to improve the performance of the model. In this paper, we conduct comparison experiments with five baseline-based methods in the official document dataset of government affairs domain. The Precision of the model is 91.8%, Recall is 90.5%, and F1 value is 91.1%, which are better than other baseline models, indicating that the proposed model is more accurate for recognizing named entities in government documents.*

**Keywords:** natural language processing; government official documents; information extraction; named entity recognition

## Introduction

In recent years, with the rapid development of information technology, the synchronization of online and offline office of government affairs has become the mainstream trend (Zhao et al., 2022). Facing the quantity and variety of governmental documents, how to extract and classify keywords among is a key research direction in the field of modern library intelligence (Tang et al., 2021; Zhang et al., 2022) and also a hot research direction in the field of Natural Language Processing (NLP) (Baksa et al., 2016;Catelli et al., 2021). Named Entity Recognition (NER) is a key technology in the fields of natural language processing (Fan et al., 2019), knowledge graph construction and recommender systems (Xie et al., 2020; Ding et al., 2022; Li et al., 2022; Zhang et al., 2022), whose main task is to identify valuable nouns or phrases from unstructured domain texts and to classify and label them (Liu et al., 2021). Although existing NER techniques have worked well in some domains, entity naming recognition applied to government affairs still faces difficulties. Firstly, text data in the field of government affairs is difficult to obtain, especially the

relatively scarce datasets that have been annotated, which makes it difficult for the model to achieve expected recognition performance (Zhao et al., 2022). Secondly, the terminological features of the naming style of government entities are distinctive and the naming structure is complex while some lack standardized naming rules and even have the phenomenon of multiple meanings of words, thus it is difficult to perform text processing operations such as word separation, annotation, and classification on the government corpus (Tang et al., 2019). Furthermore, unlike English NER task, where each word is separated by a space, it is difficult to accurately identify. Chinese entities because of the large amounts of long texts in Chinese publications and the lack of obvious boundaries between words (Li et al., 2021; Cui et al., 2020).

To solve the above problems, we propose a deep learning model for naming entities of Chinese publications in the domain of government affairs. The model uses a multiple feature representation model (Roberta-WWM-Dict) of text to construct word vectors, and uses Bi-LSTM model and multi-head attention mechanism (MHA) to extract the contextual semantic features of text, and finally uses CRF model to obtain the best prediction sequence to label entities.

## Related Work

With the development of computer software and hardware, NER research based on deep learning has been carried out successively and has gradually become the main method for solving natural language processing (NLP) problems (Qiu et al., 2019). Compared with the previous manual methods of constructing rules based on data features, containing more semantic information (Wang et al., 2020), deep learning models can reduce human intervention, and are suitable for solving serial annotation problems such as named entity recognition. Deep learning focuses more on the construction of the overall neural network model and the optimization of parameters. Some researchers used Word2vec tool to obtain word vectors as the input of the model (Jiang et al., 2021). Wang et al. used the Skip-gram model in Word2vec to pre-train the labeled corpus and transform it into word vector sequences by word embedding layer, then input the dependency information learned by Bi-LSTM (Wang et al., 2020), and finally used CRF layer to obtain the global optimal sequence. Cao et al. (2019) proposed a CNN-CRF-based named entity recognition model for Chinese electronic medical records, which uses iterative expansion convolution to process the input vectors and dropout methods to randomly discard connections, and applied CRF to correct the classification results of the network to extract five categories of entities from medical records: diseases, symptoms, body parts, examinations and treatments. However, the word vectors generated by Word2vec are static and have only a single representation, which cannot solve the problem of multiple meanings of words well.

To better extract textual feature information, techniques such as Transformer and BERT are widely used in NER tasks. Hu Wei et al. (2022) proposed a named entity recognition method for TCM medical cases based on BERT-Bi-LSTM-CRF model. The method uses BERT for text feature extraction, then Bi-LSTM algorithm to obtain the information of the context, and finally outputs the results of named entities of TCM medical cases by CRF algorithm. Lin Litao et al. (2023) constructed a canonical animal named entity recognition model based on the Siku-BERT pre-training model and validated the effectiveness of the method. Zhao et al. (2021) used domain knowledge to generate character-level candidate entities and modeled the global interdependencies among these entities based on the reference graph model. They jointly embedded the latest BERT-based character vectors and character-level candidate entities into a deep learning model, and demonstrated the effectiveness of the model using Chinese car review data as an experimental dataset. Yu et al. (2022) used a bi-directional encoder of Transformers (BERT) to extract entities from Chinese mineral literature, and integrated the transfer matrix of the conditional random field algorithm to improve the sequence tagging accuracy, and the results showed that the model can effectively identify seven classes of mineral entities.

Although the above methods alleviate the problems faced by entity naming recognition to a certain extent, they are mostly applied in one of specific domains. In order to reduce the impact of training set data quality on the recognition performance of Chinese government official document named entities and solve the problem of not being able to obtain word level semantic representation, this paper introduces the pre training model RoBERTa-WWM in named entity recognition (Liu et al., 2019). This model can obtain prior semantic knowledge from a large number of unlabeled texts to enhance semantic representation, and obtain word level semantic representation in pre-training. Meanwhile, in order to take into account

contextual information, solve the semantic connection between distant contexts in text statements, and solve the problem of gradient vanishing and exploding in the RNN model during entity recognition, this paper introduces the deep learning model BiLSTM-CRF to improve the accuracy of named entity recognition using contextual information. Therefore, combining the above two models, this paper proposes a method for Chinese government document named entity recognition based on Robertsa-WWM-Dict.

The main work of this paper is as follows.

- We construct a dictionary based on expertise in the government domain to create embedding vectors to improve model recognition performance.
- We use the constructed dictionary-enhanced Roberta-WWM-Dict embedding model, further integrated with Bi-LSTM and multi-head attention mechanism to improve the accuracy of entity extraction for government official documents.
- We construct a training corpus of governmental documents and compare it with several baseline models to verify the effectiveness of the proposed method.

## Methodologies

In this section, the various parts of the model and their implementation details are described systematically, and the overall framework is shown in Figure 1.

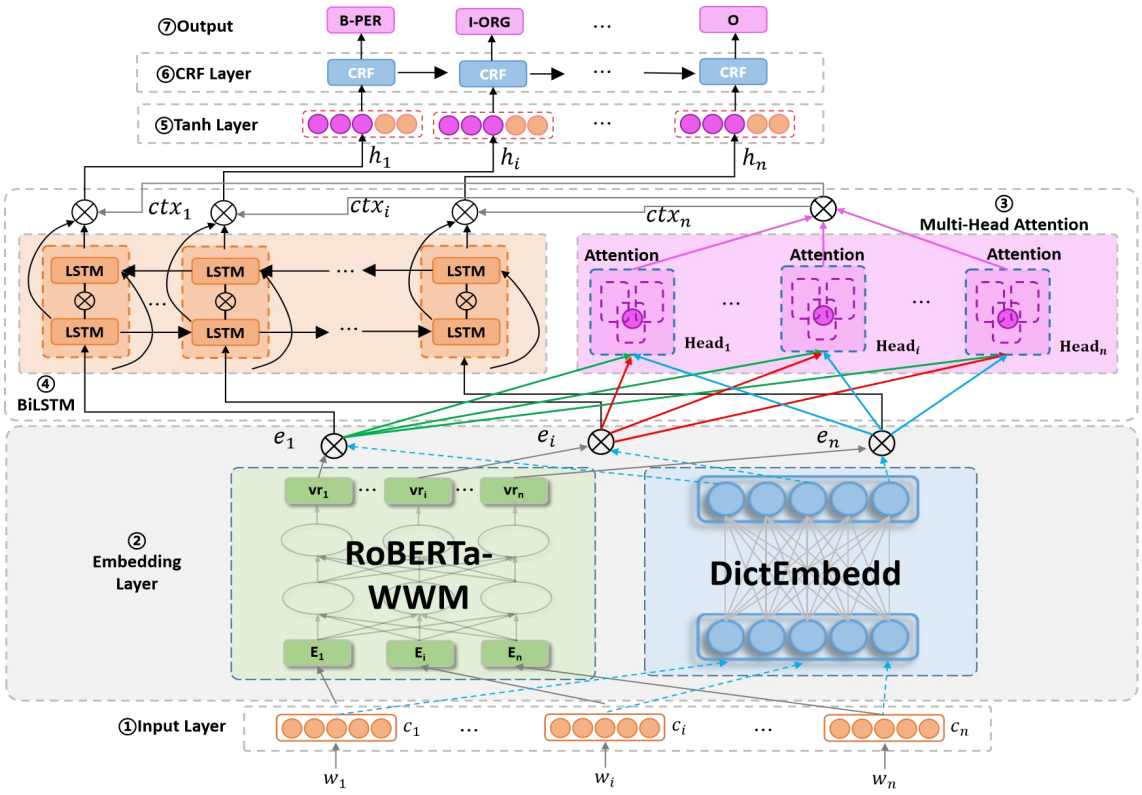


**Figure 1. Modified Research Model**

### *Input layer*

The input layer is used to encode the pre-processed texts into vectors that can be processed by the model. Given a vector of Chinese text sequences $W = [w_1, w_2, \cdots, w_n]$ , where n is the dimensionality of the words

in the training corpus. The input layer represents each word $w_i$ as a one-hot coding $C = [c_1, c_2, \cdots, c_n]$ and sends it to the embedding layer.

### *Multi-feature Embedding Layer*

Since solo-heat vector coding leads to an oversized feature space, complex composition structure of government named entities, and strong contextual association between words, direct use of solo-heat vector coding does not represent the semantic relationships between texts well. Therefore, the embedding layer in this paper maps them to more dense embedding vectors by RoBERTa-WWM model and dictionary embedding.

## RoBERTa-WWM based word embedding vector

In this paper, RoBERTa-WWM is selected to obtain the semantic representation. The RoBERTa-WWM model not only inherits the advantages of the BERT model, expressing the input sentences as the sum of word vectors, sentence vectors, and position vectors, but also uses a larger number of single training samples and more data to train the model. It also removes the Next Sentence Prediction target function and trains with a longer sequence length. At the same time, the model uses a dynamic masking mechanism to learn different language representations, In the pre-training stage, Chinese full word masking technology is used to mask all Chinese characters that make up the same word, making it more suitable for the named entity recognition task of Chinese government documents.

Static masking in BERT is to select 15% of the tokens randomly for each sequence and replace them with [MASK], and the masked tokens do not change during the pre-training process. Dynamic masking is used in RoBERTa-WWM, and the masked words are the tokens reselected in each iteration cycle. The pre-training task of the WWM-based masked language model helps to capture semantic features at the Chinese word-level, thus improving the overall performance of the model. In addition, RoBERTa-WWM deletes in the pre-training phase in addition to the NSP task, which extends the length of inputtable individual sentences to 512 characters. Eventually, the label of each word is mapped to the embedding vector $vr = [vr_1, vr_2, \ldots, vr_n]$.

## Dictionary embedding based word embedding vector

Dictionary feature embedding refers to mapping each combination of characters and their corresponding dictionary feature labels to a feature embedding vector, so that the same words mapped to different labels can represent different entities and improve the recognition success rate of terms specific to the government domain. In this section, we use the bi-directional maximum matching (Bi-MM) algorithm and dictionary mapping to segment the sentences to obtain the segmented entity sequence embedding vector, and integrate with the B-I-O tagging scheme, each character will get a feature label. The entity categories in the dictionary mainly include the names of government personnel, organizations, locations, policy documents, etc. These terms are extracted from e-government terms [GB/T 25647-2010, GB/T 34078.1-2017], e-government standard guidelines [GB/T 30850-2014, parts 1-5] and other sources, such as official government websites, Sogou cell thesaurus, etc. In addition, each English term corresponds to a Chinese name, so that in the recognition process, the English term is directly mapped to the corresponding Chinese label, preventing ambiguity in the recognition process of texts in different languages. Eventually, the label of each word is mapped to the lexical embedding vector $vs = [vd_1, vd_2, \ldots, vd_n]$ .

Finally, through each vector$vr \in \mathbb{R}^K$ , the $vd \in \mathbb{R}^K$ is connected to the final embedding vector $e_i \in \mathbb{R}^{2K}$ , the embedding layer outputs the embedding vector matrix $E = [e_1, e_2, \cdots, e_n]$ , where $e_i$ is expressed in the form shown in equation (1).

$$e_i = vr_i \oplus vd_i \quad (1)$$

### *Bi-LSTM Feature Extraction Module*

The LSTM model integrated with memory unit and gate control not only achieves long-term memory and can capture text sequence features, and its model structure is shown in Fig. 2. Bi-LSTM not only inherits

the advantages of LSTM, but also takes into account the subsequent states on this basis. Therefore, this paper uses the Bi-LSTM model to simultaneously consider the contextual information of each character in the text sequence and merge the outputs at the same moment to obtain more comprehensive semantic features of the text.

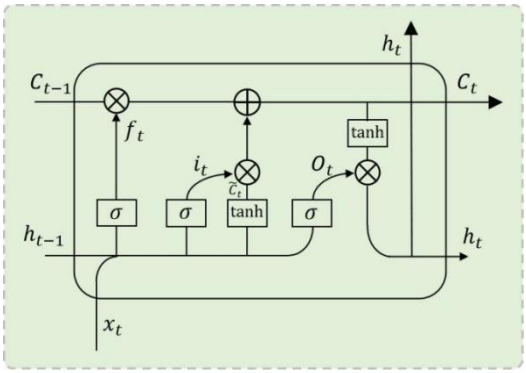


**Figure 2. LSTM model structure diagram**

Using the government text as the input sequence, after receiving the input vector matrix E from the multiple feature embedding layer, the module returns the LSTM unit to output the sequence vector $h_t = (h_1, h_2, \cdots, h_n)$ at moment t. The input vector matrix E contains n embedding vectors, and each embedding vector is represented as a $K$ dimensional vector set, which contains the corresponding location feature vector $v_p$ , the contextual semantic feature vector $v_s$ and its character embedding vector $v_w$ . At each moment t, the LSTM uses the input vector $e_t$ and the previous hidden state $h_{t-1}$ to calculate the current hidden state $h_t$. The specific implementation is shown in equation. (2)-(7).

$$i_t = \sigma(W_{ei}e_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \quad (2)$$

$$f_t = \sigma(W_{ef}e_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \quad (3)$$

$$z_t = \tanh(W_{ec}e_t + W_{hc}h_{t-1} + b_c) \quad (4)$$

$$c_t = f_t c_{t-1} + i_t z_t \quad (5)$$

$$o_t = \tanh(W_{eo}e_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \quad (6)$$

$$h_t = o_t \tanh(c_t) \quad (7)$$

Where $\sigma$ is the activation function, $b$ is the bias vector, and $W$ is the weight matrix, and $h$ denotes the hidden state, and i denotes the input gate, f denotes the forgetting gate, and o denotes the output gate, and $c_t$ denotes the update state at time t, and $z_t$ denotes the information to be added. While Bi-LSTM can integrate the outputs of the same moment, so for each t moment, it corresponds to the computation of forward and backward information. The final hidden state $h_t$ of the representation is shown in equation (8).

$$h_t = \overrightarrow{h_t} \oplus \overleftarrow{h_t} \quad (8)$$

In equation (8), the $\overleftarrow{h_t}$ denotes the forward LSTM, $\overrightarrow{h_t}$ denotes the backward LSTM, the whole Bi-LSTM can be used on each input $e_t$ The contextual information of the input sequence is obtained on each input.

## *Multi-Head Attention Mechanism*

As the length of text sequences increases, the contextual environment may not be able to preserve enough semantic information at once, which can cause Bi-LSTM to lose a large amount of important information required for the recognition task. Therefore, finding out the relationships between word entities and their contextual semantic features in text sequences is important for accurate named entity recognition. In this section, we introduce a multi-head attention mechanism as a supplement to Bi-LSTM to capture text semantic features from three levels: word, phrase and sentence to further improve the performance of named entity recognition. The multi-head attention mechanism can capture context-sensitive information

in several different subspaces to better understand the sentence structure in order to improve the performance of the NER task for non-canonical text. The multi-head attention mechanism is shown in Figure 3.

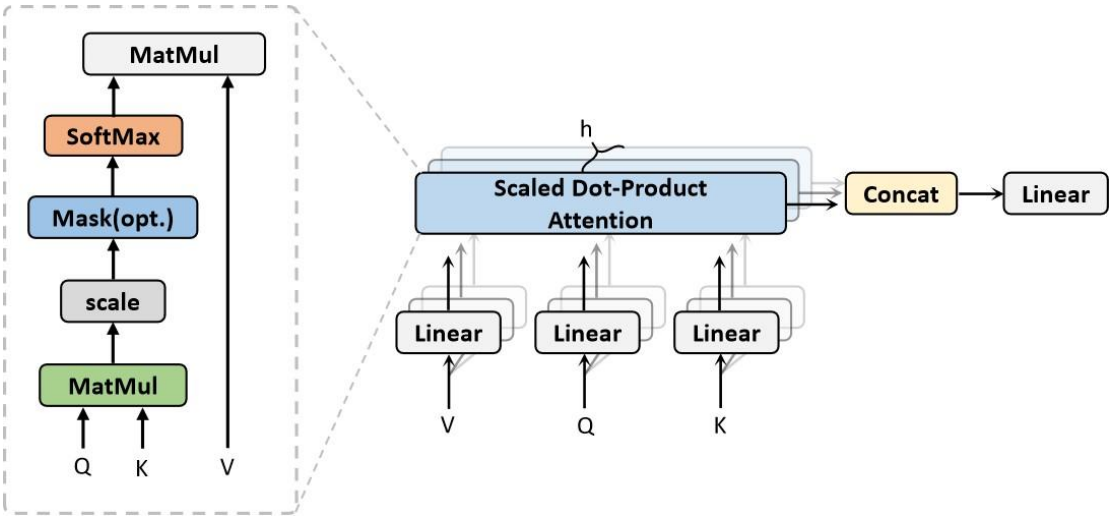


**Figure 3. Schematic diagram of scaling dot-product (left) and multi-head attention mechanism (right)**

The attention mechanism maps the input $H = [h_1, h_2, \ldots, h_n]$ representation of the query space Q, key space K and value space V to the output vector mapping process. The multi-head attention module is the process of projecting the query $Q = [Q_1, Q_2, \ldots, Q_m]$ , key-value space $(K, V) = [(k_1, v_1), (k_2, v_2), \ldots, (k_n, v_n)]$ with different parameter matrices m times for linear transformation, and then inputting to m parallel heads respectively and calculating the results for the attention function. The specific process is as follows.

Setting $a_{ij}$ represents the attention weight of the ith query vector $q_i$ and the jth input information $k_j$, and $ctx_i$ represents the attention output vector calculated from the query vector $q_i$, then $ctx_i$ is calculated as shown in equations (9)-(10).

$$a_{ij} = \text{softmax}(s(k_j, q_i)) = \frac{\exp(s(k_j, q_i))}{\sum_{d=1}^{n} \exp(s(k_d, q_i))} \quad (9)$$

$$ctx_i = \sum_{j=1}^{n} a_{ij} \cdot v_j \quad (10)$$

In equations (9) and (10), the s denotes the vector dot product, i.e. $s(h, q) = h^T q$ , and $v_j$ denotes the jth input value. Finally, the results of all query vectors are stitched together as the final result as shown in equation. (11).

$$ctx = \text{Concat}(ctx_1, ctx_2, \ldots, ctx_m) \quad (11)$$

In equation (11), the Concat denotes the vector splicing function.

## *Annotation layer*

The merged annotation layer contains the Tanh layer and the CRF layer. the Tanh layer is used to activate the output of the Bi-LSTM and the multi-headed attention layer, and is sent as input to the CRF layer, which can be expressed as shown in equation (12).

$$r_t = \text{Tanh}(ctx_t \oplus h_t), 1 \le t \le n \quad (12)$$

Although Bi-LSTM can effectively handle long-range text information in entity naming task, it cannot handle the dependency relationship between neighboring text labels. The use of CRF can compensate for the shortcomings of Bi-LSTM by obtaining the optimal prediction sequence through the relationship of

neighboring labels. In this section, we use the B-I-O named entity recognition annotation system, which includes "O", "B-PER", "I-PER", "B -LOC", "I-LOC", "B-ORG", "I-ORG", "B- POL" "I-POL". Where "B" indicates the entity name header, "I" indicates the entity name body, and "O" indicates the non-entity part. The entity type is represented as follows: "PER" for Person, "LOC" for Location, "ORG" for Organization, "POL " denotes the policy file name, and the CRF markup layer uses the annotation system to identify the named entities and output the results. After setting the input observation sequence $R = [r_1, r_2, \ldots, r_n]$ corresponds to the output annotation sequence $Y = [y, y_2, \ldots, y_n]$ , the CRF layer calculates the joint probability distribution of the whole sequence given the observation sequence to be annotated, and finally outputs a globally optimal annotation sequence, as shown in equation (13).

$$S(R, Y) = \sum_{i=1}^{n} (R_{i,y_i} + A_{y_i,y_{i+1}}) \quad (13)$$

In equation (13), the $R_{i,y_i}$ is the probability of the ith word in the sequence is predicted to be the label $y_i$. $A_{y_i,y_{i+1}}$ is the fraction generated by the process of the label $Ay_i$ in the transfer matrix transfers to the label $y_{i+1}$.

# Experiment and analysis

## *Dataset settings*

The original Chinese corpus in this paper is composed of texts crawled by web crawlers from government websites, microblogs and other public document publishing platforms. Since there are many interfering information in the original text, such as tables, pictures, special symbols, etc., which may affect the accuracy of recognition, and the text is too long that may also lead to the degradation of recognition accuracy, we need to perform data cleaning and long sentence cutting process on the text data. We first use the NLTK toolkit in Python to perform data cleaning on the official text corpus data, using regular expressions to replace all special symbols except letters and punctuation marks with a single space, and convert all words to lowercase to avoid ambiguity in the semantic extraction process. Then the long text is divided into standard single sentences by identifying sentence break symbols such as period and question mark. Finally, a combination of machine pre-processing as well as manual review and error correction annotation is used to generate a corpus of a total of 1800 government texts and 10,500 sentences, with a total of 4 categories of government entities and a total of 10,231 government entities, the training set, validation set and test set are divided according to 8:1:1, and the scale is shown in Table 1.

| East Distance | West Distance | Count |
|---|---|---|
| Person | B-PER,I-PER | 3042 |
| Location | B-LOC,I-LOC | 3364 |
| Organization | B-ORG,I-ORG | 890 |
| Policy | B-POL,I-POL | 872 |

**Table 1. Size of datasets**

## *Metrics and settings*

This paper uses Precision, Recall and F1 values as the main experimental evaluation metrics, and each metric is defined as shown in equations (14)-(16).

$$\text{Precision} = \frac{\text{TruePosit}}{\text{AllPosit}} \quad (14)$$

$$\text{Recall} = \frac{\text{TruePosit}}{\text{RecoPosit}} \quad (15)$$

$$\text{F1} = \frac{2 \times P \times R}{(P + R)} \quad (16)$$

Where TruePosit denotes the number of correctly identified government entities, AllPosit denotes the actual total number of government entities in the dataset, and RecoPosit represents the number of identified government entities.

The experiment used an Adam optimizer and a 12 layer Transformer in the RoBERTa-WWM-model; To prevent overfitting, Dropout is used in the input and output of BiLSTM, and the value is 0.5.

### *Baseline comparison experiment*

In this section, we compared the nine baseline models with our model using experimental results respectively under three metrics Precision, Recall, and F1. The IDCNNs-CRF model (Strubell et al., 2017) is based on CNN improvement with the addition of null convolution feature. The ERNIE model (Sun et al., 2019) is based on continuous learning semantic understanding pre-training framework using multi-task learning incremental construction of pre-training task. The 2nd and 3rd groups of comparison models, on the other hand, use Bert as embedding integrated with other modules, and the RoBERTa model is mainly improved for the Bert training model.

| No. | Models | P | R | F1 |
|---|---|---|---|---|
| 1 | IDCNNs-CRF | 0.814 | 0.813 | 0.813 |
| 2 | Bert+CRF | 0.873 | 0.870 | 0.871 |
| 3 | Bert+BiLSTM+Attention+CRF | 0.885 | 0.882 | 0.883 |
| 4 | ERNIE+BiLSTM+CRF | 0.901 | 0.899 | 0.899 |
| 5 | RoBERTa+BiLSTM+Attention+CRF | 0.908 | 0.906 | 0.907 |
| 6 | RobertaWWM+Dict+BiLSTM+MHA+CRF | 0.918 | 0.905 | 0.911 |

**Table 2. Comparative experiment of different models**

As can be seen in Table 2, groups 1-2 did not use Bi-LSTM with the attention mechanism, and the overall effect was poor. Groups 3-6 all use Bi-LSTM or attention mechanism, and the effect is significantly improved compared with the models of groups 1-2. Group 6 is the model proposed in this paper, which fully integrates the multiple semantic features of characters, word position semantics, and lexical semantics of the text. It not only extracts the text bi-directional semantic features by BiLSTM, but also introduces the multi-headed attention mechanism also improves the accuracy of the semantic division of sentences, and has stronger generalization ability for governmental text. Therefore, the recognition effect of the model proposed in this paper is better than other models.

### *Ablation experiment*

In order to further verify that the performance improvement of the proposed model comes from the improvement part, comparative experiments of model ablation are respectively conducted in this section under nine different sets of improvement combinations. The results of the ablation experiments under each index are shown in Table 3.

| No. | Models | P | R | F1 |
|---|---|---|---|---|
| 1 | Dict+BiLSTM+MHA+CRF | 0.836 | 0.804 | 0.819 |
| 2 | RobertaWWM+BiLSTM+MHA+CRF | 0.842 | 0.827 | 0.834 |
| 3 | RobertaWWM+Dict+BiLSTM+CRF | 0.895 | 0.878 | 0.886 |
| 4 | RobertaWWM+Dict+MHA+CRF | 0.906 | 0.892 | 0.898 |
| 5 | RobertaWWM+Dict+BiLSTM+MHA+CRF | 0.918 | 0.905 | 0.911 |

**Table 3. Ablation experimental results**

The experimental models in groups 1-2 consider word embeddings of only one feature, and groups 3-4 consider word embeddings of all features. The ablation model in group 4 does not fuse the BiLSTM

module compared to group 3, while the ablation model in group 3 does not introduce the multi-headed attention mechanism compared to group 4. Group 5 is the multiple text feature embedding model in this paper, which considers the fusion of all methods. Experiments were conducted on these nine ablation models separately under three metrics, Precision, Recall, and F1.

As shown in Table 3, the RobertaWWM+Dict+Bi-LSTM+MHA+CRF model in this paper has the best results compared to all control models. The word embedding model considering only a single feature is less effective in control groups 1 and 2, and the results of each index of the word embedding model considering two features in control groups 3 and 4 are 7.02% better than the average recognition accuracy of those considering one, but still 2.57% lower on average than the model considering all features. Although the index results of control groups 3 and 4, which are shielded with BiLSTM and MHA modules respectively, are lower than the final model 5, the highest difference in recognition accuracy is 1.8% compared to the models of groups 1-2 with reduced word embedding features. It can be seen that the performance improvement of the model in this paper mainly comes from fully considering multiple semantic features, which indicates that the model in this paper has a stronger ability to generalize entities in the field of government affairs and can effectively improve the entity naming recognition effect for government affairs.

## *Discussion*

As shown in Figure 4 amd Figure 5 is the recognition results of this research model for the four entity names of person, location, organization and policy document names in different indexes, which can be seen that the model has higher values of various indexes for location and policy document names, and other categories of entities are relatively lower. The reason is that the number of location entities is large, leading to more adequately used for building the entity dictionary training, especially some cultivar names are often composed of the way of "xx city in xx province", and these typical features improve the accuracy of this kind of entity recognition to a certain extent. Since the format of policy document named in a relatively more standardized way, generally by "document name + document number + issue + date", and the naming rules are relatively clear and standardized, the accuracy of policy document name recognition is also relatively high. The names of people are mostly random, and they even have multiple meanings in some contexts, so the model sometimes cannot distinguish them effectively. And some organization names often have the situation of nested entities with more interfering information, which will also have some influence on the recognition effect of the model. To address the problems encountered by the current model recognition, methods such as improving the accuracy of semantic segmentation of the model as a way to obtain more accurate feature information, or expanding the lexicon of the governmental domain can be used to achieve better recognition results.

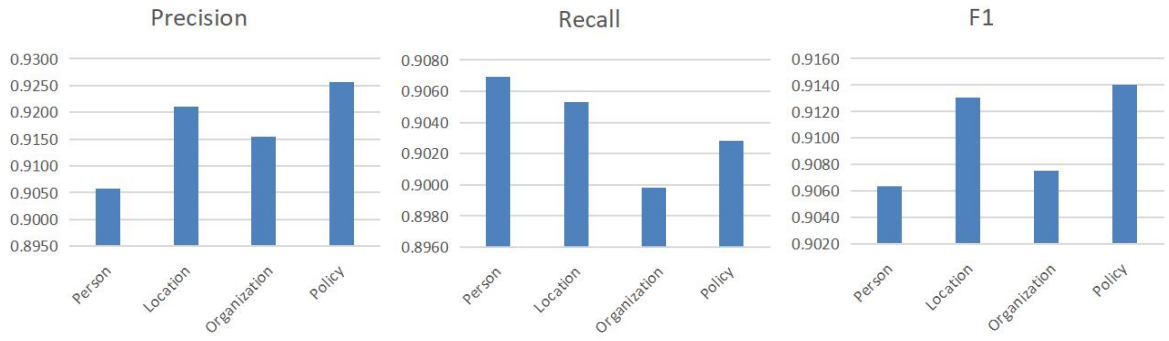


**Figure 4. The recognition effect of the model on various entities under different indicators**

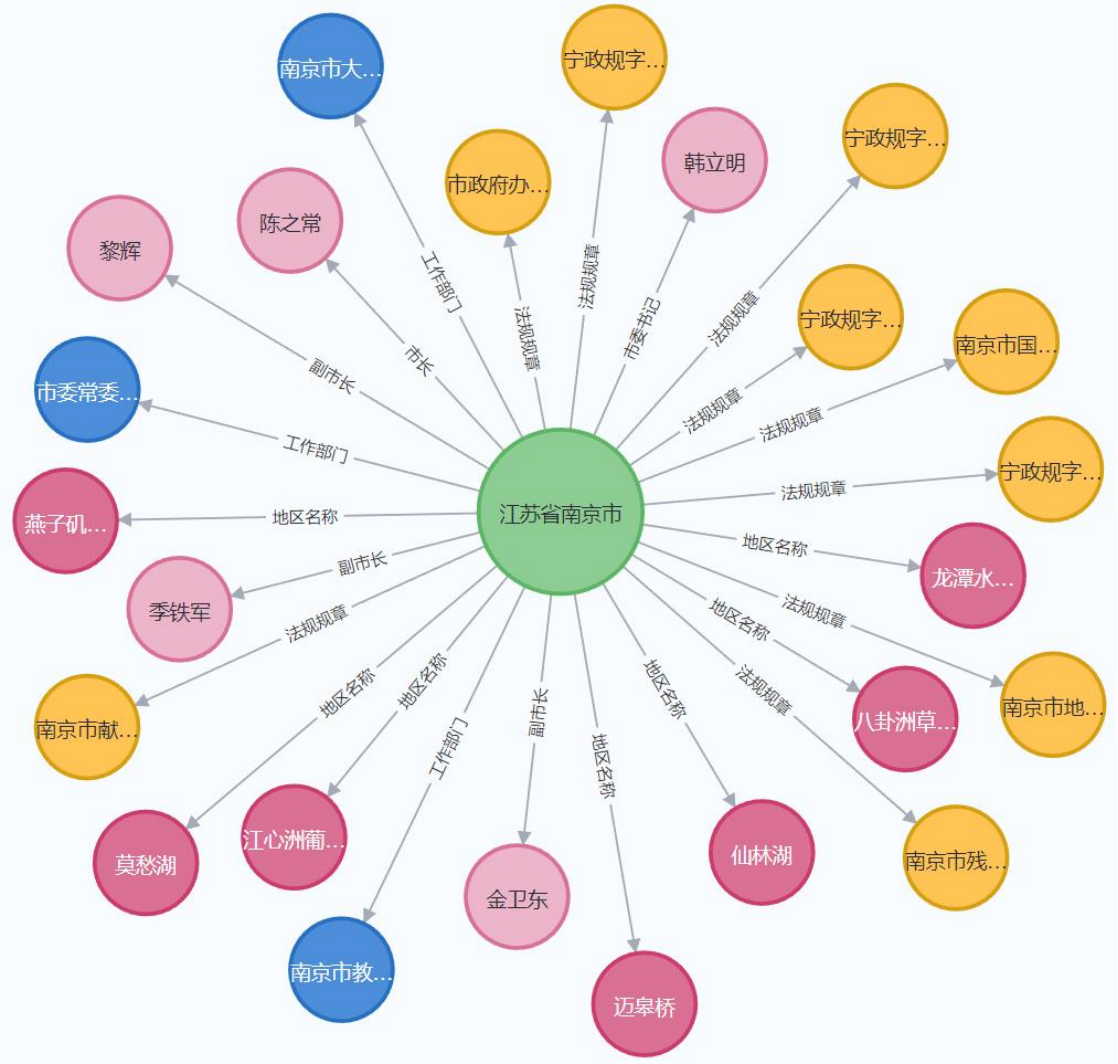


**Figure 5. Schematic Diagram of Chinese Government Official Document Entity Extraction**

## Conclusion

In this paper, we propose a recognition method for government named entities based on RobertaWWM+Dict+BiLSTM+MHA+CRF model to address the problem of low recognition accuracy due to the complicated and variable naming methods of long Chinese government named entities.By using multiple feature quantities such as combining full masked word embedding and constructing domain dictionaries as embedding layers, the contextual semantic information of characters is fully considered to provide a more complete characterization of Chinese entity features of government affairs and obtain more entity feature information.It alleviates the bias caused by semantic incompleteness of the model in named entity recognition and strengthens the model's ability to characterize Chinese semantics.Meanwhile, this paper uses BiLSTM, a bi-directional long and short term memory network, and a multi-headed attention mechanism to learn the long-range dependent information of text.Then, the conditional random field CRF is used to obtain the global optimal labeling sequence, and the Precision of the model is 91.8%, Recall is 90.5%, and F1 value is 91.1%, which is better than the baseline model and effectively improves the recognition effect of the model. Since there are problems of missing features and nested complex entities in the government entities, the future research direction will focus on further exploration of the entity recognition methods with fuzzy features. Such as adding the dictionary to the

model and extract the fuzzy relationships between different entities based on the model, while exploring the impact of more diverse annotation methods on entity recognition performance.

## Acknowledgements (optional)

This work is supported by the Major Program of the National Social Science Foundation of China "Research on the accurate construction of urban and rural community service system driven by big data"(Grant No. 20&ZD154) and Postgraduate Research & Practice Innovation Program of Jiangsu Province "Research on Topic Mining and Knowledge Graph Construction for Time Series Commentary of Online Health Information."(Grant No. KYCX23_0079)

## References

*Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012). The role of social networks in information diffusion. In Proceedings of the 21st international conference on world wide web (pp. 519–528).*

*Botnevik, B., Sakariassen, E., & Setty, V. (2020). BRENDA: Browser extension for fake news detection. In Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval (pp. 2117–2120).*

*Chieu, H. L., & Ng, H. T. (2003). Named entity recognition with a maximum entropy approach. In Proceedings of the seventh conference on natural language learning at HLT-NAACL 2003 - Volume 4 (pp. 160–163).*

*Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. Journal of Machine Learning Research, 12, 2493–2537.*

*Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., & Hu, G. (2020). Revisiting pre-trained models for Chinese natural language processing. In Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020.*

*Deping, C., Bo, W., Hong, L., Fang, F., & Run, W. (2021). Geological entity recognition based on ELMO-CNN-BILSTM-CRF model. Geoscience, 46, 3039–3048.*

*Fan, R., Wang, L., Yan, J., Song, W., Zhu, Y., & Chen, X. (2019). Deep learning-based named entity recognition and knowledge graph construction for geological hazards. ISPRS International Journal of Geo-Information, 9, 15.*

*Fang, Z. , Zhang, Q. , Kok, S. , Li, L. , Wang, A. , & Yang, S. . (2021). Referent graph embedding model for name entity recognition of chinese car reviews. Knowledge-Based Systems, 233, 107558-.*

*Han, J., Choi, D., Chun, B.-G., Kwon, T., Kim, H.-c., & Choi, Y. (2014). Collecting, organizing, and sharing pins in pinterest: Interest-driven or social-driven? InProceedings of the 2014 ACM international conference on measurement and modeling of computer systems (pp. 15–27).*

*Hu Wei, Liu Wei,&Shi Yujing (2022). A method of named entity recognition of TCM medical records based on bert-bilstm-crf Computer Age (009), 000*

*Jiang Yi, Huang Yong, Xia Yikun, Li Pengcheng,&Lu Wei (2021). Lexical function recognition of academic text -- application in automatic keyword extraction Journal of Information Technology (2), 152-162*

*Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In Proceedings of the 31st international conference on international conference on machine learning (pp. II–1188–II–1196).*

*Lin Litao, Wang Dongbo, Liu Jiangfeng, Li Bin,&Feng Minxuan (2022). The research of named entity recognition of ancient books and animals from the perspective of digital humanity -- taking sikubert pre-training model as an example Library Forum, 42 (10), 9*

*Li, W., Ma, K., Qiu, Q., Wu, L., Xie, Z., Li, S., & Chen, S. (2021). Chinese word segmentation based on self-learning model and geological knowledge for the geoscience domain. Earth and Space Science, 8,*

*Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2020). Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42,2980–2988.*

*Liu, X., Yang, N., Jiang, Y., Gu, L., & Shi, X. (2020). A parallel computing-based deep attention model for named entity recognition. The Journal of Supercomputing, 76, 814–830*

*Liu, Z. , Wang, X. , Chen, Q. , & Tang, B. . (2018). Chinese Clinical Entity Recognition via Attention-Based CNN-LSTM-CRF. 2018 IEEE International Conference on Healthcare Informatics Workshop (ICHI-W) (pp.68-69). IEEE Computer Society.*

*Lu Wei, Li Pengcheng, Zhang Guobiao,&Cheng Qikai (2020). Lexical function recognition of academic texts - research on automatic keyword classification based on bert vectorization Journal of Information Technology, 39 (12), 10*

*Qiu, J. , Zhou, Y. , Wang, Q. , Ruan, T. , & Gao, J. . (2019). Chinese clinical named entity recognition using residual dilated convolutional neural network with conditional random field. IEEE Transactions on NanoBioscience, 306-315.*

*Liu, Y. , et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." (2019).*

*Strubell, E. , Verga, P. , Belanger, D. , & Mccallum, A. . (2017). Fast and accurate entity recognition with iterated dilated convolutions.*

*Sun, Y. , Wang, S. , Li, Y. , Feng, S. , & Wu, H. . (2019). Ernie: enhanced representation through knowledge integration.*

*Tang Miaoji (2021). Analysis on the dynamic mechanism of knowledge discovery service in smart library based on data-driven Information Science, 39 (10), 7*

*Wang, J. , Xu, W. , Fu, X. , Xu, G. , & Wu, Y. . (2020). Astral: adversarial trained lstm-cnn for named entity recognition. Knowledge-Based Systems, 105842.*

*Yu Y, Wang Y, Mu J, et al (2022). Chinese mineral named entity recognition based on BERT model[J]. Expert Systems with Applications, 206: 117727.*

*Zhang Fangcong, Qin Qiuli, Jiang Yong,&Zhuang Runtao (2022). Research on Chinese electronic medical record named entity recognition based on roberta-wwm-ilstm-crf. Data Analysis and Knowledge Discovery (002), 006*

*Zhang, X., Ye, P., Wang, S., & Du, M. (2018). Geological entity recognition method based on deep belief networks. Yanshi Xuebao, (Acta Petrologica Sinica) 034,343–351.*