A Novel Approach to Named Entity Recognition in Chinese Disciplinary Inspection and Supervision Petition Cases

Jinyue Xu[1, a], Caijian Hua[1,2, b], and Yan Zhang[1,2, c]

[1] School of Computer Science and Engineering, Sichuan University of Science and Engineering, Yibin 644000, China

[2] Key Laboratory of Higher Education of Sichuan Province for Enterprise Informationalization and Internet of Things, Sichuan University of Science and Engineering, Yibin 644000, China

[a]xjy591135501@163.com, [b]hwacj@suse.edu.cn, [c]jkzhangyan@qq.com

Abstract

A high-quality named entity recognition (NER) model plays a vital role in the field of discipline inspection and supervision of petition cases by providing effective support for tasks like extracting key information and facilitating semantic retrieval. However, the irregularity of expressions in petition case texts restricts the applicability of existing general NER models or tools in this domain. To address this limitation, this article proposes an enhanced NER model, named ALBERT-BiGRU-Attention-CRF, specifically designed for the discipline inspection and supervision of reporting letters and visits. The proposed model incorporates an attention mechanism between the BiGRU and CRF models. By leveraging the dynamically trained word vectors of ALBERT, it effectively represents text features, which are then input into BiGRU to capture contextual information. The attention mechanism complements local features, while the CRF captures the dependency relationships between adjacent labels. The experimental results demonstrate that the proposed model exhibits robust extraction performance in Chinese discipline inspection and supervision petition cases, achieving an F1 score of 77.52%. Furthermore, a discipline inspection and supervision NER system is developed, distinguishing itself from conventional NER systems by incorporating a naive Bayes classifier for petition case classification, thereby attaining commendable performance.

Keywords

Disciplinary Inspection and Supervision; Petition Cases; Entity Extraction; ALBERT; Bigru; Attention Mechanism; CRF; Naive Bayes.

1. Introduction

Discipline inspection and supervision are essential functions of China's disciplinary inspection agencies and government supervision departments, which are crucial for enhancing the governance capacity of the ruling party [1]. As China's disciplinary and supervisory system reform progresses, the public's participation in the supervision process has increased. Specialized departments have been established to handle petition cases, including complaints and reports from the public. These cases generate a large volume of data daily, characterized by their wide coverage, low relevance, rapid generation speed, and the presence of numerous entities and rare words from various fields, such as medical subsidies and fund bribery. This data serves as a valuable source of information, providing work clues and insights into the political ecology, public opinion, and social issues.

Within these petition cases, crucial information, including location, subjects, and specific details of events, plays a pivotal role in case handling and building a chain of clues and evidence. This information enables discipline inspection personnel to quickly identify breakthroughs and follow clues, facilitating the construction of case recommendations, question-and-answer systems, case analysis, and character portraits. However, the increasing amount of petition case data poses a challenge, as manual extraction methods hinder the efficiency of discipline inspection and supervision personnel in handling cases. Therefore, finding a solution to swiftly extract relevant entity information becomes crucial in streamlining and simplifying the handling process of discipline inspection and supervision petition cases.

NER is an important direction of research in natural language processing [2]. As an intelligent information extraction method, NER aims to extract specific entities, relations, and entity attributes from structured, semi-structured, or unstructured data, such as names of people, places, and organizations in traditional text, to quickly and accurately locate keywords and mine deep semantic relationships. Currently, NER research has been extensively applied to various vertical fields, such as military [3, 4], medical [5, 6], and so on, but research on applying NER to the field of disciplinary inspection and supervision is rare [7].

This article proposes an ALBERT-BiGRU-Attention-CRF model for entity recognition in petition cases within the discipline inspection and supervision field. The model leverages the "attention mechanism" [8, 9], to collect context-related semantic information. To optimize training time, we use the ALBERT model [10] with fewer parameters in the encoding layer. However, this reduction in parameters may result in a loss of overall model performance. To address this, we employ two approaches. Firstly, we combine the BiGRU model with the CRF model [11] to capture dependency information and obtain the globally optimal tag sequence. Secondly, we integrate the attention mechanism to enhance the model's identification performance in handling limited information processing capabilities.

Experimental results demonstrate the effectiveness of the proposed model for entity recognition in petition cases within the discipline inspection and supervision domain. Furthermore, we utilize a naive Bayes classifier [12, 13], for text classification in petition cases, simplifying statistical work in discipline inspection and supervision. The experimental results reveal the favorable performance achieved by this classification approach.

The structure of the remainder of this paper is arranged as follows. Section 2 introduces related research. Section 3 describes a NER method for disciplinary inspection and supervision petition cases named ALBERT-BiGRU-Attention-CRF. Section 4 introduces experimental parameters and analyzes the experimental results. Section 5 constructs a NER system for disciplinary inspection and supervision cases. Section6concludes the paper.

2. Related Research

NER was officially proposed at the Message Understanding Conference (MUC-6) [14]. The research history of NER can be roughly divided into several stages.

2.1 Methods based on Rule and Dictionary

The rule-based and dictionary-based method requires the construction of a large and comprehensive knowledge base and dictionary, observation the composition of entity components, and summarizes templates based on grammar rules [15]. This method has a high recognition effect on its specific corpus, but these rules overly rely on specific domains and specific languages, which are difficult to cover all aspects and are prone to errors. Additionally, there are problems with poor portability, high cost, and a long knowledge base establishment cycle.

2.2 Methods based on Statistical Machine Learning

This method requires the integration of relevant knowledge in machine learning, statistics, and linguistics to build a model. Its essence is sequence labeling, using supervised training on human-labeled corpus texts. Compared with rule-based methods, this method has a certain level of effectiveness in entity recognition. Commonly used statistical machine learning models for Chinese entity recognition include Hidden Markov Model (HMM) [16], Maximum Entropy [17], and CRF [18].

2.3 Methods based on Deep Learning

This method utilizes the powerful nonlinear transformation, vector representation, and computational abilities of deep neural networks to obtain vector representations of words and texts, reducing the workload of feature engineering. Additionally, it learns contextual semantic information to better complete entity recognition tasks [19]. This approach has good generalization capabilities and is gradually becoming mainstream in the field.

In recent years, deep learning has been widely applied to a large number of natural language processing tasks, including NER, and has achieved significant performance improvement. Some deep learning models have also achieved good performance improvement in English NER tasks [20]. However, applying these models directly to Chinese NER tasks does not yield good experimental performance. Compared with English text, Chinese text has some unique language features: Chinese text does not have spaces as explicit boundary markers like English text, making it difficult to determine word segmentation boundaries. Chinese entities themselves have ambiguity, flexible word formation, complex language structure, and other problems, which increase the difficulty of entity recognition [21].

NER is typically addressed as a sequence labeling problem using Bi-directional Long Short-Term Memory (Bi-LSTM) and Conditional Random Field (CRF) as the main methods [22, 23, 24]. BiLSTM-CRF has revolutionized NER deep learning by effectively capturing contextual information, dependency relationships between labels, and optimizing label sequencing [25]. Miao et al. [26] proposed an improved BiLSTM-CRF model for aspect-based Chinese sentiment analysis, enhancing sentiment analysis accuracy and capturing user demand information. Tang et al. [27] developed a multi-task BERT-BiLSTM-AM-CRF model, utilizing BERT for dynamic word vector extraction and contextual information. However, incorporating different entity annotation rules into the multi-task learning model yielded inferior results compared to the single-task model. Lang et al. [28] proposed a topic-oriented automatic text extraction method using BiLSTM and CRF, effectively identifying more entity information and forming standardized structured data from unstructured text. Liu et al. [29] addressed the problem of a large amount of irrelevant information in the disciplinary inspection field by constructing a disciplinary inspection and supervision knowledge graph using a BERT-BiLSTM-CRF model. Li et al. [30] introduced an ALBERT-BiGRU-CRF-based method for Chinese word segmentation, significantly improving accuracy and generalization performance. In another study, Liu et al. [7] presented a joint extraction model using BERT-BiGRU-CRF for event identification, but its feature representation is limited and it struggles with noisy data in the disciplinary inspection domain.

3. Model Design

In the field of NER, using deep neural network models for entity recognition has become mainstream. In this article, we propose to construct an ALBERT-BiGRU-Attention-CRF model to extract named entities in the field of disciplinary inspection and supervision from petition cases.

3.1 Overall Framework

This article proposes an improved ALBERT-BiGRU-Attention-CRF neural network model for NER in the field of discipline inspection and supervision. Starting from word vectors, the model aims to improve the accuracy of entity recognition. The improved model framework is shown in Figure1, and the model process is as follows: First, the pre-trained ALBERT model is used to obtain the vector representation of all words in the text. Then, the obtained character vectors are input to BiGRU for feature extraction. The forward GRU network learns future features, while the backward GRU network learns historical features. To better capture the information output from the BiGRU network and improve the recognition ability of the model for key information, this article introduces an "attention mechanism" to compensate for the accuracy loss of the model and enhance the feature representation ability: use the mined global features (hidden state ht at time t) as the output, and use the attention mechanism to supplement local features, predicting the internal relationship between the input text sequence and the label. Finally, the CRF layer is used for decoding to output the entity annotation sequence with the highest overall probability. The attention layer is placed between the BiGRU layer and the CRF layer, mainly because the attention mechanism searches for the optimal sequence at the vector level, and both the input and output are vectors, while the output of the BiGRU layer is a sequence vector with position information.

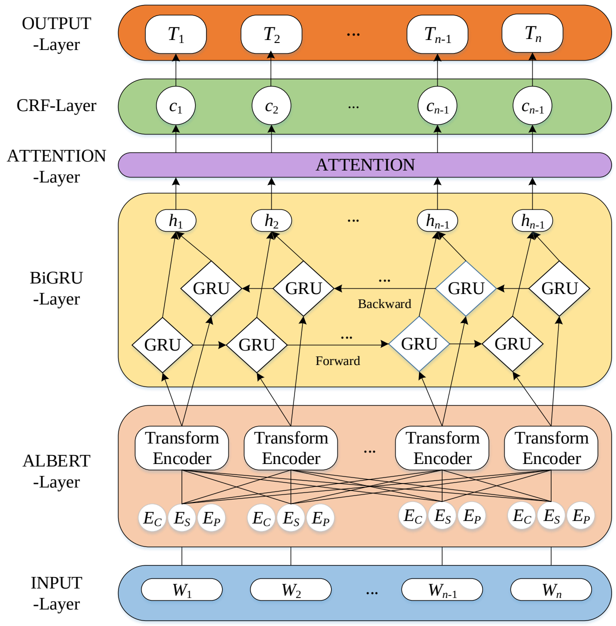


Figure 1. Overall model framework

3.2 Input Layer

Figures compiled of more than one sub-figure are presented side-by-side, or stacked. If a multipart figure is made up of multiple figure types (one part is linear, and another is grayscale or color) the figure should meet the stricter guidelines.

$$S = \{W_1, W_2, \cdots, W_n\} \tag{1}$$

When a text (or a sentence) S is inputted to the ALBERT layer, the text or sentence S is shown as (1), where Wi represents the i-th character in the sentence.

3.3 ALBERT Layer

The ALBERT pre-trained language model is a lightweight pre-trained language model based on the BERT model. It simplifies the number of parameters in the BERT model while ensuring model performance. The structure of the ALBERT model is shown in Figure 2, where En is the encoding representation of a word, Trm is the Transformer structure, and Tn is the word vector of the target word after training. ALBERT uses a bidirectional Transformer as a feature extractor, which can obtain longer contextual information compared to traditional recurrent neural networks or language models, thus improving feature extraction ability. In this article, we used the Chinese pre-trained model "ALBERT-Base" released by Google.

Firstly, each character of the input is embedded into a vector space representation, where the vector of each character E is shown in (2), where Ec represents the character vector, Es represents the sentence classification vector of the character, and Ep represents the sequence position vector of the character.

$$E = E_c + E_s + E_p \tag{2}$$

Secondly, the vector representation E of Chinese characters is trained through multiple layers of bidirectional Transformer encoders, which can read the entire text sequence at once, allowing each layer to integrate context information, and finally obtaining the feature representation {T1, T2,..., Tn} of the text. The ALBERT model only uses the Encoder part of the Transformer, which is composed of N identical network layers stacked together.

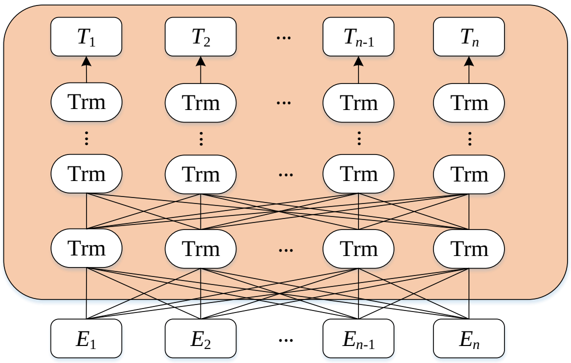


Figure 2. The structure of the ALBERT model

3.4 BiGRU Layer

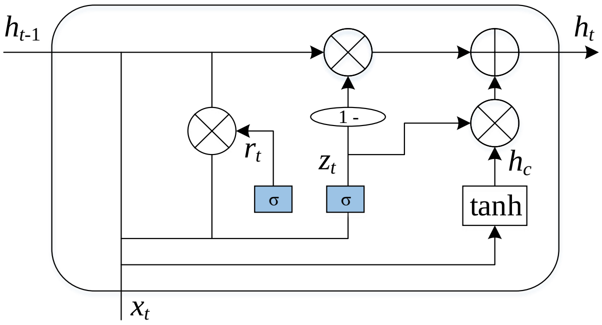


Figure 3. The unit structure of the GRU model

The text features obtained from the ALBERT model are trained using the BiGRU model. BiGRU is composed of bidirectional GRU. GRU is an improved model based on Long Short-Term Memory (LSTM) with a simplified internal structure similar to LSTM. It processes gradients through gate mechanisms to avoid memory degradation. The schematic diagram of the GRU cell structure is shown in Figure 3. Compared with the complex and redundant LSTM model, the GRU model is more concise, because it only consists of an update gate zt and a reset gate rt , one less gate than LSTM, which leads to fewer parameters and faster convergence during training. The reset gate decides whether to discard the previous information. The smaller the reset gate, the more information is discarded. The role of the update gate is to update past information to the current position, and the larger the update gate, the more information is updated. The workflow of the GRU network layer is as follows:

$$z_t = \sigma(W_z x_t + U_z h_{t-1}) \tag{3}$$

Computing the update gate zt , the mathematical expression is shown in (3).

In the above equation, zt represents the information captured by the update gate, σ denotes the sigmoid activation function, and Wz and Uz are the weight parameters of the update gate that need to be randomly initialized. The xt is the vector at the t-th position of the pre-trained word embedding sequence output in the previous step, and ht-1 is the vector at the previous position.

Computing the reset rt, the mathematical expression is shown in (4).

$$r_t = \sigma\left(W_r x_t + U_r h_{t\text{-}1}\right) \tag{4}$$

In the above equation, rt represents the information captured by the reset gate; Wr and Ur are the weight parameters of the reset gate that need to be randomly initialized.

Computing the candidate state hc, the mathematical expression is shown in (5).

$$h_c = \tanh\left(W x_t + U\left(r_t h_{t\text{-}1}\right)\right) \tag{5}$$

In the above equation, hc is the current candidate state, W and U are the weight parameters of the candidate state, which need to be randomly initialized, and tanh is the activation function.

Computing the candidate state ht, the mathematical expression is shown in (6).

$$h_t = \left(1 - z_t\right) h_{t\text{-}1} + z_t h_c \tag{6}$$

In the above equation, ht represents the current hidden state at time t, and hc is the current candidate state.

The above is the workflow of the unidirectional GRU. When transmitting information, unidirectional GRU can only propagate forward and cannot learn information beyond the current time step. Therefore, a bidirectional GRU network structure is proposed here. The character vectors obtained from the ALBERT layer are input into the forward and backward GRUs for feature extraction, which can fully utilize the feature vectors to link contextual information.

3.5 Attention Layer

The attention mechanism is introduced to address the limitations of BiGRU networks in capturing local features within the corpus. The primary objective of incorporating the attention mechanism is to enhance the NER model's ability to focus on relevant input and local information that directly contributes to the current output. This paper aims to utilize the attention mechanism to learn the dependency relationship between any two characters in a sentence, thereby obtaining crucial internal structural information about the sentence.

By applying the attention mechanism to the feature vectors (hj) generated by the BiGRU layer, the model calculates weight assignments (atj) that are utilized to obtain the jointly outputted feature vector (ct) by the BiGRU and attention layers at time t, which serves as the final output, as shown in (7), (8), and (9). This final output serves as a comprehensive representation, incorporating the contextual information captured by the BiGRU layer with the relevant information highlighted by the attention mechanism. Through this integration, the attention mechanism plays a crucial role in enriching the model's ability to capture intricate dependencies and improve its overall performance.

In the above three equations, etj is the alignment model, and v, w, and m are weight vectors.

$$c_t = \sum_{j=1}^{T} a_{tj} h_j \tag{7}$$

$$a_{tj} = \frac{\exp(e_{tj})}{\sum_{k=1}^{T} \exp(e_{tk})} \tag{8}$$

$$e_{tj} = v^T \tanh(wc_{t-1} + mh_j) \tag{9}$$

3.6 CRF Layer

Although the BiGRU layer can learn the feature information between contexts and select the label with the highest probability as output, it cannot capture the dependency relationships between output labels, which may result in consecutive identical labels. However, the CRF has transition features that can consider the orderliness of output labels. Therefore, the CRF is selected as the output layer for the BiGRU and attention mechanism.

CRF is a probabilistic graphical model used to solve sequence labeling problems. The model structure is shown in Figure 4, which can learn contextual information by considering the global information of the label sequence and adding constraints to the predicted results. It combines the global probability of the label sequence and the output layer results and predicts the label sequence with the highest probability.

For a given sentence, that is input sequence X={x1, x2, ···, xn} and corresponding output label sequence Y={y1, y2, ···, yn}, the CRF evaluation score is defined in (10).

$$\mathrm{s}(x, y) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i} \tag{10}$$

In the above equation, A and P are the transition score matrix and output score matrix, respectively. $A_{y_i, y_{i+1}}$ represents the transition score from label i to label i+1. P_{i, y_i} represents the output score of the i-th Chinese character for the label yi. The probability of the predicted sequence y is shown in (11).

$$P(x \mid y) = \frac{\exp[s(x, y)]}{\sum s(x, y)} \tag{11}$$

Taking the logarithm on both sides yields the likelihood function of the predicted sequence, as shown in (12).

$$\ln(P(y \mid x)) = s(x, y) - \ln(\sum s(x, y)) \tag{12}$$

The problems commonly encountered in the field of discipline inspection and supervision of petition cases are classified into different categories, and the petition case events are abstractly represented in each category. For example, the event “The collapse of the road south of Bamao Village”, “Yongxing Town” is abstractly classified into the field of transportation. The problem category set is Y={y1, y2, ···, yn}, the problem feature set is X={x1, x2, ···, xn}. The Bayesian formula is shown in (13).

$$P(y_i \mid X) = \frac{P(X \mid y_i) P(y_i)}{\sum_{i=1}^{n} P(X \mid y_i) P(y_i)} \tag{13}$$

Naive Bayes is a classification method based on Bayes' theorem and the assumption of feature independence. A set of keywords is used to represent the features of a certain category of problems in the data set, and this set of keywords is the feature vector X for this category of problems. After inputting the content text, the class with the maximum posterior probability P(yi|X) can be determined, which is the classification of the content text. The denominator in (13) can be considered as a constant, therefore, the Naive Bayes formula can be expressed as (14).

$$\max\left(P\left(y_i \mid X\right)\right) = \max\left(P\left(y_i\right)\prod_i^n P\left(x_i \mid y_i\right)\right) \tag{14}$$

4. Experiment

4.1 Experimental Data Construction

(1) Data description

Due to the fact that text related to disciplinary inspection and supervision petition cases is generally not publicly available, the data used in this study was provided by the Discipline Inspection Commission of a certain city, consisting of a set of internal petition case data divided into seven subcategories. The data set contains a total of 11250 cases, which are divided into three parts in an 8: 1: 1 ratio, with 9000 cases in the training set and1, 125 cases each in the validation and test sets, as shown in Table 1.

In this paper, entity annotation and recognition were performed on 11250 disciplinary inspection and supervision data. Three types of entities were set in terms of entity categories:

1) Regional name

Geographical names are the areas where the events occurred, and the name reflects the geographic location of the case.

2) Case subject name

In the text of disciplinary and supervisory petition cases, there are names of suspects, organizations, and sometimes objects such as “highway” in the term “highway collapse”. These are the main elements of the event and provide the fundamental information of the case.

3) Event name

Event names include specific report contains information of the reported subject, such as “embezzlement of public funds”, “accepting bribes”, “violating laws for personal gain”, and so on.

Table 1. Classification and quantity statistics of the data set

Case type	Training set	Testing set	Validation set	Total	Average length
Market regulation and supervision	1580	196	198	1974	92
Ecological environment protection	1553	194	194	1941	65
Transportation	1547	194	193	1934	52
Corruption in work style	1536	192	192	1920	168
Rural revitalization	1011	126	127	1264	43
Education and healthcare	954	120	119	1193	98
Housing and city construction	819	103	102	1024	110
Total	9000	1125	1125	11250	90

(2) Data annotation

Table 2. BIO annotation example

Text	Annotation	Describe
永	B-REG	The beginning of the region part entity
兴	I-REG	The middle of the region part entity
镇	I-REG	The middle of the region part entity
南	O	Non-entity
边	O	Non-entity
公	B-SUB	The beginning of the subject part entity
路	I-SUB	The middle of the subject part entity
崩	B-EVE	The beginning of the event part entity
裂	I-EVE	The middle of the event part entity
坍	I-EVE	The middle of the event part entity
塌	I-EVE	The middle of the event part entity

Data annotation is carried out by using the original annotation method, where each character in 11250 pieces of data is annotated using the BIO tagging scheme. The character “B” (Begin) represents the beginning of an entity, “I” (Intermediate) represents the middle part of an entity, and “O” (Other) represents irrelevant characters. The tagging system used in this article is the BIO labeling scheme, which includes three types of entities: REG, SUB, and EVE. Therefore, there are seven tags, namely O, B-REG, I-REG, B-SUB, I-SUB, B-EVE, and I-EVE. The annotation data style is shown in Table 2 (The meaning of this Chinese text is: “The southern road of Yongxing Town collapsed and crumbled.”).

4.2 Evaluation Metrics

The named entities involved in this paper are divided into three categories, and individual categories and overall system performance need to be evaluated. Precision (P), recall (R), and F1 are used as evaluation metrics for NER performance.

The P formula is shown in (15).

$$P = \frac{TP}{TP + FP} \tag{15}$$

The R formula is shown in (16).

$$R = \frac{TP}{TP + FN} \tag{16}$$

The F1 formula is shown in (17).

$$F_1 = \frac{2 * P * R}{P + R} \tag{17}$$

In the above equations, TP refers to true positive samples and predictions; FP refers to negative samples and positive predictions; FN refers to positive samples and negative predictions.

4.3 Experimental Environment and Parameters

The experimental environment used during the experiment is shown in Table 3. To better investigate the performance of the model, the experimental parameters were set as shown in Table 4.

Table 3. Lab environment

NAME	DESCRIBE
CPU	Intel(R) Xeon(R) Silver 4210R
GPU	RTX 3080
This paper memory	128GB
System environment	Windows10
Python	3.9.7
PyTorch	1.12.1

4.4 Experimental Result Analysis

Table 4. Experimental Parameters

PARAMETER	DESCRIBE
max_seq_length	150
train_epoch	20
train_batch_size	32
learning_rate	1e-5
GRU_Layers	2
GRU_Hidden_Size	128
dropout	0.2
Shuffle	True
optimizer	Adam

This article compares the loss values of four models, BiLSTM-CRF, ALBERT-BiLSTM-CRF, ALBERT-BiGRU-CRF, and ALBERT-BiGRU-Attention-CRF during the training process. The changes in training loss are shown in Figure 4. By comparing these three models, it can be observed that the loss values of the GRU-based models all decrease to below1at around the 17th round.

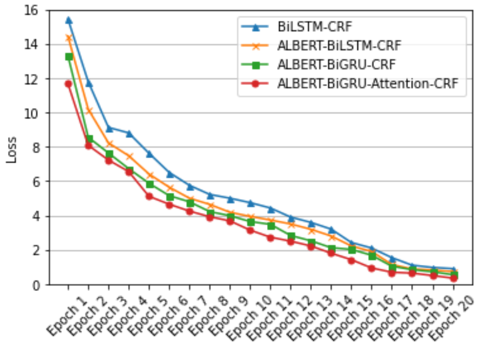


Figure 4. Loss change graph

Table 5. Comparison of model results

METHOD MODEL	Precision	Recall	F1
BiLSTM-CRF	0.7257	0.6985	0.7118
ALBERT-BiLSTM-CRF	0.7389	0.7472	0.7430
ALBERT-BiGRU-CRF	0.7616	0.7691	0.7653
ALBERT-BiGRU-Attention-CRF	**0.7705**	**0.7801**	**0.7752**

During the 20 rounds of iteration, the model with the added attention mechanism showed significantly lower loss compared to the model without the attention mechanism. The use of the Attention mechanism enabled the model to better utilize textual features, resulting in a lower loss.

To verify the performance of the proposed ALBERT-BiGRU-Attention-CRF model in the field of disciplinary inspection and supervision for petition cases, comparative experiments were conducted with several classic models on the Chinese disciplinary inspection and supervision petition case dataset, obtaining the accuracy, recall rate, and F1 value for the four models, as shown in Table 5.

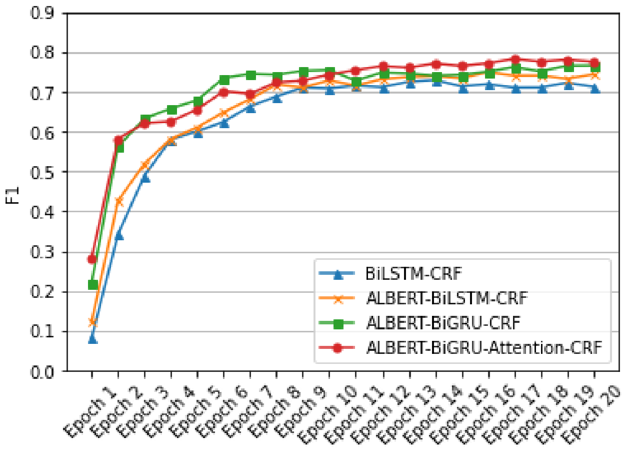


Figure 5. Comparison of model performance

Figure 5 displays a comparison of the F1 scores for the first 20 rounds of the four models. Furthermore, by analyzing and comparing BiLSTM-CRF and ALBERT-BiLSTM-CRF in conjunction with Table 5, it can be concluded that the F1 value of ALBERT-BiLSTM-CRF with the addition of pre-trained ALBERT model is 3.07% higher than that without, demonstrating a significant impact of ALBERT's domain-specific word vectors on event extraction tasks. Compared to ALBERT-BiLSTM-CRF, ALBERT-BiGRU-CRF achieves a higher F1 score of 2.23% on the disciplinary inspection and supervision petition dataset. Although both LSTM and GRU have good long sequence processing capabilities and well-designed gate mechanisms, the GRU structure requires fewer parameters and converges faster during training because it has one less gate than LSTM. Therefore, the experimental performance of ALBERT-BiGRU-CRF is slightly better than that of ALBERT-BiLSTM-CRF. Comparing the ALBERT-BiGRU-CRF model with the ALBERT-BiGRU-Attention-CRF model, it can be observed that the addition of the Attention model leads to an increase in P by 0.89%, R by 1.1%, and F1 by 0.99% on the disciplinary inspection and supervision petition dataset, indicating that adding the Attention model is more favorable than not adding it. However, it is worth noting that the ALBERT-BiGRU-Attention-CRF model proposed in this study has a slower convergence speed due to the increased self-attention mechanism, which requires more training to achieve optimal performance. Nevertheless, the integration of the attention mechanism can capture the characteristics of the text's contextual features in disciplinary inspection and supervision petition cases from multiple dimensions and can further improve model performance. Therefore, the ALBERT-BiGRU-Attention-CRF model proposed in this study has the best experimental performance compared to the other models and generally performs better than other models on the disciplinary inspection and supervision petition data set.

Table 6. Examples of intent classification for some petition cases

PETITION CASES	Entity extraction	Classification
翠屏区群众反映镇上工厂污水未处理完善往农户田里乱排乱放	region：翠屏区 case subject：工厂污水 event：乱排乱放	Ecological environment protection
长宁县乡村公路被重车压裂崩塌	region：长宁县 case subject：乡村公路 event：压裂崩塌	Transportation and logistics
群众举报江安县水务局科员违规借贷问题	region：江安县 case subject:水务局科员 event：违规借贷	Corrupt work style

Table 7. Comparison of model results

MODEL	P	R	F_1
NB	0.8276	0.8024	0.8148
SVM	0.7693	0.7045	0.7355
DT	0.7058	0.6530	0.6784

In addition, the imported petition text is classified using a naive Bayes classifier. Examples of intent classification for some disciplinary inspection and supervision petition cases are presented in Table 6. Compared with other classification algorithms, the naive bayes algorithm used in this paper for case classification has higher accuracy in the dataset. As shown in Table 7, the F1 value of the naive bayes (NB) is the highest, which is 9.3% higher than the support vector machine (SVM) [31] and 13.1% higher than the decision tree algorithm (DT) [32].

The three sample cases respectively report on: “The people in Cuiping District report that the sewage from the factory in the town is not treated properly and is discharged indiscriminately into the agricultural fields”, “Rural highway in Changning County collapsed due to heavy truck fracturing”, and “The masses report the problem of illegal lending by staff members of Jiang'an County Water Affairs Bureau”. They are classified into the categories of “Ecological environment protection”, “Transportation and logistics” and “Corruption in work style”.

5. System Implementation

In this section, we developed the discipline inspection and supervision NER system. The system is built on a B/S architecture and utilizes the Django development framework. The Django framework follows the model template view (MTV) pattern, which divides a development task into three parts: model, template, and view. The model component is responsible for connecting the business objects with the database objects. The template component mainly stores the HTML files and determines how to present the pages to the user. The View component implements the business logic through the view functions and calls model and template at appropriate times.

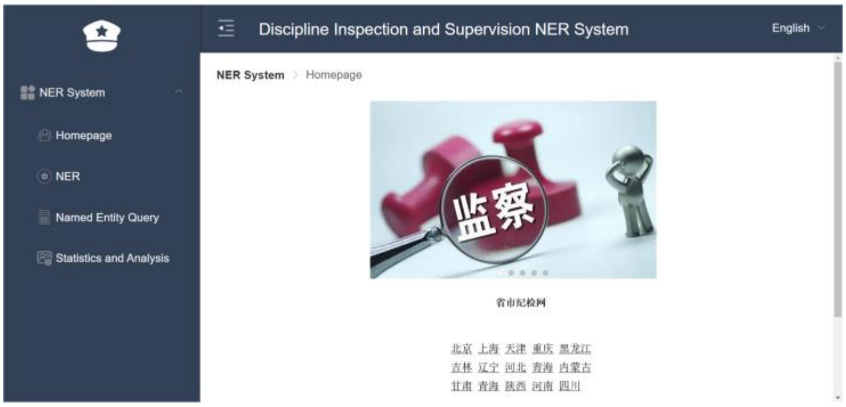
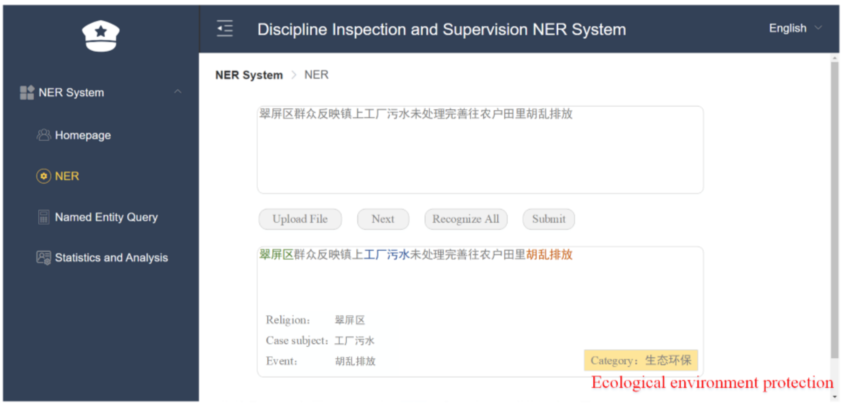


Figure 6. System Homepage

The framework used in this system has various advantages, such as complete functionality and components, including powerful database access components and a backend management system, which allows for front-end and back-end separation. This separation greatly enhances development efficiency and reduces the coupling degree between the front-end and back-end.



翠屏区群众反映镇上工厂污水未处理完善往农户田里胡乱排放

翠屏区群众反映镇上工厂污水未处理完善往农户田里胡乱排放

Religion: 翠屏区
Case subject: 工厂污水
Event: 胡乱排放

Category: 生态环保
Ecological environment protection

Figure 7. NER Result

The system homepage is illustrated in Figure 6. The left side of the page contains a functional bar, where users can execute NER and named entity query functions, and finally perform statistical analysis on the recognized entities. At the bottom of the page, users can click hyperlinks to the discipline inspection and supervision websites of various provinces and cities.

The page for performing NER is shown in Figure 7. When inputting a petition case text or uploading a petition case file, users can choose to identify and extract entities one by one or all at once. After clicking “Submit”, all named entities related to the case will be recognized and the system will highlight the identified information in a special color, which will be displayed in a standardized manner in the lower left corner. At the same time, a classification result for this text will be given in the lower right corner.

6. Conclusion

In response to the challenges of the large amount of data in the Chinese discipline inspection and supervision field, as well as the complex and tedious nature of handling petition cases, this paper proposes a deep learning-based NER model ALBERT-BiGRU-Attention-CRF to extract data information from Chinese discipline inspection and supervision petition cases, thereby improving the efficiency of discipline inspection and supervision work. The contribution of this model lies in the fusion of the attention mechanism, ALBERT, and BiGRU-CRF. Compared with the baseline model, the proposed model achieves better recognition performance in less training time. This model performs well in a Chinese discipline inspection and supervision petition case event corpus constructed in this paper. Finally, this paper also uses Naive Bayes to classify petition case texts. The Naive Bayes classifier has the advantages of fast computation speed and stable performance but also has some drawbacks, such as oversimplified assumptions and poor performance in complex situations.

This paper proposes a NER model based on deep learning, ALBERT-BiGRU-Attention-CRF, to extract data information in the field of Chinese disciplinary inspection and supervision petition cases and improve the efficiency of disciplinary inspection and supervision work. The model combines the attention mechanism, ALBERT, and BiGRU-CRF, and achieves better identification performance in less training time compared to the baseline model. This model performs well in constructing a corpus of Chinese disciplinary inspection and supervision petition cases in this paper. Finally, this paper uses naive Bayes to classify petition case texts. Although the naive Bayes classifier has advantages such as fast computation speed and stable performance, it also has some disadvantages, such as oversimplified assumptions and poor performance in complex situations.

There is still room for improvement in the proposed model, and further efforts can be made in the following aspects:

(1) The corpus of Chinese disciplinary inspection and supervision petition cases constructed in this paper has a small data scale and few entity types. The next step can expand the amount of data and entity categories, including time, disciplinary regulations, and rules for entity recognition.

(2) Since the petition case texts are submitted by ordinary people, the expression may be unclear, and the labeling work of entities is difficult due to the involvement of many fields and the complexity of Chinese semantic text. Further correction of data labeling is an effective way to improve the training effect of the model.

(3) Although the naive Bayes classifier has good classification performance, it has certain limitations. To achieve better classification results, other methods can be combined for improvement.

(4) In the future, a disciplinary inspection and supervision domain dictionary can be established to improve the accuracy of NER. Additionally, similar case comparisons, question-and-answer systems, analysis of connected cases, and character profiling can be developed to improve disciplinary inspection and supervision work.

Acknowledgments

This work was supported in part by the Innovation Fund of Postgraduate of Sichuan University of Science and Engineering under Grant Y2022174, and the Key Laboratory of Higher Education of Sichuan Province for Enterprise Informationalization and Internet of Things Plan Project under Grant 2022WYY01.

References

[1] J. Wang, J. Ren. New Era Observation of the Reform of Disciplinary Inspection and Supervision Business Work: Situation Tasks, Institutional Reform, and Performance Performance [J]. Socialism Research, (2021) No. 257, p. 79-90.

[2] S. Zhao, R. Luo, Z. Cai, Survey of Chinese Named Entity Recognition [J]. Journal of Frontiers of Computer Science & Technology, vol. 16 (2022) No. 2, p. 296.

[3] L. Ma, T. Li, A. Liu, J. Qin, Research on Named Entity Recognition of Small Datasets Based on Transfer Learning [J]. Journal of Huazhong University of Science and Technology (Natural Science Edition), vol. 50 (2022) No. 2, p. 118-123.

[4] H. Li, Y. Lin, J. Zhang, M. Lv, Fusion of Deep Learning and Machine Learning for Heterogeneous Military Entity Recognition [J]. Wireless Communications and Mobile Computing, vol. 2022 (2022) p. 1-11, 2022.

[5] L. -H. Lee and Y. Lu, Multiple Embeddings Enhanced Multi-Graph Neural Networks for Chinese Healthcare Named Entity Recognition [C]. IEEE Journal of Biomedical and Health Informatics, vol. 25 (2021) No. 7, p. 2801-2810.

[6] Q. Zhang, M. Wu, P. Lv, M. Zhang, H. Yang, Research on Named Entity Recognition of Chinese Electronic Medical Records Based on Multi-Head Attention Mechanism and Character-Word Information Fusion [J]. Journal of Intelligent & Fuzzy Systems, vol. 42 (2022) No. 4, p. 4105-4116.

[7] X. Liu, J. Chen, J. Gao, H. Fan, and J. Dong, A Method of Extracting Discipline Inspection Cases Based on Deep Learning [C]. 2022 3rd International Conference on Education, Knowledge and Information Management (ICEKIM), Harbin, China (2022) p. 385-391.

[8] K. Zheng et al., Named Entity Recognition in Electric Power Metering Domain Based on Attention Mechanism [J]. IEEE Access, vol. 9 (2021) p. 152564-152573.

[9] J. Kang, L. Zhang, M. Jiang, and T. Liu, Incorporating multi-level CNN and attention mechanism for Chinese clinical named entity recognition [J]. Journal of Biomedical Informatics, vol. 116 (2021) p. 103737.

[10] P. Zhao, W. Wang, H. Liu, and M. Han, Recognition of the Agricultural Named Entities With Multi-feature Fusion Based on ALBERT [J]. IEEE Access, vol. 10 (2022) p. 98936-98943.

[11] Q. Qin, S. Zhao, and C. Liu, A BERT-BiGRU-CRF model for entity recognition of Chinese electronic medical records [J]. Complexity, vol. 2021 (2021) p. 1-11.

[12] I. Wickramasinghe and H. Kalutarage, Naive Bayes: Applications, Variations, and Vulnerabilities: A Review of Literature With Code Snippets for Implementation [J]. Soft Computing, vol. 25 (2021) No. 3, p. 2277-2293.

[13] X. Deng, Y. Li, J. Weng, J. Zhang, Feature selection for text classification: A review [J]. Multimedia Tools and Applications, vol. 78 (2019) p. 3797-3816.

[14] Chinchor, N., Named Entity Task Definition [C]. Proceedings of the Seventh Message Understanding Conference (1998) p. 137-142.

[15] D. Li, S. Luo, X. Zhang, F. Xu, A survey of named entity recognition methods [J]. Computer Science and Exploration, vol. 16 (2022) No. 9, p. 1954.

[16] W. Hu, G. Tian, Y. Kang, C. Yuan, and S. Maybank,Dual Sticky Hierarchical Dirichlet Process Hidden Markov Model and Its Application to Natural Language Description of Motions [C]. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40 (2018) No. 10, p. 2355-2373.

[17] H. L. Chieu and H. T. Ng, Named entity recognition with a maximum entropy approach [C]. Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL (2003), p. 160-163.

[18] J. Lafferty, A. McCallum, and F. C. N. Pereira, Conditional random fields: probabilistic models for segmenting and labeling sequence data [C]. Proceedings of the Eighteenth International Conference on Machine Learning (2001) p. 282-289.

[19] H. Zheng, X. Song, H. Yu, S. Li, Y. Hao, Review of Chinese Named Entity Recognition Based on Deep Learning [J]. Journal of Information Engineering University, vol. 5 (2021) p. 590-596.

[20] K. Liu, Q. Yu, and S. Zhong, Chinese Named Entity Recognition Based on Bi-directional Quasi-Recurrent Neural Networks improved with BERT: new method to solve Chinese ner [C]. 2021 5th International Conference on Innovation in Artificial Intelligence (ICIAI) (2021) p. 15-19.

[21] P. Zhu et al., Improving Chinese Named Entity Recognition by Large-Scale Syntactic Dependency Graph [C]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30 (2022) p. 979-991.

[22] I. J. Unanue, E. Z. Borzeshi, and M. Piccardi, Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition," Journal of Biomedical Informatics, vol. 76 (2017) p. 102-109.

[23] Z. Liu et al., Entity recognition from clinical texts via recurrent neural network [C]. BMC Medical Informatics and Decision Making, vol. 17 (2017) pp. 53-61.

[24] Z. Dai, X. Wang, P. Ni, Y. Li, G. Li, and X. Bai, Named Entity Recognition Using BERT BiLSTM CRF for Chinese Electronic Health Records [C]. 2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Suzhou, China (2019), p. 1-5.

[25] S. Kwon, Y. Ko, and J. Seo, Effective vector representation for the Korean named-entity recognition [J]. Pattern Recognition Letters, vol. 117 (2019) p. 52-57.

[26] Y. Miao, W. Cheng, Y. Ji, S. Zhang, Y. Kong, Aspect-based sentiment analysis in Chinese based on mobile reviews for BiLSTM-CRF [J]. Journal of Intelligent & Fuzzy Systems, vol. 40 (2021) No. 5, pp. 8697-8707.

[27] X. Tang, Y. Huang, X. Meng, C. Long, A Multi-Task BERT-BiLSTM-AM-CRF Strategy for Chinese Named Entity Recognition [J]. Neural Processing Letters (2022).

[28] H. -X. Lang, Y. -Y. Li, Y. Wang, H. Wang, and J. Dong, An Automatic Topic-oriented Structured Text Extraction Method based on CRF and Deep Learning [C]. 2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD), Hangzhou, China (2022) p. 1408-1413.

[29] Y. Liu et al., Construction of Knowledge Graph Based on Discipline Inspection and Supervision [C]. 2021 IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), Shenyang, China (2021) p. 1467-1472.

[30] X. Li and Q. Deng, Chinese Position Segmentation Based on ALBERT- BiGRU-CRF Model [C]. 2021 International Symposium on Computer Technology and Information Science (ISCTIS), Guilin, China (2021) p. 116-120.

[31] K. Han, Q. Wei, J. Qiu, Y. Cheng, Application of support vector machine (SVM) in the sentiment analysis of Twitter dataset," Applied Science [C]. tomatic classification of emotions in news articles through ensemble decision tree classification techniques," Journal of Ambient Intelligence and Humanized Computing, vol. 12 (2021) No. 12, pp. 5709-5720.