Hindawi
Scientific Programming
Volume 2022, Article ID 7682373, 14 pages
https://doi.org/10.1155/2022/7682373

Research Article

Named Entity Recognition for Public Interest Litigation Based on a Deep Contextualized Pretraining Approach

Hongsong Dong,[1] **Yuehui Kong**,[2] **Wenlian Gao**,[1] **and Jihua Liu**[1]

[1]*Department of Computer Science, Lüliang University, Lüliang 033000, China*
[2]*Center for Information and Modern Education Technology, Lüliang University, Lüliang 033000, China*

Correspondence should be addressed to Hongsong Dong; dong_hs@126.com

Received 10 December 2021; Revised 25 March 2022; Accepted 13 September 2022; Published 11 October 2022

Academic Editor: Dongpo Xu

The named entity recognition (NER) in the field of public interest litigation can assist prosecutors in handling cases and provide them with specific entities in making legal documents. Previously, the context-free deep learning model is used to catch the semantic comprehension, in which the static word vector is obtained without considering the context. Moreover, this kind of method relies on word segmentation technology and cannot solve the error transmission caused by word segmentation inaccuracy, which brings great challenges to the Chinese NER task. To tackle the above issues, an entity recognition method based on pretraining is proposed. First, based on the basic entities, three legal ontologies, NERP, NERCGP, and NERFPP are developed to expand the named entity recognition corpus in the judicial field. Second, a variant of the pretrained model BERT (Bidirectional Encoder Representations from Transformer) called BERT-WWM (whole-word mask)-EXT(extra) is introduced to catch the text character-level word vector hierarchical and the context bidirectional features, which effectively solve the problem of task boundary division of named entities. Then, to further improve the model recognition effect, the general knowledge learned from the pretrained model is used to fit the downstream neural network BiLSTM (bi-long short-term memory), and at the end of the architecture, CRF (conditional random fields) is introduced to restrict the label relationship. Finally, the experimental results show that the proposed method is more effective than the existing methods, which reach 96% and 90% in the *F*1 index of NER and NERP entities, respectively.

1. Introduction

Named entity recognition (NER) of legal cases is to identify the legal entities with specific meaning from the legal cases in the judicial field [1, 2]. From the coarse granularity, it mainly includes the name of the person, place name, the name of the institution, and the legal entity information related to the illegal subject, the illegal institution, the place where the case happened, and so on. The purpose of NER is to extract reference information such as names of people, institutions, and organizations from unstructured legal texts, which is to provide a basis for the construction of a structured database or element extraction and other tasks. On the one hand, legal named entity recognition technology can help legal professionals master the key content from the massive documents and improve their work efficiency. On the other hand, as the basis of legal artificial intelligence, it can provide support for the construction of an intelligent court, the realization of intelligent case decision prediction, the construction of a legal case database, and the knowledge map of the judicial field.

Tremendous scientific efforts have been made on the judicial NER task. In earlier studies, scholars employ rules-based and dictionary-based methods for named entity recognition research, in which rules and dictionaries are established manually, then punctuation marks, keywords, central words, indication information, location information, etc., are selected as features, and pattern matching is used as the main means to select corresponding entities from the text. For example, Sun et al. [3] employed a dictionary-based approach, in which the dictionary consists of the name-based word frequency table and the frequency table of

surnames, probability distribution of Chinese names (including single name and double name), names dictionary (including absolutely closed, relatively closed and open), and appellation and signifying verbs table. Then, the potential name table is obtained from the input sentences by the method of maximum matching, and the Chinese name entities are obtained by probability screening and correction rules. The recall rate of 99.77% is obtained from the data of Xinhua News Agency News Corpus. This kind of method needs to establish rules manually to distinguish entity types, which is poor in portability, and the effect depends on the size of the dictionary and the pattern rules. Moreover, it is challenging to construct large-scale dictionaries, and the update of dictionaries is also time-consuming and labor-consuming.

In recent years, with the development of machine learning technology, scholars have applied the method to the named entity recognition research. On the NER task in the generic domain, at present, there are supervised machine learning algorithms such as the conditional random fields (CRF) [4], the hidden Markov model (HMM) [5], and the support vector machine (SVM) [6]. Inspired by this, many scholars apply the above methods to the research of named entity recognition in the judicial field. To solve criminal cases, Chen [7] first set 13 types of entities, including people's names, addresses, methods, number of people, frequency, institutions, time, amount, and case names, and then mark some unique and useful information according to the characteristics of the judicial field (prompt words and boundary words). The above information is used to construct the common word list of criminal cases (including basic entity and case name entity). According to the word list, the entity keywords are preliminarily labeled as the training data of the model. Eight templates are defined manually, and features are extracted. Finally, the CRF model is selected to train the data, and the basic entity and case name entity types of the predicted data are finally obtained. However, this kind of method requires the manual definition of the template and manual definition of features. The design of the template and the selection of artificial features affect the learning performance of the model, and the strong dependence on corpus also restricts the application of this kind of method.

With the development of big data and deep learning, deep neural networks have surpassed classical machine learning methods in precision in many fields. Compared with the machine learning method, deep learning methods can effectively scale with data. Because its complex network structure can learn the characteristics of data, it does not need a lot of feature engineering. Therefore, it is more adaptable and easier to migrate. At present, many experts and scholars learn from the deep learning method in the general field and apply it to the task of NER in the legal document. In the judicial field, Leitner et al. applied BiLSTM [8–10] and CRF to the German corpus in the legal field and completed seven classes of coarse-grained recognition and 19 small classes of fine-grained recognition. BiLSTM achieved 95.95% and 95.46% of $F1$ values, respectively, while CRF achieved 93.22% and 93.23%, respectively. It shows that the BiLSTM method is superior to the CRF baseline model in different granularity legal domain corpora. Yin et al. [11] proposed a method of judicial named entity recognition combining CNN and [12, 13] LSTM. First, word embedding and CNN are used to obtain character-level embedding representation, and then, BiLSTM is used as an encoder, and one-way LSTM and character-level CNN are used as a decoder. Good performance is obtained on both the Chinese court decision data set and the CoNLL-2003 data set. Cardellino et al. [14] marked the judgments of the European Court of Human Rights and obtained three levels of legal ontology: NERC, LKIF, and Yago. They adopted a "three-step" strategy; first of all, neural networks with random weights are trained to distinguish between entities and nonentities. Then, when the classifier converges, the weights obtained are used to initialize another classifier with the same number of layers and neurons to identify the six types of entities. Finally, the classifiers are trained in the same way to recognize 69 subclass entities and 358 subclass entities. The above deep learning method based on CNN or LSTM is a one-way word embedding model in essence, unable to use context information, resulting in the decline of the recognition effect.

In the general domain, natural language processing technology can develop rapidly, largely due to transfer learning through pretrained models. The essence is to train the model on a large data set and fine-tune the model on the target data set to achieve different NLP capabilities, such as text classification [15, 16], factor extraction [17, 18], text generation [19, 20], and NER. Among them, the BERT model has promising results, which is constructed by Devlin et al. in 2018 [21]. Later, scholars optimized the BERT model and obtained RoBERTa [22], ALBERT [23], etc., by adopting different training strategies. These different pretrained models have achieved different effects on different tasks in different fields. It is difficult to say which one is optimal for all tasks, and the parameters need to be refined to suit the downstream tasks. Compared with CNN or LSTM models based on word embedding, this kind of model is a bidirectional self-coding language model, which considers the context of words and is more accurate in text understanding. For NER tasks in the legal field, we refer to pretrained methods for entity identification.

For the task of named entity recognition in the judicial field, although the existing deep learning-based work has achieved certain results, two areas need to be expanded or improved. On the one hand, the existing research on named entity recognition in the judicial field lacks corpus and is mainly focused on the field of criminal law, which is relatively simple, and there are few named entity recognition in other litigation fields. On the other hand, most of the existing named entity recognition methods in the judicial field use a context-free word embedding representation method, which does not consider the context in the understanding of the text, failing to understand the entity accurately. For example, a certain word usually has multiple meanings, such as "eldest son." Whether it refers to the entity name "Zhangzi County" or the nonentity "parents' eldest son" needs to be understood with the context.

Moreover, this kind of the context-independent text comprehension method is highly dependent on word segmentation technology, which will cause the error transmission problem caused by word boundary division, thus affecting the effect of named entity recognition.

Given this, we take "environmental protection" cases in litigation cases as data, establish a three-level NER system architecture, complete the NER study of "environmental protection" cases at the sample level, and expand the named entity identification database of litigation cases. Then, the context-dependent self-coding pretrained models are explored to improve the results of the named entity recognition task in judicial domain text understanding.

In summary, the contribution of the paper is as follows:

(1) A deep contextualized pretraining approach is designed for the Chinese public interest litigation named entity recognition. We have developed a set of NER standards related to the warning of environmental violations, established the corresponding corpus data set, and expanded the corpus of named entity identification in the judicial field. Three levels of legal ontology NERP, NERCGP, and NERFPP are constructed; the standard specification can meet the business requirements in the actual scene, such as the extraction of illegal fact elements.

(2) We explored the usefulness of the variant BERT model called BERT-WWM-EXT for the Chinese legal NER task. The character-level word vectors are obtained through the embedding strategy. This text representation strategy based on characters avoids the problem of the wrong demarcation of task boundaries caused by word segmentation. The context bidirectional features are extracted by the inner Transformer structure of the model, which carries rich semantic information of the text.

(3) We established a recognition module, by which the features learned from the pretrained language model translate to knowledge related to the tags of NER, and at the end of the module, CRF is introduced to restrict the tag relationship. Experiment results show that the proposed method achieves competitive results compared with other baseline models and the entity recognition rate of the model has been greatly improved combined with the recognition module.

The remainder of this paper is organized as follows: Section 2 describes related works; Section 3 describes the proposed method in detail; Section 4illustrates theexperimental results; and finally, Section 5concludes the paper.

2. Related Works

In this section, we review some works closely related to our study, including named entity recognition based on character and transfer learning based on pretraining.

2.1. Character-Based Methods. In different fields, the named entity recognition task is basic in information extraction, and experts and scholars have proposed many methods. In the aspect of English named entity recognition, the sequential labeling model based on the neural network has become the mainstream method, among which the combination of neural networks BiLSTM [9, 10] and CRF [4] is the most representative. Drawing on the achievements in the field of English named entity recognition, many experts apply this kind of method to Chinese named entity recognition. However, due to the huge differences between Chinese and English, each word is distinguished by spaces in English, while there is no natural space in Chinese text to segment each word. According to whether word segmentation is carried out, named entity recognition in the Chinese field can be divided into word-based and character-based methods.

Among them, the word-based method first needs word segmentation technology to distinguish between the words; the commonly used word segmentation technology includes Hagongda word segmentation and Jieba word segmentation—different word segmentation technologies can be different according to different fields of text. It is necessary to choose the appropriate word segmentation technology and then carry out named entity recognition. The boundary between words after participle is also the boundary of the entity.

The model of neural network CNN [12, 13] or BiLSTM [9, 10] combined with CRF is based on the recognition model after word segmentation, and then, word2vec is used for word vector representation. To improve the effect of entity recognition, some scholars [24, 25] use semantic information to improve the word vector as a result. Although this kind of method has achieved some results, due to the influence of correct word segmentation, there will be the problem of the wrong transmission of subsequent results. The reason is that if the word segmentation is not accurate, then the named entity division based on the word segmentation must be inaccurate.

The character-based named entity recognition task does not need word segmentation. In this method, each word is treated as an independent individual, and each word is divided into named entities according to the annotation technology of named entity recognition, so there is no error transmission problem caused by word segmentation technology. Therefore, some experts and scholars proposed a Chinese named entity recognition model based on character level [26, 27]. The main shortcoming of this model is that it cannot make use of word information. In this model, to improve the recognition effect, experts pay attention to how to make better use of word information [28–30]. Current studies have found that the character-based approach is superior to the word-based approach in the field of Chinese [31].

2.2. Pretraining. Our work is closely related to the transfer learning-based pretrained model. In 2018, with the advent of the BERT model [21], the neural network method has been improved to a new height. This kind of method does not need word segmentation technology, so there is no

subsequent error transmission problem. Moreover, compared with the word2vec-based model, the word ambiguity problem can be solved due to its two-way coding method. Subsequently, experts and scholars made improvements based on BERT and obtained many BERT-based models, among which the excellent ones are RoBERTa [22], AlBERT [23], etc. In our earlier work [17], we find that RoBERTa achieves the best result on the factor element extraction task. However, these models have different effects in different areas of natural language processing, and no one model has been found to achieve overwhelming results for all tasks and all areas.

For the sake of illustration, the Chinese version is used as an example for the subsequent models. To get a better effect, scholars adopt some techniques to improve the BERT, RoBERTa, and AlBERT models. For example, [32] constructed BERT-WWM and BERT-WWM-EXT models by adopting a full-word mask strategy. Through experiments, they found that mask strategy skills and the adoption of richer pretrained corpus could improve the results of various downstream tasks. Similarly, based on RoBERTa, the joint laboratory of Xunfei has released the Chinese RoBERTa-WWM-EXT pretrained model with the combination of Chinese whole-word masking technology and RoBERTa model. RoBERTa-WWM-EXT combines the advantages of Chinese whole-word masking technology and RoBERTa model to achieve better experimental results. It is worth noting that the WWM strategy is used for masks during the pretrained phase (but no dynamic masking was used).

Although those BERT-based networks have made a series of breakthroughs in many different tasks, exploring an optimal model for our NER task in the legal domain and achieving better results are important for our work.

3. Proposed Method

A method for named entity recognition in the judicial domain based on pretraining transfer learning and deep feature extraction is proposed, and the influence of different pretrained models on the downstream named entity recognition task is explored. The performance of the model is further improved by selecting an appropriate pretrained model and feature extractor. The proposed method includes two stages: pretraining and fine-tuning. In the pretraining stage, the weight of fitting parameters containing lexical, syntactic, and semantic information is obtained by training a large number of unsupervised data sets. In the fine-tuning stage, the parameters of the pretrained model are loaded as initialization instead of random initialization. The proposed method is shown in Figure 1, including the following steps:

(i) *Self-coding pretrained model*. Different models are trained on a different large general corpus with different training strategies according to MLM (masked language modeling) or NSP (next sentence prediction) task, and weight parameters suitable for NER are constructed.

(ii) *Data set annotation*. First, according to the data of "environmental protection" in civil litigation cases, the legal entity is constructed. Then, the data are labeled entity, and the supervised sample is obtained. The details are shown in Section 4.

(iii) *Word vector embedding*. The legal text is encoded by the word vector embedding layer to obtain the input representation.

(iv) *Feature representation based on the Transformer model*. The input representation is further encoded with the Transformer module to get the feature representation with context information.

(v) *Recognition components*. The components consist of deep feature extraction based on BiLSTM and linear output layer, and CRF restricts label relationships.

3.1. Self-Coding Pretrained Model. This section mainly discusses the BERT series, RoBERTa series, and AlBERT series of several common auto-encoding models, and explores the migration ability of auto-encoding pretrained models under different mechanisms for the downstream named entity recognition task. Since the data set in this chapter is Chinese-oriented, the following mainly involves the Chinese version of the model.

Based on the self-encoding language model, the context-related bidirectional feature representation is obtained by introducing noise [MASK]:

$$p(x|\widehat{x}) \approx \prod_{i=1}^{n} m_i p(w_i|\widehat{x}). \tag{1}$$

In the above formula, $m_i$represents whether the current word is masked. In essence, it is a kind of joint probability estimation. In the training phase, instead of using the current word to predict the next word, we use [MASK] to cover the words in the sentence and use the context to predict what the masked word is.

The whole-word mask strategy is obtained based on the auto-coded language [MASK]; that is, the same MASK strategy is made for different characters belonging to a certain word in the input sample so that BERT-WWM and RoBERTa-WWM can be obtained. BERT-WWM-EXT and RoBERTa-WWM-EXT are obtained based on WWM and by changing the amount of pretraining data. Both of these strategies can improve the effect of downstream tasks.

For RoBERTa, compared with BERT, fine-tuning and training strategies are mainly carried out. The training strategies mainly include training time, training batch size, training data volume, NSP task removal, and dynamic mask. AlBERT, compared with BERT, there are two improvements: one is to reduce the memory consumption of the model by reducing parameters, and the other is to replace the NSP task with a sentence order prediction model to improve the performance of the downstream task.

3.2. Word Vector Embedding. For BERT, RoBERTa, or AlBERT models based on pretraining, to obtain the representation of sentences and the meaning of words with context, sentences are processed during design. For example,

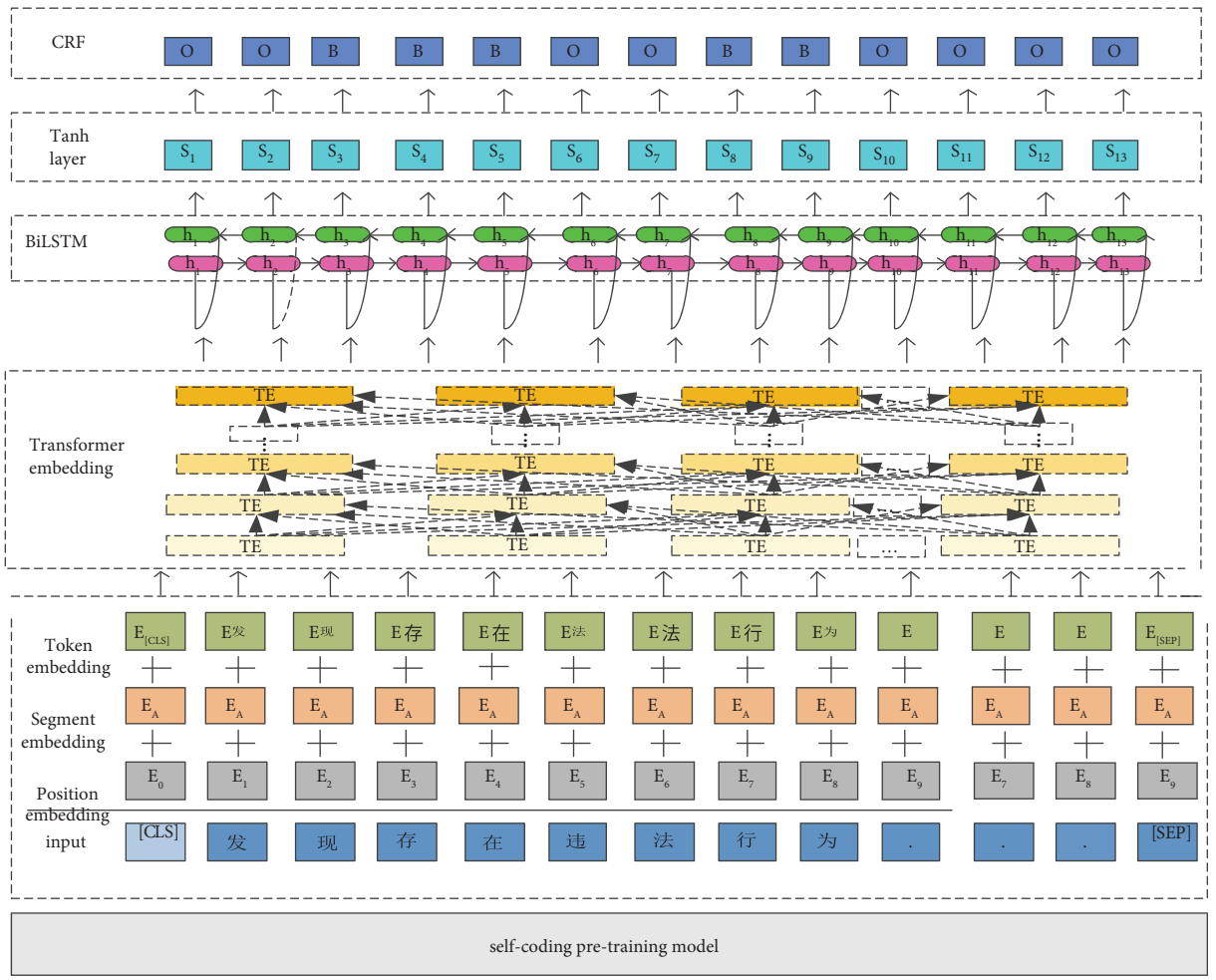

Figure 1: Method for named entity recognition in the judicial domain based on pretraining transfer learning and deep feature extraction.

identifiers [CLS] and [SEP] are added to represent the beginning and end of sentences, respectively. The word and identifier in each sentence are called "token." Sentence encoding is token encoding essentially, which involves three types of computation, namely, token embedding, segment embedding, and position embedding. The word embedding representation of this kind of model is obtained by adding three kinds of embedding. The word vector representation of the input data is essentially the shallow encoding of the text by the embedded matrix in the pretrained model. Suppose a sentence in the sample is "found to have an illegal act," then the representation of the sentence is as shown in Figure 2.

3.3. Feature Representation. After word vector embedding, the output serves as the input of the feature representation part. For the BERT, RoBERTa, and AlBERT pretrained models, the deep coding nature is based on Transformer [33]. Yet for AlBERT, the implementation process is more focused on algorithmic efficiency. The three variants of the three models (normal version, WWM version, and WWM version) based on extensible data EXT are all composed of the L-layer Transformer structure in terms of model structure.

Suppose the input text "Found the above violation. . .," which is denoted as $T = (T_1, T_2, \ldots, T_n)$, after word vector embedding $S = (S_1, S_2, \ldots, S_n)\, S = (S_1, S_2, \ldots, S_n)$ is obtained. Let $H^L = (h_1^L, h_2^L, \ldots, h_n^L)$ be the output of the Transformer layer L, as shown in the following:

$$H^L = \text{Transformer}^L\left(H^{L-1}\right). \tag{2}$$

In the above formula, when L is equal to 1, $H^1 = \text{Transformer}^1(S)$. H^Lrepresents the output of the representative word finally encoded by the Transformer model. The structural schematic diagram is shown in Figure 3.

3.4. Recognition Components. In the components, NER is taken as the three-level tag BIO sequence labeling task, BiLSTM is used to predict the tags of the sequences, and CRF defines the relationship between the tags considering the correlation of the tags. Take the output of formula (2) as the input for this component. The H^L is input to the BiLSTM layer, and BiLSTM is composed of the forward LSTM and the backward LSTM. The forward LSTM extracts the feature of the input to get the left representation h_{lt} of the current time t, while the backward LSTM extracts the feature of the

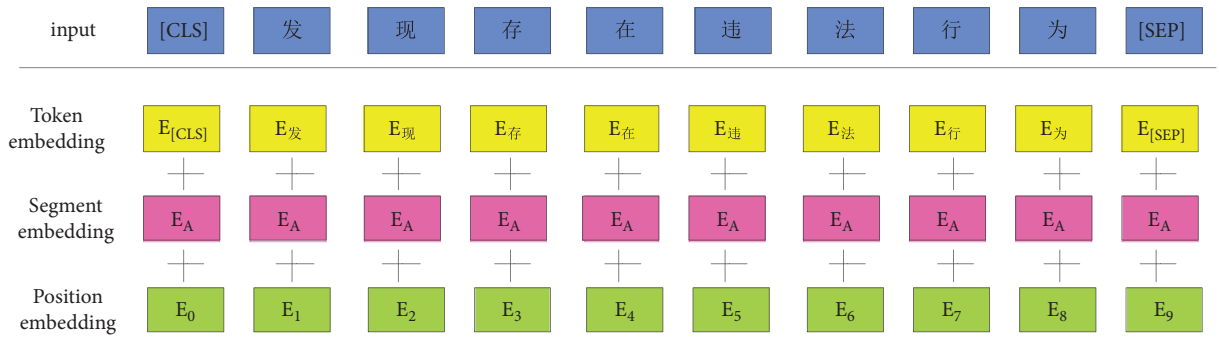

Figure 2: Word vector embedding.

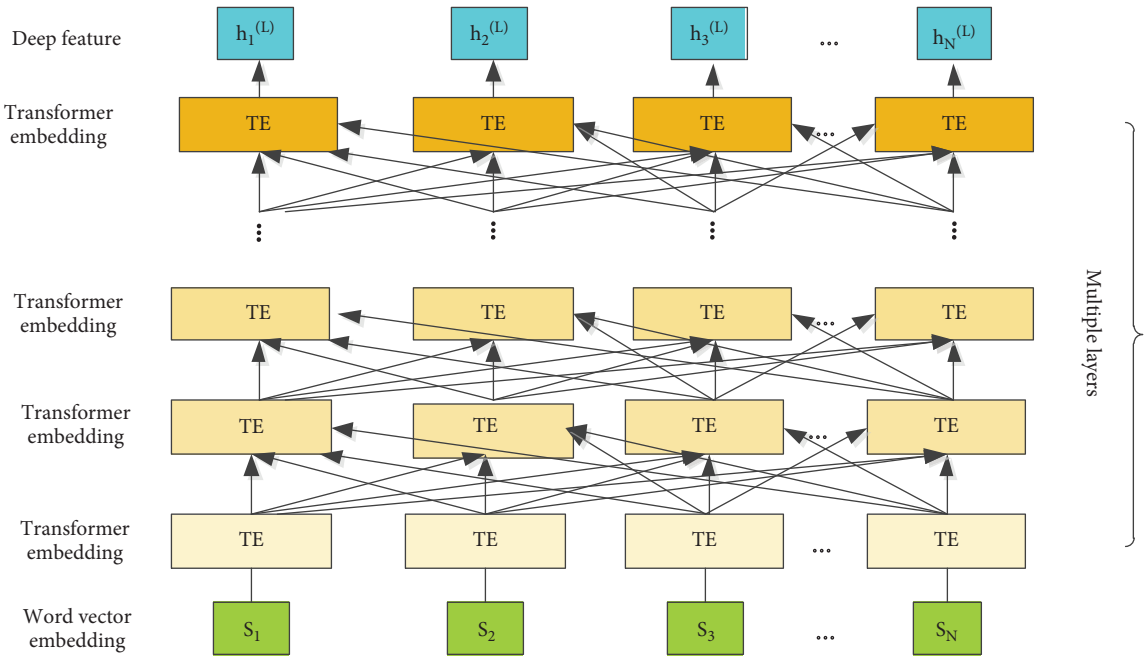

Figure 3: Transformer structure.

input to get the right representation h_{rt} of the current time t, and the output of BiLSTM is the concatenation of the two:

$$h_t = \text{concact}(h_{lt}, h_{rt}). \tag{3}$$

Finally, through the tanh linear output layer, the sequence h_t is mapped to the s dimension, where s is the number of tags contained in the tag set, where the BIO category is 3. Suppose that the final output list = $(l_1, l_2, \ldots, l_n)$, where each element $l_{i,j}$ in the column vector l_i represents the score of the ith word's tag j.

In practice, if the score value is directly mapped to the prediction probability obtained by the classification layer, the correlation between labels cannot be considered, but the local optimum cannot reach the global optimum, so it needs to consider CRF to limit it. Two elements are considered here: one is the output of the BiLSTM layer, and the other is the relationship between the outputs. Based on this, two matrices are defined: the output matrix L of BiLSTM and the label state transfer matrix Q. The element $L_{i,j}$ in L represents the score of a word $w_i w_i$ labeled j. The element $Q_{i,j}$ in Q represents the transition probability, from which the tag is transferred from tag_i to tag_j. The network layer output score and the state transition probability score are added as the final network output score, which is shown as follows:

$$s(X, y) = \sum_{i=0}^{n} Q_{y_i, y_{i+1}} + \sum_{i=1}^{n} L_{y_i, y_i}, \tag{4}$$

where $s(X, y)$ represents the score of Y for the predicted output sequence of sentence X.

After the normalized softmax function, a probability is defined for each correct sequence Y:

$$p(y|X) = \frac{e^{s(X,y)}}{\sum_{\overline{y} \in Y_X} e^{s(X,\overline{y})}}, \tag{5}$$

where Y_X represents all possible sequence cases. During the training, the correct probability is just maximized, maximizing formula (5). The logarithm is applied to the both sides of the above formula to get the following:

$$\log\left(p(y|X)\right) = \log\left(\frac{e^{s(X,y)}}{\sum_{\overline{y}\in Y_X} e^{s(X,\overline{y})}}\right) = s(X,y) - \log\left(\sum_{\overline{y}\in Y_X} e^{s(X,\overline{y})}\right). \tag{6}$$

The loss function is defined as $-\log\left(p(y \mid X)\right)$. During the training process, the maximum-likelihood estimation method is used to estimate the model parameters, and the stochastic gradient descent method is used to optimize the model parameters.

4. Experiments

4.1. Dataset and Data Preprocessing. Using datasets from the procuratorial organ and related website (https://www.itslaw.com/home), the legal document is obtained, which contains 1000 copies and mainly involves the public interest litigation cases "environmental pollution control" type, including air pollution prevention, water pollution control, and solid waste pollution prevention and control. The document is a structured legal document, including the name of the administrative organ to be supervised, the source of the case or the description of the case, the facts of the omission of the administrative organ to be identified, the reasons and legal basis for the proposed administrative authority that constitutes an illegal exercise of its powers, and the specific content of the suggestion, as shown in Figure 4.

We only use the description of the case section of the document, which includes the entities involved. The original data were in Word format and annotated in TXT after processing. The training sample, validation sample, and test sample are divided by 8 : 1 : 1. The labeled data sets include air pollution prevention and control data, water pollution prevention and control data, and solid pollution prevention and control data, with their distribution shown in Figure 5.

For the legal named entity recognition, the character-based annotation method is adopted. The annotation mode selected three markings of IOB {I, O, B}. O, I, and B represent the nonentity, the beginning of entity, and the end of entity, respectively. The main task is to add "environmental pollution prevention and control" entity recognition based on the original name recognition of people's names, place names, and institutions. There are the "air pollution prevention and control" type, "water pollution prevention and control" type, and "solid waste pollution prevention and control" type, which are the first-level NERP of the legal ontology, and the first-level entity identification has been completed. The second-level NERCGP and the third-level NERFPP of legal ontology are constructed by fine-grained division, as shown in Table 1. In terms of labeling strategy, {PER, LOC, ORG, GAS, WATER, SOLID} is proposed based on people's name, place name, and organization name in the format of People's Daily data set.

By combining the I-O-B schema with the entity name {PER, LOC, ORG, GAS, WATER, SOLID}, we can get the annotated data, for example, B-LOC for the beginning of the place name, and I-LOC for the end of the place name.

In performing its duties, the hospital found that there were rubble and bricks on the public passageway and green space at the north gate of XX, which was located within the jurisdiction of the people's Government of XX Town, XX District, XX City, and accumulated wastes and garbage, which had a serious impact on the surrounding city appearance and environment and infringed on the social and public interests.
The court conducted an investigation in accordance with the law.Ascertain now: your government is responsible for and undertake the city appearance environment health management duty within this jurisdiction specifically, exerise the power of punishment to the city appearance environment health illegal behavior.
For a period of time, there are rubble and bricks on the public passageway and green space at the north gate of XX, and the accumulation of wastes and garbage, resulting in the pollution of the city appearance and environment and the hidden danger of public safety, and the above situation has not been effectively treated.
Above fact has the evidence such as the scene photography that the procuratorial organ takes, scene investigation verifies working record to prove.
The court believes that in order to maintain the city's appearance and environmental health and prevent garbage from polluting the environment, the relevant laws and regulations have strict regulations on the construction and piling up of public place.
The above accumulation of waste and garbage on the public passageway and green space is in violation of the provisions of article 25, Paragraph 3, of the Regulations on Urban Appearance and Environmental Sanitation of XX City. Your government should order the offender to stop the illegal act, correct within a time limit and impose a fine.

Figure 4: Example in the document.

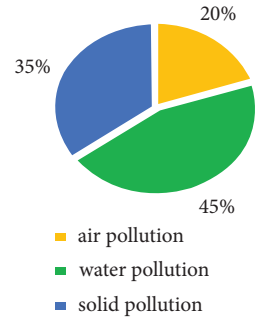

Figure 5: Distribution of data types in data sets.

After manual annotation, the character-based annotation data obtained are shown in Figure 6.

4.2. Parameter Settings and Evaluation Metrics. In the experiment, the BIO corpus and BIODSG corpus are obtained according to the annotated specification of the People's Daily data set. To ensure fairness, cross-validation is done on the validation set to determine the optimal model hyperparameters. The main superparameters of the pretrained model are shown in Table 2. For the input part, the main hyperparameter is the word vector dictionary dimension, and its value is 21128. For the model, the hyperparameter mainly involves the hidden layer dimension H of 768. For BiLSTM, the parameter is the dimension of the hidden layer, which is 384. The early stop strategy is used to select the optimal number of iterations on the verification set, and the index F is used as the evaluation method on the test set. The learning rate and other parameters of the model are shown in Table 2.

4.3. Comparison of Experimental Results. In order to verify the effectiveness of the pretrained model for subsequent

Table 1: Division of legal entities.

Entity name	NERP	NERCGP	NERFPP
PER	Name	Name	Name
LOC	Place name	Place name	Place name
ORG	Organization name	Organization name	Organization name
GAS	Air pollution	Industrial air pollution ... Air pollution caused by coal-burning ... Dust air pollution ... Agricultural air pollution ... Domestic air pollution	Refuse to accept supervision and inspection by environmental law enforcement departments ... Import and sale of coal and petroleum coke that do not meet the quality standards ... The construction site is not equipped with hard enclosures or effective dust and dust control measures ... In densely populated areas, trees, flowers, and plants are sprayed with highly toxic and highly toxic pesticides ... Set off fireworks in prohibited areas and at prohibited times ...
WATER	Water pollution	Industrial water pollution ... Urban water pollution ... Agricultural and rural water pollution ... Drinking water source pollution ... Disposal of water pollution accident	Discharge of water pollutants without a legally obtained discharge permit ... The discharge of oil, acid, lye, and highly toxic, radioactive, and pathogen-containing wastewater into water bodies ... Effluent from livestock, poultry, and aquaculture causes water pollution ... Sewage outlets shall be set up in drinking water sources ... Failing to formulate emergency plans for water pollution accidents under regulations ...
SOLID	Solid pollution	Pollution by industrial solid waste ... Construction waste pollution ... Agricultural solid waste pollution ... Hazardous waste pollution ... Domestic waste pollution	Dumping, stacking, and discarding industrial solid waste without authorization cause environmental pollution ... The construction unit has not prepared the construction waste disposal plan and put it on record ... Livestock and poultry farms and farming areas use food waste without harmless treatment to feed livestock and poultry ... Failing to formulate hazardous waste management plans or report relevant information as required ... Violate the garbage classification regulations to put household garbage ...

现 O 查 O 明 O ： O 宁 B-LOC 陵 I-LOC 县 I-LOC 清 I-LOC 水 I-LOC 河 I-LOC 老 I-LOC 道 I-LOC 城 I-LOC 郊 I-LOC 段 I-LOC 起 O 点 O 城 B-LOC 郊 I-LOC 乡 I-LOC 耿 I-LOC 庄 I-LOC 村 I-LOC 北 O ， O 终 O 点 O 至 O 城 B-LOC 郊 I-LOC 乡 I-LOC 李 I-LOC 庄 I-LOC 村 I-LOC ， O 河 O 流 O 两 O 侧 O 大 O 量 O 排 O 入 O 生 B-S 活 I-S 污 I-S水 I-S、 O 生 B-G 活 I-G垃 I-G圾 I-G。

Now O Find O out O: O Ning B-LOC ling I-LOC County I-LOC Qing I-LOC shui I-LOC River I-LOC Lao I-LOC dao I-LOC suburban I-LOC section I-LOC of the starting point O of suburban B-LOC town I-LOC Geng I-LOC Zhuang I-LOC village I-LOC north O, O The O end O point O to O suburban B-LOC Town I-LOC Li I-LOC Zhuang I-LOC village I-LOC. O Both O Sides O of O the O river O discharge O a O large O amount O of O living B-S sewage I-S and household B-G Garbage I-G.

Figure 6: Annotation data.

Table 2: Parameter setting.

Parameters	Value
Vocabulary size	21128
BERT hidden size	768
BERT attention heads	12
BERT layers	12
BiLSTM hidden size	384
LSTM layers	2
Learning rate	0.0001
Pretraining word embedding size	768

Table 3: Comparison of experimental effects between various models (NER category).

Methods	*P*	*R*	*F*1
CRF [34]	67.7	68.7	68.2
BiLSTM-CRF [34]	69.8	70.6	70.2
WL-BiLSTM-CRF [34]	70.8	71.6	71.2
CNN-BiLSTM-CRF [35]	72.2	72.2	72.2
BERT-FC [36]	75.5	71.7	73.5
BERT-WWM-FC	**77.8**	72.6	75.1
BERT-WWM-EXT-FC	75.3	**75.9**	**75.6**
RoBERTa-FC	74.3	68.3	71.2
RoBERTa-WWM-FC	74.3	69.3	71.7
RoBERTa-WWM-EXT-FC	75.8	70.0	72.8
ALBERT-FC	73.8	64.0	68.6
ALBERT-tiny-FC	68.7	52.9	59.8

The bold values mean the optimal values among all the methods.

Table 4: Comparison of results between each of the pretrained models (NER category).

Methods	*P*	*R*	*F*1
BERT-FC [36]	75.5	71.7	73.5
BERT-BiLSTM-CRF [36]	94.0	94.0	94.0
BERT-WWM-FC	77.8	72.6	75.1
BERT-WWM-BiLSTM-CRF	95.2	95.2	95.2
BERT-WWM-EXT-FC	75.3	75.9	75.6
BERT-WWM-EXT-BiLSTM-CRF (proposed method)	**96.0**	**96.0**	**96.0**
RoBERTa-FC	74.3	68.3	71.2
RoBERTa-BiLSTM-CRF	95.2	94.8	95.0
RoBERTa-WWM-FC	74.3	69.3	71.7
RoBERTa-WWM-BiLSTM-CRF	95.4	95.0	95.2
RoBERTa-WWM-EXT-FC	75.8	70.0	72.8
RoBERTa-WWM-EXT-BiLSTM-CRF	95.5	94.7	95.1
ALBERT-FC	73.8	64.0	68.6
ALBERT-BiLSTM-CRF	90.0	90.0	90.0

The bold values mean the optimal values among all the methods.

NER tasks and find out the most appropriate pretrained model, NER tests are carried out on the three types of NER entities, respectively, for 8 different models, and the experimental effects are observed in combination with BiLSTM-CRF, respectively, and compared with the BiLSTM-CRF baseline model and CRF model based on word2vec. The experimental results are shown in Tables 3 and 4. In the experiment, the same data are selected for a test to ensure the fairness of the results.

In comparison with existing methods, WL-BILSTM-CRF [34] is an improved BiLSTM-CRF model that integrates word vectors and LDA topic vectors, compared with the word2vec-based word vector model; this model can not only obtain richer semantic features but also obtain the characteristics of word sequence and topic coherence, making full use of the advantages of LSTM for serialization tasks.

Wang [35] proposed a CNN-BiLSTM-CRF model to recognize the nine key elements in the judgment document. Compared with the baseline model BiLSTM-CRF, before the overall features of the text context information are obtained, the model added a convolutional neural network to obtain the local features at the character level of the word vector, which improved the recognition rate of the model. There is a 2% improvement in the overall *F*1 value compared with that in the baseline model. Zhao et al. [36] proposed to use BERT for named entity recognition of Chinese attractions, combining BERT with BiLSTM-CRF, using BERT to obtain the vector matrix of word granularity, combining BERT with BiLSTM to extract context features, and finally using CRF to obtain the optimal tag sequence. In order to make a fair comparison, the annotated data are used to test the above models, and the comparison results are divided into four situations.

4.3.1. Comparison between Existing Models and CRF-Based Models. Table 3 shows that machine learning methods based on the CRF class exhibit the worst result in *F*1 value, and the method based on BiLSTM-CRF improves by 2.6% compared with the CRF method. This is mainly because the effect of the CRF-based method is mainly dependent on the design of features and feature templates, and the incomprehensibility of features affects the performance of the model. In the method based on BiLSTM-CRF, the semantic features of a text can be automatically learned by using the word2vec-based word embedding method, and bidirectional features of a text can be extracted by combining with BiLSTM. As the result, entity identification is improved.

Compared with the CRF model, the overall *F*1 is improved by 5.3% and 3%, respectively, by the method based on BERT-FC and RoBERTa-FC, which indicates the effectiveness of automatic feature extraction of the pretrained model. Through the introduction of transfer learning, the model's understanding of the text context is enhanced, and the ability to identify ambiguous entities is enhanced. It is worth noting that BERT-WWM-EXT-FC has the highest *F*1 value.

4.3.2. Comparison between Existing Models and BiLSTM-CRF-Based Models. Based on the baseline model BiLSTM-CRF, the CNN module is added to obtain the CNN-BiLSTM-CRF model. In Table 3, compared with the baseline model BiLSTM-CRF, the overall *F*1 value of this model is increased by 2%, indicating that the capability of model feature characterization is strengthened after the combination of CNN modules. In addition to obtaining context information, CNN obtains the local features of text characters, which reflects the complex dependence between words. The acquisition of local feature information and context overall information makes the text representation

more perfect and finally improves the result of named entity recognition.

Compared with BiLSTM-CRF, *F*1 of BERT-FC and RoBERTa-FC improved by 3.3% and 1%, respectively, which demonstrates the effectiveness of the transfer learning ability of the pretrained model and solved the problem of the decline in feature extraction ability caused by the failure of the BiLSTM model to utilize context information depending on word2vec.

4.3.3. Comparison between Eight Pretrained Models. In essence, the above models based on CRF, BiLSTM-CRF, or CNN-BiLSTM-CRF show a poor recognition effect due to the limitations of their feature extraction methods. For example, the CRF-based method relies on feature design, and the model of BiLSTM-CRF or CNN-BiLSTM-CRF essentially employs word2vec technology to obtain word vectors. The representation of word vectors is fixed, which cannot solve the poly-meaning problem of words in the text, and the ability of feature expression needs to be improved. On the contrary, to obtain word vectors, word segmentation technology is used to perform word segmentation on the text, and the inaccuracy of word segmentation directly affects the division of the boundary of named entity recognition.

From Table 3, compared with CRF, BiLSTM-CRF, and CNN-BiLSTM-CRF, the named entity method based on the BERT pretrained model improves the overall *F*1 by 5.3%, 3.3%, and 1.3%, respectively. This indicates that the feature extraction ability of the BERT pretrained model is better than that of the word embedding method, and the representation of the text is polysemy information that can represent words, which is a dynamic word vector representation method. In addition, different types of pretrained models have different results on NER tasks. First, the results of the basic version are compared with those of WWM and WWM-EXT. From Table 3, it can be found that the recognition results based on WWM-EXT are the best, followed by WWM and the basic version. For example, compared with BERT-WWM and BERT-based methods, the *F*1 value of BERT-WWM-EXT is increased by 0.5% and 2.1%, which is mainly because the pretrained model adopts a richer pretraining corpus, and the weight parameters obtained based on such rich corpus are obtained. It can better fit the downstream NER task.

On the contrary, compared with the BERT basic version, the WWM mask technology that is adopted takes into account that the smallest unit in Chinese is "word" granularity and gives the same mask to different characters in a word. This understanding based on word granularity is transferred to the downstream NER; that is, the understanding of Chinese words is grasped in the pretraining stage. The fine-tuning phase grasps the understanding of character granularity.

As for the comparison between the RoBERTa model and the BERT model, in terms of the entity recognition task in this field, the BERT model is 2.8% better than the RoBERTa model on average. Although the RoBERTa pretrained model is better than BERT, it is not overwhelmingly stronger than the BERT model in all areas of all tasks, so it can be concluded that the appropriate pretrained model should be selected according to the task and domain.

As for the comparison between the AlBERT model and the BERT model, it can be seen from Table 3 that the recognition results of AlBERT are somewhat lower compared with those of BERT. This is mainly because AlBERT mainly adopts optimization strategies aimed at reducing the training complexity and training parameters, but it does not contribute much to the improvement of downstream tasks.

Table 5: Comparison of effects between various models (NERP category).

Methods	*P*	*R*	*F*1
BERT-FC [36]	72.0	72.8	72.4
BERT-BiLSTM-CRF [36]	89.0	89.0	89.0
BERT-WWM-EXT-BiLSTM-CRF (proposed method)	**89.3**	**90.7**	**90.0**
RoBERTa-FC	72.3	67.8	70.0
RoBERTa-BiLSTM-CRF	85.2	89.5	87.3
RoBERTa-WWM-EXT-BiLSTM-CRF	88.7	87.9	88.3
ALBERT-FC	66.2	67.4	66.8
ALBERT-BiLSTM-CRF	88.0	88.4	88.2

The bold values mean the optimal values among all the methods.

4.3.4. Comparison between Eight Pretrained Models in Combination with BiLSTM-CRF. The eight pretrained models are combined with the baseline model BiLSTM-CRF, which had the best performance among all the methods (CRF, BILSTM-CRF, and CNN-BILSTM-CRF). The results are shown in Table 4. It fully demonstrates the strong transfer learning ability and the representation ability of text features of the pretrained model, which is reflected in the fine-tuning task, namely, the increase in the NER recognition rate. However, pretrained models with different strategies have different transfer learning abilities.

In the comparison of results between several different pretrained models, it can be found that the method based on BERT-WWM-EXT-BILSTM-CRF has the best effect, and its overall recognition of *F*1 reaches 96%. The recognition results based on RoBERTa-WWM-EXT-BILSTM-CRF reach 95.1%. This optimal result is due to the powerful feature representation ability of BERT-WWM-EXT, which takes into account both polysemous problems and word boundary problems. On the contrary, combining BERT-WWM-EXT with BiLSTM-CRF can enhance the understanding of contextual information and the selection of the optimal tag sequence. Finally, the optimal recognition result is obtained.

4.4. The Impact of Data Label Categories on Results. To further verify the validity of the model, in addition to conducting experiments on the three entity categories of NER, this section identifies the four entities of NERP. The identification results are shown in Table 5. Similar to the results of the identification of three types of entities of NER, the result based on BERT-WWM-EXT-BILSTM-CRF is also the best. However, compared with the identification of three

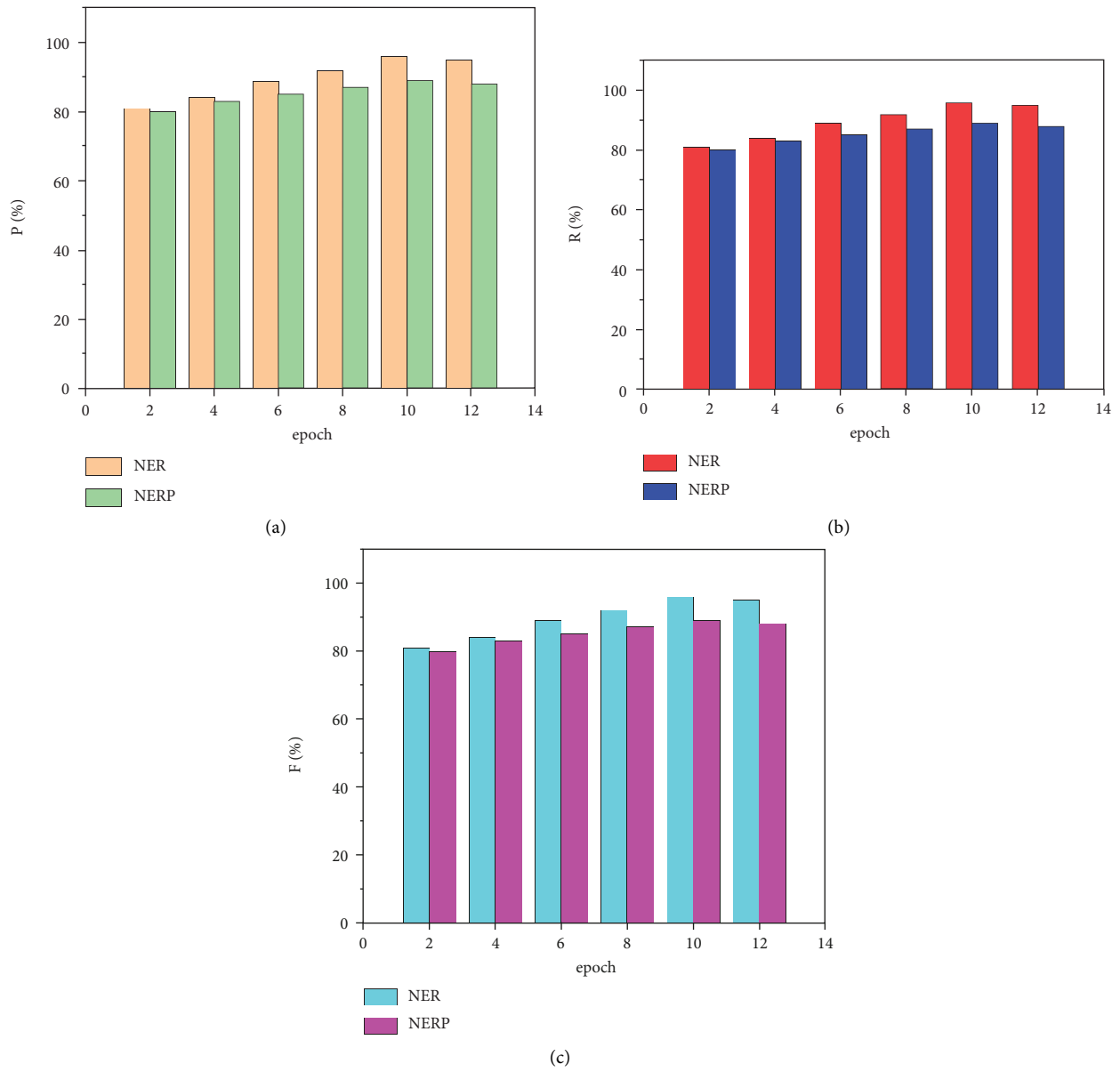

Figure 7: Changes in the NER and NERP categories of the proposed method with different epochs on *P*, *R*, and *F*1 indexes.

entities of NER, the *F*1 value of the identification results decreases in each method. For example, the method based on BERT-WWM-EXT-BILSTM-CRF improved by 1% compared with the method based on BERT-BILSTM-CRF, because the extra data and mask strategy used in the pretrained model guaranteed a stronger data migration capability. Then, BERT is used to obtain the word vector and express the text at a deeper level. Although the task is more difficult by adding entities, it still solves the polysemy problem and does not affect the feature extraction. Compared with the NER three entity recognition tasks, the NERP four entity recognition tasks added entities, which reduced the recognition results to different degrees. For example, the BERT-WWM-EXT-BILSTM-CRF method decreased by 6%, but still achieved 90% recognition results, which was still the optimal result.

4.5. The Influence of the Number of Iterations on the Results. To further illustrate the effectiveness of the pretrained model, Figure 7 shows the *P*, *R*, and *F*1 indexes of the proposed method BERT-WWM-EXT-BILSTM-CRF in NER and NERP categories with different iterations. As can be seen from the figure, during the whole iteration process, all indicators are rising and reach the optimum when the epoch is 10.

Comparing the changing process of the two categories, it is found that the recognition result of the NER class is better than that of the NERP class because NERP class adds one more environmental entity class than the NER class, which increases the task difficulty. During the whole iteration, the recognition performance of NERP class is worse than that of the NER class. However, the optimal value is also obtained at the epoch of 10.

Table 6: Result of the 10-fold cross-validation test at the NER level.

Fold	Method	P	R	$F1$
1	BERT-W-E-LSTM-CRF	95.8	95.6	95.7
2	BERT-W-E-LSTM-CRF	96.2	96.5	96.3
3	BERT-W-E-LSTM-CRF	95.4	96.4	95.9
4	BERT-W-E-LSTM-CRF	96.5	96.4	96.4
5	BERT-W-E-LSTM-CRF	95.3	95.5	95.4
6	BERT-W-E-LSTM-CRF	96.6	95.8	96.2
7	BERT-W-E-LSTM-CRF	96.3	96.8	96.5
8	BERT-W-E-LSTM-CRF	96.6	95.8	96.2
9	BERT-W-E-LSTM-CRF	95.8	96.2	96.0
10	BERT-W-E-LSTM-CRF	95.5	95.0	95.0
Ove	BERT-W-E-LSTM-CRF	96 ± 0.48	96 ± 0.52	96 ± 0.45

Notes: Ove means the overall performance (mean ± std).

Table 7: Result of the 10-fold cross-validation test at NERP level.

Fold	Methods	P	R	$F1$
1	BERT-W-E-LSTM-CRF	89.6	90.5	90.0
2	BERT-W-E-LSTM-CRF	89.9	90.6	90.2
3	BERT-W-E-LSTM-CRF	89.2	90.3	89.7
4	BERT-W-E-LSTM-CRF	88.8	90.7	89.7
5	BERT-W-E-LSTM-CRF	88.7	90.8	89.7
6	BERT-W-E-LSTM-CRF	89.6	90.6	90.1
7	BERT-W-E-LSTM-CRF	88.9	90.8	89.8
8	BERT-W-E-LSTM-CRF	89.5	91.2	90.3
9	BERT-W-E-LSTM-CRF	89.6	91.0	90.29
10	BERT-W-E-LSTM-CRF	89.2	90.5	89.8
Ove	BERT-W-E-LSTM-CRF	89.3 ± 0.38	90.7 ± 0.25	90 ± 0.24

Notes: Ove means the overall performance (mean ± std).

4.6. Statistical Test. Several statistical tests are performed to validate the performances. We listed the results of each fold of our proposed method BERT-WWM-EXT-BILSTM-CRF, which are shown in Tables 6 and 7. Table 6 shows the results of each fold of our proposed method at the NER level, from which we can find that the proposed method is steadily superior to other methods in the $F1$ value. We further conducted a t test for statistical significance tests. The performance of the proposed method is also significantly better than other models ($P < 0.05$), and the average error rate of the proposed method is smaller and the performance is the best. The experimental results demonstrate that combining BERT-WWM-EXT representations and the extracting components BiLSTM-CRF can steadily elevate the performance. Table 7 is the results of each fold of our proposed method at the NERP level. From Table 7, we can draw similar conclusions as Table 6.

5. Conclusion

The shallow feature extraction method based on the word2vec word vector is highly dependent on word segmentation technology, which cannot solve the problem of error transfer caused by word segmentation inaccuracy, and the performance of the named entity recognition task deteriorates due to text context features. In this paper, a named entity recognition method based on transfer learning for judicial case text is proposed. This method improves the ability of context bidirectional feature extraction, effectively solves the problem of task boundary division of named entities, and improves the model's ability to recognize ambiguous entities. The experimental results show that the BiLSTM-CRF method based on BERT-WWM-EXT pretraining transfer learning has the best effect. Compared with other models, the entity recognition rate of the BiLSTM-CRF method reaches 96% and 90%, respectively, in NER and NERP entity recognition.

On the contrary, because of the lack of relevant corpus for named entity identification in the judicial field, a set of named entity identification standard specifications related to environmental protection violation warnings is developed, which can meet the business needs in the actual scene, such as the extraction of illegal fact elements. The experimental results show that for different methods, the results based on NER are better than those based on NERP. However, the BERT-WWM-EXT-BiLSTM-CRF method has the best identification result regardless of the type of entity recognition.

Data Availability

The data set used to support the findings of this study can be obtained from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Hongsong Dong contributed to the conception of the study; Yuehui Kong performed the experiment; Wenlian Gao contributed significantly to analysis and manuscript preparation; Hongsong Dong performed the data analyses and wrote the manuscript; Jihua Liu helped perform the analysis with constructive discussions.

Acknowledgments

This work was supported by the Doctoral Natural Science Foundation Project of Lüliang College (Grant no. 2110150544), the Higher Education Institutions of Shanxi Province Teaching Reform and Innovation Project (Grant no. J20221132), the 2020 Science and Technology Targeted Poverty Alleviation Project of Deeply Impoverished Counties in Shanxi Province (Grant no. 2020FP-11), and the Innovation and Entrepreneurship Training Program for College Student (Grant no. 20221241).

References

[1] S. Sharafat, Z. Nasar, and S. W. Jaffry, "Data mining for smart legal systems," *Computers & Electrical Engineering*, vol. 78, pp. 328–342, 2019.

[2] M. Bruckschen, C. Northfleet, D. Silva, P. Bridi, and T. Sander, "Named entity recognition in the legal domain for ontology population," in *Proceedings of the 3rd Workshop on Semantic Processing of Legal Texts. 2010 (SPLeT 2010)*, Malta, May 2010.

[3] M. S. Sun, C. N. Huang, H. Y. Gao, and J. Fang, "Identifying Chinese names in unrestricted texts," *Journal of Chinese Information Processing*, vol. 02, pp. 16–27, 1995.

[4] J. C. W. Lin, Y. Shao, J. Zhang, and U. Yun, "Enhanced sequence labeling based on latent variable conditional random fields," *Neurocomputing*, vol. 403, pp. 431–440, 2020.

[5] D. Chopra, N. Joshi, and I. Mathur, "Named entity recognition in Hindi using hidden Markov model," in *Proceedings of the Second International Conference on Computational Intelligence & Communication Technology*, pp. 581–586, IEEE, Ghaziabad, India, February 2016.

[6] F. A. Yusup, M. A. Bijaksana, and A. F. Huda, "Narrator's name recognition with support vector machine for indexing Indonesian hadith translations," *Procedia Computer Science*, vol. 157, pp. 191–198, 2019.

[7] H. W. Chen, "Research on text information extraction of criminal cases," Master dissertation, Nanjing Normal University, Nanjing, 2011.

[8] E. Leitner, G. Rehm, and J. Moreno-Schneider, "Fine-grained named entity recognition in legal documents," in *Proceedings of the 15th International Conference Semantic Systems*, pp. 272–287, Karlsruhe, Germany, September 2019.

[9] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber, "LSTM: a search space odyssey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2222–2232, 2017.

[10] C. Chen and J. Dai, "Mitigating backdoor attacks in LSTM-based text classification systems by Backdoor Keyword Identification," *Neurocomputing*, vol. 452, pp. 253–262, 2021.

[11] X. Yin, D. Zheng, Z. Lu, and R. Liu, "Neural entity reasoner for global consistency in NER," 2018, https://arxiv.org/abs/1810.00347.

[12] M. R. Hossain, M. M. Hoque, N. Siddique, and I. H. Sarker, "Bengali text document categorization based on very deep convolution neural network," *Expert Systems with Applications*, vol. 184, Article ID 115394, 2021.

[13] Q. Li, P. Li, K. Mao, and E. Y. M. Lo, "Improving convolutional neural network for text classification by recursive data pruning," *Neurocomputing*, vol. 414, pp. 143–152, 2020.

[14] C. Cardellino, M. Teruel, L. Alemany, and S. Villata, "A low-cost, high-coverage legal named entity recognizer, classifier and linker," in *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, pp. 9–18, London, UK, June 2017.

[15] H. Dong, F. Yang, and X. Wang, "Multi-label charge predictions leveraging label co-occurrence in imbal-anced data scenario," *Soft Computing*, vol. 24, no. 23, pp. 17821–17846, 2020.

[16] S. Vashishtha and S. Susan, "Fuzzy rule based unsupervised sentiment analysis from social media posts," *Expert Systems with Applications*, vol. 138, Article ID 112834, 2019.

[17] H. Dong, F. Yang, X. Wang, and Y. Sun, "Automatic extraction of associated fact elements from civil cases based on a deep contextualized embeddings approach: kgcee," *Soft Computing*, vol. 25, no. 17, pp. 11817–11836, 2021.

[18] Y. H. Liu, Y. L. Chen, and W. L. Ho, "Predicting associated statutes for legal problems," *Information Processing & Management*, vol. 51, no. 1, pp. 194–211, 2015.

[19] J. G. Yao, X. J. Wan, and J. G. Xiao, "Recent advances in document summarization," *Knowledge and Information Systems*, vol. 53, no. 2, pp. 297–336, 2017.

[20] M. Mojrian and S. A. Mirroshandel, "A novel extractive multi-document text summarization system using quan-tum-inspired genetic algorithm: mtsqiga," *Expert Systems with Applications*, vol. 171, Article ID 114555, 2021.

[21] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association For Computational Linguistics: Human Language Technologies*, pp. 4171–4186, Minneapolis, MN, USA, June 2019.

[22] Y. Liu, M. Ott, N. Goyal et al., "RoBERTa: a robustly optimized BERT pretraining approach," 2019, https://arxiv.org/abs/1907.11692.

[23] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: a lite BERT for self-supervised learning of language representations," 2019, https://arxiv.org/abs/1909.11942.

[24] L. Yang and M. Sun, "Improved learning of Chinese word embeddings with semantic knowledge," in *Proceedings of the Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. CCL NLP-NABD*, pp. 15–25, Guangzhou, China, November 2015.

[25] X. Jian, J. Liu, L. Zhang, Z. Li, and H. Chen, "Improve Chinese word embeddings by exploiting internal structure," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, CA, USA, June 2016.

[26] X. Yang and W. Huang, "A conditional random fields approach to clinical name entity recognition," in *Proceedings of the Evaluation Tasks at the China Conference on Knowledge Graph and Semantic Computing (CCKS 2018)*, pp. 1–6, Tianjin, China, August 2018.

[27] C. Dong, J. Zhang, C. Zong, M. Hattori, and D. Hui, "Character-based LSTM-CRF with radical-level features for Chinese named entity recognition," in *International Conference on Computer Processing of Oriental Languages National CCF Conference on Natural Language Processing and Chinese Computing*vol. 10102, Cham, Springer International Publishing, 2016.

[28] Z. Hai and C. Kit, "Unsupervised segmentation helps supervised learning of character tagging for word segmentation and named entity recognition," in *Proceedings of the Sixth Sighan Workshop on Chinese Language Processing*, Hyderabad, India, January 2008.

[29] N. Peng and M. Dredze, "Improving named entity recognition for Chinese social media with word segmentation representation learning," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, vol. 2, Berlin, Germany, August 2016.

[30] N. Peng and M. Dredze, "Named entity recognition for Chinese social media with jointly trained embeddings," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, September 2015.

[31] W. Liu, T. Xu, Q. Xu, J. Song, and Y. Zu, "An encoding strategy based word-character LSTM for Chinese NER," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 2379–2389, Minneapolis, Minnesota, June 2019.

[32] W. Cui, W. Che, T. Liu, B. Qin, and Z. Yang, "Pre-training with whole word masking for Chinese BERT," 2019, https://arxiv.org/abs/1906.08101.

[33] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," 2017, https://arxiv.org/abs/1706.03762.

[34] Y. M. Lin, "Research on named entity recognition in judicial field," Master's Dissertation, Kunming, Yunnan University of Finance and Economics, Kunming, China, 2019.

[35] X. Q. Wang, "Research on key technologies of element extraction in legal instruments for smart court," Harbin Institute of Technology, Harbin, Master's Disser-tation, 2020.

[36] P. Zhao, L. Y. Sun, and Y. Wan, "Chinese scenic spot named entity recognition based on BERT+BiLSTM+CRF," *Computer Systems & Applications*, no. 06, pp. 169–174, 2020.