Attention-BLSTM-CRF Based Method for Named Entity Recognition in Judicial Domain

Chen Wang, Bo Li, Wenjing Zhang
School of Data Science and Intelligent Media, Communication University of China, Beijing 100024, China
1346639498@qq.com

Abstract. Although the texts in the judicial field are relatively standardized, the entity categories are rich and the structure is different, and the entity expression is special in some legal documents, electronic files and guiding cases. In order to improve the effect of named entity recognition in the field of justice, this paper entity type can be divided into four categories, and presents a model of named entity recognition based on attention mechanism, structure improvement of input vector fusion CNN embed mode, BLSTM neural network at the same time we constantly study the characteristics of the context, finally by CRF decoding output sequence. The experimental results show that the method is helpful to the application research in the judicial field, and the experiment on the self-built corpus test set has achieved a good entity identification effect.

1. Introduction
In the Internet information era, judicial texts (judicial documents) play an important role in the construction of a law-based society, and normative judicial texts (judicial documents) help citizens to understand the specific details of cases and the process of judicial results. Although the judgment document of the court has the characteristics of standardization and accuracy, its structure is relatively complex for entity identification, which is different from the general entity identification. For example, the name of a person, the case will include the defendant, the plaintiff, the judge, and the criminal suspect, etc., while the location may be divided into the place of the case, the place of reporting, and the place of handling a case. Time and literature are also very sensitive and important words in the judgment documents. Faced with the lack of corpus for judicial text annotation at present, we can only use manual annotation to obtain small-scale sample sets for machine learning exercise, which may affect the effect of entity identification to some extent. Obviously, the particularity of judicial text brings some difficulties to the identification of named entities.

In recent years, the research on named entity recognition has achieved good results and has been widely used in many fields, especially in some proprietary fields such as medicine, military and network media. In order to realize the correct word segmentation of traditional Chinese medicine nouns and prescription nouns in medicine, the BLSTM neural network model is used to identify named entities, which reduces ambiguity and achieves good word segmentation effect[1]; Aiming at the grammatical features of military text, a conditional random field model (CRF) is established and the method based on dictionaries and rules is used to identify military named entities[2]; Microblog, as an important platform for information exchange, has a unique content structure[3]. The method based on the combination of rules and statistics is used to identify named entities.
Compared with the application of named entity identification in medical and military fields, the combination of judicial fields has not been fully promoted. Text in the field of justice, while the relative specification, entity class rich and different structure, and in some legal documents, electronic files, and to guide the particularity of physical expression in the case, so whether add the various characteristics of CRF model, or named entity recognition method based on rules or statistics, can make effective to entities in the field of judicial identification[4].

In order to improve the judicial field named entity, especially in some entity recognition effect in criminal investigation cases, based on the characteristics of judicial field text, this paper USES the traditional entity type: no longer person names, place names, organization names, time, etc., but to set the entity justice text category for legal subject, object, legal facts and legal documents these four classes. Generally speaking, the subject of law includes natural persons and organizations, while the object of law includes things and natural achievements, etc. Legal facts refer to some legal events and legal acts. In this paper, after solving the problem of long distance dependence which cannot be solved by traditional recognition methods by using the Bidirectional Long and Short Term Memory (BLSTM)-Conditional Random Field (CRF), attention mechanism is introduced to obtain the local features of text. Experimental results show that this method can effectively improve the identification effect of named entities in judicial field.

2. Named Entity Identification Framework in the Judicial Field

This paper abandons the traditional entity classification, and establishes four specific categories for judicial texts: Legal Subject (SUB), Legal Object (OBJ), Legal Fact (FAC), and Legal Document (DOC). This paper comprehensively considers the interconnectedness of concepts in judicial texts, and sometimes there are reciprocal semantic relations among these four types of entities. Therefore, the identification of named entities in judicial field is attributed to the task of sequential annotation for research. The model framework is shown in Fig.1:

![Attention-BLSTM-CRF model](image)

Fig.1 Attention-BLSTM-CRF model

2.1. Input Vector Representation

In the process of named entity recognition based on traditional methods, the advantages of word features and part of speech features to better reflect the expected set information [5] will be utilized to help complete the task of named entity recognition., but past experience has shown that this method has certain disadvantages, which cannot be deep dig the information processing and the relationship between the word and the word has strict requirements. There are two ways for vectorization of Chinese characters in judicial texts: Singly Heat Vector Representation and Distributed Vector Representation[6]. The research shows that the distributed vector representation of these word features and part of speech features can solve the "semantic gap". This representation method is to express some large and extremely sparse vectors with low-dimensional dense real value vectors, in which each dimension represents a hidden feature, and these distributed representations can automatically learn
important information from the input text, so the distributed vector representation can improve the accuracy of the experimental results.

In this paper, attention mechanism is introduced to better obtain the local features of the text, and in order to extract the internal features of the text, we adopt the method of character vector and word vector splicing as the input vector embedding. As shown in Fig.2:

![Fig.2 Word combination vector model](image1)

As can be seen from Fig.2, a CNN model is added for the word vector. The character vector of each word is converted into matrix form, and considering that the length of each word is different and the size of the matrix is different, placeholders are introduced at both ends of the word to supplement to ensure that the size of each character-level matrix is the same[7]. Then it enters the convolutional neural network for training through the back-propagation algorithm. After feature extraction through the convolutional layer, it enters the pooling layer to aggregate the included elements. The maximum pooling layer can realize the extraction of descending and multi-feature, and finally get the character-level feature vector.

![Fig.3 Character-level CNN structure](image2)

2.2. **BLSTM Model**

LSTM, as a kind of time loop neural network[8], is similar to the traditional RNN structure, which can process the current information while transmitting the information forward. However, the difference is that the model structure of LSTM has more cell state and gate structure, which enables LSTM to remember the long-term information[9].

![Fig.4 LSTM Structure](image3)
The gating structure of LSTM mainly includes three mechanisms: forgetting gate, input gate and output gate[10]. Forgetting gate, as the name implies, is the function of "forgetting". It is used to decide what kind of old information is discarded from the cell state. The input gate is used to determine what kind of new information should be placed in the cell state. The output gate determines what kind of information the next cell state will contain. The information transmission algorithm of LSTM is[11]:

\[
\begin{align*}
    f_t &= \text{sigmoid}(W_f h_{t-1} + W_f x_t + b_f) \\
    i_t &= \text{sigmoid}(W_i h_{t-1} + W_i x_t + b_i) \\
    \tilde{C}_t &= \tanh(W_c h_{t-1} + W_c x_t + b_c) \\
    C_t &= C_{t-1} \ast f_t + i_t \ast \tilde{C}_t \\
    o_t &= \text{sigmoid}(W_o h_{t-1} + W_o x_t + b_o) \\
    h_t &= o_t \ast \tanh(C_t) \\
    y_t &= \text{sigmoid}(V h_t + e)
\end{align*}
\]

(1) (2) (3) (4) (5) (6) (7)

Compared with the traditional RNN structure, LSTM solves the problem of gradient vanishing and long-term dependence[12], but it can not use the information of context. In order to effectively use the contextual information, this paper adopts the BLSTM structure, which can not only read the sentence forward to obtain the information above, but also read the sentence backward to obtain the information below. Let \( \{x_1, x_2, \cdots, x_n\} \) be an input sequence, forward recursion and backward recursion are respectively adopted to obtain forward LSTM and backward LSTM: \( \{h_{R1}, h_{R2}, \cdots, h_{Rn}\} \) and \( \{h_{L1}, h_{L2}, \cdots, h_{Ln}\} \).

The hidden vectors obtained in the forward and backward directions are spliced to obtain \( \{[h_{R1}, h_{L1}], [h_{R2}, h_{L2}], \cdots, [h_{Rn}, h_{Ln}]\} \), namely \( (h_1, h_2, \cdots, h_n) \) and \( h_t = [h_{Rt}, h_{Lt}] \). Finally, the eigenvectors of BLSTM are output[13].

![BLSTM model structure](image)

2.3. Attention Model
The mechanism of attention comes from the study of human vision. During the cognitive process, people will choose some specific information in the visual field to focus on and read, while they will discard other visible information[14]. In the neural network model, each output data has different importance to us. By adopting the attention mechanism, the weight of the important data needed in the whole model is increased and the weight of the noise data is reduced.

In this paper, by introducing the attention mechanism and combining with the BLSTM model, each feature vector of the output splicing of the BLSTM model is given appropriate weight, and some local features are noticed by the attention mechanism while the global features are extracted, and finally the joint features are obtained[15]. The neural network model of attention mechanism introduced in the framework of named entity recognition is shown in Fig. 6:
Fig. 6 Attention machine

$c_t = \sum_{j=1}^{t} a_j h_j$ is the final output state of the attention mechanism. 

$a_{ij}$ assigns the weight of each eigenvector to the attention mechanism:

$$a_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{K} \exp(e_{ik})}$$ (8)

$e_{ij}$ is the score of the attention mechanism:

$$e_{ij} = a(C_{i-1}, h_j) = V \tanh(W C_{i-1} + U h_j)$$ (9)

$C_{i-1}$ is the output state of the attention mechanism at the last moment; 
$tanh$ is hyperbolic tangent activation function; 
$h_j$ is the eigenvector of BLSTM output; 
$V$ is the global weight transformation matrix; 
$W$ is the weight transformation matrix of the output state of the attention mechanism at the previous time; 
$U$ is the weight transformation matrix of the eigenvectors.

2.4. CRF Layer

After the BLSTM model and the Attention Layer, although the output sequence of joint features of global and local features has been obtained, the rationality between annotations cannot be taken into account in the tagging task of named entities[16]. In order to avoid illegal annotation situations such as B-SUB and B-OBJ after b-sub, CRF Layer was added at the end of the model to ensure that the predicted sequence output was the correct annotation sequence with the maximum probability[17].

Let $\{x_1, x_2, \cdots, x_n\}$ be an input sequence, and $\{y_1, y_2, \cdots, y_n\}$ as the annotation sequence.

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

$score(y)$ represents the final score of the annotation sequence; $A$ is the state transition matrix; $A_{i,j}$ represents the transition matrix from the $i_{th}$ tag to the $j_{th}$ tag; $P$ is the output state matrix of Attention layer; $P_{y_i}$ is the probability that the $i_{th}$ word is labeled $y_j$, and the probability value of the annotation sequence is $P(y)$:

$$p(y) = \frac{\exp(score(y))}{\sum_{y\in Y} \exp(score(y))}$$ (11)

And $Y=$\{All possible label sequences\}. The objective function of the model is denoted as $L$:

$$L = score(y) + \log_2 \sum_{y \in Y} \exp(score(y))$$ (12)

According to the function of CRF layer, the maximum probability of the model's output prediction sequence being correctly labeled is:

$$tag = \arg \max_{y \in Y} p(y)$$ (13)

3. Experimental Results and Analysis

3.1. Data Set and Evaluation Index
Due to the particularity of judicial texts, there is no official judicial corpus. This experiment from China's official website to download the written judgment of Beijing in 2018 the civil verdict and administrative orders, orders for compensation, execute orders each 250, a total of more than 1.201 million words, taking the artificial labeling and coding quadratic mark[4], the corpus of the experiment, and according to the proportion of 8:1:1 corpus can be divided into training, validation, corpus, test corpus. The labeling of corpus is shown in Table 1:

| Name     | Numbers |
|----------|---------|
| Texts(Part) | 1000    |
| Sentences | 22000   |
| Words    | 712000  |
| Characters | 1201000 |

In this paper, the task of named entity identification is divided into four categories: legal subject, legal object, legal fact and legal literature. In order to better evaluate the performance of the system, Precision (P), Recall (R) and F-score are used as the evaluation indexes of the named entity identification task.

3.2. Element Label
The annotation mode adopted in this experiment is BIO annotation mode. B marks the beginning part of the entity, I marks the middle part of the entity, and O represents the non-entity class. For example, the annotation of the legal subject is denoted as B-SUB/O-SUB.

3.3. Experimental Results and Analysis
In order to verify the best entity recognition effect of the model adopted in this experiment, CRF and BLSTM-CRF models were used to conduct experiments on the corpus obtained by manual annotation and coded secondary annotation in this paper. Experimental comparison results are shown in Table 2:

| Model               | P(%) | R(%) | F (%) |
|---------------------|------|------|-------|
| CRF                 | 80.31| 78.52| 79.40 |
| BLSTM-CRF           | 85.77| 86.21| 85.99 |
| Attention-BLSTM-CRF | 88.63| 87.95| 88.29 |

In a comprehensive analysis of the effect of the model, we compared the data results in table 2 to the judicial corpus of this paper for analysis, which also verified the better effect of the Attention-BLSTM-CRF model adopted in the experiment. According to the comparison results between CRF model and BLSTM-CRF model, BLSTM-CRF model has better effect, mainly because BLSTM is added to the model to fully obtain the characteristics of the context and consider the long-term dependence between features. From the results of the BLSTM-CRF model and the Attention-BLSTM-CRF model, the recognition effect of the Attention-BLSTM-CRF model is better, because the named entity recognition model adopted in this paper adds Attention mechanism. It can enrich the local features of the text according to the particularity of the judicial text, realize the acquisition of joint features, and integrate the CNN structure model into the processing of the input vector to better train the character vector, so as to achieve a better recognition effect of the model as a whole.

4. Conclusion
In this paper, by studying the particularity of judicial text and judicial named entity, a named entity recognition model based on self-attention mechanism is proposed. The internal feature of character vector is obtained by CNN structure, and the ability of BLSTM model to learn context information is better used. At the same time, CRF model is used to ensure the rationality of annotation task. Although the recognition effect of this model is better than the other two models, the accuracy and recall rate are not very ideal. This may be because our entity labeling in the judicial field is not mature enough. In
the future work, we will increase the text processing, and consider adding the BERT pre training language model to improve the effect of named entity recognition.

Acknowledgments
I would like to thank the teacher for his guidance in writing and the students for their help in processing the data. Meanwhile, thanks for the project support of "Research on Investment Risk Assessment of IP Adapted Film Project" (CUC190T02) and "Research on Judicial Concept Relation Extraction Algorithm Based on Big Data" (HG19012).

References
[1] Gong Deshan, Liang Wenyu, Zhang Bingzhu, Ma Xingang. Application of Named Entity Recognition in the Recognition of Chinese Medicine Nouns and Prescriptions[J]. China Pharmaceutical Affairs, 209,33 (06):710-716(in Chinese).
[2] Feng Yuntian, Zhang Hongjun, Hao Wenning. Named Entity Recognition for Military Texts [J]. Computer Science, 2015,42(07):15-18 + 47(in Chinese).
[3] Li Gang, Huang Yongfeng. Named Entity Recognition Method for Microblog[J]. Application of Electronic Technique, 2008,44(13)(in Chinese).
[4] Lin Yimeng. Research on Named Entity Identification in the Judicial Field [D]. Yunnan University of Finance and Economics, 2019(in Chinese).
[5] Soutnner D, Miller L. Continuous distributed representations of words as input of LSTM network language model[C]. Proc of International Conference on Text, Speech, and Dialogue,2014:150-157.
[6] Xiao Wen. Application of Distributed Representation and Combination Model in Chinese Natural Language Processing [D]. Nanjing: Southeast University, 2016(in Chinese).
[7] Li Jianlong, Wang Chuanqing, Han Qiuyu. Military Named Entity Identification Based on Bidirectional LSTM [J]. Computer Engineering and Science, 204,41(04): 713-718(in Chinese).
[8] HOCHREITER S, SCHMIDHUBER J. Long Short -Term Memory[J]. Neural Computation, 1997, 9(8):1735- -1780.
[9] GREGG K, SRIVASTAVA R K, KOUTNIK J, et al. LSTM: a search space odyssey[J]. IEEE Transactions on Neural Networks & Learning Systems(S2162-237X), 2015, 28(10): 2222-2232.
[10] Chiu J P C, Nichols E. Named Entity Recognition with Bidirectional LSTM- CNNs [J]. Computer Science, arXiv: 1511. 08308:2015.
[11] S. Hochreiter and J. Schmidhuber. Long short-term memory[J]. Neural Computation,9(8):1735–1780,1997.
[12] GRAVES A, SCHMIDHUDE J. Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures[J].Neural Networks,2005,18(5/6):602-610.
[13] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv:1508.01991,2015.
[14] MIRSAMADI S, BARSOUOM E, ZHANG C. Automatic Speech Emotion Recognition Using Recurrent Neural Networks with Local Attention[C]. IEEE International Conference on Acoustics, 2017: 2227-2231.
[15] QAMAS G K S , YIN Jize, PAN Limin, et al. Research on Named Entity Recognition Method Based on Deep Neural Network[J]. Netinfo Security,2017(10):29-35.
[16] Mulwad V, Li W, Joshi A, et al. Extracting information about security vulnerabilities from web text[C]. IEEE Computer Society, 2011.
[17] Lafferty J D, McCallum A, Pereira F C N. Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data[C]. Eighteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc. 2001 :282-289.