EURASIP Journal on Image and Video Processing

RESEARCH Open Access

Named entity recognition for Chinese judgment documents based on BiLSTM and CRF

Wenming Huang[1], Dengrui Hu[1], Zhenrong Deng[2*] and Jianyun Nie[3]

Abstract

Chinese named entity recognition (CNER) in the judicial domain is an important and fundamental task in the analysis of judgment documents. However, only a few researches have been devoted to this task so far. For Chinese named entity recognition in judgment documents, we propose the use a bidirectional long-short-term memory (BiLSTM) model, which uses character vectors and sentence vectors trained by distributed memory model of paragraph vectors (PV-DM). The output of BiLSTM is used by conditional random field (CRF) to tag the input sequence. We also improved the Viterbi algorithm to increase the efficiency of the model by cutting the path with the lowest score. At last, a novel dataset with manual annotations is constructed. The experimental results on our corpus show that the proposed method is effective not only in reducing the computational time, but also in improving the effectiveness of named entity recognition in the judicial domain.

Keywords: Named entity recognition, Judgment documents, Neural network

1 Introduction

Named entity recognition (NER), aiming to extract the words or expressions denoting specific entities from documents, is a core research topic in the fields of nature language processing (NLP) and multimedia security. It has been extensively investigated in recent years [1] and applied in various scenarios, such as information extraction, dialog system, sentence parsing, machine translation, and metadata annotation.

The general named entities studied by academic community are divided into three categories: entity, time, and number. These categories are further divided into seven sub-categories: person name, organization name, place, time, data, currency, and percentage, respectively. In different domains, we can also define named entities unique to the domain.

Named entity in text contains rich semantics and is an important semantic unit. Recognizing these named entities from the original plays an important role in natural language understanding. Recently, named entity recognition has achieved quite satisfactory results to some extent, so some scholars regard it as a solved problem. But in practical applications, there are still many problems to be solved.

At present, named entity recognition has achieved good results in some limited fields and corpus, such as news. But these methods cannot be effectively transferred to other fields, such as biology, medical, military, and judicial fields. On the one hand, because the text in different fields contains its unique named entities, for example, the judicial documents will contain the penalty, laws, and regulations and other unique entities, while the methods applied in the field of news are not universal in this field. On the other hand, due to the lack of corresponding annotation datasets in these fields, it is difficult to use large-scale datasets for model training.

*Correspondence: zhrdeng@guet.edu.cn
[2]Guangxi Key Laboratory of Intelligent Processing of Computer Image and Graphics, Guilin University of Electronic Technology, Guilin, China
Full list of author information is available at the end of the article

In addition to the above problems, the named entity recognition of different languages also has great differences. Zhu et al. [2] enhance the representation by increasing the entity-context diversity without relying on external resources and present a flexible NER framework compatible with different languages and domains. They conduct experiments on five languages, such as English, German, Spanish, Dutch, and Chinese, and biomedical fields, such as identifying the chemicals and gene/protein terms from scientific works.

In English and some other European language, there are spaces between words as division marks, and the initial of proper nouns will be capitalized. But in Chinese, there is no obvious word boundary. In Chinese NER (CNER), most research is based on available public datasets [3]. Chen et al. [4] used conditional random fields and maximum entropy model for CNER without segmentation and obtained 86.2% in F1 on Microsoft Research Asia (MSRA) dataset and 88.53% in F1 on City University of Hong Kong (CITYU) dataset. Named entities contain specific nouns with specific meanings in different fields. Zhang et al. [5] used a Lattice LSTM for CNER. Compared with the character-based method, the Lattice model uses words and word order features without segmentation error, and this produced the best results on datasets in several different fields. However, in judicial field, NER is a difficult task because of the following characteristics of judgment documents.

1. The development of artificial intelligence technology in judicial field is relatively slow due to the lack of annotated dataset. In this paper, we focus on judgment documents, which have several domain specificities. According to the classification of Supreme People's Court of the PRC, the types of documents include judgments, rulings, mediations, decisions, notices, and orders, each of which has different characteristics.
2. Judgment documents are professional descriptions of the process of a case, relevant evidence, applicable laws and regulations, and the result of a judgment. They contain a large number of special expressions such as professional terms in judgment documents. The generic types of entities designed for general NER are no longer sufficient in this field.
3. There are many nested names of organizations, laws, and regulations in judicial documents. For example, "广西壮族自治区桂林市检察院" (Procuratorate of Guilin city of Guangxi Zhuang Autonomous Region) is an organization name in which a place name "广西壮族自治区桂林市" (Guilin city of Guangxi Zhuang Autonomous Region) is nested; Similarly, there are an organization name "最高院" (Supreme court) and a law "《中华人民共和国刑事诉讼法》" (Criminal procedure law of the People's Republic of China) in a regulatory name "《最高院关于适用<中华人民共和国刑事诉讼法>的解释》" (Explanation of the Supreme court about the Criminal procedure law of the People's Republic of China). The names of laws and regulations are usually very long, and the length of the accusation and the names of laws and regulations are often uncertain, which makes it difficult to find rules.

With the development of cloud services, it provides a lot of convenience for our work. Le et al. [6] design an annotation system based on Web Ontology Language (OWL) to enrich the semantic expressivity of the model. And they also propose a Cloud Service Selection with Criteria Interaction (CSSCI) framework and a priority-based CSSCI (PCSSCI) to solve service selection problems in the case where there is a lack of historical information to determine criteria relations and weights [7].

To improve the precision of NER, we propose a method based on a character level bi-directional long-short-term memory network (BiLSTM). In the proposed method, we use character-level BiLSTM to avoid word segmentation error. To address the issue that explicit word information is not fully exploited, we fuse character vector with sentence vector which is train by distributed memory model of paragraph vectors (PV-DM). Then, we further use conditional random field (CRF) layer to tag the input. In our implementation, we also try to improve the efficiency of Viterbi algorithm by pruning some useless path, leading to reduced average prediction time. We construct an annotated dataset of judgment documents manually and tested our method on it. Data enhancement is used to avoid oneness of entity to improve the generalization ability of the model.

The remainder of the paper is organized as follows. In Section 2, we introduce the related work on named entities in different languages and fields, Section 3 presents the methods we proposed in detail, Section 4 discusses the experimental setup and results, and Section 5 concludes the work finally.

2 Related works

NER usually includes two parts: determining entity boundaries and identifying entity types. The recognition of named entities in English and other European languages is much helped by the fact that some entities follow obvious patterns. For example, personal names usually start with capital letters. Compared to Chinese, words in English are also naturally delimited.

The main methods of NER include those that are based on rules and dictionaries, statistics, combination of rules and statistics, and neural networks. The task has traditionally been solved as a sequence labeling problem. Ma et al. [8] used an LSTM-CNN-CRF which implements end-to-end NER for part-of-speech (POS) tagging task without feature engineering and data processing. They achieved 97.55% precision and 91.21% in F1 score. Lample et al. [9] achieved the state-of-the-art for English NER by integrating character information into word representation.

The early mainstream methods of named entity recognition are rule-based. A representative work is DL-Co training method proposed by Wang et al. [10]. In this study, rules are automatically discovered and generated by machine learning. First, a set of seed rules is defined, and then, more rules are obtained through iterations of unsupervised learning on the corpus. Finally, the rules are applied to entity recognition of person, place, and organization names. Their research shows that unlabeled data can help reduce the supervision requirements to only seven "seed" rules. This method effectively utilizes the natural redundancy in the data. For many named entities, the spelling of the name and the context in which the name appears are sufficient to determine the type of the entity. Mikheev et al. [11] found that combination of rules and statistic model can identify place names without using a named entity dictionary.

Despite the good result obtained, rule-based methods are inherently limited in coverage: In practice, it is impossible to define rules to cover all the cases. In addition, it is often difficult to transfer the rules designed for a domain to another domain.

Support vector machine (SVM) is widely used in text and image classification. The support vector machine recursive feature elimination (SVMRFE) method is used to improve the classification accuracy [12]. And they propose to extract forensic features to improve the classification accuracy [13]. Recently, considering the prior knowledge extracted from training samples, Lan et al. proposed an excellent representation-based classifier called PKPCRC, and the comparative results show that it achieves state-of-the-art performance [14].

More recent work on NER is usually based on machine learning methods, in particular, deep learning methods [15, 16]. It is also extended to many other languages than European languages. Hammerton [17] successfully applied LSTM for NER for the first time. Mo et al. [18] constructed a dataset for Burmese NER by dividing Burmese text into syllables and conducted experiments using various models and methods. Among them, the BiLSTM-CRF model has achieved the best performance. Anh et al. [19] proposed a method for Vietnamese Part-of-Speech Tagging (POS Tagging) and NER. Experiments were conducted using bi-word information sources, character-based word representation, and pre-trained word vectors. The precision rate was 93.52% in POS tagging tasks and 94.88% in NER tasks, reaching the state-of-the-art performance.

Compared with English, the biggest difference and difficulty of Chinese named entity recognition lies in the fact that a Chinese text does not have explicit word boundaries like in English texts. Therefore, the first step of named entity recognition is to determine word boundaries, namely word segmentation [20, 21]. Chinese named entity recognition is thus impacted by the effectiveness of Chinese word segmentation.

On Chinese word segmentation, Qiu's [22] research shows that Chinese word segmentation has high accuracy on news datasets, but the segmentation accuracy in other domains is much lower. In an attempt to develop an approach applicable to different domains, they designed a novel dual propagation algorithm, which combines named entities with common context patterns and serves as a plug-in for training model word segmentation in the source domain.

Xu et al. [23] propose a simple effective neural framework to derive the character-level embeddings for NER in Chinese text, named ME-CNER. A character embedding is derived with rich semantic information harnessed at multiple granularities, ranging from radical, character to word levels. Also a convolutional-gated recurrent unit (Conv-GRU) network is designed to capture semantic representation for a character based on its local context and long-range dependence.

Zhou et al. [24] used multiple HMMs in series for NER. Firstly, an N-gram model is used to model sentence segments. The output of a lower-layer HMM is used as input of a higher-layer HMM. A web search engine is then used to collect data to calculate the degree of association between named entities so as to identify and disaggregate synonym entities.

Wang et al. [25] proposed segment-level Chinese NER using GCNN-LSTM model with beam search algorithm. The problem is solved based on beam-search algorithm, and some low-quality information is selectively discarded by gate mechanism. For a given Chinese character input sequence, the segment information is obtained while it is segmented, and the segment is labeled by the encoder. Through analysis, the overall score of the input sequence is obtained, and the segmentation marker sequence with the highest score is selected as the final prediction result.

Devlin et al. [26] proposed Bidirectional Encoder Representations from Transformers (BERT) model and achieved improvements in many tasks including NER. Li et al. [27] pre-train BERT model on the unlabeled Chinese clinical records, which can leverage the unlabeled domain-specific knowledge. Different layers such as LSTM and CRF are used to extract the text features and decode the predicted tags, respectively. Radical features of Chinese characters are used to improve the model performance as well.

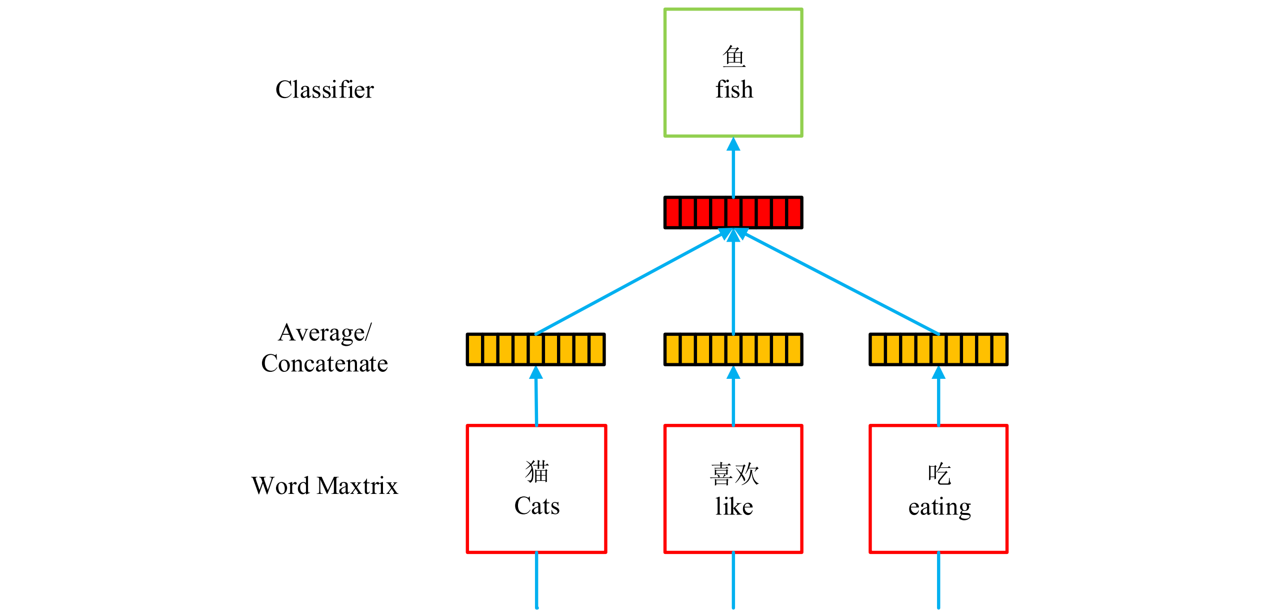

Fig. 1 PV-DM model structure. Predict that the next word is "fish" through the given words "cat likes eating"

The main purpose of Chen et al. [28] was to automatically identify Adverse Drug Reaction (ADR)-related entities from the narrative descriptions of Chinese ADE Reports (ADERs) so as to serve as supplements when evaluating the structured section of cases, which can further assist in ADR evaluation. In this paper, they employed two highly successful NER models of CRF and BiLSTM-CRF, as well as one generated Lexical Feature-based BiLSTM-CRF (LF-BiLSTM-CRF) model to conduct NER tasks respectively in the Chinese ADEPDs in this paper. Large

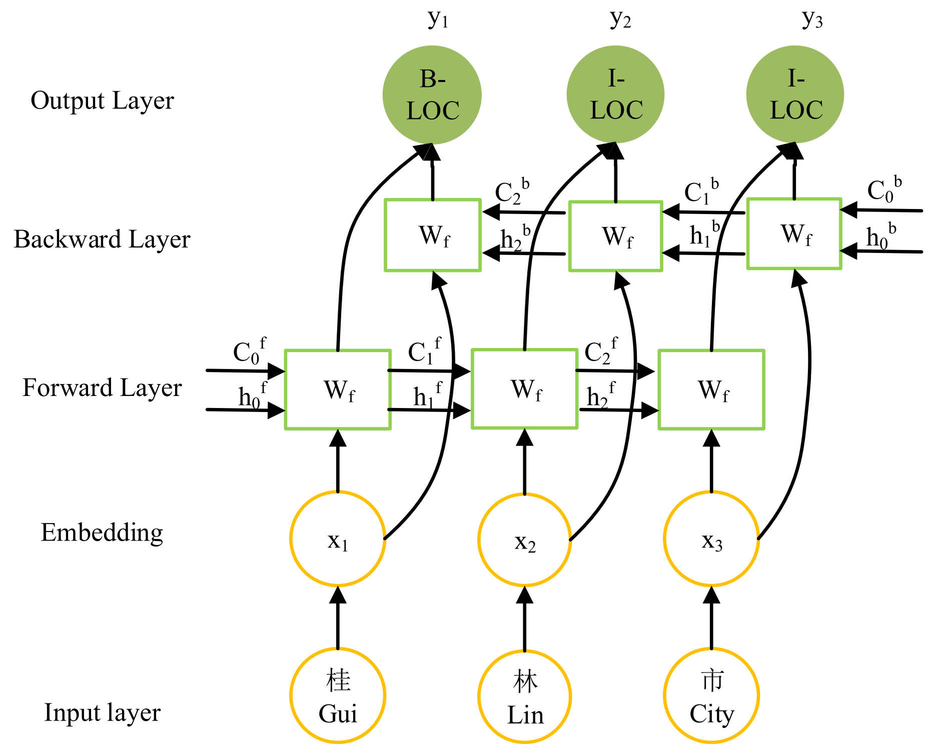

Fig. 2 Character-level BiLSTM structure. The input character sequence is vectorized and the output is the corresponding tag sequence

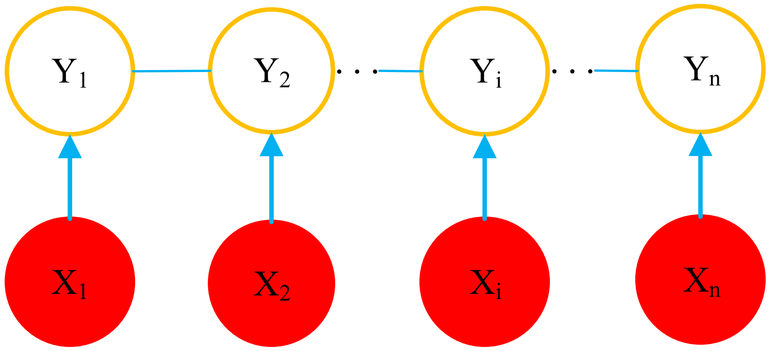

Fig. 3 Linear chain conditional random fields. Input sequence is observation sequence, while the output sequence is a label sequence. It is a conditional probability model

amounts of data were manually annotated for model training. To take full advantage of the un-annotated raw data, they also explored a semi-supervised iterative training strategy of tri-training on the basis of three established models to crosswise give un-annotated cases tags with high confidence and subsequently add the newly tagged cases into the training sets to retrain the basic models.

Because of the differences between domains, the methods mentioned above often cannot perform well in domains other than news articles. Zhu et al. [29] have studied the application of the Convolutional Attention Network (CAN) to Chinese named entity recognition. The model consists of a character-level Convolutional Neural Network (CNN) with local-attention mechanism and a Gated Recurrent Unit (GRU) of global self-attention mechanism captures information from context and adjacent characters. Experiments show that this method achieves good results when character vectors and external dictionaries have some difficulties for NER in different fields.

3 Methods

In this section, we elaborate how we fuse character vector and sentence vector. Next, we also introduce the character-level Bi-LSTM model and conditional random fields (CRF) with improved Viterbi algorithm. In addition, we present how to improve Viterbi algorithm used in CRF.

3.1 Fusion of character and sentence vector

In text vectorization, bag$-of-$words (BOW) [30, 31] is a commonly used method, but this method will lose part of the word order information and semantic information. In addition, the establishment and maintenance of vocabularies are worth considering. Too many vocabularies will lead to significant sparseness of document representation vectors. In this paper, we build character vectors and fuse them with sentence vector before feeding them to BiLSTM for training. The Chinese character vectors are independent from the specific vocabulary (words) and thus provide a robust representation of texts in different domains.

We use distributed memory model of paragraph vectors (PV-DM) [32] for sentence embedding. The method of training sentence vectors is similar to the method of training word vectors, the core idea of which is to predict the context of each word for prediction.

In the PV-DM model, each sentence is mapped to an independent vector, which is a column of the matrix; at the same time, each word is mapped to an independent vector, which is a column of the matrix. It is on average or end-to-end for the sentence vector and these word vectors to predict the next word in the text. This model can learn fixed length vector representation from variable length text segments, such as a single sentence, a paragraph, or a document. The structure of PV-DM model is shown in Fig. 1.

Table 1 The process of solving the optimal path by Viterbi algorithm

Initialization:	
	$\delta_1(j) = w \cdot F_1(y_0 = start, y_1 = j, x)$, $j = 1, 2, ..., m$
Recurrence for l = 2,3,..., n:	
	$\delta_i(l) = \max\limits_{1 \le j \le m} \{\delta_{i-1}(j) + w \cdot F_i(y_{i-1} = j, y_i = l, x)\}, \quad l = 1, 2, ..., m$
	$\varphi_i(l) = \arg\max\limits_{1 \le j \le m} \{\delta_{i-1}(j) + w \cdot F_i(y_{i-1} = j, y_i = l, x)\}, \quad l = 1, 2, ..., m$
Termination when i = n:	
	$\max\limits_{y}(w \cdot F(y, x)) = \max\limits_{i \le j \le m} \delta_n(j)$
	$y_n^* = \arg\max\limits_{i \le j \le m} \delta_n(j)$
Return path:	
	$y_1^* = \varphi_{i+1}\left(y_{i+1}^*\right), \quad i = n-1, n-2, ..., 1$

Given a training sequence composed of T words (w_1, w_2, ..., w_T), the goal is to maximize the average logarithmic probability:

$$\frac{1}{T}\sum_{t=k}^{T-K}\log P\left(w_t|w_{t-k},...,w_{t+k}\right) \tag{1}$$

Then, a multi-classifier is used for prediction:

$$P\left(w_t|w_{t-k},...,w_{t+k}\right)=\frac{e^{y_{w_t}}}{\sum_i e^{y_i}} \tag{2}$$

Each y_i represents the no-normalized log-probability of each output word i:

$$y=b+Uh\left(w_{t-k},...,w_{t+k};W\right) \tag{3}$$

where U and b are the softmax parameters. h is constructed by a concatenation or average of word vectors extracted from W. After the sentence vector is obtained by PV-DM, the sentence vector and the character vector are fused by summation, and we use tanh to prevent overflow and underflow.

$$Vec=\tanh\left(V_{sentence}+V_{character}\right) \tag{4}$$

3.2 Bi-directional LSTM based on characters

Long-short-term memory (LSTM) [33] is a special kind of Recurrent Neural Network (RNN), which can solve the problems of long-term dependence and vanishing gradient in ordinary RNN. In RNN, we think that the loop connection is very simple and lacks non-linear activation function. It can only use information from the past. With the continuous transmission of information, when the time interval is long enough, the original information will be forgotten. We call it long-term dependence. For example, “武汉市长江大桥” in Chinese can be understood as “the name of mayor of Wuhan is Jiang daqiao”, but it can be also understood as “the Yangtze River Bridge of Wuhan”. We need to find out its meaning according to its context. So, we choose BiLSTM. LSTM can spread information in sentences. BiLSTM is composed of forward and backward LSTM layers, so that it can make full use of context information to solve the problem of long-term dependence. Moreover, it is suitable for processing and predicting events with relatively long intervals in time-

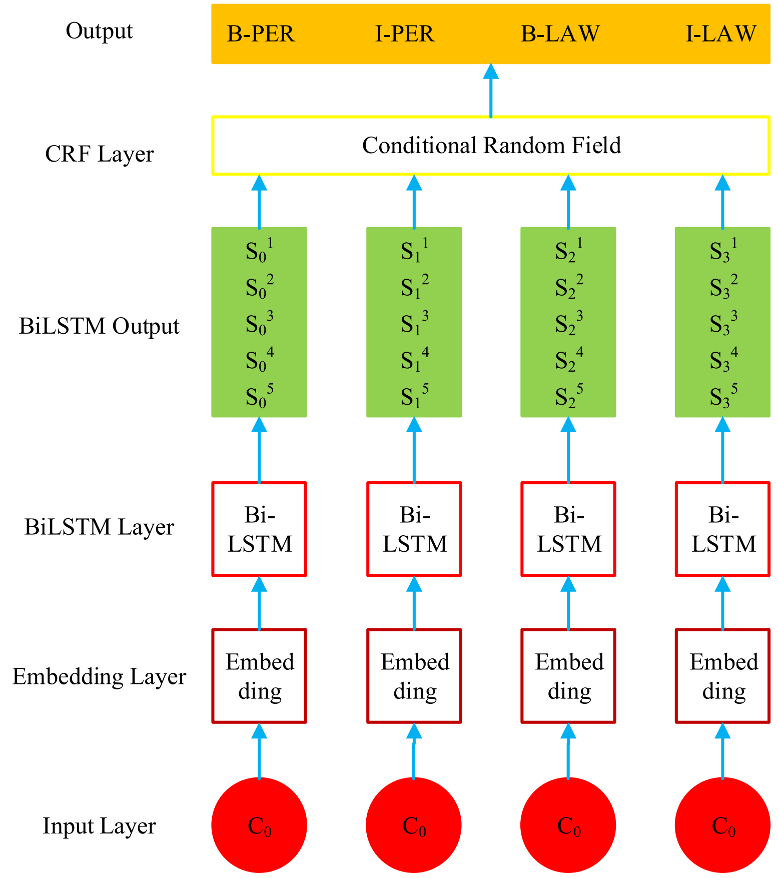
Output
B-PER
I-PER
B-LAW
I-LAW
CRF Layer
Conditional Random Field
BiLSTM Output
S_0^1 S_0^2 S_0^3 S_0^4 S_0^5
S_1^1 S_1^2 S_1^3 S_1^4 S_1^5
S_2^1 S_2^2 S_2^3 S_2^4 S_2^5
S_3^1 S_3^2 S_3^3 S_3^4 S_3^5
BiLSTM Layer
Bi-LSTM
Embedding Layer
Embedding
Input Layer
C_0

Fig. 4 Structure of BiLSTM-CRF. Take the output of BiLSTM as the input of CRF layer

series, and the prediction effect is more accurate than that of LSTM. In this paper, we use BiLSTM as the base model.

The BiLSTM architecture is shown in Fig. 2, where x is the input sequence, y is the output prediction sequence, i.e., tag sequence, c is the cell state, and h is the hidden state. In named entity recognition, BiLSTM can be further combined with conditional random field to enhance its ability to take into account the sequential information during the NER processs.

3.3 Conditional random fields with improved Viterbi

Named entity recognition is usually treated as a sequence labeling task. Conditional Random Field (CRF) is a typical machine learning model for this task. CRF is a conditional probability distribution model, which is characterized by assuming that the output random variables constitute Markov Random Fields. As shown in Fig. 3, in the conditional probability model, X and Y are both random variables, $P(Y \mid X)$ represents the probability distribution of Y given X, which represents the observation sequence to be labeled, and y is the output variable, which denotes the labeling sequence or the state sequence.

When X is x, the conditional probability that the value of Y is y is as follows:

$$P(y|x) = \frac{1}{z(x)} \exp\left(\sum_{i,k} \lambda_k t_k\left(y_{i-1}, y_i, x, i\right) + \sum_{i,k} \mu_l s_l\left(y_i, x, i\right)\right) \tag{5}$$

$$Z(x) = \sum_y \exp\left(\sum_{i,k} \lambda_k t_k\left(y_{i-1}, y_i, x, i\right) + \sum_{i,l} \mu_l s_l\left(y_{i-1} y_i, x, i\right)\right) \tag{6}$$

where t_k is the feature function defined on the edge, called transfer feature, which depends on the current and previous positions, and s_l is the feature function defined on the node, called state feature, which depends on the current position. Generally, the values of the two characteristic functions are 0 or 1. When the characteristic condition is satisfied, the value is 1; otherwise, it is 0. λ_k and μ_l are the corresponding weights. $Z(x)$ is the normalization factor, and the summation is performed on all possible output sequences.

The sequence labeling problem is the prediction problem of CRF. Given the conditional random field $P(Y \mid X)$ and the input sequence (observation sequence) x, the output sequence (labeling sequence) y^* with the greatest conditional probability is found, and the observation sequence is labeled. Therefore, the prediction problem of CRF becomes the optimal path problem with the largest

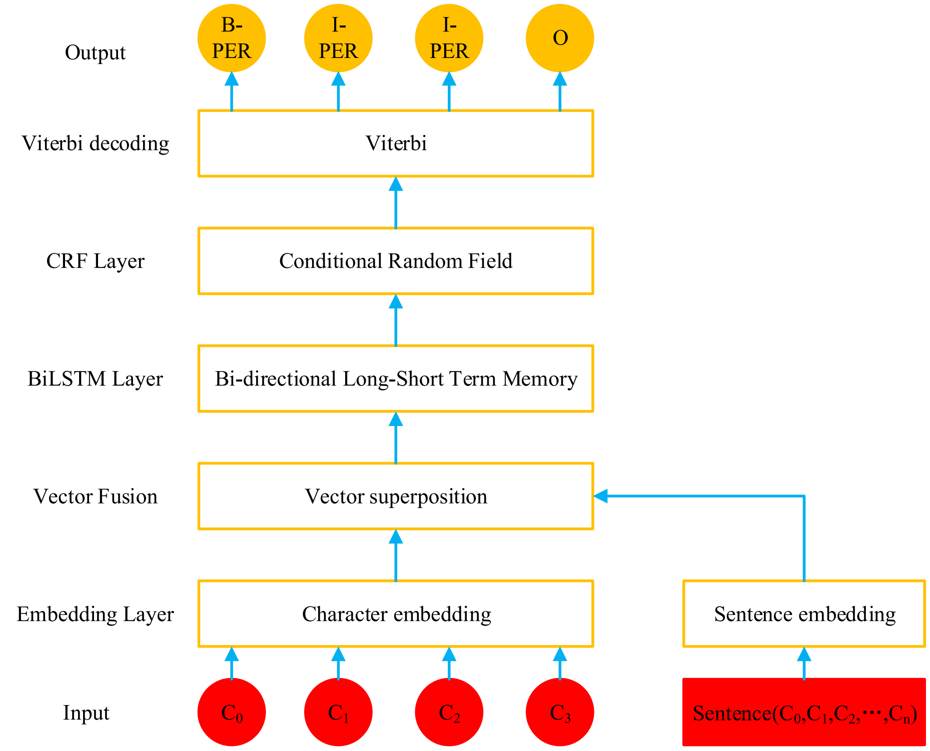

Fig. 5 Chief flow chart. The input on the left is the character, and the input on the right is the sentence containing the corresponding characters. After vectorizing the two inputs separately, the vectors are superposed as the input of the BiLSTM layer

Table 2 Person name substitution. Replace the anonymous name in the source text with another real name

Source text	虽然杨皓在作案过程中没有直接持刀，但在陈某打电话邀约杨皓报复徐某某时，已明确告知杨皓携带作案工具，杨皓也明确供述其知道陈某是让其带刀，杨皓在当时就能够预见可能发生造成人员伤亡的危害结果，故杨皓对于徐某某死亡结果的发生存在概括故意。 Although Yang Hao did not hold the knife directly in the process of committing the crime, when Chen called to invite Yang Hao to revenge Xu ,Yang Hao had clearly informed Yang Hao to carry the crime tools, and Yang Hao also clearly stated that he knew that Chen had brought the knife, and Yang Hao could foresee the possible harmful result of casualties at that time, so Yang Hao had a general intention for the occurrence of Xu 's death result.
Enhancement by named entity word substitution	虽然杨皓在作案过程中没有直接持刀，但在宋江打电话邀约杨皓报复董泽元时，已明确告知杨皓携带作案工具，杨皓也明确供述其知道甘志祥是让其带刀，杨皓在当时就能够预见可能发生造成人员伤亡的危害结果，故杨皓对于梁香莲死亡结果的发生存在概括故意。 Although Yang Hao did not hold the knife directly in the process of committing the crime, when Song Jiang called to invite Yang Hao to revenge Dong Zeyuan, he had clearly informed Yang Hao to carry the crime tools, Yang Hao also clearly stated that he knew Gan Zhixiang was to let him carry the knife, Yang Hao was able to foresee the possible result of injury and death of people at that time, so Yang Hao had a general intention for the occurrence of Liang Xianglian's death result.

un-normalized probability. CRF prediction algorithm typically uses the well-known Viterbi search to find the best solution. Each path corresponds to a state sequence. The optimal path is solved by dynamic programming, through which, the most suitable output sequence is found. The improved Viterbi algorithm is described as Table 1:

In the above algorithm description, $\delta_i(l)$ represents the cumulative output probability of each mark l = 1,2, ..., m at position i. $\varphi_i(l)$ records the previous path of each mark l at position i.

CRF can learn a constraint from training data to ensure the validity of the final prediction label, and this constraint condition is called transition score. Suppose a sentence S consisting of n characters is expressed as:

$$S = \{w_i, |i = 1, 2, ..., n\} \tag{7}$$

There are K possible sequence paths of S. We need to calculate the cumulative score for every path:

$$P_j = e^{p_j}, j \in [1, K] \tag{8}$$

$$p_j = EmissionScore_j + TransitionScore_j \tag{9}$$

$$EmissionScore_j = \sum_{i=1}^{n} s_e\left(c_i, label_m | m = 1, 2, ..., K\right) \tag{10}$$

$$TransitionScore_j = \sum_{i=1}^{n} s_t\left(label_i \rightarrow label_{i+1}\right) \tag{11}$$

where e is the natural number. The *Emission Score*, which is gained from the output of BiLSTM, means the sum of the scores of marking the character c_i as $label_m$ in sentence S, and $s_e(c_i, label_m)$ refers to the score of the ith character in mth path; $s_t\left(label_i \rightarrow label_{i+1}\right)$ represents the score of the ith label transferring to the (i+1)th label. As Viterbi algorithm needs a large amount of computation, for the purpose of improving efficiency in practical applications, it is beneficial to reduce the search space. We achieve this goal by cutting off the path with the lowest score as follows:

According to Begin-Inside-Outside (BIO) labeling rules, B-X represents the beginning of entity word X, I-X represents the middle and end of entity word X, and O is the non-entity word. As we know, the B-Person cannot be followed by the I-Org, that is, the probability of transiting from the B-Person to the I-Org is very small. So, we regard

Table 3 Penalty substitution. Replace the penalty with other penalty names that appear less in source text

Source text	经审查认为，原审生效裁判认定伍震犯故意伤害罪的事实清楚。 After examination, it is believed that the fact that Wu Zhen committed the crime of intentional injury is clear.
Enhancement by named entity word substitution	经审查认为，原审生效裁判认定伍震犯以危险方法危害公共安全罪的事实清楚。 After examination, it is believed that the fact of Wu Zhen's crime of endangering public safety by dangerous means is clear.

Table 4 Laws and regulations' substitution. Replace the common laws and regulations with other rare laws and regulations

Source text	综上，再审申请人王鹏飞再审申请不符合《中华人民共和国行政诉讼法》第九十一条规定的情形。 In conclusion, the retrial application of Wang Pengfei, the retrial applicant, does not meet the requirements of Article 91 of the administrative procedure law of the People's Republic of China.
Enhancement by named entity word substitution	综上，再审申请人王鹏飞再审申请不符合《中华人民共和国政府信息公开条例》第十一条第（三）项规定的情形。 In conclusion, the retrial application of Wang Pengfei does not meet the requirements of Article 11 (3) of the regulations of the People's Republic of China on the disclosure of government information.

this path as an impossible path and cut it off to reduce computation. The path we cut out is :

$$y'_n = \arg\min_{1\leq j\leq K} TransitionScore_j \tag{12}$$

In the BiLSTM−CRF model, the output of BiLSTM layer, that is *Emission Score*, is the input of CRF layer. And the final output is the tag sequence determined by CRF. The model structure is shown in Fig.4. Figure 5 shows the whole process of our method.

3.4 Data enhancement

As the NER model is trained on annotated data, it is critical that the latter covers well different types of named entities. This is difficult because of the large variety of named entities. In particular, in the legal domain, organizations often have very long names, and each name is only mentioned once or a few times in the dataset. In order to expose our model with various named entity samples, we use a data enhancement to enrich the training dataset of named entities. Data enhancement is processed in the following two steps:

1. The source texts are divided into sentences and then randomly recombined.
2. Named entities are randomly substituted with other entity names of the same type, collected in entity dictionary.

After the second step, we generate new instances in the training data. The enhancement approach can thus increase substantially the size of the training data. Tables 2, 3, and 4 show examples of different types of data enhancement.

Table 5 Statistic of dataset. It contains the amount of five entity types and labels, and the total amount

Entity type	Amount of entity	Amount of tags
Person name	5579	18062
Organization name	1573	5930
Laws and regulations	1224	5005
Accusation	1693	4938
Judgment	1936	6802
Total	12005	40737

4 Results and discussion

In this section, the experimental settings are first presented, including the dataset, the evaluation metrics, and the comparative methods. Then, experiments are carried out to investigate the proposed methods in the judicial field. Subsequently, the experimental results are compared with other methods. At last, a discussion on these results is given.

4.1 Experimental settings

4.1.1 Construction of dataset

The dataset used in our experiment was collected and cleaned manually by ourselves. It contains all over 260,000 words of various judicial documents obtained from the Net of Chinese Judicial Documents. The documents include criminal cases, civil cases, and administrative cases. Then, we manually annotate the obtained documents according to BIO rules. The annotated entity types include names of people, organizations, crimes, laws and regulations, and penalties. The statistic of our datasets is shown in Table 5. Seventy percent of these data are taken as training set, 10% as validation set and 20% as test set.

The original character vectors are derived from Wikidata, and then, they combined with sentence vector trained by PV-DM model using the proposed method in this paper. Due to the anonymization of some personal names in the obtained documents and the problem of

Table 6 Our method and other methods are tested in our dataset and the results are compared

Model	Precision	Recall	F1
Multiple HMMs [24]	71.64	71.06	71.34
Char baseline	68.79	60.35	64.30
+bichar+softword LSTM	74.36	69.43	71.81
Lattice LSTM [5]	76.35	71.56	73.88
Segment-level neural network [25]	71.60	69.40	70.48
BiLSTM-CRF	71.14	72.21	71.67
Our method	**77.08**	**73.69**	**75.35**

Table 7 Comparison of experimental results of five kinds of entities recognition of our method and the segment-level neural network

Methods	Entity type	Precision	Recall	F1
Our method	Person name	77.08	73.69	75.35
	Organization name	**66.69**	**61.48**	**64.10**
	Laws and regulation	**92.59**	**94.34**	**93.46**
	Accusation	76.67	71.43	73.95
	Penalty	**77.12**	**72.80**	**74.90**
	Overall	**77.08**	**73.69**	**75.35**
Segment-level neural network	Person name	**80.54**	**84.08**	**82.27**
	Organization name	62.55	69.68	65.92
	Laws and regulation	15.54	26.14	19.49
	Accusation	**82.12**	**75.12**	**78.47**
	Penalty	35.96	47.29	40.85
	Overall	71.60	69.40	70.48

over-fitting of the model caused by less mentioned names of relevant laws and regulations, a data enhancement method is adopted.

4.1.2 Compared methods and parameters

We compare our method with multiple HMMS [24], Lattice LSTM [5], and segment-level neural network with beam search [25] on our dataset.

We first created a mapping of characters and labels to obtain the corresponding index. The dimension of the character vector is set to 100. Dropout is adopted to prevent over-fitting and the dropout rate is set to 0.5. The learning rate is initially set to 0.005. The optimization algorithm adopts stochastic gradient descent (SGD) algorithm. The batch size and training epochs are set as 20 and 100, respectively. In each epoch, there are 100 iterations. The experimental environment is NVIDIA Quadro P2000 and the development language is Python3.6.

4.1.3 Evaluation metrics

The evaluation criteria used in this paper are *precision*, *recall*, and *F1-score*(*F1*), and the recognition effectiveness of each type of entity is evaluated separately. *Precision* is the proportion of correctly predicted entity tags to all predicted entity tags.

$$precision = \frac{EntityTag_{correct}}{EntityTag_{predicted}} \tag{13}$$

Recall is the proportion of correctly predicted entity labels to all entity labels in the sample.

$$recall = \frac{EntityTag_{correct}}{EntityTag_{all}} \tag{14}$$

The *F1* is the harmonic mean of precision and recall.

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \tag{15}$$

4.2 Experimental results

4.2.1 Overall effectiveness

Table 6 compares our method with several mainstream models based on character level information. It can be seen that our method performs the best. The tests on different types of judicial text samples show that the Fl score of multiple HMMs [24] can be stabilized above 71.34%. Lattice LSTM [5] obtained 76.35% for precision, 71.56% for recall, and 73.88% for Fl. Only character information is used in these methods. In our method, we added sentence information. This superior performance confirms that the combination of character vector and sentence vector is beneficial, and this makes it possible to learn deeper semantic features from the text, thus improving the effectiveness of named entity recognition in the field.

4.2.2 Effectiveness on different types of entity

The segment-level neural network model proposed by Wang et al. [25] is used in the field of legal documents. The recognition of named entities is completed by obtaining page information and assigning markers to pages as a whole. And the experiment reached 71.60% for precision, 69.40% for recall, and 70.49% for F1.

Comparing the results of our method and segment-level neural network in Table 7, it can be seen that the recognition effectiveness of the laws and regulations by the method in this paper is obviously higher than that of the segment level neural network, while the recognition effectiveness of the names is slightly lower than that of the latter. After analysis, we found that the lower recognition effectiveness on names is due to names translated from ethnic minority languages. Because these translated names and common Chinese names use different charac-

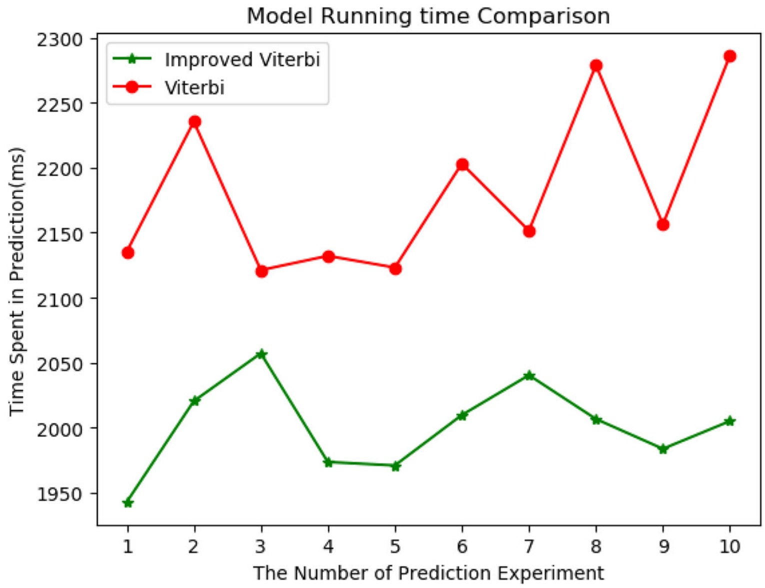

Fig. 6 Running time comparison of the model. We have carried out ten random experiments with the Viterbi algorithm and the improved algorithm, respectively. We can see that the improved algorithm takes less time and is more efficient

ter patterns, the recognition effectiveness on such names becomes poor.

4.2.3 Model performance comparison

Running time of the model. We randomly selected 70,000 words of judgment documents and conducted ten experiments using Viterbi and our method respectively to compare the prediction running time of the two methods. Figure 6 shows running time comparison between the two methods. We can see that our method takes less time. Under the same experimental conditions, when using the original Viterbi algorithm, the average prediction time is 2187.65 milliseconds, and 2009.41 milliseconds for the improved algorithm.

Comparison of evaluation metrics using our method and Viterbi. We also analyzed the influence of path pruning on the evaluation metrics of the model. We compared precision, recall, and F1 value of our method and Viterbi. As shown in Table 8, they are overall effectiveness comparison and effectiveness on different types of entity, respectively. We can see that our method does have a certain impact on the performance of the model, but the loss is not much. We believe that it is reasonable to improve the efficiency of the model with the loss of lower precision,

Table 8 Comparison of overall effectiveness of our method and Viterbi

Method	Entity type	Precision	Recall	F1
Our method	Person name	73.09	76.54	74.78
	Organization name	66.69	61.48	64.10
	Laws and regulations	92.59	94.34	93.46
	Accusation	76.67	74.43	73.95
	Penalty	77.12	72.80	74.90
	Overall	77.08	73.69	75.35
Viterbi	Person name	**74.13**	**77.02**	**75.55**
	Organization name	**68.22**	**62.58**	**65.28**
	Laws and regulations	**94.41**	**95.74**	**95.07**
	Accusation	**77.38**	**74.95**	**76.15**
	Penalty	**78.43**	**74.02**	**76.16**
	Overall	**78.13**	**74.56**	**76.30**

Table 9 An example to show the effectiveness of data enhancement. Given a paragraph containing two law names. Before using data enhancement, only one of them can be recognized. But all entities can be recognized when data enhancement is adopted

Input	依照《中华人民共和国刑法》第七十八条、第七十九条和《中华人民共和国刑事诉讼法》第二百六十二条第二款之规定，裁定如下： In accordance with the provisions of Articles 78 and 79 of the Criminal Law of the People's Republic of China and the second paragraph of Article 262 of the Criminal Procedure Law of the People's Republic of China, the judgment is as follows:
Before data enhancement	命名实体：《中华人民共和国刑法》，实体类型：法律名称 Named entity: Criminal Law of the People's Republic of China, Entity Type: Law
Data enhancement	命名实体：《中华人民共和国刑法》，实体类型：法律名称。 命名实体：《中华人民共和国刑事诉讼法》，实体类型：法律名称 Named entity: Criminal Law of the People's Republic of China, Entity Type: Law. Named entity: Criminal Procedure Law of the People's Republic of China, Entity Type: Law

recall, and F1 value. In practical application, the working efficiency of the system is as important as the accuracy.

Viterbi algorithm is a dynamic programming algorithm. When solving the state transition path, the result is good enough, but it does not mean that it is optimal. The method proposed in this paper is based on the original Viterbi algorithm to prune the path only according to the current score. So in this process, the better path may be cut off, making the performance of the model reduced to a certain extent.

4.2.4 Impact of data enhancement

As we described, data enhancement can increase the size of training data by adding artificially generated instances. The example given in Table 9 shows that the impact of data enhancement. The input sentence contains two entities. Without data enhancement, the model can only recognize one of them. The model trained using data enhancement recognizes both named entities.

4.3 Discussion

The research on named entity recognition in the legal field can lay a foundation for the related research of judicial intelligence and promote the development of intelligent multimedia devices such as information forensics and intelligent consulting service system. We can use this system to simplify the case process without manual work, so as to better protect the privacy of the parties.

Although the proposed method achieved satisfactory performance in named entity recognition, it still has following limitations. Firstly, the method proposed in this paper is only used for named entity recognition in legal field because the texts in different fields have different characteristics. If we want to use the method in other fields, we may need to further improve it to adapt to new domain characteristics. So, the domain mobility needs to be improved. Secondly, there are differences in the number of words, form, and meaning between the translation names of ethnic minorities and the general Chinese person names, so the recognition performance of the method in this paper needs to be improved. At last, the classification of legal named entity types needs further mining and refinement, such as time and location category, so as to achieve fine-grained named entity recognition.

5 Conclusion

In this paper, we investigated NER in judgment documents by taking into account their characteristics. A dataset of judgment documents is built with manual annotation of named entities. The Viterbi algorithm is improved by cutting off the path with the lowest score, and this is shown to improve the efficiency of the algorithm. The PV-DM model is used to train the sentence vector of the text, which is then combined with the character vector to make the model more capable of capturing sentence information and other features, thus making more accurate prediction. In the future work, the identification of names translated by ethnic minority languages will be studied to improve the precision of identification, as this has been found to be the main cause of relatively low recognition effectiveness on personal names.

Abbreviations

NER: Named entity recognition; CNER: Chinese named entity recognition; CRF: Conditional random field; BiLSTM: Bi-directional long short-term memory; PV-DM: Distributed memory model of paragraph vectors

Acknowledgements

The authors thank the referees and editors' careful reviewing and constructive suggestions.

Authors' contributions

Wenming Huang and Zhenrong Deng conceived the structure of the manuscript and gave analytical methods. Dengrui Hu performed the experiments and analyzed the results and write the manuscript. Jianyun Nie polished the language. The authors read and approved the final manuscript.

Authors' information

Huang Wenming is a professor in the School of Computer and Information Security, Guilin University of Electronic Technology, China. He is mainly

engaged in the research, development, and teaching of big data processing, graphics and image processing, and software engineering. In recent years, he has presided over 18 scientific research projects. As the second person in charge, he has undertaken 5 scientific research projects. Many scientific research achievements have been promoted and applied. He has published more than 60 academic papers, including 23 core journals and 27 EI journals.
Dengrui Hu received the BE degree from City College, Wuhan University of Science and Technology, China, in 2017. Now, he is studying in the Guilin University of Electronic Technology. He is majoring in nature language processing.
Zhenrong Deng is a professor in the School of Computer and Information Security, Guilin University of Electronic Technology, China. She is mainly engaged in the research of big data processing, nature language processing, and graphics and image processing. She presided over the completion of 9 scientific researches; published more than 20 papers; as the first draft of the standard, he successfully registered 1 enterprise standard; successfully obtained 2 software product registration certificates for cooperative enterprises.
Jianyun Nie is a full-time professor at the University of Montreal. His main research field is information retrieval and natural language processing. He graduated from the University of Grenoble in France with a doctor's degree. Professor Nie has published more than 150 research papers and Monographs on cross language information retrieval in international journals and conferences. Professor Nie Jianyun served as the program committee member of SIGIR, ACL, CIKM, and other well-known international conferences and also served as the president of SIGIR 2011 conference. At the same time, Professor Nie is also the editorial board member of seven international journals.

Funding

The work presented in this paper was supported by the Guangxi Science and Technology Project (Nos. AB20238013, ZY20198016, and 2018GXNSFAA294127) and by Innovation Project of Guet Graduate Education (No. 2019YCXS050) and by the Research Projects of Key Laboratory of Intelligent Processing of Computer Image and Graphics (No: GIIP2003).

Availability of data and materials

We do not open our experimental dataset.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]School of Computer Science and Information Security, Guilin University of Electronic Technology, 541004 Guilin, China. [2]Guangxi Key Laboratory of Intelligent Processing of Computer Image and Graphics, Guilin University of Electronic Technology, Guilin, China. [3]University of Montreal, Montreal, Canada.

Received: 26 February 2020 Accepted: 12 October 2020
Published online: 23 November 2020

References

1. E. F. Sang, F. De Meulder, in *Proceedings of CoNLL-2003*. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition, (Edmonton, 2003), pp. 142–147
2. H. Zhu, W. Hu, Y. Zeng, in *CCF International Conference on Natural Language Processing and Chinese Computing*. Flexner: a flexible LSTM-CNN stack framework for named entity recognition (Springer, Cham, 2019), pp. 168–178
3. D. Nadeau, S. Sekine, A survey of named entity recognition and classification. Lingvisticae Investigationes. **30**(1), 3–26 (2007)
4. A. Chen, F. Peng, R. Shan, G. Sun, in *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*. Chinese named entity recognition with conditional probabilistic models (Association for Computational Linguistics (ACL), Sydney, 2006), pp. 173–176
5. Y. Zhang, J. Yang, Chinese NER using lattice LSTM. ACL. **1: Long Papers**, 1554–1564 (2018)
6. L. Sun, J. Ma, H. Wang, Y. Zhang, J. Yong, Cloud service description model: an extension of USDL for cloud services. IEEE Trans. Serv. Comput. **11**(2), 354–368 (2018)
7. L. Sun, H. Dong, O. K. Hussain, F. K. Hussain, A. X. Liu, A framework of cloud service selection with criteria interactions. Futur. Gener. Comput. Syst. **94**, 749–764 (2019)
8. X. Ma, E. Hovy, End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. ACL. **1**, 1064–1074 (2016)
9. G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, in *Proceedings of NAACL*. Neural architectures for named entity recognition (Association for Computational Linguistics (ACL), San Diego, 2016)
10. M. Collins, Y. Singer, in *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. Unsupervised models for named entity classification, (College Park, 1999)
11. A. Mikheev, M. Moens, C. Grover, in *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics*. Named entity recognition without gazetteers (Association for Computational Linguistics, Stroudsburg, 1999), pp. 1–8
12. J. Wang, H. Wanga, J. Li, X. Luo, Y.-Q. Shi, S. K. Jha, Detecting double JPEG compressed color images with the same quantization matrix in spherical coordinates. IEEE Trans. Circ. Syst. Video Technol. **30**, 2736–2749 (2019)
13. J. Wang, T. Li, X. Luo, Y.-Q. Shi, S. K. Jha, Identifying computer generated images based on quaternion central moments in color quaternion wavelet domain. IEEE Trans. Circ. Syst. Video Technol. **29**(9), 2775–2785 (2018)
14. R. Lan, Y. Zhou, Z. Liu, X. Luo, Prior knowledge-based probabilistic collaborative representation for visual recognition. IEEE Trans. Cybern. **50**(4), 1498–1508 (2020)
15. R. Lan, L. Sun, Z. Liu, H. Lu, Z. Su, C. Pang, X. Luo, Cascading and enhanced residual networks for accurate single-image super-resolution. IEEE Trans. Cybern., 1–11 (2020). https://doi.org/10.1109/TCYB.2019.2952710
16. R. Lan, L. Sun, Z. Liu, H. Lu, C. Pang, X. Luo, MADNet: a fast and lightweight network for single-image super resolution. IEEE Trans. Cybern., 1–11 (2020). https://doi.org/10.1109/TCYB.2020.2970104
17. J. Hammerton, in *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL*. Named entity recognition with long short-term memory. vol. 4 (Association for Computational Linguistics, Edmonton, 2003), pp. 172–175
18. H. M. Mo, K. M. Soe, Syllable-based neural named entity recognition for Myanmar language. Int. J. Nat. Lang. Comput. (IJNLC). **536**, 204–211
19. D. N. Anh, H. N. Kiem, V. N. Van, in *Proceedings of the 2019 IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF), Danang, Vietnam*. Neural sequence labeling for Vietnamese POS tagging and NER, (Vietnam, 2018), pp. 1–5. arXiv preprint arXiv:1811.03754
20. F. Wu, J. Liu, C. Wu, Y. Huang, X. Xie, in *The World Wide Web Conference*. Neural chinese named entity recognition via CNN-LSTM-CRF and joint training with word segmentation (Association for Computing and Machinery (ACM), New York, 2019)
21. J. Gao, M. Li, C.-N. Huang, A. Wu, Chinese word segmentation and named entity recognition: a pragmatic approach. Computat. Linguist. **31**(4), 531–574 (2005)
22. L. Qiu, Y. Zhang, in *Twenty-Ninth AAAI Conference on Artificial Intelligence*. Word segmentation for chinese novels (Association for the Advancement of Artificial Intelligence (AAAI), Menlo Park, 2015), pp. 2440–2446
23. C. Xu, F. Wang, J. Han, C. Li, in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. Exploiting multiple embeddings for chinese named entity recognition (ACM, New York, 2019), pp. 2269–2272
24. Z. Xiaohui, Legal named entity recognition model based on implicit Markov model. J. Comput. Appl. **2**, 365–368 (2017)
25. L. Wang, Y. Xie, J. Zhou, Y. Gu, W. Qu, Segment-level chinese named entity recognition based on neural network. J. Chin. Inf. Process. **32**(3), 84–90 (2018)
26. J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding. CoRR. **abs/1810.04805**, 4171–4186 (2018). http://arxiv.org/abs/1810.04805
27. X. Li, H. Zhang, X.-H. Zhou, Chinese clinical named entity recognition with variant neural structures based on bert methods. J. Biomed. Inform. **107**, 103422 (2020)
28. Y. Chen, C. Zhou, T. Li, H. Wu, X. Zhao, K. Ye, J. Liao, Named entity recognition from chinese adverse drug event reports with lexical feature based BiLSTM-CRF and tri-training. J. Biomed. Inform. **96**, 103252 (2019)
29. Y. Zhu, G. Wang, B. F. Karlsson, *CAN-NER: convolutional attention network for Chinese named entity recognition. NAACL-HLT, Vol. 1 (Long and Short Papers)*. (Association for Computational Linguistics, Minneapolis, 2019), pp. 3384–3393. https://doi.org/10.18653/v1/N19-1342

30. D. Galvez-Lopez, J. D. Tardos, in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*. Real-time loop detection with bags of binary words (IEEE, 2011), pp. 51–58
31. D. Gálvez-López, J. D. Tardos, Bags of binary words for fast place recognition in image sequences. IEEE Trans. Robot. **28**(5), 1188–1197 (2012). Piscataway
32. Q. Le, T. Mikolov, in *International Conference on Machine Learning*. Distributed representations of sentences and documents (International Machine Learning Society (IMLS), Beijing, 2014), pp. 1188–1196
33. S. Hochreiter, J. Schmidhuber, Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.