electronics

MDPI

Article

Innovative Deep Neural Network Modeling for Fine-grained Chinese Entity Recognition

Jingang Liu [1,2,3], Chunhe Xia [2,3], Haihua Yan [1,3,*] and Wenjing Xu [3]

[1] State Key Laboratory of Software Development Environment, School of Computer Science and Engineering, Beihang University, Beijing 100191, China; ljg@buaa.edu.cn
[2] Beijing Key Laboratory of Network Technology, School of Computer Science and Engineering, Beihang University, Beijing 100191, China; xch@buaa.edu.cn
[3] School of Computer Science and Engineering, Beihang University, Beijing 100191, China; xuwenjing@buaa.edu.cn
* Correspondence: yhh@buaa.edu.cn

Received: 26 May 2020; Accepted: 11 June 2020; Published: 15 June 2020

Abstract: Named entity recognition (NER) is a basic but crucial task in the field of natural language processing (NLP) and big data analysis. The recognition of named entities based on Chinese is more complicated and difficult than English, which makes the task of NER in Chinese more challenging. In particular, fine-grained named entity recognition is more challenging than traditional named entity recognition tasks, mainly because fine-grained tasks have higher requirements for the ability of automatic feature extraction and information representation of deep neural models. In this paper, we propose an innovative neural network model named En2BiLSTM-CRF to improve the effect of fine-grained Chinese entity recognition tasks. This proposed model including the initial encoding layer, the enhanced encoding layer, and the decoding layer combines the advantages of pre-training model encoding, dual bidirectional long short-term memory (BiLSTM) networks, and a residual connection mechanism. Hence, it can encode information multiple times and extract contextual features hierarchically. We conducted sufficient experiments on two representative datasets using multiple important metrics and compared them with other advanced baselines. We present promising results showing that our proposed En2BiLSTM-CRF has better performance as well as better generalization ability in both fine-grained and coarse-grained Chinese entity recognition tasks.

Keywords: neural network; information extraction; deep learning; fine-grained; NER; NLP

1. Introduction

With the vigorous development of big data mining technology, named entity recognition [1–4] as the most important subtask of natural language processing (NLP) tasks is playing an essential role in many fields driven by data [5–7]. NER systems aim to recognize various entities from a large number of unstructured texts according to the required naming types. Early named entity recognition tasks focused on a few types of entities. As studies continue, the work based on fine-grained named entity recognition [8,9] is becoming more and more popular. However, fine-grained entity recognition is more challenging, mainly because this recognition is more difficult and there are few high-quality, fine-grained Chinese NER datasets for research [9]. Besides, the recognition of named entities in Chinese is more complicated and difficult than in English. This is mainly due to the absence of separators in a Chinese text that indicates the boundaries of words, and the recognition effect of named entities is greatly affected by the results of automatic word segmentation [10,11]. Besides, Chinese often has a phenomenon that a word has multiple meanings, which brings more challenges to the recognition of Chinese named entities.

Traditional named entity recognition methods are mainly rule-based [11] and statistical learning-based [12]. Although these traditional methods are more interpretable, their recognition effects depend to a greater extent on the modeling capabilities and rules for specific tasks. In particular, the task of fine-grained NER requires higher modeling capabilities of traditional methods. In recent years, deep learning-based NER methods [3,13–16] have become dominant and achieve state-of-the-art results, since high-performance computers can replace humans to automatically extract useful features from unstructured text. Classical deep neural networks such as long short-term memory (LSTM) networks [17] and convolutional neural networks (CNNs) [18] and their variants are very popular and perform well on some specific tasks. However, the limitations of these deep neural structures still exist. For example, unidirectional LSTM networks have difficulty extracting comprehensive contextual features. Moreover, the CNN structure has a shortcoming in the sequence labeling task, that is, it cannot remember the long-term historical information in the Chinese text except for the spatial features. To make better use of the contextual information in the text, the bidirectional LSTM (BiLSTM) [19–21] structure that can extract contextual features from the opposite direction has become the first considered method for named entity recognition tasks in recent years. In recent years, methods [21,22] combining conditional random field (CRF) [23,24] models with the LSTM networks have achieved great success, mainly because they can use rich hidden information and contextual feature information in the process of tagging each location. However, if the structures used are relatively simple, the context information of the text cannot be fully utilized. Considering that the success of an NER system greatly depends on the selection of input features, we can use the existing character semantic information to enrich the input information. The emergence of pre-training models based on NLP tasks, such as BERT [25], provides researchers with a better method of coding the representation of the input.

In this paper, we propose an innovative neural network model, named En2BiLSTM-CRF, to improve the recognition effect of fine-grained Chinese named entities. This end-to-end NER model combines the advantages of the pre-trained model, the dual BiLSTM networks, the residual connection mechanism [26], and the conditional random field. The proposed model uses a structure of three module layers to complete the initial encoding and enhanced encoding as well as decoding of the input data hierarchically, thus ensuring that sufficient semantic information and context features are used for fine-grained entity recognition tasks. In the initial encoding layer, motivated by the idea that performance gains can be achieved by initialized semantic representation provided by the pre-trained model, we use the recent ALBERT [27] model to encode our original input. In the enhanced encoding layer, we use a combination of two BiLSTM networks to process the encoded information output by the previous layer hierarchically. We use the combination of dual BiLSTM networks to process the coding information output by the previous layer hierarchically, and then use the results of two feature extractions to enhance the re-encoding ability. In the decoding layer, we use the softmax [28] algorithm and the CRF model to decode the re-encoded information to calculate the most likely recognition probability of the required entities, thus ensuring the effect of fine-grained named entity recognition.

The contributions of this paper can be summarized as follows:

(1) We propose a universal fine-grained named entity recognition model, En2BiLSTM-CRF, which includes three effective modules: the initial encoding module, the enhanced encoding module, and the decoding module. Our method can fully characterize and encode the original input information, that is, it has powerful encoding and text representation capabilities based on these advantages to finish high-precision sequence labeling tasks;

(2) We use an innovative hierarchical feature extraction and merging method to strengthen the feature weight of text sequences. Besides, this novel but uncomplicated method highlights the weight of character-level representation. These enhanced representations and weights play an essential role in the process of fine-grained entity recognition;

(3) We conducted sufficient experiments on the latest fine-grained public dataset and another classic benchmark dataset. We also used three important indicators on two datasets to compare with other advanced baselines. The experimental results show that our En2BiLSTM-CRF model

outperforms all the baselines. In particular, the proposed method is more effective in the fine-grained tasks, far surpassing other state-of-the-art baselines. Excellent performance on three metrics proves the generalization ability and robustness of the proposed model.

2. Related Work

2.1. Fine-Grained Named Entity Recognition

The main goal of the NER task is to extract as many entities of the required types as possible from a large amount of unstructured text. From a practical perspective, named entity recognition has important application value in many fields [5–7]. The coarse-grained entity recognition task usually only needs to extract a few basic entity types, such as the person name, the place names, and the organization name. Although fine-grained [8,9] NER has a wider application value, it is also more challenging. Due to the inherent characteristics of Chinese, the Chinese named NER task is more challenging than English. For example, the effect of Chinese word segmentation [10] seriously affects the effect of the NER task, while English does not require word segmentation. Early researchers mainly used rules-based and statistical-based learning methods [11,12]. However, the recognition accuracy is often not guaranteed. With the development of deep learning technologies, more and more researchers use different deep neural network models [13,14] to train entity recognition systems and have achieved obvious results. [14] once tried to combine deep learning techniques with statistics-based methods. [15,16] used to design NER systems based on convolutional neural networks (CNNs). The main principle of the NER model based on neural networks is to semantically represent a text by automatically learning the context or spatial features. In addition to deep learning models that can effectively learn useful representations and potential factors from raw data, they can also be trained in an end-to-end paradigm [3]. Since more and more deep learning methods continue to improve on Chinese NER tasks, NER research based on deep neural networks has become mainstream. In recent years, widely-used encoders of NER models based on deep learning are usually some kinds of neural network structures [29,30]. To reduce training time and training costs, transformers [31,32] are sometimes used as encoders for NLP tasks. Further, using one or several linear layers in conjunction with softmax or the conditional random field as decoders is still the most typical and popular method at present [3,33].

2.2. Bidirectional LSTM Networks

Long short-term memory networks are recurrent neural networks [34] equipped with the special gating mechanism that controls access to memory cells [35]. A typical LSTM network achieves information protection and control through three gate structures, which are the input gate, the forget gate, and the output gate, respectively. Besides, the advantages of these gate mechanisms can remember the long-term historical information and forget the unimportant information, so the LSTM structure helps solve the problem of gradient disappearance or gradient explosion during model training. The BiLSTM [36] structure is a neural network structure including a forward LSTM network and a backward LSTM network. Therefore, one of the advantages of the BiLSTM network is that it can model the context of sequence data from two directions, thus improving the semantic representation ability. However, it is difficult to learn the constraints between characters using the BiLSTM network alone, and there will be many invalid prediction labels, that is, when using BiLSTM alone, the prediction labels of each element in the sequence are independent of each other, and there is no dependency relationship, so the CRF layers will be connected later [19,22].

2.3. Pre-Trained Models

In recent years, the rapid growth of NLP technology has mainly benefited from the concept of transfer learning [37] through pre-trained models. Pre-training models refer to training the model on a large-scale corpus and then optimizing for a specific task to produce better results. However, pre-trained models are not omnipotent, but sometimes we can use their strengths, that is, we can use the output vector to replace word2vector [38] to pre-encode the input information. As the BERT [25]

model has been proposed and achieved great success, research based on pre-trained models has received increasing attention. After BERT, some more powerful pre-training models have gradually appeared in the field of natural language processing, such as Xlnet [39] and ERNIE [40]. These pre-training models have different advantages. In particular, the emergence of the ALBERT [27] model attracted widespread attention last year, and it is a new model. ALBERT, which is a lite BERT for the self-supervised learning of language representations, has the advantage of significantly reduced parameters and lower training costs based on a guaranteed performance. Although there is not much research on using ALBERT models for detailed downstream tasks, we think this is a trend.

3. Proposed Method

In this paper, we propose a deep neural network model named En2BiLSTM-CRF that enhances the semantic feature representation capability of encoded information to improve the recognition effect of Chinese fine-grained named entities. In summary, this model combines the advantages of ALBERT, two BiLSTM networks for feature extraction, and calculating the weight of semantic representation hierarchically. At the same time, the CRF model with a dense layer is also used as an information decoder because of its advantages in NER tasks. The basic framework of our En2BiLSTM-CRF model is shown in Figure 1. In the following, we will introduce the proposed method in detail from three parts: the initial encoding layer, the enhanced encoding layer, and the decoding layer.

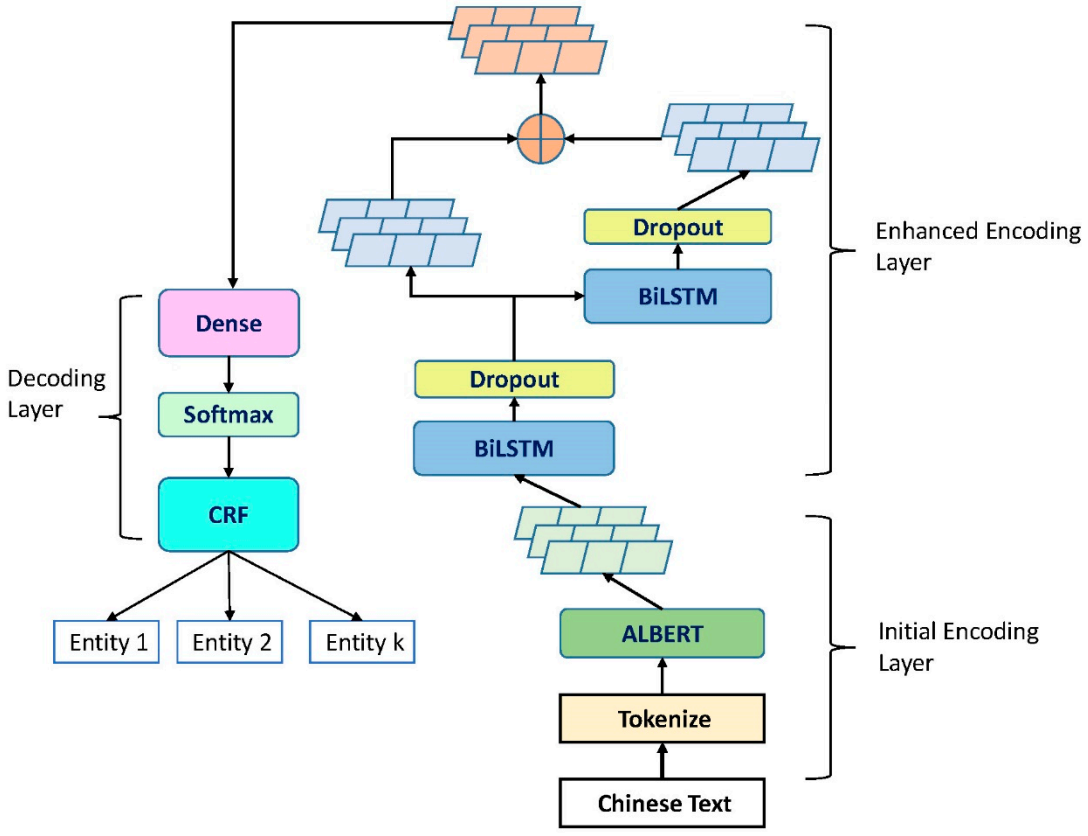

Figure 1. The basic framework of our En2BiLSTM-CRF model. It shows the main process of fine-grained entity recognition.

3.1. Initial Encoding Layer

In this layer, we use ALBERT as the precoding model to initially encode the input information. Specifically, we use the pre-trained ALBERT model as a feature extractor to initialize the characters in the original corpus, that is, use the feature weights output by this pre-trained model. The

advantage is that it provides a richer semantic representation for the input information and reduces the cost of model training in the later stage. The original input information after the initialization encoding can be directly used as the character embeddings of the enhanced encoding layer. Note that if the original input has not been preprocessed, we first need to clean the original data, cut the words, and tag the data. The schematic flow of initializing the input in this layer is shown in Figure 2. According to the design of our model, if we use function L to represent the feature characterization of the ALBERT model, and e is represented as the encoded feature vector, the relationship between them is as shown in

$$e_k = \sum_{i=1}^{l} L_i(x_j). \quad (1)$$

where x_j is the j-th character in each input sentence, k accounts for the k-th sentence in the input, i accounts for the i-th dimension in each character represented as an l-dimensional vector, and $\sum$ accounts for the expansion of the vector dimensions.

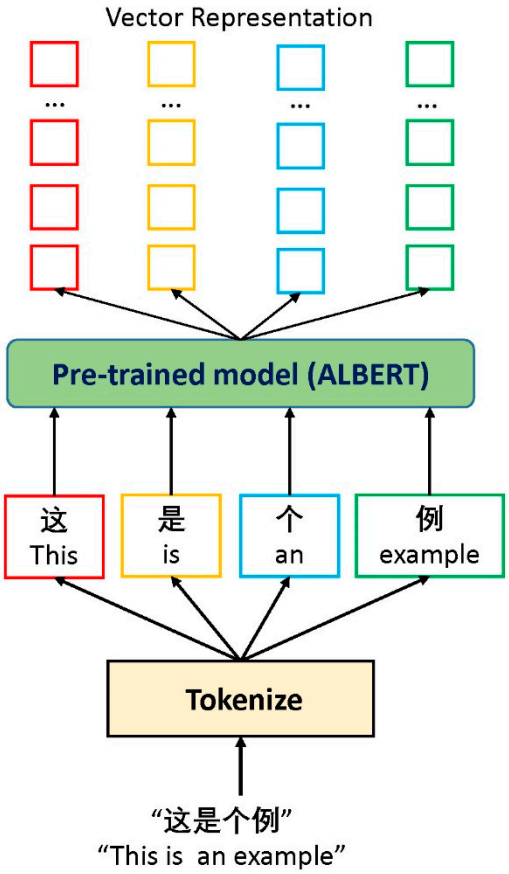

Figure 2. An example of input information being processed by the initial encoding layer in the proposed model. It shows the schematic flow of initializing the input.

The structure of ALBERT [27] is similar to that of BERT, which uses multi-layer transformers to encode the embeddings in two directions. The ALBERT model also has the cross-layer parameter sharing mechanism to significantly improve the parameter efficiency. In this work, we use a tiny version of the Chinese pre-trained ALBERT model provided by [41], which is named albert_tiny_zh. Compared with the basic BERT model, although the size of this ALBERT is only 16 MB (about 1/25 of BERT), the training speed and prediction speed are increased by nearly 10 times without a significant loss of accuracy. Specifically, the ALBERT model we used just includes four transformer layers.

When using the proposed model, the main reason we consider character-level representations as the basic input is that most pre-trained models, including ALBERT, are character-level. Besides,

current Chinese word segmentation technologies still have various problems in terms of accuracy. If the word segmentation is wrong, it will mislead the model training process. Another obvious advantage of character-level representation is that it can naturally handle out-of-vocabulary (OOV). Therefore, character-based models can infer the representation of invisible words.

3.2. Enhanced Encoding Layer

Normally, the contextual features hidden in the text have a great role in understanding the meaning and semantics of entities contained in this text. In this layer, to better contextually model the Chinese text containing the named entities to be recognized, we use two BiLSTM networks to extract contextual features and sequence information hierarchically. Each BiLSTM network in our En2BiLSTM-CRF model includes two LSTM networks in opposite directions. One of the two LSTM networks models sentences from beginning to end, while the other network models from back to front. After the two LSTMs in opposite directions complete the processing of the sentence, the two sets of feature vectors are automatically spliced. Specifically, if H_x accounts for the representative dimension of the output of the forward LSTM network, and H_y accounts for the representative dimension of the output of the reverse LSTM network, then the total dimension H of the output of this BiLSTM can be expressed as

$$\mathrm{H} = \mathrm{H}_x + \mathrm{H}_y. \quad (2)$$

Besides, a residual connection mechanism is also used in our model, which is also called a shortcut [26]. To strengthen the feature weights in semantic representation, the feature space trained by our first BiLSTM network and the feature space trained by two BiLSTM networks are merged in a point-to-point manner. In other words, the merged dimensions are unchanged, only the weight values are added. If E represents the output of the first layer and W represents the output of a BiLSTM network, then the total output E' of the second layer can be expressed as

$$E' = \mathrm{W}(E) + \mathrm{W}\big(\mathrm{W}(E)\big) \quad (3)$$

To alleviate overfitting, we follow a dropout layer behind each BiLSTM network. The basic structure of this enhanced encoding layer is shown in Figure 3. The weighted feature vectors after the residual connection using the merge operation (the shortcut) are used as the input of the decoding layer.

In this work, each BiLSTM network we use contains 128 units, which can convert the specific feature representation dimensions into 256 dimensions. In other words, no matter what the input dimension of each BiLSTM is, we use 128 units to convert it to 256 dimensions.

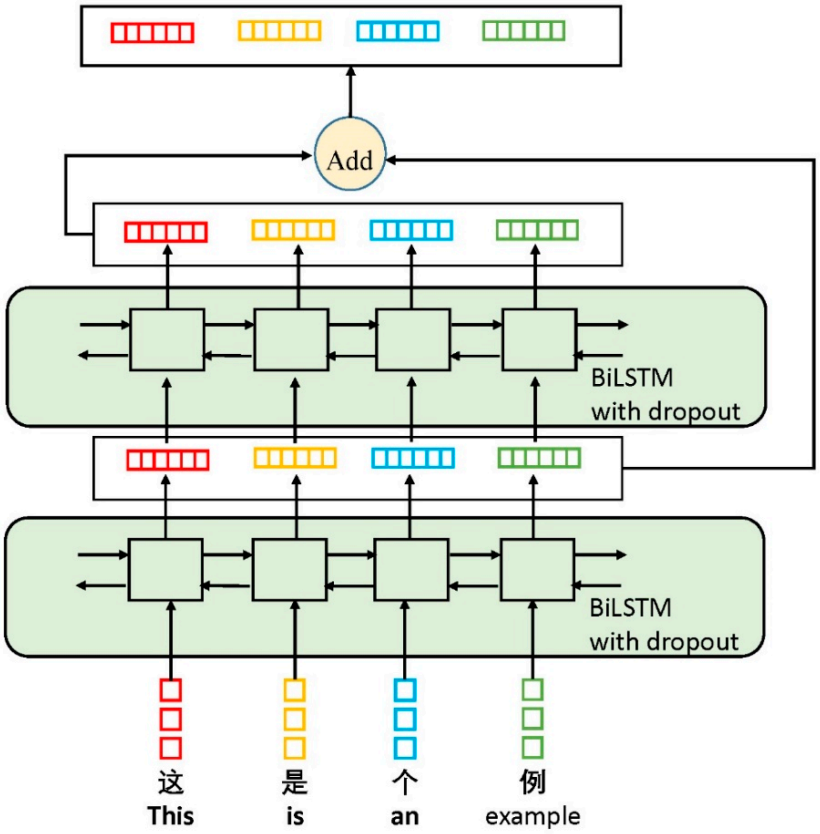

Figure 3. The basic structure of the enhanced encoding layer in the En2BiLSTM-CRF model.

3.3. Decoding Layer

In the decoding layer, when the output of the previous layer enters this layer, it is first processed by a dense neural network. The number of output units of the dense network should be the same as the number of all labeled types of the named entity to calculate the probability using the softmax function. Then, we use the CRF model to calculate the most likely categories of named entities. The CRF used in our model can add some constraints to ensure that the final prediction result is effective, and these constraints can be automatically learned by the CRF layer when training data. The input dimension of the dense layer in the enhanced encoding layer needs to match the output of the second layer. Besides, the output of the dense layer depends on the number of categories. For example, if the output dimension of the enhanced encoding layer is 256 and the number of labeled categories is 22, then the size of the input and output dimensions of the dense layer are 256 and 22.

3.4. Hyperparameters of the Proposed Model

In this section, we describe the hyperparameters used in the proposed En2BiLSTM-CRF model. We use a tiny pre-trained ALBERT model [41] (version: albert_tiny_zh) to encode the input information, and the input is represented as 312-dimensional vectors. In this model, we use two BiLSTM networks with the same parameters (128 units), and they can convert the input feature representation dimensions into 256 output dimensions. We also use a dropout rate of 0.1 and use 0.01 as the default learning rate. In addition, the idea of the viterbi algorithm is used in the CRF layer to calculate the optimal solution. To clearly show details of the proposed method, we show the input and output shapes of each processing module in the entire model in Figure 4. Where "len" accounts for the length of the input, "batchsize" accounts for the number of batches during training, and "number of tags" means the types of sequence labeling of the text in the corresponding dataset. To ensure that the output dimension of each BiLSTM is 256, we set the number of units in these two BiLSTM networks to 128.

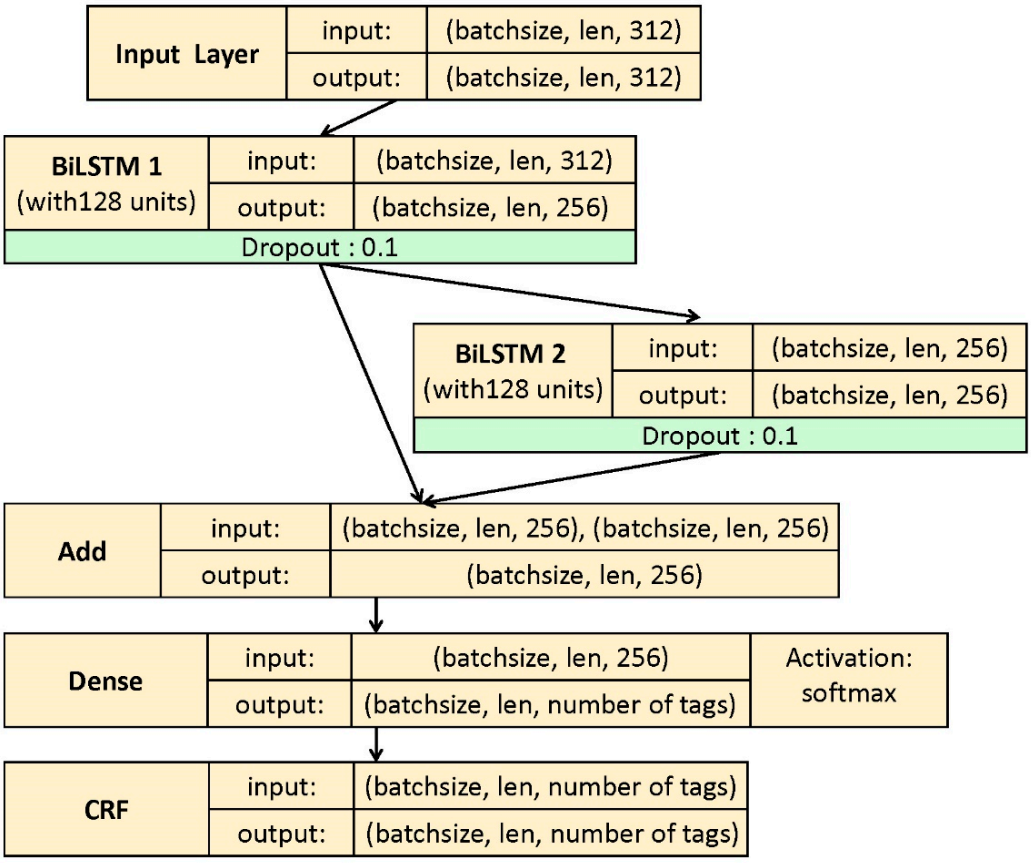

Figure 4. Input and output parameters of the proposed model.

4. Experiments and Settings

4.1. Dataset Description

In this paper, we use two very representative and convincing real-world datasets to evaluate the effect of our work. These two datasets with different characteristics can well reflect the real performance of the proposed model. In all experiments, we use the training set for model training and verify the effect and performance of models on the test set.

4.1.1. CLUENER2020

CLUENER2020 [9] is the latest fine-grained Chinese named entity recognition dataset. It is a well-defined fine-grained pubic dataset containing up to 10 named entity types, that is more challenging than other existing datasets. In particular, the corpus content of this dataset is Chinese news, which mainly includes the following ten entity types: movie, scene, book, address, position, government, company, name, organization, and game. It is easy to find that these entity types are very practical and important types in text information extraction tasks. What is commendable is that there are few such fine-grained specialized data sets. Detailed information about CLUENER2020 is shown in Table 1. The distribution of each entity included in its 1434 test set is shown in Table 2.

Table 1. Description of two representative datasets.

Dataset	Categories	Train	Test	Size
CLUENER2020	10	10,748	1,343	13 kB
People's Daily Ner Corpus	3	20,864	4,636	23 kB

Table 2. Detailed entity distribution in two test sets.

Dataset	Test	Entity	Support
CLUENER2020	1,343	movie	150
		book	152
		government	244
		company	366
		position	425
		scene	199
		organization	344
		name	451
		address	364
		game	287
People's Daily Ner Corpus	4,636	PER	1864
		LOC	3658
		ORG	2185

4.1.2. People's Daily Ner Corpus

The People's Daily Ner Corpus [42] is a very classic named entity recognition dataset. Since it is a benchmark dataset, it is often used by researchers to evaluate the effect of NER tasks. Although the People's Daily Ner Corpus only includes three types of entities (location, organization, and person name), it can still reflect the effect of the model well due to a large number of samples and the standard of data labeling. Through this dataset, we can verify whether our En2BiLSTM-CRF model still performs well in a dataset that contains fewer entity types. Detailed information about this dataset is shown in Table 1. The distribution of each entity included in its 1434 test set is shown in Table 2.

4.2. Evaluation Metrics

To verify the effect and performance from multiple aspects, we selected three important metrics (precision, recall, and F1-score) to evaluate these Chinese named entity recognition tasks.

- Precision

  Precision refers to how many positive samples are true positive samples;
- Recall

  The recall indicates how many of the positive samples in all samples were predicted correctly;
- F1-score

  The F1-score is a comprehensive indicator, which is the harmonic average of precision and recall. Therefore, it is more commonly used in various NER tasks.

If we make the following regulation: TP refers to the number of positive samples predicted as positive samples, TN refers to the number of negative samples predicted as negative samples, FN refers to the number of positive samples predicted as negative samples, and FP predicts the number of actual negative samples as positive samples; then, the calculation methods of the above three metrics can be expressed using the following formulas:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

$$\text{F1score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

4.3. Baselines

In this paper, we compare the proposed method with several representative and advanced baselines. Furthermore, we also compared the existing advanced methods reported on these two

datasets. By comparing with so many advanced methods, we can objectively evaluate the real performance of the proposed method.

4.3.1. ALBERT + BiLSTM + CRF

BiLSTM + CRF is a very classic and effective deep neural network architecture for sequence tagging tasks, that uses a single-layer BiLSTM network. To this day, the model is still widely used in various natural language processing tasks including the NER task. ALBERT + BiLSTM + CRF refers to encoding the input of the model with the output of the pre-trained model ALBERT to enhance the ability of the BiLSTM + CRF model. Like our En2BiLSTM-CRF model, it uses a dense network layer with softmax as the activation function.

4.3.2. ALBERT + CRF

This baseline uses ALBERT to encode the input and is processed by the dense network layer before using CRF to decode the information. This is also a sequence-to-sequence model.

4.3.3. ALBERT + BiLSTM

This model uses ALBERT to encode the input of the model and uses a single-layer BiLSTM network to extract features and uses softmax as the classifier after the dense layer processing to output the most probable results.

4.3.4. BiLSTM-CRF-NER

BiLSTM-CRF-NER [9] is a BiLSTM-CRF model structure [19] that does not use a pre-trained model to encode input sentences. In theory, the method of using the pre-trained model for information encoding is often better than the method without the pre-trained model.

4.3.5. BERT-NER

BERT [9,25] is a pre-trained model based on transformers, which is characterized by the pre-training mechanism and fine-tuning on specific natural language processing tasks to obtain the desired effect.

4.3.6. RoBERTa-NER

RoBERTa - NER [9,43] is also a pre-training model with excellent performance, which has been improved and retrained based on BERT. This model enhances computing power, so the effect may be better.

4.3.7. Human Performance

Human performance refers to the use of manual recognition to compare with the results of machine recognition. The related paper [9] reported the results of manual performance on the CLUENER2020 dataset. In the process of manual evaluation, the provider of this dataset [9] adopts a two-stage method like the machine learning method: the training stage and the testing stage. It is generally believed that human intelligence and cognition are better than computers. However, the results show that artificial performance is not as good as deep neural network models, which is quite different from our expectations. According to [9], the main reason for this situation is that human annotators learn too few cases compared with machines. In addition, it may be difficult for human annotators to become familiar with a large number of entity definitions in a short time.

4.3.8. BiLSTM_Model

The BiLSTM_Model [36,44] refers to a model that uses a BiLSTM neural structure to identify named entities. The baseline can be filled into the input embedding layer using a random initialization method or the encoded output of pre-trained models like BERT and ERNIE, among

others. The performance of this baseline and the BiLSTM_CRF_Model on the People's Daily Ner Corpus has been reported by [44].

4.3.9. BiGRU_Model

Similar to the previous model, this baseline [44] also uses a bidirectional structure named the gated recurrent unit (GRU) [45] network to implement the named entity recognition task. The performance of this baseline on the People's Daily Ner Corpus also has been reported by [44]. It can also use the encoding results of different pre-trained models to fill its character embedding layer.

4.3.10. BiGRU_CRF_Model

The input of this baseline [44] is provided to two GRU networks in opposite directions, and its output is determined by the two unidirectional GRU networks. Moreover, this model combines the advantages of the GRU network and the CRF probability model. The performance of this baseline on the People's Daily Ner Corpus also has been reported by [44].

4.4. Experimental Settings

The software environment we use includes Tensorflow 1.13.1, Keras 2.2.4, and Python 3.5.6. Besides, the graphics card used in our entire experiment was NVIDIA GeForce RTX 2080 Ti, and only one GPU was used. During the model training process, we set the batch size of our En2BiLSTM-CRF to 16 and set Adam (default learning rate: 0.01) as the optimizer. Additionally, we used the CRF [23,24] for loss calculation during model training. To take care of all input samples as much as possible, we set the input length as 128, and the length of the output was set to 22 and 8 on these two datasets, respectively. In addition, the batch size of the input of the model during training was set to 16. These two original datasets used in the experiment can be obtained from the corresponding references.

5. Results and Discussion

We have implemented sufficient and necessary evaluation experiments on the above two representative data sets, and the results of these experiments are valid and convincing. Experimental results using all three metrics on the datasets show that our En2BiLSTM-CRF method performs better than the advanced baseline. Furthermore, the advantages of our proposed method in the task of fine-grained named entity recognition are more obvious, showing a strong fine-grained entity recognition capability. In addition, the experimental results on another dataset with a larger sample size and fewer entity types can also reflect the powerful performance and versatility of our method on coarse-grained tasks. The specific comparative experiment results are shown in Tables 3 and 4. In this section, we will discuss the experimental results from the perspective of overall performance and detail performance.

Table 3. Comparison with different baselines on CLUENER2020.

Method	Precision	Recall	F1-score
ALBERT + BiLSTM + CRF	0.8876	0.8270	0.8555
ALBERT + CRF	0.8094	0.6120	0.6936
ALBERT + BiLSTM	0.7736	0.8132	0.7925
Our En2BiLSTM-CRF	**0.9156**	**0.8337**	**0.8720**
BiLSTM-CRF-NER	0.7106	0.6897	0.7000
BERT-NER	0.7724	0.8046	0.7882
RoBERTa-NER	0.7926	0.8169	0.8042
Human Performance	0.6574	0.6217	0.6341

Table 4. Comparison with different baselines on the People's Daily Ner Corpus.

Method	Precision	Recall	F1-score
ALBERT+BiLSTM+CRF	0.9674	0.9229	0.9446
ALBERT+CRF	0.8315	0.6974	0.7580
ALBERT+BiLSTM	0.9455	0.9096	0.9270
Our En2BiLSTM-CRF	**0.9784**	**0.9306**	**0.9538**
BiLSTM_Model (Random Init)	N/A	N/A	0.74147
BiLSTM_CRF_Model (Random Init)	N/A	N/A	0.81378
BiGRU_Model (Random Init)	N/A	N/A	0.74375
BiGRU_CRF_Model (Random Init)	N/A	N/A	0.82516
BiLSTM_Model (BERT)	N/A	N/A	0.92727
BiLSTM_CRF_Model (BERT)	N/A	N/A	0.94013
BiGRU_Model (BERT)	N/A	N/A	0.92700
BiGRU_CRF_Model (BERT)	N/A	N/A	0.94319
BiLSTM_Model (ERNIE)	N/A	N/A	0.93109
BiLSTM_CRF_Model (ERNIE)	N/A	N/A	0.94460
BiGRU_Model (ERNIE)	N/A	N/A	0.93512
BiGRU_CRF_Model (ERNIE)	N/A	N/A	0.94218

5.1. Overall Experimental Results and Discussion

The experimental results are sufficient to show that the proposed methods surpass the current advanced and representative baselines in both fine-grained entity recognition tasks and coarse-grained entity recognition tasks. Our En2BiLSTM-CRF model can recognize at least ten Chinese named entity types, showing that the proposed method has more obvious advantages in fine-grained Chinese entity recognition tasks. In particular, the results on two different types of real-world datasets can show that our method has strong encoding–decoding capabilities, deep-level feature extraction capabilities, and a wide range of generality.

On the CLUENER2020 dataset, the proposed model outperforms the ALBERT + BiLSTM + CRF model using a single-layer BiLSTM network in all three indicators and far exceeded other baselines. It can be seen from Table 3 that when using ALBERT as a pre-training model to encode the input information, the effect is also far better than when this encoding is not used. Moreover, although both RoBERTa-NER and BERT-NER are advanced models, the effect and performance of this fine-grained dataset are not as good as our En2BiLSTM-CRF model.

On the People's Daily Ner Corpus, the proposed method also outperforms all the baselines, although this advantage is not very obvious. We think the main reason is that it is a coarse-grained dataset that is less challenging than a fine-grained dataset, so most of the baselines also perform well. The results on this dataset also show that the use of ALBERT for information encoding is better than BERT and is evenly matched with ERNIE. Since some baselines have only been reported with F1-scores on this dataset, we use N/A to mark their precision and recall rates. However, this will not affect the experimental conclusions. In fact, researchers are accustomed to using only F1-score to evaluate the overall performance of the NER models because the F1-score already includes the precision and the recall, so it is more convincing and comprehensive.

Considering that the effect of using the pre-trained model alone based on experience and existing knowledge is probably not as good as the combined use of the pre-trained model and other methods, we currently did not compare BERT and RoBERTa on the second dataset. We will supplement these experiments to understand the effect of these two baselines on this dataset in future research.

5.2. Detailed Results and Discussion

We also implemented statistics and a display of the test results of each entity on each data set. These statistical recognition results are shown in Figures 5 and 6. The refined prediction results can

help us analyze the detailed performance of the proposed model when identifying entity types with different data distributions.

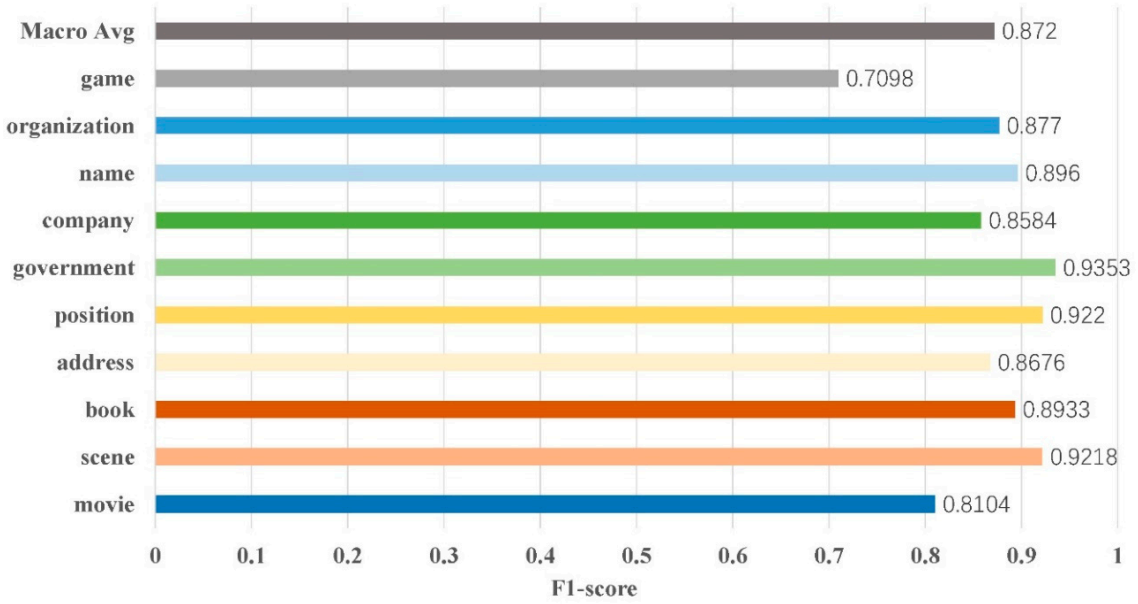

Figure 5. Statistical recognition results on CLUENER2020.

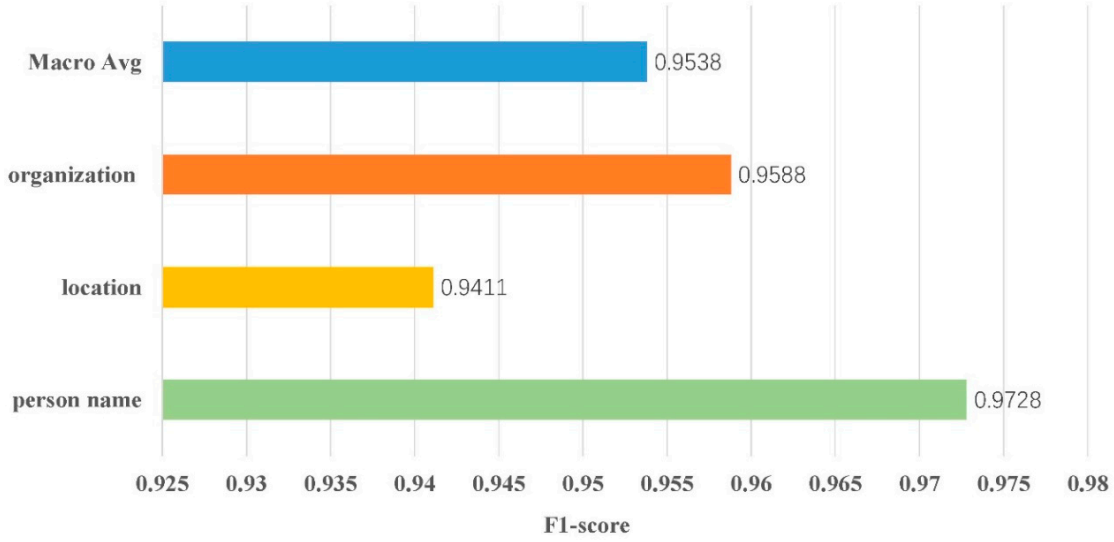

Figure 6. Statistical recognition results on the People's Daily Ner Corpus.

Involving all ten named entities on CLUENER2020, our method is very sensitive in identifying "scene", "government", and "position". However, the performance in identifying "game" is not as sensitive as other entities. We suspect that the name of many games is often random and irregular, so it is more challenging to recognize. The test results on the People's Daily Ner Corpus show that the recognition of "person" is relatively more sensitive than the recognition of "location" and "organization". This may be related to the distribution and number of samples for each entity type.

6. Conclusions

In this paper, an innovative deep neural network model is proposed for Chinese named entity recognition tasks. This model named En2BiLSTM-CRF provides a new idea for researchers to solve

the problem of fine-grained entity recognition. Besides, this model combines the advantages of some popular methods and advanced mechanisms. It can encode the information to be processed multiple times and extract features hierarchically. Further, we use three important indicators to compare with some of the most advanced baselines on two representative datasets, proving the remarkable effect of the proposed method on the fine-grained Chinese named entity recognition task. In the future, we will continue to study the effect of the proposed method on other sequence labeling tasks, and study its mechanism from the perspective of interpretability.

Author Contributions: J.L. proposed the idea, conducted the experiments, and wrote the manuscript. C.X. and H.Y. supervised the entire research. W.X. provided technical support and revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the State Key Laboratory of Software Development Environment under Grant SKLSDE-2019ZX-22, and in part by the National Natural Science Foundation of China under Grant U1636208 and Grant 61862008.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sang, E.F.; De Meulder, F. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *arXiv* **2003**, arXiv:0306050.
2. Nadeau, D.; Sekine, S. A survey of named entity recognition and classification. *Lingvisticae Investig.* **2007**, *30*, 3–26.
3. Li, J.; Sun, A.; Han, J.; Li, C. A survey on deep learning for named entity recognition. *IEEE Trans. Knowl. Data Eng.* **2020**, doi:10.1109/TKDE.2020.2981314.
4. Ritter, A.; Clark, S.; Etzioni, O. Named entity recognition in tweets: an experimental study. In Proceedings of the conference on empirical methods in natural language processing, Edinburgh, UK, 27–31 July 2011; Volume 7, pp. 1524–1534.
5. Peng, N.; Dredze, M. 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, 17–21 September 2015; Volume 9, pp. 548–554.
6. Settles, B. Biomedical named entity recognition using conditional random fields and rich feature sets. In Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP), Geneva, Switzerland, 28–29 August 2004; pp. 107–110.
7. Habibi, M.; Weber, L.; Neves, M.; Wiegandt, D.L.; Leser, U. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics* **2017**, *33*, i37–i48.
8. Lee, C.; Hwang, Y.G.; Oh, H.J.; Lim, S.; Heo, J.; Lee, C. H.; Kim, H.G.; Wang, J.H.; Jang, M.G. Fine-grained named entity recognition using conditional random fields for question answering. In Proceedings of the Asia Information Retrieval Symposium. Singapore, 16–18 October 2006; Springer: Berlin, Heidelberg,Germany, 2006; Volume 10, pp. 581–587.
9. Xu, L.; Dong, Q.; Yu, C.; Tian, Y.; Liu, W.; Li, L.; Zhang, X. CLUENER2020: Fine-grained Name Entity Recognition for Chinese. *arXiv* **2020**, arXiv:2001.04351.
10. Gao, J.; Li, M.; Huang, C.N.; Wu, A. Chinese word segmentation and named entity recognition: A pragmatic approach. *Computational Linguist.* **2005**, 31, 531–574.
11. Zhenggao, P. Research on the recognition of Chinese named entity based on rules and statistics. *Information Sci.* **2012**, *30*, 708–712.
12. Zhou, G.; Su, J. Named entity recognition using an HMM-based chunk tagger. In proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, USA, 6–12 July 2002; Volume 7, pp. 473–480.
13. Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; Dyer, C. Neural architectures for named entity recognition. *arXiv* **2016**, arXiv:1603.01360.
14. Sun, Y.; Li, L.; Xie, Z.; Xie, Q.; Li, X.; Xu, G. Co-training an improved recurrent neural network with probability statistic models for named entity recognition. In Proceedings of the International Conference on Database Systems for Advanced Applications. Suzhou, China, 27–30 March 2017; Volume 3, pp. 545–555.

15. Chiu, J.P.; Nichols, E. Named entity recognition with bidirectional LSTM-CNNs. *Transactions Assoc. Comput. Linguist.* **2016**, *4*, 357–370.
16. Strubell, E.; Verga, P.; Belanger, D.; McCallum, A. Fast and accurate entity recognition with iterated dilated convolutions. *arXiv* **2017,** arXiv:1702.02098.
17. Hammerton, J. Named entity recognition with long short-term memory. In Proceedings of the seventh conference on Natural language learning at HLT-NAACL. Edmonton, Canada, 31 May–1 June 2003; Volume 4, pp. 172–175.
18. Zhu, Q.; Li, X.; Conesa, A.; Pereira, C. GRAM-CNN: a deep learning approach with local context for named entity recognition in biomedical text. *Bioinformatics* **2018**, *34*, 1547–1554.
19. Huang, Z.; Xu, W.; Yu, K. Bidirectional LSTM-CRF models for sequence tagging. *arXiv* **2015**, arXiv:1508.01991.
20. Liu, Y.; Sun, C.; Lin, L.; Wang, X. Learning natural language inference using bidirectional LSTM model and inner-attention. *arXiv* **2016**, arXiv:1605.09090.
21. Na, S. H.; Kim, H.; Min, J.; Kim, K. Improving LSTM CRFs using character-based compositions for Korean named entity recognition. *Computer Speech Lang.* **2019**, *54*, 106–121.
22. Dong, C.; Zhang, J.; Zong, C.; Hattori, M.; Di, H. Character-based LSTM-CRF with radical-level features for Chinese named entity recognition. In *Natural Language Understanding and Intelligent Applications*, Kunming, China, 2–6 December 2016; Springer: Cham, Switzerland, 2016; pp. 239–250.
23. McCallum, A.; Li, W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In Proceedings of the seventh conference on Natural language learning at HLT-NAACL. Edmonton, Canada, 31 May–1 June 2003; Volume 4, pp. 188–191.
24. Finkel, J.R.; Kleeman, A.; Manning, C.D. Efficient, feature-based, conditional random field parsing. In Proceedings of ACL-08: HLT, Columbus, OH, USA, 15–20 June 2008; Volume 6, pp. 959–967.
25. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
26. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
27. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. Albert: A lite bert for self-supervised learning of language representations. *arXiv* **2019**, arXiv:1909.11942.
28. Gold, S.; Rangarajan, A. Softmax to softassign: Neural network algorithms for combinatorial optimization. *Journal Artif. Neural Netw.* **1996**, *2*, 381–399.
29. Liu, L.; Ren, X.; Shang, J.; Peng, J.; Han, J. Efficient contextualized representation: Language model pruning for sequence labeling. *arXiv* **2018**, arXiv:1804.07827.
30. Jia, C.; Liang, X.; Zhang, Y. Cross-Domain NER using Cross-Domain Language Modeling. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August, 2019; Volume 7, pp. 2464–2474.
31. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, L. Attention is all you need. In Proceedings of the advances in neural information processing systems, Long Beach, CA, USA, 4-9 December 2017; pp. 5998–6008.
32. Kitaev, N.; Klein, D. Constituency parsing with a self-attentive encoder. *arXiv* **2018**, arXiv:1805.01052.
33. Cui, L.; Zhang, Y. Hierarchically-Refined Label Attention Network for Sequence Labeling. *arXiv* **2019**, arXiv:1908.08676.
34. Connor, J.T.; Martin, R.D.; Atlas, L.E. Recurrent neural networks and robust time series prediction. *IEEE Trans. neural Netw.* **1994**, *5*, 240–254.
35. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780.
36. Schuster, M.; Paliwal, K.K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681.
37. Pan, S. J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. data Eng.* **2009**, *22*, 1345–1359.
38. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
39. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. Xlnet: Generalized autoregressive pretraining for language understanding. In Proceedings of the advances in neural information processing systems, Vancouver, Canada, 8–14 December 2019; pp. 5754–5764.

40. Zhang, Z.; Han, X.; Liu, Z.; Jiang, X.; Sun, M.; Liu, Q. ERNIE: Enhanced language representation with informative entities. *arXiv* **2019**, arXiv:1905.07129.
41. albert_zh (albert_tiny_zh). Available online: https://github.com/brightmart/albert_zh (accessed on 30 April 2020).
42. Kashgari (v1.1.5). Available online: https://kashgari.readthedocs.io/en/v1.1.5/index.html (accessed on 30 April 2020).
43. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O,; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. arXiv **2019**, arXiv:1907.11692.
44. Performance. report. Available online: https://kashgari.readthedocs.io/en/v1.1.5/tutorial/text-labeling.html#performance-report (accessed on 30 April 2020).
45. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv* **2014**, arXiv:1412.3555.