applied sciences

MDPI

*Article*

# Chinese Named Entity Recognition of Geological News Based on BERT Model

**Chao Huang, Yuzhu Wang *, Yuqing Yu, Yujia Hao, Yuebin Liu and Xiujian Zhao**

School of Information Engineering, China University of Geosciences, Beijing 100083, China; huangchao@cugb.edu.cn (C.H.); yuyq@cugb.edu.cn (Y.Y.); haoyj@cugb.edu.cn (Y.H.); liuyb@cugb.edu.cn (Y.L.); zhaoxj@cugb.edu.cn (X.Z.)

* Correspondence: wangyz@cugb.edu.cn

**Abstract:** With the ongoing progress of geological survey work and the continuous accumulation of geological data, extracting accurate information from massive geological data has become increasingly difficult. To fully mine and utilize geological data, this study proposes a geological news named entity recognition (GNNER) method based on the bidirectional encoder representations from transformers (BERT) pre-trained language model. This solves the problems of traditional word vectors that are difficult to represent context semantics and the single extraction effect and can also help construct the knowledge graphs of geological news. First, the method uses the BERT pre-training model to embed words in the geological news text, and the dynamically obtained word vector is used as the model's input. Second, the word vector is sent to a bidirectional long short-term memory model for further training to obtain contextual features. Finally, the corresponding six entity types are extracted using conditional random field sequence decoding. Through experiments on the constructed Chinese geological news dataset, the average F1 score identified by the model is 0.839. The experimental results show that the model can better identify news entities in geological news.

**Keywords:** BERT; named entity recognition; geological news; CRF



**Citation:** Huang, C.; Wang, Y.; Yu, Y.; Hao, Y.; Liu, Y.; Zhao, X. Chinese Named Entity Recognition of Geological News Based on BERT Model. *Appl. Sci.* **2022**, *12*, 7708. https://doi.org/10.3390/app12157708

Academic Editor: Douglas O'Shaughnessy

Received: 8 July 2022
Accepted: 28 July 2022
Published: 31 July 2022



## 1. Introduction

With the rapid development of society, artificial intelligence has been applied in all aspects of our lives [1]. In recent years, artificial intelligence has developed rapidly in various fields. Taking natural language processing (NLP) as an example, it is applied in many fields, such as information extraction, question answering systems, machine translation, and text classification [2]. Named entity recognition (NER) is an important field in NLP technology, and many studies on NLP are based on it. Currently, research on NER is mainly concentrated in the fields of finance and medical care and has not yet been seen in the field of geological news.

In early research on NER, rule-based and dictionary-based methods were mainly used [3–5]. Later, with the development of machine learning, machine learning methods were used to solve NER tasks. Common methods include hidden Markov models [6,7], maximum entropy models [8,9] and conditional random field (CRF) models [10], among others. In recent years, with the development of computer technology and deep learning, the method of using deep learning models has become a trend for solving NLP problems. In the field of geology, some scholars have used deep learning-based models for NER tasks and have achieved good results. Zhang et al. [11] proposed a geological entity recognition model based on a deep belief network, which achieved good entity recognition results on a small-scale corpus, and each evaluation index (P, R, F1) reached 90% and above. Liu et al. [12] proposed an improved lattice long- and short-term memory (LSTM) model based on a bidirectional long short-term memory conditional random field (BiLSTM-CRF). The proposed model is based on LSTM [13] and achieved good results for named entities

in the coal mining field, with an F1 score of 94.04% and an improvement of 2.1% based on the original BiLSTM-CRF.

There is a huge amount of geological news texts that contain a large amount of information. Accurately identifying effective information from them can provide important data support for related geological survey work. However, traditional manual extraction methods have problems such as high time consumption and low accuracy. As the scale of geological news text data increases, the extraction becomes more and more difficult. Therefore, it is important to realize the automatic extraction of geological news information entities, which is also the basic work of geological news knowledge graph construction.

Geological news text data are complex and contain many types of data. Related entities include time, geographic location, organization, job title, event, etc. Geological news texts are different from common news texts. Because they are news related to geology, the names of related entities have obvious characteristics, such as a professional background and application behavior—for example, an organization entity (China Natural Resources Airborne Geophysical and Remote Sensing Center). In addition, the text also has polysemy and entity nesting problems, such as China referring to either a geographical location or a country. The China Geological Survey of the Ministry of Natural Resources contains an organizational entity (Ministry of Natural Resources), a geographical location entity (China), and an organizational entity (Geological Survey). At present, there is no public dataset in the field of geological journalism. Therefore, it is challenging to construct a corpus of geological journalism before carrying out the NER task.

To accurately extract entities from geological news, this study proposes a model that combines a bidirectional encoder representations from transformers (BERT) pre-trained model and a BiLSTM-CRF model for geological news named entity recognition (GNNER). The model first uses the pre-trained word vector model BERT for semantic extraction. Compared to the traditional word vector, the GNNER Word2vec, BERT [14] can better represent semantic information in different contexts to solve the polysemy problem. After obtaining the output of the BERT model, part-of-speech analysis and chunking analysis features are added to help the model identify entity boundaries. Finally, the word vector is sent to the BiLSTM model for further training. The results of the BiLSTM model are modified using CRF, outputting the labeled sequence with the highest score. Based on the geological news texts of the China Geological Survey and according to the characteristics of geological news texts, time, name, geographic location, organization, and other information are extracted. The experimental results show that the model could better identify the entities in geological news.

The main contributions of this paper are summarized as follows:

(1) GNNER was based on the BERT model, integrating a variety of different models and extracting various types of entities from the constructed geological news corpus.
(2) This research used crawler technology to obtain geological news texts from the China Geological Survey Bureau, preprocessed the data, including long text segmentation, data cleaning, and removal of uncommon punctuation marks, and used the "BIO" named entity labeling method to label the texts to create a dataset of a certain scale in the field of geological news.
(3) The BERT-BiLSTM-CRF model was used to conduct a comparative experiment with the other five models on the geological news dataset, analyze the quality of the six models, and discuss the effects of geological news entity type, number of labels, and model hyperparameters on model evaluation.

The rest of the paper is structured as follows: Section 2 introduces the related model design methods and the dataset construction process; Section 3 presents the experimental results; and Section 4 discusses the experimental results and directions for future research.

## 2. Materials and Methods

### *2.1. Word2vec*

Before using the deep learning model to solve the NLP problem, we needed to convert the language data type into a data type that the neural network could handle. Word embedding technology was developed because of this requirement.

The Word2vec model was developed by Tomas Mikolov [15] et al. in 2013. It is an efficient model for training word vectors. Unlike the traditional language model, Word2vec assumes that there is a relationship between similar words in a sentence. It has two models: the skip-gram and CBOW. The model structure of Word2vec is shown in Figure 1. The skip-gram uses the current word to predict nearby words, whereas CBOW uses nearby words to predict the current word.

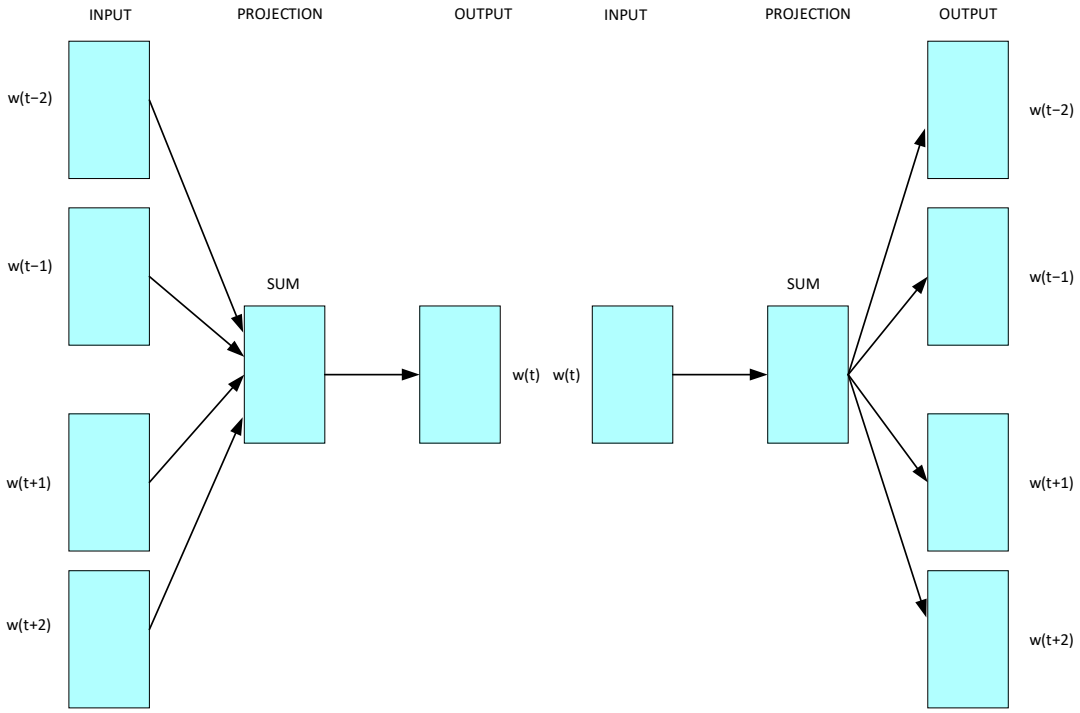


**Figure 1.** CBOW model (**left**) and skip-gram model (**right**).

Although the Word2vec model has achieved good results in word embedding, it also has some problems. For example, the word vectors trained using this method are fixed and cannot change the meaning of words in different contexts.

### *2.2. BiLSTM-CRF*

Proposed by Lample [16] et al. and based on the LSTM-CRF model, the BiLSTM-CRF model is a deep learning model that integrates feature engineering and serialization. The model structure is shown in Figure 2. It is mainly divided into a three-layer structure of a word vector input layer, a BiLSTM layer, and a CRF layer. The experimental process can be divided into three steps. First, the input of the model is a sequence of word vectors. Second, the probability vector of the corresponding label of each word is output on the BiLSTM layer. Finally, the result is corrected through the CRF, and the label sequence with the highest probability is output. These three parts are explained in detail below.

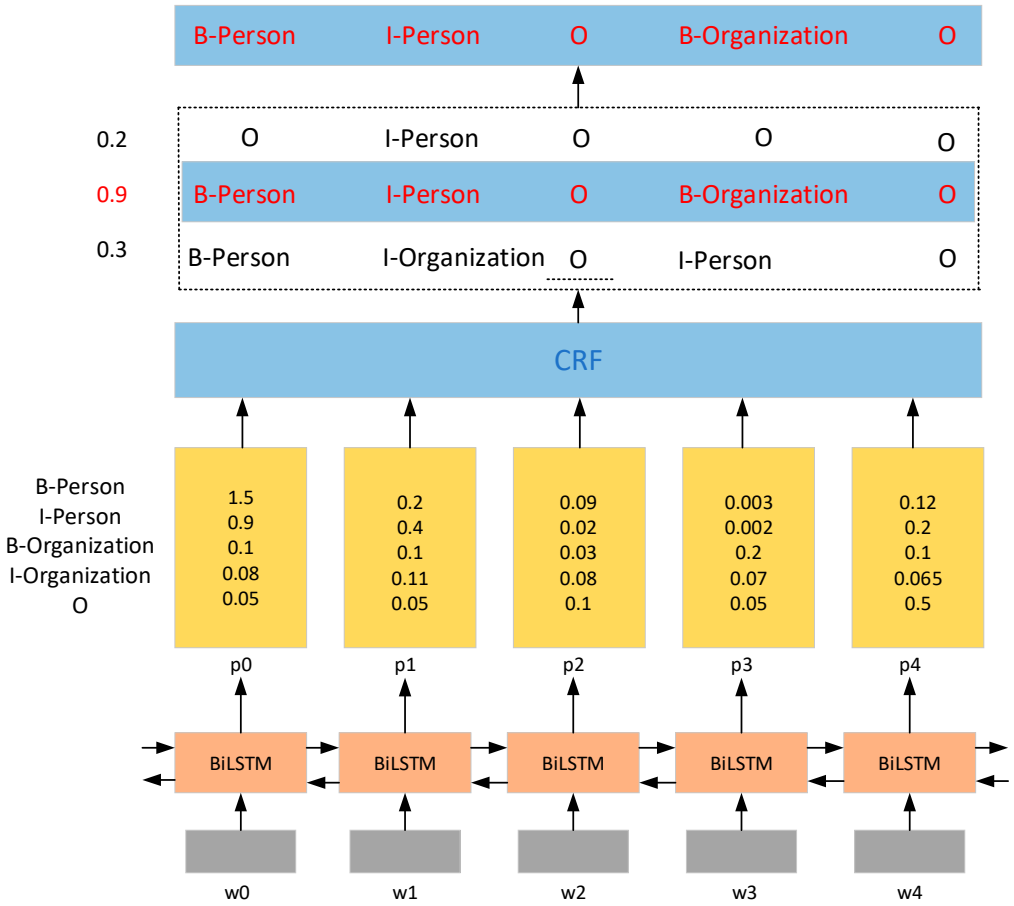


**Figure 2.** The structure of the BiLSTM-CRF model.

The first part is the word vector input layer, where the sentence is input. Assuming that the sentence is $X$, $X$ is composed of $n$ words $(X_1, X_2, \ldots, X_n)$, denoted as $X = (X_1, X_2, \ldots, X_n)$. Each word in the sentence is mapped into a word vector using the word embedding method, corresponding to $w_0, w_1, w_2, w_3, w_4$, etc. In the figure, the matrix formed by these word vectors is used as the input of the next layer.

The second part is the BiLSTM layer, which has two hidden layers: a forward LSTM layer and a backward LSTM layer. The input word vector goes through the two hidden layers, and the results of these hidden layers are spliced together to determine the state of the final hidden layer. Based on the state of the final hidden layer, a linear layer is connected to perform a mapping operation to map the hidden layer matrix from $n$ dimensions to K dimensions, where $n$ is the dimension of the matrix and K is the number of labels. Finally, the BiLSTM layer will output the resulting matrix, which is denoted as $P = (p_0, p_1, \ldots, p_n)$, where $p_i$ is the vector composed of the emission scores of $w_i$ corresponding to each label, e.g., $p_0 = [1.5, 0.9, 0.1, 0.08, 0.05]$. The matrix P serves as the input matrix of the CRF layer.

The last part is the CRF layer, which adds constraints between the labels and reduces the number of invalid predicted labels. Although the Softmax function outputs the label with the maximum probability corresponding to the word, the output labels are independent of each other. This means that the sequence is prone to unreasonable situations, resulting in a decrease in the accuracy rate, such that the time and organization entity of the "I" label may become adjacent. However, we know that according to the "BIO" labeling method, the entity tag must begin with "B". Thus, the "I" label of the time and organization entity cannot be adjacent.

After the word vector passes through the BiLSTM and CRF layers, the score of the final sequence consists of two parts: the emission score of the BiLSTM layer and the transfer score of the CRF layer, as shown in Equation (1). The $X$ on the left side of the equal sign

represents the scoring sequence, and the first part on the right side of the equal sign $P_{i,y_i}$ represents the emission score of the $y_i$ label in the $i$th word vector. The second part $A_{y_{i-1},y_i}$ represents the emission score from $y_{i-1}$ label to the $y_i$ label.

$$socre(X,y) = \sum_{i=1}^{n} P_{i,y_i} + \sum_{i=1}^{n+1} A_{y_{i-1},y_i} \tag{1}$$

After calculating the score for each possible sequence, it is normalized using Softmax. The result is shown in Equation (2), which $Y_{(x)}$ represents all possible labeled sequences.

$$p(y|X) = \frac{e^{socre(X,y)}}{\sum \widetilde{y} \in Y_{(x)} e^{socre(X,\widetilde{y})}} \tag{2}$$

In the model training process, the log-likelihood function is used to optimize the model. The result is shown in Equation (3).

$$\log(p(y|X)) = \log(\frac{e^{socre(X,y)}}{\sum \widetilde{y} \in Y_{(x)} e^{socre(X,\widetilde{y})}}) = score(X,y) - \log \sum_{\widetilde{y} \in Y(x)} e^{socre(X,\widetilde{y})} \tag{3}$$

Finally, the Viterbi algorithm [17] is used to decode the hidden state sequence to obtain the optimal label sequence. The result is shown in Equation (4).

$$y^* = \arg max_{\widetilde{y} \in Y_{(x)}} socre(X,\widetilde{y}) \tag{4}$$

### *2.3. BERT*

In Section 2.1, we discussed the need to convert the data type into a data type that the neural network can handle when dealing with NLP problems; thus, word embedding technology is needed. However, traditional word embedding technology has some problems, such as its inability to solve polysemy and dynamically optimize specific tasks. To solve these problems, Jacob Devlin et al. proposed a new pre-training model called the BERT model in 2018. BERT is a deep bidirectional language representation model pre-trained on a corpus consisting of a large number of books and Wikipedia, and its main structure is the encoder part of the Transformer model [18].

The input of the BERT model is a token sequence, which is inserted [CLS] at the beginning of each sequence to classify sentences and [SEP] at the end of the sequence to separate different sentences. Each token sequence consists of three parts: token embeddings, segment embeddings, and position embeddings.

The Transformer model was developed by Google's Vaswani et al. in 2017. This model efficiently realizes the parallelization of non-serialized models, which can greatly improve computational efficiency. Figure 3 shows a structural diagram of the Transformer model. As the structure of the BERT model is mainly the encoder part, the composition and principles of the encoder part are explained as follows.

(1) Input of Transformer

The Transformer model trains all words in the sequence at the same time. To identify the position information of each word in the sequence, it is necessary to add a position encoding (PositionEncoding) to each word vector (EmbeddingLookup ($X$)), as shown in Equation (5).

$$X = \text{EmbeddingLookup}(X) + \text{PositionalEncoding} \tag{5}$$

(2) Self-attention mechanism

Unlike the attention mechanism [19], the self-attention mechanism calculates the relationship between the elements in the input or output sequence, which is an improved method based on the attention mechanism. In the calculation process of the self-attention mechanism, the matrix Query, Key, and Value need to be used. From the perspective

of the information retrieval system, Query is the input information, Key is the content information matching Query, and Value is the information itself. The calculation process is then described in detail. Query, Key, and Value are denoted as $Q$, $K$, and $V$, respectively.

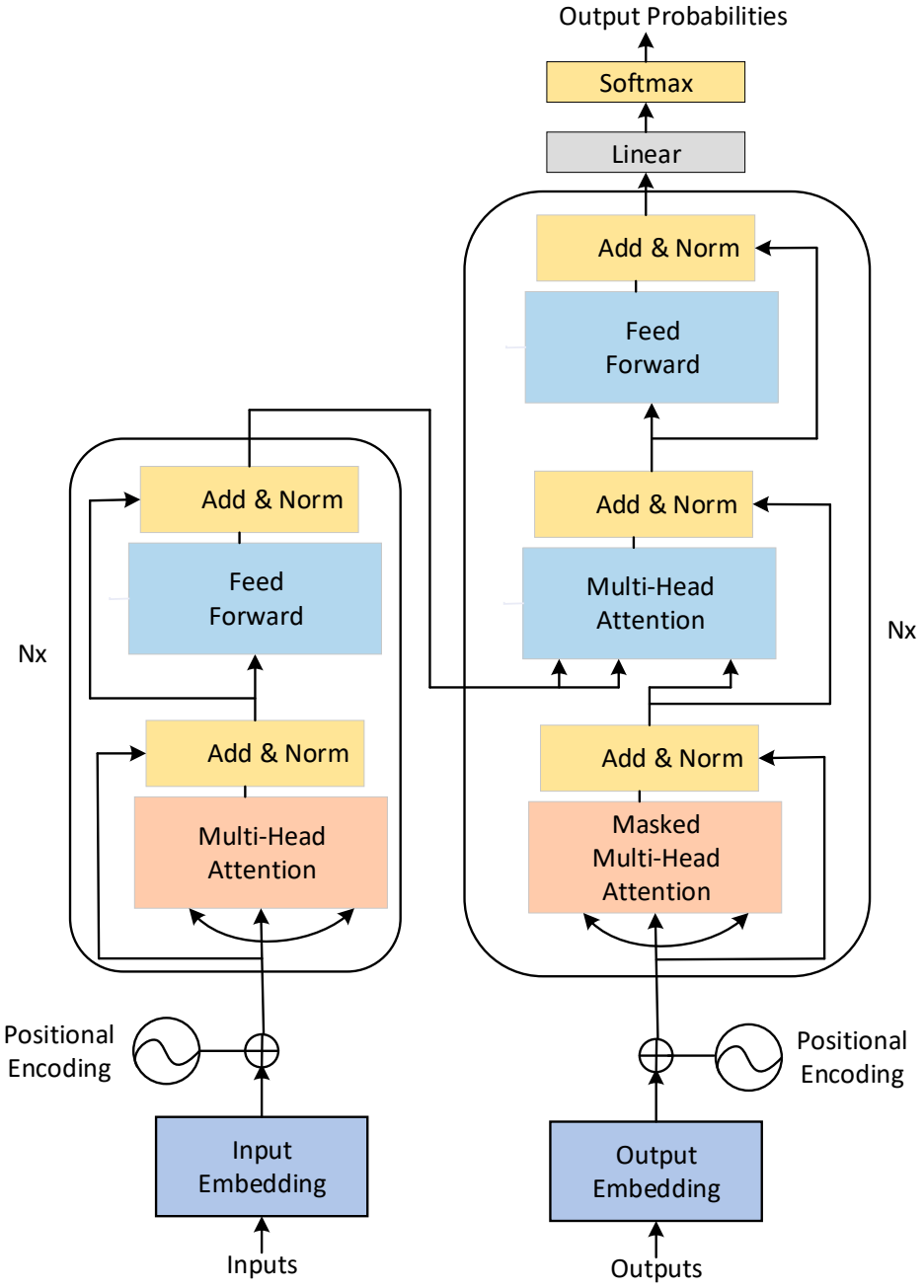


**Figure 3.** The structure of Transformer [18].

The input of the self-attention mechanism is the matrix $X$. $Q$, $K$, and $V$ are obtained by the linear transformation of $X$, as shown in Equations (6)–(8), where $W_Q, W_k, W_v$ are the three auxiliary matrices. The word vector matrix $X$ is multiplied by these auxiliary matrices to obtain the corresponding $Q$, $K$, and $V$ values for each item in the sequence. The $Q$ of the current item is multiplied by the $K$ of each item in the sequence to determine the relationship between the two. After scaling and normalizing the product using Softmax, it is multiplied by $V$, and each $V$ is added to obtain the feature representation of the current item. In Equation (9), $d_k$ is the dimension of the $Q$ and $K$ vectors.

$$Q = Linear(X) = XW_Q \tag{6}$$

$$K = Linear(X) = XW_k \tag{7}$$

$$V = Linear(X) = XW_v \tag{8}$$

$$X_{attention} = SelfAtteention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{9}$$

(3) Multi-head mechanism

In the self-attention mechanism, each item in the input sequence corresponds to a set of feature expressions: Query, Key, and Value. Conversely, the so-called multi-head mechanism establishes multiple sets of auxiliary matrices in the Transformer model and multiplies them by the word vector input matrix $X$ to obtain multiple sets of the Query, Key, and Value values. Therefore, each item in the sequence has multiple sets of feature expressions. Multiple sets of feature expressions are spliced together, and dimensionality reduction is performed using a fully connected layer.

(4) Summation and normalization

A residual connection is also required to ensure a better feature extraction effect. The residual connection adds the vector after the self-attention mechanism and the multi-head mechanism to the original input vector, as shown in Equations (10) and (11). It is necessary to normalize the hidden layer to speed up convergence.

$$X_{attention} = X + X_{attention} \tag{10}$$

$$X_{attention} = LayerNorm(X_{attention}) \tag{11}$$

After extensive training, the BERT model can be applied to various natural language processing tasks. This paper uses the BERT model instead of Word2vec to obtain word vectors that can better integrate context information and improve the accuracy of named entity recognition.

In addition, this paper also uses two improved models based on BERT for experiments, namely distilled BERT (DistilBERT) and robustly optimized BERT approach (RoBERTa). DistilBERT is a distilled version of BERT proposed by Victor Sanh [20] et al., which is smaller, faster and cheaper than the BERT model. RoBERTa is a robustly optimized BERT pre-training approach proposed by Liu [21] et al. By improving BERT, these two models enable BERT to achieve high performance on large datasets.

### *2.4. Model Design*

In the NLP task, as the BERT model has achieved good results, more and more people have begun to combine BERT with deep learning models for NER tasks. In this study, we introduce the BERT model based on the BiLSTM-CRF model and design the BERT-BiLSTM-CRF model, which is used to identify geological news entities. As shown in Figure 4, the structure of the model is mainly divided into three layers from bottom to top: the BERT layer, BiLSTM layer, and CRF layer. First, the input of the BERT model is the superposition of each word vector, including the sentence vector and the position vector. The word vector can obtain the text context features after the encoding layer of the Transformer. Therefore, the word vector output after BERT training can also be remarkable and effectively integrate the article features. Second, the BERT output result is used as the input of the BiLSTM layer, and the context information can be better integrated using the two-layer LSTM neural network before and after. Finally, the labeling sequence output by the BiLSTM layer goes through the CRF layer, and the labeling sequence is corrected by the state transition matrix. The optimal labeling sequence is finally output.

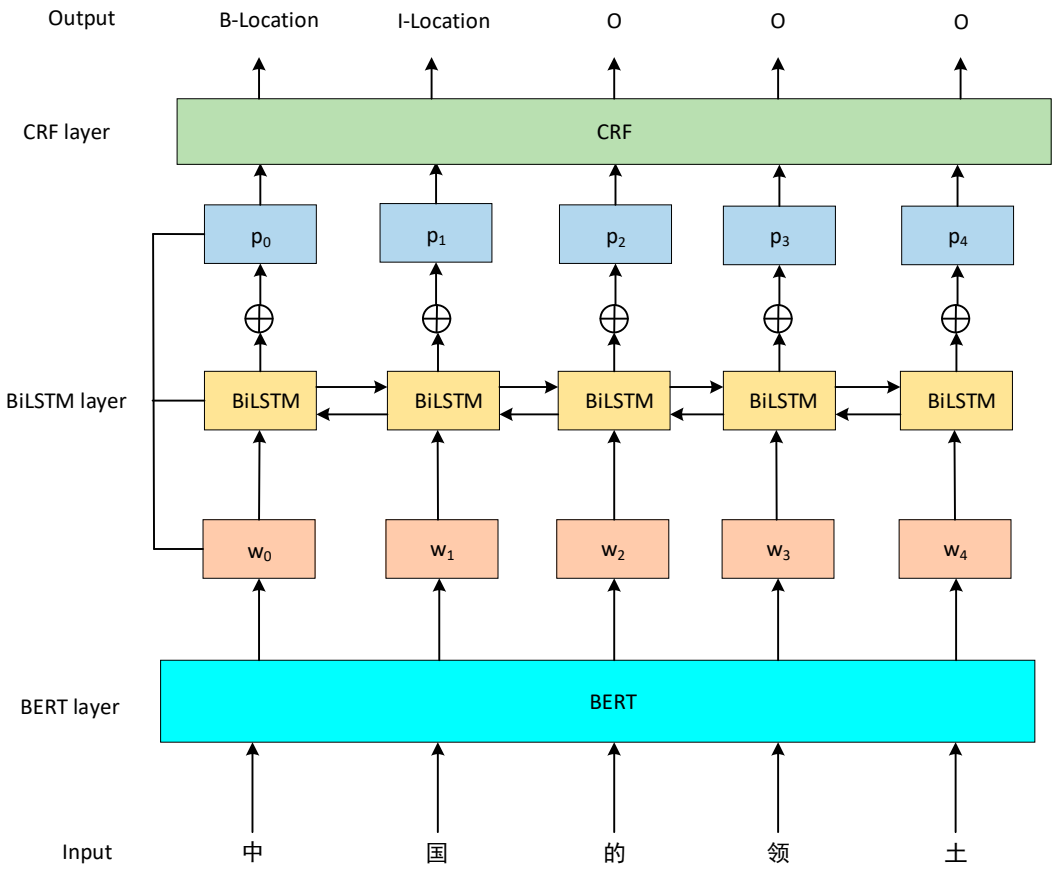


**Figure 4.** The structure of the BERT-BiLSTM-CRF model. The Chinese characters in the figure show the first to fifth characters of the territory of China expressed in Chinese.

### *2.5. Data Processing and Experimental Setup*

#### 2.5.1. Data Source and Pre-Processing

The original data of this paper are a total of 2200 geological news texts obtained from the China Geological Survey through web crawlers with about one million words. As the data have much noise, it was necessary to preprocess and clean the data to reduce the effect on the experimental accuracy. First, special symbols, such as some non-Chinese, non-English, and non-digital symbols, were removed from the text. Second, the text contains elements that have nothing to do with geological news. This kind of content has obvious signs before and after the text, so that it can be used mainly for filtering using a regular expression. News texts that were too long needed to be segmented. In this study, news texts were segmented according to the priority of punctuation. The preprocessed dataset consists of Chinese and English characters, numbers, punctuation marks and spaces, and the length of each sentence after segmentation does not exceed the max_len set in the experimental parameters.

#### 2.5.2. Text Corpus Annotation

Sequence labeling is the most important step in the construction of datasets. There are many popular labeling methods, and the "BIO" labeling method is used in this paper. In the "BIO" notation method, a paragraph is usually marked as "B-X," "I-X," and "O;". "B" means "begin," which in Chinese refers to the Chinese character at the beginning of a named entity; "I" means "Inside," which refers to the middle and end parts of the named entity; and "O" means "Other," which refers to the non-named entity part.

The task of labeling named entities is cumbersome. To reduce the intensity of the task, this study uses YEDDA [22] as an auxiliary tool for labeling corpora. After the text is

marked, according to statistics, 21,323 entities were marked, including 6 entity types: name, time, geographic location, job title, organization, and event. Through the statistics of the number of various entities in the labeled corpus, the specific labeling situation is shown in Table 1. In the labeling process, we labelled complex entities with multiple entities one by one. For example, the entities "Ministry of Natural Resources China Geological Survey", "Ministry of Natural Resources", "China" and "Geological Survey" were annotated as an organization, geographic location, and organization, respectively.

**Table 1.** The number of various entities in the annotated corpus.

| Tags | Type | Number |
|---|---|---|
| TIM | Time | 3492 |
| ORG | Organization | 5630 |
| POS | Job title | 1050 |
| EVE | Event | 2032 |
| LOC | Geographic location | 7702 |
| PER | Name | 1417 |

#### 2.5.3. Experimental Environment and Parameters

The BERT model requires strong computing power during the training process, so it has certain requirements for computer hardware. Table 2 presents the relevant information on the hardware and software used in this experiment.

**Table 2.** Experimental environment. CUDA is a parallel computing platform and programming model invented by NVIDIA, headquartered in Santa Clara, California, USA. Python is a programming language designed by Guido van Rossum in the Netherlands. Tensorflow is a symbolic math system developed by Google's artificial intelligence team, Google Brain. Pytorch is an open source Python machine learning library launched by the Facebook Artificial Intelligence Research Institute. Numpy is an extension tool for numerical computing developed by Jim Hugunin and other collaborators.

| Category | Configuration |
|---|---|
| Hardware | GPU: 4*NVIDIA Tesla K80<br>OS: CentOS 8.3<br>Video memory: 11 GB GDDR6 |
| Software | CUDA: 11.4<br>Python: 3.6<br>Tensorflow: 1.14.0<br>Pytorch: 1.4.0<br>Numpy: 1.19.2 |

In the training process of the model, the setting of hyperparameters influences the training effect. To exclude the effects of different hyperparameters on the experiment, fixed hyperparameters were used to train different models. Table 3 shows several important parameters used in the model training process. Among these parameters, an epoch is a process of training the training set once, max_len is the length of the maximum sequence, batch_size is the amount of data obtained in one training process, learning_rate is the learning rate, and drop_rate is set to prevent overfitting of the neural network.

**Table 3.** Model parameters.

| Hyper-Parameter | Parameter Values |
|---|---|
| Epochs | 8 |
| max_len | 128 |
| batch_size | 16 |
| learning_rate | $3 \times 10^{-5}$ |
| drop_rate | 0.5 |

## 3. Results

### 3.1. Experimental Evaluation Indicators

In the process of NER, evaluation indicators are needed to evaluate the quality of the model. In this study, the three evaluation indicators used in all experiments are precision rate (P), recall rate (R), and F1 score (F1).

### 3.2. Comparison of Different Models

We divided the labeled dataset into the training set, validation set, and test set according to a ratio of 8:1:1, with 17,124, 2056, and 2143 entities, respectively. In the NER task of the geological news dataset, we trained the dataset on six different models and tested it on the test set. The experimental results show that the model has achieved good results in GNNER. Table 4 presents relevant information on the precision rate, recall rate, and F1 scores of the six models for six categories of entities: name, time, geographic location, job title, organization, and event.

**Table 4.** P, R, and F1 scores of six types of entities on six models. The numbers in bold font are the three indicators (P, R, F1) with the highest corresponding scores in the experiment.

| Model | Eval | TIM | ORG | POS | EVE | LOC | PER | Avg |
|---|---|---|---|---|---|---|---|---|
| BERT | P | 0.863 | 0.820 | 0.691 | 0.820 | 0.802 | 0.887 | 0.819 |
| | R | 0.827 | 0.844 | 0.827 | 0.796 | 0.835 | 0.806 | 0.829 |
| | F | 0.845 | 0.832 | 0.753 | 0.808 | 0.818 | 0.844 | 0.824 |
| DistilBERT | P | 0.847 | 0.796 | 0.721 | 0.793 | 0.787 | 0.864 | 0.808 |
| | R | 0.804 | 0.852 | 0.803 | 0.812 | 0.824 | 0.816 | 0.821 |
| | F | 0.825 | 0.823 | 0.760 | 0.802 | 0.805 | 0.839 | 0.814 |
| RoBERTa | P | 0.865 | 0.828 | 0.725 | 0.815 | 0.808 | 0.876 | 0.823 |
| | R | 0.823 | 0.841 | 0.814 | 0.821 | 0.827 | 0.829 | 0.834 |
| | F | 0.843 | 0.834 | 0.767 | 0.818 | 0.817 | 0.852 | 0.828 |
| BiLSTM-CRF | P | 0.893 | 0.876 | 0.864 | 0.563 | 0.812 | 0.854 | 0.814 |
| | R | 0.765 | 0.803 | 0.760 | 0.697 | 0.687 | 0.616 | 0.728 |
| | F | 0.824 | 0.838 | 0.809 | 0.623 | 0.744 | 0.716 | 0.768 |
| BERT-CRF | P | 0.841 | 0.837 | 0.733 | 0.803 | 0.849 | 0.878 | 0.838 |
| | R | 0.860 | 0.847 | 0.811 | 0.808 | 0.841 | 0.840 | **0.841** |
| | F | 0.850 | 0.842 | 0.770 | 0.805 | 0.845 | 0.859 | **0.839** |
| BERT-BiLSTM-CRF | P | 0.844 | 0.844 | 0.739 | 0.853 | 0.843 | 0.827 | **0.839** |
| | R | 0.835 | 0.846 | 0.863 | 0.838 | 0.838 | 0.811 | 0.838 |
| | F | 0.839 | 0.845 | 0.796 | 0.846 | 0.840 | 0.819 | 0.838 |

By analyzing the above experimental results, we can draw the following conclusions:

(1) The six models adopted in the experiment have achieved good results in the geological news text NER task.

(2) In the geological news text NER task, the F1 scores of the BERT, DistilBERT, RoBERTa, BERT-CRF, and BERT-BiLSTM-CRF models are 0.824, 0.814, 0.828, 0.839, and 0.838, respectively. Compared to BiLSTM-CRF, the F1 scores increase by 5.6%, 4.6%, 6%, 7.1%, and 7%. This shows that as the BERT pre-training model can understand the contextual information of the text well and solve the polysemy problem, it has a good effect on the named entity recognition task.

(3) The improved DistilBERT and RoBERTa models based on BERT achieve F1 scores of 0.814 and 0.828, respectively. Compared to the BERT model, the entity recognition effect of the DistilBERT model is slightly worse, while the RoBERTa model is better.

(4) In the geological news text NER task, the P, R, and F1 scores of the BERT-CRF model improve by 1.9%, 1.2%, and 1.5%, respectively, compared to the BERT model because of a mutual constraint relationship between the tags (e.g., the tag of an entity can only start with "B" but not "I"). It can be seen after adding the CRF layer that CRF can

deal with the mutual constraint relationship between the tags and effectively solve the problem of inconsistent sequence tags.

(5) The P, R, and F1 scores of BERT-CRF are 0.838, 0.841, and 0.839, respectively, which are the best among all models. Compared to the BERT-BiLSTM-CRF model, which introduced the BiLSTM layer, the two achieved comparable results in the NER task of geological news texts. The reason for this is that the BERT model itself is effective in feature extraction, and the BiLSTM layer is introduced based on the BERT-CRF model. Overfitting occurs after this layer is trained, resulting in a decreased effect.

### *3.3. Effect of Entity Type and Quantity*

Figure 5 shows the F1 scores of the six models in the six entity categories in the form of a bar chart. Using the same model, there is a gap in the recognition effect for the different entity categories. The recognition effect is better for the entities of time, organization, name, and geographic location because the number of these entities in the corpus is large, their contextual information is more abundant, and the text features are more obvious. The recognition effect of job titles and events is poor. Two reasons account for this result: (1) the number of these two types of entities is small, especially job titles, which leads to insufficient contextual information for the neural network to learn; (2) the entities of the event class are usually nested entities, and geographic location and organization entities often appear, increasing the difficulty of identification.

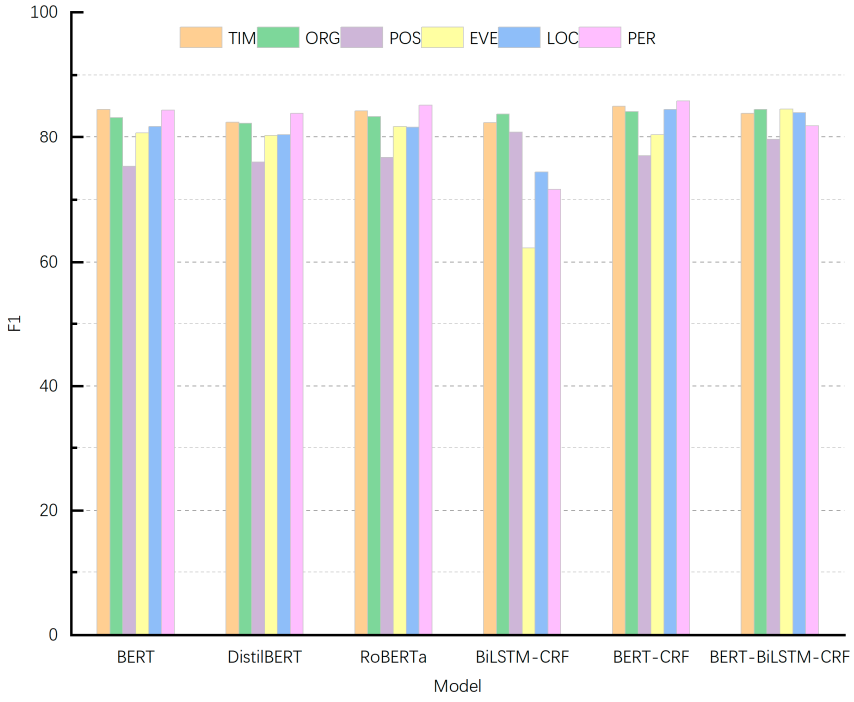


**Figure 5.** F1 scores of six types of entities on six models.

### *3.4. Influence of Model Hyperparameters*

The hyperparameters need to be set before the model is trained. The setting of hyperparameters plays a role in the training effect of the model. For different models and datasets, experiments are often required to find the most suitable hyperparameters. In the experiment, we mainly discussed the effects of the learning rate and training times on the model training effect. The F1 score of the model is used as the condition for parameter evaluation. When the F1 score is the largest, this means that the parameter is the best parameter. In deep learning, the learning rate is a common hyperparameter. This setting not only determines whether the objective function can converge but also affects the speed of convergence. When the setting is too small, the convergence speed may be too slow; when the setting is too large, it may lead to non-convergence. Therefore, choosing

an appropriate learning rate is critical during model training. Based on all models, the experiment was performed by changing the learning rate of the model. Table 5 shows the experimental results for different learning rates. The F1 scores of all models are the largest when the learning rate is $3 \times 10^{-5}$, which means that the entity recognition effect of the model is the best under this parameter condition.

**Table 5.** P, R and F1 scores of the model under different learning rates. The numbers in bold font are the three indicators (P, R, F1) with the highest corresponding scores in each set of experiments.

| Model | Learning_Rate | P | R | F |
|---|---|---|---|---|
| BiLSTM-CRF | $1 \times 10^{-5}$ | 0.795 | 0.725 | 0.758 |
| | $2 \times 10^{-5}$ | 0.804 | **0.732** | 0.766 |
| | $3 \times 10^{-5}$ | 0.814 | 0.728 | **0.768** |
| | $4 \times 10^{-5}$ | **0.817** | 0.721 | 0.766 |
| | $5 \times 10^{-5}$ | 0.806 | 0.725 | 0.763 |
| BERT | $1 \times 10^{-5}$ | 0.805 | 0.823 | 0.814 |
| | $2 \times 10^{-5}$ | 0.814 | 0.827 | 0.821 |
| | $3 \times 10^{-5}$ | 0.819 | **0.829** | **0.824** |
| | $4 \times 10^{-5}$ | **0.823** | 0.819 | 0.821 |
| | $5 \times 10^{-5}$ | 0.817 | 0.822 | 0.819 |
| DistilBERT | $1 \times 10^{-5}$ | 0.798 | 0.812 | 0.805 |
| | $2 \times 10^{-5}$ | 0.803 | 0.814 | 0.808 |
| | $3 \times 10^{-5}$ | 0.808 | **0.821** | **0.814** |
| | $4 \times 10^{-5}$ | **0.812** | 0.811 | 0.811 |
| | $5 \times 10^{-5}$ | 0.805 | 0.818 | 0.811 |
| RoBERTa | $1 \times 10^{-5}$ | 0.809 | 0.816 | 0.812 |
| | $2 \times 10^{-5}$ | 0.817 | 0.822 | 0.819 |
| | $3 \times 10^{-5}$ | 0.823 | **0.834** | **0.828** |
| | $4 \times 10^{-5}$ | **0.826** | 0.828 | 0.827 |
| | $5 \times 10^{-5}$ | 0.821 | 0.827 | 0.824 |
| BERT-BiLSTM-CRF | $1 \times 10^{-5}$ | 0.821 | 0.823 | 0.822 |
| | $2 \times 10^{-5}$ | 0.832 | 0.833 | 0.832 |
| | $3 \times 10^{-5}$ | **0.839** | 0.838 | **0.838** |
| | $4 \times 10^{-5}$ | 0.834 | **0.838** | 0.836 |
| | $5 \times 10^{-5}$ | 0.834 | 0.825 | 0.829 |
| BERT-CRF | $1 \times 10^{-5}$ | 0.823 | 0.819 | 0.821 |
| | $2 \times 10^{-5}$ | 0.829 | 0.832 | 0.831 |
| | $3 \times 10^{-5}$ | **0.838** | **0.841** | **0.839** |
| | $4 \times 10^{-5}$ | 0.836 | 0.838 | 0.837 |
| | $5 \times 10^{-5}$ | 0.832 | 0.824 | 0.828 |

Figure 6 is a line graph showing the effect of epoch times on the F1 score. As shown in the figure, the abscissa is the epoch value, and the ordinate is the F1 score. The F1 scores of the six models all increase with the increase in the number of training rounds. First, the F1 score of the BiLSTM-CRF model in the first few rounds is much lower than that of the BERT, DistilBERT, RoBERTa, BERT-CRF, and BERT-BiLSTM-CRF models. Second, as the number of training increases, the F1 scores of the six models gradually increase. The F1 score of the BiLSTM-CRF model gradually becomes close to the F1 score of the other five models that introduced BERT. At the 8th epoch, the F1 scores of the BERT-CRF model and the BERT-BiLSTM-CRF model reach the maximum. Finally, the F1 scores of the six models tend to be stable, but the F1 score of the BilSTM-CRF model lags behind those of the other five models.

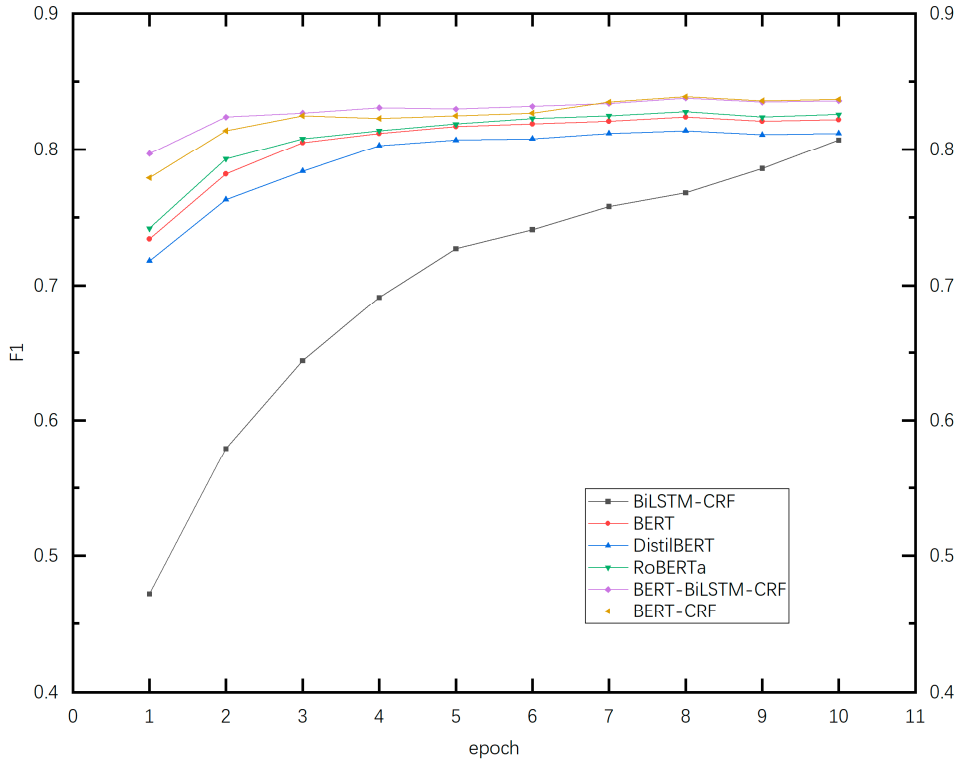


**Figure 6.** Changes in F1 score with increasing epochs.

## 4. Conclusions

This study uses a deep learning-based model to perform named entity recognition on geological news texts. As there is no public labeling training dataset for model comparison in the field of geological news, we collected geological news texts from the China Geological Survey. We annotated the data using automatic annotation and manual calibration with the open source annotation tool YEDDA to build a corpus of geological news texts. In past NER research in the geological field, Chen et al. [23] proposed the BERT- BiLSTM-CRF model to perform entity recognition in Chinese mineral texts. In the experiment, eight entity types were extracted, and their F1 scores all exceeded 95%. Xie et al. [24] obtained F1 scores of 94.65% and 95.67% on the MSRA and People's Daily corpora, respectively, in the BERT-BiLSTM-CRF model for named entities. Our experiment compares six models on the constructed geological news dataset. The F1 score of the BERT-BiLSTM-CRF model is 0.838, which achieves a high entity recognition effect, and the F1 score of the BiLSTM-CRF model is 0.768, which is less efficient than the BERT-based models. Compared to the previous study, we annotated 21,323 entities, trained based on the geological news dataset, extracted six entity types, and the F1 score of the model reached 0.839. The experimental results show that the proposed model can also achieve a good entity recognition effect in the field of geological news.

Although this research has achieved good results in NER tasks in the field of geological news, there are still some shortcomings and areas that can be further improved. First, the geological news corpus constructed in this study is relatively small, resulting in too few job titles and event entities, thus influencing the effect of entity recognition. Therefore, the dataset should be expanded in future studies. Second, as the process is cumbersome, it is inevitable that some labeling errors will occur. Moreover, the original model should be improved to enhance its performance and improve the entities' recognition effect. The information extracted from the geological news text can be applied to constructing geological news knowledge graphs.

**Author Contributions:** Conceptualization, Y.W.; methodology, C.H. and Y.W.; software, C.H.; validation, Y.H., Y.L. and X.Z.; data curation, C.H. and Y.Y.; writing—original draft preparation, C.H.; writing—review and editing, Y.W., Y.Y. and Y.H.; supervision, Y.W. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the China University of Geosciences (Beijing) College Students' Innovation and Entrepreneurship Training Program under Grant X202211415100.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Data sharing is not applicable.

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study, in the collection, analyses, or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

## References

1. Goralski, M.A.; Tan, T.K. Artificial intelligence and sustainable development. *Int. J. Manag. Educ.* **2020**, *18*, 100330. [CrossRef]
2. Khurana, D.; Koli, A.; Khatter, K.; Singh, S. Natural language processing: State of the art, current trends and challenges. *arXiv* **2017**, arXiv:1708.05148. [CrossRef]
3. Shaalan, K.; Raza, H. Arabic named entity recognition from diverse text types. In Proceedings of the International Conference on Natural Language Processing, Gothenburg, Sweden, 25–27 August 2008; pp. 440–451.
4. Alfred, R.; Leong, L.C.; On, C.K.; Anthony, P. Malay named entity recognition based on rule-based approach. *Int. J. Mach. Learn. Comput.* **2014**, *3*, 300–306. [CrossRef]
5. Shaalan, K.; Raza, H. NERA: Named entity recognition for Arabic. *J. Am. Soc. Inf. Sci. Technol.* **2009**, *60*, 1652–1663. [CrossRef]
6. Todorovic, B.T.; Rancic, S.R.; Markovic, I.M.; Mulalic, E.H.; Ilic, V.M. Named entity recognition and classification using context Hidden Markov Model. In Proceedings of the 2008 9th Symposium on Neural Network Applications in Electrical Engineering, Belgrade, Serbia, 25–27 September 2008; pp. 43–46.
7. Eddy, S.R. What is a hidden Markov model? *Nat. Biotechnol.* **2004**, *22*, 1315–1316. [CrossRef] [PubMed]
8. Och, F.J.; Ney, H. Discriminative training and maximum entropy models for statistical machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 295–302.
9. Ratnaparkhi, A. Maximum Entropy Models for Natural Language Processing. In *Encyclopedia of Machine Learning and Data Mining*; Springer: New York, NY, USA, 2017; pp. 800–805.
10. Lafferty, J.; McCallum, A.; Pereira, F.C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the 18th International Conference on Machine Learning, San Francisco, CA, USA, 28 June–1 July 2001.
11. Zhang, X.Y.; Ye, P.; Wang, S.; Du, M. Geological entity recognition method based on deep belief network. *Chin. J. Petrol.* **2018**, *34*, 343–351.
12. Liu, P.; Ye, S.; Shu, Y.; Lu, X.L.; Liu, M.M. Research on coal mine safety knowledge graph construction and intelligent query method. *Chin. J. Inf.* **2020**, *34*, 49–59.
13. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
14. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K.J.a.p.a. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
15. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
16. Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; Dyer, C. Neural architectures for named entity recognition. *arXiv* **2016**, arXiv:1603.01360.
17. Viterbi, A.J. A personal history of the Viterbi algorithm. *IEEE Signal Processing Mag.* **2006**, *23*, 120–142. [CrossRef]
18. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems 30, Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017*; Curran Associates Inc.: Red Hook, NY, USA, 2017.
19. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
20. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv* **2019**, arXiv:1910.01108.
21. 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.
22. Yang, J.; Zhang, Y.; Li, L.; Li, X. YEDDA: A lightweight collaborative text span annotation tool. *arXiv* **2017**, arXiv:1711.03759.

23. Chen, Z.L.; Yuan, F.; Li, X.H.; Zhang, M.M. Joint extraction of named entities and relations from Chinese rock description text based on BERT-BiLSTM-CRF Model. *Geol. Rev.* **2022**, *68*, 742–750. Available online: https://kns.cnki.net/kcms/detail/detail.aspx?doi=10.16509/j.georeview.2022.01.115 (accessed on 23 July 2022).
24. Xie, T.; Yang, J.A.; Liu, H. Chinese entity recognition based on BERT-BiLSTM-CRF model. *Comput. Syst. Appl.* **2020**, *29*, 48–55.