Expert Systems With Applications 206 (2022) 117727

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

# Chinese mineral named entity recognition based on BERT model

Yuqing Yu [a], Yuzhu Wang [a,*], Jingqin Mu [a,b], Wei Li [c], Shoutao Jiao [d], Zhenhua Wang [a], Pengfei Lv [e], Yueqin Zhu [f,*]

[a] School of Information Engineering, China University of Geosciences, Beijing 100083, China
[b] Department of Computer Science, Tangshan Normal University, Tangshan 063000, China
[c] School of Computer Science, The University of Sydney, Sydney, NSW 2006, Australia
[d] Development and Research Center, China Geological Survey, Beijing 100037, China
[e] National Geological Library of China, Beijing 100083, China
[f] National Institute of Natural Hazards, Ministry of Emergency Management, Beijing 100085, China

ARTICLE INFO

*Keywords:*
Named entity recognition
Mineral text
BERT
CRF

ABSTRACT

Mineral named entity recognition (MNER) is the extraction for the specific types of entities from unstructured Chinese mineral text, which is a prerequisite for building a mineral knowledge graph. MNER can also provide important data support for the work related to mineral resources. Chinese mineral text has many types of entities, complex semantics, and a large number of rare characters. To extract entities from Chinese mineral literature, this paper proposes an MNER model based on deep learning. To create word embeddings for mineral text, Bidirectional Encoder Representations from Transformers (BERT) is used. Moreover, the transfer matrix of the Conditional Random Field (CRF) algorithm is combined to improve the accuracy of sequence labeling. Finally, some experiments are conducted on the constructed dataset. The results show that the model can effectively recognize seven mineral entities with an average F1-score of 0.842.

## 1. Introduction

Chinese mineral literature contains a lot of meaningful information and rich knowledge. The extraction of mineral information based on text data will provide an important data support for the work of mineral resources. Therefore, data mining and knowledge extraction are urgently needed. The purpose of information extraction is to extract entities and their relationships from unstructured text and transform them into structured expressions. With the continuous increase of the digital scale of mineral-related data, it is very difficult to extract useful information from digital files manually. The recognition of named mineral entities based on deep learning is an important method to realize the automatic extraction of mineral information, and it is also a prerequisite for constructing a knowledge graph in the mineral field.

At present, there are few researches especially in the field of mineral named entity recognition (MNER), but some scholars have applied deep learning to geological named entity recognition and achieved certain results. For example, Qiu et al. (2019) proposed a model that combined Bi-directional Long Short-Term Memory (BiLSTM) with Conditional Random Field (CRF), which uses attention mechanism to capture the correlation information between words and extract geological entities from geological reports, such as geological history and geological structure.

The characteristics of mineral resources are complex, and related entities include geographic location, mineral type, mineral scale, geochemistry, etc. Additionally, compared with common field texts, the authors and research areas of mineral documents are different, and the text features are diverse. The mineral entities have longer characters and more rare words, such as the mineral type entity Qixiashan Layered-hydrothermal lead-zinc-silver poly-metallic mineral. In addition, there is a phenomenon of mutual nesting of mineral entities, such as Tianjiahe small placer gold deposit, which includes three entities: deposit (Tianjiahe), mineral scale (small) and mineral type (gold). At present, there are no publicly available datasets in the mineral field, so it is very challenging to identify named entities in the field of mineral text.

In order to solve the challenge of MNER, this research will propose a named entity recognition (NER) model based on deep learning in the mineral field. Based on the literature in the Chinese mineral field of China National Knowledge Infrastructure (CNKI), according to the characteristics of mineral text, we extracted the information such as the location of deposit, mineral scale and mineral type, mining areas. Compared with the previous NER model, we introduced Bidirectional Encoder Representations from Transformers (BERT). BERT (Devlin et al.,

* Corresponding authors.
*E-mail addresses:* yuyq@cugb.edu.cn (Y. Yu), wangyz@cugb.edu.cn (Y. Wang), mujq@cugb.edu.cn (J. Mu), weiwilson.li@sydney.edu.au (W. Li), jshoutao@mail.cgs.gov.cn (S. Jiao), wangzh@cugb.edu.cn (Z. Wang), lpengfei@mail.cgs.gov.cn (P. Lv), yueqinzhu@ninhm.ac.cn (Y. Zhu).

https://doi.org/10.1016/j.eswa.2022.117727
Received 17 November 2021; Received in revised form 9 May 2022; Accepted 31 May 2022
Available online 7 June 2022

2018) is a language representation model, which can pre-train deep bidirectional representations based on context from a large-scale text corpus. BERT performed well in sequence labeling tasks, which can effectively characterize the ambiguity of words and enhance the semantic representation of sentences. We merged BERT with the Long and Short-term Memory (LSTM) (Hochreiter & Schmidhuber, 1997) and CRF (Lafferty et al., 2001) to conduct comparative experiments on the MNER task.

The main contributions of this article are as follows:

(1) Our work first applied BERT to MNER task, combined BERT with other models, and realized extracting more types of mineral entities from a smaller corpus.

(2) This paper evaluated the influence of mineral entity type and the number of tags, explored optimum hyper-parameters for the models, and optimized the loss function to improve the performance of NER.

(3) This study obtained mineral literature from CNKI by web crawler, and performed text pre-processing such as data cleaning, sentence segmentation and tagging the mineral location (the location of mining area, deposit, section), mineral body, scale, mineral type, etc., then constructed a Chinese corpus in the mineral field.

The rest of this paper is structured as follows. Section 2 shows related work. Section 3 provides the proposed methodology. Section 4 introduces and analyzes the experiments. Section 5 presents our conclusions and discusses prospects for future work.

## 2. Related work

Early NER models were mainly based on rules and dictionaries (Hettne et al., 2009). Later, machine learning methods gradually emerged. Commonly used models include Maximum Entropy (ME) (Chieu & Ng, 2003; Haarnoja et al., 2018), Hidden Markov Model (HMM) (Baksa et al., 2016; Bikel et al., 1999; Zhao, 2004), and CRF (Li et al., 2008; Wang et al., 2018). In recent years, with the advancement of computer software and hardware, deep learning has evolved and gradually become the main method to solve Natural Language Processing (NLP) problems. Compared with the earlier manual approaches that build rules and dictionaries based on data features, deep learning models contain more semantic information and can reduce manual intervention, and are suitable for solving serialized annotation problems such as named entity recognition. Deep learning pays more attention to the construction of the overall neural network model and parameter optimization. After the model is built, it can be applied to related fields with only minor adjustments, and therefore has more portability.

Recurrent Neural Network (RNN) is a neural network model commonly used in the early NLP field. RNN is suitable for processing and predicting sequence data, but when the sequence is too long, the model has the problem of gradient disappearance. Based on RNN, Hochreiter and Schmidhuber (1997) proposed a LSTM neural network. The model is used to control the circulation and loss of features through gate mechanism, which effectively solves the problem of learning long-term dependencies of RNN. The current method for solving NER problems is usually to combine statistical machine learning models with deep learning models. Lample et al. (2016) proposed an LSTM-CRF model that combined Long and Short-term Memory neural networks with Conditional Random Fields, and applied it in English NER tasks. Compared with the previous model, its F1 score is increased by about 5%. Collobert et al. (2011) proposed the CNN-CRF model, which combined Convolutional Neural Networks (CNN) and CRF for NER tasks. Later, Google proposed a BERT pre-trained model, which has now become a popular model in the field of NLP (Liu et al., 2020; Mutinda et al., 2021).

Named entity recognition has been widely used in many fields, some scholars have studied geological NER and text segmentation, and mainly extract entities such as geological age, strata and geological structures in geological texts. For example, Zhang et al. (2018) formulated the tagging specifications of geological entities based on the characteristics of geological literature, and designed a geological NER model based on Deep Belief Networks (DBN). Li et al. (2021) applied BERT to the word segmentation task of geological text, and integrated BiLSTM with CRF to improve the word segmentation ability of the model in unstructured geological reports. Deping et al. (2021) used ELMo and CNN to extract word dynamic features and local features, and spliced them with pre-trained word vectors, and proposed a geological NER model of ELMo-CNN-BiLSTM-CRF, which can effectively identify entities such as geological age and geological process. Fan et al. (2019) combined the multi-branch BiGRU and CRF to extract entities in geological hazard literature, and built a knowledge graph of geological hazard.

At present, there are few researches on NER in the field of mineral. In this study, a Chinese mineral corpus was constructed and BERT was combined with the BiLSTM and CRF to recognize mineral named entities, and realized the effective extraction of seven kinds of mineral entities.

## 3. Materials and methods

In this section, we will introduce the method of NER model involved in this study in detail, and apply the fusion model to the Chinese mineral corpus.

### 3.1. Word2vec

In the NER task, we must first convert the language into the data type that can be processed by the neural network. Word embedding (Mikolov et al., 2013) technology achieves this goal by transforming words into vectors in numeric form. The Word2vec is a concrete realization method of creating word embeddings, and Zhang et al. (2015) used it for Chinese comments sentient classification. Word2vec believes that there is a certain relationship between several words that are close in a sentence. Word2vec contains two neural network language models, the structure is shown in Fig. 1. The purpose of training the CBOW and Skip-gram is to obtain the weight of the hidden layer, and the weight matrix is the expected word vector.

The CBOW predicts the central word based on the context. In the Skip-gram model, each word will be used as the central word. At the same time, all surrounding words need to be retrained. Therefore, the training time of Skip-gram is longer than CBOW. At the same time, Skip-gram training effect is better, especially for rare words and domain vocabularies. At present, Word2vec is still a widely used word embedding pre-training tool, but the model has a shortcoming that the word vectors obtained by this method are unchangeable, resulting that the meaning of words or phrases cannot be changed according to different contexts.

### 3.2. BiLSTM-CRF

The structure of BiLSTM-CRF is mainly divided into three layers, they are the word vector input layer, the BiLSTM layer and the CRF layer. The main process of the experiment is as follow. First, input the word vector sequence. Next, perform feature extraction through the BiLSTM layer, so as to obtain the probability of each word on each tag. Finally, use the CRF layer to constrain different combinations and obtain the optimal sequence labeling. BiLSTM can only predict the relationship between the text sequence and the label, and the relationship between labels can be calculated by the CRF transition matrix.

In the word vector matrix of BiLSTM, the output result of the BiLSTM hidden layer is obtained by combining the forward LSTM and the backward LSTM, and then splicing the output word vectors. Assuming that the complete hidden layer state is an n-dimensional matrix at this time, a linear layer needs to be connected at the same time to map it from n-dimensional to K-dimensional, and K is the number of tags.

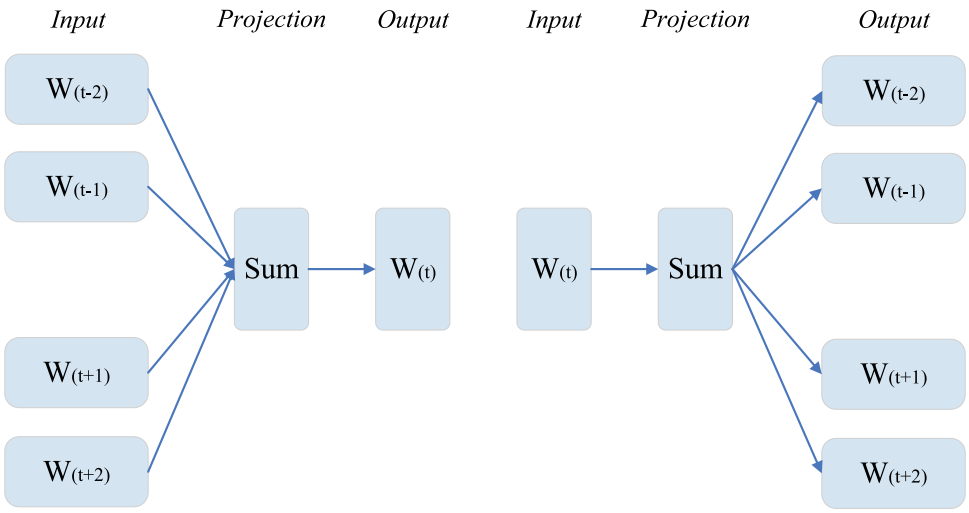


**Fig. 1.** The left is CBOW, which predicts the current word W(t) based on the context. The right is Skip-gram, which predicts surrounding words given the current word W(t).

In the CRF layer, the constraint relationships between tags are predicted by calculating the transfer score, so that the number of invalid prediction tags is greatly reduced. The formula is as follows:

$$\text{score}(X, y) = \sum_{i=0}^{n} T_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i} \tag{1}$$

For the input sequence $X = (X_1, X_2, \ldots, X_n)$, there is a corresponding output sequence $y = (y_1, y_2, \ldots, y_n)$. $T$ represents the transfer score between sequence labels. The output result of the previous layer is recorded as a $P$ matrix, where $P = (p_1, p_2, \ldots, p_n)$. CRF obtains the final output sequence $y_i$ corresponding to the input sequence $x_i$ through normalization.

In addition, in order to make the experimental results more adequate, we also compared the Gated Recurrent Unit (GRU) (Cho et al., 2014), which is a simplified and improved neural network model of LSTM. In the GRU neural network, the three gate units in LSTM are replaced by the Update gate and the Reset gate, and the parameters and tensor of the model are reduced by this way. Similar to the BiLSTM-CRF, we input the word vector obtained by Word2vec to the Bidirectional GRU layer, and output the label sequence with the highest probability through the CRF layer.

### 3.3. BERT

This article used BERT to combine BiLSTM and CRF to build BERT-CRF and BERT-BiLSTM-CRF models for MNER.

Compared with Elmo (Peters et al., 2018) and openAI-GPT (Radford & Narasimhan, 2018) models, BERT mainly uses 12 layers encoder of Transformer (Vaswani et al., 2017) as the basic unit, and then extract the context information from the previous and subsequent text to obtain the word vector. Additionally, BERT uses the Masked Language Modeling (MLM) to achieve the effect of learning the contextual semantic information of the corpus.

Transformer's frame structure is shown in Fig. 2. It trains all the words in the sequence at the same time, and adds the positional encoding to each word embedding, to identify the priority of each word in the sequence. Transformer's single encoder unit consists of Multi-Head Attention and Feed Forward. In the structure of BERT, Trm is the encoding module of Transformer. When BERT is training, the encoding layer will extract contextual information based on the input of the previous layer. Besides, due to the self-attention mechanism of Transformer, each Trm can obtain text features of all words in the sequence, so as to achieve the effect of bidirectional extraction of features.

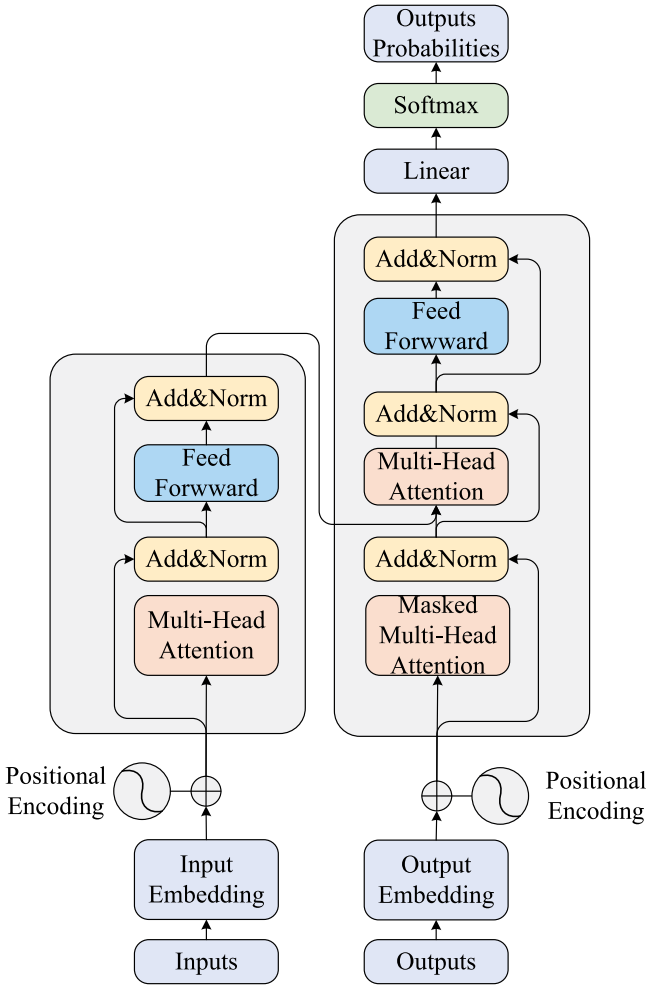


**Fig. 2.** The structure of Transformer (Vaswani et al., 2017).

In the self-attention mechanism of BERT, Q (Query) is an element in the target sequence, K (key) is an element in the input sequence, V (Value) can be obtained by calculating the correlation between Q and K. The Multi-Head Attention in the Trm is composed of multiple Scaled

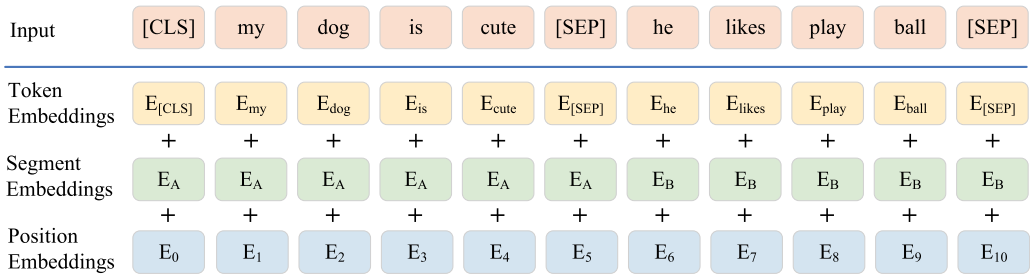


**Fig. 3.** Input representation of BERT (Devlin et al., 2018).

Dot-Product Attentions. By setting multiple sets of auxiliary matrices, the current item has multiple Q, K, and V, so that each item has multiple feature expressions to achieve the ability to make the model focus on different aspects of information.

BERT constructs the input representation of the token by three embedding layers. The final embeddings, the sum of token, segment, and position embeddings are input to the BERT's encoder layer. The formula is as follows:

$$E = TokenEmbeddings + SegmentEmbeddings + PositionEmbeddings \quad (2)$$

The structure of BERT Embedding is shown in Fig. 3. The [CLS] mark indicates that this sentence used for classification, and the [SEP] is used to segment different sentences. The first layer of token embedding uses WordPiece tokenization to transform each word piece token into a 768-dimensional vector. The second layer of segment embedding generates fixed tokens to determine which sentence the token belongs to. Position embeddings encodes the position of the vector according to the diverse positions of the words in the sentence.

### 3.4. MNER based on BERT

This research first applied BERT to create word embeddings on the mineral text, and then used the word vectors obtained as the input of CRF and BiLSTM-CRF. The specific process of BERT-CRF module is shown in Fig. 4.

The experiments used the "BERT-Base, Chinese" as base model, which is the official BERT pre-trained model for Chinese published by Google. The data input to the model is the superposition of the vector of each word, which contains the sentence vector and the position vector. As mentioned earlier, Word2vec was the main model for creating word embeddings in the early days. But Word2vec cannot change the trained word vectors, so it is difficult to deal with the common ambiguity problems in Chinese. Since many Chinese place names in mineral naming entities have long characters and may have ambiguities. For example, Jinping Copper Factory is a place name, but "Jin" also means the gold in Chinese, so it may be identified as two types of minerals: gold and copper. Using BERT training method can update the word vector in the pre-trained model, thereby improving the recognition effect for mineral text. Finally, BERT layer will output a 128-dimensional vector matrix as the input of BiLSTM-CRF or being directly inputted to the CRF layer.

The CRF layer will formulate rules for the annotation of the whole sequence, and score the sequence labeling results through the transfer matrix. As the final output decoding layer, CRF layer is used to obtain the most practical prediction of mineral named entities.

After word embeddings trained by BERT are used as the input of the BiLSTM layer, the model will output a matrix. Since BiLSTM usually works between text sequences and labels, it does not pay attention to the relationship between the labels. Generally, the label of a word cannot be judged by the emission score alone. It needs to be combined with the information near the current item to finalize it. Therefore, the final label can be determined by CRF.

## 4. Experimental evaluation

This section introduces the experimental process of MNER based on deep learning, including data processing, entity tagging, experimental environment, training parameter settings, and finally results analysis and model effect comparison.

### 4.1. Text pre-processing

The raw data are 200 mineral resources-related documents obtained by web crawlers from CNKI, totaling about 890,000 words. There are many problems with the raw data so that they cannot be used directly. For example, interference information, such as title number, table, picture, which will make text recognition very difficult. In addition, too long text will lead to a decrease in recognition accuracy, so the text data needs to be pre-processed. In this study, the text pre-processing stage was mainly performed for data cleaning and sentence segmentation. To obtain text that can be tagged with entities, the main steps of text pre-processing include handling missing values, checking format and content, correcting logical errors, cutting long text. The data cleaning task is to divide the continuous original text into a sequence of tokens containing only words, punctuation marks, numbers, and spaces. Sentence segmentation divides the cleaned text into standardized sentences by identifying sentence boundaries such as periods and question marks.

### 4.2. Text annotation

Text annotation is the basic work for constructing a corpus of pre-trained models, and it is also a prerequisite for the realization of NER. Text annotation refers to tagging the entities and non-entity data in the text separately. We use the BIO format to annotate the entities in the sentence. The first word of each entity is annotated as B and the remaining words are tagged as I, and the non-entity words are annotated as O. After text annotation, the computer can identify the category of each word and process it. We have annotated a total of 12,837 entities, including seven main features of mineral resources: mineral location (mining area, deposit, section), mineral body, scale, mineral type, geologic occurrence, mineral alteration and weathering degree, as shown in Table 1. For complex entities, we annotate multiple entities separately. For example, the entity "Pengjiakuang large gold mine" (蓬(B-LOC)家(I-LOC)矜(I-LOC)大(B-SCA)型(I-SCA)金(B-MT)矿(I-MT)), "Pengjiakuang", "large", "gold mine" are tagged as Location, Scale and Mineral Type, respectively, and in the experiments the complex entities are also matched individually.

The specific tagging process has the following steps: Calling the "jieba" word segmentation tool to process the text data, obtaining the starting position of a sentence in the text and the length information of the sentence in the text, tagging entities according to the rules of mineral characteristics. The rest of the non-named entity parts are tagged as O.

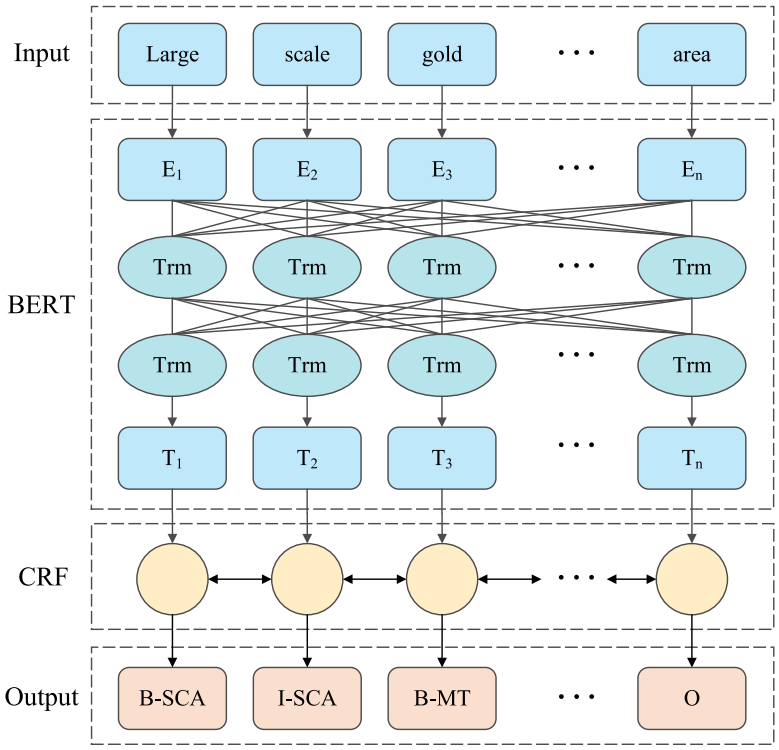


**Fig. 4.** BERT-CRF model structure. Mineral text will be input, and processed by BERT layer and CRF layer to finally get the label of each word.

**Table 1**
Type and number of mineral entity. The first column "Tags" are abbreviations used in the tagging process.

| Tags | Type | Number |
|---|---|---|
| LOC | Mineral area/deposit/section | 3867 |
| MB | Mineral body | 837 |
| SCA | Scale | 583 |
| MT | Mineral type | 4463 |
| GO | Geologic occurrence | 571 |
| MA | Mineral alteration | 2429 |
| WD | Weathering degree | 87 |

**Table 2**
Experimental environment.

| Category | Configuration |
|---|---|
| Hardware | CPU: Intel(R) E5-2620 v4 @ 2.10 GHz<br>GPU: 4*NVIDIA Tesla K80<br>OS: CentOS 8.3<br>Video memory: 11 GB DDR6 |
| Software | CUDA: 10.1<br>Python: 3.6<br>Pytorch: 1.6.0<br>Tensorflow: 1.14<br>Numpy: 1.19.2 |

**Table 3**
Hyper-parameters of the experimental models.

| Hyper-Parameter | Parameter values |
|---|---|
| max_seq_length | 128 |
| epoch | 10 |
| batch_size | 32 |
| bert_learning_rate | 3e−5 |
| lstm_learning_rate | 0.001 |
| dropout | 0.5 |
| clip | 5.0 |
| bilstm_size | 128 |
| gru_size | 128 |

### 4.3. Training

The dataset is divided into training, validation and test sets at a ratio of 8:1:1, containing 10152, 1336 and 1349 entities, respectively. The pre-trained model used in the experiment is "BERT-Base, Chinese", which contains 12-layer, 768-hidden, 12-heads and 110M parameters. The word embeddings output by BERT comes from the weighted average of these 12-layer networks. Finally the word embeddings of 768 dimensions will be input to the next layer of model. For comparison with other research works, we also input embeddings into the BiGRU-CRF layer in our experiments.

The hardware configuration and software environment of the experiments are shown in Table 2.

In order to ensure the credibility of the experimental results, we will train with fixed experimental hyper-parameters. The specific hyper-parameters are shown in Table 3. *max_seq_length* is the length of the maximum sequence; an *epoch* is a period that the training set is completely passed through the neural network and returned once; *batch_size* is the amount of data processed each time; *bert_learn_rate* is the initial learning rate set in BERT; *lstm_learn_rate* is the learning rate set in BiLSTM; *dropout* refers to the proportion of neurons discarded during training to prevent overfitting; *clip* refers to the gradient threshold set to prevent gradient explosion; *bilstm_size* refers to the number of hidden layers of the BiLSTM network. *gru_size* refers to the number of hidden layers of the BiGRU network.

### 4.4. Evaluation metrics

The three evaluation indicators used in all experiments in this article are precision, recall, and F1 score, which are widely used in NER evaluation criteria (Lample et al., 2016; Liu et al., 2020). The above three evaluation indicators are derived from the confusion matrix of the classification results. As is shown in Table 4, the four values are all composed of two letters, where $T$ and F represent the correctness of the prediction, which are correct and wrong respectively. P and $N$ represent the predicted category, which are positive examples and

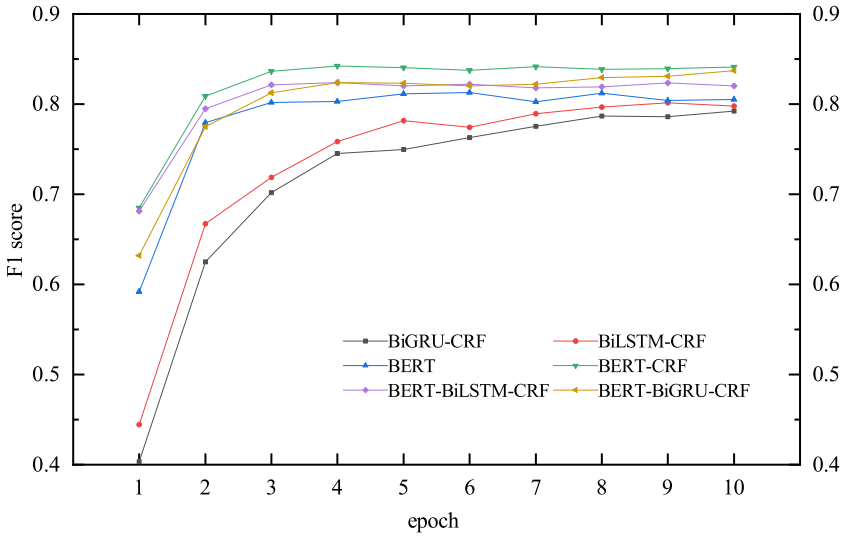


**Fig. 5.** Changes in F1 score as epoch increases.

**Table 4**
Confusion matrix of classification results.

| Actual Values | Predicted Values | |
|---|---|---|
| | Positive | Negative |
| Positive | TP | FN |
| Negative | FP | TN |

negative examples, respectively, and the result of adding these four values is the total number of samples.

The precision represents the proportion of the number of correctly named entities recognized by the model to the number of all identified named entities. Recall represents the proportion of the number of correctly named entities identified by the model to the total number of named entities in the sample. In the actual experimental results, there may be conflicts between the Precision and Recall. In this case, this article also sets the F1 score to comprehensively evaluate the experimental results.

### 4.5. Experimental results and analysis

We conduct the NER task on the mineral dataset in six different models, and applied the pre-trained model on test set to evaluate its actual performance. The results show that the model can effectively extract mineral entities. Tables 5 and 6 show examples of extracting entities from Chinese mineral text. In the following, we will analyze the experimental results from the perspectives of models and entities.

#### 4.5.1. Comparison of different models

The results of the NER model in this experiment are shown in Table 7, and the results of the experiment are analyzed as follows:

(1) In the NER task of mineral text, BiLSTM-CRF's Precision, Recall and F1 reached 0.806, 0.797, 0.801, respectively, BiGRU-CRF's Precision, Recall and F1 reach 0.792, 0.793, 0.792, respectively.

(2) The F1 scores of BERT, BERT-CRF, BERT-BiLSTM-CRF and BERT-BiGRU-CRF are 0.813, 0.842, 0.824, and 0.837, respectively. The results show that compared with BiLSTM-CRF and BiGRU-CRF, the four models introduced with BERT have improved the average F1 score, confirming the effectiveness of introducing BERT in the MNER task.

(3) The P, R and F1 of BERT-CRF reached 0.833, 0.852, 0.842, respectively, which performed best among all models. The precision is significantly ahead of BERT-BiLSTM-CRF 0.791, but the recall rate is equivalent. Through the comparison of BERT-CRF and BERT-BiLSTM-CRF, it is found that after adding BiLSTM and then CRF constraint, the effect is reduced. The reason is that the word vector of the pre-trained model in BERT is already relatively accurate, and over-fitting occurred after the two-way LSTM training, which leads to a decrease in the effect. The CRF layer in the model is responsible for scoring the labeling results of the sequence through the transition matrix and finally selecting the labeling sequence that is best meet the actual situation. Thereby reducing the number of invalid tag predictions, and combining with BERT can further improved the precision and recall rate.

#### 4.5.2. Effects of type and quantity of entities

In the recognition of seven types of mineral entities, all models basically show that the recognition effect of MA entities is good, and the F1 score can reach more than 90%. The reason is that mineral alteration is more distinguishable in Chinese than other entities, and the number of tags reaches 2429. Secondly, the three entities with better effect are Mineral Type, Mineral Body, and Location. These entity types also have a large number of tags. In addition, for the improved BERT models, the precision of mineral body recognition is the highest, and the recall rate of mineral alteration recognition is the highest. Geologic occurrence entities are often represented by Chinese characters mixed with numbers in the text, which is difficult to recognize. However, since the number of entity annotations is the smallest, the extraction of weathering degree is not as effective as other types. After retaining four kinds of entities including mineral alteration, mineral type, mineral body and location, we conducted comparative experiments, and the results were shown in Table 8. The results showed that the effect of the four BERT-based models was improved, and the F1-score in BERT-CRF reached 0.856.

#### 4.5.3. Influence of the model hyper-parameters

Hyper-parameters need to be set before training, which sometimes will affect the performance of the model. For different models and datasets, it usually finds the most suitable hyper-parameters by numerous experiments. The learning rate is an important parameter in the deep learning model, which determines the convergence speed of the objective function and whether it can converge to a local minimum. As for the adjustment of the learning rate of BERT-CRF, the experimental results are shown in Table 9. The best performance was achieved when the learning rate was set to 3e−5.

The performance of the models with the increase in epoch is shown in Fig. 5. The F1 score of BiLSTM-CRF and BiGRU-CRF in the first few epochs was significantly lower than that of BERT. As the number of training rounds increased, the F1 score of the model gradually

**Table 5**
The samples of MNER in Chinese.

| 原始文本 | 实体识别 |
|---|---|
| 断层内赋存蓬(B-LOC)家(I-LOC)夼(I-LOC)<br>大(B-SCA)型(I-SCA)金(B-MT)矿(I-MT)。 | 位置: 蓬家夼; 规模: 大型;<br>矿物类型: 金矿 |
| 主矿体3(B-MB)0(I-MB)号(I-MB)、3(I-MB)1(I-MB)<br>号(I-MB) 呈向(B-OC)北(I-OC)陡(I-OC)倾(I-OC)。 | 矿体：30号、31号;<br>产状：向北陡倾 |
| 硅(B-MA)化(I-MA)作用<br>广(B-WD)泛(I-WD)发布于矿区中间。 | 蚀变类型: 硅化;<br>发育程度：广泛 |

**Table 6**
The samples of MNER in English. Raw text is Chinese, and the samples are shown after translation.

| Original text | Entity recognition |
|---|---|
| Pengjiakuang large<br>gold mine occurs in the fault. | Location: Pengjiakuang; Scale: large;<br>Mineral Type: gold mine |
| Main mineral bodies No. 30 and<br>No. 31 dip steeply to the north. | Mineral Body: No. 30, No. 31<br>Geologic Occurrence: dip steeply to the north |
| Silicification is widely developed<br>in the middle of the deposit. | Mineral Alterations: silicification<br>Weathering Degree: widely developed |

increased. Among them, BiLSTM-CRF and BiGRU-CRF had the best effect and tend to be stable after 6~7 rounds, and the improved BERT models such as BERT-CRF and BERT-BiLSTM-CRF achieved the best effect and stabilized after 3~4 rounds.

#### 4.5.4. Replacement of pre-trained models

"BERT-wwm, Chinese" and "BERT-wwm-ext, Chinese" are Chinese pre-trained models published by Joint Laboratory of HIT and iFLYTEK Research (HFL) (Cui et al., 2020). Compared with "BERT-Base, Chinese", "BERT-wwm, Chinese" introduces whole word masking (wwm) strategy, and "BERT-wwm-ext, Chinese" additionally increases the training steps and training dataset size.

We replaced the base model of BERT-CRF and kept the other parameters unchanged, the experimental results are as follows: The P, R and F1 of "BERT-wwm, Chinese" reached 0.856, 0.823, 0.839, respectively, "BERT-wwm-ext, Chinese" reached 0.859, 0.828, 0.843, respectively. Compared to the official base model "BERT-wwm, Chinese", the effects of these two base models are close but show higher Precision and lower Recall.

#### 4.5.5. Optimization of the loss function

In the above experiments we use the Cross-Entropy loss function, which is widely applied in classification problems. There are two methods to optimize the loss function in this study. The first is Label Smoothing (Müller et al., 2019), which is a regularization technique that reduces the word errors and avoids the model's overfitting by adding noise to the true distribution. The second is Focal Loss (Lin et al., 2020). Focal Loss was originally designed to solve the class imbalance problem in the object detection task by adding a factor to the Cross-Entropy function to make the loss steeper, thus allowing the classifier to focus its learning purpose on samples that are difficult to classify. We use Cross-Entropy Loss and Focal Loss as the loss function, respectively, and combine them with Label Smoothing. The experimental results are shown in Table 10.

The results show that Label Smoothing has almost no improvement on recognition, while there is some improvement after replacing Cross-Entropy Loss with Focal Loss, with an F1-score of 0.851. However, the effect decreased after adding label smoothing, which may be caused by addition of noise that leads to underfitting of the model.

### 4.6. Discussion

This research was based on deep learning model to realize named entity recognition of mineral text, but there was no public dataset of the same type for the model to compare. In the past related work, there were some researches of geological NER, such as Xie et al. (2021) proposed a BERT-BiGRU-Attention-CRF model to extract five kinds of geological named entities, and the F1 score of the model was up to 0.840. Since the authors did not provide source code and transformer model has a self-attention mechanism, we compared the BERT-BiGRU-CRF model in our experiments and the results showed an F1 score of 0.837. Qiu et al. (2019) proposed a Att-BiLSTM-CRF module for geological NER too, who constructed a dataset of 5 kinds entities including geological, structure, rock, geological history, stratum and toponym, which were tagged with a total of 37311 entities. The F1-score of Att-BiLSM-CRF in the overall performance reached 0.896. We also compared the BiLSTM-CRF model, and the result showed that the F1 score was 0.801, which was less effective compared to the BERT-based model.

Compared with the previous study, we annotated a total of 12,837 named entities of seven mineral types, and trained the model based on a smaller corpus. The average F1 score of MNER reached 0.842 and reached up to 0.851 after optimizing the loss function, which achieved accurate recognition of more entity types.

## 5. Conclusions and future work

Our work was mainly to build an NER model based on deep learning to extracted mineral named entities from a large number of mineral documents, and provided a data support for constructing mineral knowledge graphs. The paper combined BERT into the previous NER model to extract seven entities in the mineral literature, with the average precision of 0.833, the average recall rate of 0.852, the average F1-score of 0.842, and the F1-score reaches 0.851 after model is optimized. The following conclusions are drawn through experimental results: (1) The F1 score was increased after introducing BERT in the MNER task. (2) The best effect was obtained by directly inputting the word vector trained by BERT to the CRF layer, while the overfitting phenomenon would occur when adding BiLSTM, which led to the decline of the model's performance. (3) The effect of entity recognition was better when the meaning of Chinese entities could be clearly distinguished and the number of tags was adequate. (4) Replacing the loss function from Cross-Entropy Loss to Focal Loss further improved the effectiveness of the model.

Although the recognition of mineral named entities is valid in this study, there are still some aspects that can be improved in future research work: (1) The recognition effect of some entities with a small number of annotations is relatively poor. We plan to increase the number of mineral entities in dataset. (2) We will construct a high-quality Chinese mineral corpus with more type of entity, such as thickness, length of deposit and stratigraphic age. (3) We plan to make further improvements to the model to improve the performance of the model, and apply the model to extract information from the mineral texts and build a domain knowledge graph.

**Table 7**
Experimental results of models. P, R, and F indicate our evaluation criteria precision, recall, and F1 score. LOC, MB, SCA, MT, GO, MA, WD are entity tags, and Avg represents the average score. The numbers in bold font are the best performance of the three indicators.

| Model | Evaluate | LOC | MB | SCA | MT | GO | MA | WD | Avg |
|---|---|---|---|---|---|---|---|---|---|
| BiLSTM-CRF | P | 0.810 | 0.834 | 0.627 | 0.859 | 0.577 | 0.832 | 0.643 | 0.806 |
| | R | 0.758 | 0.824 | 0.775 | 0.792 | 0.691 | 0.919 | 0.900 | 0.797 |
| | F | 0.783 | 0.829 | 0.693 | 0.824 | 0.629 | 0.873 | 0.750 | 0.801 |
| BiGRU-CRF | P | 0.844 | 0.833 | 0.677 | 0.833 | 0.725 | 0.732 | 0.600 | 0.792 |
| | R | 0.802 | 0.755 | 0.637 | 0.755 | 0.716 | 0.907 | 0.900 | 0.793 |
| | F | 0.822 | 0.792 | 0.657 | 0.792 | 0.721 | 0.810 | 0.720 | 0.792 |
| BERT | P | 0.850 | 0.731 | 0.727 | 0.791 | 0.811 | 0.863 | 0.705 | 0.811 |
| | R | 0.817 | 0.817 | 0.800 | 0.836 | 0.814 | 0.755 | 0.797 | 0.814 |
| | F | 0.833 | 0.772 | 0.762 | 0.813 | 0.813 | 0.805 | 0.748 | 0.813 |
| BERT-CRF | P | 0.851 | 0.924 | 0.695 | 0.834 | 0.709 | 0.853 | 0.708 | 0.833 |
| | R | 0.803 | 0.801 | 0.836 | 0.825 | 0.742 | 0.932 | 0.803 | **0.852** |
| | F | 0.841 | 0.871 | 0.776 | 0.844 | 0.741 | 0.906 | 0.769 | **0.842** |
| BERT-BiLSTM-CRF | P | 0.834 | 0.861 | 0.713 | 0.789 | 0.580 | 0.799 | 0.571 | 0.791 |
| | R | 0.885 | 0.874 | 0.807 | 0.823 | 0.729 | 0.940 | 0.800 | 0.857 |
| | F | 0.859 | 0.868 | 0.757 | 0.806 | 0.646 | 0.864 | 0.667 | 0.824 |
| BERT-BiGRU-CRF | P | 0.810 | 0.692 | 0.686 | 0.929 | 0.890 | 0.795 | 0.900 | **0.840** |
| | R | 0.681 | 0.900 | 0.800 | 0.750 | 0.840 | 0.847 | 0.885 | 0.833 |
| | F | 0.740 | 0.783 | 0.739 | 0.830 | 0.864 | 0.820 | 0.893 | 0.837 |

**Table 8**
The experimental results of retaining four entities: MA, MT, MB, LOC. The numbers in bold font are the highest F1 score.

| Model | P | R | F |
|---|---|---|---|
| BERT | 0.830 | 0.803 | 0.816 |
| BERT-BiGRU-CRF | 0.838 | 0.850 | 0.844 |
| BERT-BiLSTM-CRF | 0.826 | 0.835 | 0.830 |
| BERT-CRF | 0.835 | 0.877 | **0.856** |

**Table 9**
Model performance under different BERT learning_rate. The numbers in bold font are the highest F1 score.

| learning rate | P | R | F |
|---|---|---|---|
| 1e−5 | 0.819 | 0.817 | 0.818 |
| 2e−5 | 0.828 | 0.831 | 0.830 |
| 3e−5 | 0.833 | 0.852 | **0.842** |
| 4e−5 | 0.832 | 0.842 | 0.837 |
| 5e−5 | 0.833 | 0.813 | 0.823 |

**Table 10**
Model performance under different loss function. CEL stands for Cross-Entropy Loss, FL stands for Focal Loss. The numbers in bold font are the highest F1 score.

| Loss | P | R | F |
|---|---|---|---|
| Cross-Entropy Loss | 0.833 | 0.852 | 0.842 |
| CEL with Label Smoothing | 0.848 | 0.842 | 0.845 |
| Focal Loss | 0.835 | 0.868 | **0.851** |
| FL with Label Smoothing | 0.845 | 0.840 | 0.842 |

## CRediT authorship contribution statement

**Yuqing Yu:** Methodology, Software, Writing – original draft. **Yuzhu Wang:** Supervision, Conceptualization, Methodology, Writing – review & editing. **Jingqin Mu:** Writing – review & editing. **Wei Li:** Writing – review & editing. **Shoutao Jiao:** Writing – review & editing. **Zhenhua Wang:** Writing – review & editing. **Pengfei Lv:** Writing – review & editing. **Yueqin Zhu:** Supervision, Writing – review & editing.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 41872253, in part by the National Key Research and Development Program of China under Grant 2018YFC1505501, and in part by the GHFUND B of China under Grant ghfund202107021958.

## References

Baksa, K., Golović, D., Glavaš, G., & Šnajder, J. (2016). Tagging named entities in Croatian tweets. *Slovenščina 2.0 Empir. Appl. Interdiscip. Res.*, *4*, 20–41. http://dx.doi.org/10.4312/slo2.0.2016.1.20-41.

Bikel, D. M., Schwartz, R., & Weischedel, R. M. (1999). Algorithm that learns what's in a name. *Machine Learning*, *34*, 211–231. http://dx.doi.org/10.1023/a:1007558221122.

Chieu, H. L., & Ng, H. T. (2003). Named entity recognition with a maximum entropy approach. In *Proceedings of the seventh conference on natural language learning at HLT-NAACL 2003 - Volume 4* (pp. 160–163). http://dx.doi.org/10.3115/1119176.1119199.

Cho, K., van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, eighth workshop on syntax, semantics and structure in statistical translation* (pp. 103–111). Doha, Qatar: Association for Computational Linguistics, http://dx.doi.org/10.3115/v1/W14-4012.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, *12*, 2493–2537.

Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., & Hu, G. (2020). Revisiting pre-trained models for Chinese natural language processing. In *Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020*. http://dx.doi.org/10.18653/v1/2020.findings-emnlp.58.

Deping, C., Bo, W., Hong, L., Fang, F., & Run, W. (2021). Geological entity recognition based on ELMO-CNN-BILSTM-CRF model. *Geoscience*, *46*, 3039–3048.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. http://dx.doi.org/10.48550/arXiv.1810.04805, arxiv.org.

Fan, R., Wang, L., Yan, J., Song, W., Zhu, Y., & Chen, X. (2019). Deep learning-based named entity recognition and knowledge graph construction for geological hazards. *ISPRS International Journal of Geo-Information*, *9*, 15. http://dx.doi.org/10.3390/ijgi9010015.

Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement. In *Icml 2018*: Vol. 80. (pp. 1861–1870).

Hettne, K. M., Stierum, R. H., Schuemie, M. J., Hendriksen, P. J. M., Schijvenaars, B. J. A., Mulligen, E. M. V., Kleinjans, J., & Kors, J. A. (2009). A dictionary to identify small molecules and drugs in free text. *Bioinformatics*, *25*, 2983–2991. http://dx.doi.org/10.1093/bioinformatics/btp535.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*, 1735–1780.

Lafferty, J. D., McCallum, A., & Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning* (pp. 282–289). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc..

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 260–270). San Diego, California: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N16-1030, https://aclanthology.org/N16-1030.

Li, W., Ma, K., Qiu, Q., Wu, L., Xie, Z., Li, S., & Chen, S. (2021). Chinese word segmentation based on self-learning model and geological knowledge for the geoscience domain. *Earth and Space Science*, *8*, http://dx.doi.org/10.1029/2021EA001673, e2021EA001673.

Li, D., Savova, G., & Kipper-Schuler, K. (2008). Conditional random fields and support vector machines for disorder named entity recognition in clinical texts. In *Proceedings of the workshop on current trends in biomedical natural language processing* (pp. 94–95). Columbus, Ohio: Association for Computational Linguistics, https://aclanthology.org/W08-0615.

Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2020). Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*, 2980–2988. http://dx.doi.org/10.1109/TPAMI.2018.2858826.

Liu, X., Yang, N., Jiang, Y., Gu, L., & Shi, X. (2020). A parallel computing-based deep attention model for named entity recognition. *The Journal of Supercomputing*, *76*, 814–830. http://dx.doi.org/10.1007/s11227-019-02985-5.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Müller, R., Kornblith, S., & Hinton, G. E. (2019). When does label smoothing help? *Advances in Neural Information Processing Systems*, *32*.

Mutinda, F. W., Yada, S., Wakamiya, S., & Aramaki, E. (2021). Semantic textual similarity in Japanese clinical domain texts using BERT. *Methods of Information in Medicine*, *60*, e56–e64. http://dx.doi.org/10.1055/s-0041-1731390.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *NAACL HLT 2018 - 2018 conference of the North American chapter of the association for computational linguistics: Human language technologies - Proceedings of the conference* (pp. 2227–2237). New Orleans, Louisiana: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/n18-1202.

Qiu, Q., Xie, Z., Wu, L., Tao, L., & Li, W. (2019). *Earth Science Informatics*, *12*, 565–579. http://dx.doi.org/10.1007/s12145-019-00390-3.

Radford, A., & Narasimhan, K. (2018). Improving language understanding by generative pre-training.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *2017-December*.

Wang, C., Ma, X., Chen, J., & Chen, J. (2018). Information extraction and knowledge graph construction from geoscience literature. *Computers & Geosciences*, *112*, 112–120. http://dx.doi.org/10.1016/j.cageo.2017.12.007.

Xie, X., Xie, Z., Ma, K., Chen, J., Qiu, Q., Li, H., Pan, S., & Tao, L. (2021). Geological named entity recognition based on BERT and BiGRU-attention-CRF model. *Geological Bulletin of China*, 1–13.

Zhang, D., Xu, H., Su, Z., & Xu, Y. (2015). Chinese comments sentiment classification based on word2vec and SVMperf. *Expert Systems with Applications*, *42*, 1857–1863. http://dx.doi.org/10.1016/j.eswa.2014.09.011.

Zhang, X., Ye, P., Wang, S., & Du, M. (2018). Geological entity recognition method based on deep belief networks. *Yanshi Xuebao*, (Acta Petrologica Sinica) *034*, 343–351.

Zhao, S. (2004). Named entity recognition in biomedical texts using an HMM model. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications* (pp. 87–90).