

Earth and Space Science



RESEARCH ARTICLE

10.1029/2021EA002166

Key Points:

- A named entity recognition model based on BERT-BiGRU-CRF is proposed and a set of detailed experimental tests on domain data set
- Our model is compared in detail with four other mainstream models, and experiments demonstrate that our model obtains better performance
- We share the source code of BERT-BiGRU-CRF and the annotated test data

Correspondence to:

Q. Qiu,
qiuqinjun@cug.edu.cn

Citation:

Lv, X., Xie, Z., Xu, D., Jin, X., Ma, K., Tao, L., et al. (2022). Chinese named entity recognition in the geoscience domain based on BERT. *Earth and Space Science*, 9, e2021EA002166. <https://doi.org/10.1029/2021EA002166>

Received 4 DEC 2021

Accepted 24 JAN 2022

Author Contributions:

Conceptualization: Zhong Xie, Kai Ma, Qinjun Qiu

Data curation: Zhong Xie

Formal analysis: Yongsheng Pan

Investigation: Yongsheng Pan

Methodology: Zhong Xie, Xiangguo Jin, Kai Ma, Qinjun Qiu

Project Administration: Dexin Xu

Resources: Zhong Xie

Software: Xiangguo Jin, Liufeng Tao

Supervision: Dexin Xu

Validation: Liufeng Tao, Qinjun Qiu, Yongsheng Pan



Visualization: Dexin Xu

Writing – original draft: Xia Lv, Qinjun Qiu

© 2022 The Authors. Earth and Space Science published by Wiley Periodicals LLC on behalf of American Geophysical Union.

This is an open access article under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Chinese Named Entity Recognition in the Geoscience Domain Based on BERT

Xia Lv^{1,2}, Zhong Xie³, Dexin Xu⁴, Xiangguo Jin³, Kai Ma⁵, Liufeng Tao³ , Qinjun Qiu^{3,6} , and Yongsheng Pan⁷

¹National Engineering Research Center of Geographic Information System, Wuhan, China, ²Development and Research Center (National Geological Archives of China), Beijing, China, ³School of Computer Science, China University of Geosciences, Wuhan, China, ⁴Wuhan Geomatics Institute, Wuhan, China, ⁵College of Computer and Information Technology, China Three Gorges University, Yichang, China, ⁶Hubei Key Laboratory of Intelligent Geo-Information Processing, China University of Geosciences, Wuhan, China, ⁷Wuhan Zondy Cyber Science & Technology Co., Ltd., Wuhan, China

Abstract Geological reports are frequently used by geologists involved in geological surveys and scientific research to record the results and outcomes of geological surveys. With such a rich data source, a substantial amount of knowledge has yet to be mined and analyzed. This paper focuses on automatically information extraction from geological reports, namely, geological named entity recognition. Geological named entity recognition has an important role in data mining, knowledge discovery and Knowledge graph construction. Existing general named entity recognition models/tools are limited in the domain of geoscience due to the various language irregularities associated with geological text, such as informal sentence structures, several domain-geoscience words, large character lengths and multiple combinations of independent words. We present Bidirectional encoder representations from transformers (BERT)-(Bidirectional gated recurrent unit network) BiGRU- (Conditional random field) CRF, which is a deep learning-based geological named entity recognition model that is designed specifically with these linguistic irregularities in mind. Based on the pretrained language model, an integrated deep learning model incorporating BERT, BiGRU and CRF is constructed to obtain character vectors rich in semantic information through the BERT pretrained language model to alleviate for the lack of specificity of static word vectors (e.g., word2vec) and to improve the extraction capability of complex geological entities. We demonstrate our proposed model by applying it to four test datasets, including a geoscience NER data set from regional geological reports, and by comparing its performance with those of five baseline models.

Plain Language Summary Geological named entity recognition has an important role in information extraction and knowledge discovery. This paper presents BERT-BiGRU-CRF, which is a deep learning-based geological named entity recognition model that is designed specifically with these linguistic irregularities in mind. We hope that our approach will serve as an alternative method that deserves further study.

1. Introduction

Geological reports are rich sources of information. Geological report records geological conditions and geological events within a certain area (Qiu, Xie, Wu, Tao, & Li, 2019; Qiu, Xie, Wu, & Tao, 2020). They cover a variety of topics about geological survey, ranging from ore/deposits, mineral, remote sensing, to geology, hydrology. Many types of entities are also mentioned in various geological reports, which include not only significant locations, but also rocks, minerals, to stratigraphic units, locations, geological timescales (Fan et al., 2020; Qiu, Xie, Wu, Tao, & Li, 2019; Qiu, Xie, Wu, & Tao, 2019; Wang et al., 2021; Wu et al., 2017). Named entity recognition (NER) has received much attention from the academic field and industry for many years (Nieh et al., 2021; Qiu, Xie, Wu, Tao, & Li, 2019; Qiu, Xie, Wu, & Tao, 2019; Zhou et al., 2021). The basis of geological domain knowledge graph construction is named entity recognition, that is, the recognition of specific categories of proper noun entities from unprocessed geological domain texts, and its accuracy directly affects the results of multiple natural language processing techniques in the geological domain (Ma et al., 2020, 2021; Wang et al., 2018).

A substantial amount of effort has been dedicated to recognizing named entities from texts. Early named entity recognition tasks mainly use rule- and lexicon-based methods (Goyal et al., 2018; Rouhou et al., 2021; Santoso

et al., 2021). Such approaches depend on hand-crafted pattern-matching-based rules for supporting the recognition and extraction of target entities from geological textual data. Outside of the geoscience domain, several rule-based NER techniques/models have been proposed (Appelt et al., 1993; Elsebai et al., 2009; Lehnert et al., 1992). As the amount of data increases, the workload of rule extraction increases, the difficulty of maintaining rule consistency increases, and rule-based and dictionary-based methods cannot address the heterogeneity and complexity of text and the thus achieve high GNER performance (Qiu, Xie, Wu, Tao, et al., 2019; Santoso et al., 2021). Compared with rule-based methods, the statistical learning methods can learn from large amount of annotating training datasets to guide the recognition and extraction NER (Liu et al., 2022; Molina-Villegas et al., 2021; Peng et al., 2021). The popular deep learning methods for named entity recognition generally base on word embeddings, which are able to learn similar representations for semantically or functionally similar words (Fang et al., 2021; Santoso et al., 2021; Tian et al., 2021). Researchers started to construct models without complicated feature engineering and to minimize the reliance on NLP toolkits for feature acquisition.

However, the recognition of geological named entities still faces some difficulties and challenges (Qiu, Xie, Wu, Tao, & Li, 2019; Qiu, Xie, Wu, & Tao, 2019): (a) compared with general domain texts, geological named entities have (a) large character lengths, (b) many remote words, (c) multiple combinations of independent words, and (d) nested entities. For example, “garnet diopside silica” indicates a kind of silica with garnet and diopside as the main mineral components, and this geological naming entity contains 10 characters. The entity “garnet diopside silica” includes garnet, diopside, and silica. The geological nomenclature of “garnet diorite” includes garnet, diorite, diorite, and garnet diorite silica. There are nesting relationships among geological naming entities of different conceptual levels; for example, a certain stratigraphic entity contains several rocks, and rocks contain several minerals. Presently, there is a lack of a large-scale annotated corpus in the geological field; there is a lack of samples for training the model; and the recall and accuracy are not sufficient. Therefore, geological named entity recognition becomes a challenging task.

The popular deep learning approaches for NER generally based on word embeddings for learning similar representations for semantically or functionally similar words. One drawback of these approaches is that the embedding for the same word in different semantic sentences are identical. A large-scale annotated corpus is required for model training to prevent parameter underfitting or overfitting, while the production of an annotated corpus is costly, and building a large-scale annotated data set is challenging for most natural language processing tasks. As an advanced and widely employed language expression model BERT (Devlin et al., 2018) extracts bidirectional semantic features of Chinese utterances from large-scale texts by unsupervised learning, further enhances the generalization ability of the word vector model, and can better characterize syntactic and semantic information in different contexts, which can, to a certain extent, free the dependence of supervised learning on large-scale annotation data and effectively reduce the labor intervention of manual annotation. Therefore, the integration of BERT into deep learning models will become a new way to improve the performance of Chinese, geological named entity recognition.

To address the abovementioned issues, in this paper, a character-level embedding-based BERT-BiGRU-CRF model is proposed for extracting Chinese geological named entities in response to the characteristics of named entity recognition tasks in the geological domain and to the problems of small sample size and poor results for some target entities (Note that BERT can be used for other languages, while this paper we focus on Chinese GNER). The model consists of an encoder, a BiLSTM neural network layer and a conditional random field (CRF) layer from the bottom up: the encoder is a character-level, Chinese, BERT-based model that maps the input Chinese entity characters into a low-dimensional, dense real space to mine the potential semantics embedded in Chinese entity elements; the BiLSTM neural network layer takes the character vector transformed by the encoder as input to capture the Chinese entity sequences; the BiLSTM neural network layer takes the character vectors transformed by the encoder as input to capture the forwards (left-to-right) and backwards (right-to-left) bidirectional features of the Chinese entity sequences; and the conditional random field layer takes the upstream BiLSTM extracted bidirectional features as input to generate the labels that correspond to each character of the geological entities in combination with the BIOES annotation specification. The experimental results demonstrate that our presented model outperforms previous models and that our approach achieves state-of-the-art performance on the constructed datasets.

The main contributions of this research are summarized as follows:

1. Based on the characteristics of Chinese geological domain text, we propose BERT-BiGRU-CRF model for guiding the recognition and extraction of target information from unstructured textual data
2. Our model is compared in detail with four other mainstream models, and experiments demonstrate that our model obtains higher performance on both the general domain data set and geological domain data set
3. We share the source code of BERT-BiGRU-CRF and the annotated test data at <https://doi.org/10.5281/zenodo.5758466>

The remainder of the article is structured as follows: Section 2 discusses related work in the area of geological named entity recognition. Section 3 introduces the proposed approach, and Section 4 provides details on the experiments and results. Some concluding remarks are described in Section 5.

2. Related Work

2.1. Text Mining in Geoscience

In the domain of geosciences, various systems and applications, such as mineral exploration (Holden et al., 2019; Shi et al., 2018), paleontological studies (Peters et al., 2014, 2017; Wang et al., 2018), and geological text mining and application in Chinese (Qiu, Xie, Wu, Tao, & Li, 2019; Qiu, Xie, Wu, & Tao, 2019; Wang et al., 2018), have been developed and constructed.

Holden et al. (2019) analyzed 25,419 mineral exploration reports in a targeted manner using NLP pipeline analysis methods (word recognition, keyword extraction, etc.), focusing on searching and summarizing the geology-related content in the reports, and developed a related system named the GeoDocA system. This system retrieves geology-related subject terms based on a dictionary of pre-customized entity groups (geological timescales, mineralogy, host rock types and alteration types), calculates the co-occurrence of these terms, and ultimately creates a summary map for retrieval. Shi et al. (2018) utilized a mature text mining algorithm for direct application to geological text data in Chinese, performed a case study of the Lara copper deposit in Sichuan Province, China, for retrieval, and used convolutional neural networks (CNNs) to classify geological, geophysical, geochemical and remote sensing-related text data for geological texts. Entity co-occurrence and frequency statistics of the deposit analysis were performed. Enkhsaikhan et al. (2018) used a form of word embedding to analyze the semantic similarity among geology-related phrases and applied an analogical solver to establish the semantic relationships among terms to investigate mineral exploration through a semantic analysis approach. The final results demonstrate the potential application of semantic relationships among entities in the domain.

For the domain of paleontological text mining, Peters et al. (2014) developed PaleoDeepDive, a paleontology-oriented knowledge mining platform based on the original platform named DeepDive (De Sa et al., 2016), which was primarily utilized to generate historical and generic turnover rates of taxonomic diversity. In their recent work (Peters et al., 2017), they automated the analysis of stratigraphic databases using a computer reading system to automatically extract information about the three phases of occurrence of stromatolites arranged on a geological time scale, as well as predictors of stromatolite prevalence.

In the domain of geosciences in Chinese, Qiu, Xie, and Wu (2018) proposed a self-learning-based word segmentation method to segment meaningful words from geoscience reports written in Chinese, in response to problems such as the drastic performance degradation of generic domain, word segmentation methods applied to the geological domain. To address the lack of a massive corpus in the Chinese geology domain, Qiu, Xie, Wu, et al. (2018) proposed a method to generate a corpus based on a random combination of word and word frequency information to segment words related to geology by using a BiLSTM model. Wang et al. (2018) segmented mineral deposit domain texts based on CRFs and then used the segmented texts for keyword extraction and co-occurrence word statistics to construct a knowledge graph for visual analysis. Their algorithm was able to segment both generic domain words and geological domain words. Li et al. (2021) constructed a Chinese word segmentation algorithm based on a geological domain ontology assisted by a self-loop approach to better segment geological domain texts. Ma et al. (2021) employed a deep learning model to train journal abstracts and titles in the field of Chinese geology. The resulting model was trained to perform automatic abstract construction.

All of these studies have explored and visualized geological text but have not addressed the more fine-grained information (e.g., NER, keywords and relation extraction) contained in it. Instead, this paper focuses on the

extraction of named entities from geological texts to build a knowledge graph of the geological domain and to mine and discover hidden knowledge in the geological domain.

2.2. Previous Research on Geological Named Entity Recognition

Geological named entity recognition is a domain-specific named entity recognition that aims to identify some important concepts in geology, including geological age, geological formations, stratigraphy, rocks, minerals and locations (Qiu, Xie, Wu, & Tao, 2019). Several scholars have conducted research on geological named entity identification. Zhang et al. (2018) developed a classification system for elements of geological entity information and an annotation specification based on the linguistic features of geological texts and applied a deep belief network model to geological entity information recognition. Ma (2018) combined a geological domain ontology to carry out the named entity recognition task using a mature BiLSTM-CRF model after preprocessing operations of word separation and deactivation for geological texts, such as geological reports, geological domain journal papers and geological professional websites. Qiu, Xie, Wu, and Tao (2019) preprocessed the corpus with reference to the word2vec model using a large amount of unlabeled data to train word embeddings in the geological domain, used a recurrent neural network approach based on BiLSTM with an attention mechanism for semantic encoding of sentences, and finally combined it with conditional random fields to achieve geological entity recognition. Chu et al. (2021) fused multiple methods, such as ELMO (embeddings from language models), CNN, and Bi-LSTM-CRF, to extract geological entities after word vectorization representation using a CNN to add character features and ELMO to extract word dynamic features for input distributed representation. The commonality of these studies is that they take advantage of the ability of deep learning models to learn deep nonlinear features among words to tackle geological named entity recognition tasks.

In this paper, we focus on named entity recognition in the Chinese geological domain, compare the performance of the current mainstream deep learning models on geological named entity recognition, compare the performance of the mainstream models on a standard data set in the general domain and a constructed data set in the domain of geosciences.

3. Proposed Methods

3.1. BERT-BiLSTM-CRF

In the BERT-BiLSTM-CRF model, the BERT model is selected as the feature representation layer for word vector acquisition. The BiLSTM model is employed for deep learning of full-text feature information for specific geological domain named entity recognition. The output sequence of the BiLSTM model is processed in the CRF algorithm layer, and combined with the CRF algorithm, a global optimal sequence is obtained based on the labels between two neighbors, as shown in Figure 1.

3.2. BERT-BiGRU-CRF

The model framework consists of a BERT pretrained language model, BiGRU network, and CRF layer; a diagram of the model architecture is shown in Figure 2. First, the input sequence is input into the BERT layer for pretraining to obtain context-dependent representations, which are utilized to solve the key problem of entity recognition with many rare characters and nested entities. Second, the vectors obtained from the BERT layer are input into the BiGRU layer to solve the problem of long-term text memory and long text dependency, which can be applied to solve the key problem of large-length geological entity characters. Last, decoding is performed by the CRF layer to obtain the output label sequence.

3.3. Basic Components

3.3.1. BERT Layer

The RNNs and CNNs have certain shortcomings when addressing NLP tasks: the circular network structure of RNNs is not parallelized, and training is slow; the inherent convolutional operation of CNNs is not well suited to serialized text. The transformer model (Vaswani et al., 2017) is a new architecture for textual serial networks that is based on a self-attention mechanism, where arbitrary units interact and there is no length limi-

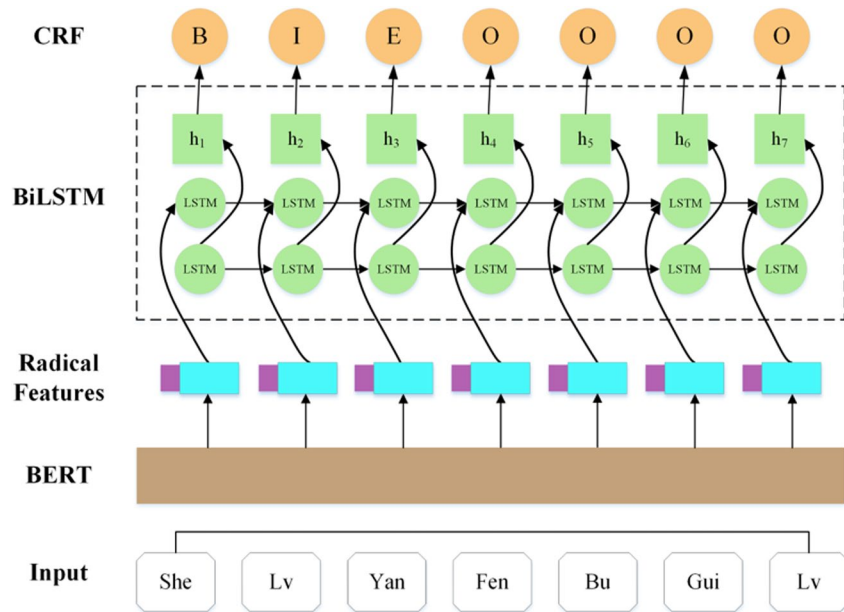


Figure 1. Architecture of BERT-BiLSTM-CRF.

tation problem. The BERT model uses a multilayered bidirectional transformer encoder structure while being constrained by left-right and right-left contexts and is better able to include rich contextual semantic information than the ELMo model (Peters et al., 2018), where left-to-right and right-to-left LSTM connections are trained independently. In addition, the transformer uses positional embedding to add temporal information in response to the problem that the self-attention mechanism cannot extract temporal features. The BERT input representation is stitched by three embeddings, namely, token embedding (word embedding), sentence embedding and positional embedding, which can clearly represent a single text sentence or a pair of text sentences in a token sequence, as shown in Figure 3.

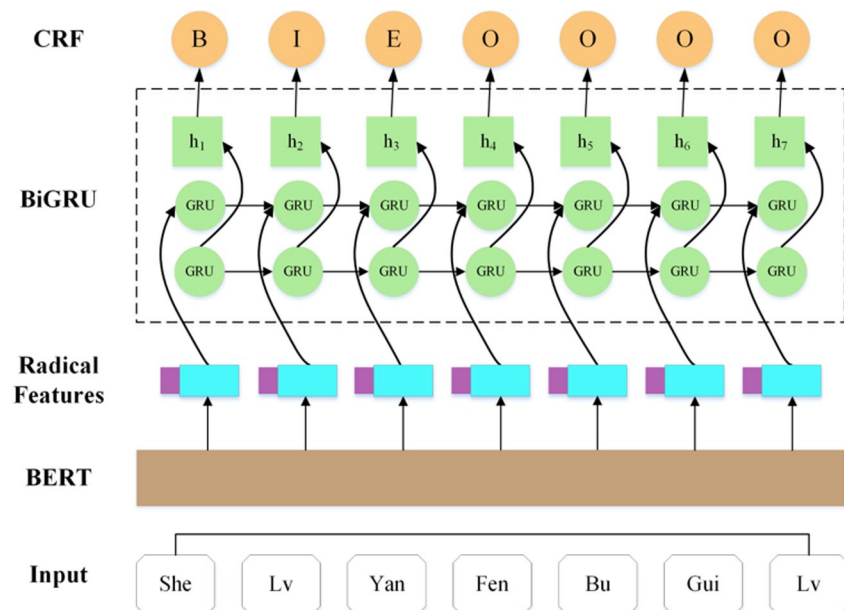


Figure 2. Architecture of BERT-BiGRU-CRF.

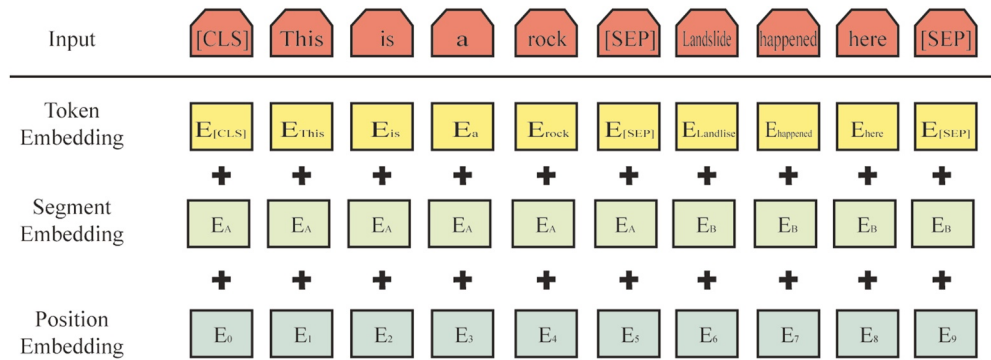


Figure 3. Input embedding representation of BERT.

In addition, the BERT pretrained language model also captures word-level and sentence-level representations through two tasks, the masked language model and next sentence prediction, respectively, and is trained jointly. The masked language model is designed to train a deep bidirectional language representation vector by randomly masking certain words in a sentence and then predicting the masked words. In contrast to standard language models (e.g., word2vec) that can only predict the target function in one direction from left to right or from right to left, masked language models can predict the masked words from any direction.

3.3.2. BiLSTM Layer

LSTM networks were proposed by Hochreiter and Schmidhuber (1997) to solve the phenomenon of gradient disappearance or gradient explosion that occurs in earlier recurrent neural networks after multilayer network propagation; they comprise a special kind of recurrent neural network. LSTM is widely applied in text information processing tasks as it can well capture temporal information and handle information with backwards and forwards dependencies (Chiu & Nichols, 2016). The standard LSTM can only accept antecedent information and only considers the impact of antecedent information on the current moment, disregarding the following information. Considering the close contextual connection of Chinese text, this paper adopts a bidirectional LSTM. BiLSTM is a further development of LSTM by adding a backwards LSTM that combines the forwards hidden layer and the backwards hidden layer to obtain two different vector representations of the input information at the current moment in a recursive operation, combining them as the vector representation of the input information at the current moment, which can access both preceding information and following information. The detailed encoding pattern of BiLSTM is shown in Figure 4.

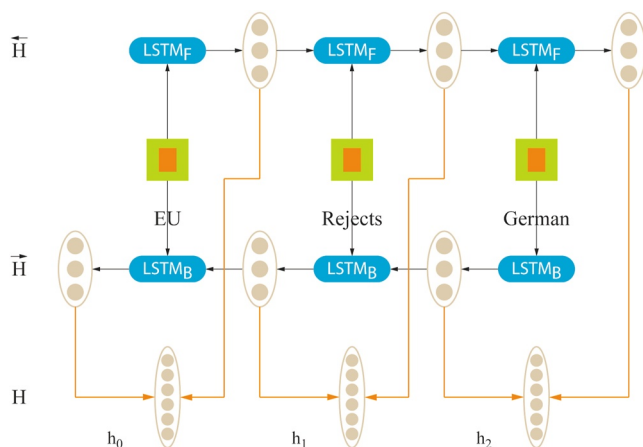


Figure 4. Architecture of BiLSTM.

3.3.3. GRU Layer

A gated recurrent unit (GRU) is designed to solve the problem of RNN, long-term memory and back-propagation gradient disappearance (Cho et al., 2014). The performance effect of a GRU is similar to that of LSTM, and its advantages are reflected in fewer parameters, lower hardware and time cost, and better generalization ability effect on small sample datasets. The internal structure of GRU is shown in Figure 5.

The GRU combines the current node input x^t with the state $h^{(t-1)}$ transmitted from the previous node to derive the output y^t of the current node and the hidden state h^t passed to the next node. The network internal parameter transfers and update equations are shown in (1)–(4).

$$r^t = \sigma \left(w^{rx} x^t + w^{rh} h^{t-1} + b^r \right) \quad (1)$$

$$z^t = \sigma \left(w^{zx} x^t + w^{zh} h^{t-1} + b^z \right) \quad (2)$$

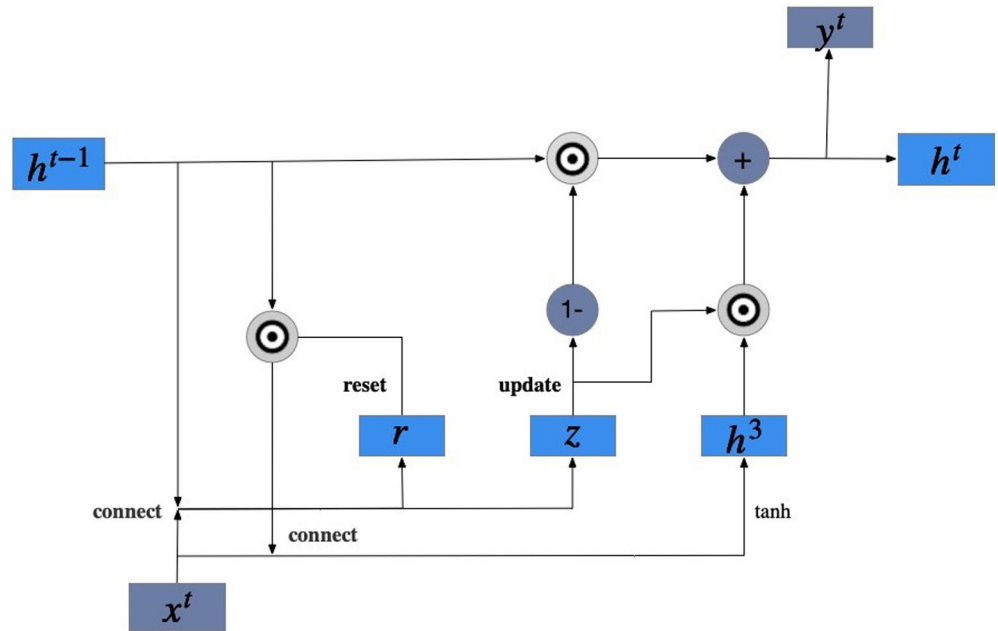


Figure 5. Internal structure of GRU.

$$h' = \tanh(w^{xh} x^t + r^t \odot w^{hh} h^{t-1}) \quad (3)$$

$$h^t = (1 - z^t) \odot h^{t-1} + h' \odot z^t \quad (4)$$

where σ is the sigmoid function, which is utilized as a gating signal. The closer the gating signal is to 1, the more data are remembered, and conversely, the more data are forgotten. r is the gating to control the reset, and z is the gating to control the update. h' refers to the candidate hidden state. w^{rx} , w^{rh} , etc. Are the weight matrices, and b^r , b^z , etc. Are the bias quantities. \odot is the Hadamard product, that is, the corresponding elements in the matrix are multiplied.

3.3.4. Attention Mechanism Layer

GRU can solve the problem of long-term memory to a certain extent and extract global features. However, it is difficult to solve the problem of long-distance dependency in geological text, and it is difficult to retain local detail information in long text. To compensate for the shortcomings of BiGRU in extracting local features (there are many solutions, and in this paper we choose the attention mechanism strategy), this paper introduces the attention mechanism (Bahdanau et al., 2014) to extract the degree of association between different characters in a sentence and the context, which is conducive to solving the long-distance dependency problem caused by the large character length of geological named entities. The attention mechanism adds feature weights to the semantics related to geological named entities to improve the effect of local feature extraction.

The attention mechanism layer assigns weights to the feature vectors h^t output by the previous layer and calculates the common output feature vector c_t of the previous layer and the attention layer at time t .

$$c^t = \sum_{i=1}^n a^{t,i} h^i \quad (5)$$

$$a^{t,i} = \frac{\exp(\text{score}(s^{t-1}, h^i))}{\sum_{i'=1}^n \exp(\text{score}(s^{t-1}, h^{i'}))} \quad (6)$$

$$\text{score}(s^t, h^i) = v \tanh(w[s^t, h^i]) \quad (7)$$

Table 1
Statistical Analysis of the Datasets in Our Work

Data set	^a Word	^b Sentence	Text (%)	Count	Max length	Avg length
MSRA	2,169,756	43,906	3.26	70,673	35	3.300
Boson	542,825	9,930	4.07	22,098	57	3.987
PeopleNER	23,641,001	45,5726	3.17	74,9316	26	2.928
GeoNER2021	1,013,491	16,894	2.61	26,472	57	4.221

Note. Text (%) = the number of entities/number of words in the data set; count = the number of entities on the data set; Max length = the maximum length of the entity on the data set; Avg length = the average length of the entity on the data set.

^aWord = the number of the words on the data set. ^bSentence = the number of sentences on the data set.

where a^{ti} is the attention function. The score function is the alignment model, which assigns scores based on how well the inputs and outputs match at moment i , defining how much weight each output gives to each input hidden state.

3.3.5. CRF Layer

The named entity recognition task can generally be considered a sequence annotation problem, and usually the output of BiLSTM can be applied for sequence annotation by adding a softmax layer to the top layer for judgment and outputting the label with the highest probability to complete the annotation task of the input sequence. Although the BiLSTM solves the problem of contextual linkage, it lacks constraints on the output label information. The softmax layer is based on the judgment of the information at the current moment and does not consider the overall linkage, and the output result is only the optimal solution of the information at the current moment, that is, the local optimal solution. This approach may produce an invalid sequence

of labels, which are sequences of outputs such as “B-ROC, I-TIME...”. In the case of named entity identification tasks, the BIOE method (Lample et al., 2016) is selected for labeling so that B-ROC followed by I-TIME is not possible.

Given the set of random variables X as the observation sequence and the output sequence Y , the CRF model is described using the conditional probability $P(Y/X)$. For a sentence of text, $X = \{x_1, x_2, \dots, x_n\}$ denotes its observation sequence, and for the output label sequence $Y = \{y_1, y_2, \dots, y_n\}$, its score is computed as follows:

$$\text{score}(X, y) = \sum_1^m Q_{i, y_i} + \sum_0^m A_{y_i, y_{i+1}} \quad (8)$$

where Q is a matrix of scores $m \times k$ output by the attention mechanism, where m is the length of the sentence and k is the number of tags with different entity types. $Q_{i, j}$ denotes the score for the j th tag of the i th word. A is a matrix of transferred scores of size $k + 2$, where $A_{y_i, y_{i+1}}$ denotes the scores transferred from label i to label $i + 1$.

Y_X is the sequence of all possible annotations for sentence X . The final decoding is performed by the Viterbi algorithm to obtain the predicted tag sequence with the highest score as follows:

$$y^* = \arg \max(\text{score}(X, y)) \quad (9)$$

4. Experiments and Results

A set of primary experiments are conducted to evaluate the presented approach on the gold NER datasets and GNER datasets. First, we introduce the experimental environment and parameters. Second, we presented the evaluation metrics based on precision, recall and F1-score. Third, we compare a set of models/algorithms with our proposed models. Last, we demonstrate the recognition results by a case study and analyze the illustrative errors.

4.1. Datasets

Our presented approach is evaluated on the MSRA, Boson, PeopleNER and GeoNER2021 datasets, all of which are preprocessed. Among these datasets, GeoNER2021 is the domain of geoscience that was developed by human annotation; the other datasets comprise the gold NER datasets in the generic domain. Table 1 lists the statistics of these datasets.

MSRA: A data set provided by Microsoft Research Asia, with a test set of approximately 50,000 sentences and a training set. The main entities consist of names, place names and organization names.

Table 2
Number Distribution of Geological Entity Categories

Entity type	Training set	Development set	Test set	Count
GTM	645	220	215	1,080
GST	1,284	459	428	2,171
STR	2,850	990	950	4,790
ROC	6,846	2,298	2,282	11,426
MIN	2,040	699	680	3,419
PLA	2,226	780	742	3,748
Average	2,648.5	907.7	882.8	4,439

Boson: This dataset was provided by the Boson Chinese Semantic Open Platform; it includes 2,000 sentences in total. There are 6 entity categories: time, location, person name, organization name, company name and product name.

PeopleNER: This data set is divided into two data files, with a total of approximately 286,000 sentences. The first data set includes the original text after splitting, and the second data set includes the corresponding tags one by one after splitting, with a total of 286,000 sentences. The main entity types include location, organization, person name and time.

GeoNER2021: The data in this paper were obtained from the National Geological Archive of China Geological Survey (NGAC) website, with a total of 43 regional geological survey reports. The collected geological reports were manually labeled into six geological named entity categories GTM, GST, STR, ROC, MIN and PLA.

4.2. Experimental Environment and Parameter Settings

The model was trained and tested in Python 3.7.3 and TensorFlow 1.1. The experiments were conducted using the BERT-Base model, which contains 12 transformer layers, a 768-dimensional hidden layer and a 12-head multi-head attention mechanism. The GRU network has a 128-dimensional hidden layer. The attention mechanism layer is set to 50 dimensions, and the maximum sequence length is set to 256. The optimization function is Adam; the learning rate is set to $5E-5$; and the dropout layer is set to 0.5. The dropout layer is set to 0.5. All models were trained on a single GTX 3090 GPU (Table 2).

4.3. Evaluation Metrics

The performance is measured with the precision (named P), recall (named R) and F1-score. The detailed formula is calculated as follows:

$$P = \frac{TP}{TP + FP} \quad (10)$$

$$R = \frac{TP}{TP + FN} \quad (11)$$

$$F1 = \frac{2PR}{P + R} \quad (12)$$

where TP refers to a positive sample and a positive prediction; FP refers to a negative sample and a positive prediction; and FN refers to a positive sample and a negative prediction.

In this work, each experiment is repeated five times, and we report the average F1-score as the final result (Table 3).

Table 3
Data Set Annotation Examples on the GeoNER2021 Data Set

Ying	An	Yan	Ya	Lei	Chu	Lu	Yu	Qu
B-ROC	I-ROC	E-ROC	O	O	O	O	O	B-STR
Shen	La	Zu	De	Shang	Bu	Ceng	Wei	,
I-STR	I-STR	E-STR	O	O	O	O	O	O
Zhu	Yao	Fen	Bu	Yu	Chong	Chang	,	Yu
O	O	O	O	O	B-PLA	E-PLA	O	B-PLA
Xiong	Deng	Di	。					
E-PLA	O	O	O					

Table 4
Performance of Different Models on the MSRA Data Set

Model	Precision	Recall	F1-score
IDCNN	0.988	0.860	0.871
IDCNN2	0.990	0.872	0.897
BiLSTM-CRF	0.989	0.870	0.863
BiLSTM-Attention-CRF	0.987	0.850	0.848
BERT-BiLSTM-CRF	0.996	0.954	0.950
BERT-BiGRU-CRF	0.997	0.966	0.981

Note. Bolded indicates the best performance indicator, highlighting.

4.4. Comparative With Other Algorithms

We selected six mainstream algorithms for testing on three generic domains, named entity recognition datasets. The experimental results are shown in Table 4, which show that all the deep learning models obtained more than 85% performance on these datasets. Among them, BERT-BiGRU-CRF obtained the highest performance on all three datasets, with F1 values of 98.1%, 99%, and 99.1%. This finding further demonstrates that the model has a good generalization capability.

Also, six comparative experiments are conducted to validate the different deep learning-models on the Boson data set, which is a relatively small test data set, and the result is reported in Table 5. As seen in Table 5, the algorithm that achieves the best performance among all the models is BERT-BiGRU-CRF. As we expected, BERT-BiGRU-CRF has good ability to generalize

for identifying named entity. In addition, increasing comprehensive and representative gold data set can help BERT-BiGRU-CRF obtain better performance, and further illustrating the ability of BERT-BiGRU-CRF model to obtain higher performance even on small sample datasets.

Further, a series of experiments were designed to validate the performance of the six algorithms on the PeopleNER data set that contains a large number of named entities. Table 6 demonstrates the results. BERT-BiGRU-CRF achieves the best performance, reaching a precision, recall and F1-score of 0.981, 0.983 and 0.982, respectively. This again validates that the model exhibits superior performance compared to mainstream deep learning models and is capable of better recognition in large scale corpora.

We compared six mainstreams, named entity recognition models on the geological named entity recognition data set that we constructed. The experimental results, which are shown in Table 7, suggest that the BERT-BiGRU-CRF model achieves the best performance among all deep learning-based models on the GEONER2021 data set. The second best model is the BERT-BiLSTM-CRF model. The experimental results demonstrate that the BERT model has better input sequence characterization, while the GRU module has better modeling results than BiLSTM.

The above three experimental results show that the BERT model can fully extract word-level, word-level and sentence-level features, and the pre-trained word vectors can better characterize the contextual semantics and enhance the generalization ability of the model, especially for the recognition of domain entities with small data set size, and also for the types of geological entities that consist of Chinese and numeric combinations, it can avoid the problem of word separation errors affecting the recognition of entities. It is also possible to avoid the problem that entity recognition is affected by word separation errors. Among the experiments, it is also found that the training time has increased significantly after incorporating the BERT model.

4.5. Impact of the Size of the Training Data Set

A set of primary experiments was conducted to determine how many training data from the corpus could affect the GNER model performance, so an optimal number of training data could be determined and utilized for improved performance. A total of 45 controlled experiments were conducted with training data whose proportion ranged from 10% to 100% (the step size for increasing is 10%). The experimental results, in terms of average F1-score, are demonstrated in Figure 6.

The experimental results show that the number of training datasets can improve the GNER performance. As shown in Figure 6, increasing the number of training datasets can increase the average F1-score. For instance, the BERT-BiGRU-CRF model achieved the least satisfactory performance with a 10% training data set, and increasing this proportion to 90% improved the average F1-score by 72%. Although increasing the number of training datasets could improve the performance of the model, optimal GNER performance cannot be achieved if the number of training datasets is solely increasing. For instance, when the number of training datasets was set to 100%, the

Table 5
Performance of Different Models on the Boson Data Set

Model	Precision	Recall	F1-score
IDCNN	0.628	0.704	0.661
IDCNN2	0.639	0.76	0.691
BiLSTM-CRF	0.565	0.554	0.559
BiLSTM-Attention-CRF	0.595	0.694	0.636
BERT-BiLSTM-CRF	0.814	0.829	0.819
BERT-BiGRU-CRF	0.821	0.832	0.826

Note. Bolded indicates the best performance indicator, highlighting.

Table 6
Performance of Different Models on the PeopleNER Data Set

Model	Precision	Recall	F1-score
IDCNN	0.962	0.967	0.964
IDCNN2	0.966	0.969	0.968
BiLSTM-CRF	0.942	0.968	0.954
BiLSTM-Attention-CRF	0.948	0.934	0.941
BERT-BiLSTM-CRF	0.972	0.977	0.974
BERT-BiGRU-CRF	0.981	0.983	0.982

Note. Bolded indicates the best performance indicator, highlighting.

average F1-score dropped by 2% compared to the best performance (demonstrated with a 90% proportion).

Figure 6 also demonstrates that the BERT-BiGRU-CRF is sensitive to relatively minor changes in the number of training datasets, with a range of 10%–90%. For instance, increasing the number of training datasets from 50% to 90% did affect the average F1-score. Based on the experimental results, the optimal number of training datasets ranges between 80% and 90%.

4.6. Impact of Different Architectures

To evaluate the effects of different features and components on the overall performance of the presented methods, we conducted an additional set of experiments to assess the variants of BERT-BiGRU-CRF without different

features and components, as illustrated in Table 8. These results demonstrate that each enhancement incorporated into the proposed model contributes to the improvement of the overall performance. The following model settings were evaluated:

1. The entire architecture (BERT-BiGRU-CRF), as introduced in Section 3.2
2. The entire architecture proposed in this study with a word embedding of low dimensionality (i.e., vectors of 150 dimensions instead of 200)
3. The entire architecture proposed in this study with a word embedding of high dimensionality (i.e., vectors of 300 dimensions instead of 200)
4. The entire architecture proposed in this study apart from the BERT representation layer
5. The entire architecture proposed in this study is separated from the CRF layer

As is evident from Table 8, all aforementioned variants of BERT-BiGRU-CRF outperformed the traditional matching method, although slight differences were observed in scores obtained via the various approaches. The first two variants were assessed to determine the effect of lower or higher dimensionalities of the word embeddings in the architecture. The results demonstrate that the quality of the results depends on dimensionality. An appropriate dimension setting of the word embedding is necessary for this DNN model, as the use of a higher dimensionality requires a greater number of computational resources, producing slightly inferior results. The two final variants were assessed to determine the effects of specific layers within the proposed model. As expected, in either case, the performance of the model was slightly degraded in terms of each metric, indicating the important impact of the two layers on the performance of the overall model.

4.7. Extraction Results and Error Analysis

In this research, a set of experiments is conducted to validate the extracted NER results and analyze the errors. Some illustrative examples of geological NER results are demonstrated in Table 9. As shown in Table 9, basic geological nomenclature entities, such as the stratigraphic unit entities “Upper Cretaceous” and “Neoproterozoic”

and rock entity “red molasses”, can be effectively identified. In addition, the model can accurately identify the long entity with the remote geographic nomenclature “Nima County Zhang'en-Shenzha County Kargol”, which is a nested entity composed of several independent words: Nima County, Zhang'en, Shenzha County, and Kargol. The result of identifying Nima County, Zhang'en-Shenzha County, and Kargol as geological named entities is output directly.

We also summarized and analyzed the model recognition errors, as shown in Table 10, and discovered the following main problems: (a) We cannot accurately recognize some consecutive entity characters separated by symbols only. For example, in the entity “Raja-Kanru Fault”, “Raja” is not recognized as the character of the geological structure, but the “-” character followed by “Kanru” is employed as the starting character of this geological structure.

Table 7
Performance of Different Models on the GeoNER2021 Data Set

Model	Precision	Recall	F1-score
IDCNN	0.63	0.60	0.60
IDCNN2	0.64	0.66	0.64
BiLSTM-CRF	0.54	0.58	0.55
BiLSTM-Attention-CRF	0.63	0.63	0.61
BERT-BiLSTM-CRF	0.71	0.61	0.64
BERT-BiGRU-CRF	0.75	0.68	0.71

Note. Bolded indicates the best performance indicator, highlighting.

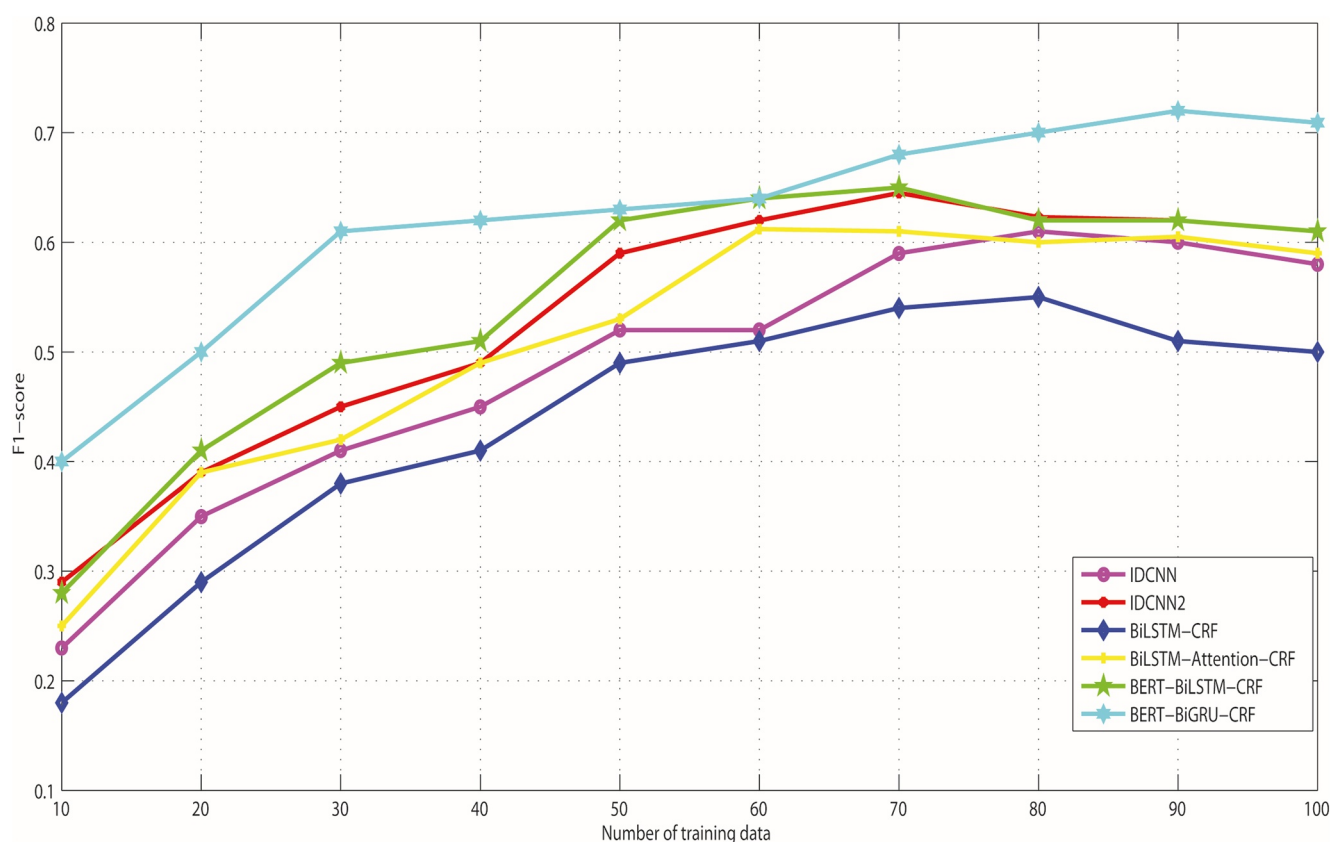


Figure 6. F1-score on the test set in terms of the size of the training data set on the GeoNER2021 data set.

(b) The model only recognizes local information. For example, the rock entity of “medium basal volcanic rock” only recognizes “volcanic rock”, and “medium basal” is marked as other characters.

The solution for Problem 1 can be constructed by constructing regular expressions to judge the entities on the left and right sides of the symbol “-”. If both sides match the entity type, then the entity is considered as a whole. For Problem 2, regular expressions can also be employed to judge and fuse the lexical properties before identifying the entity, and of course, it is also possible to improve the generalization ability of the model by extending the data set.

5. Conclusions and Future Work

Geological named entity recognition is the basis step for acquiring information and extracting knowledge from massive geological reports or documents and enables further relationship extraction and construction of geological knowledge graphs. In this paper, we study and compare the mainstream, deep learning-based, named entity

Table 8

Values of the Metrics Corresponding to Different Variants of the Proposed DNN Architecture

Method	Accuracy	Precision	Recall	F1-score
BERT-BiGRU-CRF	0.724	0.799	0.669	0.728
Word-embedding of dimensionality 150	0.687	0.701	0.61	0.652
Word-embedding of dimensionality 300	0.655	0.681	0.609	0.643
Without BERT representation	0.632	0.658	0.59	0.62
Without CRF layer	0.714	0.744	0.621	0.68

Note. Bolded indicates the best performance indicator, highlighting.

Table 9
Illustrative Examples of Recognition Results

No.	Sentence	Named entity ground truth	Prediction label
1	Upper Cretaceous red millstone construction and Neoproterozoic intermountain basin complex land debris construction	Upper Cretaceous- STR red millstone- ROC Neoproterozoic- STR	Upper Cretaceous- STR red millstone- ROC Neoproterozoic- STR
2	Three sections of the Devonian system are exposed in the areas of Zhang'en-Shenzha County, Kargol, and Nima County	Devonian system- STR Nima County- PLA Zhang'en-Shenzha County Kargol- PLA	Devonian system- STR Nima County- PLA Zhang'en-Shenzha County Kargol- PLA
3	Further delineated by the Raja-Kanru Fracture	Raja-Kanru Fracture- GST	Raja-Kanru Fracture- GST
4	Primary rock is clastic-medium basal volcanic rock	Primary rock- ROC clastic-medium basal volcanic rock- ROC	Primary rock- ROC clastic-medium basal volcanic rock- ROC

recognition methods in the geological domain to address the current challenges of named entity recognition in the geological domain and the descriptive characteristics of geological domain texts. An annotated corpus for geological domain, named entity recognition is constructed, and a named entity recognition model for geological domain texts based on the BERT, pretrained language model is proposed. We applied the proposed BERT-BiGRU-CRF model in four different datasets, and evaluated its performances in terms of precision, recall, and F1-score. The experimental results show that the method significantly outperforms the baseline model and other deep learning models in terms of performance evaluation metrics, such as precision, recall and F1-score for text named entity recognition in the geological domain.

The contributions of this research can be seen from two perspectives. From the perspective of methodology, this work presents a deep learning model (named BERT-BiGRU-CRF) approach. As indicated by the experiment results, the BERT-BiGRU-CRF has a better performance in extracting named entity than other deep learning models. From the perspective of application, this paper develops domain-geosciences NER data set for extracting geological NER and further knowledge discover.

Future research will focus on two aspects: (a) the geological domain corpus is an open domain corpus, and the construction of the corpus is a process of continuous updating and improvement. We will further improve and optimize the geological domain named entity annotation corpus, expand the classification of geological domain named entities, and enrich the corpus by various means such as offline collection; (b) research advanced deep learning models, explore the application of new deep learning models in the geological domain, named entity recognition task, optimize the existing model structure, design a more suitable named entity recognition model for the geological domain, and obtain better performance.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Table 10
Illustrative Examples of Recognition Error Results

No.	Sentence	Named entity ground truth	Prediction label
1	Strongly slippery petrified serpentized pyroxene peridotite	Strongly slippery - ROC petrified serpentized pyroxene peridotite- ROC	Strongly slippery petrified serpentized pyroxene peridotite- ROC
2	Mainly distributed in Nari Gonma - Ziqu south of the East Foot Chung - East Mo Chung	Nari Gonma - PLA Ziqu- PLA East Foot Chung - PLA East Mo Chung - PLA	Nari Gonma - Ziqu- PLA East Foot Chung - East Mo Chung - PLA

Data Availability Statement

We share the annotated test data at <https://doi.org/10.5281/zenodo.5758466>.

Acknowledgments

This study was financially supported by the National Natural Science Foundation of China (42050101, 41871311, 41871305), the China Postdoctoral Science Foundation (No.2021M702991), Wuhan Multi-Element Urban Geological Survey Demonstration Project (WHDYS-2020-004), Open Research Project of The Hubei Key Laboratory of Intelligent Geo-Information Processing (No. KLIGIP-2021A01), and the Fundamental Research Funds for the Central Universities, China University of Geosciences (Wuhan; No. CUG2106116).

References

- Appelt, D. E., Hobbs, J. R., Bear, J., Israel, D., & Tyson, M. (1993). Fastus: A finite-state processor for information extraction from real-world text[C]/IJCAI. 93, 1172–1178.
- Bahdanau, D., Cho, K., & Bengio, Y. (2014). *Neural machine translation by jointly learning to align and translate*[J]. arXiv preprint arXiv:1409.0473.
- Chiu, J. P. C., & Nichols, E. (2016). Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4, 357–370. https://doi.org/10.1162/tacl_a_00104
- Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv preprint arXiv:1406.1078.
- Chu, D. P., Wan, B., Li, H., Fang, F., & Wang, R. (2021). Geological entity recognition based on ELMO-CNN-BiLSTM-CRF model. *Earth Science India*, 46(8), 3039–3048. <https://doi.org/10.3799/dqkx.2020.309>
- De Sa, C., Ratner, A., Ré, C., Shin, J., Wang, F., Wu, S., & Zhang, C. (2016). Deepdive: Declarative knowledge base construction[J]. *ACM SIGMOD Record*, 45(1), 60–67. <https://doi.org/10.1145/2949741.2949756>
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805.
- Elsebai, A., Meziane, F., & Belkredim, F. Z. (2009). A rule based persons names Arabic extraction system[J]. *Communications of the IBIMA*, 11(6), 53–59.
- Enkhsaikhan, M., Liu, W., Holden, E. J., & Duuring, P. (2018). Towards geological knowledge discovery using vector-based semantic similarity. In *International Conference on Advanced Data Mining and Applications* (pp. 22–237). Springer, Cham.
- Fan, J., Shen, S., Erwin, D. H., Sadler, P. M., MacLeod, N., Cheng, Q.-m., et al. (2020). A high-resolution summary of Cambrian to Early Triassic marine invertebrate biodiversity[J]. *Science*, 367(6475), 272–277. <https://doi.org/10.1126/science.aax4953>
- Fang, Z., Zhang, Q., Kok, S., Li, L., Wang, A., & Yang, S. (2021). *Referent graph embedding model for name entity recognition of Chinese car reviews*[J]. Knowledge-Based Systems. 107558. <https://doi.org/10.1016/j.knosys.2021>
- Goyal, A., Gupta, V., & Kumar, M. (2018). Recent named entity recognition and classification techniques: A systematic review[J]. *Computer Science Review*, 29, 21–43. <https://doi.org/10.1016/j.cosrev.2018.06.001>
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Holden, E. J., Liu, W., Horrocks, T., Wang, R., Wedge, D., Duuring, P., & Beardsmore, T. (2019). GeoDocA—fast analysis of geological content in mineral exploration reports: A text mining approach[J]. *Ore Geology Reviews*, 111, 102919. <https://doi.org/10.1016/j.oregeorev.2019.05.005>
- Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). *Neural architectures for named entity recognition*. arXiv preprint arXiv:1603.01360.
- Lehnert, W., Cardie, C., Fisher, D., McCarthy, J., Riloff, E., & Soderland, S. (1992). *University of Massachusetts: MUC-4 test results and analysis*[R]. Massachusetts Univ Amherst Dept of Computer Science.
- 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[J]. *Earth and Space Science*, 8(6), e2021EA001673. <https://doi.org/10.1029/2021ea001673>
- Liu, P., Guo, Y., Wang, F., & Li, G. (2022). Chinese named entity recognition: The state of the art[J]. *Neurocomputing*, 473, 37–53. <https://doi.org/10.1016/j.neucom.2021.10.101>
- Ma, K. (2018). *Research on the key technology of geological big data representation and correlation*[D]. China University of Geosciences.
- Ma, K., Tian, M., Tan, Y., Xie, X., & Qiu, Q. (2021). What is this article about? Generative summarization with the BERT model in the geosciences domain[J]. *Earth Science India*, 1–12. <https://doi.org/10.1109/tcss.2021.3088506>
- Ma, X., Ma, C., & Wang, C. (2020). A new structure for representing and tracking version information in a deep time knowledge graph[J]. *Computers & Geosciences*, 145, 104620. <https://doi.org/10.1016/j.cageo.2020.104620>
- Molina-Villegas, A., Muñoz-Sánchez, V., Arreola-Trapala, J., & Alcántara, F. (2021). Geographic named entity recognition and disambiguation in Mexican news using word embeddings[J]. *Expert Systems with Applications*, 176, 114855. <https://doi.org/10.1016/j.eswa.2021.114855>
- Nieh, E. H., Schottorf, M., Freeman, N. W., Low, R. J., Lewallen, S., Koay, S. A., et al. (2021). Geometry of abstract learned knowledge in the hippocampus[J]. *Nature*, 1–5.
- Peng, Q., Zheng, C., Cai, Y., Wang, T., Xie, H., & Li, Q. (2021). Unsupervised cross-domain named entity recognition using entity-aware adversarial training[J]. *Neural Networks*, 138, 68–77. <https://doi.org/10.1016/j.neunet.2020.12.027>
- Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., et al. (2018). *Deep contextualized word representations*[J]. arXiv preprint arXiv:1802.05365.
- Peters, S. E., Husson, J. M., & Wilcots, J. (2017). The rise and fall of stromatolites in shallow marine environments[J]. *Geology*, 45(6), 487–490. <https://doi.org/10.1130/g38931.1>
- Peters, S. E., Zhang, C., Livny, M., & Ré, C. (2014). A machine reading system for assembling synthetic paleontological databases[J]. *PLoS One*, 9(12), e113523. <https://doi.org/10.1371/journal.pone.0113523>
- Qiu, Q., Xie, Z., & Wu, L. (2018). A cyclic self-learning Chinese word segmentation for the geoscience domain[J]. *Geomatica*, 72(1), 16–26. <https://doi.org/10.1139/geomat-2018-0007>
- Qiu, Q., Xie, Z., Wu, L., & Li, W. (2018). DGeoSegmenter: A dictionary-based Chinese word segmenter for the geoscience domain[J]. *Computers & Geosciences*, 121, 1–11. <https://doi.org/10.1016/j.cageo.2018.08.006>
- Qiu, Q., Xie, Z., Wu, L., & Tao, L. (2019). Gner: A generative model for geological named entity recognition without labeled data using deep learning[J]. *Earth and Space Science*, 6(6), 931–946. <https://doi.org/10.1029/2019ea000610>
- Qiu, Q., Xie, Z., Wu, L., & Tao, L. (2020). Automatic spatiotemporal and semantic information extraction from unstructured geoscience reports using text mining techniques[J]. *Earth Science India*, 13(4), 1393–1410. <https://doi.org/10.1007/s12145-020-00527-9>
- Qiu, Q., Xie, Z., Wu, L., Tao, L., & Li, W. (2019). BiLSTM-CRF for geological named entity recognition from the geoscience literature[J]. *Earth Science India*, 12(4), 565–579. <https://doi.org/10.1007/s12145-019-00390-3>
- Rouhou, A. C., Dhiaf, M., Kessentini, Y., & Salem, S. B. (2021). *Transformer-based approach for joint handwriting and named entity recognition in historical document*[J]. Pattern Recognition Letters. <https://doi.org/10.1016/j.patrec.2021.11.010>

- Santoso, J., Setiawan, E. I., Purwanto, C. N., Yuniarno, E. M., Hariadi, M., & Purnomo, M. H. (2021). Named entity recognition for extracting concept in ontology building on Indonesian language using end-to-end bidirectional long short term memory[J]. *Expert Systems with Applications*, 176, 114856. <https://doi.org/10.1016/j.eswa.2021.114856>
- Shi, L., Jianping, C., & Jie, X. (2018). Prospecting information extraction by text mining based on convolutional neural networks—a case study of the Lala copper deposit, China[J]. *IEEE access*, 6, 52286–52297. <https://doi.org/10.1109/access.2018.2870203>
- Tian, Y., Sun, X., Yu, H., Li, Y., & Fu, K. (2021). Hierarchical self-adaptation network for multimodal named entity recognition in social media[J]. *Neurocomputing*, 439, 12–21. <https://doi.org/10.1016/j.neucom.2021.01.060>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). *Attention is all you need*[C]//Advances in neural information processing systems (pp. 5998–6008).
- Wang, C., Hazen, R., Cheng, Q., Stephenson, M., Zhou, C., Fox, P., et al. (2021). The deep-time digital Earth program: Data-driven discovery in geosciences. *National Science Review*, 8. <https://doi.org/10.1093/nsr/nwab027>
- Wang, C., Ma, X., Chen, J., & Chen, J. (2018). Information extraction and knowledge graph construction from geoscience literature[J]. *Computers & Geosciences*, 112, 112–120. <https://doi.org/10.1016/j.cageo.2017.12.007>
- Wu, L., Xue, L., Li, C., Lv, X., Chen, Z., Jiang, B., et al. (2017). A knowledge-driven geospatially enabled framework for geological big data[J]. *ISPRS International Journal of Geo-Information*, 6(6), 166. <https://doi.org/10.3390/ijgi6060166>
- Zhang, X. Y., Ye, P., Wang, S., & Du, M. (2018). Geological entity recognition method based on Deep Belief Networks. *Acta Petrologica Sinica*, 34(2), 343–351.
- Zhou, C., Wang, H., Wang, C., Hou, Z., Zheng, Z., Shen, S., et al. (2021). Geoscience knowledge graph in the big data era. *Science China Earth Sciences*, 64(7), 1105–1114. <https://doi.org/10.1007/s11430-020-9750-4>