Journal of Earth Science, *Vol. 35*, *No. 3*, *p. 1035−1043*, *June 2024* *ISSN 1674-487X*
Printed in China
https://doi.org/10.1007/s12583-023-1944-8

Low Resource Chinese Geological Text Named Entity Recognition Based on Prompt Learning

Hang He[1, 2], **Chao Ma**[*1, 2], **Shan Ye**[3], **Wenqiang Tang**[1, 2], **Yuxuan Zhou**[1, 2], **Zhen Yu**[1, 2], **Jiaxin Yi**[1, 2], **Li Hou**[1, 2], **Mingcai Hou**[1, 2]

1. *State Key Laboratory of Oil and Gas Reservoir Geology and Exploitation*, *Institute of Sedimentary Geology*, *College of Computer Science and Cyber Security*, *Chengdu University of Technology*, *Chengdu* 610059, *China*
2. *Key Laboratory of Deep-Time Geography and Environment Reconstruction and Applications of Ministry of Natural Resources*, *Chengdu University of Technology*, *Chengdu* 610059, *China*
3. *School of Information Engineering*, *China University of Geosciences* (*Beijing*), *Beijing* 100083, *China*

Hang He: https://orcid.org/0000-0002-9742-7814; Chao Ma: https://orcid.org/0000-0002-0122-5696

ABSTRACT: Geological reports are a significant accomplishment for geologists involved in geological investigations and scientific research as they contain rich data and textual information. With the rapid development of science and technology, a large number of textual reports have accumulated in the field of geology. However, many non-hot topics and non-English speaking regions are neglected in mainstream geoscience databases for geological information mining, making it more challenging for some researchers to extract necessary information from these texts. Natural Language Processing (NLP) has obvious advantages in processing large amounts of textual data. The objective of this paper is to identify geological named entities from Chinese geological texts using NLP techniques. We propose the RoBERTa-Prompt-Tuning-NER method, which leverages the concept of Prompt Learning and requires only a small amount of annotated data to train superior models for recognizing geological named entities in low-resource dataset configurations. The RoBERTa layer captures context-based information and longer-distance dependencies through dynamic word vectors. Finally, we conducted experiments on the constructed Geological Named Entity Recognition (GNER) dataset. Our experimental results show that the proposed model achieves the highest F1 score of 80.64% among the four baseline algorithms, demonstrating the reliability and robustness of using the model for Named Entity Recognition of geological texts.

KEY WORDS: Prompt Learning, Named Entity Recognition (NER), low resource geological text, text information mining, big data, geology.

0 INTRODUCTION

Scientific research has advanced from the traditional scientific research paradigm based on experiment, theory, and simulation to the fourth scientific research paradigm driven by big data (Kitchin, 2014). In the field of geoscience, a large-scale data-driven research mode has emerged, facilitated by various structured thematic databases, such as PetDB (Lehnert et al., 2000), EarthChem (Walker et al., 2005), GeoChron (Bowring et al., 2015), Paleobiology Database (Peters and McClennen, 2016), LiPD (McKay and Emile-Geay, 2016), Macrostrat (Peters et al., 2018), Neotoma (Williams et al., 2018), StraboSpot (Walker et al., 2019), and Sparrow (Ye et al., 2023; Quinn et al., 2021). The emergence of these geological databases has provided new opportunities for geological research and supported new field workflows (Shipley and Tikoff, 2019; Vieira et al., 2014). However, due to the lack of a mature and unified framework for data extraction, there is still a large amount of data hidden in the literature, causing the waste of data or the unrealizing of existing datasets (Yan et al., 2020). Therefore, some geology-related long-standing hypotheses and debates are still unsolved due to the lack of data or the unrealizing of existing datasets (e.g., Raja et al., 2022; Ye, 2022; Chan et al., 2016; Cutcher-Gershenfeld et al., 2016). It has been recognized that the massive size of geoscientific data brings potentials and serves as a solid foundation for further geoscience studies (Zhu et al., 2023). Therefore, in the era of big data, geoscience researchers collect as much data as possible from geological reports and papers to advance geoscience exploration.

Geological reports are crucial sources of information for understanding Earth's structure, composition, and history. However, extracting useful information from these reports is a challenging and time-consuming task due to the high complexi-

*Corresponding author: machao@cdut.edu.cn

Manuscript received July 14, 2023.
Manuscript accepted September 13, 2023.

He Hang, Ma Chao, Ye Shan, Tang Wenqiang, Zhou Yuxuan, Yu Zhen, Yi Jiaxin, Hou Li, Hou Mingcai, 2024. Low Resource Chinese Geological Text Named Entity Recognition Based on Prompt Learning. *Journal of Earth Science*, 35(3): 1035–1043. https://doi.org/10.1007/s12583-023-1944-8. http://en.earth-science.net

ty of geological terminology and the high diversity of textual data required for analysis (Qiu et al., 2023). The increasing volume of geological big data is transforming the way geoscientific research is conducted, creating new opportunities as well as challenges in the field of geosciences. To address this challenge, text mining is an effective way to collect a large amount of valid data. This process involves several key stages, including data acquisition, text pre-processing, analysis, and visualization (Wang et al., 2022).

Recent advancements in Artificial Intelligence (AI) have enabled the automatic processing and analysis of large volumes of text using Natural Language Processing (NLP) techniques to extract useful information or knowledge from unstructured text data in order to obtain large datasets and reveal meaningful insights (Allahyari et al., 2017).

Big data have many applications in various domains of geology. In the field of paleobiology, for example, large datasets have been immensely helpful in understanding the history of biodiversity and macroevolution patterns of organisms. To effectively compile the vast amount of relevant paleobiological data sourced from literature, over 150 scientists and their students have collaboratively established the Paleobiology Database (PBDB) over a period of about 30 years, making it one of the most successful databases in paleobiology (Peters and McClennen, 2016; Peters, 2014; Peters et al., 2014). However, recent studies have revealed that data in the PBDB only cover a portion of available literature (Raja et al., 2022). Many more publications containing fossil data, particularly those from regions outside of North America and western Europe, as well as those written in non-English languages, have not been well processed during data collection, and as a result, macroevolutionary studies based on PBDB may introduce potential biases in fossil sampling (Ye and Peters, 2023; Raja et al., 2022). Likewise, the Macrostrat database relies heavily on stratigraphic columns and maps found in geological reports (Peters et al., 2018). Its coverage is primarily limited to North America and New Zealand, representing only about 15% of the continental crust, and this is mainly due to the challenges associated with accessing regional geological knowledge in other parts of the world (Peters and Husson, 2018). Similar situations may be widespread in other geoscientific datasets or databases, highlighting the importance of establishing a method to access geological knowledge from different regions globally, especially non-English-speaking regions.

The open database can meet the needs of geological researchers in a part of hotspots, but it cannot solve the problem of data needs in different languages, different regions and different subdivisions (Wang et al., 2023). Then in the traditional way, it is very time-consuming and labor-intensive for experts to extract by making rules and manual assistance, so we need to use NLP technology to automatically extract useful entities from the text by understanding geological natural language text (Guo et al., 2021). Named Entity Recognition (NER) is a technique for categorizing entities into predefined groups, aiming at extracting key entities from unstructured text. Commonly named entities are diverse in form, semantically divergent, and have blurred contextual boundaries, such as Apple, which in the field of electronic devices represents a smartphone produced by Apple, while it could also represent a fruit (Fan et al., 2019). Geological Named Entity Recognition (GNER) techniques are essential for extracting proprietary geological entities (Consoli et al., 2020), which are the basis for extracting geological relationships and constructing knowledge maps (Lü et al., 2022). The NER methods for different languages can perform very differently, so non-English NER in particular deserves the attention of the geological community to invest in. In this paper, we use four Chinese regional surveys to extract the geological entity data inside by NER techniques.

Geological Named Entity Recognition technology is essential for extracting proprietary geologic entities such as geologic features, lithologies, deposits and stratigraphic units from unstructured text (Enkhsaikhan et al., 2021). These entities are identified and classified by computer technology to create structured data used to support various geological analyses and form a structured knowledge map of geology (Chu et al., 2021). This paper investigates the application of NER with a few-sample geological text dataset, focusing on its ability to extract information relevant to mineral exploration. Specifically, Prompt Learning can activate large language models with prior knowledge, exploit the rich contextual information capability of pre-trained models without changing the overall model structure (Liu et al., 2023), stimulate the gap-filling capability of pre-trained models, and fully utilize the effective knowledge of small-sample data, thus obtaining a better contextual representation capability under low-resource datasets. This study shows in detail how Prompt Learning can be used to extract entity related to deposits, rock types and geological structures from geological reports and compare it with traditional deep learning methods.

The paper presents the following contributions: Introducing a RoBERTa-Prompt-NER model for Geological Named Entity Recognition task applicable to the Chinese language. The model is based on Prompt Learning which enables it to recognize specific characteristics of geological named entities. Our proposed model shows better performance than traditional deep learning methods as it leverages pre-trained language models and prior knowledge. The proposed model learns from low-resource annotated data and can identify geological named entities.

1 RELATED WORK

1.1 Geological Information Extraction

In the field of geosciences, text mining techniques are becoming increasingly popular for tasks that support data analysis and visualization. Geoscientists recognize the potential of natural language processing (NLP) techniques for their research projects (Holden et al., 2019). Information extraction is a set of NLP techniques that extract factual information, such as entities, relationships, and events, from natural language text and finally structured data output (Piskorski and Yangarber, 2013). In particular, it is used for Geological Named Entity Recognition.

1.2 Geological Named Entity Recognition

The Geological Named Entity Recognition (GNER) task involves using Natural Language Processing (NLP) techniques to extract geological information from unstructured geological

texts to support data analysis and interpretation. The task aims to identify and classify relevant entities such as geological places, geological time, mineral names, and geological formations (Qiu et al., 2019). Named entity recognition is gaining traction in the field of geology, particularly with the proposal of the Deep-Time Digital Earth (DDE) Big Science Program (https://www. ddeworld. org) which combines digital technology with Earth system science to establish a comprehensive understanding of the Earth system for sustainable development. Despite being a considerable challenge, GNER and its related technologies hold significant and far-reaching implications for geological research. Currently, there are four methods for Chinese Named Entity Recognition tasks: (1) rule-based approach, (2) static word vector-based approach, (3) pre-trained model-based approach, and (4) Prompt Learning-based approach.

Based on statistical principles, He et al. (2015) introduced geographic rule-related dictionaries to improve the recall and accuracy of model recognition in Geographic Named Entity Recognition based on Conditional Random Field (CRF) model using statistical principles. This approach requires significant labor costs and specific geoscientists to create a large dictionary of geoscientific keywords. Qiu et al. (2019) utilized attention-based Bidirectional Long-Short Term Memory (Att-BiLSTM) and CRF layer to extract geological named entities from geological texts via vectorizing geological texts with static word vectors, achieving automatic extraction with good performances. However, one drawback of the static word vector approach is its inability to capture semantic context representation. To address this, BERT (Devlin et al., 2018) is a language model that generates dynamic word vectors by combining contextual semantics during training. Lü et al. (2022) proposed a BERT-BiGRU-CRF model based on character-level embedding for extracting Chinese geological named entities, which incorporates contextual information to overcome the limitation of the static word vector approach. Figure 1 shows the English and Chinese Geological Named Entity Recognition example.

1.3 Prompt Learning

Traditional NER models require sufficient labeled sample data for learning and obtaining good recognition results by learning numerous textual features. However, in Named Entity Recognition tasks in geoscience, due to the scarcity of labeled data (Huang et al., 2022) and the challenge of identifying unknown types of entities, it is difficult to learn rich features from a small number of samples. Prompt Learning involves reformatting a downstream task by processing input textual information according to a specific template that leverages the processing capabilities of pre-trained language models. This new paradigm deviates from the traditional approach of fine-tuning after pre-training. Instead of adapting the pre-trained model to the downstream task through goal engineering, the approach reformulates the downstream task to resemble the Masked Language Modeling (MLM) training task using textual prompting (Yao et al., 2021; Shin et al., 2020). As traditional fine-tuning may not be consistent with the downstream task, solving the downstream task based on pre-training entails adjusting the parameters to adapt to the downstream task.

The core idea of Prompt Learning is to transform the input text information based on the pre-trained model task without altering the overall model structure, utilizing prior knowledge of the pre-trained model, and mitigating the gap between pre-training and fine-tuning targets by using masked LM targets to align the pre-trained model and downstream task more closely (Huang et al., 2022). Traditional NLP supervised learning systems predict the output x based on the model $P(y|x;\theta)$, where y can be a label, text, or various other outputs, x is the input, usually text, and θ is a trained parameter vector of this model, and we use a dataset containing (input x output y) and train the model to predict this conditional probability. Prompt-based learning methods for NLP attempt to circumvent this problem by learning an LM that models the probability $P(x;\theta)$ of the text x itself and uses this to predict the probability of y, reducing or avoiding the need to label large amounts of text.

Basic Prompt Learning is performed in three steps to predict the highest scoring y. First, a prompt template is added with

$$x' = f_{\text{prompt}}(x) \tag{1}$$

x' being the prompt output from the text after f function. For example, "The central rise of the Ordos Basin is a tectonic feature", the template may take x' = "[x], [x_1] is a [z] entity", followed by an answer search to fill [z] and find the truth value, and finally output the answer y*. Prompt Learning can effectively identify geological entities such as Chinese Cretaceous, Jurassic, etc. The main advantage of this approach is that a language model can be trained unsupervised and used to solve various tasks with the help of suitable prompts. As a result, research in this area focuses on prompt engineering to mine the most appropriate prompts template for addressing specific tasks. The model training is shown in Figure 2. The traditional approach will use a fine-tuning approach where the input text is encoded in a pre-training model BERT to obtain a dynamic embedding word vector, feature extraction through BiLSTM, and finally label classification output through the CRF layer, while the prompt-based learning approach treats the Named Entity Recognition task as a completion task and constructs the downstream task as a pre-training task form to give full play to the prior knowledge of the pre-training model and thus obtain more semantic features.

2 THE PROPOSED METHOD

In this section, we focus on Named Entity Recognition under low-resource conditions, the structure of the RoBERTa-Prompt-Tuning-NER model, and the structure and roles of the layers.

2.1 Low-Resource Geological Named Entity Recognition

Previous work treated NER as a sequence tagging task and fine-tuned it on rich resource datasets using deep learning structures such as BERT, RNN, CNN, etc. The output prediction labels were encoded using CRF, SoftMax, etc. However, traditional manual geological tagging is challenging and requires expert annotation (Singer, 2021). Our goal is to train an excellent model under a low-resource dataset configuration using Prompt Learning (Li et al., 2022).

Geological text input sentence $X = [x_1, x_2, \cdots x_N]$, where x_i

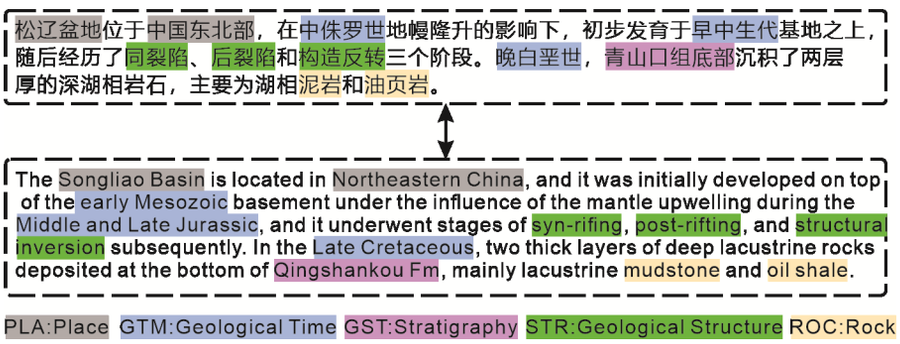
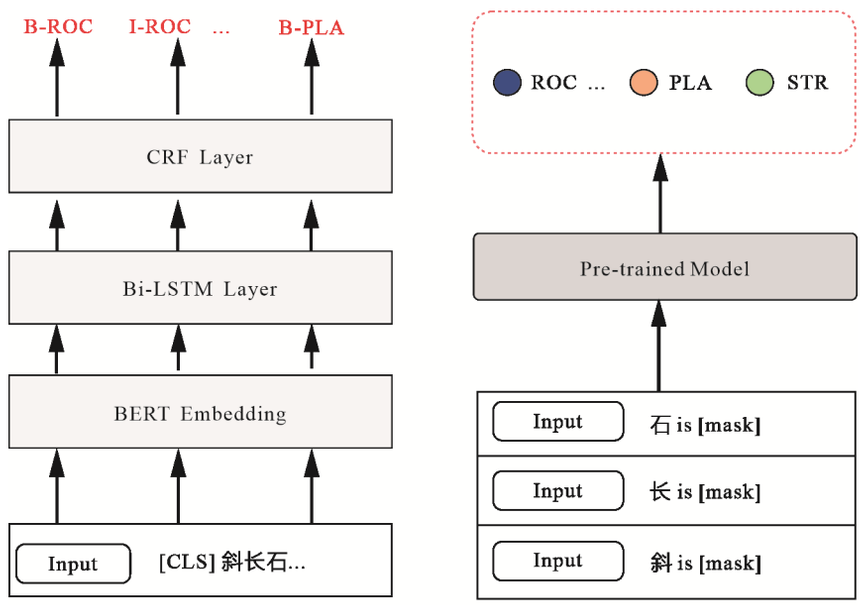


Figure 1. An example of Chinese sentence containing geological named entities, including tag types used in this study and its corresponding English translation.



(a) Traditional method (b) Prompt method

Figure 2. Traditional method and Prompt-based method. Traditional methods can use the generic representation already learned by BERT to finetune the model parameters on the target downstream task to obtain better prediction performance. Prompt-based methods replace certain words in the input sequence with a special [MASK] token and ask the MLM to predict these replaced words.

represents the word after text segmentation, and the output is $Y = [y_1, y_2, \cdots y_N]$, where y_i represents the sequence label $Y = [O$, B-min, I-min$\cdots]$, $y_i \in Y$ corresponding to each notation of X. To simulate a low-resource dataset, we randomly sample from the geological dataset.

2.2 RoBERTa-Prompt-Tuning-NER

This paper proposes the RoBERTa-Prompt-Tuning-NER method for Geological Named Entity Recognition in low-resource datasets, utilizing Prompt Learning and RoBERTa as our pre-trained model. Prompt-Tuning-NER involves fine-tuning the model using Prompt Learning to accomplish NER tasks. The model comprises a coding representation layer with RoBERTa, a RoBERTa-MLM layer for prompt completion, two fully connected layers for feature-to-sample space mapping, and a label prediction layer. The RoBERTa layer captures context-based information and longer distance dependencies through dynamic word vectors, while the RoBERTa -MLM layer selects the most fitting word from a group of options to finish the prompt task. Finally, the output is generated through the two fully connected layers that map feature space to sample space, followed by label prediction. Figure 3 illustrates the model architecture.

2.3 Basic Components

Firstly, we preprocess the data by partitioning the whole corpus into individual sentences, each sentence is truncated and padded according to the maximum string length. The tag word V corresponds to the tag space $Y \rightarrow V$ via the mapping function M. The prompt template is a text string, '$[X]$, X_i is $[Z]$', with two filled slots. For each input sentence $[X] = (x_1, x_2, \cdots x_N)$, the $[Z]$ slot is filled in (X_i is [MASK]) and spliced after $[X]$. For example, the 'calcium hornblende group of....., calcium is [MASK]'. Finally, the model processes the data.

$$H_{\text{embedding}} = \text{emb}([x, \tilde{x}]) \tag{2}$$

where x represents the original input sentence, $\tilde{x}$ represents the prompt template, and $H_{\text{embedding}}$ represents the vector generated after the text has been passed through RoBERTa.

Prompt-based learning uses the MLM to complete the

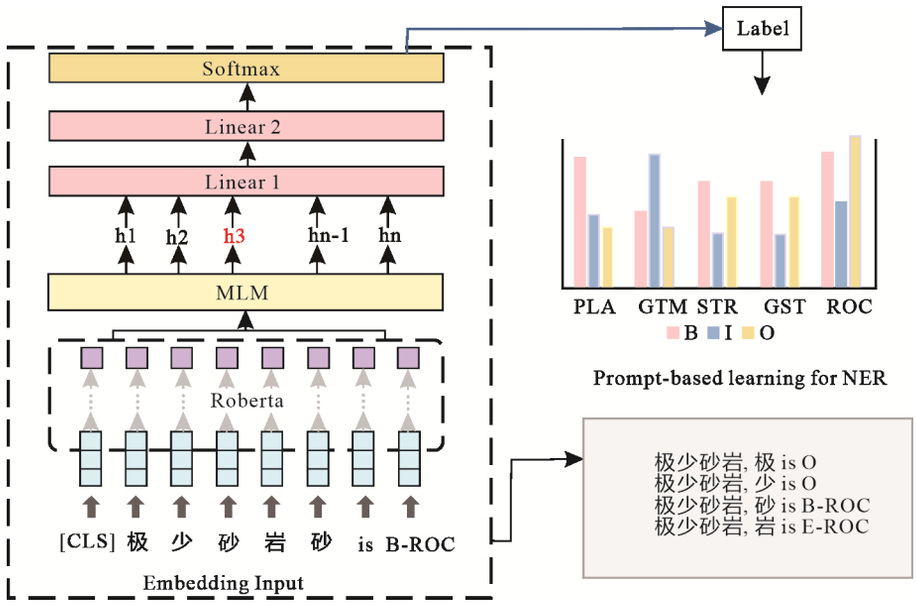


Figure 3. Architecture of RoBERTa-Prompt-Tuning-NER

prompt task, with the [MASK] in the example template representing the probability associated with a set of candidate words. Subsequently, softmax is used for label prediction after passing through two fully connected layers.

The set of candidate words is the whole word list V:{V_0,V_1, ····· , V_m}, and for each sentence [X], M([X]/V) denotes the score of the language model at [MASK] position W_t, and $\boldsymbol{W}_{lm}$ are the parameters of the pre-trained LM head. The score corresponding to each candidate word is

$$P([\text{MASK}] = M([x, \tilde{x}]|V)) = \text{Soft}_{\max}(\boldsymbol{W}_{lm} \cdot \boldsymbol{h}_{[\text{MASK}]}) \quad (3)$$

After two fully connected layers, the feature space is mapped to the sample token space, which is then transformed into a probability distribution through the Softmax function. The argmax function is subsequently used to fill the vacancy with the word that has the highest probability.

3 DATASET AND EXPERIMENT

3.1 Dataset

The experimental data comes from four regional geological survey reports obtained from the National Geological Archives of the China Geological Survey. The data set underwent manual annotation by geologists in four rounds, with consistency checks performed at each stage. The annotated data set includes six types of entities: Geological Time (GTM), Geological Structure (STR), Stratigraphy (GST), Rock (ROC), Mineral (MIN), and Place (PLA) (Table 1). A total of 10 803 sentences were labeled, comprising 1 000 106 labeled words and 598 406 unlabeled words (Ma et al., 2022). The labeled data was divided into training, validation, and test sets at an 8 : 1 : 1 ratio. In the low-resource dataset, statistical random sampling was used to extract data from the training set. Precision, recall, and F1 scores were used to evaluate the models by comparing our proposed method with several baseline models on both rich and low resources. Table 2 shows the prompt template on the low and rich resource dataset.

Table 1 Chinese and English cross-references of the six entities of Geological NER

Dataset	Entities	Example	
		Chinese	English
Geological NER	ROC	花岗岩	Granite
	PLA	日喀则	Shigatse
	STR	多尼组	Doni Formation
	GTM	侏罗纪	Jurassic
	MIN	斜长石	Plagioclase
	GST	斑点状构造	Speckled structure

3.2 Experimental Parameter Settings

Our proposed prompt-based model was trained using Python 3.7 and PyTorch 1.12 as the training framework. The RoBERTa layer employed RoBERTa-wwm-ext-Chinese, a 12-layer bidirectional transformer structure with a hidden layer dimension of 768. Dropout was used to prevent overfitting, and all experiments were conducted on an NVIDIA 4090. During training, the RoBERTa-based pre-trained model had a learning rate of 0.000 01, while the non-pre-trained model had a learning rate of 0.001. The fixed step decay method was utilized for learning rate updates, and parameter updates implemented the Adamw gradient optimization algorithm. Table 3 shows the generic experimental hyperparameters on the low and rich resource dataset.

3.3 Evaluation Metrics

We evaluated the model's performance using Conlleval.py, the Python version of the industry-recognized CoNLL-2003 (Sang and De Meulder, 2003) Named Entity Recognition dataset used for sequence annotation evaluation. The algorithmic metrics for generic NER on public datasets consist of P (precision) and R (recall) together, and a single metric is not enough to prove the performance of the modeling algorithm.

Table 2 Labeled data and training templates

Chinese sentences	Token	Sentences template	[MASK]	Label
松辽盆地位于中国东北部。 (The Songliao Basin is located in northeastern China)	松	松辽盆地位于中国东北部,松是 [MASK]	0	B-PLA
	辽	松辽盆地位于中国东北部,辽是 [MASK]	1	I-PLA
	盆	松辽盆地位于中国东北部,盆是 [MASK]	2	I-PLA
	地	松辽盆地位于中国东北部,地是 [MASK]	3	E-PLA
	位	松辽盆地位于中国东北部,位是 [MASK]	4	O

The entity "The Songliao Basin" is extracted from the Chinese text "The Songliao Basin is located in northeastern China". The [MASK] represents the sequence position of the token in the Chinese text. The Label represents the real tag sequence of the token.

Table 3 Generic hyper-parameters

Hyper-parameters	Values
Epoch	20
Batch size	64
Learning rate	1e-5
Hidden size	768

Therefore, in this paper, F1 score is chosen as the optimal performance evaluation tool.

$$P = \frac{TP}{TP + FP} \tag{4}$$

$$R = \frac{TP}{TP + FN} \tag{5}$$

$$F_1 = \frac{2PR}{P + P} \tag{6}$$

3.4 Rich-Resource Geological NER

The models were trained on the training set of the dataset and evaluated on the same test set. Table 4 presents the overall entity recognition scores for various models utilizing different algorithms. We found that prompt based method achieved the highest precision in recognizing multiple entities, including ROC, PLA, and STR. It performed particularly well in recognizing PLA and GST entities, with F1 scores of 74.20% and 70.88%, respectively. Our use of a MLM allowed us to narrow the gap between pre-training and fine-tuning when training on different targets. As a result, our prompt-based learning approach was able to adapt more quickly to downstream tasks than traditional fine-tuning methods, leading to improved performance.

3.5 Domain Transfer for Low-Resource NER

We conducted experiments using scientific random sampling of training data to simulate the Geological Named Entity Recognition task in a low-resource scenario. We evaluated the performance (F1) of various models on datasets containing 10%, 20%, 30%, 50%, and 80% of the overall dataset (Table 5). The results demonstrate that our proposed prompt tuning method based on RoBERTa outperforms the fine-tuning method when using low-resource data and small sample sizes. Although the difference between the two methods decreases as the sample size increases, the prompt tuning method still performs better than other methods under the same dataset conditions. Figure 4 shows the F1 score visualization of results from different models.

4 DISCUSSION

4.1 Generalizability and Error Analyses

In this subsection, we elaborate on the advantages proposed model in comparison with other models, and the error analysis of the geological entities.

In this paper, four common algorithmic models for Chinese Geological Entity Recognition are selected, namely IDCNN + CRF, BILSTM + CRF, BERT + IDCNN + CRF and BERT + BILSTM + CRF, where the first two models are based on static word vectors and the latter two are pre-trained models based on dynamic word vectors. The perceptual field of text features is obtained by convolutional neural networks, BLISTM is introduced to learn more contextual information, and finally they are connected to CRF (Conditional Random Field), which can learn the constraint relationship between tags and ensure the reasonableness of the predicted tag sequence. After experiments, we found that static word vectors are not as semantically rich as dynamic word vectors, so using pre-trained language models (e.g., BERT) as word embedding layers can extract more semantic features, and the experimental recognition effect is further improved. Our algorithm achieves recognition rates of 79.89%, 70.88%, 74.20% and 82.66% for GTM, GST, PLA and ROC entities, respectively. Compared with other models, our model incorporates the hinting idea in the recognition of toponymic entities and makes full use of the prior knowledge of the pre-trained model to achieve a better recognition of toponymic entities, where the best results can be seen in the location and geological age. The reason is that the entities of location and geological age are then more common in the generic prior data and are learned into the pre-trained model.

We analyze each tag and find that some entities are not recognized well. For example, in the Chinese text "biotite", the whole fragment is tagged as an entity, but only the granite entity is recognized by the model. Additionally, some other strata with location names are also incorrectly recognized because of the text boundary problem. Therefore, more tagging data are needed for correction to obtain better recognition. We also tested the recognition of geological named entities for the English dataset and found that the recognition is not satisfactory, because Chinese is character-based while English is word-based. When Chinese does the NER task, one character corresponds to one label, while English is one word to one label. Our

Table 4 Performance of different models on the rich-resource data set

Method	Rock			Place			Stratum			Mineral			Geological time			Geological structure		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
IDCNN + CRF	73.70	76.73	75.18	59.75	56.88	58.28	69.73	71.58	70.64	85.13	84.61	84.87	67.75	77.13	72.14	58.75	45.81	51.48
BiLSTM + CRF	74.59	79.25	76.85	62.79	62.29	62.54	71.89	72.76	72.32	85.62	85.95	85.79	68.18	79.97	73.61	58.62	45.49	51.23
BERT + IDCNN + CRF	74.23	78.94	76.52	61.80	56.56	59.07	71.49	71.83	71.66	85.82	85.46	85.64	66.39	78.38	71.89	62.88	49.10	55.14
BERT + BILSTM + CRF	75.21	79.08	77.10	63.02	61.53	62.27	74.10	74.37	74.23	87.38	85.64	86.50	67.84	77.31	72.26	60.71	52.84	56.50
RoBERTa_Prompt	81.05	84.33	82.66	76.83	71.75	74.20	79.48	75.77	77.58	85.63	80.69	83.08	81.05	78.76	79.89	71.17	70.59	70.88

The best F1 scores are presented in boldface. P, R, F1 are all measured in %.

Table 5 Comparison of model recognition F1-score (%) on low resource datasets

Method	10%	20%	30%	50%	80%
IDCNN + CRF	53.32	59.65	62.25	67.44	71.40
BiLSTM + CRF	56.83	62.95	62.95	69.43	73.90
BERT + IDCNN + CRF	56.84	62.86	66.67	70.74	73.66
BERT + BILSTM + CRF	60.26	65.06	65.06	72.17	75.33
RoBERTa_Prompt	77.14	78.76	80.45	79.52	80.64

Random sampling by 10%, 20%, 30%, 50%, 80% of the overall dataset. The best results are presented in boldface.

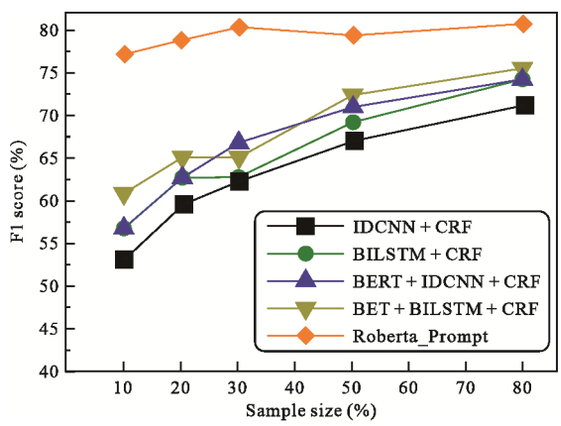


Figure 4. F1 score results of different models.

prompt template is carefully constructed for the Chinese dataset to accommodate common Chinese geological text representations, such as rock entities, which often end in rock characters in Chinese. Hence, more prompt template construction experiments must be conducted for the English dataset, and then find the optimal template.

Given the existing sample bias in current geoscientific datasets, particularly the limited synthesis of non-English data, our model helps to bridge this gap within China by optimizing the NER process for Chinese geological literature. Over the years, Chinese geological literature accumulated since the early 20th century contains abundant geological knowledge. Scholars have estimated that from 1985 to 2006, over 300 000 Chinese geoscience documents, including journals, monographs, and conference papers, covering various fields such as stratigraphy, paleontology, petrology, geochronology, and geochemistry, which are included in the Chinese Geological Literature Database. The China Geological Survey holds nationwide geological survey reports, which also contain rich geological information. It is evident that Chinese literature represents an untapped wealth of geoscientific information, which can make a crucial contribution to global geoscientific data synthesis and the construction of geoscientific knowledge graphs. Our model significantly improves the accuracy of geoscience-related NER tasks, making the process of mining geoscientific knowledge from Chinese literature more precise and efficient. This has positive implications for the consolidation and reuse of global geoscientific knowledge.

4.2 Limitations and Future Work

However, there are some limitations to this study. The first limitation is that deep learning is essentially a generalization of statistical laws for a large number of samples, called statistical learning, and the proposed approach is due to the need to obtain more textual information features in the case of low-resource datasets in order to obtain a priori knowledge of the pre-trained model. The second limitation is that we used some Chinese pre-training models, including Chinese-BERT, ERNIE, RoBERTa, etc. Finally, we chose RoBERTa because it increases the randomness and diversity of pre-training data, and RoBERTa can better learn the contextual information of the text, thus improving the accuracy of various natural language processing tasks. Our recognition results will be better if we train with a domain pre-training model based on a large geological corpus.

Therefore, it is an exceptionally good topic for future work to reduce the labor cost to annotate the large amount of geological data. Weakly supervised and remotely supervised based strategies to reduce the labor cost and use large language models for labeling data and building a richer domain corpus.

Future work will focus on three aspects: (a) At the paradigm level, we aim to optimize existing algorithms for Geological Named Entity Recognition tasks and design better templates to improve results. (b) At the data level, we plan to build a richer domain corpus under low-resource data conditions, potentially using large language models such as ChatGPT for conversational annotation, to enrich our geological text domain corpus. (c) At the algorithmic level, we will explore the use of large language models such as GPT-3 and ChatGPT for zero-sample information extraction in geological contexts.

5 CONCLSION

In this paper, we propose a novel approach to incorporate prompting ideas into geoscientific Named Entity Recognition under low resource datasets. Our approach transforms the downstream task into a pre-trained language model downstream task without changing the model structure, and uses the prior knowledge of the pre-trained model to achieve better recognition of geological named entities. Compared with traditional deep learning methods, our model captures more features of geological entities and performs better under low-resource dataset conditions. Experimental results show that the model proposed in this paper achieves the best F1 score of up to 80.64% under several low-resource data configurations, and the prompt-based learning approach can improve the performance of model recognition.

ACKNOWLEDGMENTS

This study was supported by the National Natural Science Foundation of China (Nos. 42488201, 42172137, 42050104, and 42050102), the National Key R&D Program of China (No. 2023YFF0804000), and Sichuan Provincial Youth Science & Technology Innovative Research Group Fund (No. 2022JDTD0004). This study is a contribution to the Deep-time Digital Earth (DDE) Big Science Program and IGCP 739. The data-sets used and analyzed in the current study are available from https://www.geodoi. ac. cn/edoi. aspx? DOI=10.3974/geodb. 2021.09.04.V1. The final publication is available at Springer via https://doi.org/10.1007/s12583-023-1944-8.

Conflict of Interest

The authors declare that they have no conflict of interest.

REFERENCES CITED

Allahyari, M., Pouriyeh, S., Assefi, M., et al., 2017. A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. arXiv: 1707.02919. http://arxiv.org/abs/1707.02919

Bowring, J. F., McLean, N. M., Walker, J. D., et al., 2015. Advanced Cyberinfrastructure for Geochronology as a Collaborative Endeavor: A Decade of Progress, A Decade of Plans. American Geophysical Union, Fall Meeting 2015. IN23E-03

Chan, M. A., Peters, S. E., Tikoff, B., 2016. The Future of Field Geology, Open Data Sharing and CyberTechnology in Earth Science. *The Sedimentary Record*, 14(1): 4–10. https://doi. org/10.2110/sedred. 2016.1.4

Chu, D. P., Wan, B., Li, H., et al., 2021. Geological Entity Recognition Based on ELMO-CNN-BiLSTM-CRF Model. *Earth Science*, 46(8): 3039–3048. https://doi. org/10.3799/dqkx. 2020.309 (in Chinese with English Abstract)

Consoli, B., Santos, J., Gomes, D., et al., 2020. Embeddings for Named Entity Recognition in Geoscience Portuguese Literature. Proceedings of The 12th Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France. 4625–4630

Cutcher-Gershenfeld, J., Baker, K. S., Berente, N., et al., 2016. Build It, but will They Come? A Geoscience Cyberinfrastructure Baseline Analysis. *Data Science Journal*, 15: 8. https://doi.org/10.5334/dsj-2016-008

Devlin, J., Chang, M. W., Lee, K., et al., 2018. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. arXiv: 1810.04805. http://arxiv.org/abs/1810.04805

Enkhsaikhan, M., Holden, E. J., Duuring, P., et al., 2021. Understanding Ore-Forming Conditions Using Machine Reading of Text. *Ore Geology Reviews*, 135: 104200. https://doi. org/10.1016/j. oregeorev. 2021.104200

Fan, R. Y., Wang, L. Z., Yan, J. N., et al., 2019. Deep Learning-Based Named Entity Recognition and Knowledge Graph Construction for Geological Hazards. *ISPRS International Journal of Geo-Information*, 9(1): 15. https://doi.org/10.3390/ijgi9010015

Guo, C., Xu, Q., Dong, X. J., et al., 2021. Geohazard Recognition and Inventory Mapping Using Airborne LiDAR Data in Complex Mountainous Areas. *Journal of Earth Science*, 32(5): 1079–1091. https://doi.org/10.1007/s12583-021-1467-2

He, Y. X., Luo, C. W., Hu, B. Y., 2015. Geographic Entity Recognition Method Based on Crf Model and Rules Combination. *Computer Applications and Software*, 32(1): 179–185, 202. https://doi. org/10.3969/j. issn.1000-386x.2015.01.046 (in Chinese with English Abstract)

Holden, E. J., Liu, W., Horrocks, T., et al., 2019. GeoDocA—Fast Analysis of Geological Content in Mineral Exploration Reports: A Text Mining Approach. *Ore Geology Reviews*, 111: 102919. https://doi.org/10.1016/j.oregeorev.2019.05.005

Huang, G. H., Zhong, J., Wang, C., et al., 2022. Prompt-Based Self-Training Framework for Few-Shot Named Entity Recognition. Knowledge Science, Engineering and Management. Proceedings of 15th International Conference, KSEM 2022. August 6–8, 2022, Singapore. 91–103. https://doi.org/10.1007/978-3-031-10989-8_8

Kitchin, R., 2014. Big Data, New Epistemologies and Paradigm Shifts. *Big*

Data & Society, 1(1): 205395171452848. https://doi.org/10.1177/2053951714528481

Lehnert, K., Su, Y., Langmuir, C. H., et al., 2000. A Global Geochemical Database Structure for Rocks. *Geochemistry*, *Geophysics*, *Geosystems*, 1(1): 1012. https://doi.org/10.1029/1999gc000026

Li, D. F., Hu, B. T., Chen, Q. C., 2022. Prompt-Based Text Entailment for Low-Resource Named Entity Recognition. arXiv: 2211.03039. http://arxiv.org/abs/2211.03039

Liu, P. F., Yuan, W. Z., Fu, J. L., et al., 2023. Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*, 55(9): 195. https://doi.org/10.1145/3560815

Lü, X., Xie, Z., Xu, D. X., et al., 2022. Chinese Named Entity Recognition in the Geoscience Domain Based on BERT. *Earth and Space Science*, 9(3): e02166. https://doi.org/10.1029/2021ea002166

Ma, K., Tian, M., Tan, Y. J., et al., 2022. Named Entity Recognition Dataset for Four Regional Geological Survey Reports by Data Mining Methodology. *Journal of Global Change Data & Discovery*, 6(1): 78–84. https://doi.org/10.3974/geodp.2022.01.11

McKay, N. P., Emile-Geay, J., 2016. Technical Note: The Linked Paleo Data Framework—A Common Tongue for Paleoclimatology. *Climate of the Past*, 12(4): 1093–1100. https://doi.org/10.5194/cp-12-1093-2016

Peters, S. E., Husson, J. M., 2018. We need a Global Comprehensive Stratigraphic Database: Here's a Start. *The Sedimentary Record*, 16(1): 4–9. https://doi.org/10.2110/sedred.2018.1.4

Peters, S. E., Husson, J. M., Czaplewski, J., 2018. Macrostrat: A Platform for Geological Data Integration and Deep-Time Earth Crust Research. *Geochemistry*, *Geophysics*, *Geosystems*, 19(4): 1393–1409. https://doi.org/10.1029/2018gc007467

Peters, S. E., McClennen, M., 2016. The Paleobiology Database Application Programming Interface. *Paleobiology*, 42(1): 1–7. https://doi.org/10.1017/pab.2015.39

Piskorski, J., Yangarber, R., 2013. Information Extraction: Past, Present and Future. Multi-source, Multilingual Information Extraction and Summarization. Springer, Berlin, Heidelberg. 23–49. https://doi.org/10.1007/978-3-642-28569-1_2

Qiu, Q. J., Xie, Z., Wu, L., et al., 2019. GNER: A Generative Model for Geological Named Entity Recognition without Labeled Data Using Deep Learning. *Earth and Space Science*, 6(6): 931–946. https://doi.org/10.1029/2019ea000610

Qiu, Q. J., Tian, M., Xie, Z., et al., 2023. Extracting Named Entity Using Entity Labeling in Geological Text Using Deep Learning Approach. *Journal of Earth Science*, 34(5): 1406–1417. https://doi.org/10.1007/s12583-022-1789-8

Quinn, D., Linzmeier, B., Sundell, K., et al., 2021. Implementing the Sparrow Laboratory Data System in Multiple Subdomains of Geochronology and Geochemistry. EGU General Assembly Conference Abstracts. EGU21-13832. https://doi.org/10.5194/egusphere-egu21-13832

Raja, N. B., Dunne, E. M., Matiwane, A., et al., 2022. Colonial History and Global Economics Distort our Understanding of Deep-Time Biodiversity. *Nature Ecology & Evolution*, 6(2): 145–154. https://doi.org/10.1038/s41559-021-01608-8

Sang, E. F., De Meulder, F., 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003. Edmonton, Canada. Association for Computational Linguistics, Morristown, NJ, USA. https://doi.org/10.3115/1119176.1119195

Shin, T., Razeghi, Y., Logan IV, R. L., et al., 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. arXiv: 2010.15980. http://arxiv.org/abs/2010.15980

Shipley, T. F., Tikoff, B., 2019. Collaboration, Cyberinfrastructure, and Cognitive Science: The Role of Databases and Dataguides in 21st Century Structural Geology. *Journal of Structural Geology*, 125: 48–54. https://doi.org/10.1016/j.jsg.2018.05.007

Singer, D. A., 2021. How Deep Learning Networks could be Designed to Locate Mineral Deposits. *Journal of Earth Science*, 32(2): 288–292. https://doi.org/10.1007/s12583-020-1399-2

Vieira, D. A., Mookerjee, M., Matsa, S., 2014. Incorporating Geoscience, Field Data Collection Workflows into Software Developed for Mobile Devices. AGU Fall Meeting Abstracts. IN41A-3641

Walker, J. D., Tikoff, B., Newman, J., et al., 2019. StraboSpot Data System for Structural Geology. *Geosphere*, 15(2): 533–547. https://doi.org/10.1130/ges02039.1

Walker, J., Lehnert, K., Hofmann, A., et al., 2005. EarthChem: International Collaboration for Solid Earth Geochemistry in Geoinformatics. AGU Fall Meeting Abstracts. IN44A-03

Wang, B., Ma, K., Wu, L., et al., 2022. Visual Analytics and Information Extraction of Geological Content for Text-Based Mineral Exploration Reports. *Ore Geology Reviews*, 144: 104818. https://doi.org/10.1016/j.oregeorev.2022.104818

Wang, Q. Y., Li, Z. H., Tu, Z. P., et al., 2023. Geotechnical Named Entity Recognition Based on BERT-BiGRU-CRF Model. *Earth Science*, 48(8): 3137–3150. https://doi.org/10.3799/dqkx. 2022.462 (in Chinese with English Abstract)

Williams, J. W., Grimm, E. C., Blois, J. L., et al., 2018. The Neotoma Paleoecology Database, a Multiproxy, International, Community-Curated Data Resource. *Quaternary Research*, 89(1): 156–177. https://doi.org/10.1017/qua.2017.105

Yan, H., Yang, N., Peng, Y., et al., 2020. Data Mining in the Construction Industry: Present Status, Opportunities, and Future Trends. *Automation in Construction*, 119: 103331. https://doi.org/10.1016/j. autcon. 2020. 103331

Yao, Y., Zhang, A., Zhang, Z. Y., et al., 2021. CPT: Colorful Prompt Tuning for Pre-Trained Vision-Language Models. arXiv: 2109.11797. http://arxiv.org/abs/2109.11797

Ye, S., 2022. A Quantitative Investigation of Large Geoscientific Datasets: How Records of Geochronology and Macroevolution are Distorted by Paleoclimate, Paleoenvironment, and Sediment Preservation: [Dissertation]. The University of Wisconsin-Madison, Madison

Ye, S., Cuzzone, J. K., Marcott, S. A., et al., 2023. A Quantitative Assessment of Snow Shielding Effects on Surface Exposure Dating from a Western North American 10Be Data Compilation. *Quaternary Geochronology*, 76: 101440. https://doi.org/10.1016/j.quageo.2023.101440

Ye, S., Peters, S. E., 2023. Bedrock Geological Map Predictions for Phanerozoic Fossil Occurrences. *Paleobiology*, 49(3): 394–413. https://doi.org/10.1017/pab.2022.46

Zhu, Y. Q., Sun, K., Hu, X. M., et al., 2023. Research and Practice on the Framework for the Construction, Sharing, and Application of Large-Scale Geoscience Knowledge Graphs. *Journal of Geo-information Science*, 25(6): 1215–1227. https://doi.org/10.12082/dqxxkx. 2023. 210696 (in Chinese with English Abstract)