Computers & Geosciences 168 (2022) 105229

Contents lists available at ScienceDirect

Computers and Geosciences

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

Understanding geological reports based on knowledge graphs using a deep learning approach

Bin Wang [a], Liang Wu [b], Zhong Xie [b], Qinjun Qiu [b,*], Yuan Zhou [c], Kai Ma [d], Liufeng Tao [b]

[a] School of Geography and Information Engineering, China University of Geosciences, Wuhan, 430074, China
[b] School of Computer Sciences, China University of Geosciences, Wuhan, 430074, China
[c] National Engineering Research Center of Geographic Information System, Wuhan, 430074, China
[d] College of Computer and Information Technology, China Three Gorges University, Yichang, 443002, China

ARTICLE INFO

Keywords:
Geological reports
Geological entities and relations
Deep learning
Geological knowledge graphs

ABSTRACT

Geological reports aid in understanding exploration by providing valuable information on rock formation , evolution and the geological environment in which deposits formed. Querying and extracting the critical information from these enormous historical geological report data helps in understanding the exploration risks in different geological settings. However, large amounts of unstructured text data occur in geological reports; therefore, it is challenging to obtain valuable information from them without performing information extraction and processing. This study proposed an automated method for extracting information from geological reports through triple extraction, then automatically constructs a geological knowledge graph from the extracted entities and relations. Simultaneously, due to the lack of samples, this study used multiple geological reports to construct a corpus of jointly extracted geological entities and relations. The proposed model reached an F1-score of 90.05% in the experimental results on the constructed corpus. Finally, a knowledge graph was constructed based on the extracted results to demonstrate the application value of the proposed method. The results showed that the structured information helps better represent the content of the source report and matches well with the geological domain knowledge. The proposed method can quickly and robustly convert textual data into a structured form that is convenient for reasoning and querying geological entities and relations.

1. Introduction

Geological prospecting work is clearly important for promoting mineral surveys, human geographic activities, and the social economy (Zhao et al., 2011; McManus et al., 2021; Liang et al., 2021). Through continuous geological surveys and scientific research development, a variety of geological data have gradually accumulated, including geological database data, texts, images, voices, etc. (Jianping et al., 2016; Wang et al., 2021). These data contain rich geoinformation to be explored. Text is an important medium for recording geological work. Among these data, geological documents such as geological reports and scientific documents contain considerable information. However, most of these data are unstructured and consist of complex types, making it difficult to use directly to understand and mine new knowledge (Wei et al., 2014, 2021). An important source of such information is historical geological reports that are publicly available but not widely used, including geologic exploration reports, such as the vast number of geologic reports of various types available in China's National Geological Archives (http://www.ngac.cn). These geological reports detail work done in the study area as well as observations and detailed interpretations of geological phenomena. For example, determining the conditions and environment under which a deposit occurs requires a detailed understanding of the geographic location of the deposit, the geological structure, the age of the mineralization, and the complex correlations between the host rock, the deposit and the minerals coexisting with the mineralization. In addition to the mineralizing environment of the deposit in the area, the mineralized zone can be evaluated and analyzed based on the above-described information on mineralization-related properties (Enkhsaikhan et al., 2021a; Qun et al., 2021; Wang et al., 2021; Zhou et al., 2021).

Text is an important means of communication. However, considerable textual data written in natural language are extracted from complex

* Corresponding author.
E-mail addresses: wangbin@cug.edu.cn (B. Wang), wuliang@cug.edu.cn (L. Wu), xiezhong@cug.edu.cn (Z. Xie), qiuqinjun@cug.edu.cn (Q. Qiu), zhouyuan@cug.edu.cn (Y. Zhou), makai@ctgu.edu.cn (K. Ma), taoliufeng@cug.edu.cn (L. Tao).

https://doi.org/10.1016/j.cageo.2022.105229
Received 22 January 2022; Received in revised form 28 July 2022; Accepted 3 September 2022
Available online 9 September 2022

and diverse content, which is a time-consuming task and bureaucratic procedure when performed manually. Artificial intelligence (AI) has boosted the efficiency of natural language processing (NLP) technology in processing large amounts of text corpora. By processing unstructured data with NLP technology, more abundant patterns and knowledge can be found through the recognition of entities, attributes, and relations, which is called text mining. Numerous geological concepts are related to geographical information in geological documents, including rocks, geological structure features, geological time, geological orientation, and geological place names. These content words and their relations can fully reflect the humanistic, economic and geographical characteristics of the survey area (Ma, 2021).

Compared to other information extraction efforts, the extraction of Chinese geological entities and relations is an immensely complex task for numerous reasons: (1) First, geological reports are highly variable in terms of text characteristics and patterns because they are typically written by many different writers/inspectors from different=national and provincial, municipal and county agencies. (2) Second, compared to domain-general text, geological reports exhibit domain-specific uniqueness that involves high levels of technical detail, complex concept identification and relationship association (i.e., identifying complex technical concepts about geological structures, rock minerals, etc., and their associated relations), while existing models and techniques cannot deal with such complexities and variabilities with high precision and recall performance.

Knowledge graphs (KGs) effectively store the mined structured information and can be widely used in many query and retrieval applications, both domain-specific and domain-independent (Paulheim, 2017; Fensel et al., 2020). A KG is a representation based on a graph structure that takes critical information from one or more sources, integrates it into a knowledge base, establishes connections between data objects through constructed nodes and relations, and enables the generation of new knowledge through the inference techniques. A KG represents entities and relations/attributes in the form of nodes and edges to encode semantic relations about the mined data. Geological named entity recognition (NER) and relation extraction (RE) are critical links in knowledge exploration tasks and are also essential elements for constructing a knowledge graph. The methods used by such tasks include 1) methods based on rules and dictionaries (Sari et al., 2010; Ravikumar et al., 2017); 2) methods based on traditional machine learning (Sobhana et al., 2010; Minard et al., 2011); and 3) methods based on deep learning (Huang et al., 2021; Lun et al., 2021). The first two approaches require manual participation or the construction of considerable features, resulting in time-consuming, labor-intensive and inefficient procedures. In recent years, big data have driven the development of deep learning technology, and researchers have sought to use deep learning technologies for information extraction tasks. Existing research typically uses an open-source tool, such as the extensible knowledge extraction toolkit DeepKE (Zhang et al., 2022), to make relation extraction. These open-source tools are designed to recognize not only different named entities but also entity relations. In contrast to DeepKE purely extracting entities or relations, we propose a joint entity and relation extraction model for the geological domain. In the geosciences domain, much work has been done to enable single tasks in NER or RE, but few studies have focused on combining two subtasks to construct triples. Triples integrate named entities and relations in sentences in the form (*Entity 1, Relation/Attribute, Entity 2/Attribute Value*). Triples not only enable text to be stored in a structured form but also support the construction of knowledge graphs, which is an interesting research topic. In recent years, most studies have used the sequence annotation mode (Wan et al., 2019) for corpus tagging, followed by NER and RE tasks in turn. However, this pipelined method can lead to error accumulation because errors that occur in the identified candidate named entities will be propagated to the relation extraction, causing the triplet extraction task to fail (Li and Ji, 2014). Therefore, researchers began to study joint extraction methods, i.e., using a model to identify named entities and relations simultaneously (Sui et al., 2020). Generally, the extraction results of joint models are better than those of pipeline extraction. However, there are fewer corpora in the domain of geology, the density of the geological entity distribution is high, and considerable overlapping relations exist. Therefore, both pipelined and joint extraction methods for specific information extraction in the geosciences domain face certain challenges.

The abovementioned problems need to be solved urgently. This study proposed a new joint extraction model for geological texts, named GeoERE-Net, which automatically extracts triples from geological literature, and then triples were applied to build geological knowledge graphs. GeoERE-Net focuses more on global context information to better account for the long and short differences, and semantic differences of geological terms. GeoERE-Net mainly uses the RoBERT model (Liu et al., 2019) to encode sentences and proposes two different structures that are used during the entity and relation extraction stages of the joint extraction model. In the entity extraction stage, axial attention based on a bidirectional long short-term memory (BiLSTM) model was introduced, which considers the context dependence of large areas or even global areas. In the relation extraction phase, a graph convolution unit (GCU) was designed to obtain multiscale context information to alleviate multiple relation overlap.

The main contributions of this study are as follows:

(1) From a methodological perspective, this study was the first to propose the application of a joint extraction model in the domain of geological text, and we proposed a new model, GeoERE-Net.
(2) The axial attention based on BiLSTM and GCU modules was designed to consider the contextual information in geological texts to improve the integrity of triplet extraction.
(3) From an application perspective, this study proposed a new idea for using extracted geological text triples to construct a geological KG that transforms considerable textual data into a graph-based knowledge representation that aids in geological knowledge discovery and relation reasoning.

2. Related works

Mining information from geological documents has focused primarily on keyword extraction (Wang et al., 2018), geological NER (Consoli et al., 2020; Enkhsaikhan et al., 2021b) and geological RE (Deng et al., 2021), which are used to mine the subject content of documents, realize the association of multiple document contents, and construct geological knowledge graphs, respectively.

2.1. Keyword extraction

Keyword extraction tasks are mostly conducted using unsupervised methods. Generally, the extracted words express the topic of a document (Li et al., 2021; Zhao and Xie, 2021). Specifically, Wang et al. (2018) established a mixed corpus and trained a word segmentation model using conditional random fields (CRFs). Then, the obtained model was used for word segmentation of Chinese documents. The content words were obtained after removing stop words. Finally, the semantic relationship between content words was analyzed by a statistical method, and a knowledge graph of the content words and their relations was visualized. Several researchers have conducted similar works aimed at different geological applications. Shi et al. (2018) proposed a text mining method based on a convolutional neural network (CNN) to extract prospecting information, which was visualized using word frequency statistics, co-occurrence matrix statistics and the term frequency-inverse document frequency (TF–IDF) method. These methods process these classified texts based on words, sentences and paragraphs. Qiu et al. (2021) proposed an enhanced graph-based error-feedback propagation in keyword extraction; among them, words in geological reports form the nodes of the graph, and the symbiotic relations between words form the edges. Then, the candidate

content words were graded and sorted by a sorting algorithm, and the final keywords are obtained by repeated error-feedback correction. However, all of these keyword extraction methods were only implemented by statistical methods such as frequency, TF-IDF and cooccurrence relationships between words, and frequently extract large numbers of keywords unrelated to the document topic. Therefore, intervention rules typically need to be designed manually. Overall, keyword methods tend to extract only superficial information from geological documents.

2.2. Named entity recognition and relation extraction

Manually designing rules or constructing dictionaries for information extraction can be highly accurate but cannot cover every entity and relation, and it becomes inefficient when faced with a large number of task conditions. Traditional machine learning methods, such as support vector machines (SVMs), CRFs and decision trees (DTs), learn mostly shallow features, require considerable feature engineering and are not generalizable when applied to different data scenarios (Li et al., 2004; Zhu et al., 2005; Jyothi et al., 2008). In the big data era, deep learning provides opportunities for information extraction tasks. The NER method based on deep learning does not require manual rules or complex features and can easily extract hidden features from the input corpus. Qiu et al. (2019) proposed an attention-based BLSTM model with a CRF layer network (Att-BiLSTM-CRF) to identify named entities in geoscience documents. This study used 17 regional geological reports to manually annotate geological named entities samples. Similarly, Fan et al. (2020) proposed a pattern-based method for constructing an NER corpus for geological hazards that eliminated the complexity of manually constructing a corpus. Based on the constructed corpus, a bidirectional gated recurrent unit (BIGRU) and a CRF were used to construct a multibranch model (BiGRU-CRF) used to identify the named entity for a geological disaster. Importantly, the geological RE also reflects the temporal and spatial evolution of geological features. At the word level, Luo et al. (2017) introduced a new highway network into the BIGRU model by using an attention mechanism to capture additional semantic features between words; this model achieved good geological RE performances on geological datasets. However, simple word-level features were insufficient for learning the relationships of complex contents well. Therefore, end-to-end RE neural networks containing word features, sentence features, category features and location information were successively proposed (Li et al., 2013; Gupta et al., 2019; Qin et al., 2021). However, although individual geological NER or RE tasks can perform well, they cannot fully reflect the information or knowledge in geological texts; therefore, constructing knowledge graphs that reflect deep geological knowledge is still challenging.

2.3. Triple extraction

Geological text information extraction has rarely been studied from the aspect of triple extraction. However, triple extraction in general fields is a hot topic, and many researchers have achieved successes. Early triple extraction mostly involved pipeline extraction methods, which propagate errors and can have unsatisfactory results (Yu et al., 2019). Joint extraction methods link the two submodels through various dependencies, including integer linear programming (Yang and Cardie, 2013) and global probability graph models (Singh et al., 2013). In later studies, the successive proposals of table-filling methods (Zhang et al., 2017), multilevel-labeling methods (He et al., 2020) and reinforcement-learning methods (Wang and Zhang, 2021) solved the problem of extracting triples under complex relations to a certain extent. Therefore, triple extraction from geological texts to capture their rich temporal and spatial attributes was necessary.

3. Dataset and methodology

3.1. Dataset

The geological object relation description includes the spatial relation description and semantic relation description. (1) The spatial relation description involves the topological relation, orientation relation and metric relation description. (2) Semantic relation description, mainly for the description of the relation between concepts, between concepts and instances, and between instances. The most basic semantic relations include parent/child relations, whole/part relations, mutually exclusive relations, equivalence relations, and instance relations, which are the relations between concepts and their instances. We have supplemented and extended the relations between geological text entities at the ontology concept level.

The ontology prototype (named GeoOnto) was built upon the geological expert knowledge expressed in Qiu et al. (2019), which provides a global and broad hierarchical terminology that models geoscience data (i.e., geological structure, engineering geology, paleobionts, environmental geology, and geochemical exploration). GeoOnto was developed into 23 top-level hierarchies, including a total of 49,406 unique concepts. Finally, based on the constructed ontology, we develop and form a total of 24 geological entity relations.

The annotated corpus adopted in this study consisted of 8 Chinese regional survey reports from different regions. Each report details the physical geography and economics of the area along with the stratigraphy and different rock distributions. The corpus we annotated focuses more on temporal information, spatial information and attribute information in geological texts. Therefore, the corpus can be divided into three categories, namely, geological temporal features, geological spatial features and geological attribute features.

The specific annotation rules are shown in Table 1. Finally, 4,268 statements were annotated. The constructed dataset was called the Geosciences Information Extraction Dataset (GIED). We divided the training, validation and test sets at a ratio of 6:2:2; that is, the training set contained 2,560 statements, and the validation and test sets each contained 854 statements. It should be noted that since we were studying triplet extraction from Chinese geological reports, the English translation does not completely conform to the Chinese grammatical structure. Therefore, the sentences illustrated in the following examples and the triplet extraction results were in accordance with Chinese grammar.

3.2. The GeoERE-Net framework

The GeoERE-Net network proposed in this study is an improvement based on the CasRel network (Wei et al., 2019) and is primarily divided into two submodules: subject tagger and object tagger. The network architecture is shown in Fig. 1. In the encoder stage, we replaced the BERT model used in the CasRel network with the pretrained RoBERT model (Liu et al., 2019) to encode the input sentence. First, the subjects were identified; then, for each candidate subject, all possible objects that satisfy the specific relation are checked, that's, the corresponding relation and object pairs are checked under the given condition. As shown in Fig. 1, we introduced the BiLSTM to establish a long-term dependency on the sentence after RoBERT encoding and used its output as the input to the axial attention (see Section 3.2.1 for details) to strengthen the global context information. We named this structure the axial attention based on BiLSTM. In addition, during the object tagging stage of a specific relation, the encoding information of the subjects (V_{sub}) and the sentence vector information (h) encoded by RoBERT were added to the input and fed to a designed GCU (see Section 3.2.2 for details) that considers long-distance context information.

Specifically, regarding the tagging strategy of the subject, we obtained the vector h' after conducting the RoBERT encoding, BiLSTM operation and axial attention operation. Through a binary classification

Table 1
Geosciences information extraction dataset (GIED) annotation rules.

Main categories	Subclass relations	Examples	Triplets
Geological temporal information	Age	The age of the survey area is early middle Permian.	(Survey area, Age, Early middle Permian)
Geological spatial information	Exposed in	Carboniferous strata in the survey area are exposed in the Ganglong-Shenya area.	(Carboniferous, Exposed in, Ganglong-Shenya area)
	Located in	The Hongliugou gold-copper mine is located in the Hongliugou area in the western part of the survey area.	(Hongliugou gold-copper mine, Located in, The Hongliugou area in the western part of the survey area)
	Conformable contact	The Bakun Formation is in conformity with the underlying Borila Formation.	(Bakun Formation, Conformable contact, Borila Formation)
	Unconformable contact	The Ordovician Elantag Formation does not peak, and the lower is unconformized on the Cambrian monzonitic granite.	(Ordovician Elantag Formation, Unconformable contact, Cambrian monzonitic granite)
	Disconformable contact	Tongziyan Formation — — — — — — — — parallel unconformity — — — — — — — — Maokou Formation.	(Tongziyan Formation, Disconformable contact, Maokou Formation)
	Fault contact	Grey medium-thick bedded medium-grained lithic quartz sandstone fault Geren volcanic rock.	(Grey medium-thick bedded medium-grained lithic quartz sandstone, Fault contact, Geren volcanic rock)
	Intrusive contact	The geological background is Jurassic biotite granite outlying section, which is in intrusive contact with pyroclastic strata.	(Jurassic biotite granite, Intrusive contact, Pyroclastic rock strata)
	Swallowed	The copper ore was emplaced and swallowed by the intermediate-acid complex containing disseminated pyrite in the later stage.	(Copper ore, Swallowed, Intermediate-acid complex)
Attribute information	Distribution pattern	The Carboniferous strata are zoned east–west.	(Carboniferous, Distribution pattern, east–west zonal distribution)
	Geotectonic location	The Carboniferous tectonics area is located in the northern margin of the Gangdise tectonic belt.	(Carboniferous, geotectonic location, northern margin of Gangdise tectonic belt)
	Stratigraphic regionalization	The Carboniferous stratigraphy is divided into Gangdis-Tengchong stratigraphy and Bangor-Batsu stratigraphy.	(Carboniferous, Stratigraphic regionalization, Gangdis-Tengchong stratigraphy and Bangor-Batsu stratigraphy)
	Administrative division	The administrative division of the survey area belongs to Dingqing County of the Tibet Autonomous Region.	(Survey area, Administrative division, Dingqing County of Tibet Autonomous Region)
	Exposed strata	The exposed strata of the Carboniferous are the Lower-upper Carboniferous Yongzhu Formation and the Upper Carboniferous-Lower Permian Laga Formation.	(Carboniferous, Exposed strata, Lower-upper Carboniferous Yongzhu Formation); (Carboniferous, Exposed strata, Upper Carboniferous-Lower Permian Laga Formation)
	Lithology	The Yongzhu Formation is characterized by black mudslate.	(Yongzhu Formation, Lithology, black mudslate)
	Thickness	Light bluish-grey mudslate 25.30 m.	(Light bluish-grey mudslate, Thickness, 25.30 m)
	Area	The exposed area of the Carboniferous is approximately 1,500 km^2.	(Carboniferous, Area, 1,500 km^2)
	Coordinate	The starting point coordinates of the profile are: N31°33′25″, E87°16′37".	(Profile, Coordinate, N31°33′25″); (Profile, Coordinate, E87°16′37″)
	Length	The Siribeng fault zone has a total length of approximately 20 m.	(Siribeng fault zone, Length, 20 m)
	Elevation	The elevation of the profile is 4,943 m.	(Profile, Elevation, 4,943 m)
	Contain	The phenocrysts mainly consist of plagioclase and a small amount of quartz biotite.	(Phenocrysts, Contain, plagioclase); (Phenocrysts, Contain, quartz biotite)
	Develop	A fine horizontal bedding is developed in the slate.	(Slate, Develop, horizontal bedding)
	Paleontology	Hecosmilia sp. scabbard coral was collected from limestone.	(Limestone, paleontology, hecosmilia sp. scabbard coral)
	Belong to	The survey area belongs to the high cold area of the Qinghai-Tibet Plateau.	(Survey area, Belong to, The high cold area of the Qinghai-Tibet Plateau)

operation, the start and end positions of the subject are marked as 1, and the remaining positions are marked as 0. The specific calculation formulas are shown in Equations (1) and (2):

$$p_i^{start\_s} = \sigma(W_{start} x_i + b_{start}) \tag{1}$$

$$p_i^{end\_s} = \sigma(W_{end} x_i + b_{end}) \tag{2}$$

where $p_i^{start\_s}$ and $p_i^{end\_s}$ represent the probability that the i-th token is about the start and end of the subject, respectively. A threshold was set in the experiment such that if the probability exceeds the set threshold, it was assigned a 1; otherwise, it was assigned a 0. Here, we set the threshold to 0.5. In addition, the i-th token vector representation x_i was through RoBERT encoding, BiLSTM and axial attention operations, among them, $x_i = h'[i]$. W_{start} and W_{end} were, respectively, expressed as the weights obtained by model training. b_{start} and b_{end} represent the bias values, and σ represents the sigmoid binary classification function.

The strategy to tag an object for a specific relation is similar to the strategy for tagging the subject but differs by directly decoding the output vector h'. This stage also considers the characteristics of the identified subject V_{sub}^k. Then, the model performs GCU processing on the summed vectors of h and V_{sub}^k to obtain the final object representation. The formulas are as follows:

$$p_i^{start\_o} = \sigma\left(G\left(W_{start}^r\left(x_i + V_{sub}^k\right)\right) + b_{start}^r\right) \tag{3}$$

$$p_i^{end\_o} = \sigma\left(G\left(W_{end}^r\left(x_i + V_{sub}^k\right) + b_{end}^r\right)\right) \tag{4}$$

where $p_i^{start\_o}$ and $p_i^{end\_o}$ represent the probability of the i-th token being the start and end positions of the object, respectively; G represents the GCU operation; Additionally, the i-th token vector representation x_i was encoded by RoBERT, among them, $x_i = h[i]$; and W_{start}^r and W_{end}^r are, respectively, expressed as the weights obtained by model training. V_{sub}^k represents the encoding representation of the k-th subject, which is the average vector representation of all tokens between the start and end of the k-th subject, and b_{start}^r and b_{end}^r represent bias values.

3.2.1. The axial attention based on BiLSTM

The long short-term memory network (LSTM) (as shown in Fig. 2) is a network derivative that helps solve vanishing gradient and gradient explosion problems by introducing a gating mechanism. An LSTM consists of multiple structures: an input word x_t, a cell state C_t, a temporary cell state $\widetilde{C}_t$, a hidden layer state h_t, a forget gate f_t, a memory gate i_t, and an output gate o_t. Generally, an LSTM first calculated and obtains the results f_t, i_t, and $\widetilde{C}_t$; among them, $\widetilde{C}_t$ selected the information, some were forgotten and some were remembered. Second, it calculated and

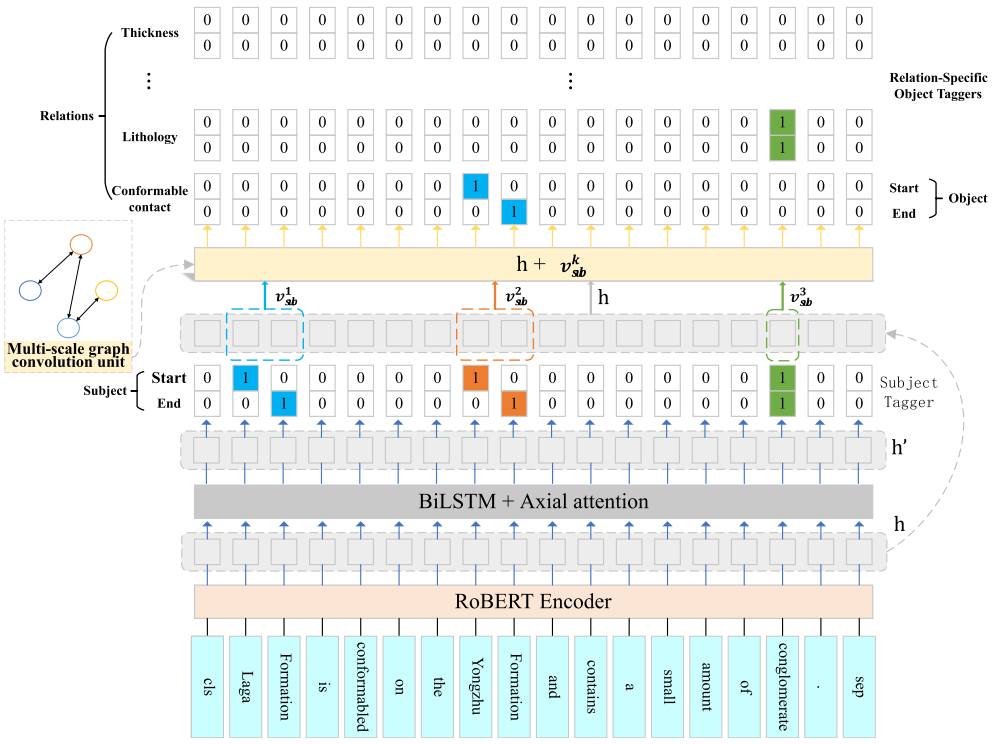

Fig. 1. GeoERE-Net framework. The encoder uses RoBERT to encode sentences, and the decoder uses cascaded decoding to first decode the candidate subjects and then decode the objects under a specific relation.

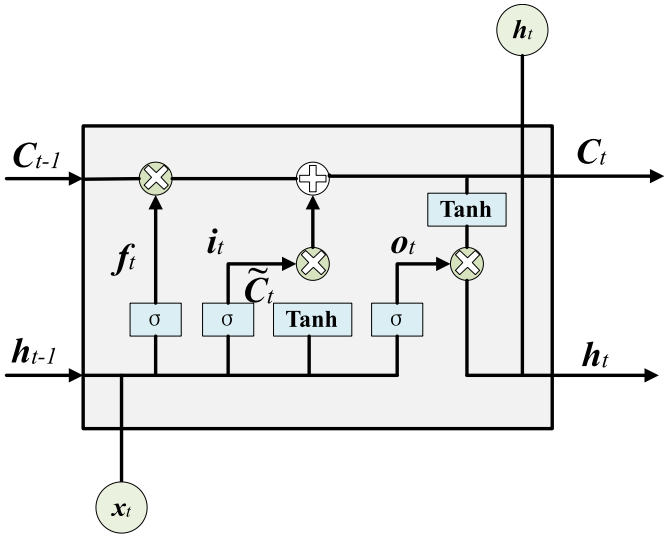

Fig. 2. The LSTM framework.

obtained the current cell state C_t. Finally, the output gate o_t and the current hidden layer state h_t were obtained. The BiLSTM was formed by combining a forward LSTM unit and a backward LSTM unit. Assuming that the forward output after the LSTM operation is $\overrightarrow{h_i}$ and the reverse output is $\overleftarrow{h_i}$, the vector obtained after the BiLSTM operation for the *i*-th word is $h_i = [\overrightarrow{h_i} \oplus \overleftarrow{h_i}]$. Therefore, the hidden state sequence *H*: [h_1, h_2 …, *ht*], corresponding to each word embedding, is generated after the BiLSTM operation. The feature vector *H* is then input to the next layer, the axial attention layer.

In this study, the vector features processed by the BiLSTM were converted into two-dimensional vectors according to the length term so that attention was applied along a single axis of the two-dimensional tensor. Compared with standard self-attention, axial attention achieves significant savings in computational memory. If the shape of a *d*-dimensional tensor is $N = N^{1/d} \times \cdots \times N^{1/d}$, the complexity of axial attention was $O(N^{(d-1)/d})$ complexity compared to that of the standard self-attention mechanism. Axial attention was divided into row-oriented attention and column-oriented attention, and it can conduct attention operations over a larger area or even over a global area, as shown in Fig. 3.

In Fig. 3, the BiLSTM output forms the input of the axial attention, that is, $I \in \mathbb{R}^{L \times C1}$ is applied as the input. First, we converted $I \in \mathbb{R}^{L \times C1}$ into a two-dimensional tensor $I' \in \mathbb{R}^{C1 \times H \times H}$ along the length dimension, where *L* represents the length of the text after encoding, C_i represents the channel dimension, and $H = L^{1/2}$. Specifically, for row-oriented attention, $I' \in \mathbb{R}^{C1 \times H \times H}$ first undergoes a convolution with a convolution kernel of 1 × 1, and three vectors are obtained, called Query ($Q \in \mathbb{R}^{C2 \times H \times H}$), Key ($K \in \mathbb{R}^{C2 \times H \times H}$) and Value ($V \in \mathbb{R}^{C1 \times H \times H}$). Second, these three vectors were reshaped to obtain $Q \in \mathbb{R}^{N \times H \times C2}$, $K \in \mathbb{R}^{N \times C2 \times H}$ and $V \in \mathbb{R}^{N \times C1 \times H}$, where *N* represents the product of batch size and *H* during model learning. Then, $Q \in \mathbb{R}^{N \times H \times C2}$ and $K \in \mathbb{R}^{N \times C2 \times H}$ were subjected to the dot product operation, yielding the row-oriented attention matrix $A \in \mathbb{R}^{N \times H \times H}$. The specific calculation formula is shown in Equation (5). Then, the softmax activation function was applied to normalize the attention matrix $A \in \mathbb{R}^{N \times H \times H}$ as shown in Equation (6), and its weight result was transposed; then, the dot product calculation between the transposed weight result and $V \in \mathbb{R}^{N \times C1 \times H}$ was conducted as shown in Equation (7), and the result was reshaped as $X \in \mathbb{R}^{C1 \times H \times L0}$. For row-oriented attention, L_0 is 2. The output vector $O' \in \mathbb{R}^{C1 \times H \times H}$ was obtained by multiplying $X \in \mathbb{R}^{C1 \times H \times L0}$ with the identity matrix α and adding the result to the feature $I' \in \mathbb{R}^{C1 \times H \times H}$. Finally, the vector $O' \in \mathbb{R}^{C1 \times H \times H}$ was transformed into $O \in \mathbb{R}^{L \times C1}$ as the output. For column-oriented attention, the operation was similar, but the dimension of the vector X has changed, which is $\mathbb{R}^{C1 \times H \times H}$, that is, $L_0 = H = L^{1/2}$.

$$A = f(Q, K_i) = Q \cdot K_i \tag{5}$$

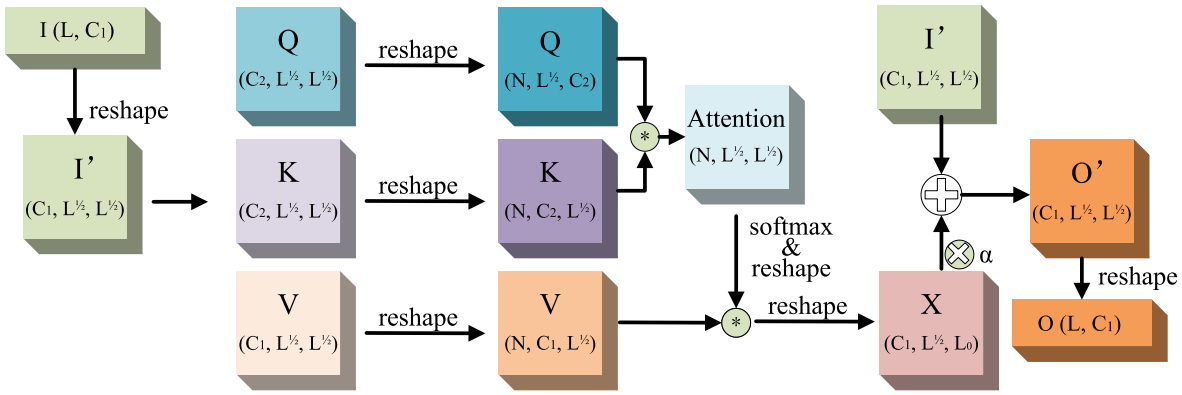

Fig. 3. Schematic diagram of axial attention.

$$a_i = \sigma(f(Q, K_i)) = \frac{\exp(f(Q, K_i))}{\sum_j \exp(f(Q, K_i))} \tag{6}$$

$$f\left(Q, K, V\right) = \sum_i a_i \cdot V_i \tag{7}$$

where i represents the i-th element of the vector $K \in \mathbb{R}^{N\times C2\times H}$, $i \in (0, L)$. The result was the similarity between the $Q \in \mathbb{R}^{N\times H\times C2}$ and $K \in \mathbb{R}^{N\times C2\times H}$ matrices, so the weight coefficients constitute the matrix $A \in \mathbb{R}^{N\times H\times H}$. Here, σ represents the softmax activation function, which normalizes the obtained matrix, $A \in \mathbb{R}^{N\times H\times H}$, and a_i represents the weight score of each element.

In general, combining row-oriented attention and column-oriented attention to form an axial attention mechanism is better able to integrate the global information, which is conducive to strengthening the dependencies in long text passages. In this study, the feature vectors generated by the BiLSTM were used as input, the row-oriented attention was calculated first, and then, the results were used as input of column-oriented attention to improve the accuracy of NRE in the geological field.

3.2.2. Graph convolution unit

In this study, a GCU was designed to enhance the expression of long-distance dependencies between entities in geological texts by capitalizing on the fact that graphs can express regional context dependence, as shown in Fig. 4.

Specifically, the feature $T \in \mathbb{R}^{L\times C}$, which is the fusion of the encoded position vector and the RoBERT encoded vector, was first applied as the input of the GCU. L is the length of the encoded vector, and C was the number of channel dimensions. We transform $T \in \mathbb{R}^{L\times C}$ into a two-dimensional vector $T' \in \mathbb{R}^{C\times H\times H}$, where $H = L^{1/2}$. Therefore, the regional context association is realized in a two-dimensional space, and long-distance dependencies with surrounding tokens can be strengthened. Then, $T' \in \mathbb{R}^{C\times H\times H}$ was decomposed into multiple different regions, where the different regions serve as different nodes of the graph. V represents the number of decomposed regions. As shown by the blue line of the grid in Fig. 4, V is divided into four regions, that is, $V = 4$. The probability that each token belongs to a given node in the two-dimensional feature graph is shown in Formula (8); thus, the probability matrix $Q \in \mathbb{R}^{H\times H\times C}$ was constructed.

$$q_{ij}^k = \frac{\exp\left(-\left|\left|(x_{ij} - w_k)\middle/\sigma_k\right|\right|_2^2/2\right)}{\sum_k \exp\left(-\left|\left|(x_{ij} - w_k)\middle/\sigma_k\right|\right|_2^2/2\right)} \tag{8}$$

where q_{ij}^k is the probability of each token belonging to region (node) k, $k \in (1,V)$, x_{ij} represents the feature vector of the token in the i-th row and j-th column of the 2D feature graph, and w_k was the feature vector of the k-th node. In addition, σ_k represents the variance of all the dimensions of node k, normalized to (0, 1) by the sigmoid function.

Furthermore, we considered the weighted average of the residual as the difference between each token and each node, as shown in Equation (9). The L2 regularization shown in Equation (10) was employed to obtain the encoded feature of each node. In these results, a smaller residual represents greater consistency between the encoded nodes and the actual situation:

$$z_k' = \frac{1}{\sum_{ij} q_{ij}^k} \sum_{ij} q_{ij}^k (x_{ij} - w_k) \Big/ \sigma_k \tag{9}$$

$$z_k = \frac{z_k'}{||z_k'||_2} \tag{10}$$

where z_k represents the encoded feature of the k-th node. Therefore, all z_k constitute the encoded matrix $Z \in \mathbb{R}^{C\times V}$.

In addition, we applied $A = Z^T Z$ to calculate an adjacency matrix, where each element in $A \in \mathbb{R}^{V\times V}$ expresses the cosine similarity between Z^T and Z. This matrix captures the long-distance dependencies between different nodes. Then, we employed the constructed transpose matrix Z^T

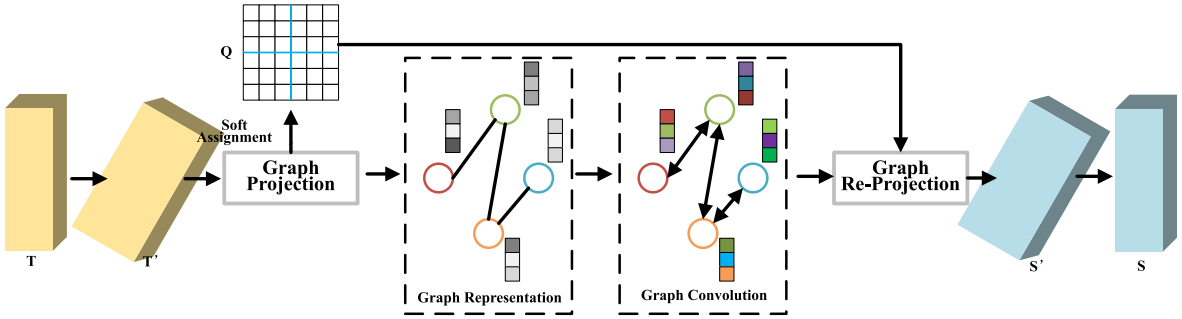

Fig. 4. The structure of GCU.

and adjacency matrix A to spread the feature information through graph convolution. The formula is shown in Equation (11):

$$\widetilde{Z} = f\left(AZ^T W_g\right) \tag{11}$$

where $\widetilde{Z}$ refers to the features obtained after the GCU operation, and f refers to the nonlinear activation function. We adopted a rectified linear unit (RELU) activation function in this study. W_g was a shape parameter of size $\mathbb{R}^{C\times C}$.

Finally, the feature after graph convolution processing was reprojected into a two-dimensional feature map S', as shown in Equation (12). Finally, the feature map was reconstructed into a one-dimensional feature $S \in \mathbb{R}^{L\times C}$:

$$\widetilde{S'} = \mathscr{C}\widetilde{Z}^T \tag{12}$$

In this study, to comprehensively integrate the local and global information, the number of nodes V was set to 4, 8, 16, 32, and these output features were integrated to more comprehensively establish long-distance dependency relationships at multiple scales.

4. Experimental results and discussion

4.1. Evaluation metrics

In this study, three evaluation metrics were adopted to measure the model performances: precision (P), recall (R) and F1-score (F1), whose formulas are shown in Equations 13–15.

$$P = \frac{TP}{TP + FP} \tag{13}$$

$$R = \frac{TP}{TP + FN} \tag{14}$$

$$F_1 = \frac{2PR}{P + R} \tag{15}$$

where *TP* represents true positives, *FP* represents false-positives, *FN* represents false negatives, and *TN* represents true negatives.

4.2. Experimental setup

The deep learning framework used in this study was PyTorch 1.7. We used Python3.7 and an RTX 2080Ti graphics card, and its memory is 11 GB. The initial learning rate of the model was set to 0.001 and reduced by 10 every 100 epochs. The loss function employed cross entropy, we used the Adam adaptive optimization gradient descent optimizer, and the weight attenuation was 10^{-5}. In the experiment, training was conducted for 300 epochs, and the batch sizes in both the training and validation stages were set to 10.

4.3. Experimental results and analysis

4.3.1. Comparison experiments

To verify the performance of GeoERE-Net in extracting triples from geological texts proposed in this study, this study conducted comparative experiments with a number of advanced network models proposed by prior researchers, including the pipeline model, the pipeline model using ALBERT-BiLSTM-CRF models during the geological NER stage and the ALBERT-BiGRU-CRF models during the geological RE stage. Additionally, the five existing nonpipeline-based methods showed their performance, including the multihead selection network (Bekoulis et al., 2018), NTS (Zheng et al., 2017), Seq2RDF (Liu et al., 2018), CasRel network (Wei et al., 2019) and TPLinker network (Wang et al., 2020).

Table 2 shows the quantitative description results of the proposed GeoERE-Net and of several advanced triple extraction methods. In terms of quantitative description results, the GeoERE-Net proposed in this study achieved the best result on the GIED dataset, with an F1-score 23.18% above that of the traditional pipeline method. Fig. 5 shows the counts of annotated triplet truth values as well as the predicted triplet results for each of the seven models; the results were recorded in terms of each specific relation.

Table 2
Results of different models based on the GIED dataset.

Methods	Precision (%)	Recall (%)	F1-score (%)
Pipeline	65.52	68.41	66.87
NTS	67.78	66.11	66.93
Seq2RDF	71.22	70.33	70.77
Multihead selection	85.78	83.48	84.62
CasRel	84.26	84.68	84.47
TPLinker	83.28	85.43	84.34
GeoERE-Net	90.23	89.87	90.05

The pipeline method is an easily implementable classical approach for extracting triples in which the NER and RE subprocesses run independently of each other, eventually forming a combination result. However, Table 1 shows that the pipeline method triplet extraction results were not excellent; its precision was 65.52%, its recall was 68.41%, and its F1-score is only 66.87%. However, due to the NER and RE subprocesses, the method results have great advantages. This method adopted the BIO annotation mode in which the identified subject entity precision in the validation set reaches 99.52%, and the identified object entity precision reaches 98.49%. During the RE subprocess, the precision of the validation set reached 80.37%. However, the precision of RE in this process was not particularly high, and error propagation can occur when combining the results into triples with the NRE results because of error accumulation. Simultaneously, Fig. 5 also clearly shows that the extraction results of triples with the relation "Exposed in" were an important reason for error accumulation, leading directly to low quantitative evaluation scores. For example, the actual triplet **("Great Cyanosis Formation", "Thickness", "1,400–1,600 m")** was incorrectly predicted as **("Great cyanosis Formation", "Exposed in", "1,400–1,600 m")**.

We also compared our proposed framework with nonpipeline-based methods. For NTS and Seq2RDF, our proposed model had a better performance of 23.12% when compared to NTS. Additionally, a higher F1-Score value of 19.28% was achieved when compared with Seq2RDF. These results proved that the proposed architecture is better than the NTS and Seq2RDF architectures at recognizing entities and extracting relations jointly. The two models have one thing in common. For each sentence, the models can only extract a triplet of the sentence instead of fully extracting multiple triples, so they cannot solve the problem of overlapping relations and intersection of geological entity relations. For example, "Yunkai Formation has a large exposed area, about 924.27 m^2, accounting for 5.46% of the total area of the survey area, and is widely distributed in Luoding-Xinyi-Yangchun area". The statement has a single-entity overlap problem, and the reasoning performance is weak in these two models. Both of them can only extract the triple structure of **("Yunkai Formation", "Area", "924.27m^2")**, but cannot extract the triple result of the relation "Exposed in".

The multihead selection model views entities and relations extraction as a multihead selection problem. In this study, multihead means that any specific entity may own multiple relations with others. The model contains multiple layers, including embedding layer, BiLSTM, CRF, label embedding, sigmoid, and head relations layers. After Chinese word segmentation and encoding, the BiLSTM layer was used to obtain contextual information to construct a more complicated vector format for each Chinese word. Second, the CRF layer and sigmoid layer were used to generate the output results, which included the entity recognition label (such as I-LOC, which expressed the type of entity as position), the set of the subject entity, and relations between two entities, such as (survey area, age). Then, the final triplet representation was obtained on this basis. As shown in Fig. 5, the model is not particularly friendly when

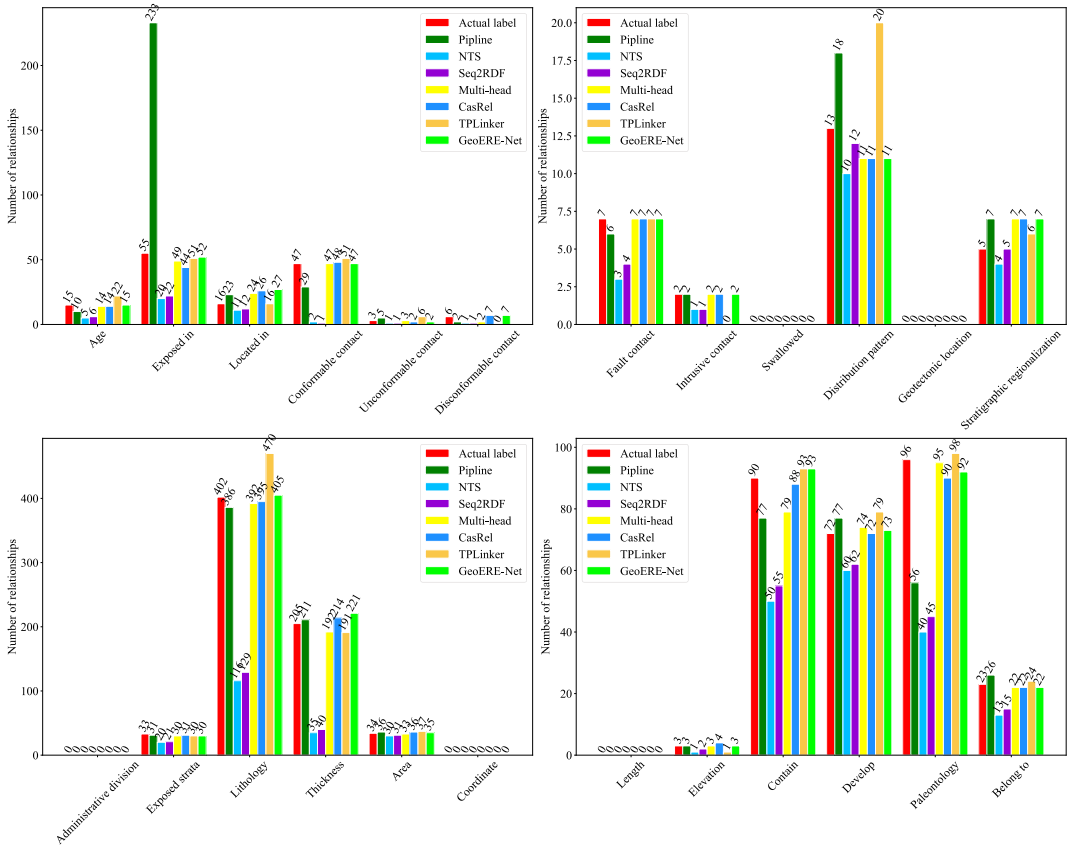

Fig. 5. The number of triples recognized by different models under specific relations.

dealing with relation overlap, which can be seen from the triplet results of "Lithology", "Thickness" and "Containing", which had a large number of missed prediction results. Additionally, the recall score of the quantitative analysis result is 83.48%, which indicated missed triplets. For example, the statement "There are Qingbaikou-Cambrian, Devonian-Carboniferous, Cretaceous and Quaternary strata in the survey area of Yangchun County," should be extracted as **("Yangchun County survey area", "Contain", "Qingbaikou-Cambrian strata")**, **("Yangchun County survey area", "Contain", "Devonian-Carboniferous strata")**, **("Yangchun County survey area", "Contain", "Cretaceous strata")**, **("Yangchun County survey area", "Contain", "Quaternary strata")**. However, some omissions occurred during the actual prediction, which lacks the triplet extraction of "Quaternary strata"; therefore, this approach fails to solve extracting the overlapping relations in geological texts.

The CasRel cascade network was proposed for overlapping triples. The main idea underlying this model was to use BERT to perform vector encoding for the input statements and then use the sigmoid function to identify the head and tail positions of the subject entity. Then, the object entities under the specific relation are judged based on the identified subject entities, so the relation and object entities were output as a set. This model effectively alleviates the overlapped triple problems that were found in text in the geological domain: "The Yongzhu Formation is characterized by black mudslate and siltstone". This was a statement involving multiple object entities that have a specific relation with a single subject entity. The model can accurately extract the two triples **("Yongzhu Formation", "Lithology", "Black mudslate")** and **("Yongzhu Formation", "Lithology", "Siltstone")**. Table 1 shows that the CasRel network was effective at extracting a large number of overlapping triples in geological texts that were similar to the above examples. Its precision score of 84.26% and recall score of 84.68% indicated that the model not only accurately extracts geological text triples but also ensured the integrity of triplet extraction.

TPlinker's annotation framework was designed to address the problem of relation overlap and exposure bias. Exposure bias referred to discrepancies between the prediction of RE using labeled entities during the training stage and the model's prediction of relations using entities during the inference stage. These differences between the training and inference processes result in error accumulation. In this study, all the quantitative evaluation indices of TPlinker were high, and its recall value reached 85.43%. Additionally, Fig. 5 shows that the model predicts more triples than the true number of annotated triples for almost every specific relation in the model. Therefore, TPlinker greatly improves the triplet extraction integrity of statements and reduces the possibility of missed points. However, its precision score of 83.28% indicated that TPlinker lacks high precision. For example, in the statement "Lanweng Formation, developed banded structure, showing the characteristics of the BaoMa sequence," the two extracted triples should be **("Lanweng Formation", "Develop", "Banded structure")**, **("Lanweng Formation", "Develop", "BaoMa sequence")**, but TPlinker extracted one extra result, **("Lanweng Formation", "Develop", "Banded structure, showing the characteristics of the Baoma sequence")**.

The GeoERE-Net network proposed in this study establishes long-distance dependencies and global information among the input statement vectors through axial attention based on the BiLSTM. This module was applied to enhance subject entity recognition. In addition, context dependencies were established through GCUs at different scales. This module was employed during the object entity recognition stage for a specific relation. As seen in Table 2 and Fig. 5, the GeoERE-Net model proposed in this study achieves high precision and recall rates (90.23% and 89.87%, respectively) when constructing geological text triplets, which was highly useful in the construction of knowledge graphs.

4.3.2. Ablation experiments

To verify the enhancement effect of axial attention based on a BiLSTM and the multiscale GCU introduced in this study, successive ablation experiments were conducted on the proposed structures. The results are shown in Table 3. We assessed the impact on the model's performance after removing specific individual modules.

The ablation experimental results are shown in Table 3. The F1-scores of the proposed GeoERE-Net with different module combinations were high, reaching 90.05%, 87.61%, 85.90% and 84.47% for the various combinations. GeoERE-Net achieves the best experimental results when the full proposed model was applied to triplet extraction from geological text. When we remove a module, the triplet extraction result accuracy decreases. In this study, the precision decreased by 3.42% and the recall decreased by 1.44% when the GCU module was removed. As Fig. 6 shows, the main reason for this accuracy reduction lies in the triplet extraction results that include "Exposed in", "Located in", "Thickness", "Area" and "Containing" relations, which may result in incorrect relation classifications. Triples such as **("Bali Formation", "Exposed in", "Yangchun—East of Yunan Town, Helang")** were incorrectly classified as **("Bali Formation", "Exposed in", "Yangchun—East of Yunan Town, Helang, near Donghu Reservoir, south of Enping")**. The error occurs mainly due to a lack of dependence on text context information, leading to the redundant classification "near Donghu Reservoir, south of Enping". This result also indicated that the GCU module plays a significant role during the object entity extraction phase for a specific relation. The GCU considers the multiscale context information in the encoded vector, which can strengthen the long-distance dependencies between tokens. In addition, when both the GCU module and axial attention based on the BiLSTM module are removed simultaneously, the model results fluctuated substantially; precision and recall decrease by 1.96% and 1.45%, respectively. These results indicated that axial attention based on the BiLSTM module can be employed to strengthen the dependence between token vectors and enhance the global information during subject entity extraction; therefore, triplets can be extracted more accurately and completely. Additionally, we also compared the influence of RoBERT and BERT encoding methods on the experimental results. As seen from Table 2 and Fig. 6, the integrity of extracted triplets can be improved by replacing BERT with RoBERT encoding, which improved recall by 2.3%. In general, the two modules proposed in this study offer help that significantly improves the performance of the model.

4.4. Case study

To show the generalizability and practical application value of the proposed GeoERE-Net for extracting triplets from geological texts, we selected a regional geological survey report conducted in Nima County, Tibet, China. We selected a section of Chapter 2 that describes the strata as a case study. This section mainly introduces Carboniferous strata. The Carboniferous strata in the survey area are in the Ganglong-Shenya and spread in an east–west direction. The geotectonic location is located at the northern edge of the Gangdese structural belt. The stratigraphic division belongs to the Bangor-Basu strata in the Gangdese-Tengchong stratigraphic area. The exposed strata are the Lower-Upper Carboniferous Yongzhu Formation and Upper Carboniferous-Lower Permian Laga Formation. The lithology consists of clastic rocks, the thickness of the strata is approximately 2,725 m, and the exposed area is approximately 1,500 km^2.

Table 3
Results of ablation experiments on the GIED dataset.

Methods	Precision (%)	Recall (%)	F1-Score (%)
GeoERE-Net	90.23	89.87	90.05
-Graph convolution unit	86.81	88.43	87.61
-Axial attention	84.85	86.98	85.90
-BERT(RoBERT)	84.26	84.68	84.47

After applying GeoERE-Net to extract triples from the strata description texts, knowledge graphs of the strata can be constructed. Through the knowledge graph, not only can the strata distribution structure, rock development and other related geological information be visually displayed in a structured form, but simple retrieval and query can also be carried out. By comparing the original geological report texts and including a small amount of manual intervention, a knowledge graph structure of the Carboniferous strata was finally presented. Neo4j (https://neo4j.com/) is currently the mainstream graph database (compared to titan, OrientDB, JanusGraph, HugeGraph, Trinity) with rich features and excellent performance. The single-node server can host hundreds of millions of nodes and relations with high traversal efficiency. As shown in Fig. 7, we used the neo4j database to store and display knowledge graphs. The red rectangles in Fig. 7(a) intuitively showed that the three nodes of the "Carboniferous", "Yongzhu Formation" and "Laga Formation" are the primary nodes and have the most connections. This was because the "Carboniferous" stratum is composed of the "Yongzhu Formation" and the "Laga Formation". Therefore, most of the nodes in the graphs were related to these three nodes. Taking the "Yongzhu Formation" as the central node, it was primarily connected to nodes that express attributes such as "paleontology", "rocks", "length", and "thickness". Additionally, the adjacency relation "Conformable contact" between the adjacent "Laga Formation" strata and the exposed location of this group, "Nima County", can be accurately expressed in the knowledge graph.

In addition, through the geological knowledge graph, we can conduct query and retrieval operations such as:

(1) Query all nodes that have the specific relation ("Exposed in").
Query sentence: MATCH p = () - [r: "Exposed in"]->() RETURN p
Result: Fig. 7(b)

Significance: This approach specifically counts all the entities with this relation that exist, allowing a better understanding of all the entity types that have this relation. In this study, it can be judged that the entity types under the relation "Exposed in" were the strata node and the location node.

(2) Query the surrounding nodes and relations for a certain node ("horizontal wormhole").
Query sentence: MATCH (n: Node {name: "Horizontal wormhole"})-[r*.2]-(m) RETURN n, r, m
Result: Fig. 7(c)

Significance: Further reasoning can obtain hidden content that is difficult to find in the original unstructured text. For example, by searching for the "horizontal wormhole" node, we can determine which rocks contain such features; then, combined with geological knowledge, we can compare and infer the relationships between these fossils and their evolution characteristics.

5. Discussion

5.1. Benefits

Natural language processing and deep learning are effective in rapidly extracting domain-specific information from a large amount of texts. This study presented a computational framework for extracting entities and related relations from geological reports using a joint extraction model. The proposed framework can automatically 'read' a large number of geological reports via recognize entities and relations for understanding large-scale content analysis. The framework can also perform graph-based knowledge representations and was useful for applications that explore, search and recommend information. KGs were automatically generated by avoiding possible human fatigue, bias, or

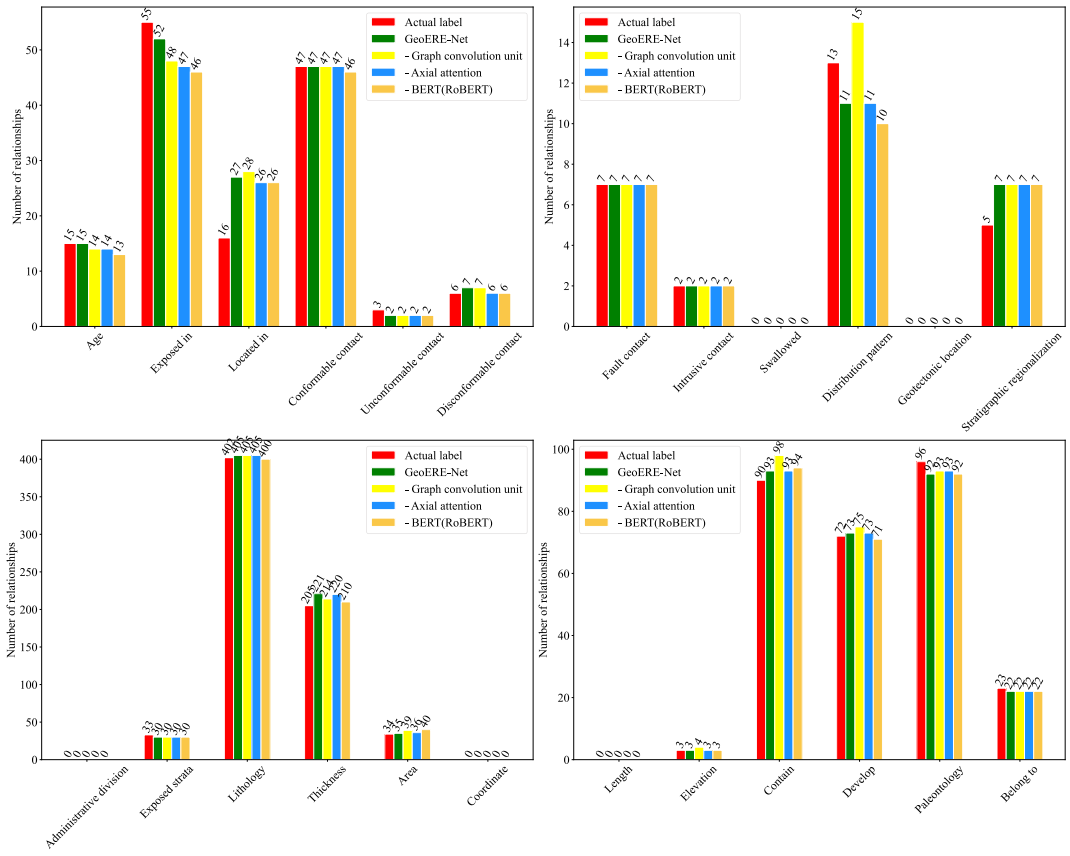

Fig. 6. The number of triples recognized by different module combinations under specific relations.

misunderstandings associated with reading and comprehending lengthy and complex contents in geological reports.

5.2. Limitations

Although the experiments demonstrated that the proposed approach provides better results, as the results showed, some failures still occur when processing the test dataset because sparse data will affect the initial training using the proposed method because it lacks the necessary training information. In some cases, GeoERE-Net failed to identify unpredetermined geological entity relations in a regular sentence. One simple way to address the issue might be to use a statistical approach as prior knowledge. For example, the lexical, location and distance characteristics of words were counted by the bootstrapping technique to calculate the word weights in the context, whereby the keywords describing the relationships of geological entities were determined, and the relationships were identified by the keywords.

Therefore, there were also some limitations to this work, such as that we only experimented on the geological dataset. For future work, we can also plan to test our proposed deep neural models, on the task of recognizing entities and relations collected from other datasets/languages. These can include datasets of geological reports collected from resources such as the USGS (https://www.usgs.gov/), Department of Mines, Industry Regulation and Safety of Government of Western Australia (https://www.dmirs.wa.gov.au/), or British Geological Survey (https://www.bgs.ac.uk/). Experiments with these data would allowed us to assess the degree to which the inferred models can properly generalize to different contexts.

In this study, entities and relations are currently extracted from geological texts using a deep learning model, and disambiguation of the extracted entities is required in the future. A similar strategy could be applied to compare the original entity vector to the existing vectors in a semantic space that has been created by using the same embedding models. Using the cosine and Euclidean distances, we established a ranking where the existing entities are sorted according to their distances to the original entity. In addition, there were different representations of temporal and spatial information in geological texts. A simple strategy was to construct a unified ontological framework for temporal and spatial representation. This framework contained the time scale, time granularity, time type, etc., as well as the reference system, space type, and conversion between coordinates. By constructing a unified spatiotemporal mapping function to transform, map and store the spatiotemporal information of geological text.

6. Conclusion and future work

This study proposed a new joint entity and relation extraction model that can be applied to extract triplet information from geological texts and finally constructed a knowledge graph for geological findings. First, we used multiple Chinese regional geological survey reports to construct a study corpus. Second, we designed an axial attention mechanism based on BiLSTM and GCU modules to solve the existing problems of the current methods, which enhances the completeness and accuracy of triplet extraction. Finally, we applied another regional geological report to verify the performance of our proposed model. The constructed knowledge graph suggested that it has strong application value for the proposed model. However, this study does not fully consider the technical terminology of geological texts. Generally, geological texts are relatively long. Therefore, this model outputs incorrectly a small number of long geological entities as shorter-entity content. Additionally, the influence of entity disambiguation was not considered when constructing the knowledge graph. In future work, strengthening the constraints of geological terminology and geological knowledge is an urgent consideration. We will also consider entity disambiguation to extract

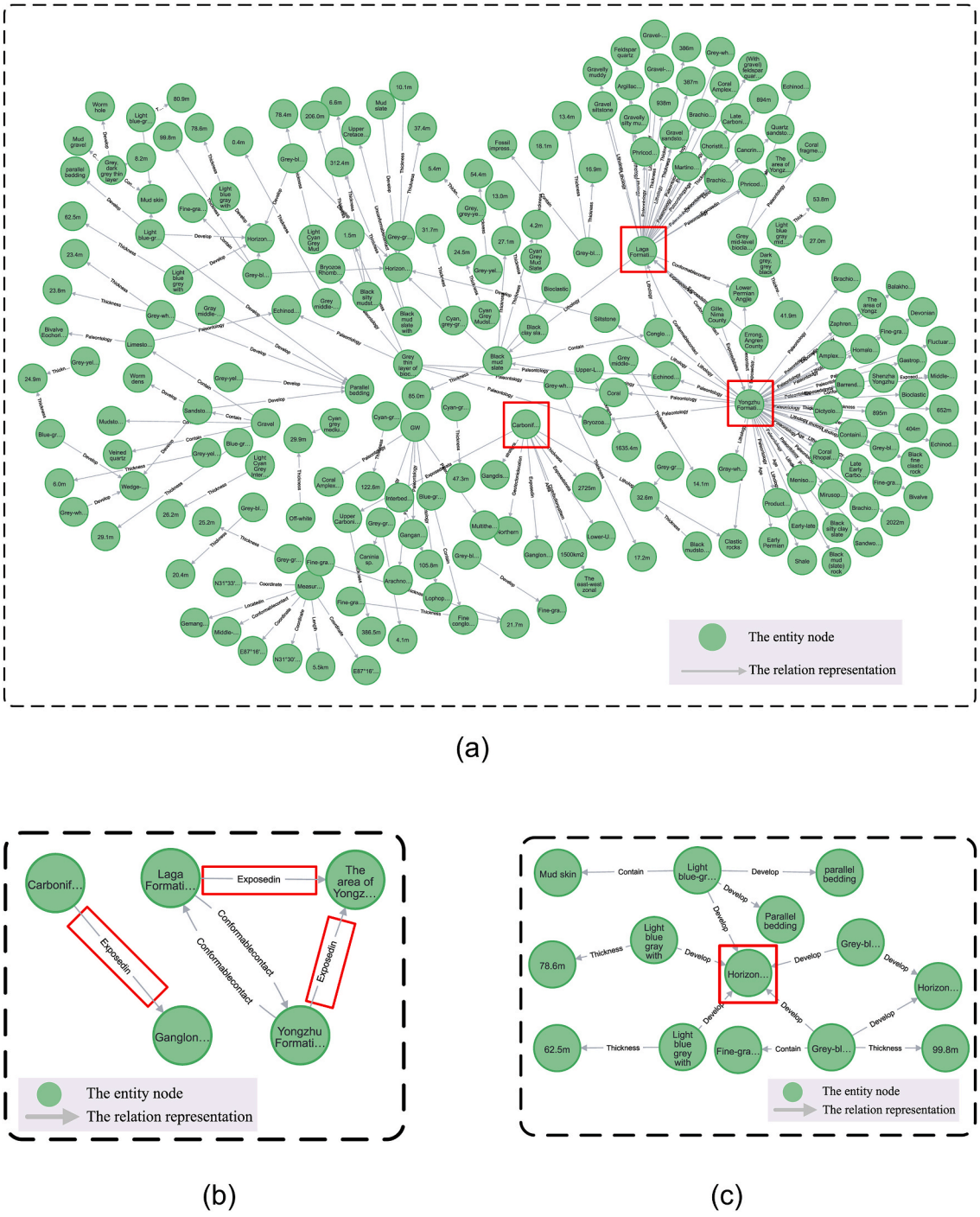

Fig. 7. Knowledge graphs constructed from descriptive texts of Carboniferous strata. The green circular nodes represent entity nodes, and the arrows with text descriptions represent relations between entities. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

triplets more accurately and completely to build large-scale knowledge graphs of the geological domain.

Code availability

https://github.com/CUGGISer/GeoERE-Net.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgments

This work is supported by the Deep-time Digital Earth (DDE) Big Science Program. This study was financially supported by the National Natural Science Foundation of China (42050101, 41871311,

41871305), the China Postdoctoral Science Foundation (No.2021M702991), Open Research Project of The Hubei Key Laboratory of Intelligent Geo-Information Processing (No. KLIGIP-2021A01).

References

Bekoulis, G., Deleu, J., Demeester, T., Develder, C., 2018. Joint entity recognition and relation extraction as a multi-head selection problem. Expert Syst. Appl. 114, 34–45.

Consoli, B., Santos, J., Gomes, D., Cordeiro, F., Vieira, R., Moreira, V., 2020. Embeddings for named entity recognition in geoscience Portuguese literature. In: Proceedings of the 12th Language Resources and Evaluation Conference, pp. 4625–4630.

Deng, C., Jia, Y., Xu, H., Zhang, C., Tang, J., Fu, L., Zhang, W., Zhang, H., Wang, X., Zhou, C., 2021. GAKG: a multimodal geoscience academic knowledge graph. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, pp. 4445–4454.

Enkhsaikhan, M., Holden, E.-J., Duuring, P., Liu, W., 2021a. Understanding ore-forming conditions using machine reading of text. Ore Geol. Rev., 104200

Enkhsaikhan, M., Liu, W., Holden, E.-J., Duuring, P., 2021b. Auto-labelling entities in low-resource text: a geological case study. Knowl. Inf. Syst. 63, 695–715.

Fan, R., Wang, L., Yan, J., Song, W., Zhu, Y., Chen, X., 2020. Deep learning-based named entity recognition and knowledge graph construction for geological hazards. ISPRS Int. J. Geo-Inf. 9, 15.

Fensel, D., Şimşek, U., Angele, K., Huaman, E., Kärle, E., Panasiuk, O., et al., 2020. Introduction: what is a knowledge graph?. In: Knowledge Graphs. Springer, Cham, pp. 1–10.

Gupta, P., Rajaram, S., Schütze, H., Runkler, T., 2019. Neural relation extraction within and across sentence boundaries. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 6513–6520.

He, Y., Li, Z., Yang, Q., Chen, Z., Liu, A., Zhao, L., Zhou, X., 2020. End-to-end relation extraction based on bootstrapped multi-level distant supervision. World Wide Web 23, 2933–2956.

Huang, X., Zhu, Y., Fu, L., Liu, Y., Tang, K., Li, J., 2021. Research on a geological entity relation extraction model for gold mine based on BERT. J. Geomechanics 27, 391–399.

Jianping, C., Jie, X., Qiao, H.U., Wei, Y., Zili, L., Bin, H.U., Wei, W., 2016. Quantitative geoscience and geological big data development: a review. Acta Geologica Sinica-English Edition 90, 1490–1515.

Jyothi, P.B.B., Jyothi, S., Sekar, K., 2008. Knowledge extraction using rule based decision tree approach. IJCSNS 8, 296.

Li, H., Wu, X., Li, Z., Wu, G., 2013. A relation extraction method of Chinese named entities based on location and semantic features. Appl. Intell. 38, 1–15.

Li, Q., Ji, H., 2014. Incremental joint extraction of entity mentions and relations. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, vol. 1. Long Papers, pp. 402–412.

Li, R., Zhang, X., Li, C., Zheng, Z., Zhou, Z., Geng, Y., 2021. Keyword extraction method for machine reading comprehension based on natural language processing. In: Journal of Physics: Conference Series. IOP Publishing, p. 012072.

Li, Y., Bontcheva, K., Cunningham, H., 2004. An SVM based learning algorithm for information extraction. Mach. Learn. 1.

Liang, R., Tang, P., Xiong, G., Liu, Z., Yu, D., 2021. A review on sustainable development of geological exploration technology and risk management. Recent Pat. Eng. 15, 45–52.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. Roberta: A Robustly Optimized Bert Pretraining Approach arXiv preprint. 1907.11692.

Liu, Y., Zhang, T., Liang, Z., Ji, H., McGuinness, D.L., 2018. Seq2RDF: an End-To-End Application for Deriving Triples from Natural Language Text arXiv preprint. 1807.01763.

Lun, C.H., Hewitt, T., Hou, S., 2021. Extracting knowledge with NLP from massive geological documents. In: 82nd EAGE Annual Conference & Exhibition. European Association of Geoscientists & Engineers, pp. 1–5.

Luo, X., Zhou, W., Wang, W., Zhu, Y., Deng, J., 2017. Attention-based relation extraction with bidirectional gated recurrent unit and highway network in the analysis of geological data. IEEE Access 6, 5705–5715.

Ma, X., 2021. Knowledge Graph Construction and Application in Geosciences: A Review.

McManus, S., Rahman, A., Coombes, J., Horta, A., 2021. Uncertainty assessment of spatial domain models in early stage mining projects–a review. Ore Geol. Rev. 133, 104098.

Minard, A.-L., Ligozat, A.-L., Grau, B., 2011. Multi-class SVM for relation extraction from clinical reports. In: Recent Advances in Natural Language Processing.

Paulheim, H., 2017. Knowledge graph refinement: a survey of approaches and evaluation methods. Semantic Web 8 (3), 489–508.

Qin, H., Tian, Y., Song, Y., 2021. Relation extraction with word graphs from N-grams. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 2860–2868.

Qiu, Q., Xie, Z., Wu, L., Tao, L., Li, W., 2019. BiLSTM-CRF for geological named entity recognition from the geoscience literature. Earth Science Informatics 12, 565–579.

Qiu, Q., Xie, Z., Xie, H., Wang, B., 2021. GKEEP: an enhanced graph-based keyword extractor with error-feedback propagation for geoscience reports. Earth Space Sci. 8 e2020EA001602.

Qun, Y., Linfu, X., Zeyu, L., Xin, G., Rui, W., Junhao, D., 2021. Construction of deposit model-oriented knowledge graph. In: IOP Conference Series: Earth and Environmental Science. IOP Publishing, p. 012034.

Ravikumar, K.E., Rastegar-Mojarad, M., Liu, H., 2017. BELMiner: adapting a rule-based relation extraction system to extract biological expression language statements from bio-medical literature evidence sentences. Database 2017.

Sari, Y., Hassan, M.F., Zamin, N., 2010. Rule-based pattern extractor and named entity recognition: a hybrid approach. In: 2010 International Symposium on Information Technology. IEEE, pp. 563–568.

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. IEEE Access 6, 52286–52297.

Singh, S., Riedel, S., Martin, B., Zheng, J., McCallum, A., 2013. Joint inference of entities, relations, and coreference. In: Proceedings of the 2013 Workshop on Automated Knowledge Base Construction, pp. 1–6.

Sobhana, N., Mitra, P., Ghosh, S.K., 2010. Conditional random field based named entity recognition in geological text. Int. J. Comput. Appl. 1, 143–147.

Sui, D., Chen, Y., Liu, K., Zhao, J., Zeng, X., Liu, S., 2020. Joint Entity and Relation Extraction with Set Prediction Networks arXiv preprint. 2011.01675.

Wan, Z., Xie, J., Zhang, W., Huang, Z., 2019. BiLSTM-CRF Chinese named entity recognition model with attention mechanism. In: Journal of Physics: Conference Series. IOP Publishing, p. 032056.

Wang, B., Wu, L., Li, W., Qiu, Q., Xie, Z., Liu, H., Zhou, Y., 2021. A semi-automatic approach for generating geological profiles by integrating multi-source data. Ore Geol. Rev., 104190

Wang, C., Hazen, R.M., Cheng, Q., Stephenson, M.H., Zhou, C., Fox, P., Shen, S., Oberhänsli, R., Hou, Z., Ma, X., 2021. The Deep-Time Digital Earth program: data-driven discovery in geosciences. Natl. Sci. Rev. 8 nwab027.

Wang, C., Ma, X., Chen, Jianguo, Chen, Jingwen, 2018. Information extraction and knowledge graph construction from geoscience literature. Comput. Geosci. 112, 112–120.

Wang, Y., Jin, L., Zhu, Y., Bai, M., Bao, X., 2018. Development of keyword extraction algorithm for geoscience unstructured document based on big data. Prog. Geophys. 33, 1274–1281.

Wang, Y., Yu, B., Zhang, Y., Liu, T., Zhu, H., Sun, L., 2020. Tplinker: Single-Stage Joint Extraction of Entities and Relations through Token Pair Linking arXiv preprint. 2010.13415.

Wang, Y., Zhang, H., 2021. BIRL: bidirectional-interaction reinforcement learning framework for joint relation and entity extraction. In: International Conference on Database Systems for Advanced Applications. Springer, pp. 483–499.

Wei, D., Jiang, B., Zhang, J., 2021. Research on content storage method of unstructured geological data. Northwest. Geol. 54, 266–273.

Wei, D., Li, C., Naheman, W., Wei, J., Yang, J., 2014. Organizing and storing method for large-scale unstructured data set with complex content. In: 2014 Fifth International Conference on Computing for Geospatial Research and Application. IEEE, pp. 70–76.

Wei, Z., Su, J., Wang, Y., Tian, Y., Chang, Y., 2019. A Novel Cascade Binary Tagging Framework for Relational Triple Extraction arXiv preprint. 1909.03227.

Yang, B., Cardie, C., 2013. Joint inference for fine-grained opinion extraction. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, vol. 1. Long Papers, pp. 1640–1649.

Yu, B., Zhang, Z., Shu, X., Wang, Y., Liu, T., Wang, B., Li, S., 2019. Joint Extraction of Entities and Relations Based on a Novel Decomposition Strategy arXiv preprint. 1909.04273.

Zhang, M., Zhang, Y., Fu, G., 2017. End-to-end neural relation extraction with global optimization. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 1730–1740.

Zhang, N., Xu, X., Tao, L., Yu, H., Ye, H., Xie, X., et al., 2022. Deepke: A Deep Learning Based Knowledge Extraction Toolkit for Knowledge Base Population arXiv preprint. 2201.03335.

Zhao, H., Xie, Q., 2021. An improved TextRank multi-feature fusion algorithm for keyword extraction of educational resources. In: Journal of Physics: Conference Series. IOP Publishing, p. 012021.

Zhao, Z., Du, J., Zou, C., Hu, S., 2011. Geological exploration theory for large oil and gas provinces and its significance. Petrol. Explor. Dev. 38, 513–522.

Zheng, S., Wang, F., Bao, H., Hao, Y., Zhou, P., Xu, B., 2017. Joint Extraction of Entities and Relations Based on a Novel Tagging Scheme arXiv preprint. 1706.05075.

Zhou, C., Wang, H., Wang, C., Hou, Z., Zheng, Z., Shen, S., Cheng, Q., Feng, Z., Wang, X., Lv, H., 2021. Prospects for the research on geoscience knowledge graph in the Big Data Era. Sci. China Earth Sci. 1–11.

Zhu, J., Nie, Z., Wen, J.-R., Zhang, B., Ma, W.-Y., 2005. 2d conditional random fields for web information extraction. In: Proceedings of the 22nd International Conference on Machine Learning, pp. 1044–1051.