Earth and Space Science

RESEARCH ARTICLE
10.1029/2019EA000610

GNER: A Generative Model for Geological Named Entity Recognition Without Labeled Data Using Deep Learning

Qinjun Qiu[1,2], Zhong Xie[1,2], Liang Wu[1,2], and Liufeng Tao[1,2]

[1]School of Information Engineering, China University of Geosciences, Wuhan, China, [2]National Engineering Research Center for GIS, Wuhan, China

Key Points:
- Geological named entities are extracted from unstructured Chinese geoscience reports using deep learning
- The proposed framework can be easily extended to other subject domains through fine-tuning
- A training data set is constructed based on both domain-specific entities and domain-general words

Correspondence to:
L. Wu,
wuliang@cug.edu.cn

Citation:
Qiu, Q., Xie, Z., Wu, L., & Tao, L. (2019). GNER: A generative model for geological named entity recognition without labeled data using deep learning. *Earth and Space Science*, *6*, 931–946. https://doi.org/10.1029/2019EA000610

Received 27 FEB 2019
Accepted 30 APR 2019
Accepted article online 9 MAY 2019
Published online 14 JUN 2019

Abstract A variety of detailed data about geological topics and geoscience knowledge are buried in the geoscience literature and rarely used. Named entity recognition (NER) provides both opportunities and challenges to leverage this wealth of data in the geoscience literature for data analysis and further information extraction. Existing NER models and techniques are mainly based on rule-based and supervised approaches, and developing such systems requires a costly manual effort. In this paper, we first design a generic stepwise framework for domain-specific NER. Following this framework, domain-specific entities and domain-general words are collected and selected as seed terms. Normalization and grouping processes are then applied to these seed terms for further analysis. A random extraction algorithm based on a unigram language model is used to generate a large-scale training data set consisting of probabilistically labeled pseudosentences. Each generated sentence is then used as input to the self-training and learning algorithm. Experimental results on two constructed data sets demonstrate that the proposed model effectively recognizes and identifies geological named entities.

Plain Language Summary Existing entity recognition and classification methods are less functional for automatic geological name entity recognition. In this paper, we propose a stepwise unsupervised approach to geological name entity recognition in geological reports that requires no labeled data. This approach dynamically adapts to extract and identify unseen instances. We hope that our approach will serve as an alternative method that deserves further study.

1. Introduction

As engagement in the geoscience domain continues to grow within governmental agencies and scientific organizations and the number of related computational models expands, an overwhelming amount of geoscience text is becoming available in a variety of digital forms and in numerous languages (Lima et al., 2017; Wu et al., 2017; Xiao et al., 2016; Zheng et al., 2015). Such free text is often recorded in either structured or unstructured forms (e.g., technical reports, geological reports, books, and other types of reports), thus posing challenges for engineers and scientists who need to effectively manage, share, analyze, and reuse all these online data (Cernuzzi & Pane, 2014; Ma, 2017; Wang et al., 2018). As a key collection of open data, the geoscience literature is a rich resource that can facilitate in knowledge discovery and information extraction because this literature contains voluminous meaningful information and expertly defined data that can be applied to train new models and enrich our understanding (Cracknell & Reading, 2014; Qiu et al., 2018; Wang et al., 2015).

The first step in automatically extracting potential geological information from the vast number of extant geological reports is to develop a named entity recognition (NER) system. This is a crucial part of various extraction systems and applications. NER is important in fields such as text clustering (Chen et al., 2018), information extraction (Huang et al., 2018), information retrieval (Dai et al., 2018), and automatic text summarization (Enríquez et al., 2017). Geological named entity recognition (GNER; also referred to as geological concept and element identification or the mapping of geological concepts and elements with similar properties) is a key process in geological language processing in which relevant terms (including single words and multiple-word phrases) are used to identify and classify content into predefined sematic categories. Such information and knowledge can enable scientists and engineers to participate more actively in geological investigations. In recent years, several research efforts have been devoted to developing standard scalable and flexible knowledge bases and terminologies to facilitate information extraction and reasoning using raw data. The bottleneck of geological information processing has shifted from finding free resources and data to making full use of those free raw resources and developing standard and scalable models to process the fast-growing collection of available text corpora (Shi et al., 2018; Tran et al., 2017; Zhu & Iglesias, 2018).

Consequently, there is a need to develop flexible and scalable models and approaches that enable the automatic recognition and identification of key information/data concerning geological named entities from a large number of unstructured textual geological reports. Then, the extracted important information should be transformed into a structured form to enable further flexible analysis. However, existing techniques and models for automatic GNER and classification are limited in their capabilities compared to other NER efforts involving geological reports for three reasons (Goyal et al., 2018; Liu & El-Gohary, 2017; Zhou & El-Gohary, 2017). First, compared to domain-general narrative text, geological reports are highly ambiguous and variable, in terms of both their textual characteristics and patterns because these reports are typically written by different inspectors/experts from a variety of scientific organizations and governmental agencies. Unstructured geological reports include domain-specific characteristics and unique aspects; they contain enormous amounts of high-level complex concepts and identified technical details (e.g., geophysical surveys, geochemical tests, and rock parameters). Second, the existing approaches to NER (e.g., those based on rules or supervised learning) often require considerable manual effort to construct a comprehensive and robust set of representative features/patterns or to annotate training data to address the high variability in free text patterns. The main limitation of the supervised training approach is that it requires excessive amounts of human labor to manually label the training data set. However, in real-world situations (e.g., in the geoscience domain), such labeled data sets are not available. The process of data set annotation is challenging and time-consuming because the annotators require not only specialized domain knowledge but also a background in natural language processing. Third, compared with English-language models, the processing of geoscience reports in Chinese in particular faces greater challenges because of the lack of spaces between words in the Chinese language, making it difficult to automatically identify the meanings of words/phrases. Hence, there is a need to develop an unsupervised model that can automatically extract and identify geological named entities from complex and variable-free texts.

To tackle these challenges, we propose a stepwise unsupervised approach to GNER in geological reports that requires no labeled data. This approach dynamically adapts to extract and identify unseen instances. To train the generative model, the method takes as input a collection of domain-specific entities and domain-general words along with their corresponding frequencies (referred to as sample seeds). The generative model then randomly chooses among the seeds to generate new sentences using an attention-based bidirectional long short-term memory (Bi-LSTM) model (Fernando et al., 2018; Wang et al., 2019). The process of generating sentences based on a unigram language model is iteratively repeated until the algorithm converges (Guo et al., 2017; Qiu et al., 2018; Quijano-Sánchez et al., 2018).

The main contributions of the approach proposed in this paper are threefold. (1)We address the GNER issue specifically for the geoscience domain, which is an understudied but important problem. To our knowledge, this is the first study to extract geological named entities from unstructured Chinese geoscience reports using deep learning (DL). (2)We propose a theoretical framework for domain-specific GNER that incorporates DL methodology and algorithms into GNER. This framework can be easily extended to other subject domains through fine-tuning. (3)We propose a novel generative model/methodology for constructing a training data set based on both domain-specific entities and domain-general words. This method dynamically generalizes the pseudosentences used as the input training data in a concave DL model.

The remainder of this paper is organized as follows. Section 2 details the related research. Section 3 presents the preliminaries and the problem formulation. Section 4 describes the algorithm, and section 5 reports and discusses the experimental results. Section 6 presents a discussion. Finally, section 7 offers conclusions and prospects for future work.

2. Related Work

The task of NER is to identify terms or concepts that have similar properties from a given set of data. Existing NER approaches can be grouped into three main categories: rule-based approaches, statistics-based approaches, and hybrid approaches. In this subsection, we review some related works from the above perspective.

2.1. Rule-Based Approaches

Early rule-based NER approaches mostly depended on sets of various handcrafted rules or matching patterns that supported extraction and recognition of target entities from unstructured textual data (Liu & El-Gohary, 2017; Nadeau & Sekine, 2007). Early systems relied on lexical resources and heuristic rules, including information from programs such as WordNet and gazetteers, to recognize and classify target named entities. Such rule-based systems can be regarded as highly efficient because they explore and use large amounts of language-related knowledge (Sarawagi, 2008). The existing approaches exploit domain-specific syntactic and semantic text features to achieve high precision performance.

Shaalan (2010) proposed the idea of a local grammar combined with a set of names for identifying entities with Arabic names. The goal of a filtering technique is to correct the output of an NER by filtering out incorrect named entities. Riaz (2010) defined a detailed set of rules and patterns for Urdu NER to address issues such as the agglutinative nature of the language, lack of capitalization, and spelling variations. The authors presented a cross-domain analysis concerning the transference of information from Hindi to Urdu, but these two languages are closely related. The work of Singh et al. (2012) presented a method for using various rules and patterns to extract named entities in Urdu. These authors constructed a group of dictionaries used to find and identify different entities in data sets from different domains. Rahem and Omar (2015) studied drug-related entities using a rule-based approach. In their method, many heuristics and grammatical rules were constructed for identifying various classes, including the prices of some drugs, types of drugs, and numbers of drugs. Quimbaya et al., (2016) defined a set of dictionaries for extracting named entities from electronic free health records. They applied stemmed matching and fuzzy matching approaches to identify relevant named entities such as treatments and diagnoses.

However, some disadvantages of rule-based approaches are that they require a large effort to develop comprehensive and representative knowledge bridging the language and the topic domain as well as programming skills to foster further development (Shaalan, 2010). Consequently, rule-based approaches are impossible to transfer across domains: rule-based NER techniques developed for one domain cannot be extended to other domains or adapted to meet the needs of researchers using other approaches (e.g., machine learning [ML] methods).

2.2. Statistics-Based Approaches

2.2.1. Supervised Learning

Supervised learning approaches focus on the idea of learning from labeled training examples, including both positive and negative examples. Using this approach, adaptive features can be developed from a large number of examples, and appropriate algorithms can be constructed to distinguish between positive and negative examples based on combinations of these features and the identification of similar information from unseen instances (data).

Supervised learning algorithms heavily rely on training data that are manually labeled by domain experts or researchers, a task that is both labor intensive and time-consuming. The annotated data are then input into algorithms for training, and a model is obtained that is further applied to classify and recognize named entities in the training data set or in test data. The features considered in these algorithms play an important role in supporting the multidimensional aspect of free text forms; these features are applied by the learning algorithms to produce models. A generative model has the ability to identify patterns that can be used to recognize similar data and classify examples (both positive and negative data).

Selecting appropriate features for a learning algorithm is a crucial task. Various learning models and techniques have been widely used by NER researchers, such as the conditional random field (CRF) models (Liu & Zhou, 2013; Majumder et al., 2012), support vector machine (SVM) models (Saha et al., 2010), hidden Markov models (HMMs; Wang et al., 2014), logistic expression models (Ekbal & Saha, 2011), and maximum entropy Markov models (Saha et al., 2010).

2.2.2. Semisupervised Learning

Semisupervised learning methods aim to learn from a substantial set of both labeled and unlabeled data to extract units of target information (e.g., named entities). The existing models and techniques rely on information-theoretic regularization, bootstrapping strategies, and robust representations of unannotated data as inputs (Liu & El-Gohary, 2017). The generated results are then input into the system to produce more labeled examples.

Ekbal et al. (2012) presented a new approach for automatic data annotation using a ML technique. This method iteratively extracts meaningful sentences from a pool of unannotated documents, labels them, and adds them to the training data set. An example can be found in the work of Küçük (2015), who proposed a novel method of automatically compiling language resources from a large number of Turkish Wikipedia articles.

2.2.3. Unsupervised Learning

Unsupervised learning models and techniques learn how to extract and recognize units of target information (e.g., named entities) without requiring annotated training data. These systems/methods apply only unlabeled data for decision making. The key goal of an unsupervised learning method is to produce a model that fully captures the distributional and structural features of the data to achieve better learning (Goyal et al., 2018).

Traditional unsupervised learning methods are often based on clustering or association rule-based methods. The goal of clustering-based methods is to group similar named entities into the same cluster by applying similarity measures. Association rule-based approaches attempt to extract associations among items that are buried in large databases.

Zhang and Elhadad (2013) presented an unsupervised method for biomedical entity recognition using a stepwise solution that included seed-term selection, inverse document frequency (IDF) information filters, and semantic similarity calculations. Konkol et al. (2015) proposed a novel method for NER that utilized semantic features including automatic gazetteer construction to explore word-similarity features.

2.3. Hybrid Approaches

Hybrid approaches combine the advantages of rule-based approaches and statistics-based approaches. These methods focus on achieving high performance by utilizing the results of various statistics-based techniques (e.g., ML techniques) and large numbers of handcrafted rules. Recently, several hybrid NER approaches have been proposed by different researchers.

In the biomedical domain, a ML approach (e.g., CRF) and a postprocessing technique have been used to find and extract biomedical domain information (Li et al., 2009). A combination of transformation-based learning, CRFs, and human-created rules and patterns was developed for Chinese NER (Zeng et al., 2009). Classifier ensemble techniques have been studied to identify relevant entities from different types of data (Ekbal & Saha, 2011). Some hybrids of supervised methods (e.g., linear CRFs) and unsupervised methods (e.g., cluster-based approach) have been constructed in an attempt to extract various entities from English-language tweets (Liu & Zhou, 2013).

Atkinson and Bull (2012) studied a multistrategy method for identifying biomedical entities without using any external lexical features (e.g., a dictionary or ontology). This algorithm mainly focused on using preprocessing techniques such as tokenization, POS tagging, and stop word removal, all of which enhanced the extraction performance when the two-model hybrid (HMM and SVM) was utilized. Küçük and Yazıcı (2012) presented a hybrid Turkish NER method that applied the same features (e.g., pattern bases and lexical resources) used in a rule-based recognizer (Dalkilic et al., 2010). This approach shows a strong ability to extend its feature resources by learning from annotated data. Saha and Ekbal (2013) presented an NER classifier ensemble. Several supervised learning techniques, including the CRF, decision tree, maximum entropy, HMM, and SVM approaches, were combined to produce different models in an effort to obtain improved results. Munkhjargal et al. (2015) conducted a study of the extraction of Mongolian named entities based on a classifier ensemble. This is a rather challenging task because of the complex structure and agglutinative morphology of the Mongolian language. Three types of features generated via the maximum entropy, CRF, and SVM approaches were used to enhance the extraction performance.

3. Preliminaries and Problem Formulation

3.1. Definition of GNER

The term “named entity” refers to a word or phrase that allows elements that have similar properties to be recognized. A named entity can be either a rigid designator or a member of a semantic class that may vary based on the domain of interest. In the domain general, NER focuses on recognizing and identifying different types of names, including names of persons, organizations, reports, and locations. To date, various entities have been identified in various languages and in different domains and used with different types of approaches. To the best of our knowledge, the problem of extracting named entities from textual geoscience reports has been less well studied, hence motivating us to shed light on this domain.

Our goal is to recognize four different types of named entities (see Table 1): geological history, geological structure, rock, and stratum. Examples of each named entity type are listed in Table 1. These geological named entities, together with their representative meanings, are described as follows:

Geological history (GH) relates to the time during which various geological events occurred, including both relative and absolute ages.

Geological structure (GS) refers to the deformation or displacement of a rock or stratum under tectonic action.

Rock (RK) is a solid substance formed by geological processes in the Earth's crust and consists of a collection of minerals or rock debris with certain structural characteristics and rules governing its modifications.

Stratum (SM) is related to the general term for strata or combinations of strata with common characteristics or attributes formed during a particular geological period.

The goal of NER is to annotate words in input sentences as named entities. More formally, an input sentence is represented as S=[w_1, w_2, … ,w_N], and the corresponding prediction labels are denoted by Y=[y_1, $y_{\ 2}$, … , y_N]. Following the standard procedure of sequence processing (Reyes-Galaviz et al., 2017; Zheng et al., 2017), the BIO scheme is used for annotation and evaluation in our task, where *B* represents "begin," *I* represents "inside," and *O* represents "other."

3.2. Attention-Based Bi-LSTM Model

We describe an attention-based, Bi-LSTM model, for the GNER extraction task. Most traditional deep-learning-based NER approaches (e.g., LSTM and Bi-LSTM models) are limited in their ability to address tagging inconsistency problems (Luo et al., 2017). A natural approach for overcoming this issue is to apply an attention mechanism (Vaswani et al., 2017). Attention-based models have recently become popular in fields such as image recognition, natural language processing (NLP), and speech recognition (X. Li et al., 2018; Qu et al., 2018; Zhou et al., 2018).

Therefore, we design a Bi-LSTM model together with an attention mechanism to automatically address the GNER task. Figure 1 shows the network architecture of this model; the embedding features are provided as the input to the first layer, and then a fixed vector representation is formed by the subsequent layers. The final layer calculates a score for each possible information unit (e.g., named entity class) and predicts unseen instances using these scores. In the following subsections, we briefly describe each core component of the model.

3.2.1. Embedding Layer

The embedding layer applies a process in which each discrete feature is mapped into a real-valued vector representation based on a lookup matrix. Given an embedding matrix M^i that represents the ith feature, each column of M^i is a vector representing the feature values for the ith feature.

Let a^i_j denote the one hot vector for the jth feature value of the ith feature; then, the final output x^i of the embedding layer can be computed as follows:

$$x^i = f^i_1 \oplus f^i_2 \oplus f^i_3, \tag{2}$$

where ⊕ denotes the concatenation operation, $x^i \in \mathrm{R}^{(\mathrm{n1+n2+n3})}$ denotes the feature vector for the ith word of the sentence, and n^k represents the vector dimensions of the kth feature. We use pretrained vectors for the word embedding matrix.

3.2.2. Bi-LSTM Layer

A recurrent neural network (RNN) is a network with loops. RNNs have gained attention due to their ability to model sequential tasks and data (Osipov & Osipova, 2018; Morchid, 2018) by allowing relevant information to persist throughout the sequence. However, existing RNN models and techniques are limited in their ability to handle instances with longer sentences because they may suffer from problems of vanishing or exploding gradients (Pascanu et al., 2012). An LSTM model can be used to address this issue by means of

Table 1
Sample List of Geological Named Entities (Ordered Alphabetically)

Named entity	Example in English translation
Geological history	Cenozoic Era, Quaternary Period
Geological structure	Crest of Fold, Trough of Fold
Rock	Aiounite, Talzastite
Stratum	Precambrian Section in Jixian, Sinian Section in Three Gorges area

applying the forget gate and memory mechanisms. An LSTM layer has a novel structure that includes several memory cells (denoted by as c^t) and contains three well-known types of gates: input gates i^t, outputs gate o^t, and forget gates f^t. These three gates constitute a sigmoid activation function and are used to regulate and manage cell information.

Let $\{x^1, x^2, \ldots ,x^{\mathrm{m}}\}$ denote the sequence of word vectors from a sentence, where m denotes the length of the sentences and x^i represents the feature vector that is formed by concatenating all word vectors for the ith word. The previous hidden state and the cell state are denoted by h_l^{t-1} and c_l^{t-1}, respectively; then, the current hidden state h_l^t, cell state c_l^t, and the output of the Bi-LSTM model (denoted byz^t) can be presented as follows:

$$\begin{aligned}
i_l^t &= \sigma\left(U_l^i x^t + W_l^i h_l^{t-1} + b_l^i\right),\\
f_l^t &= \sigma\left(U_l^f x^t + W_l^f h_l^{t-1} + b_l^f\right),\\
o_l^t &= \sigma\left(\mathrm{U}_l^o x^t + W_l^o h_l^{t-1} + b_l^o\right),\\
g_l^t &= \tanh\left(\mathrm{U}_l^g x^t + W_l^g h_l^{t-1} + b_l^g\right),\\
c_l^t &= c_l^{t-1} * f_l^t + g_l^t * i_l^t,\\
h_l^t &= \tanh\left(c_l^t\right) * o_l^t.
\end{aligned} \tag{3}$$

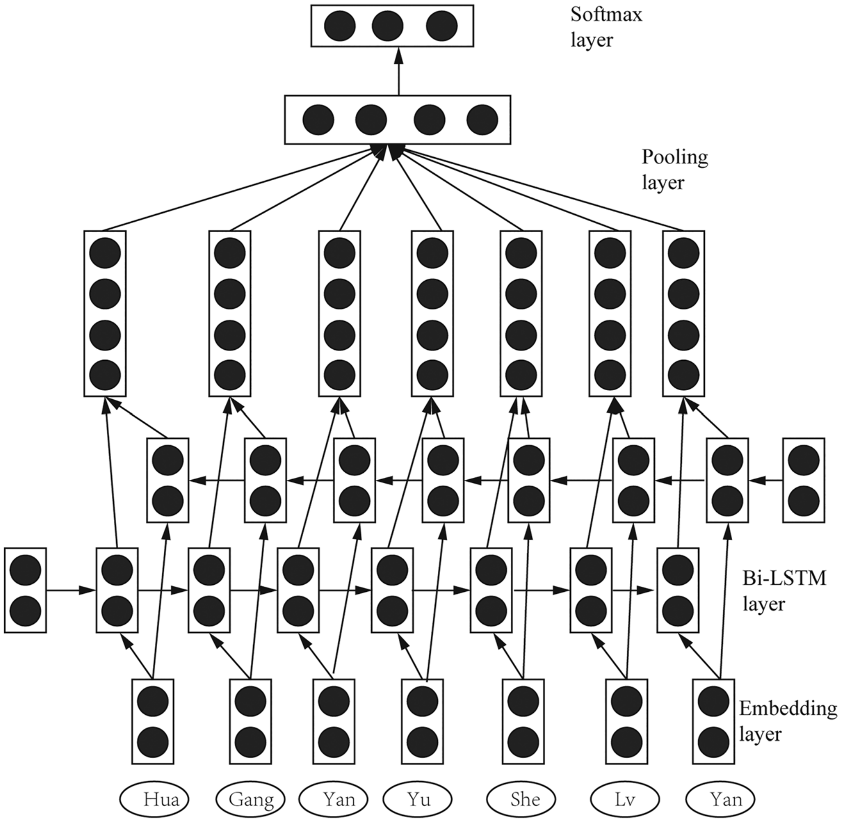


Figure 1. Attention-based bidirectional long short-term memory (Bi-LSTM) model.

In equation (3), σ denotes the sigmoid activation function, * is used to represent the elementwise product, and the other previously undefined symbols denote the learning parameters of the Bi-LSTM model. Here, d denotes the dimensions of the input feature vector, and N denotes the hidden layer size. h_r^t is computed similarly to h_l^t but with the order of the words in a given sentence reversed. The final output for the ith word from the Bi-LSTM layer is presented as follows:

$$z^t = h_l^t \oplus h_r^t. \tag{4}$$

3.2.3. Pooling Layer

The purposed of a pooling layer is to derive a fixed feature vector from a varied group of word features. In this paper, motivated by Y. Li et al. (2018), max pooling is used throughout entire complete sequence. The objective of a max pooling layer is to derive a fixed feature vector from an input sequence; it takes the maximum value from the input sequence based on the assumption that all relevant and important information is accumulated in the corresponding position. Let $\{z^1, z^2, \ldots, z^m\}$ denote the sequence of vectors formed by concatenating the previous and subsequent LSTM outputs for all words. This sequence of vectors is defined as follows:

$$z = \max_{1 \le i \le m} \left[z^i \right], \tag{5}$$

where z denotes the dimension wise maximum among all z^i.

3.2.4. Fully Connected and Softmax Layer

The output of the max pooling layer is a fixed feature vector. We use an activation function, namely, the nonlinear tanh function, to process this fixed feature vector and take the newly resulting vector as the input to a fully connected layer, as follows:

$$\begin{aligned} h^3 &= \tanh\left(h^2\right) \\ \mathrm{p}(y|x) &= \mathit{softmax}\left(h^{3T}\mathrm{W}^o + b^o\right), \end{aligned} \tag{6}$$

where h^2 denotes the output of the max pooling layer, $W^o \in R^{N \times \mathrm{C}}$ and $b^o \in R^{\mathrm{C}}$ represent the parameters of the fully connected layer, and C denotes the final number of categories in our designed model. Finally, the softmax function is used in the fully connected layer to generate a normalized probability score for each category.

4. Proposed Approach

To tackle the abovementioned needs and challenges, the authors present a novel, generative, unsupervised, DL-based, GNER methodology. The proposed GNER methodology generates training data sets (sets of generative pseudosentences) that depend entirely on the words and their corresponding frequencies in a machine-learning function based on an n-gram language model. The method dynamically adapts to unseen instances (entities) and classifies them into predefined entity categories by learning from a large number of training data sets—thus eliminating the need for human annotation.

Figure 2 depicts our proposed GNER framework. A set of unlabeled seeds (e.g., domain-specific entities and domain-general words) is initially provided as the training input. In this study, the process of GNER is performed in four steps.

As in a typical GNER approach, collecting a set of seeds is the first step. This process involves collecting comprehensive and representative terms suitable for generating the pseudosentences described in the third step. The aim of the seed collection process is to select pairs of domain-specific entities and their corresponding frequencies as well as domain-general words and their corresponding frequencies. The second step of the GNER process is to normalize and group the selected seeds (terms) based on their corresponding frequencies. In this step, all the frequencies are first normalized using an integral function, and then these seeds (terms) are grouped to speed up algorithm training. In the third step (random selection of a diverse set of seeds for generating sentences), entities and words are randomly selected from the previously formed groups to generate a set of pseudosentences to be used as input training data for a DL model (the attention-based Bi-LSTM model in this study). A randomized and automatic seed selection process is an important step of

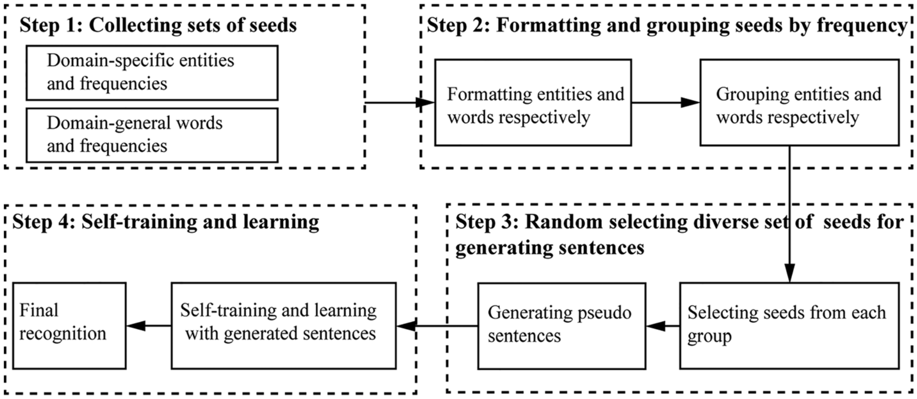

Step 1: Collecting sets of seeds
Domain-specific entities and frequencies
Domain-general words and frequencies
Step 2: Formatting and grouping seeds by frequency
Formatting entities and words respectively
Grouping entities and words respectively
Step 3: Random selecting diverse set of seeds for generating sentences
Selecting seeds from each group
Generating pseudo sentences
Step 4: Self-training and learning
Self-training and learning with generated sentences
Final recognition

Figure 2. Geological name entity recognition with the proposed approach.

automatically generating sentences to be used as training data. In the last step (self-training and learning), the DL model is trained with each generated sentence, and the trained model is then used to predict the final results.

The following sections, the proposed GNER methodology and its components are described in more detail.

4.1. Step 1: Collecting a Highly Diverse Set of Seeds

Most existing approaches to NER utilize supervised ML models and techniques to train and generate a model for classifying pairs of terms as either matches or nonmatches. The main limitation of such methods is that they require labeled training examples/data sets. However, in real-world applications (e.g., in the geoscience domain), such annotated data sets are difficult to acquire. In addition, constructing a data set of a sufficient size through manual labeling is often impossible because of the sensitivity of the data. Therefore, developing a novel unsupervised learning technique is necessary.

The first step of our approach is to collect a set of representative seed terms for different types of geological entity classes suitable for generating sample sentences in the third step of the GNER process. The seed-term set (the input to our model) is gathered from external concepts and terminologies. To enhance the automated extraction of geological named entities, the selected seed terms are mainly derived from corresponding formally defined semantic types, domain ontologies, and specific groups that utilize these domain-specific terminologies; consideration of these domain-specific terms is required because of the complexity and heterogeneity of the free texts found in geoscience reports. We call these domain-specific terminologies chosen for inclusion in the seed set “domain-specific entities.” These are paired with their corresponding frequencies and are combined with domain-general words and their corresponding frequencies.

After the selection procedure, in this study, we obtained 11,892 and 665,785 seed terms corresponding to domain-specific entities and domain-general words, respectively. Note that the domain-specific entities include various types of geological entities, including geological history, geological structure, rock, and stratum entities. Table 2 presents simple examples of the two types of seed terms, each consisting of a term and its corresponding frequency.

Table 2
Example of the Two Types of Seed Terms

Domain-specific entity	Domain-general words
Quaternary period, 6689	Reporters Center, 164
Coniacian Age, 7741	Silk-like cotton, 129
Upright fold, 2235	Rouse, 124
Glauconitic quartzarenite, 4478	Country, 119

4.2. Step 2: Normalizing and Grouping Seeds by Frequency

After collecting the highly diverse seed set, the next step is to normalize and group the seeds by frequency. Two text/term processing steps were applied in this phase: term normalization and term grouping. Term normalization involves applying a rounding function to map the frequencies of the seed terms into a form suitable for further analysis. Term grouping involves dividing the seed terms into a sequence of groups based on their frequencies. In our approach, frequency is a representative and important factor aiding in geological named entity extraction.

Formally, given a word w in a seed term set S, where G={g_1, g_2, ... ,g_n} is the final set of groups, let F={f_1,f_2, ... , f_N} denote the frequency of each group g. Here, θ denotes the contribution factor and is defined as follows:

$$\theta_i = \frac{f_i}{\sum_{j=1}^{N} f_j}. \tag{6}$$

We normalize the frequency for each term in the set and obtain the extraction probability for each term by averaging all groups. Note that θ represents a probability computed based on the weight of each word in the groups. The weight is an indicator that determines the probability of random extraction from among the seed terms. A greater probability means that a term is statistically more important when generating sentences. The reason for using this strategy for each seed term rather than another approach is that this strategy speeds up the training time.

Figure 3 shows an example of the normalization and grouping processes.

4.3. Step 3: Randomly Selecting a Subset of Seeds for Generating Sentences

After preprocessing the collection of seed terms, the next step is to apply a novel generative method based on a statistical language model to automatically construct a training data set, thus eliminating the need to manually annotate a training data set. We refer to these generated examples (for inclusion in the training data set) as “pseudosentences.” First, terms and their corresponding frequencies from the seed set are randomly and automatically selected and extracted from the different groups based on their extraction probabilities (described in section 4.2). A computational function is used to optimize the extraction probability and automatically generate a large number of pseudosentences along with appropriate annotation labels.

To generate the pseudosentences, a unigram language model is primarily used to extract and identify geological named entities, under the assumption that sentences consist of independent words (Peng et al., 2016; Qiu et al., 2018; Tripathy et al., 2016). A language model (Brown et al., 1992) is a statistical model that computes the probability of a sentence by helping to predict the next word in a given sequence for different language types. From a generative perspective, every natural sentence is composed of a set of words strung together with a probability equal to the product of a group of conditional probabilities.

For example, given a sentence s={t_1,t_2, ... ,t_n}, the sentence probability is defined as follows:

$$p(s) = p(t_1, t_2, ..., t_n) = p(t_1)p(t_2|t_1)p(t_3|t_1t_2)...p(t_n|t_1t_2..t_{n-1}). \tag{7}$$

The language model given in equation (7) represents a probability distribution over a sequence of many tokens (symbols) drawn from a finite symbol set. The chain rule of conditional probability is used to generate a sequence s.

Initially, we consider an n-gram of size 1, a unigram. Unigrams are the words that independently constitute a sentence. Thus, equation (7) is computed as follows:

$$p(s) = p(t_1, t_2, ..., t_n) = p(t_1)p(t_2)p(t_3)...p(t_n) = \prod_{i=1}^{N} p(t_i), \tag{8}$$

where p (t_i) denotes the probability of the single word t_i. The probability estimates p (t_i) is required to either compute the probability of a given sequence (sentence) or to make predictions. A natural way to calculate this probability is to apply the maximum likelihood estimate, which can be computed by counting the number of occurrences of t_i in a corpus (denoted by N (t_i)) and then dividing that value by the total number of tokens (sequences; denoted by M) in the corpus:

$$p(t_\mathrm{i}) = \frac{N(t_\mathrm{i})}{M}. \tag{9}$$

The assumption of a probability estimate is important because it directly simplifies the problem of estimating a language model from a limited set. More importantly, in the unigram model, words are assumed to be independent and exchangeable, and word order information is irrelevant.

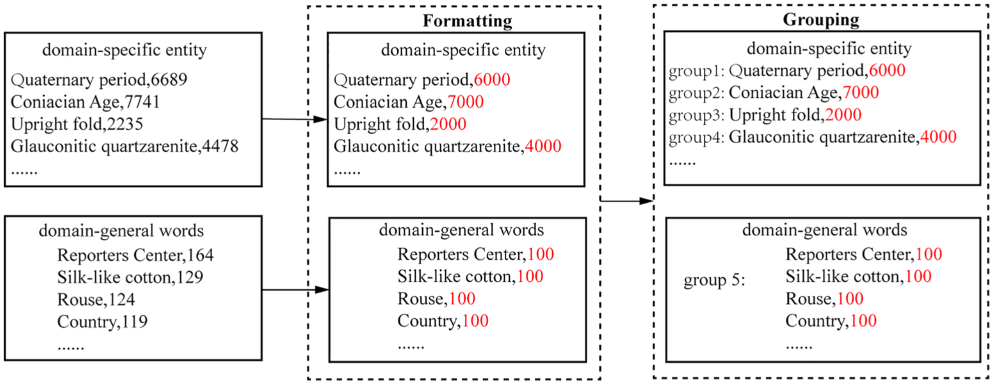


Figure 3. An example of normalization and grouping.

Because the words in a sentence in the unigram model are regarded as independent and unordered, by simply using words and their corresponding frequencies (the extraction probabilities defined in section 4.2), we can theoretically generate and produce new sentences based on the unigram language model. The corresponding frequencies are calculated and then considered as the probability $p\ (t_i)$; then, words with their corresponding probabilities can be iteratively input into an optimization function to generate pseudosentences. For this study, we constructed the following probability function as the optimization object:

$$S_{\max} = argmaxp(t_1)p(t_2)...p(t_n). \tag{10}$$

In other words, the goal is to capture the highest possible probability $p(t_1)p(t_2)$ ... $p\ (t_n)$ via a dynamic programming scheme. DL offers strong self-learning capabilities for optimizing this objective function.

Given a random selection algorithm, each iteration focuses on automatically generating a sentence accompanied by corresponding labels. Sentence generation continues until the algorithm reaches its maximum number of iterations; then, the generated sentences are added to the input training data set.

4.4. Step 4: Self-Training and Learning for GNER

The proposed attention-based Bi-LSTM model (presented in section 3.2) is trained using the set of pseudosentences generated as described above. Given a large training data set, many existing ML models and techniques can be used in the self-training and learning process. However, the computational complexity of traditional ML approaches limits their ability to support training on large data sets. To improve the training efficiency of the self-training and learning process, we use a DL algorithm to train the classifier. In addition, GNER can be regarded as a sequence-based problem. DL approaches have become prevalent for application to such sequence problems (Salaken et al., 2018; Yang & Chen, 2018; Yuan et al., 2018). Following the self-training and learning process, a classification model is generated. Then, the model is used to generate a prediction for each unseen instance (item from the test data set), determining it to be a match or nonmatch.

5. Experimental Evaluation

We conducted five primary experiments to fine-tune the parameters of the unsupervised GNER algorithm and verify its performance for automatic NER from geological reports. The first experiment was conducted to identify and select the best parameters (e.g., the word embedding dimensionality and different frequencies). The second experiment was conducted to evaluate the overall performance of the proposed GNER algorithm. The third experiment was conducted to validate the performance for new entity detection. The fourth experiment was conducted to select the most appropriate input data size (e.g., the number of seed terms from among the selected training terms). Finally, the fifth experiment compared our proposed method with other methods on two constructed data sets. The details of the experimental setup, the final results, and the performance of the proposed GNER algorithm are presented and discussed in the following sections.

Table 3
General Statistics of the Two Data Sets Used in Our Experiments

Property	GJP data set	RGR data set
# sentences	8,975	7,936
# words	18,985	25,774
# entity mentions	6,598	9,901
# test documents	500	14

Note. The “#” symbol represents “number of.”

5.1. Experimental Setup

To evaluate the final performance of the proposed GNER methodology, we constructed two experimental data sets. The first data set, named the GJP corpus, consists of abstracts of various geological journal papers published between 2014 and 2018. These data were sourced from the largest Chinese academic database CNKI (http://www.cnki.net/). Fourteen regional geological reports were selected to construct the second experimental data set, which is called the RGR corpus. A summary of these two data sets, including the numbers of sentences, entities, and words, is shown in Table 3. A step-by-step illustration of the application of the proposed GNER methodology is presented in Figure 4.

For both data sets, we followed the standard manual annotation procedure to develop gold-standard sequences of entity categories for testing purposes. The goal of manual annotation is to assign a true label to each token (term) in a sentence. Five human labelers with experience in both geoscience and NER were asked to annotate the selected documents/reports to form the gold-standard corpus.

Three metrics derived from the field of information retrieval, namely, the precision, recall, and F1-measure, were selected as the evaluation metrics. The *precision*, defined in equation (10), measures the percentage of the overall number of relevant extracted entities out of all extracted entities. The *recall*, defined in equation (11), measures the percentage of the overall number of relevant extracted entities out of all relevant entities. To quantify the a balance between the precision and recall, the third evaluation metric, the F1-measure, defined in equation (12), represents the harmonic mean of precision and recall. When both the precision and recall have fairly high values, the F1-measure will also have a high value and vice versa.

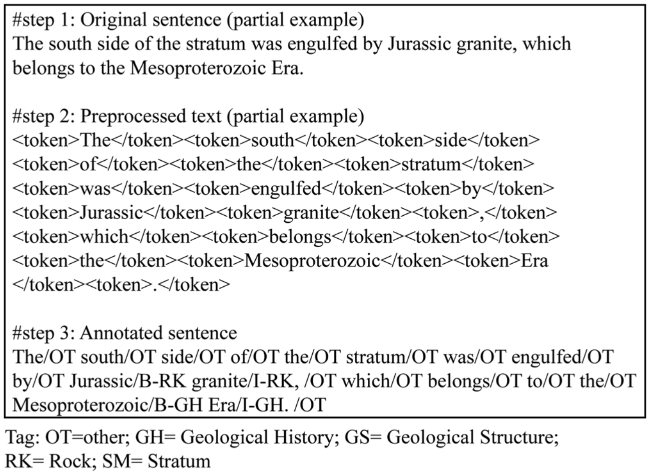
$$Precision = \frac{number\ \ of\ \ correctly\ \ extracted\ \ entities}{number\ \ of\ \ extracted\ \ entities}, \tag{10}$$

$$Recall = \frac{number\ \ of\ \ correctly\ \ extracted\ \ entities}{number\ \ of\ \ extracted\ \ entities}, \tag{11}$$

$$F1{-}measure = 2\cdot\frac{Precision\times Recall}{Precision + Recall}. \tag{12}$$

5.2. Parameters and Sensitivity Analysis

We conducted a variety of experiments to investigate how different key hyperparameters, including the vector dimensionality *d* and the terms frequency was *n*, could affect the GNER performance. The proposed model's results demonstrate the impacts of these hyperparameters in terms of the F1-measure. Extensive experiments were conducted on both the GJP and RGR data sets.



#step 1: Original sentence (partial example)
The south side of the stratum was engulfed by Jurassic granite, which belongs to the Mesoproterozoic Era.

#step 2: Preprocessed text (partial example)
<token>The</token><token>south</token><token>side</token><token>of</token><token>the</token><token>stratum</token><token>was</token><token>engulfed</token><token>by</token><token>Jurassic</token><token>granite</token><token>,</token><token>which</token><token>belongs</token><token>to</token><token>the</token><token>Mesoproterozoic</token><token>Era</token><token>.</token>

#step 3: Annotated sentence
The/OT south/OT side/OT of/OT the/OT stratum/OT was/OT engulfed/OT by/OT Jurassic/B-RK granite/I-RK, /OT which/OT belongs/OT to/OT the/OT Mesoproterozoic/B-GH Era/I-GH. /OT

Tag: OT=other; GH= Geological History; GS= Geological Structure; RK= Rock; SM= Stratum

Figure 4. Illustrative example with gold-standard annotations.

Figure 5a shows the impact of the vector dimensionality *d*. In six experiments, the vector dimensionality was varied from 50 to 300 with a step size of 50. The results suggest that the vector dimensionality affects the overall performance. Increasing the vector dimensionality *d* results in a higher-dimensional representation that includes more dependent information and can more precisely capture the relevant relationships from various features. However, a larger *d* also adds complexity and necessitates longer training time. Thus, based on the results, we set the vector dimensionality to *d=300* as a trade-off between GNER performance and training time.

As mentioned in section 4.2, the frequencies (also interpreted as the probabilities) of seed terms control their likelihood of random extraction for generating pseudosentences for training. To verify the influence of this parameter, we conducted a total of nine experiments using terms with frequencies ranging from 500 to 4,000 with a step size of 500. The final results

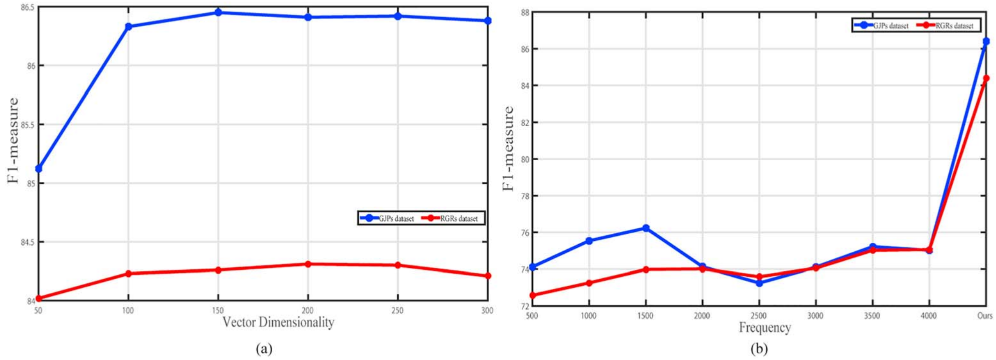


Figure 5. Impacts of Vector Dimensionality (a) and Different Term Frequencies (b).

are illustrated in Figure 5b in terms of the F1-measure. Note that a frequency of n=500 means that the frequency of all domain-specific seed terms used is 500. However, we observe that varying the frequency information in the proposed model does not result in GNER performance improvements as n increases. The proposed GNER model achieves its best performance, with F1-measure values of 86.41% and 84.31% on the GJP and RGR data sets, respectively, when the frequency information is based on a statistical strategy. Therefore, we employ a statistical strategy to set the frequencies for the GJP and RGR data sets.

5.3. Overall Performance

A set of baseline DL models (e.g., RNN, LSTM, Bi-LSTM, and Bi-LSTM-CRF) were constructed and compared with the proposed GNER models. These baseline DL models were configured with their optimal hyperparameters. Table 4 summarizes the results for the performance of our proposed GNER model on both the GJP and RGR data sets. As shown in Table 4, compared with the four baseline methods, our proposed approach yields the best results on both data sets, with average precision, recall, and F1-measure scores of 86.74%, 86.05%, and 86.39%, respectively, on the GJP data set and 84.85%, 83.75%, and 84.30%, respectively, on the RGR data set. These results illustrate the effectiveness of DL models and techniques.

5.4. New Entity Detection Capability

To test the new entity detection capability of our proposed GNER model, we chose to calculate the R_{OOV} and R_{IV} metrics on both the GJP and RGR data sets. R_{OOV} denotes the percentage of out-of-vocabulary (OOV) terms and is an indicator that can be generalized to a new field. R_{IV} denotes the percentage of in-vocabulary (IV) terms that are correctly recognized and reflects model's predictive ability with respect to the training data.

The dictionary-based matching approach (abbreviated as MM) for GNER was chosen as a basic and representative approach. All collected domain-specific entities were regarded as the dictionary in our experiments. As suggested by the R_{OOV} rates and R_{IV} rates reported in Table 5 it is evident that detecting new entities is indeed a difficult task. The R_{OOV} value for the MM approach is only 5.4% because the MM approach lacks the self-learning power to identify new entities. By contrast, the proposed method yields a significant improvement, with R_{OOV} values of 48.1% and 68.1% on the GJP and RGR data sets, respectively,

Table 4
Performance of the Proposed GNER Algorithm on the GJP and RGR Data Sets

	GJP			RGR		
Model	Precision (%)	Recall (%)	F1-measure (%)	Precision (%)	Recall (%)	F1-measure (%)
RNN	80.25	79.12	79.68	78.12	78.02	78.07
LSTM	81.93	79.99	80.95	79.15	79.05	79.10
Bi-LSTM	82.14	80.65	81.39	80.35	80.03	80.19
Bi-LSTM-CRF	83.45	82.69	83.07	81.25	81.11	81.18
Ours	86.74	86.05	86.39	84.85	83.75	84.30

Table 5
Performance of the Proposed GNER Algorithm for Identifying OOV and IV Terms

Model	R_{OOV}	Improvement	R_{IV}	Improvement
MM	0.054		0.793	
Att-BiSTM$_{GIP}$	0.535	↑48.1%	0.863	↑7.0%
Att-BiSTM$_{RGR}$	0.735	↑68.1%	0.832	↑3.9%

Note. GNER = geological name entity recognition; MM = matching approach; OOV = out of vocabulary; IV = in-vocabulary.

indicating that the proposed DL model has substantial predictive power thanks to its capabilities of self-learning and self-correction. The R_{IV} values suggest that our proposed approach performs well for NER. These results indicate that the proposed stepwise method is effective at addressing the domain-specific NER problem.

5.5. Varying the Training Data Set Size

We also conducted controlled experiments to investigate the optimal size of the training data set (as defined in section 4.1). We chose the F1-measure metric to demonstrate the impacts when the training data set size was varied from 10% to 100% with a step size of 10%; the results are shown in Figure 6. Five independent runs were performed for every split, and the average results were calculated.

As shown in Figure 6, the proposed GNER model benefits from a larger number of seed terms, resulting in steadily rising curves. Initially, the model makes full use of the training set and increasingly benefits from larger amounts of training data. Eventually, however, the performance of the proposed GNER model tends to converge as the training data set size increases. For example, the F1-measure on the GJP data set starts to fall after the training data set size reaches 90% of the total collected data set, and the average F1-measure shows a converging trend as the size increases. This finding indicates that we must select representative and comprehensive training data to allow the model to automatically learn from these data to correct mistakes and address different GNER extraction cases.

5.6. Comparison with Other Methods

We compared our proposed model with the following five related methods:

IDF_NER (Zhang & Elhadad, 2013) is an unsupervised method that identifies named entities in free text without requiring any rules or annotated data. This approach involves two main steps: detecting entity boundaries and classifying entity types.

C_NER (Alicante et al., 2016) uses several traditional NLP tools to extract entities. It is an unsupervised approach that uses clustering to group entity pairs.

SwellShark (Fries et al., 2017) is a framework for performing NER without requiring manually labeled data. It applies a generative model to achieve supervision through the generation of large-scale labeled data sets.

DPT (Gupta et al., 2018) uses a hybrid approach that incorporates dependency-based parse trees and semantics to extract named entities.

CAAEE (Cai & Wu, 2018) is an unsupervised model based on content-aware attributed entity embedding for the discovery of named entities.

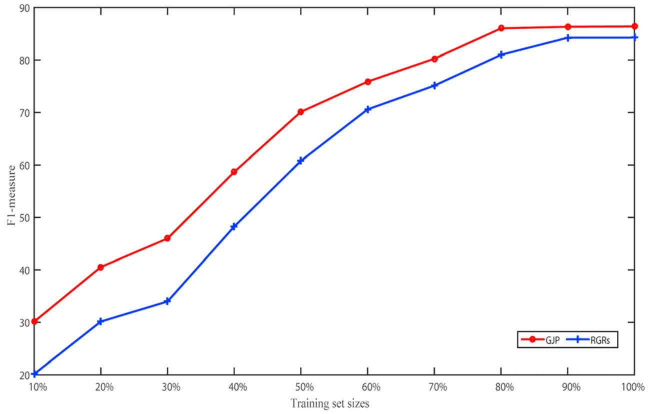


Figure 6. F1-measure values for the proposed algorithm on both the GJP and RGR data sets when varying the training data set size.

From Table 6, we can observe that the traditional approaches (e.g., IDF_NER and C_NER) are inferior to the other methods. Using our proposed method, the precision, recall, and F1-measure are improved by 15.89%, 14.8%, and 15.34%, respectively. These results suggest that DL offers a superior predictive ability compared to traditional approaches.

Our proposed model also significantly outperforms the other ML approaches (SwellShark, DPT, and CAAEE), indicating that our proposed GNER model can randomly select domain-specific entities and domain-general words to form new sentences for training DL models that exhibit the capability of self-correction.

6. Discussion

This paper presents and discusses the use of unsupervised information extraction techniques for the extraction and recognition of geological named entities from geological reports written in Chinese. In particular, it is aimed at the extraction of domain-relevant entities. To this end, a pipeline system with a four-step structure is proposed, where the four

Table 6
Results of Comparisons With Other Models

	GJP			RGR		
Model	Precision (%)	Recall (%)	F1-measure (%)	Precision (%)	Recall (%)	F1-measure (%)
IDF_NER	70.85	71.25	71.05	69.58	68.25	68.91
C_NER	75.12	75.58	75.35	74.12	74.96	74.54
SwellShark	78.21	78.68	78.44	76.56	76.58	76.57
DPT	80.21	81.05	80.63	79.89	78.29	79.08
CAAEE	83.12	84.11	83.61	82.15	82.19	82.17
Ours	**86.74**	**86.05**	**86.39**	**84.85**	**83.75**	**84.30**

Note. The best results are presented in boldface.

steps include collecting a set of seeds, normalizing and grouping the seeds by frequency, randomly selecting diverse sets of seeds for generating pseudo sentences, and self-training and learning. The proposed algorithm achieves average F1-measure values of 86.39% and 84.30% on the GJP and RGR data sets, respectively.

With the advent of the big data era, geological data services face the dual demands of digitization and socialization for both institutions and the public. Currently, large amounts of fragmented and unstructured data are buried in geological reports and go unused. In particular, textual data make up an important part of these unstructured geological data, and consequently, the automatic extraction of entities from textual data has become an important research direction. As a core element of geological entity recognition, the extraction of geological named entities will provide new opportunities to mine this wealth of geological text in the geoscience domain. In a given report, the arrangement of the words and chapters is organized around certain topics. The recognition of geological entities can not only enable effective identification of the basic information units in such a text and assist in correctly understanding the meaning of that text but also provide comprehensive support for information extraction, information retrieval, machine translation, abstract generation, and other tasks in generalized text data mining based on the extracted geological knowledge.

To this end, we can use the entities extracted by the proposed algorithm to represent the key information contained in a sentence. In general, doing so can quickly provide a basic overview of a long sentence/document quickly, avoiding the need to read the document word by word. For example, in the sentence shown in Figure 4, the domain-specific words (i.e., Jurassic and granite) are informative.

7. Conclusions

GNER is a challenging task in geological NLP. In this study, we designed and proposed an unsupervised framework that provides a pipeline solution for GNER, including colleting a highly diverse set of seeds, normalizing and grouping the seeds by frequency, randomly selecting sets of seeds to generate sentences for training, and self-training and learning for the extraction of named entities. In our framework, the selection of high-quality seed terms (e.g., domain-specific entities and domain-general words) is a key factor in building a successful model. The approaches for normalizing and grouping the seed terms are candidates for further analysis. Following the automatic and random seed selection process, we use an attention-based bidirectional long short-term memory model to generate various pseudosentences. Each generated set of sentences is then input into the self-training and learning algorithm. Our proposed approach does not depend on any heuristics or rules or require any labeled data, making it suitable for application to various domains and tasks. We conducted extensive experiments to validate the performance of the proposed model and compare it to other approaches. The experimental results obtained on two constructed data sets demonstrate the effectiveness of our method for GNER.

In future work, we plan to focus on combining our proposed method with other supervised methods to improve the ultimate GNER performance.

Acknowledgments
We would like to thank the anonymous reviewers for carefully reading this paper and their very useful comments. This study was financially supported by the National Key Research and Development Program (2018YFB0505500, 2018YFB0505504, 2017YFB0503600, and 2017YFC0602204) and the National Natural Science Foundation of China (41871311, 41671400, and 41871305). All original data can be found in the GitHub (https://github.com/CGGeoNLP/GNER). Qinjun Qiu, Liang Wu, and Zhong Xie conceived and designed the experiments; Qinjun Qiu, Liang Wu, and Zhong Xie performed the experiments; Qinjun Qiu, Liang Wu, and Zhong Xie analyzed the data; Qinjun Qiu, Liang Wu, Zhong Xie, and Liufeng Tao wrote the paper. The authors declare no conflict of interest.

References

Alicante, A., Corazza, A., Isgrò, F., & Silvestri, S. (2016). Unsupervised entity and relation extraction from clinical records in Italian. *Computers in Biology and Medicine*, *72*, 263–275. https://doi.org/10.1016/j.compbiomed.2016.01.014

Atkinson, J., & Bull, V. (2012). A multi-strategy approach to biological named entity recognition. *Expert Systems with Applications*, *39*(17), 12968–12974. https://doi.org/10.1016/j.eswa.2012.05.033

Brown, P. F., Desouza, P. V., Mercer, R. L., Watson, T. J., Pietra, V. J. D., & Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, *18*(4), 467–479.

Cai, D., & Wu, G. (2018). Content-aware attributed entity embedding for synonymous named entity discovery. *Neurocomputing*. https://doi.org/10.1016/j.neucom.2018.10.055

Cernuzzi, L., & Pane, J. (2014). *Toward open government in Paraguay*, *IT Professional* (Vol. 16, pp. 62–64).

Chen, H., Wei, B., Liu, Y., Li, Y., Yu, J., & Zhu, W. (2018). Bilinear joint learning of word and entity embeddings for Entity Linking. *Neurocomputing*, *294*, 12–18. https://doi.org/10.1016/j.neucom.2017.11.064

Cracknell, M. J., & Reading, A. M. (2014). Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. *Computers & Geosciences*, *63*, 22–33. https://doi.org/10.1016/j.cageo.2013.10.008

Dai, H., Tang, S., Wu, F., & Zhuang, Y. (2018). Entity mention aware document representation. *Information Sciences*, *430-431*, 216–227. https://doi.org/10.1016/j.ins.2017.11.032

Dalkilic, F. E., Gelisli, S., & Diri, B. (2010). Named entity recognition from Turkish texts. 2010 IEEE 18th Signal Processing and Communications Applications Conference. https://doi.org/10.1109/siu.2010.5653553

Ekbal, A., & Saha, S. (2011). A multiobjective simulated annealing approach for classifier ensemble: Named entity recognition in Indian languages as case studies. *Expert Systems with Applications*, *38*(12), 14760–14772. https://doi.org/10.1016/j.eswa.2011.05.004

Ekbal, A., Saha, S., & Singh, D. (2012). Active machine learning technique for named entity recognition. Proceedings of the International Conference on Advances in Computing, Communications and Informatics – ICACCI'12. doi:https://doi.org/10.1145/2345396.2345427

Enríquez, J. G., Domínguez-Mayo, F. J., Escalona, M. J., Ross, M., & Staples, G. (2017). Entity reconciliation in big data sources: A systematic mapping study. *Expert Systems with Applications*, *80*, 14–27. https://doi.org/10.1016/j.eswa.2017.03.010

Fernando, T., Denman, S., McFadyen, A., Sridharan, S., & Fookes, C. (2018). Tree Memory Networks for modelling long-term temporal dependencies. *Neurocomputing*, *304*, 64–81. https://doi.org/10.1016/j.neucom.2018.03.040

Fries, J, Wu, S, Ratner, A, & Ré, C. (2017). Swellshark: A generative model for biomedical named entity recognition without labeled data.

Goyal, A., Gupta, V., & Kumar, M. (2018). Recent named entity recognition and classification techniques: A systematic review. *Computer Science Review*, *29*, 21–43. https://doi.org/10.1016/j.cosrev.2018.06.001

Guo, J., Han, Q., Ma, G., Liu, H., & van Hooland, S. (2017). Tunable discounting and visual exploration for language models. *Neurocomputing*, *269*, 73–81. https://doi.org/10.1016/j.neucom.2016.08.145

Gupta, A., Banerjee, I., & Rubin, D. L. (2018). Automatic information extraction from unstructured mammography reports using distributed semantics. *Journal of Biomedical Informatics*, *78*, 78–86. https://doi.org/10.1016/j.jbi.2017.12.016

Huang, C., Zhu, J., Huang, X., Yang, M., Fung, G., & Hu, Q. (2018). A novel approach for entity resolution in scientific documents using context graphs. *Information Sciences*, *432*, 431–441. https://doi.org/10.1016/j.ins.2017.12.024

Konkol, M., Brychcín, T., & Konopík, M. (2015). Latent semantics in named entity recognition. *Expert Systems with Applications*, *42*(7), 3470–3479.

Küçük, D. (2015). Automatic compilation of language resources for named entity recognition in Turkish by utilizing Wikipedia article titles. *Computer Standards & Interfaces*, *41*, 1–9. https://doi.org/10.1016/j.csi.2015.02.003

Küçük, D., & Yazıcı, A. (2012). A hybrid named entity recognizer for Turkish. *Expert Systems with Applications*, *39*(3), 2733–2742. https://doi.org/10.1016/j.eswa.2011.08.131

Li, L., Zhou, R., & Huang, D. (2009). Two-phase biomedical named entity recognition using CRFs. *Computational Biology & Chemistry*, *33*(4), 334–338.

Li, X., Fang, J., Zhang, W., & Li, H. (2018). Finite-time synchronization of fractional-order memristive recurrent neural networks with discontinuous activation functions. *Neurocomputing*. https://doi.org/10.1016/j.neucom.2018.08.003

Li, Y., Liu, T., Hu, J., & Jiang, J. (2018). Topical co-attention networks for hashtag recommendation on microblogs. *Neurocomputing*. https://doi.org/10.1016/j.neucom.2018.11.057

Lima, L. A., Görnitz, N., Varella, L. E., Vellasco, M., Müller, K.-R., & Nakajima, S. (2017). Porosity estimation by semi-supervised learning with sparsely available labeled samples. *Computers & Geosciences*, *106*, 33–48. https://doi.org/10.1016/j.cageo.2017.05.004

Liu, K., & El-Gohary, N. (2017). Ontology-based semi-supervised conditional random fields for automated information extraction from bridge inspection reports. *Automation in Construction*, *81*, 313–327. https://doi.org/10.1016/j.autcon.2017.02.003

Liu, X., & Zhou, M. (2013). Two-stage NER for tweets with clustering. *Information Processing & Management*, *49*(1), 264–273. https://doi.org/10.1016/j.ipm.2012.05.006

Luo, L., Yang, Z., Yang, P., Zhang, Y., Wang, L., Lin, H., & Wang, J. (2017). An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. *Bioinformatics*, *34*(8), 1381–1388. https://doi.org/10.1093/bioinformatics/btx761

Ma, X. (2017). Linked Geoscience Data in practice: Where W3C standards meet domain knowledge, data visualization and OGC standards. *Earth Science India*, *10*(4), 429–441.

Majumder, M., Barman, U., Prasad, R., Saurabh, K., & Saha, S. K. (2012). A novel technique for name identification from Homeopathy Diagnosis Discussion Forum. *Procedia Technology*, *6*, 379–386. https://doi.org/10.1016/j.protcy.2012.10.045

Morchid, M. (2018). Parsimonious memory unit for recurrent neural networks with application to natural language processing. *Neurocomputing*. https://doi.org/10.1016/j.neucom.2018.05.081

Munkhjargal, Z., Bella, G., Chagnaa, A., & Giunchiglia, F. (2015). Named entity recognition for Mongolian language.

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, *30*(1), 3–26.

Osipov, V., & Osipova, M. (2018). Space-time signal binding in recurrent neural networks with controlled elements. *Neurocomputing*, *308*, 194–204. https://doi.org/10.1016/j.neucom.2018.05.009

Pascanu, R, Mikolov, T, & Bengio, Y. (2012). Understanding the exploding gradient problem. Arxiv Preprint Arxiv.

Peng, J., Choo, K.-K. R., & Ashman, H. (2016). Bit-level n -gram based forensic authorship analysis on social media: Identifying individuals from linguistic profiles. *Journal of Network and Computer Applications*, *70*, 171–182. https://doi.org/10.1016/j.jnca.2016.04.001

Qiu, Q., Xie, Z., Wu, L., & Li, W. (2018). DGeoSegmenter: A dictionary-based Chinese word segmenter for the geoscience domain. *Computers and Geosciences*. https://doi.org/10.1016/j.cageo.2018.08.006

Qu, J., Ouyang, D., Hua, W., Ye, Y., & Li, X. (2018). Distant supervision for neural relation extraction integrated with word attention and property features. *Neural Networks*, *100*, 59–69. https://doi.org/10.1016/j.neunet.2018.01.006

Quijano-Sánchez, L., Liberatore, F., Camacho-Collados, J., & Camacho-Collados, M. (2018). Applying automatic text-based detection of deceptive language to police reports: Extracting behavioral patterns from a multi-step classification model to understand how we lie to the police. *Knowledge-Based Systems*, *149*, 155–168. https://doi.org/10.1016/j.knosys.2018.03.010

Quimbaya, A. P., Múnera, A. S., Rivera, R. A. G., Rodríguez, J. C. D., Velandia, O. M. M., Peña, A. A. G., & Labbé, C. (2016). Named entity recognition over electronic health records through a combined dictionary-based approach. *Procedia Computer Science*, *100*, 55–61. https://doi.org/10.1016/j.procs.2016.09.123

Rahem, K. R., & Omar, N. (2015). Rule-based named entity recognition for drug-related crime news documents. *Journal of Theoretical and Applied Information Technology*, *77*(2).

Reyes-Galaviz, O. F., Pedrycz, W., He, Z., & Pizzi, N. J. (2017). A supervised gradient-based learning algorithm for optimized entity resolution. *Data & Knowledge Engineering*, *112*, 106–129. https://doi.org/10.1016/j.datak.2017.10.004

Riaz, K. (2010). Rule-based named entity recognition in Urdu.

Saha, S., & Ekbal, A. (2013). Combining multiple classifiers using vote based classifier ensemble technique for named entity recognition. *Data & Knowledge Engineering*, *85*, 15–39. https://doi.org/10.1016/j.datak.2012.06.003

Saha, S. K., Narayan, S., Sarkar, S., & Mitra, P. (2010). A composite kernel for named entity recognition. *Pattern Recognition Letters*, *31*(12), 1591–1597. https://doi.org/10.1016/j.patrec.2010.05.004

Salaken, S. M., Khosravi, A., Nguyen, T., & Nahavandi, S. (2018). Seeded transfer learning for regression problems with deep learning. *Expert Systems with Applications*, *115*, 565–577. https://doi.org/10.1016/j.eswa.2018.08.041

Sarawagi, S. (2008). Information extraction. *Foundations and Trends Databases*, *1*(3), 261–377.

Shaalan, K. (2010). Rule-based approach in Arabic natural language processing. *International Journal of Information and Communication Technology*, *3*(3), 11–19.

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* (pp. 1–1). China: IEEE Access. https://doi.org/10.1109/access.2018.287020

Singh, U., Goyal, V., & Lehal, G. S. (2012). Named entity recognition system for Urdu. International conference on computational linguistics, 2507–2518.

Tran, V. C., Nguyen, N. T., Fujita, H., Hoang, D. T., & Hwang, D. (2017). A combination of active learning and self-learning for named entity recognition on Twitter using conditional random fields. *Knowledge-Based Systems*, *132*, 179–187. https://doi.org/10.1016/j.knosys.2017.06.023

Tripathy, A., Agrawal, A., & Rath, S. K. (2016). Classification of sentiment reviews using n-gram machine learning approach. *Expert Systems with Applications*, *57*, 117–126. https://doi.org/10.1016/j.eswa.2016.03.028

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need.

Wang, C., Ma, X., Chen, J., & Chen, J. (2018). Information extraction and knowledge graph construction from geoscience literature. *Computers & Geosciences*, *112*, 112–120.

Wang, H., Zheng, J. G., Ma, X., Fox P., Ji, H., (2015). Language and domain independent entity linking with quantified collective validation. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP2015). 10 pp. https://doi.org/10.18653/v1/D15-1081

Wang, W., Velswamy, K., Hao, K., Chen, L., & Pedrycz, W. (2019). A hierarchical memory network-based approach to uncertain streaming data. *Knowledge-Based Systems*, *165*, 1–12. https://doi.org/10.1016/j.knosys.2018.11.011

Wang, Y., Yu, Z., Chen, L., Chen, Y., Liu, Y., Hu, X., & Jiang, Y. (2014). Supervised methods for symptom name recognition in free-text clinical records of traditional Chinese medicine: An empirical study. *Journal of Biomedical Informatics*, *47*, 91–104. https://doi.org/10.1016/j.jbi.2013.09.008

Wu, L., Xue, L., Li, C., Lv, X., Chen, Z., Jiang, B., & Xie, Z. (2017). A knowledge-driven geospatially enabled framework for geological big data. *ISPRS International Journal of Geo-Information*, *6*(6), 166. https://doi.org/10.3390/ijgi6060166

Xiao, F., Chen, Z., Chen, J., & Zhou, Y. (2016). A batch sliding window method for local singularity mapping and its application for geochemical anomaly identification. *Computers & Geosciences*, *90*, 189–201. https://doi.org/10.1016/j.cageo.2015.11.001

Yang, H.-F., & Chen, Y.-P. P. (2018). Hybrid deep learning and empirical mode decomposition model for time series applications. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2018.11.019

Yuan, Y., Xun, G., Suo, Q., Jia, K., & Zhang, A. (2018). Wave2Vec: Deep representation learning for clinical temporal data. *Neurocomputing*. https://doi.org/10.1016/j.neucom.2018.03.074

Zeng, G., Chuang, Z., & X Bo, L. Z. (2009). CRFs-based chinese named entity recognition with improved tag set. In *World Congress on Computer Science & Information Engineering* (Vol. 5, pp. 183–280). Washington, DC: IEEE.

Zhang, S., & Elhadad, N. (2013). Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. *Journal of Biomedical Informatics*, *46*(6), 1088–1098. https://doi.org/10.1016/j.jbi.2013.08.004

Zheng, J., Fu, L., Ma, X., & Fox, P. (2015). SEM+: Tool for discovering concept mapping in Earth science related domain. *Earth Science Informatics*, *8*(1), 95–102. https://doi.org/10.1007/s12145-014-0203-1

Zheng, S., Hao, Y., Lu, D., Bao, H., Xu, J., Hao, H., & Xu, B. (2017). Joint entity and relation extraction based on a hybrid neural network. *Neurocomputing*, *257*, 59–66. https://doi.org/10.1016/j.neucom.2016.12.075

Zhou, P., & El-Gohary, N. (2017). Ontology-based automated information extraction from building energy conservation codes. *Automation in Construction*, *74*, 103–117.

Zhou, P., Xu, J., Qi, Z., Bao, H., Chen, Z., & Xu, B. (2018). Distant supervision for relation extraction with hierarchical selective attention. *Neural Networks*. https://doi.org/10.1016/j.neunet.2018.08.016

Zhu, G., & Iglesias, C. A. (2018). Exploiting semantic similarity for named entity disambiguation in knowledge graphs. *Expert Systems with Applications*, *101*.