Chinese Word Segmentation Based on Self-Learning Model and Geological Knowledge for the Geoscience Domain

Wenjia Li1,2, Kai Ma1, Qinjun Qiu1,2, Liang Wu1,2, Zhong Xie1,2, Sanfeng Li4, and Siqiong Chen1,2

1National Engineering Research Center for GIS, Wuhan, China, 2School of Geography and Information Engineering, China University of Geosciences, Wuhan, China, 3College of Computer and Information Technology, China Three Gorges University, Yichang, China, 4Wuhan Zondy Cyber Science & Technology Co. Ltd., Wuhan, China

Abstract
Chinese word segmentation (CWS) is the foundational work of geological report text mining and has an important influence on various tasks, such as named entity recognition and relation extraction. In recent years, the accuracy of the domain-general CWS model has been limited by the domain and large scale of the training corpus, especially data on Chinese geological texts. Training these CWS models also requires much manually annotated data, which takes a large amount of time and effort. When applying these existing models/methods directly to the geoscience domain, the segmentation accuracy and performance will drop dramatically. To address this problem, we pretrain the Bidirectional Encoder Representations from Transformer (BERT), which can leverage unlabeled domain-specific knowledge, on unlabeled Chinese geological text and then input them into a Bidirectional long short-term memory and Conditional random field (BiLSTM-CRF) model for extracting text features. Finally, the predicted tags are decoded by the CRF. The experimental results show that the F1 score of the proposed model reaches 96.2% on the constructed test set of geological texts. Additionally, experiments illustrate that our proposed model achieves comparable performance to that of other state-of-the-art models, and the proposed cyclic self-learning strategy can be further extended to other domains.

Plain Language Summary
The supervised word segmentation model commonly lacks specialized knowledge in the training data set and has poor adaptability to the domain. This study proposes a sequential annotation model for geoscience text, which automatically constructs domain training-corpus and realizes word segmentation taking into account the long-distance dependence of sentences. We hope that our approach will serve as an alternative method that deserves further study.

1. Introduction
In the era of big data, the China National Geological Archives have accumulated massive amounts of geological data. Facing these large amounts of geological data, especially unstructured geological data, it is necessary to further establish big data and quantitative methods to obtain “data resources” and the formation of the core “data knowledge” new ways of thinking. Due to the hybridism, variability, robustness, and correlation of geological data varying with time, space and geological body, how to make full use of the rich knowledge and information contained in geological data has become the key to the study of geological big data (G. Liu et al., 2020; C. Wu et al., 2016). It means an urgent need to improve information extraction, knowledge mining, and knowledge association in heterogeneous geological data (Li & Shao, 2009). Recent studies have shown that the performance of downstream tasks for text mining, such as part-of-speech tagging, retrieve text from images (Shao et al., 2020; Zhou et al., 2017) and named entity recognition (Ma et al., 2018; Qiu et al., 2018, 2019; L. Wu et al., 2017), strongly depends on high precision geological word segmentation methods. This is because all of these downstream tasks, beside the extraction of multi-level and multi-dimensional image features required in retrieving text from images, require the system to have a good understanding of the text, which is the cornerstone of Chinese word segmentation (CWS). Natural language processing (NLP) (Dauphin et al., 2016; Kim, 2014) and development frameworks (such as TensorFlow and Keras) based on deep learning models (Chung et al., 2015; D. Li, 2019; Shi et al., 2016), which have been rapidly developed recently, provide new research paths for geological report processing tasks (Tian et al., 2020; C. Wu et al., 2020).
There are a large number of domain-specific terms, such as geomorphology, stratigraphic distribution, lithology, structure and geological history, in geological report text (Chen et al., 2020; Zhang & Liu, 2019). When the existing CWS method is directly applied to the geological field, due to the limitation of dictionary integrity and timeliness, there will be ambiguity and Out of Vocabulary (OOV) words, which will result in a low recall rate. In addition, a change in the context of the in-vocabulary (IV) geological word may lead to a change in the meaning of the word, resulting in a decrease in recognition performance (Qiu et al., 2018). Table 1 shows the main segmentation difficulties of geological data. For example, the word segmentation result of the sentence is “The strata is mainly distributed in the Geri Tu fire prevention station (Correct word segmentation results: the Geri Tu fire prevention station. Geri Tu fire prevention station is an exclusive place name in China) and Dundehabuqil/ south mountain, and invaded by Chagan Chulu coarse-grained biotite granite (Correct word segmentation results: Chagan Chulu coarse-grained biotite granite)”. In this result, the slash “/” in parentheses indicates the segmentation result of the corresponding word. When a large number of annotated domain corpora are added to train the existing models, the efficiency of the word segmentation results can be improved to a certain extent (Li & Guo, 2018; J. Liu et al., 2019; H. Zhang et al., 2020), but such an annotated corpus is currently lacking in the geological field. The core of the annotation of domain corpus is to identify domain terms, and the current research shows that the domain ontology contains rich domain terms and their constraint rules (Shao et al., 2014), while word similarity can find out the similar words to the target words, so this study considers the corpus construction task using word similarity and ontology constraints.

At present, CWS models can be divided into two types: supervised and unsupervised (Deng et al., 2016). Supervised learning requires a large-scale, high-quality annotated corpus for model training to achieve a high recognition accuracy (Graves, 2012). However, trained models have a strong dependence on the training corpus, poor portability, and poor recognition ability for new vocabulary. In summary, it is necessary to study unsupervised or weakly supervised segmentation methods for geological reports.

This study proposes a joint self-learning strategy assisted model for geological ontology to complete the task of word segmentation in geological report texts, focusing on the lack of a mature corpus and time-consuming and laborious manual geological report text annotation. The model uses four-digit labeling (B, M, E, and S). Pretraining language model (Bidirectional Encoder Representations from Transformer, BERT) and the bidirectional long short-term memory (BiLSTM) neural network structure are introduced into the automatic annotation process. The embedding vector of BERT’s pretrained words in the training process can better express the semantic information of words in different contexts. The experimental results show that the self-learning strategy, the development of domain knowledge and the addition of BERT in the pre-training model can effectively improve the effect of CWS in the geoscience domain, and the model reaches an F1 score of 96.2% in the constructed geological domain corpus. In summary, the contributions of this study are as follows:

| Table 1 |
| --- |
| **The Main Difficulties and Examples of CWS in Geology Material** |
| Category | Examples | Examples |
| **Toponym** | The thematic/work area is located at the southern of the central/ Gangdise Mountains/ and south of the Yarlung/Zangbo River. | The thematic/work area is located at the southern of the central/ Gangdise Mountains/ and south of the Yarlung/Zangbo River. |
| **Dimensional orientation** | The whole formation is distributed in a long and narrow strip to nearly east-west. | The whole formation is distributed in a long and narrow strip to nearly east-west. |
| **Landform** | Clastic/karst, loess and clay/karst, thermally melted karst and lava/karst in volcanic rocks, etc. | Clastic karst, loess and clay karst, thermally melted karst and lava karst, etc. |
| **Geologic age** | Late/Jurassic epoch, Early/Cretaceous Epoch and Middle/Cretaceous/Neotethys/Evolution | Late Jurassic epoch, Early Cretaceous Epoch and Middle Cretaceous/Neotethys/Evolution |
| **Geologic structure** | It is located in the south side of the compound syncline of Han Wangshan. | It is located in the south side of the compound syncline of Han Wangshan. |
We proposed the self-learning word segmentation model assisted by ontology in the geological domain, which implements CWS in the geoscience taking into account the abundant word level features, grammatical structure features and semantic features in sentences;

We combine self-learning strategies with domain knowledge to automatically construct the domain training corpus without manual intervention;

A set of experiments demonstrate the effectiveness of the proposed method in this study on existing manually constructed hybrid datasets.

The remainder of this study is structured as follows: Section 2 describes the current related work of geological segmentation. Section 3 proposes and describes the proposed method for CWS in geological survey reports. The results and analysis are reported in Section 4. Section 5 provides the conclusion and suggestions for future research.

2. Related Work

Text segmentation is indispensable basic research in the field of NLP. Many researchers have devoted themselves to the study of this domain. General word segmentation methods can be divided into rule-based and statistical word segmentation methods. The former is mainly based on word-based rule matching for a given constructed dictionary, such as positive maximum matching rules, reverse maximum matching rules (Luo et al., 2018) and bidirectional matching rules (Huang et al., 2015; Yunita et al., 2010). The latter is trained on annotated Chinese text to obtain different models: Hidden Markov models (HMMs) and Conditional Random Fields (CRFs), statistical machine learning models (Du et al., 2018; Huang et al., 2017; Liang et al., 2019; Y. Liu et al., 2014; Zhang & Li, 2016), deep learning models (Xu & Sun, 2016; Zhao et al., 2020), etc. Based on the trained model, the text of the unknown label is segmented. Most of these pretrained statistical models are character-based tagging approaches (Chen et al., 2015; Xu & Sun, 2016; Yuan et al., 2020), which treat word segmentation as a word-based sequential labeling problem. Thus, the model can treat the recognition of dictionary words and unregistered words equally. At present, BiLSTM is widely used in Chinese text segmentation. However, when learning longer sentences, it may be difficult for the model to acquire some important information. Therefore, a CRF is usually added to extract the local features of the current word before model training (Jia & Xu, 2018; Wang et al., 2017) or adding a pretrained language model, such as Embeddings from Language Models (ELMo) (Chu et al., 2020; Peters et al., 2018), generative pretraining (GPT) (Radford & Salimans, 2018), BERT (Devlin et al., 2018), etc.

BERT, which has good performance in many existing NLP tasks considering the contextual bidirectional characteristics of words (Ke et al., 2020), can generate a series of cross-layer word embedding vectors. Therefore, the BERT-BiLSTM-CRF model is selected in this study to complete the geological CWS task. The training of this model requires a large number of annotated corpora. When the People’s Daily Corpus and the SIGHAN-Bakeoff corpus are directly used for word segmentation of geological reports, the segmentation is usually inconsistent. For example, when the trained model treats the sentence, “This set of strata in the Upper Jurassic Manketou Ebo Group is widely distributed in the surveyed area” with CWS, the result is “Upper/Jurassic/Manchurian/ketou OBO formation/this/set/stratum/in/survey area/within/widely distributed.” However, the “Upper Jurassic Manketou Ebo Group” is an objective stratigraphic object and should not be separated in Chinese. This limitation is mainly due to the lack of domain knowledge in the trained models, and the computer relies on only the People’s Daily Corpus and SIGHAN-Bakeoff corpus to learn the naming rules of stratum entities. Domain dictionary rules and supervised methods are not effective in solving OOV words and ambiguity in domain-specific word segmentation.

To overcome the lack of a mature corpus in a domain, the literature (Chen et al., 2018) used two mature corpora, the geological and mineral corpus after manual annotation and the People’s Daily Annotated Corpus, as experimental corpora, to solve the problem of the identification of some OOV words. (Ma et al., 2018) used terms in the geological dictionary and the Geological and Mineral Resources Dictionary to match terms and label Chinese geosciences literature in the knowledge network based on characters to obtain an annotated corpus for word segmentation model training. In (J. D. Zhang et al., 2011), Medical Subject Headings (Lipscomb CE, 2000) is used as a controlled vocabulary thesaurus for corpus generation along with literature abstracts from MEDLINE. These studies show that the use of hybrid corpus can better alleviate
the identification of OOV words by considering both generic domain text features and specific domain text features, so this study also employs a hybrid corpus for model training. The literature (Han & Chang, 2015) introduced circular self-learning and collaborative strategies into domain corpus tagging, further improving the adaptability of the traditional word segmentation model in the professional domain. However, in the cyclic iteration self-learning strategy model of convergence performance, both artificial ways are used to select correct results from the annotation results to extend the training corpus. This strategy can, to a certain extent, improve the adaptability of different segmentation models in the field, but the artificial selection process has certain subjective factors. In the process of artificial selection, unregistered words, domain-specific vocabularies, place names, institutions and other complexities easily cause omissions. In addition, in the process of artificial selection, a large amount of domain-specific knowledge in the field is required. Otherwise, it is easy to make mistakes in the selection, which will make the training model worse. Therefore, this study proposes a domain corpus tagging method based on a self-learning strategy assisted by domain ontology, which uses the similarity of domain knowledge and domain vocabulary in ontology to replace the process of artificial selection and reduce the influence of subjective factors on the model.

3. Self-Learning Word Segmentation Model Assisted by Ontology in the Geological Domain

3.1. Automatic Generation of a Geological Corpus

In the training corpus of the existing supervised word segmentation model, the word frequency of geological vocabulary is low, and the number of words is very limited. Moreover, the change in context due to the change in domain will also lead to the deterioration of the recognition performance of a great deal of words in vocabulary. Therefore, this study proposes a self-learning training strategy assisted by the domain ontology (Figure 1) to automatically generate the domain corpus and strengthen the word segmentation performance and effect with the part-of-speech model in the geoscience domain.
At present, the data set used in the training of the general CWS model is usually an open corpus, such as The People’s Daily Corpus or SIGHAN-Bakeoff (Luo et al., 2018). There are few contents in the corpus involving geology, so a large number of geological corpora should be added to assist the training of the mature word segmentation model. In this study, 5,870 geological terms (Figure 2a) are obtained through the processing of the “geological dictionary”, and 2,175 field terms (Figure 2b) are obtained from the existing geological ontology (C. Li, 2010, 2013).

For self-learning aided by the domain ontology model, the previous manual “deletion” process (Han & Chang, 2015; Qiu et al., 2018), namely, the deletion of labeling errors in the corpus, is adjusted to the reservation process of labeling results with ontology-based domain constraints or similarity judgment as “Yes.” In other words, with the assistance of existing knowledge, the model is allowed to select the correct annotation results that meet the conditions from the model as the intermediate corpus for retraining and obtaining the intermediate model. The process is repeated times until the model converges, resulting in an improved geological word segmentation model. When the first judgment’s result is “Yes” in Figure 1, “Whether it has the domain vocabulary constraint in the ontology?,” the model returns the annotation result to the words set with the correct annotation (i.e., intermediate corpus) as the input of the intermediate segmentation model to continue training the model. When the judgment result is “No,” the model will make a second judgment to calculate the similarity between the domain named entity obtained by the model annotation and the existing entity or vocabulary. The annotation is considered correct only if the similarity is greater than the given threshold value θ, and the annotation result is returned to the intermediate corpus for further model training. The contents of the two judgments in the model are as follows.

**Stepped Landform** Multilayered terrain. Refers to a region, from the lowest water surface to the highest mountain nearby, in the horizontal distance of the terrain transverse section, by the multilevel coastal terrace, lakeside terrace, river terrace and multilevel mountain planation surface as well as multi-layer karst cave, multi-layer diluvial fan composition of the terraced terrain. In many parts of China’s coastal and inland areas, the layered terrain is very striking. He shows that the crust of the region has undergone either upward or downward or relatively stable neotectonic movements. It is of great significance to study the history of terrain development and tectonic movement in a region. Layered topography is a geomorphic surface formed in different periods under the action of internal and external dynamic geology, such as terrace surface and planation surface. Some of them were unified planation planes, which formed planes of different heights after structural changes in the later period.

**Correlative Sediments** In the process of the formation of denudation landform, the denudated detritus accumulates in the adjacent low-lying area almost at the same time. Such sediments are called the related sediments of the denudation landform. The geomorphic age of denudation landforms can be inferred from the stratigraphic age of sediments. This method of geomorphic age estimation is called correlative sedimentation.

**Method of Correlative Sediments** The method of inferring the corresponding age of denudation geomorphology according to the sedimentary age of the sediments is called the related sedimentary method. See Correlative sediments.

Figure 2. (a) Geological vocabulary and (b) Geological ontology.
(1) Domain vocabulary constraints in geological ontology

The geological domain ontology describes the concepts of the geological domain entity, the geological domain entity attributes, the interaction between geological domain entities and the formal description of the characteristics and laws of the geological domain. In this study, by using the open-source semantic framework Jena (https://jena.apache.org/), organized by Apache, the domain vocabularies are extracted, and the corresponding domain constraint rules are constructed from the geological ontology constructed by the China Geological Survey. The study transforms it into an ontology model based on resource description framework (RDF) expression. RDF expressions have strong semantic descriptive power, with each triplet corresponding to a logical expression or statement of reality. This study implements the extraction of 2,175 concepts in 7 categories, including paleontology, strata, the rock types, the extraction and formalized expression of geological entity naming rules and geological concept relations, such as epistatic relations, inferior relations and equivalent relations. These named entity rules and the correlations between concepts can help the model autonomously recognize and judge whether the annotation results are correct and keep the annotation correct results. The extraction of the naming rules of geological entities here needs to be determined according to the attributes of the naming rule of the entities (see Table 2). The regional stratigraphic unit “group” contains attribute information in the ontology “naming conventions (hasNamingConventions): “group” can name after geographical names and rock properties, but “group” cannot omit.” This attribute,
Earth and Space Science

As text attribute information cannot be identified by the computer and cannot be included in the constraint rules for judgment. Therefore, this study transforms it into reliable and intelligible expressions, expressed as "^(geographic name)(rock properties)(group)$".

In addition, since the contents of the geological ontology are stored in an Access database, the upper and lower relationships are well expressed, but the equivalent relationship is not clearly expressed (as shown in Figure 3). For example, in the original database, as shown in Figure 3, “YS1: Magmatic rocks” has the attribute “Equivalent” (alias): “Igneous rock”. This simple relational storage scheme in the original database makes it difficult for the computer to understand the relationship between the heterogeneous geological terms “igneous rock” and “magmatic rock”. This study converts the relationship into an RDF-based expression, known as a triple: (“magmatic rocks”, isEquivalent, “igneous rock”). When the labeled result of the initial model is “igneous rock”, the system can judge the marked result by this triple and return it as the correct labeled result to the intermediate model ($M_{\text{md}}$) for model training.

According to the constraint rules in the existing ontology, the label correctness of the extracted domain entity can be judged well. However, this judgment method has some limitations, which is mainly due to the domain entities with rule constraints in geological ontology are limited, such as “Pleistocene cohesive soil” as a kind of objective geological entity does not conform to rule R5 (Table 2). Therefore, after judging

| Domain-named entity | Text attributes | Constraint rule | Examples |
|---------------------|-----------------|-----------------|----------|
| **Group (Regional stratigraphic unit)** | Rock properties may be appended to geographical names, but “group” cannot be omitted | $R1: ^\text{(geographical names)(rock properties)(group)$} | Mount Emei basalt group |
| **Section (Regional stratigraphic unit)** | It may be named by geographical names or by lithology without geographical names | $R2: ^\text{(geographical names)(rock properties)(section)$} | Wanwangou section, Limestone section |
| **Fine clastic rocks** | The names of fine clastic rocks are as follows: cement + structure + basic name | $R3: ^\text{(cement)(structure)(basic name)$} | Argillaceous siltstone |
| **Mesoclastic rocks** | The names of middle clastic rocks are as follows: cement + structure + clastic composition + basic name | $R4: ^\text{(cement)(structure)(clastic composition)(basic name)$} | Medium feldspar quartz sandstone |
| **Argillaceous rocks** | The names of argillaceous rocks are as follows: color + mixture + mineral composition in clay + basic name | $R5: ^\text{(color)(mixture)(mineral composition in clay)(basic name)$} | Kaolinite clay or Kaolinite claystone |

Table 2: Examples of Domain Named Entity Constraints

![Figure 3. Some examples of geological ontology in the Access database (provided by the Development Research Center of the China Geological Survey).](figure-3.png)
the ontology constraints, this study also makes a judgment 2). Similarity of domain words to improve the model’s judgment of the correctness of labels.

(2) Similarity of domain words

The word similarity calculates the similarity among multiple words in a quantitative way. The calculation methods can be divided into four types: string-based similarity calculations, corpus-based similarity calculations, knowledge-based similarity calculations and other methods. Among them, corpus-based similarity calculations can be divided into three methods based on different models: bag-of-words (BOW) models, neural networks and search engines (Gomaa & Fahmy, 2013; Pradhan et al., 2015). The most typical method based on neural networks is the word2vec model (Mikolov et al., 2017). Through network training in the word2vec model, the low-dimensional vector corresponding to each word can be trained from the unstructured text, and the value of each dimension represents a word characteristic with certain semantic and grammatical interpretation. This representation brings similar words closer to the cosine distance in the same vector space. Therefore, this study takes advantage of this model to measure the similarity between word pairs. The specific calculation formula is shown in Equation 1:

$$\text{Similarity}(x_i, y_i) = \cos(x_i, y_i) = \frac{\sum_{t=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{t=1}^{n} (x_i)^2} \sqrt{\sum_{t=1}^{n} (y_i)^2}}$$  

(1)

Similarity($x_i, y_i$) indicates the similarity between the marked word $x_i$ and the existing domain term $y_i$. The higher the value is, the higher the similarity between the two words will be; otherwise, the lower the similarity will be. Similarity($x_i, y_i$) ranges from −1 to 1. When the Similarity($x_i, y_i$) value is greater than the given threshold value (see Section 4.3.1 for the influence of different thresholds on the segmentation results in this study), this study believes that the entities marked by the model are similar to known entities; that is, the results marked by the model are correct. Otherwise, the marked result is wrong.

3.2. Word Segmentation Method Based on the BERT-BiLSTM-CRF Model

In recent years, pretraining models have always been a research hotspot for the upstream task of CWS. BERT, as an advanced language preprocessing representation model, can effectively obtain a high-quality word vector representation mode, which is more beneficial to downstream tasks than the upstream task (Alsentzer et al., 2019; Tenney et al., 2019). In this model, two new unsupervised prediction tasks are used to perform pre-training on BERT, namely, the masked language model (Masked LM) and the fine-tuning task. Masked LM uses bi-directional self-attention in its transformer structure, so that each word that is randomly masked can pay attention to the words before and after it and obtain the word’s bi-directional representation. The latter adds task-specific parameters to the already trained model to obtain a new task-based model. In this study, the BERT-BiLSTM-CRF structure is adopted in geological word segmentation. The overall structure diagram is shown in Figure 5, and the BiLSTM-CRF structure is shown in Figure 4.

The BERT-BiLSTM-CRF structure mainly consists of three modules. First, the annotated corpus is input into the BERT pretraining language module to obtain the corresponding word vector representation. Then, the generated word vector is input into the BiLSTM network for subsequent processing. Finally, the output of the BiLSTM layer is decoded and predicted by the CRF layer, and then the whole process of CWS is carried out through the obtained prediction label.

The greatest advantage of the architecture presented in this study is to use BERT model in the geological report text segmentation task, which is the difference between traditional training word vectors in advance, and only the sequence of sentences is needed by preprocessing the BERT model. It can automatically extract
word-level features, grammatical structure features and rich semantic features in sequential sentences. For the BERT model, acquisition and learning at each level are progressive. The bottom level is mainly represented by learning phrases, the middle level is represented by learning the internal structure of the syntax in the sequence, and the last level captures the rich semantic information in the whole sequence (Jawahar et al., 2019).

The traditional word vector training model (such as the BiLSTM-CRF model) mainly obtains representation information at the character and phrase levels but has difficulty capturing the representation at the syntactic internal structure and sentence levels. The BERT model can well represent semantic word vectors, especially for long distance-dependent sequences. In this study, multilayer vectors are selected as input text features for further training.

4. Experiment and Results Analysis

4.1. Experimental Environment and Evaluation Parameters

The training environment and configuration for the BERT-BiLSTM-CRF architecture proposed in this study are shown in Table 3.

In this study, the Adam optimizer is selected for the experiment, and the learning rate is set to 0.001. In the network architecture, the LSTM dimension is 200, the batch size is 64, and the maximum sequence length (max_seq_len) is 128. To prevent overfitting, dropout is adopted in the experiment and set to 0.05.

| Table 3 | Training Environment and Configuration of the BERT-BiLSTM-CRF Model |
|---------|---------------------------------------------------------------------------------|
| Operating system | Ubuntu 14.04 64bit |
| CPU | 2 * Intel(R) Xeon(R) CPU E5-2620 v2 @ 2.10GHz |
| GPU | 2 * Nvidia Tesla K20 |
| Memory | 96GB |
| Python | 2.7 |
| Deep learning framework | Keras |
When evaluating the CWS model, the precision (P), recall (R) and harmonic mean of precision and F1 score 
$$(F = \frac{2PR}{P + R})$$ are commonly used as evaluation indexes. $R_{\text{OOV}}$ and $R_{\text{IV}}$ represent the percentage 
of OOV words and the percentage of IV words (i.e., those that are contained in the corpus), respectively, that 
are correctly segmented. For a given training model, $R_{\text{OOV}}$ represents the ability to extend the model to other, 
genre domains, and $R_{\text{IV}}$ represents the ability to use the training model for prediction. The values of these 
five indicators are all within [0, 1]. The higher the value is, the better the performance of the segmentation 
model. The results of all the tests were tested with 10-fold cross-validation in this study, and the average 
score of 10 independent runs was calculated as the test result.

### 4.2. Data Set

In the experiment, 43 open geological reports are collected from the National Geological Archive as the 
corpus. These geological reports are representative and were written and recorded by different geological 
experts, which describe the geological survey information in different areas at different times. Moreover, 
these reports also have domain complexity characteristics from simple text to complex text. A standard test 
set is constructed with reference to the general domain segmentation criteria by manual annotation for 
the collected geological reports. Based on the standard of the 2005 SIGHAN-Bakeoff corpus created in the 
genre domain, we invite experts with background knowledge on earth science and CWS to mark the corpus 
to avoid semantic deviation. The resulting corpus is named GEO Corpus with a size of 7.8M. Previous 
studies have shown that mixed corpus can improve the recognition efficiency of OOV (Chen et al., 2018; Ma 
et al., 2018), and the standard general corpus SIGHAN-Bakeoff covers a variety of topics, including a small 
number of common geological nouns and a large number of words in the general domain, so this study 
uses mixed corpus for model training. In the experiment, we chose the representative Microsoft Research 
(MSR) and Peking University (PKU) Simplified Chinese versions as the training and test sets, respectively, 
with the size of 12.2M/5.63M respectively. Table 4 provides a detailed description of the data set, including 
the sentence length, data set size, and the proportion of OOV words. The 
collected geoscience corpus includes the Regional Geology Report (RGR), Environmental Geology Report (EGR), Mineral Geology Report (MGR), Hydrogeology Report (HGR), Geophysical Geology Report (GGR) and 
Remote Sensing Geology Report (RSGR).

### 4.3. Experimental Analysis and Discussion

#### 4.3.1. Influence of Self-Learning Strategies With Domain Knowledge

(1) Similarity threshold ($\theta$)

As stated in Section 3.1, the similarity threshold is a very important parameter in automatic corpus generation. This subsection verifies the
effect of the variation of this parameter on the experimental results through experiments. As illustrated in Table 5, with the decrease in the similarity threshold θ, the precision (P), recall (R) and F1 values of the word segmentation results in the geological reports all showed a trend of first increasing and then decreasing. The reduction in the similarity threshold θ first improves the recognition ability of the system. However, when it is reduced to 0.6, the recognition ability of the system reaches its peak. As θ continues to decrease, the system tolerance for domain vocabulary increases because the given threshold is too small, which may lead to a decrease in domain word recognition. For example, the similarity between the error participle result “sedimentary architecture” and the existing geological term “sedimentary structure” is ~0.57. When the threshold is set to 0.5, the system will send the result, which is mistaken as the correct word segmentation, into the model for further training. The introduction of false segmentation results will increase the error rate of the final word segmentation model during model iteration training and may even cause the model to fail to converge.

(2) Domain constraint rules

To verify that the domain constraint rules proposed in this study have a positive impact on the word segmentation results, a set of comparative experiments, namely, strategy A and strategy B, are carried out and shown in Table 6. Group A deletes the domain constraint rules in this study, and strategy B is the method proposed in this study. The similarity threshold (θ) in the experiment was selected as 0.6 according to the performance in Table 5. Table 6 shows that the domain constraint rules can significantly improve the system in terms of the word segmentation accuracy and recall, which is mainly due to these rules making up for the shortage of the word similarity to some extent, and make the system have a certain ability to recognize the error entity, such as “E’meishan basalt group”. By the naming rule R1: ‘(geographic name) (rock properties) (group)’ in Table 2, “E’meishan basalt group” and “Basalt group” can be directly identified as geological entities.

4.3.2. Different Network Structures Under Self-Learning Strategies

To verify that the BERT-BiLSTM-CRF model used has a better performance in geological CWS than other deep neural networks and their hybrid models, a comparative experiment on the segmentation results of the geological reports based on different network structures is carried out in this study. The experimental results are shown in Table 7. The sequential labeling results for CWS from the BiLSTM model are better than those from the LSTM model. The key part of BERT’s pretrained language model is the transformer structure, which can combine context for pretraining and learning phrase-level information representations as well as rich linguistic and semantic information features. The BERT-BiLSTM-CRF model outperforms the other models, and the precision and recall are improved by 0.037 and 0.04, respectively, compared with the BiLSTM-CRF model.

4.3.3. Influence of the Number of Iterations on the Word Segmentation Results

Figure 6 shows the segmentation effect obtained by the self-learning strategy for the corpus of geological reports. It can be seen from Figure 6 that the BERT-BiLSTM-CRF model can reach a high level and make further improvements as quickly as possible at the beginning of training and finally maintain at a relatively high level. However, the results obtained by the traditional deep learning models are relatively low at the beginning of training and can be increased only to a certain level after several iterations. From the perspective of the results of self-learning, with the increase in the number of iterations, the precision, recall and F1 score for the OOV words are constantly increasing and finally stabilize. This finding indicates that with the continuous iteration of the corpus, cyclic self-learning begins to fulfill its role. After a certain number of iterations, the F1 score fluctuates, that is, the cyclic self-learning effect begins to decrease. From the trends in the F1 scores in Figure 6, the introduction of the self-learning strategy makes the CWS method achieve a better per-

| Table 6 | The Segmentation Results With and Without Domain Constraints When the Similarity Threshold is 0.6 |
|---------|--------------------------------------------------|
| Strategy | P   | R   | F1  |
| strategy A (has no domain constraint rules) | 0.911 | 0.905 | 0.908 |
| strategy B (has domain constraint rules) | 0.963 | 0.961 | 0.962 |

| Table 7 | Word Segmentation Results for the Geological Reports With Different Network Structures |
|---------|---------------------------------|
| Model   | P     | R     | F1   |
| LSTM-CRF | 0.91  | 0.891 | 0.900 |
| BiLSTM  | 0.918 | 0.905 | 0.911 |
| BiLSTM-CRF | 0.924 | 0.914 | 0.919 |
| BiLSTM-CNN-CRF | 0.939 | 0.927 | 0.933 |
| BERT-BiLSTM-CRF | **0.963** | **0.961** | **0.962** |
formance in geological reports, with the F1 score reaching 0.967; in addition, the improvement effect of the recall of OOV words reaches 0.881.

4.3.4. Error Analysis of the Segmentation Results

Table 8 analyzes several word segmentation cases that cannot be handled by the proposed method in this study: ① There are inconsistencies in the word segmentation results in terms of context. For example, a “metallogenic belt” can sometimes be correctly segmented but sometimes cannot. The word segmentation model lacks sufficient global semantics, so additional constraints are needed to ensure the context segmentation consistency. ② Words that have no significant difference in the context feature expression, such as “nonferrous metals”, cannot be separated. In future work, we will try to add global constraints to improve the accuracy of high-score words.

Table 8
Examples of Common Word Segmentation Errors for Geological Data

| Description | Examples | Error analysis |
|-------------|----------|----------------|
| ① The results of word segmentation in context are inconsistent and the global character of word segmentation is not strong enough. | The ore-forming/background/of/the survey area/is located in/level II/metallogenic/belt/, Daxinganling Metallogenic Province/, Inner Mongolia | “metallogenic belt” should be used as a word. When doing word segmentation, some places should be separated, while some places cannot be separated. |
| ② There is no significant difference in the characteristic expression of words in context. | With/nonferrous metals precious metals/and/other/recent/economic benefits/of/minerals | “nonferrous metals” and “precious metals” in this context have no significant differences in characteristics, but should be divided into “nonferrous metals” and “precious metals”. |
5. Conclusion and Future Work

CWS plays an important role in many applications, from supporting named entity recognition to extracting relations from texts. This study proposed a method of CWS based on BERT-BiLSTM-CRF. Aimed at corpus construction, a self-learning strategy is adopted to enrich the domain-mature corpus, which is different from traditional methods. Considering that it does not rely on domain knowledge and manual design features, the BERT-BiLSTM-CRF deep learning model can better represent semantic characteristics at the character level than the traditional word vector representation. The experimental results show that the method proposed in this study can effectively enhance the word segmentation ability in unstructured geological reports and can identify a large number of geological OOV words.

The main contributions can be summarized from two perspectives. From the perspective of methodology, the proposed approach can be easily extended to other domains and support domain-specific CWS. From the perspective of the application, this study proposes a novel self-learning strategy to enrich the domain-mature corpus for developing training datasets. As suggested by the experimental results, combining the self-learning strategy with the deep learning model has achieved a better performance in segmenting Chinese words for the geoscience domain than applying the approaches based on linguistic features alone.

However, the method in this study also has some shortcomings, such as not taking into account the semantic environment in which the text is located, and it is difficult to distinguish the word meanings in different text scenes, so the model has some randomness for the separation of ambiguous words. In the future, we will consider how to enhance the domain adaptability of the training model by adding text scene or text topic features in addition to learning the features at the word, sentence and context levels. At present, when CWS is performed in both general and specialized domains, sentence-level features and context features are often added as semantic features. When specialized domain model training is carried out, a mixed corpus is often used to train the model, which often ignores the characteristics of the domain text scene where the field-specific text is located. Moreover, this method will be extended to other specialized fields to test the generalizability of the self-learning strategies and CWS methods proposed in this study.

Conflict of 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 study.

Data Availability Statement

All original data can be found in the Zenodo (https://github.com/cugdeeplearn/OntologyCWS).

References

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323.

Chen, J., Chen, J., & Wang, C. (2018). Research on segmentation of geological mineral text using conditional random field. China Mining Magazine, 101(9), 69–74. https://doi.org/10.12075/j.issn.1004-4051.2018.09.035

Chen, Q., Liu, G., He, Z., Zhang, X., & Wu, C. (2020). Current situation and prospect of structure-attribute integrated 3D geological modeling technology for geological big data. Bulletin of Geological Science and Technology, 39(4), 51–58.

Chen, X., Qiu, X., Zhu, C., & Huang, X. (2015). Gated recursive neural network for Chinese word segmentation. In Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (pp. 1744–1753). Association for Computational Linguistics.

Chu, D., Wan, B., & Hong, L. (2020). Geological entity recognition based on ELMo-CNN-BiLSTM-CRF Model. Earth science, 1–22.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2015). Gated feedback recurrent neural networks. In Proceedings of the 32nd international conference on international conference on machine learning. (pp. 2067–2075). Proceedings of Machine Learning Research.

Dauphin, Y. N., Fan, A., Auli, M., & Grangier, D. (2016). Language modeling with gated convolutional networks. Deng, K.-X., Peter, B., Li, K., & Liu, J. S. (2016). On the unsupervised analysis of domain-specific Chinese text[C]. Proceedings of the National Academy of Sciences, 113, 6154–6159. https://doi.org/10.1073/pnas.1516510113

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding.

Du, L., Xia, C., Deng, Z., Lu, G., Xia, S., & Ma, J. (2018). A machine learning based approach to identify protected health information in Chinese clinical text. International Journal of Computer Applications, 116, 24–32. https://doi.org/10.1007/s10147-018-05010

Gomaa, W. H., Fahmy, A. A. (2013). A Survey of text similarity approaches. International Journal of Computer Applications, 68(13), 13–18.

Graves, A. (2012). Studies in Computational Intelligence. In Supervised Sequence labelling with recurrent neural networks. Springer.

https://doi.org/10.1007/978-3-642-24797-2
Han, D., & Chang, B. (2015). A domain adaptive approach to Chinese word segmentation model. *Chinese Journal of Computers*, 38, 272–281.

Huang, L., Du, Y., & Chen, G. (2015). GeoSegmenter: A statistically learned Chinese word segmenter for the geoscience domain. *Computers & Geosciences*, 76, 11–17. https://doi.org/10.1016/j.cageo.2014.11.005

Huang, S., Sun, X., & Wang, H. (2017). Addressing domain adaptation for chinese word segmentation with global recurrent structure. In Proceedings of the eighth international joint conference on natural language processing (Volume 1: Long papers). (pp. 184–193). Asian Federation of Natural Language Processing.

Jawahar, A., Sagot, B., & Seddah, D. (2019). What does BERT learn about the structure of language? In Proceedings of the 57th annual meeting of the association for computational linguistics. Association for Computational Linguistics.

Jia, Y., & Xu, X. (2018). Chinese named entity recognition based on CNN-BiLSTM-CRF. In 2018 IEEE 9th international conference on software engineering and service science (ICSESS). (pp. 1–4). IEEE.

Ke, Z., Shi, L., Meng, E., Wang, B., Qiu, X., & Huang, X. (2020). Unified multi-criteria Chinese word segmentation with BERT. arXiv.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *Eprint Arxiv*.

Li, C. (2010). The technical framework of geological survey information grid. In 2010 18th international conference on geoinformatics. IEEE.

Li, C. (2013). Geological domain ontology and its application. In China geological survey information Grid-Technology & methodology (pp. 27–51). Beijing: Geological Publishing House.

Li, D. (2019). The intelligent processing and service of spatiotemporal big data. *Journal of Geo-Information Science*, 21(12), 1825–1831.

Li, D., & Shao, Z. (2009). The new era for geo-information. *Science in China Series F: Information Sciences*, 7, 1233–1242. https://doi.org/10.1007/s11432-009-0122-9

Li, S., & Guo, Y. (2018). Biomedical named entity recognition with CNN-BiLSTM-CRF. *Journal of Chinese Information Processing*, 32, 116–122.

Liang, Y., Yang, M., Zhu, J., & Yi, S. M. (2019). Out-domain Chinese new word detection with statistics-based character embedding. *Natural Language Engineering*, 25, 1–17. https://doi.org/10.1017/S1351324918000463

Lipscomb, C. E. (2000). Medical subject headings (MeSH). *Bulletin of the Medical Library Association*, 88, 265–266.

Liu, G., Wu, C., & He, Z. (2020). Data model for geological spatiotemporal big data expression and storage management. *Bulletin of Geological Science and Technology*, 39(1), 164–174.

Liu, J., Wu, F., Wu, C., Huang, Y., & Xie, X. (2019). Neural Chinese word segmentation with dictionary knowledge. *Neurocomputing*, 338, 46–54. https://doi.org/10.1016/j.neucom.2019.01.085

Liu, Y., Zhang, Y., Che, W., Liu, T., & Wu, F. (2014). Domain adaptation for CRF-based Chinese word segmentation using free annotations. In Proceedings of the 2014 conference on empirical methods in natural language processing (pp. 864–874). EMNLP.

Luo, G., Fei, H., & Dai, Y. (2018). Research on Chinese word segmentation based on anti-order dictionary. *Computer technology and development*, 18, 80–83.

Ma, X., Carranza, E. J. M., Wu, C., & van der Meer, F. D. (2018). Ontology-aided annotation, visualization, and generalization of geological time-scale information from online geological map services. *Computers & Geosciences*, 40, 107–119. https://doi.org/10.1016/j.cageo.2011.07.018

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2017). *Advances in pre-Training Distributed word representations*. arXiv preprint arXiv:1712.01504.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 2227–2237). Association for Computational Linguistics.

Pradhan, N., Gyanchandani, M., & Wadhvani, R. (2015). A review on text similarity technique used in IR and its application. *Journal of Computer Applications*, 6, 11–17. https://doi.org/10.1016/j.cageo.2014.11.005

Qiu, Q., Xie, Z., & Wu, L. (2018). A cyclic self-learning Chinese word segmentation for the geoscience domain. *Geomatica*, 72, 16–26. https://doi.org/10.1139/geomat-2018-0007

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.

Radford, A., & Salimans, T. (2018). Improving Language understanding by generative pre-training. *Shao, Z., Zeng, W., Zhou, X., & Liu, J. (2014). Semantic graph construction for 3D geospatial data of multi-versions. *Optik-International Journal for Light and Electron Optics*, 125(6), 1730–1734.*

Shao, Z., Zhou, W., Deng, X., Zhang, M., & Cheng, Q. (2020). Multilabel remote sensing image retrieval based on fully convolutional network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13(13), 318–328. https://doi.org/10.1109/jstars.2019.2961634

Shi, Y., Yao, K., Chen, H., Hu, D., Pan, Y. C., & Hwang, M. Y. (2016). Recurrent support vector machines for slot tagging in spoken language understanding. In *HLT-NAACL* (pp. 393–399). The Association for Computational Linguistics.

Tenney, I., Das, D., & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. arXiv preprint arXiv:1905.05950.

Tian, Y., Wu, C., & Weng, Z. (2020). Technologies of geological big data visualization. *Bulletin of Geological Science and Technology*, 39(4), 29–36.

Wang, X., Xu, R., He, Y., & Chen, T. (2017). Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. *Expert Systems with Applications*, 72, 221–230.

Wu, C., Liu, G., Zhang, X., He, Z., & Zhang, Z. (2016). Discussion on geological science big data and its applications. *Chinese Science Bulletin*, 61(16), 1797–1807. https://doi.org/10.1360/n732015-01035

Wu, C., Liu, G., & Zhou, Q. (2020). Fundamental problems of integrated application of big data in geoscience. *Bulletin of Geological Science and Technology*, 39(4), 1–11.

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

Xu, J., & Sun, X. (2016). Dependency-based gated recursive neural network for Chinese word segmentation. In Meeting of the association for computational linguistics (pp. 567–572). Association for Computational Linguistics.

Yuan, Z., Liu, Y., Yin, Q., Li, R., Feng, X., Zhang, G., & Yu, S. (2020). Unsupervised multi-granular Chinese word segmentation and term discovery via graph partition. *Journal of Biomedical Informatics*, 110. https://doi.org/10.1016/j.jbi.2020.103542

Yunita, S., Hassan, M. F., & Zamin, N. (2010). Rule-based pattern extractor and named entity recognition: A hybrid approach. *Information Technology IEEE*. 

---

**Earth and Space Science**

10.1029/2021EA001673

**LI ET AL.**

14 of 15
Zhang, H., Zong, Y., & Chang, B. (2020). Medical entity annotation standard for medical text processing. In Proceedings of the 19th Chinese national conference on computational linguistics.

Zhang, J. D., Geer, L. Y., Bolton, E. E., & Byrant, S. H. (2011). Automated annotation of chemical names in the literature with tunable accuracy. *Journal of Cheminformatics*, 3(1), 1–10. https://doi.org/10.1186/1758-2946-3-52

Zhang, Q., & Liu, X. (2019). Big data: New methods and ideas in geological scientific research. *Big Earth Data*, 3(1), 1–7. https://doi.org/10.1080/20964471.2018.1564478

Zhang, Y., & Li, Z. (2016). Chinese Word Segmentation method based on Gauss word Length Characteristics. *Chinese Journal of Information Technology*, 30(5).

Zhao, L., Zhang, A., Liu, Y., & Fei, H. (2020). Encoding multi-granularity structure information for joint Chinese word segmentation and POS tagging. *Pattern Recognition Letters*, 138, 163–169. https://doi.org/10.1016/j.patrec.2020.07.017

Zhou, W., Shawn N, S., Li, C., & Shao, Z. (2017). Learning low dimensional convolutional neural networks for high-resolution remote sensing image retrieval. *Remote Sensing*, 9(5), 489. https://doi.org/10.3390/rs9050489