Word segmentation of Chinese texts in the geoscience domain using the BERT model

Dongqi Wei[1], Zhihao Liu[2], Dexin Xu[3], Kai Ma[4], Liufeng Tao[5], Zhong Xie[5], qinjun qiu[5], and Shengyong Pan[6]

[1]National Engineering Research Center of Geographic Information System
[2]National Engineering Research Center of Geographic Information System, Wuhan 430074, China
[3]Wuhan Geomatics Institute, Wuhan 430074, China
[4]Unknown
[5]China University of Geosciences
[6]Wuhan Zondy Cyber Science & Technology Co., Ltd., Wuhan, China

November 26, 2022

Abstract

Unlike English, in Chinese texts there is no natural separator-like space between words, which makes Chinese word segmentation a difficult information processing problem. At present, geological texts contain a large number of unregistered geological terms, and the existing rule-based methods, machine-learning and deep-learning algorithms still cannot solve the problem of word separation in geology, especially for the large number of unregistered words. In this paper, we explore a dual-corpus, deep learning model-based approach to geological text dictionaries and compare it with the general domain dictionary and single-corpus deep learning model dictionary methods. Our experiments show that the proposed method is significantly better than other methods in open testing, with a precision of 92.56%, recall of 91.44% and F1 of 92.00%. In this paper, the Chinese word segmentation of geological text can identify unregistered geological terms effectively and ensures the recognition rate of common words, which lays the foundation for natural language processing in the domain of geoscience.

Hosted file

Word segmentation of Chinese texts in the geoscience domain using the BERT model

Dongqi Wei[1,2], Zhihao Liu[1], Dexin Xu[7], Kai Ma[5,6], Liufeng Tao[1,3,4], Zhong Xie[1,3,4], Qinjun Qiu[1,3,4,*], and Shengyong Pan[8]

1. National Engineering Research Center of Geographic Information System, Wuhan 430074, China

2. Xi'an Center of Geological Survey, CGS, Xi'an 710054, China;

3. School of Computer Science, China University of Geosciences, Wuhan 430074, China

4. Hubei Key Laboratory of Intelligent Geo-Information Processing, China University of Geosciences, Wuhan 430074, China

5. Hubei Key Laboratory of Intelligent Vision Based Monitoring for Hydroelectric Engineering, China Three Gorges University, Yichang 443002, Hubei, China6. College of Computer and Information Technology, China Three Gorges University, Yichang 443002, China

7.Wuhan Geomatics Institute, Wuhan 430074, China

8.Wuhan Zondy Cyber Science & Technology Co., Ltd., Wuhan, China

* Correspondence: qiuqinjun@cug.edu.cn

Abstract: Unlike English, in Chinese texts there is no natural separator-like space between words, which makes Chinese word segmentation a difficult information processing problem. At present, geological texts contain a large number of unregistered geological terms, and the existing rule-based methods, machine-learning and deep-learning algorithms still cannot solve the problem of word separation in geology, especially for the large number of unregistered words. In this paper, we explore a dual-corpus, deep learning model-based approach to geological text dictionaries and compare it with the general domain dictionary and single-corpus deep learning model dictionary methods. Our experiments show that the proposed method is significantly better than other methods in open testing, with a precision of 92.56%, recall of 91.44% and F1 of 92.00%. In this paper, the Chinese word segmentation of geological text can identify unregistered geological terms effectively and ensures the recognition rate of common words, which lays the foundation for natural language processing in the domain of geoscience.

Key Points:

- A BERT-BiLSTM-CRF model of Chinese geographic text word segmentation with fused language models is proposed.
- BERT captures the truly meaningful contextual information and is able to learn the relationships between successive text fragments.

- A set of experiments to verify the effectiveness of the proposed method on three available manually constructed datasets.

1. Introduction

In the field of geology, many types of geological data collections have accumulated over a long period of time due to the diversity of technical methods and research directions. In terms of data composition structure, the massive geological data repositories include a large amount of structured data and unstructured data, especially textual data and geological map data (Wu et al., 2017; Qiu et al., 2018a,2018b; Ma et al., 2021). Currently, a large number of geological reports are accumulated during geological investigations, each containing information on different geological topics, such as rocks, minerals, or hydrology, and the contents of these reports are usually stored in different formats, such as.doc.pdf.jpg.tiff, and in spatial data files (Qiu et al., 2019a,2019b; Wang et al.,2021; Salloum et al., 2022; Wang et al. al.,2022). In addition, these reports consist of a large amount of structured and unstructured data. Structured data are usually stored and managed using relational or spatial databases; however, a large amount of unstructured data, such as geological survey reports and work records, have not been fully utilized and mined. The unstructured data contain multiple data types, and the fragmented data contain richer information and have a greater potential value than the structured data (Li et al., 2021; Qiu et al., 2020; Qiu et al., 2021).

Extracting information from unstructured geological data to obtain knowledge has become a hot issue in the current research era that is exploring the big data of geological texts. Chinese word separation is an important step in Chinese geological text information extraction and knowledge discovery (Liang et al., 2019). As an important fundamental task in natural language processing (NLP), Chinese word separation has been widely used in many fields, such as information retrieval, text classification, machine translation and intelligent question and answer, and the accuracy of the word separation task directly affects the performance of subsequent tasks (Deng et al., 2016; Qiu et al., 2018a,2018b; Yuan et al., 2020). However, unlike other languages such as English, Chinese consists of a set of Chinese characters written consecutively and does not have obvious division marks between words such as spaces in English, and the division between words is more ambiguous, which makes Chinese word separation extremely difficult (Huang et al., 2015; Shu et al., 2017; Liu et al., 2019).

Among the current mainstream Chinese word separation methods, the supervised character-based tagging approach has good word separation results (Mota et al.,2018; Wei et al.,2021; Üstün et al., 2021). However, this method requires a large number of well-labeled corpora, and the word separation effect is generally better when the training corpus is tested with the same domain corpus. According to the ACL SIGHAN evaluation data, the F value of the supervised word separation approach can reach above 0.95 using the unified domain test corpus. However, extensive practical experience shows that when the trained model is switched to other domains, the accuracy of word separation is not satisfactory due to the domain and size limitation of the corpus.

At present, the main difficulties in the research of Chinese text syllabification in the geological field are; (1) the existence of a large number of unregistered geological terms. There is a large amount of information about spatial orientation, geomorphology, stratigraphic distribution, lithology, tectonics, production, geological history, analysis and evaluation in the geological report data, and the traditional Chinese word separation methods do not have the ability to learn the connection between coding and subwords independently, which leads to a large number of ambiguity problems and a low OOV (out of vocabulary) recall rate. Additionally, because of OOV issues, it is unrealistic to use traditional methods for an exhaustive enumeration of geological data due to the existence of a large number of place names and institution names in them; (2) Lack of a standardized corpus. However, it is very expensive to build a professional corpus, and it takes a long time to train the learning model, so the effect of migration to other fields is often not satisfactory; (3) There are a large number of cross-disciplinary domain words. The presence of a large number of cross-disciplinary terms in the geological report text will have a certain impact on the word separation, such as for “hierarchical analysis”, “factor analysis”, and “nonlinear revolution”; (4) Mixing Chinese and English numbers. The mixture of Chinese and English numbers and abbreviations. For example, “Cu”, “EH4”, “SE” and so on. The constructed word classification model often directly classifies them incorrectly for such cases, which will directly affect the subsequent information extraction and postprocessing; (5) Nested professional terms. There are many nested terms in geological reports, such as “Alpine”, “Alpine trough”, “siltstone”, “clay-sand silt” and “clay-sand silt”. This means that the problem of particle size is also an issue to be considered.

The BERT-bidirectional recurrent neural network-conditional random field (BERT-BiLSTM-CRF) model is proposed to address the problems of the non-accurate representation of static word vectors used in existing word separation models and poor model adaptation in specialized domain word separation. Based on the BiGRU-CRF model, the BERT pretrained language model is first introduced to enhance the generalization ability of the word vector model and capture long-range contextual information; then, the BiLSTM is introduced to fully exploit the global and local features of the text. Experiments are conducted within a self-built geographic domain subword corpus, and the model is compared with several sets of traditional models. The results show that the F1 value (the summed average of precision and recall) of the model is 92%, and the performance of the geological domain subword recognition is better than other traditional models; the BiLSTM is projected to be an effective method for geological big data mining.

The main research contributions of this paper are listed as follows:

(1) To solve the problem of multiple meanings of a word in the feature representation of geographic text, a BERT-BiLSTM-CRF model of Chinese geographic text word segmentation with fused language models is investigated, and the

model achieves an F1 value of 92% on the dataset we constructed.

(2) We conduct a comparative analysis of a series of models on the current mainstream Chinese word segmentation dataset to demonstrate the effectiveness of our presented method.

The remainder of the article is structured as follows: Section 2 discusses related work in the area of CWS in the geoscience domain. Section 3 presents the proposed approach. Section 4 introduces the details of the experiments and the results. Concluding remarks and prospects for future work are presented in Section 5.

2. Related work

Chinese word segmentation is the foundation of natural language processing research and is the basis for subsequent tasks such as lexical annotation, named entity recognition, keyword extraction, question-and-answer systems and text mining, as well as for subsequent tasks of deep semantic analysis. At present, the main research methods are divided into three major categories, dictionary-based word segmentation methods, statistical-based word segmentation model methods and deep neural network model-based word segmentation methods.

Dictionary-based approach. This approach, also called the string matching-based method or mechanical word separation method, is an earlier Chinese word segmentation method that mainly compares the input sentences with the manually constructed dictionaries to identify the words contained in the dictionaries and then conducts a slice and dice of the sentences. According to the scanning methodology, it is mainly divided into two-way scanning, forward matching and reverse matching methods. The advantage of this method is that it is more targeted and has a higher accuracy rate in the face of the words contained in the dictionary but the disadvantages are also obvious; it cannot deal well with unregistered words and ambiguity, it is less adaptable for different domains and the cost of maintaining the dictionary manually is high.

Statistical-based approach. The statistical model-based approach usually treats Chinese word segmentation as a sequence annotation problem, solves the problem of identifying unregistered words and is gradually becoming mainstream. Xue et al. (2003) first proposed labeling the corpus as a four-labeled set (B, M, E, S) and implemented Chinese word segmentation based on the maximum entropy (ME) model. (2008) implemented Chinese word segmentation based on conditional random fields (CRFs) for subcategorization. Zhao et al. (2006) provided more labeling options based on CRF to improve word segmentation. However, these statistical model-based methods require a large number of statistical features, and the word segmentation performance is heavily dependent on the goodness of fit of the manually designed features. With an increase in features, the training is easily overfit, the generalization ability is poor and the training time grows.

The more widely applied sequence annotation models include the hidden Markov model (HMM) (Roy et al., 2016), maximum entropy Markov model (MEMM) (Asahara et al., 2005) and conditional random field model. (Chen et al. (2015) proposed a bilingual semisupervised method for Chinese word segmentation, which first uses conditional random field training to obtain an alignment submodel based on the characters before segmentation and then combines the two models on the basis of another training model to recognize unlogged words for recognition.

Deep learning-based approach. Along with the SIGHAN international Chinese word separation measurement Bakeoff, the current mainstream approaches consider Chinese word separation as a typical sequence annotation problem. After Hinton et al. (2006) proposed the concept of deep learning in 2006, learning the representation of words, etc., from a large corpus has been proven to be effective in identifying unregistered words, lexical annotations (Chen et al.,2015) and dependency analyses (Chen et al.,2016). Deep learning models have been widely used in Chinese word sorting. Zheng et al. (2013) first applied deep learning models to Chinese word sorting research and proposed a perceptron-based algorithm to speed up the training of the model with a relatively small performance loss. Chen et al. (2015) proposed an LSTM network model based on improved memory units applied to the Chinese word separation task and obtained relatively good results in a general Chinese word separation corpus and subsequently used the improved gated recursive neural network model (GRNN) (Chen et al.,2015) and generative adversarial networks (GANs, generative adversarial networks) (Chen et al.,2017) to conduct research on Chinese word separation tasks.

In the field of geosciences, Huang et al. (2015) proposed a framework for Chinese word segmentation. Based on this framework, GeoSegmenter, a statistical sequence learning framework based on conditional random fields, was constructed. Specifically, GeoSegmenter first used a general-purpose word splitter to identify general terms; then, it obtained good performance in geological report word splitting by learning and applying models to identify geological terms. Chen et al. (2017) proposed a dual-corpus CRF method for geological and mineral texts by adding a large number of normalized specialized vocabulary, including geological dictionaries and geological and mineral terms, to the model, combining the geological corpus with a general domain corpus to train the word separation rules; and the experimental results showed that a good performance was obtained. Wang et al. (2018) used a conditional random field model CRF for geological reports, extracted key information from geological reports and visualized geological report information using knowledge graphs and chord diagrams.

Research on processing Chinese geoscience domain literature is complicated by the difficulty of computers in recognizing Chinese boundaries. The word separation of Chinese texts is based on dictionaries and word frequency statistics. Qiu et al. (2018a) used dictionaries and the corresponding word frequencies to generate a geological domain corpus and combined them with Bi-LSTM neural networks based on Chinese geological reports for word separation. A comparative study can improve the quality of word separation results. DGeoSegmenter achieved an average F1-score of 86.3%. Qiu et al. (2018b) proposed a self-training-based geographic domain word separation method, which implements the recognition and update operation of a geographic corpus by using an existing word separation corpus and a deep learning model combined with a self-training mechanism and achieves the word separation of a geographic corpus through multiple iterations. Li et al. (2020) proposed a deep learning sequence annotation model based on self-learning to automate the annotation of the corpus, which first uses bidirectional encoder representations from transformers (BERT) to generate word vectors with word-level features and grammatical structure features. The model first generates word vectors with word-level features and grammatical structure features using bidirectional encoder representation (BERT), then feeds the word vector sequences into a bidirectional long short-term memory neural network (BiLSTM) for bidirectional encoding, and finally annotates the word separation results by conditional random fields (CRF).

3. The proposed approach

3.1. Overall framework

The presented model consists of the BERT model, BiLSTM and CRF modules. The overall model is shown in Figure 1. First, the BERT model is used to obtain word vectors and extract important features of text; then, BiILSTM is used to deeply learn the contextual feature information for named entity recognition; finally, the CRF layer processes the output sequence of BiLSTM and combines the state transfer matrix in CRF to obtain a global optimal sequence based on the labels between neighbors.

The first layer of the model is initialized with a pretrained BERT language model to obtain word vectors in the input text information as sequence X=(x_1,x_2,...,x_n). The obtained word vectors can be used to effectively extract features from the text by using the interrelationships between words.

The second layer of the model is the bidirectional LSTM layer. The n-dimensional word vector obtained from the first layer is used as the input to each time step of the bidirectional long- and short-term memory neural net, and the hidden state sequences $\overrightarrow{h_t}$ (for forward) and $\overleftarrow{h_t}$ (for backward) of the bidirectional LSTM layer are obtained. When the forward and backward directions are all processed, the complete hidden state sequences are obtained by splicing each hidden state sequence according to the position and are denoted as h_t=(h_1,h_2,...,h_n) $R^{\mathrm{n}^*\mathrm{m}}$. Then, the linear output layer maps the complete hidden site sequences. Nexg, the linear output layer maps the complete hidden state sequence to s dimensions (s dimension is the number of label categories in the labeled set), and the extracted sentence features are all the sequences after the mapping as the matrix L=(l_1,l_2,...,l_n) $R^{\mathrm{n}^*\mathrm{s}}$,$l_i$ R^{s}. Each dimension $l_{i,j}$ corresponds to its word x_i corresponding to the score value of each category label y_i. If we directly classify the score values of each position independently at this time and select the highest score of each to obtain the output result directly, we cannot consider the information between adjacent sentences and cannot obtain the global optimum, and the classification result is not satisfactory. Therefore, the last layer of the model is introduced.

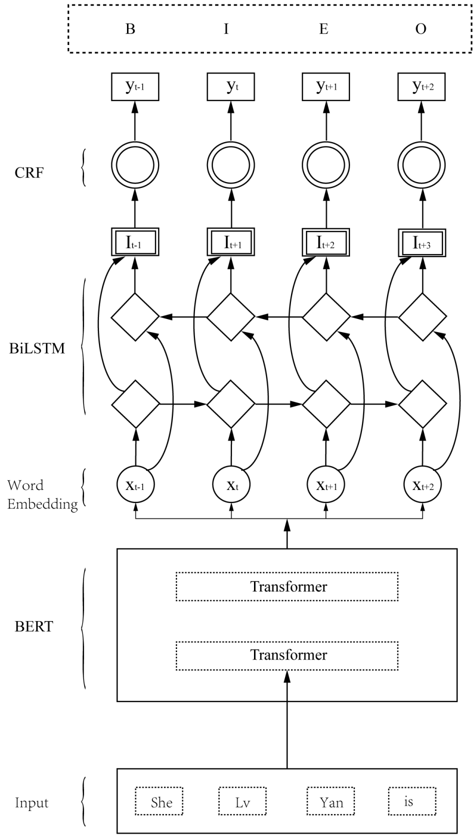

Figure 1. BERT+BiLSTM+CRF model diagram.

3.2. BERT model

BERT is a natural language processing pretrained linguistic representation model (Devlin et al., 2019; Lv et al., 2022; Ma et al., 2022). BERT is able to compute the interrelationships between words and extract important features in the text using the computed relational adjustment weights. The structure of the self-attentive mechanism is used for pretraining, based on all layers fusing the left- and right-side contexts to pretrain the deep bidirectional representations (Vaswani et al., 2017; Liu et al., 2022). Compared to previous pretraining models, it captures the truly meaningful contextual information and is able to learn the relationships between successive text fragments. The model pretraining structure diagram is shown in Figure 2.

In Figure 2, *Trm* denotes the self-attentive mechanism (transformer)-encoding converter, E_1, E_2...,E_N denotes the input of the model as a word vector, and T_1, T_2...,T_N denotes the output of the model. Since the general language model cannot understand the relationship between sentences well and the semantic relationship between sentences is very important in named entity recognition, the BERT model splices sentences *L* and *M* and predicts whether *M* lies after *L* in the original text. The pretraining of the language model can solve the problem of multiple meanings of words during the text feature extraction and then can improve the task of named entity recognition, so this paper combines the BERT language model into the task of named entity recognition and achieves significant results.

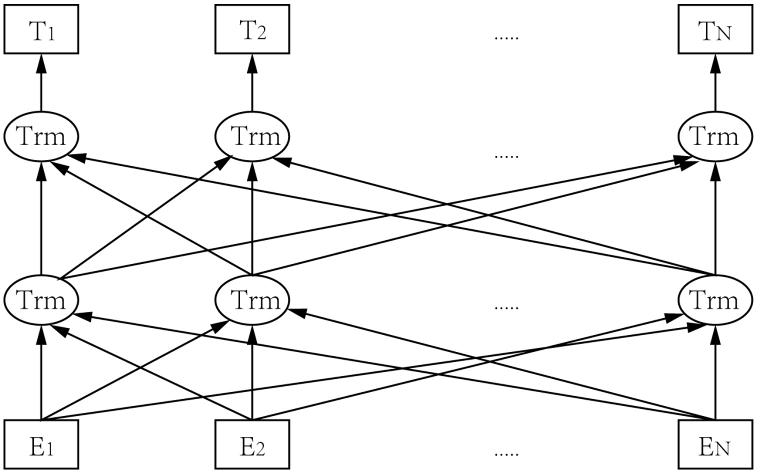

Figure 2. BERT-based pretraining model.

3.3. BiLSTM layer

The long- and short-term memory neural net was proposed in 1997 and is the most popular recurrent neural network; it is not only more sensitive to short-term inputs but also better able to preserve long-term states (Rumelhart et al. 1986; Elman 1990; Werbos 1988). The LSTM is mainly executed by three switches to control the input and output of the unit.

(1) Forget gate: the cell state c_{t-1} is retained until the current moment c_t of the decision, calculated as in Equation (1):

(1)

where W_{fh} corresponds to the input term h_{t-1}; W_{fx} corresponds to the input term X_t; W_{fh} and W_{fx} form the weight matrix W_f of the forgetting gate, b_f is the bias top, and is the activation function.

(2) Input gate: The current input X_t is saved to c_t's decision, and calculated as in Equation (2):

(2)

where W_i is the weight matrix and b_i is the bias term.

The cell state of the current input is represented by c_t, determined by the last output and the current input, as in Equation (3):

(3)

The current moment cell state c_{t-1} is given by Equation (4):

(4)

where c_{t-1} denotes the cell state of the previous gate and f_t is the forgetting gate. The symbol denotes multiplication by element.

(3) Output gate: Calculated as in Equation (5):

(5)

The input gates and unit states determine the output of the long- and short-term memory neural network as in Equation (6):

(6)

The neural network can automatically extract features based on the distributed representation of words in the text and the BiLSTM-CRF model of the word vectors. After the BiLSTM output prediction, the globally optimal labeling sequence is found by the CRF layer using the labels already predicted by the context, and the experimental comparison analysis is shown in Part IV of the text.

3.4. CRF layer

CRF is used to segment and label sequential data to predict the corresponding state sequences based on the input observation sequences, taking into account the current state features of the input and the individual labeled class transfer features, and is widely used in NER problems. CRF is applied to NER problems mainly to find the sequence that makes the objective function optimal based on the predicted output sequence of the BiLSTM model.

Two random variables X and Y, given X, if the conditional probability of each satisfying the future state is independent of the past state condition, as in Equation (7):

(7)

Then, (X, Y) is a CRF. A commonly used first-order chain structure CRF is shown in Figure 3.

CRF applied to NER is modeled by conditional probability P(yx) given the text sequence X={x_1,x_2...,x_n} to be predicted according to the output prediction sequence Y={y_1,y_2...,y_n} of the BERT-BiLSTM model; then, we have Equation (8):

(8)

where i denotes the index of the current node in x. m,n denotes the total number of feature functions on the current node i. t_n denotes the node feature function, which is only related to the current position. m denotes the local feature function, which is only related to the current position and the previous node position. $_n$ $_m$ denotes the weight coefficients corresponding to the feature functions t_n and $_m$, respectively, which are used to measure the trust of the feature function. $z(x)$ is the normalization factor, as in Equation (9):

(9)

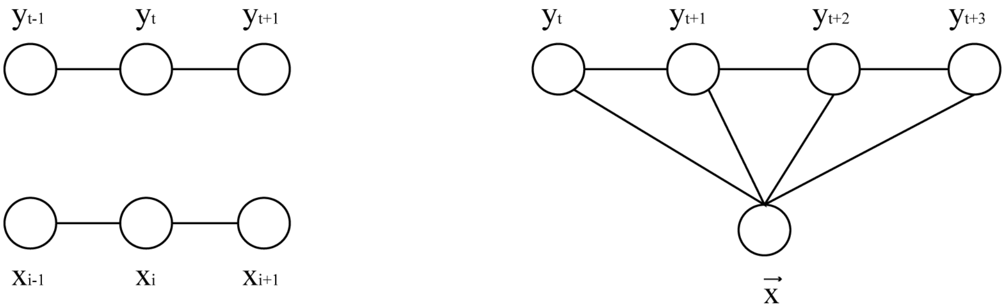

Figure 3. First-order chain structure of the conditional random field.

4. Experiments and Results

4.1. Datasets

To obtain enough data volume to get the best results of the syllogism model, the training corpus of the training set is mainly from the MSR corpus and PKU corpus; MSR is annotated by the Microsoft Asia Research Institute and PKU is annotated by Fujitsu of Peking University based on the People's Daily. Both MSR and PKU are manually annotated by the data annotators. To make the model learn the features of geographic domain participles better, we build our own geographic domain corpus (GeoCWS) to normalize the training. Table 1 summarizes the details of each corpus in the training set.

Table 1. Dataset used in this paper.

Name	**Size**	**Development data**
MSR	2,368,391 words, 4,050,469 characters	2005
PKU	1,109,947 words, 1,826,448 characters	2005
GeoCWS	887,557 words, 908,448 characters	2018

The results of the above three word separation models are evaluated by the accuracy (Precision), recall (Recall), and composite index F-measure, which are calculated by the formulas.

(10)

(11)

(12)

4.2. The experimental setting and evaluation metrics

The BERT-BiLSTM-CRF architecture proposed in this paper was used to conduct the experiments with the environment configuration shown in Table 2.

Table 2. Training Environment and Configuration of BERT-BiLSTM-CRF Model.

Operating System	Ubuntu 14.04 64bit
CPU configuration	2 * Intel(R) Xeon(R) CPU E5-2620 v2 @ 2.10 GHz
GPU configuration	2 * Nvidia Tesla K20
Memory	96GB
Python	2.7
Deep learning framework	Keras

In this paper, the Adam optimizer is selected in the process of experimentation, the learning rate is set to 0.001, the LSTM dimension in the network architecture is set to 200, the batch_size is set to 64, the max_seq_len is set to 128, and dropout is used in the experiments to prevent overfitting, which is set to 0.05.

4.3. The experimental results on various datasets

In this paper, six models, BiLSTM-CRF-Bigram, BiLSTM-CRF-Unigram, BiLSTM-CRF, BERT-CRF, BERT-Softmax and BERT-BiLSTM-CRF, were used to experiment on the PKU dataset, and three metrics, accuracy P, recall R and F1 value, were used. The performance of word separation was evaluated, and the experimental results are shown in Table 3. The experiments show that these six models achieve better performance on the PKU dataset, and their word separation accuracy P, recall R and F1 values all reach above 98%, among which the accuracy P, recall R and F1 values of the BERT-BiLSTM-CRF model reach 99.3%, 99.3% and 99.3%, respectively, which are significantly higher than the other five models, indicating that the BERT-BiLSTM-CRF model and BiLSTM-CRF model can be better for word separation of the PKU dataset.

Table 3. The performance of different models on the PKU dataset.

Model	**P**	**R**	**F1**
BiLSTM-CRF-Bigram	0.981	0.987	0.984
BiLSTM-CRF-Unigram	0.991	0.990	0.991
BiLSTM-CRF	0.992	0.991	0.992
BERT-CRF	0.992	0.992	0.992
BERT-Softmax	0.992	0.991	0.992
BERT-BiLSTM-CRF	0.993	0.993	0.993

Similarly, we conducted experiments using the MSRA dataset on six models, BiLSTM-CRF-Bigram, BiLSTM-CRF-Unigram, BiLSTM-CRF, BERT-CRF, BERT-Softmax and BERT-BiLSTM-CRF, and the experimental results are shown in Table 4. The experimental results show that the accuracy P, recall R and F1 values of these six models on the MSRA dataset all reach more than 95%, among which the BERT-BiLSTM-CRF model achieves the best scores on the MSRA dataset, and the accuracy P, recall R and F1 values can reach 97.9%, 98.1% and 98%, respectively. Compared with the other five models, the MSRA dataset achieved the best word separation results.

Table 4. The performance of different models on the MSRA dataset.

Model	**P**	**R**	**F1**
BiLSTM-CRF-Bigram	0.968	0.972	0.970
BiLSTM-CRF-Unigram	0.978	0.977	0.977
BiLSTM-CRF	0.961	0.951	0.956
BERT-CRF	0.987	0.976	0.981
BERT-Softmax	0.954	0.955	0.954
BERT-BiLSTM-CRF	0.979	0.981	0.980

In this experiment, we constructed our own disambiguation dataset GeoCWS. Similarly, we used the disambiguation dataset GeoCWS on BiLSTM-CRF-Bigram, BiLSTM-CRF-Unigram, BiLSTM-CRF, BERT-CRF, BERT-Softmax and BERT-BiLSTM- CRF on six models, and the experimental results are shown in Table 5. The experimental results show that the accuracy P, recall R and F1 values of our own constructed GeoCWS dataset can reach more than 88%, among which the BERT-BiLSTM-CRF model obtains the best results, and its accuracy P, recall R and F1 values reach 92%, 92.3% and 92.1%, respectively. Comparing the experimental results in Table 3 and Table 4, we also found that the effect of our own constructed dataset is slightly lower than that of the MSRA dataset and PKU dataset, which may be because our constructed dataset has strong domain characteristics and involves many professional names, which are more difficult to identify. Second, compared with the MSRA dataset and PKU dataset, the format of our own constructed dataset is less regular, resulting in low indices compared with other datasets. The loss curve of the BERT-BiLSTM-CRF model training is shown in Figure 4.

Table 5. The performance of different models on the GeoCWS dataset.

Model	**P**	**R**	**F1**
BiLSTM-CRF-Bigram	0.891	0.882	0.886
BiLSTM-CRF-Unigram	0.893	0.904	0.898
BiLSTM-CRF	0.911	0.902	0.906
BERT-CRF	0.919	0.904	0.911
BERT-Softmax	0.903	0.892	0.897
BERT-BiLSTM-CRF	0.92	0.923	0.921

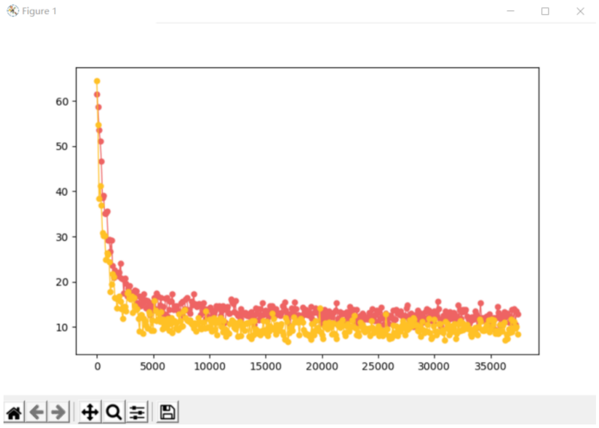

Figure 4. Loss curve of the BERT-BiLSTM-CRF model training.

4.4. Performance with different combination strategies

To effectively verify the capability of the algorithm presented in this paper, 30% of the geographic corpus is extracted as the test corpus, 70% as part of the training corpus, and the other part of the training corpus is the PKU and MSRA annotated corpus.

Three deep-learning word separation models with different corpora were used to compare and analyze the test corpus: (1) Strategy 1: Deep-learning models were constructed using the PKU corpus to separate the test corpus; (2) Strategy 2: Deep-learning word separation models were constructed using the Geo training corpus to separate the test corpus; (3) Strategy 3: Deep-learning word separation models were constructed using the MSRA corpus to separate the test corpus; (4) Strategy 4: Deep-learning models were constructed using the MSRA corpus to separate the test corpus. (4) Strategy 4: Combining the GPR corpus and the PKU corpus to build a deep-learning word separation model for the test corpus. (5) Strategy 5: Combining the GPR corpus and the MSRA corpus to build a deep-learning word separation model for the test corpus. (6) Strategy 6: Combine the geographic training corpus with the MSRA and PKU corpus and build a deep learning word separation model to classify the test corpus.

As shown in Table 6, DL-Geo, DL-PKU+GEO, DL-MSRA+GEO and DL-PKU+MSRA+GEO obtained better results than DL-PKU and DL-MSRA, among which DL-PKU+GEO obtained the best results with an F1 value of 92%. The experimental results show that the DL-PKU+GEO word separation model is more accurate than the other word separation models, and it is able to recognize the geological terminology effectively, and accurately classify the common terms in the geological text.

The reason is that DL-Geo is not accurate enough to classify common words in the geological texts. The DL-PKU+GEO word separation model trained on the geological corpus with the general domain corpus compensates for the low accuracy of DL-Geo in classifying common words. The DL-PKU+GEO and DL-Geo word-sorting models trained on the geological corpus have significantly higher accuracy, recall and F values than the DL-PKU and DL- MSRA word-sorting models trained on the general domain corpus only because the DL-PKU and DL-MSRA take into account the characteristics of geological terms in word-sorting and have a high recognition rate of geological terms, such as CRF-pku cutting "early Yanshan" into "Yanshan" and "early", while DL-PKU+GEO and DL-Geo correctly identified "early Yanshan" as "early Yanshan". DL-PKU+GEO and DL-Geo correctly identify "early Yanshan" as one word.

Table 6. The performance of the model on the different combination datasets.

Strategy	**P**	**R**	**F1**
DL-PKU	81.22	80.03	80.62
DL-Geo	91.33	90.23	90.78
DL-MSRA	80.56	79.11	79.83
DL-PKU+GEO	92.56	91.44	92.00
DL-MSRA+GEO	89.05	88.79	88.92
DL-PKU+MSRA+GEO	91.88	91.09	91.48

The main reason for the task of word separation is the recognition of geological terms, and the secondary reason is the recognition of common words in geological texts. The DL-PKU+GEO word separation model used in this experiment can effectively identify geological terms and common words in modern Chinese and accurately separate geological texts into Chinese words.

4.5. Capability of new word detection in the geoscience domain

From Table 7, we can see that unregistered word recognition is indeed a major difficulty in Chinese word separation. In particular, the unregistered word recognition rate of the MM-based model is only 5.8%, which is because MM has no self-learning ability to discover new words. From the table, it can be seen that BERT-BiLSTM-CRF has a better ability to recognize new words than MM, which indicates that the model has a better ability to discover new words through the training of the mixed corpus.

The random combination of data from general and geographic domains has an important impact on the detection of OOV in sentences. For the BERT-BiLSTM-CRF model, the OOV recall rate is significantly improved, and the new word recognition rates of GEOMSRA and GEOPKU are improved by 33.3% and 31.6%, respectively. R_{IV} demonstrates the capability of the model in word separation, and it can be seen from the table that our BERT-BiLSTM-CRF model can handle unregistered words well.

In summary, the BERT-BiLSTM-CRF network model has powerful autonomous feature learning capabilities, and our approach does not rely on any predesigned features and is effective in domain adaptation.

Table 7. Performance of different word splitters in identifying unregistered words versus words present in the corpus.

Type	Segmenter	R_{OOV}	R_{IV}
Baseline	MM	0.058	0.918
BERT-BiLSTM-CRF	GEO_{GEO}	**0.715**	**0.918**
	GEO_{GPKU}	**0.708**	**0.887**
	GEO_{GMSRA}	**0.711**	**0.906**

The deep learning models have the function of new word discovery and can identify unlisted words well. During the experiment, we found that the three models can recognize some unlisted geological terms to different degrees. Geo-gpku can recognize some unlisted geological terms, such as “Indo-Chinese”, but Geo-gpku has the disadvantage of a low recognition rate of geological terms in general domain word separation methods; for example, Geo-gpku can divide “caravan ditch” into “caravan” and “ditch”. However, Geo-gpku has the disadvantage of a low recognition rate of geological terminology by the general domain lexicography; for example, Geo-gpku cuts “caravan ditch” into three parts, “caravan”, “house” and “ditch”, and similar words, so Geo-pku cannot learn its rules. The geological text contains a large number of toponyms and nested words, such as “Tarim Basin” and “silver cave deposit”, which Geo-gpku cannot recognize well. Additionally, Geo-gpku cannot recognize geological terms such as “metal sulfide”, “mud alteration” and “gravity gradient zone”, which are composed of multiple words. The Geo-geo and Geo-gpku word separation models can recognize the unregistered geological terminology correctly. Adding a geological corpus can help to identify unlisted geological terms.

From the results of the above experiments, the Chinese word separation model based on conditional random fields can significantly improve the performance of geological texts by combining the general domain corpus and the geological corpus, which not only solves the problem of the low recognition rate of geological terms by the general domain word separation method but also has a high recognition capability for common words.

5. Conclusions and future work

In this paper, we design a BiLSTM+CRF geoscience domain word segmentation method incorporating a new language model, BERT. The BERT language model can be used to solve the problem of multiple meanings of words in text feature representations, combining the features of the BiLSTM deep learning method to fully learn the contextual information and the CRF machine learning method to extract the global optimal annotation sequence to obtain the word separation labels. The proposed model is validated in experiments, and the P value, R-value and F value are above 90%, which shows a better performance than the classical word separation model.

The recognition rate of geological terminology is significantly better than that of the general-purpose domain word separation method but the accuracy of the proposed model can be further improved. The effect of the conditional random field model is affected by the size of the training corpus, and the larger the size of the training corpus is, the better the result. Also, the larger the size of the training corpus is, the better the result of word separation. The geological corpus used is small, and there is room to improve the word segmentation accuracy.

Acknowledgments: This study was financially supported by the National Natural Science Foundation of China (42050101), the China Postdoctoral Science Foundation (No.2021M702991), Wuhan Multi-Element Urban Geological Survey Demonstration Project (WHDYS-2021-014), and the Open Research Project of The Hubei Key Laboratory of Intelligent Geo-Information Processing (No. KLIGIP-2021A01). All original data can be found in the Zenodo (https://doi.org/10.5281/zenodo.6460770).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Asahara, M., Fukuoka, K., Azuma, A., Goh, C. L., Watanabe, Y., Matsumoto, Y., & Tsuzuki, T. (2005). Combination of machine learning methods for optimum chinese word segmentation. In Proceedings of the fourth SIGHAN workshop on Chinese language processing.

2. Chen Jingwen, Chen Jianguo, Wang Chengbin, Zhu Yueqin. A study of geological and mineral text sub-word based on conditional random field[J]. China Mining magazine,2018,27(09):69-74+101.

3. Chen, W., Zhang, M., Zhang, Y. J. I. T. o. A., Speech,, & Processing, L. (2015). Distributed feature representations for dependency parsing. 23(3),

451-460.

4. Chen, W., Zhang, M., Zhang, Y., & Duan, X. J. A. I. (2016). Exploiting meta features for dependency parsing and part-of-speech tagging. 230, 173-191.

5. Chen, X., Qiu, X., Chenxi, Z., Liu, P., & Huang, X. (2015). Long Short-Term Memory Neural Networks for Chinese Word Segmentation.

6. Chen, X., Qiu, X., Zhu, C., & Huang, X. (2015). Gated Recursive Neural Network for Chinese Word Segmentation. Paper presented at the ACL (1). http://dblp.uni-trier.de/db/conf/acl/acl2015-1.html#ChenQZH15

7. Chen, X., Shi, Z., Qiu, X., & Huang, X. (2017). Adversarial Multi-Criteria Learning for Chinese Word Segmentation.

8. Collobert, R., & Weston, J. (2008, July). A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th international conference on Machine learning (pp. 160-167).

9. Deng, K., Bol, P. K., Li, K. J., & Liu, J. S. (2016). On the unsupervised analysis of domain-specific Chinese texts. Proceedings of the National Academy of Sciences, 113(22), 6154-6159.

10. Devlin J, Chang M, Lee K et al (2019) Bert: pre-training of deep bidirectional transformers for language understanding [C]. Proc of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. ACL, Stroudsburg, 4171-4186

11. Elman JL (1990) Finding structure in time. Cogn Sci 14(2):179–211

12. Hinton, G. E., & Salakhutdinov, R. J. S. (2006). Reducing the dimensionality of data with neural networks. 313(5786), 504-507.

13. Huang, L., Du, Y., & Chen, G. (2015). GeoSegmenter: A statistically learned Chinese word segmenter for the geoscience domain. Computers & geosciences, 76, 11-17.

14. Li, W., Ma, K., Qiu, Q., Wu, L., Xie, Z., Li, S., & Chen, S. (2021). Chinese Word Segmentation Based on Self-Learning Model and Geological Knowledge for the Geoscience Domain. Earth and Space Science, 8(6), e2021EA001673.

15. Liang, Y., Yang, M., Zhu, J., & Yiu, S. M. (2019). Out-domain Chinese new word detection with statistics-based character embedding. Natural Language Engineering, 25(2), 239-255.

16. Liu, H., Qiu, Q., Wu, L., Li, W., Wang, B., & Zhou, Y. (2022). Few-shot learning for name entity recognition in geological text based on GeoBERT. Earth Science Informatics, 1-13.

17. Liu, J., Wu, F., Wu, C., Huang, Y., & Xie, X. (2019). Neural Chinese word segmentation with dictionary. Neurocomputing, 338, 46-54.

18. Lv, X., Xie, Z., Xu, D., Jin, X., Ma, K., Tao, L., ... & Pan, Y. Chinese named entity recognition in the geoscience domain based on BERT. Earth and Space Science, e2021EA002166.

19. Ma, K., Tan, Y., Tian, M., Xie, X., Qiu, Q., Li, S., & Wang, X. (2022). Extraction of temporal information from social media messages using the BERT model. Earth Science Informatics, 1-12.

20. Ma, K., Tian, M., Tan, Y., Xie, X., & Qiu, Q. (2021). What is this article about? Generative summarization with the BERT model in the geosciences domain. Earth Science Informatics, 1-16.

21. Mota, P., Eskenazi, M., & Coheur, L. (2018). MUSED: A multimedia multi-document dataset for topic segmentation. Natural Language Engineering, 24(6), 921-946.

22. Qiu Q, Xie Z, Wu L, et al. DGeoSegmenter: A dictionary-based Chinese word segmenter for the geoscience domain[J]. Computers & Geosciences, 2018a: 1-11.

23. Qiu Q, Xie Z, Wu L. A cyclic self-learning Chinese word segmentation for the geoscience domain[J]. Geomatica, 2018b, 72(1): 16-26.

24. Qiu, Q., Xie, Z., Wu, L., & Li, W. (2019a). Geoscience keyphrase extraction algorithm using enhanced word embedding. Expert Systems with Applications, 125, 157-169.

25. Qiu, Q., Xie, Z., Wu, L., Tao, L., & Li, W. (2019b). BiLSTM-CRF for geological named entity recognition from the geoscience literature. Earth Science Informatics, 12(4), 565-579.

26. Qiu, Q., Xie, Z., Wu, L., & Tao, L. (2020). Automatic spatiotemporal and semantic information extraction from unstructured geoscience reports using text mining techniques. Earth Science Informatics, 13(4), 1393-1410.

27. Qiu, Q., Xie, Z., Xie, H., & Wang, B. (2021). GKEEP: An Enhanced Graph-Based Keyword Extractor With Error-Feedback Propagation for Geoscience Reports. Earth and Space Science, 8(5), e2020EA001602.

28. Roy, P. P., Bhunia, A. K., Das, A., Dey, P., & Pal, U. J. P. R. (2016). HMM-based Indic handwritten word recognition using zone segmentation. 60, 1057-1075.

29. Rumelhart D, Hinton G, Williams R (1986) Learning representations by back-propagating errors. Nature 323:533–536

30. Salloum, W., & Habash, N. (2022). Unsupervised Arabic dialect segmentation for machine translation. Natural Language Engineering, 28(2), 223-248.

31. Shu, X., Wang, J., Shen, X., & Qu, A. (2017). Word segmentation in Chinese language processing. Statistics and its Interface, 10(2), 165-173.

32. Sun, X., Zhang, Y., Matsuzaki, T., Tsuruoka, Y., Tsujii, J. J. I. P., & Management. (2013). Probabilistic Chinese word segmentation with non-local information and stochastic training. 49(3), 626-636.

33. Üstün, A., & Can, B. (2021). Incorporating word embeddings in unsupervised morphological segmentation. Natural Language Engineering, 27(5), 609-629.

34. Vaswani A, Shazeer N, Parmar N et al (2017) Attention is all you need [C]. Advances in Neural. Information Processing Systems 30. Curran Associates, New York, pp 5998–6008

35. Wang, B., Wu, L., Li, W., Qiu, Q., Xie, Z., Liu, H., & Zhou, Y. (2021). A semi-automatic approach for generating geological profiles by integrating multi-source data. Ore Geology Reviews, 134, 104190.

36. Wang, B., Ma, K., Wu, L., Qiu, Q., Xie, Z., & Tao, L. (2022). Visual analytics and information extraction of geological content for text-based mineral exploration reports. Ore Geology Reviews, 104818.

37. Wang, C., Ma, X., Chen, J., & Chen, J. (2018). Information extraction and knowledge graph construction from geoscience literature. Computers & Geosciences, 112, 112-120. doi:10.1016/j.cageo.2017.12.007

38. Werbos PJ (1988) Generalization of backpropagation with application to a recurrent gas market model. Neural Netw 1(4):339–356

39. Wei, W., Cao, X., Li, H., Shen, L., Feng, Y., & Watters, P. A. (2021). Improving speech emotion recognition based on acoustic words emotion dictionary. Natural Language Engineering, 27(6), 747-761.

40. 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.

41. Xue, N. (2003, February). Chinese word segmentation as character tagging. In International Journal of Computational Linguistics & Chinese Language Processing, Volume 8, Number 1, February 2003: Special Issue on Word Formation and Chinese Language Processing (pp. 29-48).

42. Yuan, Z., Liu, Y., Yin, Q., Li, B., Feng, X., Zhang, G., & Yu, S. (2020). Unsupervised multi-granular Chinese word segmentation and term discovery via graph partition. Journal of Biomedical Informatics, 110, 103542.

43. Zhao, H., Huang, C., Li, M., & Lu, B. L. (2006, November). Effective tag set selection in Chinese word segmentation via conditional random field modeling. In Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation (pp. 87-94).

44. Zheng, X., Chen, H., & Xu, T. (2013). Deep Learning for Chinese Word Segmentation and POS Tagging. Paper presented at the empirical methods in natural language processing.