Computers and Geosciences 121 (2018) 1–11

Contents lists available at ScienceDirect

Computers and Geosciences

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

DGeoSegmenter: A dictionary-based Chinese word segmenter for the geoscience domain

Qinjun Qiu[a,b], Zhong Xie[a,b], Liang Wu[a,b,*], Wenjia Li[a,b]

[a] School of Information Engineering, China University of Geosciences, Wuhan, 430074, China
[b] National Engineering Research Center of Geographic Information System, Wuhan, 430074, China

ARTICLE INFO

Keywords:
Chinese word segmentation
Geoscience reports
Unigram language model
Natural language processing

ABSTRACT

Larger numbers of geoscience reports create challenges and opportunities for data analysis and knowledge discovery. Segmenting texts into semantically and syntactically meaningful words is known as the Chinese word segmentation (CWS) problem because there is no space between words in the Chinese language. CWS is a crucial first step toward natural language processing (NLP). Although the available generic segmenters can process geoscience reports, their performance degrades dramatically without sufficient domain knowledge. Hence, developing effective segmenters remains a challenge and requires more work.

This inspired us to build a segmenter for the geoscience subject domain. By integrating the unigram language model and deep learning, we propose a weakly supervised model: DGeoSegmenter. DGeoSegmenter is trained with words and corresponding frequencies. We built DGeoSegmenter using the bi-directional long short-term memory (Bi-LSTM) model, which randomly extracts words and combines them into sentences. Our evaluation results using geoscience reports and benchmark datasets demonstrate the effectiveness of our method, DGeoSegmenter can segment both geoscience terms and general terms. Since manually labeled datasets and hand-crafted rules are not necessary for this proposed algorithm, it can easily be applied to various domains including information retrieval and text mining.

1. Introduction

The explosive growth of geological reports has caused their accumulation during geological survey procedures. The reports include various geological topics, such as rocks, minerals, and hydrology. In addition, large amounts of unstructured data are difficult to manage and store via virtual applications. For unstructured geological data, they contain more abundant information and have more potential value than structured data (Wu et al., 2017). In recent years, considerable research on mathematical geoscience efforts has been devoted to discovering new knowledge about georeferenced quantitative data (Cracknell et al., 2014; Lima et al., 2017; Wang et al., 2018). For many geological reports, new information can be discovered and obtained through data analysis and geological interpretation, and we can enrich our understanding based on comparing and connecting relevant work. Recent developments in the fast automatic processing of information extraction from textural geoscience data are far from sufficient. In particular, extracting information from geoscience reports in Chinese is more difficult due to CWS problem, because the Chinese language is written in continuous sequences of characters with no explicit delimiters (Huang et al., 2015).

According to several excellent reviews, the CWS methods can be broadly categorized as domain-independent or domain-specific methods according to how their texts are obtained, whether from a specific subject domain or not. Considerable research efforts have been devoted to the former, especially using deep learning techniques. It should be noted that very little optimization work has been conducted on the latter. For geoscience reports, lacking of a word segmenter adversely impacts other subsequent tasks such as information extraction.

As shown in Table 1, we used a domain-general segmenter to segment a sentence from our experimental datasets. There are no word boundaries for the original text. This is consistent with the observation from manually annotated words and English translations. For example, the four characters in *Chaganchulu* (查干楚鲁) mean *check* (查, cha), *do/work* (干, gan), *clear*(楚, chu) and *surname* (鲁, lu). In addition, *Chagan* (查干) means white, and *chulu* (楚鲁) means rock. Whereas together they form the name of a place in Inner Mongolia. If the characters in a word are treated individually rather than together, they do not represent linguistically meaningful and intended words. Nevertheless, a good word segmenter can accurately extract information from the raw

* Corresponding author. School of Information Engineering, China University of Geosciences, Wuhan, 430074, China.
E-mail addresses: qiuqinjun@cug.edu.cn (Q. Qiu), xiezhong@cug.edu.cn (Z. Xie), wuliang@cug.edu.cn (L. Wu), liwenjia@cug.edu.cn (W. Li).

https://doi.org/10.1016/j.cageo.2018.08.006
Received 27 January 2018; Received in revised form 16 August 2018; Accepted 31 August 2018
Available online 07 September 2018
0098-3004/

Table 1
Example of Chinese geoscience word segmentation.

Original text	被查干楚鲁粗粒黑云母花岗岩侵入,该套地层主要分布在格日吐防火站和敦德哈布其勒南山
English translation	It was invaded by *Chaganchulu* coarse grained biotite granite, the strata are mainly distributed in Lattice vomiting fire station and London Dehabuqile Nanshan
The correct word segmentation	被/查干楚鲁/粗粒/黑云母/花岗岩/侵入/,/该套/地层/主要/分布/在/格日吐/防火站/和/敦德哈布其勒/南山/
Segmentation made by a general domain segmenter	被查/干楚鲁/粗粒/黑云母/花岗岩/侵入/,/该套/地层/主要/分布/在/格日/吐/防火/站/和/敦德哈布/其勒/南山/

sentences and reports. Hence, correct word segmentation is a pre-processing step for Chinese geoscience reports. Without effective word segmentation, it is difficult to obtain information because significant ambiguities are present in deciphering the meaning of Chinese words.

One of the most significant challenges in the geoscience domain is limited domain knowledge (Gao et al., 2004; Yue et al., 2013; Huang et al., 2015). To address the CWS issue, the domain-generic segmenter is still limited. Because the target domains are considerably different, the capability decreases dramatically between domains (Yue et al., 2013; Liu et al., 2014; Qiu et al., 2015). As shown in Table 1, the segmentation obtained by a general domain segmenter makes some mistakes, such as "格日吐(Geritu)" and "敦德哈布其勒(Dundehabuqile)". Another interesting finding is that inconsistent words are geared toward the geoscience domain.

Motivated by the pressing needs, one method uses a comprehensive dictionary and trains the corpus for the specific segmenter. Additionally, *Dizhi Da Dictionary* (*Encyclopedia of Geology*) lists 11,000 distinct Chinese geoscience terms, and the Chinese vocabulary is an open set. However, one obvious drawback to this approach is that existing geoscience dictionaries and word libraries in China are scarce. Furthermore, even if they work well, these resources are insufficient to produce an accurate segmenter. Notably, the simple reason is that geoscience reports are also closely related to general terms.

Considering the above challenges, we designed a weakly supervised framework for domain-specific CWS, and built a geoscience segmenter based on a dictionary: the DGeoSegmenter. Based on the framework, we first categorize the words based on the word frequency, which is used to count the importance of the words. Then, we randomly select words according to the frequency of the words and combine them into sentences. Finally, the sentences are fed into models for training. In contrast to previous methods, asking humans to manually annotate word boundaries, is tedious and expensive. We obtain our corpus by applying deep learning to automatically build with word and frequency, and its advantage over the alternative solutions increases expansibility when the domain changes. We conducted some preliminary experiments to evaluate DGeoSegmenter's performance in supporting automated CWS from geological reports. Compared to all baseline segmenters, DGeoSegmenter achieved a maximum performance increase of 22.4%.

The contributions of our work are as follows. First, we address the CWS problem for the geoscience domain based on a deep learning and unigram language model, which can capture domain information and maintain useful domain information. It avoids a large selection of handcrafted features. The study also shows that some traditional algorithms can be optimized using deep learning. Second, for domain-specific CWS, a weakly supervised training framework for domain-specific CWS with a dictionary is proposed using a deep learning model and methodology, which can easily be scaled/transferred to other subject domains. Third, to our knowledge, this is the first study to segment geoscience texts from unstructured geoscience reports in Chinese via deep learning.

2. Related work

In practice, approaches for processing Chinese texts focus on word segmentation and often assume that either a comprehensive dictionary or a larger training corpus is available. Additionally, texts are manually segmented and labeled from new articles. These methods can be classified into three categories: dictionary-based, statistics-based and hybrid approaches.

Dictionary-based methods for CWS depend on large-scale lexicons, which are built based on several basic mechanical segmentation approaches with maximum matching. Without a large comprehensive dictionary, the success rate of this approach would be degraded. In dictionary-based methods, both predefined dictionaries and handcrafted heuristic rules are applied to segment input sentences (Palmer, 1997). Often, dictionary-based segmenter implementations rely on a maximum-match strategy for segmentation. This method is dictionary-based, where text is segmented by the longest word matches using a pre-specified dictionary. A sequence is regarded as a single character if no matching word exists. These methods represent domain knowledge, which are the mainstream approaches for domain-specific CWS (Zeng et al., 2011). One obvious drawback to this approach is that dictionary-based methods heavily depend on a large amount of high-quality, manually segmented data and cannot identify new words if they are not in the given corpus.

Statistics-based approaches regard word segmentation as a sequence of labeling problems (Xue, 2003). In other words, the best segmentation sequences in a word are predicted through statistics modeling such as conditional random fields (CRF) (Chang et al., 2008; Arnab et al., 2016) or support vector machines (SVM) (Tsochantaridiset al., 2005; Nawroth al., 2015). In essence, statistics-based approaches are principled and flexible. For instance, it remains challenging for dictionary-based methods to recognize new words that are not in the given dictionary, whereas statistical methods can use learning techniques to acquire knowledge from both training and test datasets to automatically improve segmentation performance. Despite being the current state-of-the-art, statistics-based approaches rely heavily on a large amount of high-quality, manually segmented data, which are difficult to handle. Furthermore, they are domain specific and involve heavily time and resource-consuming retraining when the domain is transformed. However, once the target texts are changed different from training corpora, performances of these methods drop dramatically.

As a result, statistical modeling has become the mainstream because of deep learning, whereas there is much less concern regarding domain-specific text. Recently, the neural network model has been widely used in the NLP task for replacing manual discrete features with automatic real-value features. In particular, convolution neural networks (CNNs) (Krizhevsky et al., 2012), tensor neural networks, recurrent neural networks (RNNs)(Mikolov et al., 2010; Zaremba et al., 2014), and long short-term memory (LSTM)(Graves et al., 2005; Hochreiter et al., 2012; Wang et al., 2015) have been used to extract deep level features from the input words. Yue et al. (2013) collected patent terms by a CRF-based term identifier and applied them to a generic segmenter. Huang et al. (2015) developed a CRF-based GeoSegmenter for domain specific CWS and proposed a generic two-step framework. This model can recognize geoscience terms based learning and applying it, which can transform the initial segmentation into the goal segmentation. Wang et al. (2018) used CRF model with hybrid corpus combining the generic and geology terms to CWS from geology dictionaries.

The third type of method focuses on hybrid approaches, which integrate both dictionary-based and statistics-based approaches to improve the overall performance of word segmentation. Previous research has shown that word segmentation has a substantial impact on parsing accuracy in the pipeline method. Additional data were used to improve

CWS, which resulted in a significant improvement in the parsing task using the pipeline framework.

Unlike previous studies on domain-specific CWS, the proposed model DGeoSegmenter is a weakly supervised method, which does not rely on manual annotation. Our method is based on a statistical strategy termed the "word dictionary", which only requires compiling words and frequencies. Although the dictionary is not a new, effective and scalable method, we can incorporate a dictionary with deep learning to learn and apply a model that can construct and train automatically.

3. Algorithm and computation

There are five main stages in DGeoSegmenter:

(1) *Corpus construction*: The corpus from domain-generic and domain-specific texts is collected and constructed.
(2) *Words grouped*: Each word is grouped based on frequency and a ranking algorithm.
(3) *Random extraction and combination*: Each group of words in the previous step are extracted and joined together randomly.
(4) *Training*: With the previous processing, sentences are formed via combination based on deep learning.
(5) *Testing and output*: The resulting segmentation is post-processed and output.

The main contributions of our work lie in the *Corpus construction* (Section 3.1) and *Random extraction and combination* (Section 3.3) stages. To a great extent, this study focuses on the hard case of the problem where no labels from the target domain are available, aiming at exploring an effective way of applying vocabulary to enhance segmentation.

Fig. 1 shows the DGeoSegmenter's framework, which consists of five phases. This framework exhibits the following advantages: (1) it has been designed to be modular and easily configurable and to work stably for domain-specific Chinese texts, especially with a lack of annotated target domain data; (2) it is inexpensive because it uses previously designed models; and (3) it is flexible and can be applied to other subject domains. Therefore, it avoids the basic requirements for building a comprehensive and rich training data splitter.

3.1. Corpus collection and construction

Manually annotating the geoscience corpus by domain experts is a time-consuming and error-prone process. Although some manually segmented datasets from domain generics have been constructed, once the target texts are considerably different from the training corpora or the actual vocabulary has a significant portion outside the given dictionary, the performance of CWS decreases dramatically.

Starting from this observation, this study focuses on random corpus construction for training that integrates both domain-generic and domain-specific vocabulary in the optimization process to build a corpus without any hand-crafted rules. For the domain-generic, we analyze the corpora obtained from the SIGHAN Bakeoff (http://www.sighan.org), which are the main corpora in Chinese NLP. The contents of the corpora are carefully selected, and we statistically analyzed words from the PKU and MSR corpora, including 1.07M/19.5M training words, respectively.

Unlike domain-generic text, the geoscience domain is limited by the annotated words. We collected words from the geoscience subject category of the National Geological Archives of China (NGAC), the *Encyclopedia of Geology* from the Geology Press and Wikipedia. This resulted in 615,268/467,332 words for domain-generic and domain-specific sets, respectively. Fig. 2 displays some randomly selected training sets from the reports.

3.2. Words grouped

The task of word grouping is to divide the word into groups that are ready for further analysis using a splitting strategy. A natural choice ranks the words from the corpus by their frequencies. The following preprocessing steps are conducted in this research: first, the function of rounding down converts the frequencies into a format, which greatly speeds up the algorithm with little impact on the quality of the final results. Then, the words are grouped by the predefined function, resulting in a significance score assigned to the words. The utilization of the score, which represents the contributing factor of the probability, calculates the weight of each word with an indicator function.

For each word w in group $G=\{g_1, g_2,.., g_N\}$, let $F=(f_1,f_2,..,f_N)$ be the average frequency for each group, let θ_i be the significance extraction score for word w_i as follows:

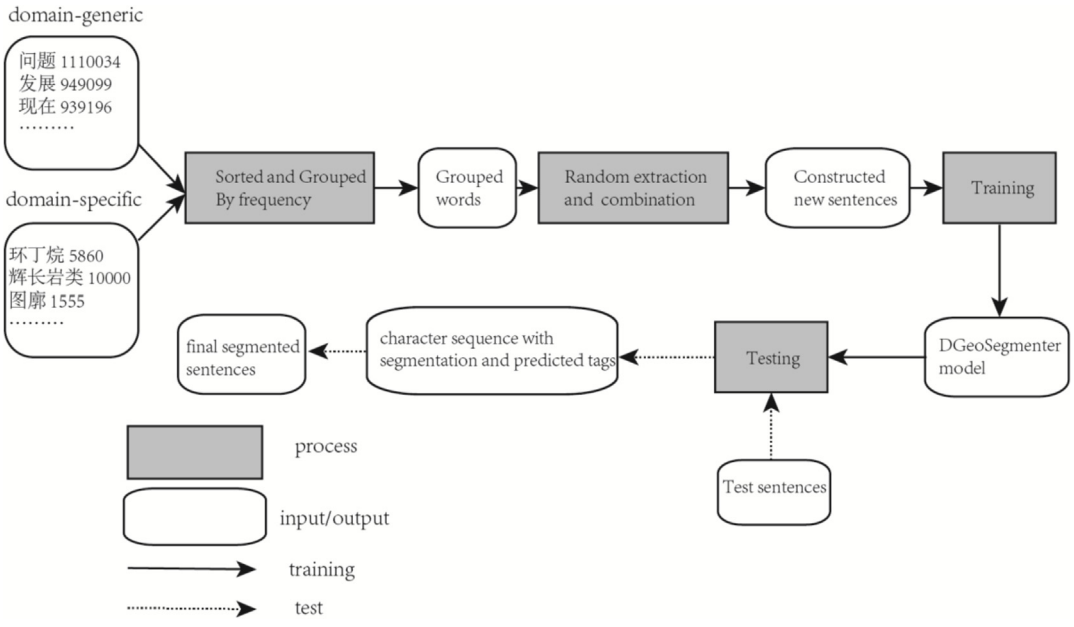


Fig. 1. The DGeoSegmenter framework: collections of words are shown in rectangles, the processes are shown in gray rectangles, and the input/output are shown in white ovals.

domain-generic		domain-specific	
问题 1110034	发展 949099	张紧装置 10000	辉长岩类 10000
现在 939196	内容 622522	环脂化合物 8903	排出线路 8874
文化 621600	力 620122	辉长岩状的 7865	环丁烷 5860
其他 619623	专业 619215	环化聚合作用 5552	地图编绘 4781
方法 618145	分享 403800	淡辉长细晶岩 4478	环烷金属化合物 3890
只是 403163	各种 402161	平面图褶积 3569	泵的反作用面积 2589
.....		抽油井动力装置 1596	图廓 1555
		.....	

Fig. 2. Fragments of the training set.

$$\theta_i = f_i / \sum_{j=1}^{N} f_j \tag{1}$$

A large θ_i means that word w_i is statistically important for the DGeoSegmenter to fit the target texts. Note that the lower frequency words can also be selected to construct the new sentences with many iteration processes based on deep learning, although one may feel that a more frequently occurring word seems to have a better chance.

3.3. Random extraction and combination

To guide the corpus collection process, our attempt here is to present a novel and effective approach based on a unigram language model automatically to enrich the corpus in the hopes of enhancing word segmentation performance. Words and corresponding frequencies in the word set, together with random extraction and a combination strategy, are integrated into an optimization function to optimize the probability scores of words. These optimized words generate sentences accompanied by labels, aiming to serve as a training corpus.

The underlying theory in our word segmentation system is the unigram language model in which words in a sentence are assumed to occur independently (Peng et al., 2016; Tripathy et al., 2016). A statistical language model is a probability $p(w_1,w_2,\ldots,w_n)$ distribution over sequences of words. Given such a sequence $w_1,w_2,\ldots,w_n$, say of length n, it assigns a probability to the whole sequence. The unigram model splits the probabilities of different terms in contents as follows:

$$p_{uni}(w_1, w_2, \ldots w_n) = p(w_1)p(w_2)\ldots p(w_n) = \prod_{i=1}^{N} p(w_i) \tag{2}$$

In probability terms, $p(w_n)$ represents the probability of the word w_n. A sentence is a sequence of basic characters of a language, but is read and understood through higher-order units, i.e., words, phrases, idioms, and regular expressions, which in our context are all broadly defined as "words." Let $A = \{a_1,a_2,\ldots,a_p\}$ be the basic set of "characters" of the language of interest. Specifically, in Chinese, it is the set of all distinct characters appearing in the text, including tens of thousands of words. A word w is defined as a sequence of elements in A, i.e. $w = a_{i1}, a_{i2},\ldots,a_{il}$. Let $D = \{w_1,w_2,\ldots,w_n\}$ be the vocabulary for the texts. DGeoSegmenter builds each sentence S as a concatenation of words drawn randomly from D with an extraction probability(score) θ_i for word w_i. $\theta=(\theta_1, \theta_2,\ldots, \theta_N)$ represents the word extraction probability, in other words, the higher the probability, the more likely the word is selected, and vice-versa. The probability of generating an N-word (segmented) sentence $S = w_{i1},w_{i2},\ldots,w_{in}$ is as follows:

$$P(S|D, \theta) = \prod_{n=1}^{N} \theta_{i_n} \tag{3}$$

This model for handling CWS can be traced back to Deng et al. (2016). Theoretically, we can generate a random sequence of words by sampling from the vocabulary with probability $p(w_1)p(w_2)\ldots p(w_n)$. Importantly, however, is it possible to produce new sentences (corpus) for further training models in an unsupervised manner? A possibility for this certainly lies in how well different frequencies of words can be put together to generate sentences via repeated iterations. Since deep learning algorithms are known to have particularly strong independent learning capabilities. The DGeoSegmenter algorithm builds words and obtains the sentences by capturing the highest probability $p(w_1)p(w_2)\ldots p(w_n)$ with a dynamic programming scheme. That is, a sentence is built by maximizing the probability of those important features constructed based on each candidate word. Formally, word selection can be obtained by maximizing the probability function of $p(w_1)p(w_2)\ldots p(w_n)$ as follows:

$$S' = \arg\max P(S|D, \theta) = \arg\max p(w_1)p(w_2)\ldots p(w_n) \tag{4}$$

Compared with the other complexities and subtleties of natural language models, our model is clearly a rough approximation. Though neglecting the long-distance dependencies among words, this research shows that using random extraction and then a combination strategy is computationally efficient and works well in practice for two reasons. First, segmentation in Chinese language processing does not necessarily depend complement on context-dependent factors to determine the boundary position of words. The important task is finding the rules and patterns of inner-words. In this regard, it is only the (grammar) correctness of words that is needed to address the problem of CWS, even if the sentence is not grammatically correct. Second, deep learning, as a learner, is used to simulate this extracting procedure and automatically learn the patterns of the texts, and aims at heightening the learner's ability to identify OOV words at the cost of sacrificing as few in-vocabulary words as possible.

Fig. 3 presents an example of word extraction and sentence generation.

Random extraction and combination algorithms are illustrated in

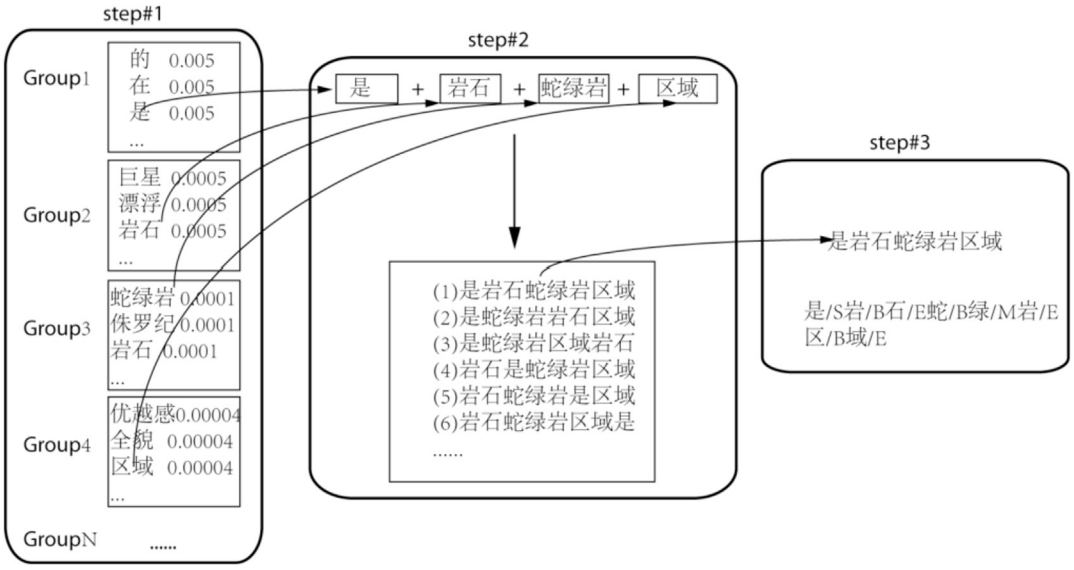


Fig. 3. An example of word extraction and sentence generation. Step #1 groups the words. Step #2 randomly combines the selected words into sentences. Step #3 tags the newly formed sentences.

Algorithm 1. Each iteration aims to produce a suitable sentence, and the newly selected words are added to the sentence until the convergence condition is satisfied. When the number of iterations reaches the condition, the algorithm stops.

Algorithm1 Random extraction and combination algorithm

Input: a set of words and homologous frequency, maxlen M
Output: a set of sentences(sentences)
corpus $C \leftarrow$ construct_corpus(words,freuencies)
groups G, extraction probability $\theta \leftarrow$ group_words(C)
do
  repeat
    sentences $S \leftarrow$ extract_random(G, θ)
    $M' \leftarrow$ calculate_length(S)
  until $M' > M$ **return** output_combination (sentences)
until the maximum number of iterations N reached

3.4. Domain-specific term segmentation as sequence labelling

We investigated the use of deep neural networks (DNNs) because they showed better performance in wide areas for NLP. RNNs are an extension of a conventional feed-forward neural network. The sequence of variable length can be handled by these models based on a recurrent hidden unit, whose activation at each time step is dependent on that of the previous one. However, these models fail due to gradient vanishing/exploding problems (Bengio et al., 2002).

The LSTM model can handle these gradient vanishing problems. However, one main shortcoming of the LSTM is that it can only make use of the previous context, knowing nothing about the future. An efficient approach to address this problem is by using Bi-LSTM (Mike et al., 1997). The model can be used to process each sequence forwards and backwards, which results in into two separate hidden states to capture past and future information.

The geoscience word segmentation can be regarded as a sequence labelling problem. After obtaining the corpus, we chose the max-margin criterion (Taskar et al., 2005) to train our model. As illustrated in (Kummerfeld et al., 2015), the max-margin criterion applies both the likelihood and perception method. The parameter set of the DGeoSegmenter model is θ.

Given a character sequence $x^{(i)}$ from the hybrid corpus, with the correctly segmented sentence for $x^{(i)}$ denoted as $y^{(i)}$, y is the label that DGeoSegmenter aims to predict. let $\Delta(y^{(i)},y)$ be a structured margin loss as follows:

$$\Delta(y^{(i)}, y) = \sum_{i=1}^{m} \mu 1\{y^{(i),t} \neq y^{t}\} \tag{5}$$

where m is the length of a given character sequence $x^{(i)}$, and μ denotes the discount parameter. The proportion of loss is the number of incorrectly segmented characters.

For a given training set Ω, the loss function $J(\theta)$ is the regularized objective function, which includes an l_2 norm term:

$$J(\theta) = \frac{1}{\Omega} \sum_{(x^{(i)},y^{(i)})\in\Omega} l_i(\theta) + \frac{\lambda}{2}\|\theta\| \tag{6}$$

$$l_i(\theta) = \max(s(y, \theta) + \Delta(y^{(i)}, y) - s(y^{(i)}, \theta)) \tag{7}$$

where $s(\cdot)$ is the sentence score. With the hinge loss, which is differentiable for the objective function, we use a subgradient method (Ratliff et al., 2007) for computing a gradient-like direction. Following (Socher et al., 2013), the diagonal variant of AdaGrad (Duchi et al., 2011) with minibatches was used to minimize the objective.

The update for the i-th parameter at time step t is as follows:

$$\theta_{t,i} = \theta_{t-1,i} - \frac{\alpha}{\sqrt{\sum_{\tau=1}^{t} g_{\tau,i}^2}} g_{t,i} \tag{8}$$

where α denotes the initial learning rate, and $.g\tau \in \mathrm{R}^{|\theta i|}$ is the

Table 2
Details of the test datasets.

Corpus	categories	#sent.	#testSize(words).	OOV Rate
domain-generic	MSR	181K	12.2M	2.60%
	PKU	708K	5.63M	4.30%
domain-geoscience	RGR	18K	1.9 M	75.2%
	EGR	15K	1.1 M	80.2%
	MGR	12K	0.9 M	73.2%
	HGR	14K	1.3 M	74.1%
	GGR	17K	1.2 M	82.6%
	RSGR	20K	1.4 M	79.7%

subgradient at time step τ for parameter θ_i.

3.5. Testing and output

With the new sentences, we employ the Bi-LSTM approach to model the correspondence between the input character sequence and the output label sequence. Specifically, each character in the sequence is tagged as one of {*B*, *E, M, S*}, representing *Begin*, *Middle*, and *End*, and *S* represents a single character segmentation (Xue, 2003). Take the following example:

A sentence $c = c_1 \ldots c_T$ is defined as a sequence of elements. Let $s = s_1 \ldots s_T$ be the segmentation locators of c. For instance, consider the 14-character sentence c = 曹家埠金矿赋存于新城焦家断裂 "Caojiapu gold deposit in the new city of Jiaojia fracture" where $T = 14$, c_1 = 曹 (cao), c_2 = 家 (jia),..., c_{14} = 裂 (lie), and $s = BMEBEBESBEBEBE$. Therefore, the linguistically meaningful segmentation is/曹家埠/金矿/赋存/于/新城/焦家/断裂/. The 4-tag segmentation rule is effective in achieving segmentation accuracy and computation efficiency, which are two important and desirable properties in NLP (Shu et al., 2017).

4. Experiment and discussion

4.1. Setup

Datasets:We collected 43 geoscience reports as the test corpus. The 43 reports are considered to be representative because they (1) record geological information from different years by different writers, (2) come from different regions, and (3) present domain-specific complexities with varying text from simple to complex. A manual annotation process was followed to build the gold standard segmentation sequence for testing purpose. Based on the same criteria used to create the benchmark corpus from the SIGHAN CWS Bakeoff, we invited some annotators who satisfied the background knowledge of both geoscience and CWS to avoid any annotation bias. This dataset was represented as the GEO corpus. This resulted in 7.8M segments. The benchmark corpus from the SIGHAN CWS Bakeoff consisted of new documents with a wide range of topics. We used simplified Chinese versions of the test corpora created by MSR and PKU. The resulting MSR and PKU corpora have 12.2M/5.63M segments, respectively. Table 2 illustrates the details of the test datasets. Out-of-vocabulary (OOV) words denote the words that were not in the training set but emerged in the test set. The content in the test set was broadly represented, including regional geological reports (RGR), environmental geological reports (EGR), mineral geological reports (MGR), hydrogeology geological reports (HGR), geophysical geological reports (GGR), and remote sensing geological reports (RSGR).

Performance:The five standard metric criteria were used to compute the segmenter performance. ***Precision*** denotes the percentage of all predicted words whose true words were labeled by a human labeler. ***Recall*** denotes the percentage of all true words that were correctly predicted. An evenly weighted ***F1-score*** indicates the overall performance as follows: $F1 = 2{*}P{*}R/(P+R)$. Recall of OOV (R_{OOV}) and recall of IV (R_{IV}) are the percentages of the OOV and in-vocabulary (IV) words, respectively, that are correctly segmented. Given a learned model, R_{OOV} indicates how well it can be generalized to a new domain, whereas R_{IV} suggests its predictive power over the training data. The range of the indicators, ***Precision***, ***Recall*** and ***F1-score***, are within [0, 1]; the greater the value is, the better the performance that is indicated. The 10-fold cross-validation is used to test, and we reported the average score of 10 independent runs.

4.2. Overall performance of closed and cross-domain segmenters

We conducted experiments for three baseline generic segmenters. We used the PKU and MSR corpora in our experiment to investigate the proposed model. Depending on whether resources other than geoscience words were used, segmenters were classified as a cross-domain segmenter or a closed-domain segmenter. This resulted in four GenSegmenter versions. Let DGeoSegmenter be the hybrid segmenter incorporating all words. All of the experiments were trained by the Bi-LSTM model with the same parameter setting.

Table 3 shows the performance of our model as well as previous state-of-the-art systems. From the table we can see that these models achieved better results for the generic domain; however, once the target domains are considerably different from the training corpora, such as those geoscience documents accumulated throughout ancient China that contain many unregistered technical words, the performance of these models trained with domain-generic datasets decreased dramatically. This could be caused by lacking of sufficient annotated data. The key is to increase the comprehensive and representativeness of labeled data, so that these algorithms can learn how to address different segmentation cases. However, as discussed, constructing such a labeled dataset is rather challenging and time-consuming.

Closed-domain segmenter: The maximum matching method (denoted as *MM*) based on the dictionary was the first closed-domain baseline. This method required a given corpus as a reference and was a representative method. We used all the training data in our experiment to investigate the *MM* method. The text was segmented based on the longest match strategy. When a new word for which there was no matching word appeared, the approach segmented it as a single character.

The Bi-LSTM-based segmenter is the second closed-domain baseline that was built directly from generic and geoscience data, denoted as GES_{MSR}, GES_{PKU}, and GES_{GEO}. The hyper-parameters of the neural model significantly impacted the performance of the algorithm. The character embedding size was chosen as a trade-off between speed and performance. The number of hidden units was set to 128, as was the character embedding size. We set the maximum word length to 4 in our experiments. We dropped the input layer of the model with a dropout rate of 20%, which is a popular technique for improving the performance of neural networks by reducing overfitting. All of these models were trained in the following environment: CPU:2 * Intel(R) Xeon(R) E5-2620 v2 @ 2.10 GHz, GPU: 2 * NVidia, Tesla K20, Memory: 96 GB. The operating system used was Ubuntu 14.04 64 bit.

Cross-domain segmenter:We initiated the four cross-domain baseline segmenter. Using the same criteria as for building *MM,* the first two were Bi-LSTM based, built from the MSR and the PKU corpora, defined as GEO_{GMSR} and GEO_{GPKU}. The third segmenter was developed using all the training data, including the MSR, PKU and GEO corpora.

A comparison of the results in Table 4 shows the effectiveness of the proposed domain adaptation method. Although the segmenter MM did not use the learning technique, it performed well. The segmentations were more accurate than those of any of the closed-domain segmenters, which did not use deep learning techniques and skills.

Nevertheless, the dictionary-based method is not necessarily better than the deep learning method. A comparative study of Table 4 indicates that the deep learning method exhibits a significantly higher capability to segment. Compared with *MM*, DGeoSegmenters outperformed their corresponding baselines by at most 22.4% in the *F1-*

Table 3
Performances on PKU, MSR and GEO test sets with different models. PKU/GEO indicate that the training and testing sets are PKU and GEO, respectively.

models	PKU/PKU			MSR/MSR			PKU/GEO			MSR/GEO		
	P	R	F	P	R	F	P	R	F	P	R	F
Zheng et al., 2013	92.8	92.0	92.4	92.9	93.6	93.3	70.5	70.1	70.3	70.8	71.5	71.2
Pei at al., 2014	93.7	93.4	93.5	94.6	94.2	94.4	70.9	71.0	70.1	71.2	72.1	71.7
Huang et al., 2015	91.8	92.8	92.3	92.5	91.4	91.9	69.5	68.9	69.2	70.8	70.9	70.8
Zhang et al., 2017	96.4	96.8	96.6	97.5	97.7	97.6	73.8	73.6	73.7	74.1	73.5	73.8
Huang et al., 2017	95.1	95.6	95.3	96.8	96.4	96.6	71.1	71.0	71.0	71.8	70.6	71.2
Bi-LSTM	96.0	95.7	95.9	96.3	96.1	95.4	72.6	73.0	72.8	73.5	73.6	73.5

score. The segmenters built based on the same corpus, MM and $GES_{GEO,}$ outperformed those built using different corpora such as GES_{MSR} and $GES_{PKU.}$ Note that GES_{GEO} achieved a lower *F1-score* on its testing data. This is not surprising because it was built using geoscience words, which only utilize geoscience information and lack the generic domain, and it was therefore difficult to segment words with a generic domain.

The cross-domain segmenters GEO_{GMSR}, GEO_{GPKU} and GEO_{GEO} are useful. With their stronger predictive power, the DGeoSegmenter performance was increased by 20% compared to the baseline methods. Furthermore, due to the use of neural network models, which have a strong learning power over the training data, DGeoSegmenter was able to learn the best adaptation model for each generic segmenter.

These results and findings suggest that general terms and geoscience terms are recognized separately and compensate for each other. Identical conclusions were obtained in previous studies, showing that the proposed DGeoSegmenter approach based on Bi-LSTM is an effective approach for solving the domain-specific word segmentation problem. Meanwhile, both generic terms and geoscience terms were recognized.

Table 5 presents the performance of DGeoSegmenter and four other advanced CWS methods: GCRF (Huang et al., 2015), TopWords (Deng at al.,2016), HyperLSTM (Zhang et al., 2017) and GRS (Huang et al., 2017), in terms of the precision, recall and F1-score. In the tables, the best (the highest) results were obtained by the HyperLSTM method, and the second best results were obtained by a particular DGeoSegmenter. However, this does not necessarily mean that DGeoSegmenter is inferior to the HyperLSTM method. This is not surprising, because the HyperLSTM approach is built with not only labeled data but also extensive dictionaries. In contrast, our proposed method relies solely on dictionaries and is a weakly supervised approach. These results and findings suggest that the proposed weakly supervised method is indeed an effective in addressing the domain-specific CWS problem.

4.3. The performance of new word detection

As shown in Table 6, it is difficult to identify new words. In particular, the OOV rate of the *MM* method was only 5.8%. This is not a surprising result since *MM* does not have the self-learning ability to detect new words. We achieved significant improvement in R_{OOV} when we applied GenSegmenter, demonstrating that the use of the Bi-LSTM model can efficiently obtain segmenting information. GES_{GEO} achieved the best score, of 65.3%. Based on our findings, it can be concluded that GenSegmenter has stronger predictive capabilities.

Table 5
Segmentation performance of different methods.

Model	P	R	F
GCRF	0.865	0.854	0.859
TopWords	0.845	0.84	0.842
HyperLSTM	0.889	0.872	0.88
GRS	0.758	0.765	0.761
DGeoSegmenter	0.861	0.871	0.866

Table 6
Performance of different segmenters on recognizing OOV and IV terms.

Type	Segmenter	R_{OOV}	R_{IV}
Baseline	MM	0.058	0.918
GenSegmenter	GES_{GEO}	0.653	0.816
	GES_{MSR}	0.405	0.776
	GES_{PKU}	0.417	0.853
DGeoSegmenter	GEO_{GEO}	**0.715**	**0.918**
	GEO_{GPKU}	**0.708**	**0.887**
	GEO_{GMSR}	**0.711**	**0.906**

DGeoSegmenter performed better than GenSegmenter. This result demonstrates that DGeoSegmenter is more powerful than GenSegmenter and that the hybrid data produce effective results.

The generic and geoscience domain data randomly combined to have a key influence on the detection of OOV in a sentence. When we trained with DGeoSegmenter, the recall rates of the OOV terms increased dramatically, resulting in an increase of 33.3% and 31.6% in R_{OOV} for GES_{MSR} and GES_{PKU}, respectively. The R_{IV} score demonstrates that our model performed well on word recognition. Again, as demonstrated in the R_{OOV} score, our model has a certain ability to address OOV words.

Our approach does not depend on any predesigned features due to the strong ability of the Bi-LSTM network in automatic feature learning. This can be attributed to the effective capability of DGeoSegmenter in domain adaptation. The performance of OOV and IV increases significantly when using the LSTM unit.

Table 4
Segmentation performance of different segmenters.

Type	Segmenter	P		R		F	
Baseline	MM	0.796		0.805		0.800	
GenSegmenter	GES_{GEO}	0.650		0.663		0.656	
	GES_{MSR}	0.620		0.646		0.633	
	GES_{PKU}	0.631		0.649		0.640	
	GES_{CRF}	0.715		0.711		0.713	
DGeoSegmenter	GEO_{GEO}	**0.861**	↑21.1%	**0.871**	↑20.8%	**0.866**	↑21%
	GEO_{GPKU}	0.844	**↑22.4%**	0.849	↑20.3%	0.846	**↑21.3%**
	GEO_{GMSR}	0.847	↑21.6%	0.853	↑20.4%	0.850	↑21%

Table 7
Results of the models on the test sets of three CWS datasets. **CNN**: consists of an input and an output layer, as well as multiple hidden layers. The hidden layers of the CNN consisted of convolutional layers, pooling layers, fully connected layers and normalization layers in our experiment. **RNN**: a class of artificial neural networks in which connections between units form a directed cycle. For the CWS task, an RNN typically has three LSTM layers. **LSTM**: a simple LSTM network with three LSTM layers. **Bi-LSTM**: Bi-LSTM means a bidirectional LSTM network with three LSTM layers.

Models	P	R	F
CNN	81.2	82.5	81.8
RNN	77.6	78.2	77.9
LSTM	81.3	82.0	81.7
Bi-LSTM	85.1	84.2	84.7

4.4. Different baseline neural networks

We conducted experiments with different baseline neural networks. We used the following generic models for comparison. Complete details of this case will be provided at a later time.

We compared these methods with the two top performers in the GEO corpus. As shown in Table 7, the experimental results of the models on the test sets of three CWS datasets with different neural networks illustrate the effectiveness of the proposed domain adaptation method. DGeoSegmenter outperformed the corresponding baseline by at most 22.1% in terms of the *F1-score*.

In particular, the proposed model with CNN and LSTM attains 81.8% and 81.7% in the F-measure, respectively, which delivers a comparable performance against the RNN performers. Specifically, the Bi-LSTM model achieves 85.1% in precision and delivers the highest recall with 84.2% and the highest F-measure with 84.7%. This result shows that the model with Bi-LSTM performs the best against the other methods at the expense of computation time. This occurs because Bi-LSTM can capture information from the sequence dataset and maintain contextual features from the past and future. In particular, the Bi-LSTM network integrates the forward and backward passes of each layer and can memorize the information of all sentences in two directions based on the propagation of two parallel layers.

4.5. Varying the training dataset

Considering that the amount of training data may affect the performance of the trained models, we conducted a set of experiments to demonstrate that the amount of training data can affect the quality of the learning models. First, we divided the hybrid dataset into disjoint subjects and randomly extracted training data. Then, the training data were increased from 10% to 100% in increments of 10%. We performed ten independent runs for each and computed the average performance results. The experimental results, in terms of *F1-score*, are shown in Fig. 4.

It can be seen from Fig. 4 that the training data have a key influence on both the cross-domain and closed-domain. This is not surprising since our approach was trained using only words.

Both the GenSegmenter and DGeoSegmenter versions increasingly benefited from more training data. We observed that using 10% of the corpus for training resulted in a low F-score. The reason for this result may be that the segmenter trained using deep learning requires not only a large amount of generic data to learn but also geoscience data, in particular, GES_{GEO}, GES_{MSR} and GES_{PKU} in the closed-domain at the same level of performance.

Domain-specific vocabulary is irreplaceable. For the DGeoSegmenter built using hybrid words, although the generic segmenter made many mistakes, the randomly collected words enabled learning of how to correct these mistakes.

DGeoSegmenter learned well by incorporating a random word collection strategy, which can be useful as training data. In other words, although the baseline segmenter made more mistakes on the geoscience documents, the proposed strategy learned how to correct these mistakes. On the other hand, DGeoSegmenter can obtain the ability of an adaptation procedure. This is because according to adaptation, what is important is the ability of the baseline segmenter to correct the mistakes that were made in the initial segmentation. This demonstrates the effectiveness of our deep learning-based technique for domain-specific CWS.

4.6. Diversifying the frequency for geoscience words

As mentioned in Sections 3.1 and 3.2, the frequency can directly influence the random extraction and model complexity, so that an optimal frequency can be determined and applied for improved performance. A total of 6 controlled experiments were conducted, with the same frequency ranging from 100 to 10000 for the domain-specific words. The experimental results, in terms of *F1-score*, are illustrated in Table 8.

The experimental results suggest that the frequency is important in determining the overall performance. As shown in Table 8, increasing the frequency leads to the improvement of the *F1-score*. For example, when the frequency was set from 100 to 500, the average *F1-score* improved by 17% and 12.8%, respectively. Although increasing frequency can improve performance, the optimal performance cannot be achieved if the frequency is set with the same value. When the frequency was collected from the statistical strategy, the average *F1-score* achieved 86.6% and 73.5%, respectively. In other words, DGeoSegmenter was greatly impacted by the frequency. This is because, the comprehensive and representativeness of the frequency is the key factor, so that the algorithm can learn how to address extraction cases.

This again demonstrates that the proposed approaches can learn from words and frequency because they can make better use of datasets.

4.7. Ratio of domain-generic and domain-specific texts

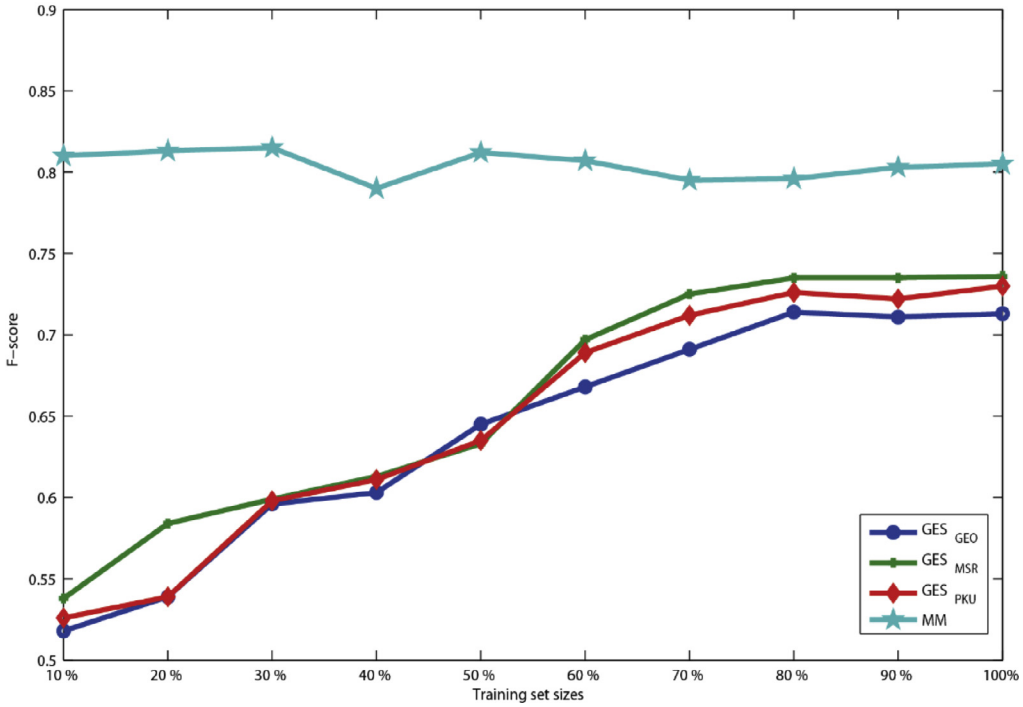
To investigate the ratio of domain-generic to domain-specific texts, a set of experiments was conducted to study the impact on the corpus. A total of 18 controlled experiments were conducted with a ratio ranging from 10% to 90% (with a step size of 10%). The experimental results, in terms of average precision, recall and *F1-score*, are shown in Table 9.

As shown in Table 9, increasing the ratio of domain-generic or domain-specific texts leads to the improvement of average precision, recall and *F1-score*. For example, when the ratio of domain-generic and domain-specific texts was set to 100/90, the DGeoSegmenter achieved an average *F1-score* of 86.3%. More substantially, increasing the ratio of domain-specific outperformed the ratio of domain-generic texts. Based on the experimental results, the proposed algorithm benefits from increasing the ratio of domain-specific texts.

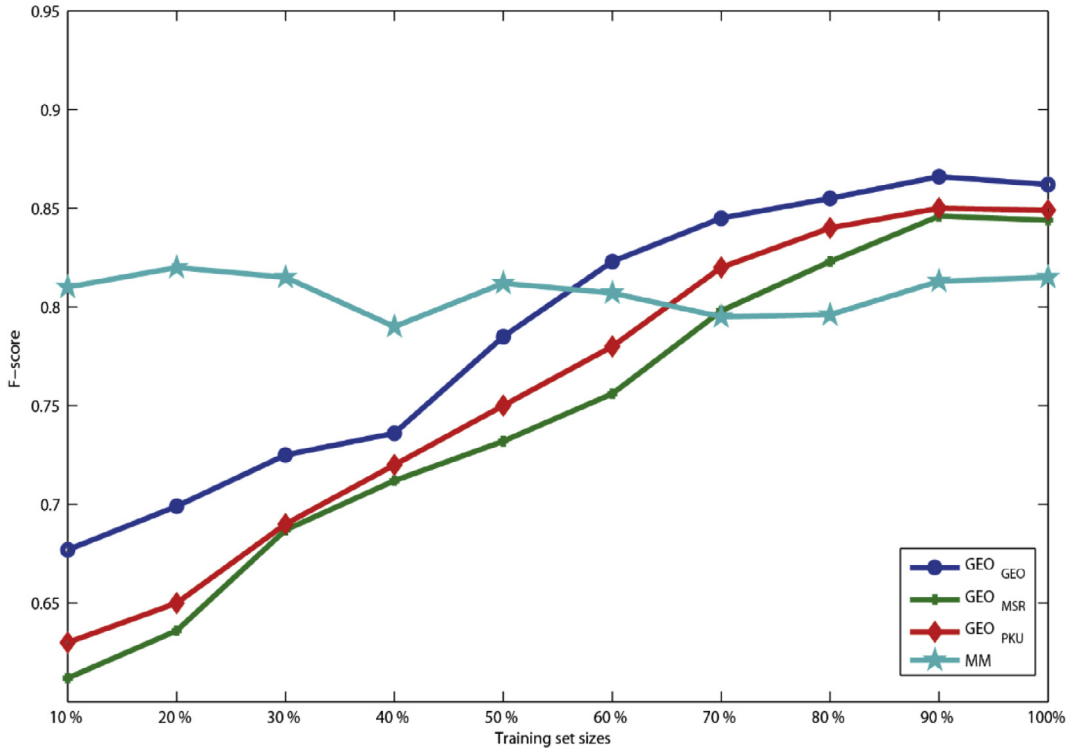
4.8. Different categories of reports

We further investigated the benefits of the proposed models by comparing different categories of reports, as shown in Table 10. Based on the results, we were able to observe that DGeoSegmenter makes progress gradually on different categories of reports.

DGeoSegmenter achieved an average *F1-score* of 85.55%. This is because it uses learning methods. DGeoSegmenter has strong learning and predictive power. As a result, compared to the baseline approaches, DGeoSegmenter for environmental geological reports outperformed other reports, achieving an *F1-score* of 87.11%. This is not surprising because most of the environmental geological reports were built using general words, with less specific vocabularies. In contrast, remote sensing geological reports included more geoscience terms, making it difficult to capture every word.



(a)closed-domain



(b)Cross-domain

Fig. 4. Performance of the (a) closed-domain and (b) cross-domain segmenters with varying training dataset sizes.

5. Conclusion

Domain-specific CWS is a difficult task due to the lack of sufficient annotated reports and the costliness of manually generated segmentation rules. In this research, we employed deep learning and unigram language model to CWS from geoscience reports. We proposed a weakly supervised training framework for domain-specific CWS with only dictionary-based deep learning. The proposed framework builds upon a theoretical unigram language model, and thus is generalizable. We leveraged multilevel features using the following steps: First, we

Table 8
F1-score performance at different frequencies.

	Cross-domain	Closed-domain
100	70.8	62.3
225	80.5	72.8
500	82.3	81.2
1000	75.1	70.2
5000	74.1	68.3
10000	65.3	62.0
ours	**86.6**	73.5

Table 9
Performance of different proportions.

Generic/ Specific	P	R	F	Generic/ Specific	P	R	F
10/100	0.581	0.512	0.544	100/10	0.648	0.652	0.650
20/100	0.591	0.558	0.574	100/20	0.665	0.678	0.671
30/100	0.647	0.603	0.624	100/30	0.689	0.690	0.689
40/100	0.741	0.712	0.726	100/40	0.735	0.762	0.748
50/100	0.765	0.742	0.753	100/50	0.775	0.783	0.779
60/100	0.798	0.784	0.791	100/60	0.802	0.813	0.807
70/100	0.825	0.819	0.822	100/70	0.835	0.839	0.837
80/100	0.848	0.840	0.844	100/80	0.858	0.860	0.859
90/100	0.855	0.843	0.849	100/90	0.861	0.865	0.863

Table 10
Comparison of different categories of documents for the performance of DGeoSegmenter. *RGR* denotes a regional geological report, *EGR* denotes an environmental geological report, *MGR* denotes a mineral geological report, *HGR* denotes a hydrogeology geological report, *GGR* denotes a geophysical geological report, and *RSGR* denotes a remote sensing geological report. Here, *P*, *R*, and *F* indicate the precision, recall, and *F* value, respectively.

	P	R	F
Baseline	79.6	80.5	80.0
RGR	85.26	85.73	85.49
EGR	87.15	87.08	87.11
MGR	85.67	85.81	85.74
HGR	85.48	85.31	85.39
GGR	86.82	85.73	86.27
RSGR	83.17	83.45	83.31
Avg	85.59	85.52	85.55

categorized the words based on the word frequency, which was used to determine the importance of words. Then, we randomly selected words according to the frequency of the words and combined them into sentences. Finally, the sentences were fed into a neural network model for training.

In our future work, we plan to perform experiments with different extraction and combination algorithms. Additionally, some supervised methods can be integrated into the proposed approach, which improves the performance. Another challenge worth investigating in the future is addressing the upper-bound limit on performance due to mistakes. Finally, we will also expand our method to other domains, such as biomedicine and history.

Author contributions

Liang Wu. Conceived and designed the experiments: Qinjun Qiu, Liang Wu, and Zhong Xie; Performed the experiments: Qinjun Qiu, Liang Wu, and Zhong Xie; Analyzed the data: Qinjun Qiu, Liang Wu, and Zhong Xie; Wrote the paper: Qinjun Qiu, Liang Wu, Zhong Xie and Wenjia Li.

Acknowledgments

We would like to thank the anonymous reviewers for carefully reading this paper and their very useful comments. This study was financially supported by the National Key Research and Development Program (Grant No: 2017YFB0503600), the National Key Research and Development Program (Grant No: 2017YFC0602204), the National Natural Science Foundation of China (41671400) and the National Science and Technology Major Project of China (Grant No: 2018YFB0505500).

References

Arnab, A., Jayasumana, S., Zheng, S., Torr, P.H., 2016. Higher order conditional random fields in deep neural networks. In: Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016. Springer, Cham, Switzerland, pp. 524–540.
Bengio, Y., Simard, P., Frasconi, P., 2002. Learning long-term dependencies with gradient descent is difficult. IEEE Trans. Neural Network. 5 (2), 157–166.
Chang, P.C., Galley, M., Manning, C.D., 2008. Optimizing Chinese word segmentation for machine translation performance. In: The Workshop on Statistical Machine Translation. Association for Computational Linguistics, pp. 224–232.
Cracknell, M.J., Reading, A.M., 2014. Geological mapping using remote sensing data: a comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. Comput. Geosci. 63 (1), 22–33.
Deng, K., Bol, P.K., Li, K.J., et al., 2016. On the unsupervised analysis of domain-specific Chinese texts. Proc. Natl. Acad. Sci. Unit. States Am. 113 (22), 6154–6159.
Duchi, J., Hazan, E., Singer, Y., 2011. Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. 12 (Jul), 2121–2159.
Gao, J., Wu, A., Li, M., et al., 2004. Adaptive Chinese word segmentation. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, pp. 462–469.
Graves, A., Schmidhuber, J., 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Network. 18 (5–6), 602–610.
Hochreiter, S., Schmidhuber, J., 2012. Long short-term memory. Neural Comput. 9 (8), 1735–1780.
Huang, L., Du, Y., Chen, G., 2015. GeoSegmenter: a statistically learned Chinese word segmenter for the geoscience domain. Comput. Geosci. 76, 11–17.
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. vol. 1. pp. 184–193 (Volume 1: Long Papers).
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105.
Kummerfeld, J.K., Berg-Kirkpatrick, T., Dan, K., et al., 2015. An empirical analysis of optimization for max-margin NLP. In: Conference on Empirical Methods in Natural Language Processing, pp. 273–279.
Lima, L.A., Görnitz, N., Varella, L.E., et al., 2017. Porosity estimation by semi-supervised learning with sparsely available labeled samples. Comput. Geosci. 106, 33–48.
Liu, Y., Zhang, Y., Che, W., et al., 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 (EMNLP), pp. 864–874.
Mike, Schuster, Kuldip, K Paliwal, 1997. Bidirectional recurrent neural networks. IEEE Trans. Signal Process. 45 (11), 2673–2681.
Mikolov, T., Karafiát, M., Burget, L., et al., 2010. Recurrent neural network based language model. In: Conference of the International Speech Communication Association. vol. 2. pp. 3.
Nawroth, C., Schmedding, M., Brocks, H., et al., 2015. Towards cloud-based knowledge capturing based on natural language processing. Procedia Comput. Sci. 68, 206–216.
Palmer, D.D., 1997. A trainable rule-based algorithm for word segmentation. In: Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 321–328.
Pei, W., Ge, T., Chang, B., 2014. Max-margin tensor neural network for Chinese word segmentation. In: Meeting of the Association for Computational Linguistics. vol. 1. pp. 293–303.
Peng, J., Choo, K.K.R., Ashman, H., 2016. Bit-level N-gram based forensic authorship analysis on social media: identifying individuals from linguistic profiles. J. Netw. Comput. Appl. 70 (C), 171–182.
Qiu, L., Zhang, Y., 2015. Word Segmentation for Chinese Novels. AAAI, pp. 2440–2446.
Ratliff, N.D., Bagnell, J.A., Zinkevich, M.A., 2006. Subgradient Methods for Structured Prediction. pp. 380–387 (Online).
Shu, X., Wang, J., Shen, X., Qu, A., 2017. Word segmentation in Chinese language processing. Stat. Interface 10 (2), 165–173.
Socher, R., Bauer, J., Manning, C.D., et al., 2013. Parsing with compositional vector grammars. In: Meeting of the Association for Computational Linguistics, pp. 455–465.
Taskar, B., Chatalbashev, V., Koller, D., et al., 2005. Learning structured prediction models: a large margin approach. In: International Conference on Machine Learning. ACM, pp. 896–903.
Tripathy, A., Agrawal, A., Rath, S.K., 2016. Classification of sentiment reviews using N-

gram machine learning approach. Expert Syst. Appl. 57, 117–126.
Tsochantaridis, I., Joachims, T., Hofmann, T., et al., 2005. Large margin methods for structured and interdependent output variables. J. Mach. Learn. Res. 6 (2), 1453–1484.
Wang, P., Qian, Y., Soong, F.K., et al., 2015. A unified tagging solution: bidirectional LSTM recurrent neural network with word embedding. arXiv Preprint arXiv:1511.00215.
Wang, C., Ma, X., Chen, J., et al., 2018. Information extraction and knowledge graph construction from geoscience literature. Comput. Geosci. 112, 112–120.
Wu, L., Xue, L., Li, C., et al., 2017. A knowledge-driven geospatially enabled framework for geological big data. Int. J. Geo-Inf. 6 (6), 166.
Xue, N., 2003. Chinese word segmentation as character tagging. Comput. Ling. Chin. Lang. Process. 8 (1), 29–48.
Yue, J., Jinan, X.U., 2013. Chinese word segmentation for patent documents. Acta Sci. Nauralium Univ. Pekin. 49 (1), 159–164.
Zaremba, W., Sutskever, I., Vinyals, O., 2014. Recurrent neural network regularization. Eprint Arxiv 1409.2329, 2014.
Zeng, D., Wei, D., Chau, M., et al., 2011. Domain-specific Chinese word segmentation using suffix tree and mutual information. Inf. Syst. Front 13 (1), 115–125.
Zhang, Q., Liu, X., Fu, J., 2018. Neural Networks Incorporating Dictionaries for Chinese Word Segmentation.
Zheng, X., Chen, H., Xu, T., 2013. Deep learning for Chinese word segmentation and POS tagging. In: Conference on Empirical Methods in Natural Language Processing.