applied sciences

MDPI

*Article*

# Candidate Term Boundary Conflict Reduction Method for Chinese Geological Text Segmentation

**Yu Tang [1,2], Jiqiu Deng [1,2,*] and Zhiyong Guo [1,2]**

[1] Key Laboratory of Metallogenic Prediction of Nonferrous Metals and Geological Environment Monitoring, Central South University, Ministry of Education, Changsha 410083, China; yutang@csu.edu.cn (Y.T.)
[2] School of Geosciences and Info-Physics, Central South University, Changsha 410083, China
* Correspondence: csugis@csu.edu.cn; Tel.: +86-13-8749-50729

**Abstract:** Though Chinese word segmentation (CWS) relies heavily on arithmetic power to train huge models and human work to label corpora, models and algorithms are still less accurate, especially for segmentation in a specific domain. In this study, a high-degree-of-freedom-priority candidate term boundary conflict reduction method (HFCR) is proposed to solve the problem of manually setting thresholds on segmentation based on information entropy. We quantify the uncertainty of left and right character connections of candidate terms and then arrange them in descending order for local comparisons to determine term boundaries. Dynamic numerical comparisons are adopted instead of setting a threshold manually and randomly. Experiments show that the average F1-value of CWS for Chinese geological text is higher than 95% and the F1-value for Chinese general datasets is higher than 87%. Compared with representative tokenizers and the SOTA model, our method performs better, which solves the term boundary conflict problem well and has excellent performance on single geological text segmentation without any samples or labels.

**Keywords:** Chinese word segmentation; information entropy; degree of freedom of terms; zero-sample; single geological text

check for updates

**Citation:** Tang, Y.; Deng, J.; Guo, Z. Candidate Term Boundary Conflict Reduction Method for Chinese Geological Text Segmentation. *Appl. Sci.* **2023**, *13*, 4516. https://doi.org/10.3390/app13074516

Academic Editor: Yu-Dong Zhang

Received: 6 March 2023
Revised: 24 March 2023
Accepted: 31 March 2023
Published: 2 April 2023

**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

Driven by technologies such as big data and cloud computing, the paradigm of geological research has changed from the previous empirical paradigm and theoretical paradigm to the data-intensive paradigm. Geological big data are mainly divided into structured data and unstructured data. Processing models of structured data have become mature [1]. There is still a large space during the development and utilization of unstructured data. Among unstructured data, text data are widespread and easy to obtain, which usually has high original production costs and contains rich domain knowledge. It is essential to make accurate and fully digital use of this knowledge [2]. Chinese word segmentation (CWS) is a basic task, and it directly affects the processing accuracy and performance of downstream tasks such as text understanding and knowledge fusion. So, CWS is an essential step during the utilization of Chinese text data [3,4], which is also a basic work for data mining, geological named entity recognition, and providing knowledge services in the geological field [5].

Two major difficulties in the current development of general word segmentation are ambiguity recognition and out-of-vocabulary term (OOV) recognition, which also exists in geological text segmentation. Because geological text contains lots of geological professional words, it is difficult to achieve the correct segmentation of these specialized expressions. Existing CWS systems mostly rely on a large amount of human work and arithmetic power to improve the accuracy such as high-quality labels, excellent feature selection, and the construction of complex models or proper domain dictionaries for specific domain [3,5–7]. However, domain dictionaries and corpus cannot remain unchanged and need to be updated

continuously and manually. Therefore, CWS tasks need to be free from manual consumption and high-performance equipment.

To address the above-mentioned issues, a high-degree-of-freedom-priority candidate term boundary conflict reduction method (HFCR) is proposed. The idea of this method is analogous to the process of human reading. A person who first starts reading a text without a knowledge background will gradually confirm word meanings and word groupings by gradually increasing the number of occurrences and contexts of proper nouns. Theoretically, the richer the left and right connections of a string and the more independently it can appear in the text, the higher the probability of the string becoming a term. We use information entropy to quantify the uncertainty of left and right character connections of candidate terms and to determine term boundaries. The experiments demonstrate that the HFCR algorithm achieves an average F1 value of more than 95% for single geological text. Compared with the representative tokenizer and SOTA, the HFCR algorithm has obvious advantages and achieves single geological text segmentation without any samples, labels, or high-performance equipment. At the same time, our algorithm also performs well in the general field.

The rest of this article is organized as follows. Section 2 describes some related research work. Section 3 introduces our ideas and algorithm. Section 4 discusses the experimental data and results. Section 5 summarizes the advantages and disadvantages of the HFCR algorithm and puts forward suggestions for future research and improvement.

## 2. Related Work

### 2.1. Technology of CWS in the General Field

Throughout the development of word segmentation technology, approaches to CWS can be mainly divided into four categories according to different principles: dictionary-based approaches, statistics-based approaches, deep-learning-based approaches, and multi-mode integration approaches.

Dictionary-based approaches, also known as mechanical segmentation, refer to the mechanical segmentation of Chinese text based on a relatively complete dictionary and matching segmentation method to obtain term segmentation sequences. Sun Maosong compared the spatial efficiency and processing speed of binary-seek-by-word, TRIE indexing tree, and binary-seek-by-characters. The experimental results proved that the processing speed of binary-seek-by-characters was increased by more than 15 times and the effect was obvious [8]. Yang Yan proposed a mechanical statistical word segmentation based on hash structure, which combined the advantages of mechanical word segmentation and statistical word segmentation and had a strong recognition ability of OOV. However, there are significant differences between the speed and accuracy of word segmentation when using dictionaries with different amounts of words [9]. Mo Jianwen proposed an improved positive maximal matching algorithm with a double hash structure, but there were still some deficiencies in the recognition of ambiguous terms and OOV [10]. Zheng Mugang improved the recognition of OOV and ambiguous term through a bidirectional matching algorithm, but the covering ambiguity could not be dealt with [7]. Dictionary-based approaches have the advantages of simplicity, quickness, and easy implementation, and they can improve OOV recognition and ambiguity recognition to some extent. Although high-quality dictionaries are often used in related fields, there are some limitations to their use. With the emergence of large amounts of new words, a solidified dictionary has obvious defects in the performance of specific word segmentation, which requires constant updating and manual maintenance, and it is difficult to keep a balance between the accuracy rate and manual cost.

Statistics-based approaches use statistics to obtain the mutual information of words based on the adjacent relationships between words and establish a word segmentation model or algorithm, so as to recognize new words and cut sentences into words. Its main advantage is that it makes use of semantic connection relationships, balances the recognition of different characters, and has a good recognition effect on OOV. Machine learning is a

typical representative of statistical methods. The widely used algorithms are the Hidden Markov Model (HMM) [11], Maximum Entropy Markov Model (MEMM) [12,13], the Conditional Random Field (CRF) [14], and their improved algorithms. For example, Zhou J S used basic segmentation and named entity recognition based on CRF to generate initial segmentation results and corrected the word segmentation results with statistical methods and grammatical rules, which achieved good word segmentation results [14]. Although statistics-based approaches are effective, their limitation is that the machine learning model relies on excellent feature engineering defined manually and a lot of work is required to verify the effectiveness of these features, which is relatively laborious.

With the development of machine learning, segmentation algorithms based on neural networks and deep learning have emerged, which can train a neural network to automatically learn features. Therefore, the labor cost and time costs will be greatly reduced in the feature engineering stage and the efficiency of feature engineering extraction will be improved. In 2011, Collobert introduced a deep learning algorithm into the natural language processing task for the first time [15]. This method can learn original characteristics and context representations through the final word segmentation training set. Subsequently, deep learning models such as CNN [16], LSTM [17], and BiLSTM [18] were introduced into CWS tasks, especially the Bert pre-training model [19] proposed by Devlin et al., in 2018, which showed strong feature extraction ability in various natural language processing tasks. Although deep-learning-based approaches reduce labor costs, they cannot completely remove human supervision from the whole process. The operation of complex models relies on high-performance computing processing and large-scale corpuses and requires staff with a domain knowledge background to debug parameters. The performance of word segmentation in different fields is different under the influence of domain knowledge.

In addition to the three categories of word segmentation methods with clear principles, some researchers have utilized many integrations and combinations which attempt to integrate the advantages of different principles and the realization of specific problems. Based on CRF and transfer learning, Zhang Xin effectively combined the improved Bert model and BiLSTM-CRF word segmentation system in the architectural field through training corpus after tokenization [20]. Liu Shuangqiao adopted the unsupervised learning segmentation method based on SentencePiece and built a word segmentation model applicable to the field of traditional Chinese medicine [21]. The experimental verification accuracy of the model in the test set was 84%, and the recall rate was 83%. These methods combine the advantages of different principles and techniques and have achieved good results. According to the scope of application, CWS can be divided into the general field and specific fields. The word segmentation method in the general field is flexible and not limited by domain knowledge. The current mainstream word segmentation system can achieve good results. In the specific field, a large number of professional words limit the accuracy of word segmentation.

### *2.2. CWS Technology in the Geological Field*

Digital utilization of geological texts is mostly based on the recognition and extraction of geological named entities and knowledge services are provided by building knowledge graphs. However, lots of context information is lost in the process of extracting knowledge entities, which usually contain more domain knowledge. It is very necessary to achieve both knowledge extraction and named entity recognition through word segmentation.

In the geographical field, Chen Jingwen, relying on a geological dictionary, carried out word segmentation research on mineral texts based on the characteristic of the dictionary and CRF model [3]. Zhang Xueying designed a geological entity recognition model based on the Deep Belief Network, which can achieve better performance on the basis of a small-scale corpus [22], but there are still some text information recognition errors in the existing experiments. Wang Hong built a basic geological dictionary based on the prime string to improve the adaptability of statistics-based segmentation methods in geological text segmentation, but he could not recognize the new words that appeared only once and in

the exactly same context [23]. Chu Deping et al., established the ELMO-CNN-BiLSTM-CRF model, which improved the accuracy of geological entity recognition in small-scale corpus and could effectively identify long geological entity words and geological polysemous words [24], but misclassification and missing characters occurred. Generally relying on geological dictionaries, statistics-based approaches and deep-learning-based approaches have been continuously applied to geological texts. Based on the small-scale corpus, good word segmentation results have been achieved, which can basically recognize terms and common words in the geological field. In general, CWS in the geological field has long relied on and lagged behind the development of word segmentation technology in the general field.

### 2.3. Zero-Sample Text Segmentation

Based on the problem that the existing word segmentation technology relies on a large amount of human work, some scholars have carried out excellent theoretical work and developed different word segmentation technology without training samples. Under the condition of unlabeled document resources, Liu Yin used BERT full-word masking technology to complete word prediction, which reduced the overall requirements for data conditions [25]. However, serial frequency statistics technology and structured data are combined to update the dictionary in the electric power field. Gu Chun extracted candidate words from text preprocessing and then combined them with the BERT model, candidate words, and text similar semantics to extract keywords, which improved the problem of Chinese single-text keyword extraction and performed better in the short text, but the accuracy was lower than 40% [26]. Based on the character connectivity of ancient Chinese medicine texts, Zhang Suhua proposed the ConnectRank algorithm, which fully realized unsupervised word segmentation, but the accuracy was only 67.6% [27]. A meaningful research attempt has been made to recognize OOV without human work from various angles, but there is still much room for improvement in accuracy. To fully realize zero-sample text segmentation in specific fields and truly realize segmentation without human work, we must start from the text itself to achieve an effective balance between zero-sample text and accuracy.

### 2.4. CWS Technology Based on Information Entropy

Reviewing the development of CWS, many researchers have solved the task of CWS from the perspectives of word, grammar, and sentence dependency relationships based on the preconceived knowledge system they have learned, which naturally opened the way for dictionary-based or deep-learning-based model word segmentation technology. When knowledge thinking and external information input are discarded, zero-sample geological text word segmentation is transformed into a word recognition problem only through the own text information of geological text data. Text is composed of several characters and is the result of the connection and composition of characters. Characters exist before the grammatical structure and words, and words themselves are also the results of the combination and connection of characters. The word segmentation method based on texts and characters transforms the problem of term recognition into the problem of character connection judgment, focusing on quantifying the uncertainty of the adjacency before and after the candidate term. As a classical uncertainty theory in information theory, information entropy is a very good quantitative indicator.

In thermodynamics, entropy is a physical quantity that represents the degree of disorder in the molecular state, and the theory of information entropy was established by Shannon. Information entropy was first introduced into the field of CWS when Zhang Min [28] carried out research on word and character relationships and filtered words by a given empirical threshold. Later, Ren He [29] extracted high-frequency words from the text by comparing information entropy with the manually set threshold, which achieved good results. Although information entropy is a classical uncertainty theory, the research on information entropy* is mainly distributed in information resource management [30],

hydrological simulation [31], no-reference image quality evaluation technology [32], and link prediction [33]. In recent years, Gao Jiaqi et al., discovered new words from a large number of classical literary corpus based on left–right entropy and mutual information to build a segmentation dictionary of ancient Chinese, but the accuracy rate, recall, and F-value of new-word discovery were low [34]. Based on the text mining theory of information entropy, Zhang Guandong analyzed the praise and criticism tendencies of the overall text information [35]. Existing research and experiments have proved the feasibility of information entropy theory in CWS. There is little research on information entropy in CWS, and it is hardly applied in geological word segmentation. This provides a good opportunity to introduce information entropy into geological text word segmentation tasks.

## 3. Candidate Term Boundary Conflict Reduction Method Based on Information Entropy

### *3.1. Technical Route of Word Segmentation Algorithm for Single Geological Text*

Analogously to the process of human reading, when a string appears frequently in a text with a fixed connection, it is probably a correct word. By observing the writing features and characteristics of geological texts, it can be found that proper nouns are expressed as fixed connections in terms of character relationships. All the Chinese words in this article are replaced by Pinyin and explained in English. For example, "Shan Chang Yan (diorite)" in Chinese becomes a word with complete basic meaning through the connection of three characters: "Shan (flash)", "Chang (long)", and "Yan (stone)". This word is correct and is allowed to appear freely in other positions of the text as a fixed whole. Therefore, there should be abundant possible items for the left and right connecting characters of this term, and the left and right degree of freedom of the correct term is relatively high. The tightness of internal character connection and richness of external character connection of one term are meaningful and regular. This research uses information entropy to quantify the left and right degrees of freedom of candidate terms.

Although there are no natural space separators as there are in English, Chinese text has natural punctuation marks such as commas, periods, semicolons, exclamation marks, and question marks, which can be initially cut off through preprocessing. The term combination cases of a text are exhaustive, and candidate terms are all combinations of characters of indefinite length. According to the text reading habit, all character combinations are obtained by traversing the preprocessed text from left to right and calculating information entropy; terms with an information entropy of 0 (that is, the degree of freedom is 0) were deleted as preliminary filtering to obtain the candidate term set.

In the actual word segmentation, some overly active characters such as "De (of)", "Shi (is)", and "Le (the)" in Chinese will freely appear in connection with other characters in the text. To a certain extent, these special combinations will affect the result of word segmentation. These active characters are called stop words in many word segmentation studies, which need to be skipped in advance. In order to further avoid manual definition, information entropy can be used to quantify the activity of a single character. After performing single-character experiments on numerous texts, the first part of the single character with the highest activity (that is, the highest information entropy) can be summarized and extracted as stop words.

There will be certain conflicts between candidate terms such as inclusive conflicts between "He Nan (Henan)" and "He Nan Sheng (Henan province)" and intersecting conflicts between "Hui Se (gray)" and "Se Zhong (color middle)". These are conflicts about the term boundary, and the information entropy of the determining term boundary is an index to quantify the internal tightness and external degree of freedom. A term with high information entropy is represented as a whole with close connections and a high degree of freedom in the text. For example, words such as "Yan Shi (rock)", "Kuang Qu (mining area)", "Di Zhi (geology)", and "Zuan Kong (borehole)" show extremely high information entropy. Compared with using a fixed threshold to filter candidate terms, it is more in line with the cognition of the meaning of information entropy to filter the value of information

entropy on the conflict side. Thus, the technical route flow chart of Chinese geological text word segmentation based on information entropy can be obtained, as shown in Figure 1 below.

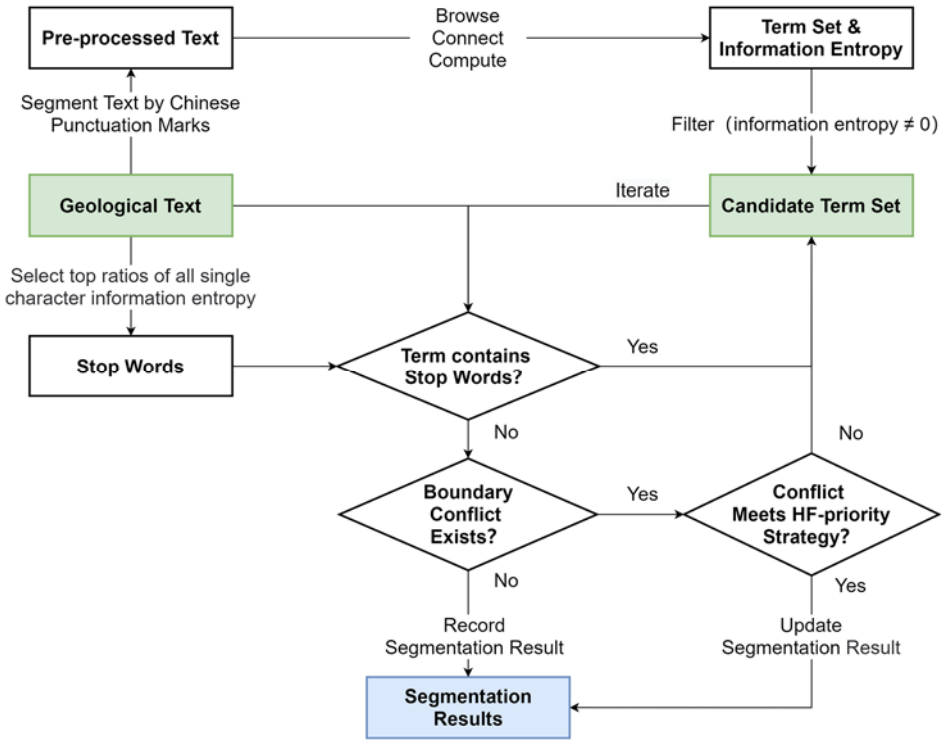



**Figure 1.** Technical route of Chinese geological text segmentation based on information entropy.

### *3.2. Information Entropy of Single Geological Text Candidate Terms*

Entropy quantifies chaotic relationships of molecular states in thermodynamics; information entropy represents uncertainty theory in information theory. In the field of word segmentation, it can be used to quantify the internal tightness and external degree of freedom of terms. Its mathematical expression is as follows:

$$H = -\sum P(x)\log P(x) \tag{1}$$

Based on the character connection relationships and meaning of the formula, the mathematical expression of the left and right information entropy converted to a string is as follows:

$$H_l(s) = -\sum_{a \epsilon A} P(s_{la}/s)\log(P(s_{la}/s)) \tag{2}$$

where, for the candidate term $s$, $H_l(s)$ is the information entropy on the left side of the term $s$. $A$ is the set of all Chinese characters on the left side of $s$. $s_{la}$ represents combinations formed of the Chinese character $a$ on the left side and $s$. $P(s_{la}/s)$ represents the conditional probability that the Chinese character $a$ appears on the left side of $s$ under the premise that the term $s$ appears in the text.

$H_l(s)$ reflects the average uncertainty of Chinese characters appearing on the left side of string $s$. The larger $H_l(s)$ is, the more uncertain the Chinese character collocation on the left side of $s$ is, that is, the higher the left degree of freedom of the term s is and the greater the probability of inferring that the string combination $s$ should be disconnected on the left side as the term boundary. Taking "Di Zhi (Geology)" and "Suan Yan (-ate, such as a suffix of carbonate)" as examples, the approximate calculation process of their information entropy is shown in Figure 2 below. The left information entropy of "Di Zhi (Geology)" is 4.8414, while the left information entropy of "Suan Yan (-ate)" is only 0.1274.

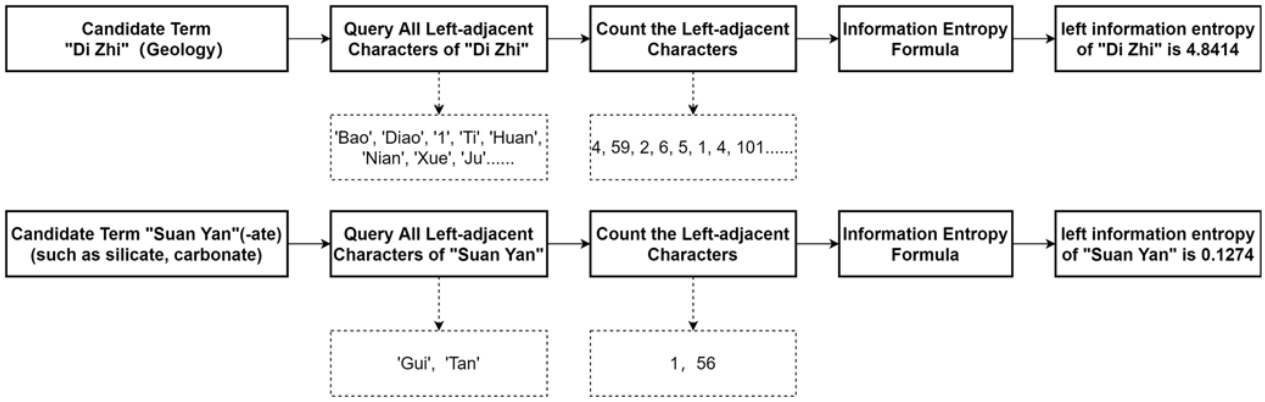



**Figure 2.** Flow chart of information entropy calculation.

Similarly, the mathematical expression of information entropy on the right side of candidate term $s$ can be obtained:

$$H_r(s) = -\sum_{\alpha \epsilon A} P(s_{ra}/s)\log(P(s_{ra}/s)) \quad (3)$$

Information entropy is a quantification of the internal tightness and external degree of freedom of a character combination. From the numerical results, the information entropy of a certain character combination can be zero, which means that the left and right connections of the character combination in the current calculation space are fixed without uncertainty and a degree of freedom. It can be non-zero with uncertainty and a degree of freedom, indicating that the character combination basically constitutes a term boundary. The information entropy can be very high, indicating that when the character combination appears as a whole, there are many uncertain possibilities for the left and right connections, that is, the degree of freedom is large. Then, we believe that the combination appears in the text as a whole in enough occurrences and the probability of it being a correct term is high. It may be very low, indicating that the uncertainty is small, the string group has fewer external character collocations, or the string itself is less; then, the probability of being a correct term is small.

### *3.3. Boundary Conflict Reduction for Candidate Term*

In the actual process of word segmentation, word segmentation conflicts of candidate terms will be encountered. Through a large number of example analyses and induction, the conflict area is mainly the text character range covered by the conflict terms, which can be divided into two types of conflicts: inclusive conflict and intersecting conflict.

Inclusive conflicts are conflicts in which one candidate term contains another candidate term, resulting in doubtful segmentation. As shown in Figure 3 below, inclusive conflict can be subdivided into same-starting-point inclusive conflict, same-ending-point inclusive conflict, and inner inclusive conflict. Among them, the conflict of the same starting point and the same ending point only causes doubt in the boundary of one side of the term, whilst the boundary of another side is certain. Since there are different segmentation requirements in different texts, inclusive conflict sometimes does not involve absolute errors, and appropriate term boundaries need to be selected according to granularity requirements. In the sentence example "Chang Zhou Gou Zu Chu Lu Yu Ce Qu Xi Bu Yang Er Zhuang (The Changzhougou Formation is exposed in Yangerzhuang in the west of the survey area)" shown in the following figure, the line segment composed of the left and right information entropy represents the left and right boundary degree of freedom of each candidate term. According to the logic of information entropy to quantify the internal tightness and external degree of freedom, when an inclusive conflict occurs, the term will have high information entropy regarding the conflict boundary, that is, a high degree of freedom should be selected as the result of word segmentation, that is, "Chang Zhou Gou

Zu (Changzhougou Formation)", "Chu Lu (exposed)", "Ce Qu Xi Bu (in the west of the survey area)" as a result of conflict resolution. In a large number of case verifications, this solution meets the expectation of word segmentation results.

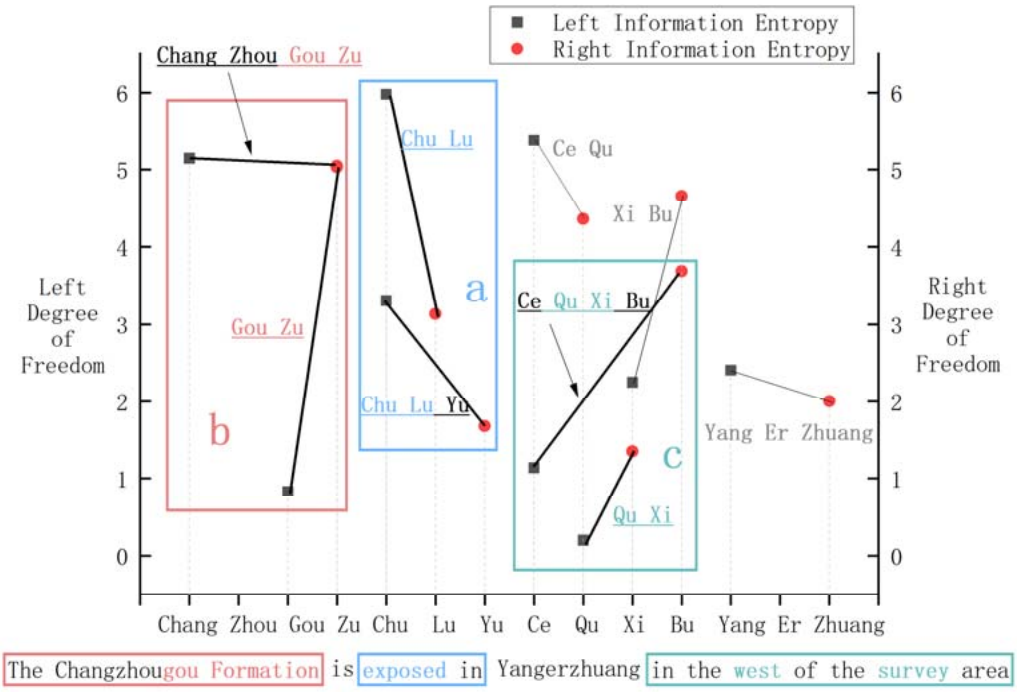



**Figure 3.** Logic visualization of inclusive conflict. (**a**) Same-starting-point inclusive conflict; (**b**) same-ending-point inclusive conflict; (**c**) inner inclusive conflict. Note: the color of the Chinese conflict is the same as in English.

Intersecting conflicts are conflicts in which a candidate term is connected to another candidate term and the attribution of some characters on the related side is uncertain. Intersecting conflict usually involves absolute errors in the attribution of some characters, which must be solved. Observing the examples of the candidate term on the uncertain side, in the sentence "Zhong Ceng Zhuang Tan Suan Yan Jiao Jie Xi Li Shi Ying Sha Yan Jia Zi Se Ye Yan (medium layered carbonate cemented fine-granularity quartz sandstone with purple shale)" shown in Figure 4 below, there is an intersecting conflict between "Xi Li (fine-granularity)" and "Li Shi (granularity stone)", "Sha Yan (sandstone)" and "Yan Ji (stone with)". The resolution of the intersecting conflict is to solve the boundary problem of the conflict characters belonging to the left or to the right. As shown in the figure, the right-side information entropy of "Xi Li (fine-granularity)" is higher than the left-side information entropy of "Li Shi (granularity stone)", which shows that the character "Li (granularity)" has a higher internal tightness and external degree of freedom when it becomes the right boundary than the left boundary. It is better to choose "Xi Li (fine-granularity)" as the word segmentation result. Similarly, the right boundary of "Sha Yan (sandstone)" has a higher degree of freedom, which yields a better word segmentation result.

There is no benchmark value for information entropy, and the numerical distribution intervals of specific fields and different texts are also different. It is difficult to artificially set a suitable threshold based on knowledge to adapt to different word segmentation requirements. The research on word segmentation based on information entropy cannot simply start from the numerical value, and it needs to comprehensively compare the change and distribution of information entropy in different positions of the text. According to the logic of quantifying the internal tightness and external degrees of freedom by information entropy, the solution to different types of conflicts is to adopt the strategy of giving priority to high degrees of freedom in conflict areas. Table 1 below summarizes

the conflict reduction methods and shows more conflict instances where the underlined candidate term and values are better results of two segmentation conflicts.

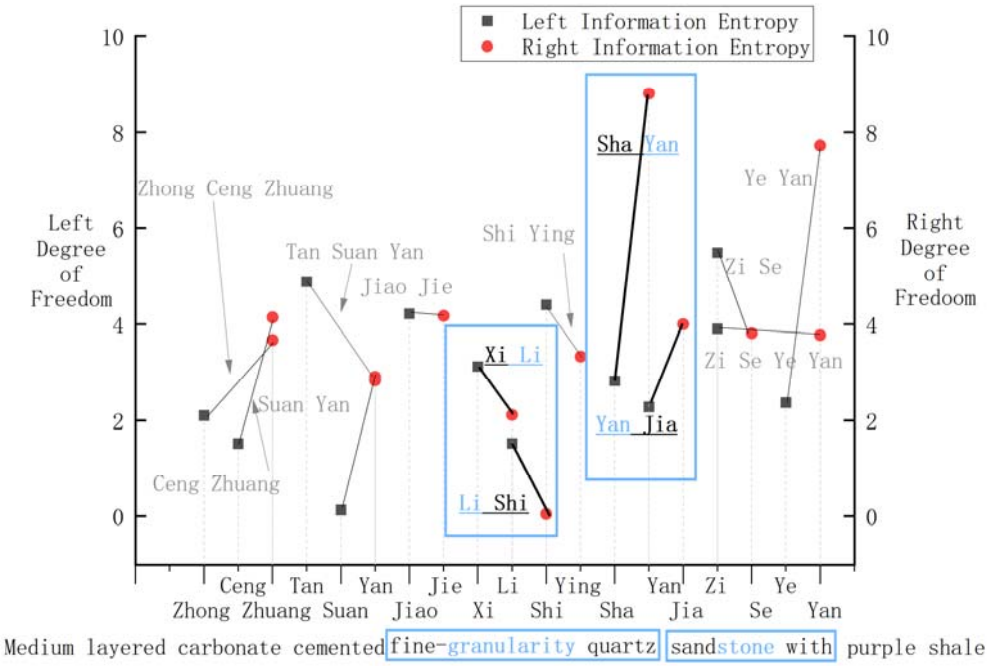



**Figure 4.** Logic visualization of intersecting conflict.

**Table 1.** High-degree-of-freedom priority strategy based on information entropy.

| Type of Conflict | Side | Candidate Term 1 | Information Entropy 1 | Candidate Term 2 | Information Entropy 2 |
|---|---|---|---|---|---|
| same-starting-point inclusive | right | Zhong Yuan Gu (Mesoproterozoic) | 0.27 | Zhong Yuan Gu Jie (n.Mesoproterozoic) | 3.24 |
| | | Zan Huang (Zanhuang) | 0.42 | Zan Huang Yan Qun (Zanhuang Group) | 5.27 |
| same-ending-point inclusive | left | Sha Zhi Ye Yan (sandy shale) | 2.86 | Zhi Ye Yan (quality shale) | 1.35 |
| | | Li Shi Ceng (gravel layer) | 3.65 | Jia Li Shi Ceng (with gravel layer) | 1.00 |
| inner inclusive | both | Jia Gou (jiagou) | (0.82, 0.53) | Ma Jia Gou Zu (Majiagou Formation) | (4.89, 3.87) |
| | | Shi Ying (quartz) | (4.40, 3.33) | Yu Shi Ying Xiang Jian (alternate with quartz) | (0.91, 2.25) |
| intersecting | connected | Hui Se (grey) | (7.74, 3.30) | Se Zhong (color middle) | (2.13, 1.40) |
| | | Ju Di (office No.) | (1.76, 1.36) | Di Qi (seventh) | (2.16, 1.49) |

## 4. Experiments and Discussion

### *4.1. Data and Evaluation Indicators*

To avoid contingency, three geological reports were used in this study to verify the segmentation effect of the HFCR algorithm. Experimental datasets include the Rencun Geological Report (hereinafter referred to as Rencun), the Xiaoshan Geological and Mineral Report (referred to as Xiaoshan), and the Geological and Mineral Results Report in Zhenfeng and Funing areas of Nanpanjiang Metallogenic Area (referred to as Nanpanjiang). This kind of report is the summative material of geological work, which has abundant reserves in the whole geological field and is very easy to obtain. It can be used as the representative data set for the task of geological text segmentation.

The evaluation system of CWS mainly contains four widely used indicators, namely Accuracy, Precision, Recall, and F-score. Accuracy is the proportion of correct cases to all cases. Precision is the proportion of correct segmentations to all segmentation cases. Recall

is the proportion of correct segmentations to the actual segmentation cases. F-value is a comprehensive indicator with a compromise between Precision and Recall. Usually, the F-value is also called the F1-value when $\beta$ is 1. In this study, the comprehensive index F1-value is used as the judgment of the effect of word segmentation.

In this study, the evaluation of segmentation performance is focused on determining whether the segmentation operation of geological text is correct or not. In particular, it should be noted that our standard text may be slightly different from other studies due to no uniform scheme in the field of CWS in terms of the understanding of lexical terms. Our judgment criterion is that the segmentation constitutes the boundary of a correct term with a full meaning.

### *4.2. Experiment and Results*

Before the start of word segmentation, the activity of single characters should be quantified to avoid the manual definition of stop words. The information entropy of single characters for different texts will be calculated and then different proportions of single characters are selected according to the descending order of information entropy. Then, the appropriate proportion of the single characters in the full text can be compared and concluded. Finally, the top 1% of all single characters is selected as the optimal proportion of stop words. The examples of some stop words for a text are extracted, such as "De (of)", "Wei (for)", "He (and)", "Yu (with)", "Huo (or)", "Zai (on)", "Shi (is)", and "Yi(by)".

By verifying the performance of the HFCR algorithm in different fields and different texts, this study designed two experiments to compare the text segmentation effects of our algorithm and other tokenizers, respectively, such as Baidu LAC, Pkuseg, Hanlp, and Jieba.

The first experiment is to compare the performance of the HFCR algorithm and other tokenizers in geological texts. The selected part of specific experimental results is shown in Table 2. It can be seen that the word segmentation results of the HFCR algorithm meet expectations and can be correctly segmented in terms of age, rock, place name, numerical value, and non-professional expression, while other tokenizers all have partial semantic segmentation errors.

**Table 2.** Result display of the HCRF algorithm for geological text segmentation.

| Description | Text |
|---|---|
| Result in Pinyin | /Lin Zhou Shi/Ren Cun Zhen/Mu Qiu Quan/Zan Huang Yan Qun/Shi Ce/Di Zhi/Pou Mian/: Pou Mian/Qi Dian/Zuo Biao/X … … /Hui Hei Se/Shi Bian/Hui Chang Hui Lü Yan/,/Zhu Yao Kuang Wu/Cheng Fen:/Xie Chang Shi/50%,/Tou Hui Shi/44% |
| Result in English | /Geological profile/measured in/Muqiuquan/Zanhuang Rock Group/,/Rencun Town/,/Linzhou City/:The profile/starting coordinates/X … … /Gray-black/altered gabbro, main mineral composition:/plagioclase/50%,/diopside/44% |

Experiments show that the average F1-value of the HFCR algorithm in geological text word segmentation can reach more than 95%, which is superior to four representative tokenizers. The specific accuracy comparison is shown in Table 3 below.

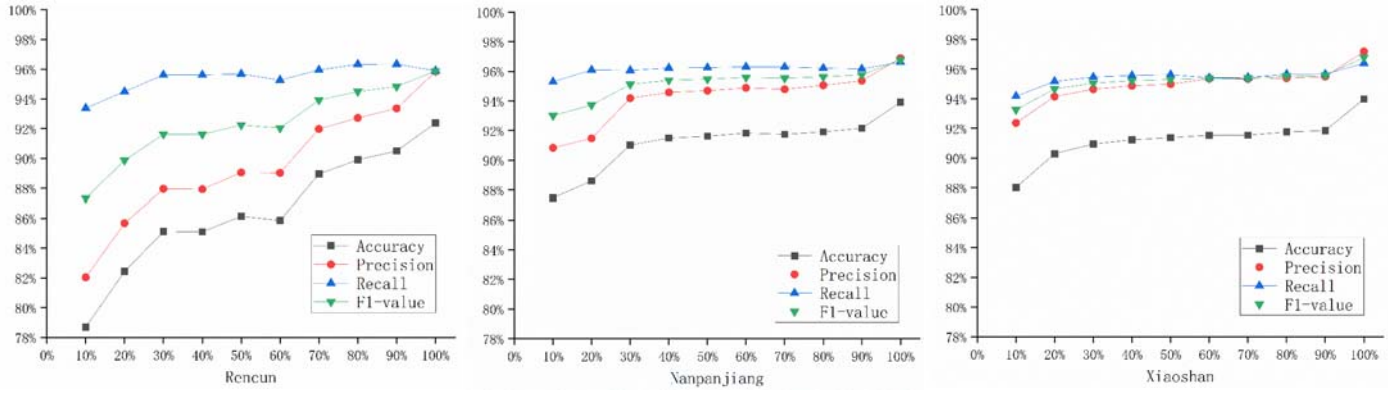
In Table 3, the text length is the length of characters. Our initial observation is that the word separation accuracy improves with increasing text character length, as shown in Table 3. Therefore, we verified the impact of candidate term sets calculated for different percentages of text on the text segmentation results. The results are shown in Figure 5 below. For the single geological independent text, the accuracy gradually increases as the range of the candidate term set is calculated, reaching the highest level when calculating the candidate term set for the full text. This result is consistent with how we read specific texts in reality. When we read part of a specific text, we may not understand it; however,

when we read more, our understanding will be deeper and clearer. For three geological texts, the accuracy is higher for longer text and slightly lower for shorter text.

**Table 3.** Comparison of accuracy between HFCR algorithm and other tokenizers (geological text).

| Data | Text Length | Tools | Accuracy | Precision | Recall | F1-Value |
|---|---|---|---|---|---|---|
| Rencun | 241k | HFCR | 92.38% | 95.85% | 95.90% | 95.88% |
| | | Baidu LAC | 80.63% | 86.62% | 89.86% | 88.21% |
| | | Pkuseg | 83.35% | 95.55% | 83.91% | 89.35% |
| | | Hanlp | 81.49% | 88.23% | 89.18% | 88.70% |
| | | Jieba | 78.95% | 87.94% | 84.99% | 86.44% |
| Nanpanjiang | 346k | HFCR | 93.91% | 96.89% | 96.62% | 96.76% |
| | | Baidu LAC | 78.13% | 80.12% | 95.77% | 87.25% |
| | | Pkuseg | 86.18% | 97.60% | 86.07% | 91.48% |
| | | Hanlp | 79.31% | 83.00% | 92.90% | 87.69% |
| | | Jieba | 83.99% | 91.46% | 89.27% | 90.35% |
| Xiaoshan | 576k | HFCR | 93.98% | 97.18% | 96.39% | 96.78% |
| | | Baidu LAC | 80.62% | 86.47% | 90.06% | 88.23% |
| | | Pkuseg | 82.72% | 92.55% | 86.04% | 89.18% |
| | | Hanlp | 80.73% | 89.20% | 86.62% | 87.89% |
| | | Jieba | 77.28% | 87.62% | 82.17% | 84.80% |

The underlined part indicates the optimal segmentation performance.




(**a**) (**b**) (**c**)

**Figure 5.** Variation under the different percentages of candidate term calculation range of the full text. (**a**) Rencun; (**b**) Nanpanjiang; (**c**) Xiaoshan. Note: The *x*-axis is the percentage of the currently calculated candidate term set in the full text, and the *y*-axis is the variation in accuracy.

The second experiment is to compare the performance of the HFCR algorithm and other tokenizers in general text. Datasets of general text segmentation experiment are PKU and MSR datasets, which are standard datasets for testing Chinese general word segmentation tools in academia. The selected part of the specific experimental results is shown in Table 4 below. It can be seen that the segmentation granularity of long terms such as "Song Nuo Si Yi Yuan (Sunnass Hospital)" and "Ao Si Lu Da Xue (University of Oslo)" are too short. The specific experimental accuracy is shown in Table 5 below. The experiment proves that the F1-value of the HFCR algorithm in general text segmentation is higher than 87%, which is slightly better than Jieba.

**Table 4.** Result display of the HCRF algorithm for general text segmentation.

| Description | Text |
| --- | --- |
| Result in Pinyin | /Zhe Xie/Te Shu/De/She Ji/,/Dui Zhe/Li De/Bing Ren/Lai Shuo/Shi Tie/Xin De/Guan Huai/./Zuo Wei/Ao Si Lu/Da Xue/De Yi/Bu Fen/,/Song Nuo Si/Yi Yuan/Dan Fu Zhe/Shu Xiang/Zhong Yao/De/Can Ji Ren/Kang Fu/Ke Yan/Gong Zuo/,/Bing Shi/Nuo Wei/Yan Jiu/Wai Ke/Yong Yao/He/Kang Fu/Yi Liao/De/Zhu Yao/Yi Yuan/. |
| Result in English | /These/special/designs/,/for/the patients/here/is intimate/care/./As/part/of the/University of/Oslo/,/Sonnos/Hospital/is responsible for/several/important/research/on the rehabilitation of/persons with disabilities/and is/the main/hospital/in Norway/for research/on surgical medicine/and/rehabilitation/. |

**Table 5.** Comparison of accuracy between HFCR algorithm and other tokenizers (general text).

| Data | Tools | Accuracy | Precision | Recall | F1-value |
| --- | --- | --- | --- | --- | --- |
| MSR | HFCR | 79.37% | 86.63% | 87.51% | 87.07% |
| | Baidu LAC | 86.46% | 90.41% | 94.35% | 92.38% |
| | Pkuseg | 87.40% | 94.81% | 90.54% | 92.63% |
| | Hanlp | 85.53% | 90.99% | 92.21% | 91.60% |
| | Jieba | 78.63% | 87.85% | 84.50% | 86.15% |
| PKU | HFCR | 78.93% | 83.20% | 91.86% | 87.32% |
| | Baidu LAC | 82.51% | 84.00% | 97.35% | 90.18% |
| | Pkuseg | 90.25% | 92.41% | 97.18% | 94.73% |
| | Hanlp | 86.43% | 88.20% | 97.32% | 92.54% |
| | Jieba | 77.17% | 83.22% | 88.20% | 85.64% |

The underlined part indicates our performance and the optimal segmentation performance.

### *4.3. More Comparisons with SOTA Model*

In this paper, more SOTA models including GPT-3 and WMSEG are considered for comparison, except for representative tokenizers.

GPT-3 is a powerful third-generation language model developed by OpenAI and has 175 billion parameters, which is a huge and complex model. It is considered to be one of the most advanced language models available, and tasks are implemented with very simple instruction. In our experiments, we used the Davinci language model of GPT-3 and found that there is great uncertainty between the results for the same instruction. There are also restrictions on the number and length of long-text word segmentation. For the uncertainty of a short geological text result, experiments show that six types of different segmentation results can be generated in twenty random segmentations, among which three results are wrong, and the accuracy rate is about 85%. The average accuracy of word segmentation for twenty repeated experiments is shown in Table 6 below.

WMSEG is a neural framework, which uses memory networks to incorporate word-hood information with popular encoder–decoder combinations for CWS [6]. It belongs to the integrated approach and achieves state-of-the-art performance on all general datasets. However, its BERT-based pre-trained model underperforms in our experiments for geological text and has limitations on the data format and text length. The specific experimental accuracy is shown in Table 6 below.

**Table 6.** Comparison of accuracy between HFCR algorithm and SOTA models.

| Data | Tools | Accuracy | Precision | Recall | F1-Value |
|---|---|---|---|---|---|
| Rencun | HFCR | 92.38% | 95.85% | 95.90% | 95.88% |
| | WMSEG | 68.34% | 79.27% | 72.68% | 75.84% |
| | GPT-3 | 92.45% * | | | |
| Nanpanjiang | HFCR | 93.91% | 96.89% | 96.62% | 96.76% |
| | WMSEG | 80.03% | 92.83% | 81.33% | 86.70% |
| | GPT-3 | 94.59% * | | | |
| Xiaoshan | HFCR | 93.98% | 97.18% | 96.39% | 96.78% |
| | WMSEG | 76.25% | 86.05% | 82.16% | 84.06% |
| | GPT-3 | 92.24% * | | | |

The underlined part indicates the optimal segmentation performance and the asterisk * indicates the average accuracy of running 20 repeated experiments.

### *4.4. Discussion and Analysis*

Through experiments, it is found that the HFCR method based on information entropy has some outstanding features for both geological and general texts.

(1) The F1-value in the geological text word segmentation is higher than 95%, as shown in Tables 3 and 6, which is obviously better than the SOTA model and other representative tokenizers. Moreover, the segmentation results for general texts are also slightly better than Jieba.
(2) Our experiments solve the problems of ambiguity recognition and OOV recognition without any samples, labels, or high-performance devices. As shown in Table 2, place names such as "Lin Zhou Shi (Linzhou City)" and "Ren Cun Zhen (Rencun Town)", rock names such as "Zan Huang Yan Qun (Zanhuang Group)" and "Xie Chang Shi (plagioclase)", numerical values, and common words received the correct segmentation.
(3) Our algorithm uses dynamic numerical comparison instead of setting a fixed threshold manually, which is good to avoid the randomness in the segmentation based on information entropy. Based on the rapid advantages of statistical methods and the cost advantage without manual labor, large-scale informatization of geological text can be realized in a short time.

In our experiments, we found that the text length has an effect on accuracy, and two objective facts are of interest.

(1) As the text length increases, most of the words are cut correctly and the accuracy will increase; however, more cases of identical end-word concatenation increase too. For example, the forward connection richness of the suffix "Jia Zhuang (Jiazhuang)" is richer than that of complete place names such as "Zhao Jia Zhuang (Zhao Jiazhuang)", "Bai Jia Zhuang (Bai Jiazhuang)", and "Guo Jia Zhuang (Guo Jiazhuang)", resulting in error segmentation of "Zhao/Jia Zhuang (Zhao/Jiazhuang)", "Bai/Jia Zhuang (Bai Jiazhuang)" and "Guo/Jia Zhuang (Guo Jiazhuang)".
(2) There are always terms that appear only once or are monotonically connected in a text of any length, but they appear more often in short texts, which may account for the lower accuracy of shorter texts. For example, sentences such as "Lue Xian Ya Bian La Chang (slightly flattened and elongated)" and "Bu Yi Qu Xiao Huo Geng Ming (not suitable for cancellation or renaming)" have only a fixed type of connection and are easy to be regarded as fake long terms and be segmented incompletely. However, some long words such as "He Nan Sheng/Guo Tu/Zi Yuan/Ting (Henan Provincial Department of Land and Resources)" and "Bei Jing/Da Di/Xi Yuan/Kuang Ye/Guan Li/Ji Shu/You Xian/Gong Si (Beijing Dadi Xiyuan Mining Management Technology Co. LTD)" will be segmented correctly.

## 5. Conclusions and Future Work

In this study, we propose a high-degree-of-freedom-priority candidate term boundary conflict reduction method (HFCR), which solves the problem of OOV recognition and avoids large-scale manual labor costs. Information entropy introduced into HFCR quantifies the internal tightness and external degree of freedom of the candidate term. Dynamic numerical comparisons are adopted instead of setting a threshold manually and randomly, which has the advantage of being fast and avoiding artificial complexity and inaccuracy well. Under the condition of zero-sample, no-label, and no-complex training, experiments show that HFCR achieves an F1-value of more than 95% for single geological text, and the F1-value for Chinese general datasets is higher than 87%.

The fundamental reason for limitations such as identical end-word concatenation and inaccurate segmentation of some sentences is that there are not sufficiently rich connections between candidate terms in one single text. This is an objective problem that is difficult to solve. The future improvement plan mainly focuses on the fusion of multi-text knowledge and is expected to gradually solve the problem by adding more text. Similar to the process of human beings absorbing knowledge continuously, the information entropy of correct terms is constantly updated and adjusted by integrating the domain knowledge of texts so as to achieve the target of correcting word segmentation.

**Author Contributions:** Conceptualization, Y.T. and J.D.; data curation, Y.T.; formal analysis, Y.T. and J.D.; funding acquisition, J.D.; investigation, Y.T.; methodology, Y.T. and J.D.; project administration, J.D.; resources, Y.T.; software, Y.T.; supervision, J.D.; validation, Y.T.; visualization, Y.T.; writing—original draft, Y.T.; writing—review and editing, Y.T., J.D. and Z.G. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Natural Science Foundation of China, grant number 42172330.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The dataset used in this paper can be accessed by https://geocloud.cgs.gov.cn (accessed on 30 March 2023). General datasets can be accessed by https://github.com/tangyudawn/HFCR (accessed on 30 March 2023).

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

## References

1. Wu, C.L.; Liu, G.; Zhang, X.L.; He, Z.W.; Zhang, Z.T. Discussion on geological science big data and its applications. *Chin. Sci. Bull.* **2016**, *61*, 1797–1807. [CrossRef]
2. Ma, K. Research on the Key Technologies of Geological Big Data Representation and Association. Ph.D. Thesis, China University of Geosciences, Wuhan, China, 2018.
3. Chen, J.W.; Chen, J.G.; Wang, C.G.; Zhu, Y.Q. Research on segmentation of geological mineral text using conditional random fields. *China Min. Mag.* **2018**, *101*, 69–74.
4. Wang, M.G.; Li, X.G.; Wei, Z.; Zhi, S.T.; Wang, H.Y. Chinese Word Segmentation Based on Deep Learning. In Proceedings of the 10th International Conference on Machine Learning and Computing (ICMLC 2018), New York, NY, USA, 26–28 February 2018; pp. 16–20. [CrossRef]
5. Xie, X.J.; Xie, Z.; Ma, K.; Chen, J.G.; Qiu, Q.J.; Li, H.; Pan, S.Y.; Tao, L.F. Geological Named Entity Recognition based on Bert and BiGRU-Attention-CRF Model. *Geol. Bull. China* **2021**, 1–13. Available online: http://kns.cnki.net/kcms/detail/11.4648.p.20210913.1040.002.html (accessed on 30 March 2023).
6. Tian, Y.H.; Song, Y.; Xia, F.; Wang, Y.G. Improving Chinese Word Segmentation with Wordhood Memory Networks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020.
7. Zheng, M.G.; Liu, M.L.; Shen, Y.M. An improved Chinese word segmentation algorithm based on dictionary. *Softw. Guide* **2016**, *15*, 42–44.
8. Sun, M.S.; Zuo, Z.P.; Huang, C.N. An experimental Study on Dictionary Mechanism for Chinese Word Segmentation. *J. Chin. Inf. Process.* **2000**, *14*, 1–6.
9. Yang, Y. Mechanical Statistical Word Segmentation System Based on Hash Structure. Master's Thesis, Central South University, Changsha, China, 2005.

10. Mo, J.W.; Zheng, Y.; Shou, Z.Y.; Zhang, S.L. Improved Chinese word segmentation method based on dictionary. *Comput. Eng. Des.* **2013**, *34*, 1802–1807.
11. Rabiner, L.R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proc. IEEE* **1989**, *77*, 257–286. [CrossRef]
12. Low, J.K.; Ng, H.T.; Guo, W.Y. A maximum entropy approach to Chinese word segmentation. In Proceedings of the 4th Sighan Workshop on Chinese Language Processing, Jeju Island, Republic of Korea, 14–15 July 2005; pp. 181–184.
13. McCallum, A.; Freitag, D.; Pereira, F. Maximum Entropy Markov Models for Information Extraction and Segmentation. In Proceedings of the 17th International Conference on Machine Learning, Stanford, CA, USA, 29 June–2 July 2000; pp. 447–454.
14. Zhou, J.S.; Dai, X.Y.; Ni, R.Y.; Chen, J.J. A Hybrid Approach to Chinese Word Segmentation around CRFs. In Proceedings of the 4th Sighan Workshop on Chinese Language Processing, Jeju Island, Republic of Korea, 14–15 July 2005.
15. Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; Kuksa, P. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* **2011**, *12*, 2493–2537.
16. Vincent, P.; Larochelle, H.; Lajoie, I.; Bengio, Y.; Manzagol, P.A. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *J. Mach. Learn. Res.* **2010**, *11*, 12.
17. Cai, D.; Zhao, H. Neural Word Segmentation Learning for Chinese. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; pp. 409–420.
18. Jin, C.; Li, W.H.; Chen, J.; Jin, X.; Guo, Y. Bi-directional Long Short-term Memory Neural Networks for Chinese Word Segmentation. *J. Chin. Inf. Process.* **2018**, *32*, 29–37.
19. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
20. Zhang, X. Research on Chinese Word Segmentation for Construction Domain. Master's Thesis, Beijing University of Engineering and Architecture, Beijing, China, 2021.
21. Liu, S.Q.; Zhou, L.; Li, C.Y.; Zhang, Y.Z.; Li, Y.D. Research on Modeling of Traditional Chinese Medicine Word Segmentation Model Based on Sentencepiece. *World Chin. Med.* **2021**, *16*, 981–985+990.
22. Zhang, X.Y.; Ye, P.; Wang, S.; Du, M. Geological entity recognition method based on Deep Belief Networks. *Acta Petrol. Sin.* **2018**, *34*, 343–351.
23. Wang, H.; Zhu, X.L.; Zeng, T.; Qiao, D.Y.; Guo, J.T. A method of Geologic Words Identification Based on Statistics. *Softw. Guide* **2020**, *19*, 211–218.
24. Chu, D.P.; Wan, B.; Li, H.; Fang, F.; Wang, R. Geological Entity Recognition Based on ELMO-CNN-BILSTM-CRF Model. *Earth Sci.* **2021**, *46*, 3039–3048.
25. Liu, Y.; Zhang, K.; Wang, H.J.; Yang, G.Q. Unsupervised Low-resource Name Entities Recognition in Electric Power Domain. *J. Chin. Inf. Process.* **2022**, *36*, 69–79.
26. Gu, C.; Yu, C.H.; Yu, Y.; Guan, W.W. Unsupervised keyword extraction model for Chinese single text based on BERT model. *J. Zhejiang Sci.-Tech. Univ. Nat. Sci. Ed.* **2022**, *47*, 424–432.
27. Zhang, S.H.; Ye, Q.; Cheng, C.L.; Zou, J. Domain Adaptive Unsupervised Word Segmentation for Traditional Chinese Medicine Ancient Books. *Softw. Guide* **2022**, *21*, 96–100.
28. Zhang, M.; Li, S.; Zhao, T.J. Algorithm of n-gram Statistics for Arbitrary n and Knowledge Acquisition Based on Statistics. *J. China Soc. Sci. Tech. Inf.* **1997**, *16*, 1.
29. Ren, H.; Zeng, J.F. A Chinese Word Segmentation Algorithm Based on Information Entropy. *J. Chin. Inf. Process.* **2006**, *5*, 40–43+90.
30. Chen, G.Q.; Liu, G.N. An Overview on Information Gain-Based GARC-Type Association Classification. *J. Inf. Resour. Manag.* **2011**, *2*, 4–9.
31. Gong, W. Watershed Model Uncertainty Analysis Based on Information Entropy and Mutual Information. Ph.D. Thesis, Tsinghua University, Beijing, China, 2012.
32. Zheng, J.S. An Image Information Entropy-Based Algorithm of No-Reference Image Quality Assessment. Master's Thesis, Beijing Jiaotong University, Beijing, China, 2015.
33. Meng, Y.Y.; Guo, J. Link prediction algorithm based on information entropy improved PCA model. *J. Comput. Appl.* **2022**, *42*, 2823.
34. Gao, J.Q.; Zhao, Q.C. Study on word Segmentation Method of classical Literature Based on New Word Discovery. *Comput. Technol. Dev.* **2021**, *31*, 178–181+207.
35. Zhang, G.D.; Yang, C.; Zhan, X.L.; Fang, H.; Wang, J.F. An Identification Method Text of Overall Commendatory and Derogatory Tendency of Sentences Based on Information Entropy. *Microcomput. Appl.* **2021**, *37*, 12–15.


**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.