Earth Science Informatics (2022) 15:21–36
https://doi.org/10.1007/s12145-021-00695-2

RESEARCH ARTICLE

What is this article about? Generative summarization with the BERT model in the geosciences domain

Kai Ma[1] · Miao Tian[1] · Yongjian Tan[1] · Xuejing Xie[2] · Qinjun Qiu[2,3]

Received: 24 May 2021 / Accepted: 27 August 2021 / Published online: 22 September 2021

Abstract

In recent years, a large amount of data has been accumulated, such as those recorded in geological journals and report literature, which contain a wealth of information, but these data have not been fully exploited or mined. Automatic information extraction offers an effective way to achieve new discoveries and pursue further analysis, which is of great significance for users, researchers or decision makers to aid and support analysis. In this paper, we utilize the bidirectional encoder representations from transformers (BERT) model, which is fine-tuned and then applied to automatically generate the title of a given input summarization based on the collection of published literature samples. The framework contains an encoder module, decoder module and training module. The core stages of summary generation involve the combination of encoder and decoder modules, and the multi-stage function is then used to connect modules, thus endowing the text summarization model with a multi-task learning architecture. Compared to other baseline models, our proposed model obtains the best results on the constructed dataset. Therefore, based on the proposed model, an automatic geological briefing generation platform is developed and used as an online platform to support the excavation of key areas and a visual presentation analysis of the literature.

Keywords Geological domain · Fine-tuned BERT model · Automatic text summarization · Briefing generation framework

Introduction

The exponential growth of natural language text data in the domain of geosciences (i.e., geological documents/reports and a variety of journal literature) has produced a rich data source for geological research. Both the U.S. Geological Survey and the Chinese Geological Survey have accumulated a large amount of geological data, and a large amount of unstructured data, such as geological survey reports and work records, has not been fully utilized and mined (Qiu et al. 2018a). They cover a variety of geological topics, ranging from descriptions of geological literature, rocks, minerals, and geological age to evolutionary patterns and geological significances (Qiu et al. 2019). For example, to attract attention from potential readers, most journalists' articles and geological documents/reports attempt to report and present the core content and information. However, the rich geological data that they contain remain underutilized but could be leveraged in several applications to offer better services and usages (Ma et al. 2020; Wang et al. 2021).

Automatic text summarization (ATS), as an important category of information extraction, is a challenging task in natural language processing (NLP) (Narayan et al. 2018a, 2018b, 2018c). It aims to facilitate the task of quickly reading and searching for information in large documents/reports by generating reduced documents without loss of meaning. This enables researchers to find relevant information quickly and accurately in massive amounts of data to address the contradiction between the existence of a very large amount of data with low content density and the needs of users to obtain efficient and accurate information. In the domain of geosciences, with the rapid growth of textual data, high-quality textual data analysis methods have become necessary to remedy information content overload since it is impossible to manually obtain text summaries from massive

Communicated by: H. Babaie

✉ Qinjun Qiu
qiuqinjun@cug.edu.cn

[1] College of Computer and Information Technology, China Three Gorges University, Yichang 443002, China

[2] National Engineering Research Center of Geographic Information System, Wuhan 430074, China

[3] School of Geography and Information Engineering, China University of Geosciences, Wuhan 430074, China

geological text data (Hunter et al. 2012). Such approaches can be used to allow users to make critical decisions by querying and finding the most salient core information quickly without examining the whole document/report.

Existing research on this issue, however, offers only partial solutions. First, most existing text summarization approaches are not suitable for Chinese text data. The grammar and semantics of Chinese texts are much more complex than those of English texts, especially in the field of geosciences, where the presence of a large number of specialized words further increases the difficulty. The Chinese language is not based on words, while English is word-based; moreover, in Chinese, there are no separators between words. In Chinese, determining what constitutes a word is difficult; therefore, word segmentation is a necessary first step to determine the boundaries between consecutive words (Qiu et al. 2018a, b). Therefore, ambiguity exists in Chinese texts. Second, the majority of text summarization methods depend on traditional bag-of-words representations that involve high-dimensional and sparse data, and it is difficult to represent and capture relevant information (Hou et al. 2017; Liu 2019; El-Kassas et al. 2020). Third, the recent studies of Chinese text summarization are still in their early stages; the literature that addresses this subject area in Chinese is fairly scant and has only recently become comparable to that on other languages. In addition, Chinese text summarization has not yet reached the same level of maturity and reliability as that for English. Therefore, there is an urgent need to develop systems/models for the analysis and summary of large amounts of Chinese texts that continue to grow.

In recent years, people have attempted to use deep learning to model and have achieved good results in a variety of NLP tasks. The current popular deep language representation models are embeddings from language models (ELMo) (Peters et al. 2018), universal language model fine-tuning (ULMFiT) (Howard and Ruder 2018), generative pre-training (GPT) (Radford et al. 2018) and bidirectional encoder representations from transformers (BERT) (Devlin et al. 2018). ELMo is based on a language model. Specifically, it is a bidirectional language model and a bidirectional long short-term memory (LSTM) structure. UMLFiT, like ELMo, uses a recurrent neural network. The pre-training process of GPT is similar to that of ELMo, but the difference is that the feature extractor of GPT uses a transformer, and ELMo can use context to predict words, while GPT uses the above to predict words. The BERT model overcomes the shortcomings of ELMo, ULMFiT and GPT. It uses a two-way converter and a simple fully connected network based on a specially designed attention mechanism to replace the complex convolutional neural network (CNN) and recurrent neural network (RNN). It can not only greatly reduce the two-way modelling of the text but also effectively improve the network performance. The BERT model has the following state-of-the-art performance: (1) When using the BERT model, our text does not require labels and the transformer encoder stack can be used to directly train a large amount of text. Additionally, such an iterative method of generating representations could allow the model to obtain a substantial amount of language information. (2) In the case of BERT, task-specific layers are added on top of BERT, and the entire model is fine-tuned based on task-specific labelled datasets. (3) BERT is based on bidirectional and context-sensitive representations. Since the BERT model learns a large amount of linguistic information during unsupervised pre-training, it can be fine-tuned even on small datasets. Therefore, it performs better than CNN- or RNN-based models that need to be trained from scratch.

In this paper, we use BERT to develop an efficient and effective automatic summarization system. We present a new Chinese text summarization method based on the bidirectional LSTM (BiLSTM) model. Figure 1 shows the addition of the collected corpus to the BERT pre-training model, which makes full use of context information to enhance the semantic representation of the word vector and obtain a better word vector representation, the input of the generated word vector into the seq2seq model for training, and finally the formation of an automatic text summary. We use the bilingual evaluation understudy (BLEU) and recall-oriented understudy for gisting evaluation (ROUGE) indicators to evaluate the experimental performance of the model. The experimental results show the effectiveness of the proposed model in automatically generating the title of a given input summarization. To the best of our knowledge, this is the first work to provide text summarization in the domain of geosciences based on a deep learning approach.

The main contributions of this research are summarized as follows:

(1) From the perspective of the algorithm, the BERT model is fine-tuned and then applied to automatically generate the title of the input summarization based on the collection of published literature samples. Compared to a list of baseline models, the proposed model obtains the best results considering the evaluation metrics.
(2) From an application perspective, we design a framework for the automatic generation of journal presentations that provides timely and effective data support to decision makers to formulate more specific decisions and to users to quickly navigate and search for information.
(3) We build a large-scale corpus in the domain of geosciences by collecting four major journals: Geological Review, Geological Journal, Mineral Deposit Geology, and Chinese Geology.

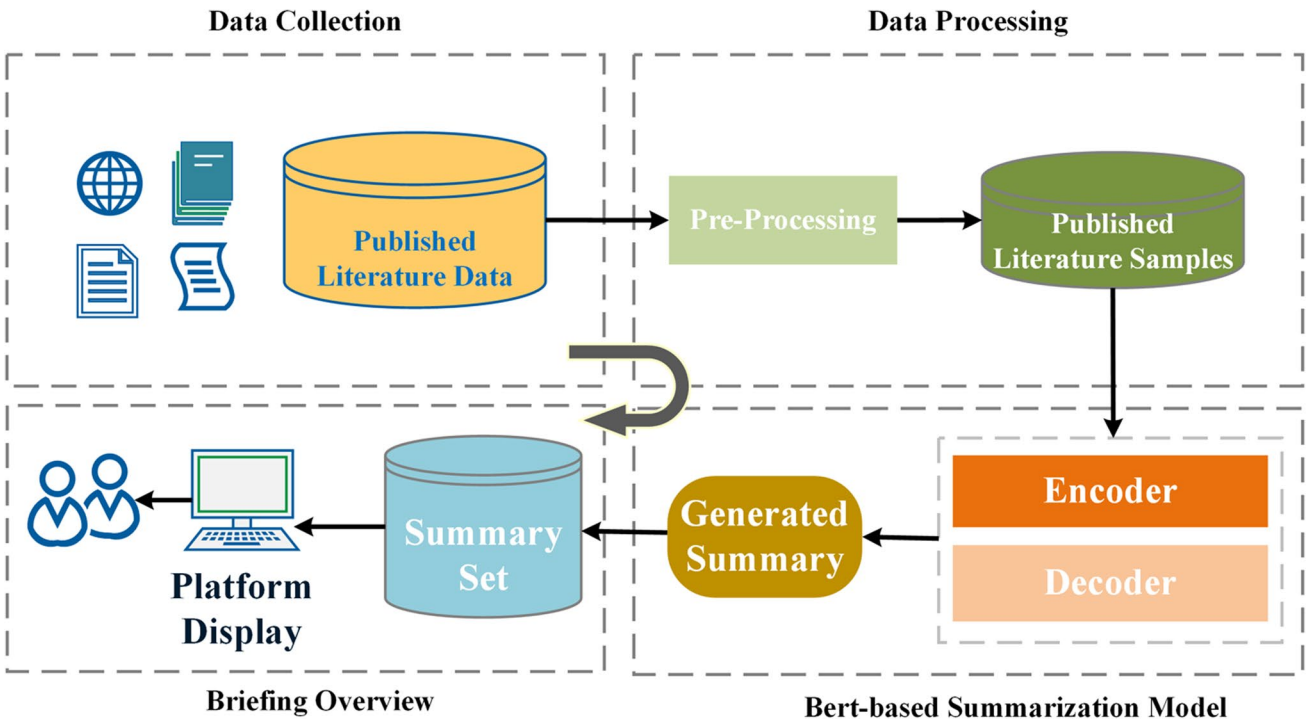

Fig. 1 Overall framework of text summarization classified into four stages: data collection, data processing, BERT-based summarization model and briefing overview. The data collection stage focuses on collecting published literature data from Chinese journals; the data processing stage aims to filter and clean the collected published data for further analysis; the BERT-based summarization model is used to generate a geoscience domain-specific summary based on the input published data; and the briefing overview stage aims to combine the generated summaries to develop an overview, which is displayed to users. The U-shaped arrow indicates the overall framework flow, including data collection, data processing, BERT-based summarization model and briefing overview

The remaining chapters of this article are organized as follows: in Section 2, we briefly introduce the method of automatic abstract generation and the development of automatic abstract generation in various fields. In Section 3, we mainly introduce the constructed dataset and the model used in the experiment. In Section 4, we present the content and results of the experiment. Finally, Section 5 demonstrates the application and platform, and Section 6 draws conclusions and discusses future work.

Related work

Approaches to text summarization can be classified into three categories: extractive text summarization approaches, abstractive text summarization approaches and hybrid text summarization approaches.

Extractive text summarization

Extractive text summarization evaluates the importance of a sentence by considering the position of the sentence in the original text, word frequency, keywords, etc., and then selects the sentence with a high importance from the original text to form the abstract. At present, most extractive summaries use sentences as the extraction unit to achieve better summarization effects (Ceylan et al. 2010), mainly considering the relevance of the summary and the redundancy of sentences (Carbonell and Goldstein 1998; Radev 2004). Relevance mainly refers to whether the generated abstract is consistent with the meaning of the source document, and redundancy is considered to evaluate the amount of redundant information among the candidate sentences.

Figure 2 shows the architecture system of extractive text summarization, which is mainly divided into three parts: (1) pre-processing of the input text (El-Kassas et al. 2020); (2) post-processing, such as reordering the extracted sentences; and (3) the following processing tasks: first, a suitable representation is created of the input source text to facilitate text analysis (Joshi et al. 2018), the sentence is then scored and sorted according to the input text representation (Nenkova and Mckeown 2012), and the most important sentences are finally selected and connected to create a new abstract (Shuai et al. 2017). In recent years, extractive text summarization has been applied in many fields and has achieved good results (Zhou et al. 2018). Mao et al. (2019) proposed a single-document summary extraction method that combines supervised learning and unsupervised learning. Wang et al.

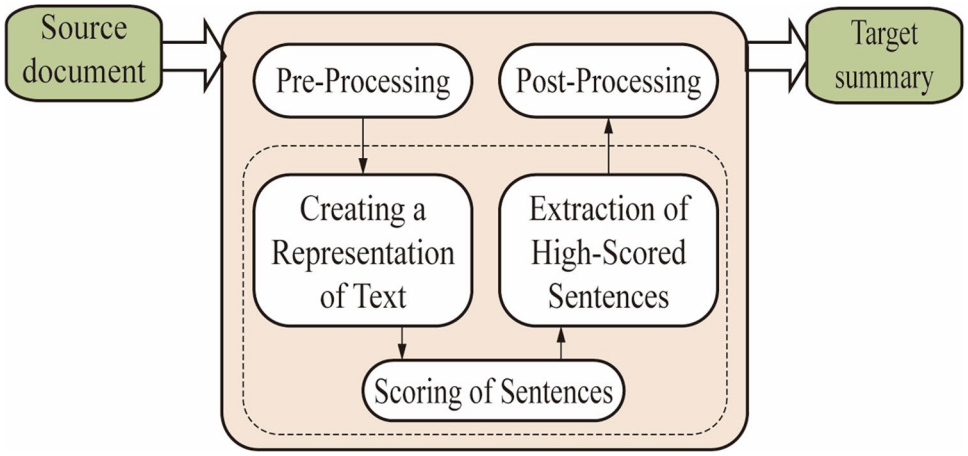

Fig. 2 Architecture of the extractive text summarization system

(2021) faced difficulties in extracting long text summaries and proposed an extraction method for long Chinese documents. Their experiment showed that the presented BERT method improves the accuracy and reduces the redundancy in the extractive summarization process of long Chinese texts.

The extractive text summarization methods are faster and simpler than the abstractive methods because such methods are based on the direct extraction of sentences from the original texts, and the extracted terms exist in the texts. However, this approach is far from the approach designed by human experts. The main reasons include redundancy from some extracted sentences, lack of semantics and cohesion and inability to cover important information (El-Kassas et al. 2020).

Abstractive text summarization

Abstractive text summarization requires a deeper analysis of the input source text. Abstractive text summarization is mainly based on understanding the semantics of a given article, compressing and refining sentences at the word level, and finally generating abstracts with fewer words and clearer language (Al-Abdallah and Al-Taani 2017). Summary generation has the ability to generate new sentences rather than simply copying sentences from the source document (Bhat et al. 2018).

Figure 3 shows the structural system of abstractive text summarization, which is mainly divided into three parts: (1) pre-processing: (1) of the source text, (2) post-processing, and (3) the following processing tasks: an internal semantic representation is first built, and natural language generation technology is then applied to generate a summary that is more in line with the understanding of people (El-Kassas et al. 2020).

With the advent of deep learning, an increasing number of deep learning methods have been used in abstractive text summarization. Sutskever first proposed the seq2seq architecture (Sutskever et al. 2014). The basic structure is composed of an encoder and a decoder, and both the encoder and the decoder are implemented in a neural network. At this stage, abstractive text summarization mainly relies on the seq2seq architecture. Nallapati et al. (2016a, 2016b) integrated the attention mechanism into a model, which can assign different weights to the input vector representation at each moment. Hu et al. (2015) proposed an RNN-based seq2seq architecture with an attention mechanism using a large-scale Chinese short text summary dataset based on

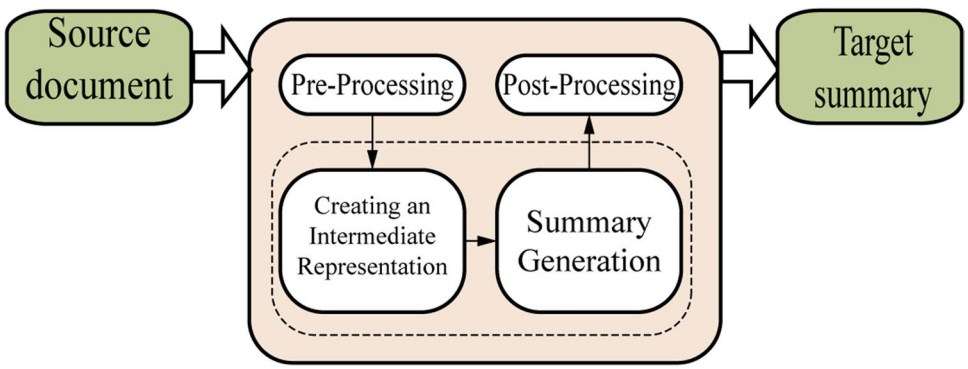

Fig. 3 Architecture of the abstractive text summarization system

Sina Weibo and achieved good results, providing a basis for further research on short text summaries. Siddiqui and Shamsi (2018) developed a local attention mechanism based on the seq2seq architecture and achieved good results in solving the problem of repeated words. Shashi Narayan et al. (2018a, 2018b, 2018c) used the seq2seq architecture based on convolutional neural networks to automatically generate abstracts of BBC online articles. TAN et al. (2017) proposed a graph-based attention mechanism neural model, which has achieved good results in the automatic generation of abstracts. Liang et al. (2020) aimed to generate a short version of a given sentence and proposed a selective reinforcement seq2seq attention model to abstract social media text summaries and combined cross-entropy and reinforcement learning strategies to directly optimize the ROUGE score.

However, there are also many problems in abstractive text summarization. If the length of the source text is very long, a length-dependent problem occurs, and information may be lost, resulting in the generated summary not being sufficiently accurate. Therefore, Rush et al. (2015) used a convolution model to encode the original document and applied a context-sensitive attention feedforward neural network to generate summaries. Adding the attention mechanism can generate more targeted summaries. Zhang et al. (2019) proposed a seq2seq model based on the BERT model, which integrates the BERT model into the summary generation task so that better summary information is generated.

Hybrid text summarization

Extractive text summarization uses sentences as the extraction unit. The generated abstracts are relatively smooth but can only use sentences from the original text and contain redundancy, while abstractive text summarization can generate abstracts on the basis of understanding the semantics of the text, but part of the information may be lost in the process of generating the abstract, resulting in an inaccurate abstract. In view of the advantages and disadvantages of these two text summarization methods, hybrid text summarization has been proposed (Chen and Bansal 2018). Aiming at the advantages and disadvantages of extractive and abstractive text summarization methods, Qiu et al. (2019) proposed a hybrid text summarization model based on reinforcement learning, which combines the advantages of extractive and abstractive text summarization and ensures the language fluency and simplicity of the final summaries. Lu et al. (2020) proposed a three-stage composite text summarization model based on pre-training by combining the extraction method and the generation method. Combining the extraction method and the generative method, the two-way contextual information word vector generated by the source text after pre-training is obtained by the sigmoid function to obtain the sentence score to extract the key sentence, and the key sentence is rewritten as a cloze task in the abstract generation stage to generate the final abstract and achieve very good results.

Methods

Pre-training with BERT

The BERT model is a new language representation model that can be used to perform unsupervised pre-training using a large number of texts (Devlin et al. 2018). Over the past few years, the BERT model has performed relatively well in natural language processing. The BERT model contains a two-way transformer encoding layer, and the model is designed to pre-train the deep bidirectional representation of unlabelled text by conditional pre-processing on the upper left and right of all layers (Liu 2012), thus enabling modification. It can better capture the two-way relationship in a given sentence.

To cope with different task input requirements, the BERT model can simply enter a sentence or combine two sentences. The input structure is shown in Fig. 4. The input of the BERT model is mainly divided into three layers, namely, the token embedding layer, segment embedding layer and position embedding layer. The token embedding layer converts each word into a fixed-dimensional vector. In the BERT model, each word is converted into a 768-dimensional vector representation, where the first input text must be tokenized before being sent to the token embedding layer, while the method used for tokenization is WordPiece tokenization. The segment embedding layer is used to distinguish two kinds of sentences. In addition to the masked language model, pre-training must also perform a classification task of judging the order of any two sentences. Each token of the previous sentence is represented by 0; each token of the latter sentence is represented by 1, etc. The position embedding layer is obtained through training in BERT. In particular, an input sequence of length n yields three different vector representations:

1. Token embedding, (1, n, 768). Vector representation of words.
2. Segment embedding, (1, n, 768). Assists BERT to distinguish the vector representation of the two sentences in a given sentence pair.
3. Position embedding, (1, n, 768). BERT enables learning the sequential attributes of the input.

The BERT model is an unsupervised NLP pre-training model, and self-attention is an important idea of BERT. Combined with position coding, this concept solves the problem of temporal correlation of text data. In layman

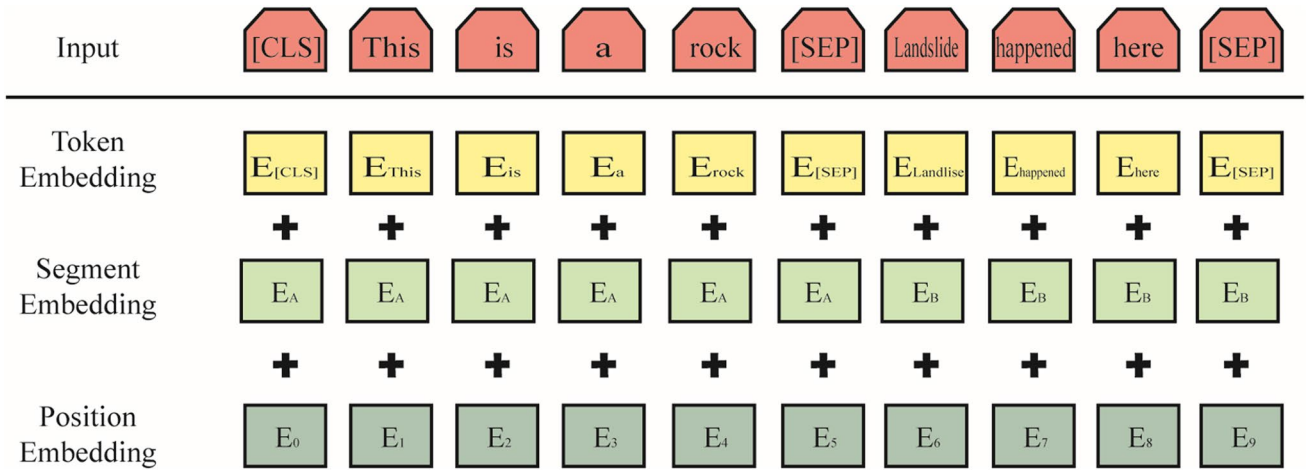

Fig. 4 Input to the BERT model

terms, self-attention is a way of dynamically calculating weights when information is propagated forward. The implementation steps are as follows:

1. x^1, x^2, x^3, *and* x^4 represent 4 input sentences. Each input sentence is embedded and multiplied by a matrix to become a^1, a^2, a^3, *and* a^4:

$$a^i = Wx^i \tag{1}$$

2. Each a^i is multiplied by three different transformations to generate three vectors, which are q, k, and v, and the dimensions of these three vectors are the same.

$$q^i = W^q a^i \tag{2}$$

$$k^i = W^k a^i \tag{3}$$

$$v^i = W^v a^i \tag{4}$$

3. Each q is considered to pay attention to each k. The scaled dot product is used in the attention algorithm in the self-attention layer.

$$a_{1,k} = q^1 \bullet k^i / \sqrt{d} \tag{5}$$

In the above equation, d is the dimension of vectors q and v.

1. $a_{1,1}$ to $a_{1,4}$ are normalized through the softmax layer to obtain $\hat{a_{1,1}}$ to $\hat{a_{1,4}}$.
2. v^1 to v^4 are multiplied and added to $\hat{a_{1,1}}$ to $\hat{a_{1,4}}$ to obtain the first output vector b_1.
3. The above steps are repeated to obtain b_2, b_3, and b_4.

In addition, the BERT model also introduces multi-head self-attention. The equation is as follows:

$$head_i = Attention\left(QW_i^Q, KW_i^k, VW_i^v\right) \tag{6}$$

$$Multi - head(Q, K, V) = Concat_i\left(head_i\right)W^o \tag{7}$$

Overall architecture

In this research, the overall architecture of text summarization is shown in Fig. 5, which contains an encoder module, decoder module and training module. The core stages of summary generation are the combination of encoder and decoder modules, and the multi-stage function is then used to connect modules, which endows the text summarization model with a multi-task learning architecture.

Encoder

As shown on the left side of Fig. 5, the encoder stage is composed of four parts: a fine-tuned BERT-embedding component, a BiLSTM component, a set of convolutional gated units, and a self-attention mechanism. The object of the fine-tuned BERT-embedding component is to initialize the value of the input sequence based on word embedding. With the input sequence, the BiLSTM component is applied to encode the input received from the previous layer. The convolutional gated units focus on retraining the core information based on the previous output at each time step. The self-attention mechanism aims to explore the links between notes and further intensifies the global information thereafter.

Let $X_{1:T} = (x_1 \oplus x_2 \oplus x_3 \ldots x_T) \in R^{1 \times T}$ be the input sequence, where the length is T and x denotes the basic unit (i.e., word) in the input sequence and $\oplus$ denotes the

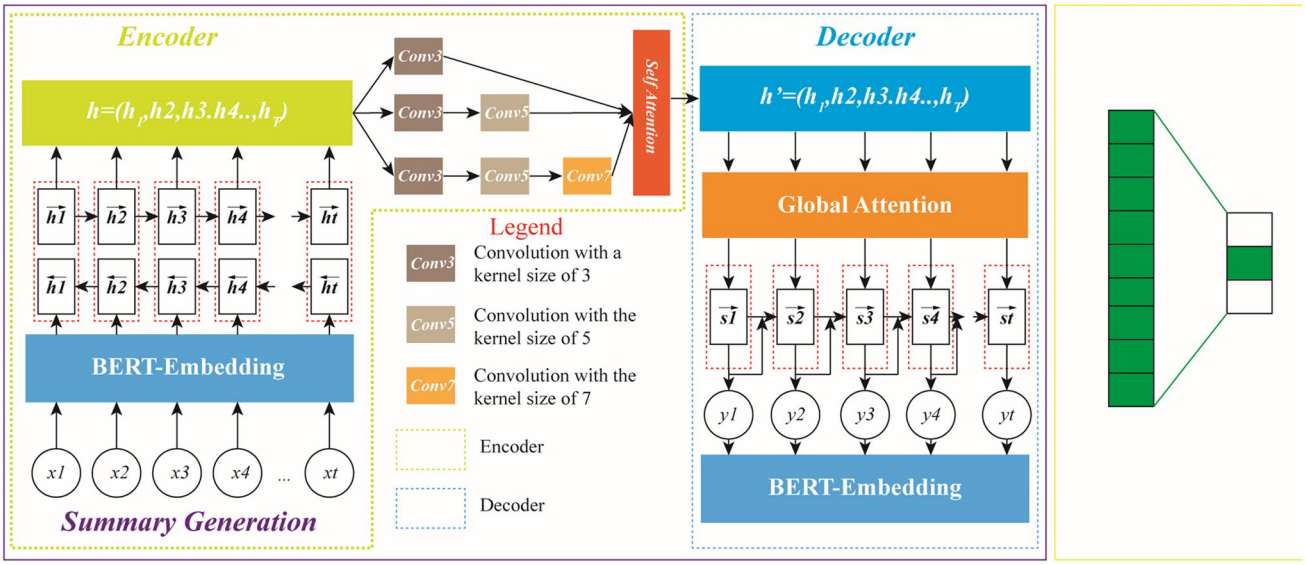

Fig. 5 Overall structure of the proposed text summarization model, which consists of two core modules: encoder and decoder modules

concatenation operator. With the source text as the input, the basic units are converted into vectors based on word embedding operations. This process is initialized by using the fine-tuned BERT model, which is further fine-tuned in the training process. Herein, we use the "BERT-Base Chinese" model to fine-tune the input sequence.

Let $h = GLU(h_1, h_2, h_3, \ldots h_t) \in R^{T \times \dim}$ be the content output of the processing step of the bidirectional LSTM from left to right and from right to left. GLU(.) denotes the gated linear unit, and h_i denotes the hidden state of the encoder with time step t, which can be represented as $h_i = [\overrightarrow{h_i}; \overleftarrow{h_i}] \in R^{1 \times 2\dim}$.

In regard to the words in the input sentence, a local link among words exists in natural language. In this research, a set of convolution operations is applied to capture these links, specifically obtaining the n-gram features. Moreover, filters and the receptive field are applied to grasp the richer local link features based on considering the average length of the input sequence X. Then, the two vectors $q \in R^{|q|}$ and $w^k \in R^{k \in \{3,5,7\}}$ are convolved to form a feature map $m \in R^{|q|-k+1}$, and the calculation can be expressed as follows:

$$m_i = f\left(w^k * q\right)_i = f\left(w^k * q_{[i:i+k-1]}\right) = f\left(\sum_{j=1}^{i+k-1} w_j^k q_j\right) \quad (8)$$

where f denotes the rectified linear unit (ReLU) function.

With the output retrieved from the CNN, a self-attention mechanism is used to mine and capture the above global links. Based on the self-attention mechanism, this enables the model to learn long-term dependencies, which does not consume excessive computational resources. Therefore, the mechanism is used to develop the link between global information and annotations at each time step as follows:

$$self - Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}V\right) \quad (9)$$

where Q and V denote the matrices generated by the CNN, and K denotes a learnable matrix.

The CNN module can be used to obtain the n-gram features of the input sequence X, and the self-attention mechanism can capture the long-term dependencies of the input sequence X. Herein, a gated unit is employed to carry out global encoding, which generates the output of the encoder. The content is output, which is expressed as follows:

$$\tilde{h} = h \odot \sigma(g) \in R^{T \times \dim} \quad (10)$$

Decoder

The context vector c output by the encoder section encodes the entire input sequence, and the decoder section decodes the information contained in context variable c to generate an output sequence. As shown on the right side of Fig. 5, the decoder stage is composed of three parts: global attention, bidirectional LSTM, and fine-tuned BERT model.

Global attention considers the hidden state of the ALL encoder to define an attention-based context vector at each decoder step, which ensures that the sequence encodes the encoder into a different c at each time step, and when decoding, combines each different c to decode the output to yield a more accurate result. The BiLSTM can decode

the sentences output by the encoder, predict the next word according to the given semantic vector c and the output sequence, and obtain the output sequence, while the fine-tuned BERT model can effectively capture the contextual relationship among sentences, generating a complete context at every moment, which makes the generated summary more accurate.

The main idea of global attention is to consider the hidden layer state of all encoders, where a_t is an alignment vector with a variable length, and the length is that of the encoder part of the time sequence. This approach compares the current hidden state h_t of the decoder to the hidden layer state $a_t(s) = align\left(h_t, \overline{h_s}\right) = \frac{\exp\left(score\left(h_t, \overline{h_s}\right)\right)}{\sum_{s'} \exp\left(score\left(h_t, \overline{h_{s'}}\right)\right)}$

Here, the score is a content-based function, which can be implemented in the following three ways:

$$score = {h_t}^T \overline{h_s} \tag{12}$$

$$score = {h_t}^T W_a \overline{h_s} \tag{13}$$

$$score = {v_a}^T \tanh\left(W_a\left[h_t; \overline{h_s}\right]\right) \tag{14}$$

All $a_t(s)$ values are integrated into a weight matrix, W_a is obtained, and the following is calculated:

$$a_t = softmax\left(W_a h_t\right) \tag{15}$$

A weighted average operation is performed on a_t to obtain the context vector c_t.

The BiLSTM can be a combination of a forward LSTM and a backward LSTM. The detailed calculation process is as follows:

1. Calculate the forget gate and select the information to be forgotten

Input: the hidden layer state h_{t-1} at the previous moment, and the input value at the current moment is X_t.

Output: the value f_t of the forget gate.

$$f_t = \sigma\left(W_f \bullet \left[h_{t-1}, x_t\right] + b_f\right) \tag{16}$$

2. Calculate the memory gate and select the information to be memorized

Input: the hidden layer state h_{t-1} at the previous moment and input word X_t at the current moment.

Output: memory gate value i_t and the temporary cell state.

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right) \tag{17}$$

$$\tilde{C}_t = \tanh\left(W_c \cdot \left[h_{t-1}, x_t\right] + b_C\right) \tag{18}$$

3. Calculate the cell state at the current moment

Input: memory gate value i_t, forget gate value f_t, temporary cell state C_t, and cell state C_{t-1} at the previous moment.

Output: current cell state C_t.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{19}$$

4. Calculate the output gate and the state of the hidden layer at the current moment

Input: the hidden layer state h_{t-1} at the previous moment, the input word X_t at the current moment, and the cell state C_t at the current moment.

Output: the value of the output gate is o_t, and the state of the hidden layer is h_t.

$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right) \tag{20}$$

$$h_t = o_t * \tanh\left(C_t\right) \tag{21}$$

Experiments and results

In this subsection, a set of primary experiments is conducted to validate the effectiveness of the proposed text summarization model. First, quantitative evaluation based on an evaluation metric is used to evaluate the performance of the proposed model. Second, we conduct qualitative evaluation experiments to test the performance of generative text summarization.

Dataset and evaluation metric

Dataset

The dataset used in our experiments in this research is mainly retrieved from published literature data that include the Geological Bulletin of China, Acta Geological Sinica, Geological Review and Mineral Deposits. We collected abstracts and titles from the published literature and then cleaned the collected data by using the process of data pre-processing. Therefore, the open-access Jieba (fxsjy 2018) segmentation library was applied to decompose the original Chinese texts into word sequences for further analysis.

Regarding data pre-processing, a list of rules is developed to clean and format the corpus and then develop the dataset. For example, author names and publication information are filtered out directly, such as Wang et al. 2021, 67 (1): 1–12. DOI: https://doi.org/10.16509/j.georeview.2021.01.001". Since the length of the abstract and title may influence the performance of the model, we obtained statistics on the length distribution of the abstracts and titles, as shown in Fig. 5. Any data beyond the range length are not shown in the figure. Additionally, the corresponding histogram of the dataset of the collected titles and abstracts is shown in Fig. 6,

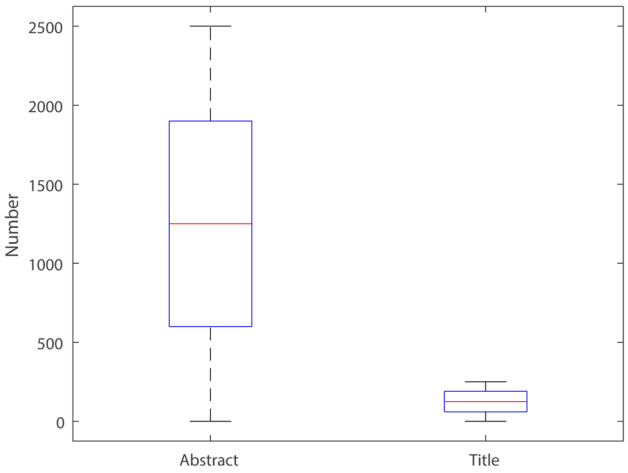

Fig. 6 Text length distributions of the abstracts and titles as a box plot

in which the length of the title is $len_{title} \in [10,70]$ and the relevant abstract length is $len_{abs} \in [200,1100]$. In this paper, the intent of the simple characterization statistics for the dataset is to visualize the characteristics of the constructed dataset, and in future work, we will test and analyse headings of different lengths (Fig. 7).

An example of the developed dataset after the above pre-processing is provided in Table 1, and we finally obtain approximately 3315 records, of which 80% of the dataset is applied for training, 10% of the dataset is applied for validation, and the remaining 10% of the dataset is applied for testing.

Evaluation metric

In this research, the ROUGE metric (Lin 2004) is applied to evaluate the text summary generation performance. This method calculates the ratio between the generated summary ($\hat{Y}$) and the original summary (Y) and assesses the similarity between them. Three metrics, namely, ROUGE-1, ROUGE-2 and ROUGE-L, are used to evaluate the effectiveness of the proposed model.

ROUGE-1 aims to calculate the unigram overlap score between $\hat{Y}$ and Y, and ROUGE-2 is used to compute the bigram overlap score between $\hat{Y}$ and Y. In this research, the recall of ROUGE-1 and ROUGE-2 is selected and calculated as follows:

$$ROUGE-N = \frac{\sum_{S \in Y} \sum_{gram_n} Count_{match}(gram_n)}{\sum_{S \in Y} \sum_{gram_n \in S} Count(gram_n)} \tag{22}$$

ROUGE-L is computed based on the longest overlap subsequence between the generated summary ($\hat{Y}$) and the original summary (Y), and it is calculated by the following equation:

$$R_{lcs} = \frac{LCS(Y, \hat{Y})}{|Y|} \tag{23}$$

$$P_{lcs} = \frac{LCS(Y, \hat{Y})}{|\hat{Y}|} \tag{24}$$

$$F_{lcs} = \frac{(1+\beta^2) R_{lcs} P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}} \tag{25}$$

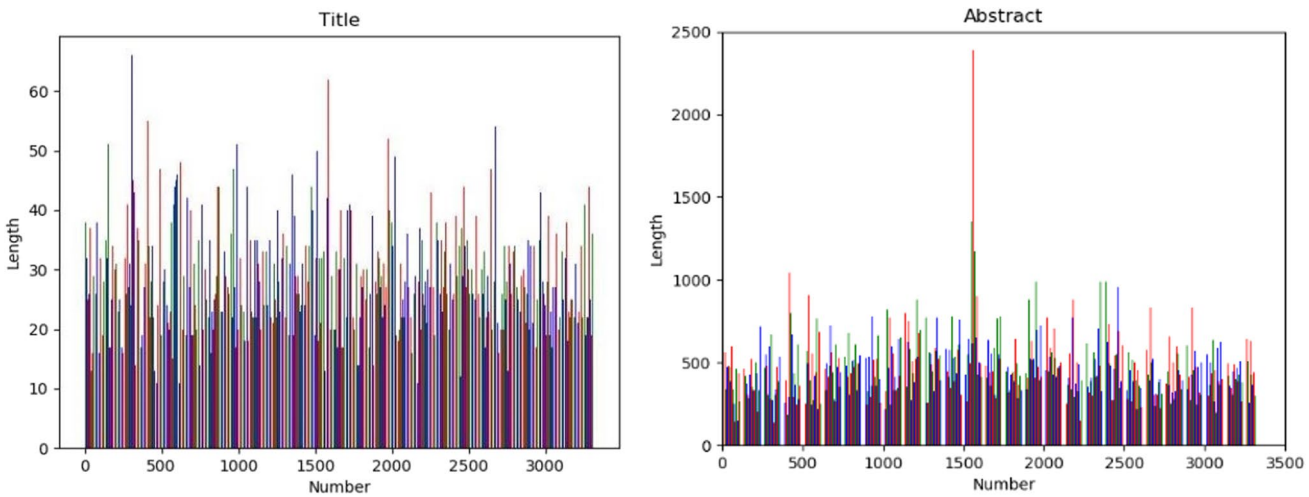

Fig. 7 Histogram of the text length of the collected titles and abstracts in the dataset

Table 1 Example from our extreme summarization dataset showing a document and its one-line summary

No.	Source	Title
1	In this study, based on the geological and geophysical studies carried out by the authors in this region for more than ten years, we found that an anomalous zone with a low velocity and high Poisson's ratio occurs in the lower crust of the eastern edge of the Tibetan Plateau, and the anomalous body is associated with soft fluvial thermal material stemming from the upwelling of the Tibetan Plateau. The anomalies converge with the soft fluvial thermal material originating from the upper Tibetan Plateau, resulting in significant changes in the deep structure of the lower crust and upper mantle from the western edge of Yangzi to the Tibetan Plateau. Along the Longmenshan fault zone, there are interlocking low-velocity (high-Poisson ratio) and high-velocity (low-Poisson ratio) regions in the middle and lower crusts, and these deep structural distribution features are in good agreement with surface deformation and foreland basin uplift and depression patterns.	Deep structure and crustal deformation in the eastern margin of the Qinghai-Tibet Plateau-Yanzi Tethys tectonic domain
2	To explore and improve the comprehensive airborne physical prospecting deep-searching model, on the basis of the latest airborne data in the south Qinling study area, combined with the characteristics of lithology, tectonics and surrounding mineralization points, the upper Gaojiazhuang magnetic and discharge composite anomaly was selected according to certain conditions for key research. The selected anomaly is located in the north of Shiquan County, Ankang city, Shaanxi Province, and is known to be a metamorphic sandstone on the northeast side and a mica-quartz schist on the southwest side based on its ground magnetic, γ-energy spectrum, rock (ore) properties and other profiling and basic geological work. It is assumed that the uranium anomaly is mainly caused by metamorphic sandstone, and the magnetic anomaly is caused by the regional metamorphic schist (the original rock is volcanic) occurring at a certain burial depth.	Geological genesis and mineralization potential of the aerial physical anomaly at Gaojiazhuang in the South Qinling Mountains
3	The Li Dangjia-Mashan paleobulge is a pre-Laiyang Group sedimentary-phase bulge in the Jiao Lai Basin, which is notably covered by the Quaternary System and has been a weak link in the study of the Jiao Lai Basin. This feature has long failed to reveal its stratigraphic and tectonic aspects. In this paper, through a large number of surface geological surveys and the interpretation of comprehensive logging data from the Jiao-Sen 2 and Jiao-Sen 3 wells above the Li-Danjia-Mashan paleo-bump, we applied the principles of stratigraphic stratigraphy, sedimentary phase analysis and petrographic paleogeography to describe and compare the sedimentary filling sequences in the study area, clarified the Cretaceous stratigraphic sequence framework in the study area, and recovered the Li-Danjia-Mashan paleo-bump Cretaceous stratigraphy. The Cretaceous lithological paleogeography of the Li-Danjia-Mashan paleobulge was recovered. The Lidanjia-Mashan paleoconvex was not deposited during the sedimentary period of the Laiyang Group, whereas the Qingshan Group was deposited during the sedimentary period, and the Houkuang and Bamudi Groups and the Xingezhuang Group of the Wang Group were deposited continuously. The study area is located in a volcanic depression, and the whole comprises a set of giant-thick terrestrial fine clastic sediments, interspersed with a small amount of combined volcanic rock-volcanic clastic sediments, in which the volcanic sedimentation indicates four phases of volcanic activity.	Stratigraphic sequence and petrographic paleogeography of the Cretaceous rocks in the Li-Danjia-Mashan paleo-bump in the central Jiao Lai Basin

Table 1 (continued)

No.	Source	Title
4	The Earth experienced at least two global-scale ice ages in the late Neogene, namely, the Sturtian and Marinoan ice ages. The South China System Gucheng and Nantuo formations in the Walking Horse area of West China are Sturtian and Marinoan ice age deposits, respectively, and the Datangpo Formation is an interglacial deposit. In this paper, we studied the elemental geochemical characteristics of fine clastic rock samples of the Datangpo Formation retrieved from the core of borehole ZK701 in the Walking Horse area and calculated paleoclimatic proxies, such as the chemical alteration index (CIA), chemical weathering index (CIW), n(K)/n(Na), n(Mg)/n(Ca), and Rb/Sr, to explore the paleoclimatic evolution during the South China Age interglacial period in the source area of the study area. The results showed that the climate in the source area was cold and dry during the late interglacial period of the ancient city (CIA values of 57.1 and 58.1 for two samples), remained cold during the early interglacial period of Datangpo (the CIA values ranged from 56.5 to 64.6, mean value of 59.8), and returned to warm and humid conditions during the middle and late interglacial periods of Datangpo (the CIA values ranged from 69.8 to 78.8, mean value of 75.5). The CIW, n(K)/n(Na K)/n(Na), n(Mg)/n(Ca), and Rb/Sr values reflect a paleoclimatic evolution process consistent with the CIA values.	Paleoclimatic study on the interglacial period of the South China Age Datang Slope in the Walking Horse area of West China
5	The Little Ice Age was one of the important events of global climate change in the past one thousand years. There are many comparative studies on climate change in the Chinese monsoon and westerly influenced regions during the Little Ice Age, but there is a lack of studies on the spatial and temporal differences in precipitation patterns in southern China, which makes it difficult to understand the precipitation change patterns in southern China. To systematically explain the complexity of precipitation in southern China during the Little Ice Age, this paper divides southern China into three regions, namely, southeast-south China coastal region, central region, and southwest region, and a total of 19 high-resolution paleoclimate records was selected for comparative study, with the following main understandings: compared to the Medieval Warm Period, the precipitation patterns in southern China during the Little Ice Age may be related to the prolonged lingering time of rainbands in southern China and the enhanced influence of typhoons on the coastal areas. The difference is that the southern Qinling Mountains and the Shennongjia Alpine Forest region in central China were mainly cold and wet during the Little Ice Age, and the difference is that the southern Qinling Mountains were mainly wet in the middle and late Little Ice Age periods, which is different from the cold and dry pattern in other regions in central China. This regional difference may be attributed to the complexity of the topography and atmospheric circulation.	Differences in precipitation patterns in southern China during the Little Ice Age

where LCS($\hat{Y}$,Y) denotes the length of the longest overlap subsequence between $\hat{Y}$ and Y, β denotes a default constant, and F_{lcs} denotes the F-measure between R_{lcs} and P_{lcs}.

Additionally, the BLEU metric is used to evaluate the performance of the proposed methodology. This metric measures the precision of the output sentence result by analysing the degree of N-element co-occurrence of the output sentence and the reference sentence. The BLEU score ranges from 0 to 1. The closer the score is to 1, the higher the quality of the generated sentences. The calculation equation is summarized as follows:

$$BLEU = BP \times \exp\left(\sum_{n=1}^{N} w_n \log P_n\right) \tag{26}$$

$$BP = \begin{cases} 1 & if \ \ c > r \\ e^{1-r/c} & if \ \ c \le r \end{cases} \tag{27}$$

BLEU calculates the precision of output sentences 1-g, 2-g..., N-gram. In the equation, P_n is the precision rate of the n-gram, and W_n is the weight of the n-gram, which is generally set to a uniform weight, i.e., $W_n = 1/N$ for any n.

The penalty factor is the brevity penalty (BP), which is less than 1 if the length of the output sentence is smaller than that of the shortest reference sentence. The 1-g accuracy of BLEU indicates how faithful the output sentences are to the original text, while the other n-grams indicate the fluency of the output sentences.

Experiment settings

All the experiments in this research are based on an NVIDIA GTX 2080TI GPU and a Linux operating system. In our task, the maximum length of the sentence is 256, and the batch size is 16. The hidden unit of the bidirectional LSTM is 512, which is the same dimensional setting of the hidden unit of the unidirectional LSTM. We use the adaptive moment estimation (Adam) optimizer (Kingma and Ba 2014) to optimize the loss function, and the learning rate is 0.00001. The dimension of the word embedding operation of the encoder and decoder modules is 768.

Baseline methods

To validate the performance of our proposed models, we select a set of state-of-the-art models as comparison algorithms as follows:

RNN-context This algorithm combines the basic RNN and the context-based RNN methodology to accomplish text summarization based on the basic seq2seq structure (Hu et al. 2015). In contrast to the basic RNN model, the context-based RNN model utilizes the attention mechanism to enhance the performance of the algorithm.

Super-AE This is an autoencoder model that applies an assistant supervisor for capturing the semantic representation and then generates a more reasonable and high-quality summary (Ma et al. 2018).

CGU This applies a convolutional gated unit (CGU) module, which captures the local and global relations of each word from a given sequence (Lin et al. 2018).

Quantitative evaluation

We choose the current mainstream Chinese word embedding methods, the word2vec and global vectors for word representation (GloVe) word vector models, to verify the effectiveness of the word embedding methods proposed in this research. A Chinese pre-training model is used for the comparison experiments. To reduce the training cost, we selected only a simple linear classifier for training and then extracted summary sentences via binary classification.

As indicated in Table 2, compared to the other five word embedding models, we first chose a simple linear classifier for training and then a method of extracting summary sentences through binary classification. ROUGE-1 reached 0.814 and ROUGE-2 reached 0.711. Through these two datasets, we find that nearly 80% of the words in the standard abstract have been captured in the machine-generated automatic abstract, but it is very likely that the machine-generated abstract is very long. Most of the words in the abstract are useless, making the generated abstract unnecessarily

Table 2 Quantitative evaluation results compared to the other word embedding models

Model	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
Word2Vec + Classifier	0.705	0.611	0.625	0.513
GloVe+Classifier	0.713	0.623	0.641	0.531
GloVe+RL	0.744	0.649	0.669	0.599
GloVe+BiLSTM	0.775	0.691	0.701	0.609
GloVe+BiLSTM+Attention	0.789	0.711	0.725	0.631
Ours	0.814	0.781	0.846	0.676

long. ROUGE-L reaches 0.846. This indicator considers all word pairs arranged in word order, which can better reflect the sentence-level word order, and it is found that the generated abstract is more fluent. BLEU is mainly based on the accuracy. Compared to the other five word embedding models, BLEU reaches 0.676. It can be concluded that the accuracy of abstract generation is higher than that of the other five word embedding models.

To validate the performance of our proposed models, we selected a set of the most advanced algorithm models for comparison and analysis. As shown in Table 3, ROUGE-1 of our model reached 0.814, and ROUGE-2 reached 0.781. Through these two sets of data, it is observed that most of the words in the abstract generated by the target are the same as the words in the original abstract, but the summary may contain redundancy. ROUGE-L of our model reached 0.846, and we also found that using the chosen model smoothed the final generated summary. This indicates that our model achieves better results than the current state of the art models and enhances the performance of Chinese summarization systems.

As suggested by the results in Table 3, it is obvious that our approach outperforms all other approaches because our system makes full use of the BERT model. The BERT model learns much of the language information during unsupervised pre-training, and it can be used to represent the input more semantically. Additionally, such results clearly demonstrate that when the information comes from several sources, the presented model can generate an effective and meaningful summary.

Table 3 Quantitative evaluation results compared to the other models

Model	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
RNN	0.738	0.705	0.781	0.621
RNN-context	0.744	0.721	0.797	0.633
Super-AE	0.781	0.743	0.805	0.642
CGU	0.792	0.765	0.821	0.655
Ours	0.814	0.781	0.846	0.676

Qualitative evaluation

We conducted a series of tests on a real dataset based on the model that was trained, and the generated headlines were compared directly to the original text summary headlines. The results are listed in Table 4. As shown in the figure, the headlines generated by our trained model are better and can cover the core information in the original text, which further confirms the effectiveness of the algorithm proposed in this study. For example, the important information of the original title "*Suggestions for countermeasures of national core science and technology resource sharing mechanism*" is "rock cores" and "sharing". Our model generates this core content information while expressing the key messages of the original text.

Convergence evaluation

In the process of training the entire model, we should pay attention to the size of the target value (loss) at any time. The loss in a normal model should slowly decrease with an increasing number of epochs and should eventually stabilize. At the initial stage of model training, the change in loss may be relatively obvious, but as long as the amount of data is sufficient, the model is correct, and the number of epochs is large enough, the model will eventually reach a state of convergence, close to the maximum value or a local maximum will be reached. In this experiment, the cross-entropy is used as the loss, and the input part is masked. The cross-entropy is the true probability distribution. The reason why the cross-entropy is used as the loss function is to achieve maximum likelihood estimation, i.e., the predicted distribution obtained by the model should be as close as possible to the actual distribution of the data. This experiment has a total of 100 epochs. We visualized

Table 4 Comparison of the original title and the title generated by the trained model

No.	Original Title	Generated Title
1	Suggestions for countermeasures of national core science and technology resource sharing mechanism	Suggestions related to core resource sharing countermeasures
2	Reflections on the development of physical geological data science in the new era - taking the natural resources physical geological data centre as an example	Reflections related to the development of geological data science for geological data centre
3	Deep-seismic bathymetry-based deep dynamics research methods and applications	Study on the dynamics related to seismic bathymetry
4	Geochemical characteristics and geological significance of rare earth elements in mud shales of the Cretaceous Madongshan Formation in the Liupanshan Basin	Geochemical characteristics of the Cretaceous and the basic significance of its study

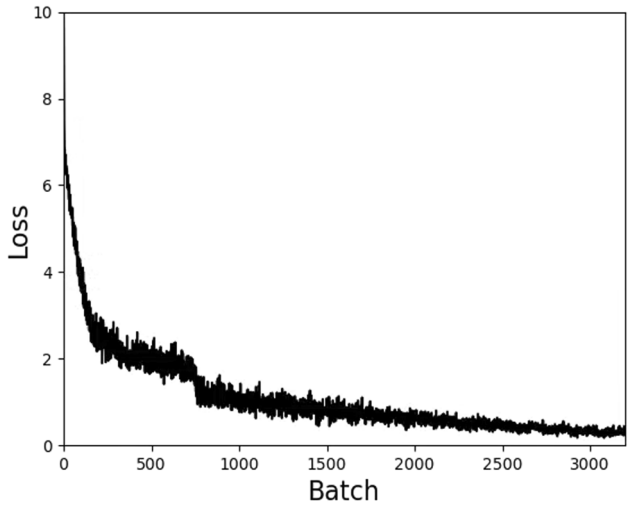

Fig. 8 Convergence evaluation results for automatic summary generation

the loss value changes during the first 20 epochs and represented them as a line graph, as shown in Fig. 8.

Application and platform

A geological public hotspot extraction and mining platform is constructed and used as an important part of the information extraction and knowledge discovery process in the domain of geoscience. Herein, the geological briefing is the core module of the mining platform, as shown in Fig. 9. The data shown in Fig. 9 are collected from 28 April 2020, to 29 April 2020, and they are regarded as an example to support decision-making. Based on this platform, the time of crawling and mining can be freely chosen. The platform can be classified into three stages: (1) the briefing list; (2) proportion of hotspots; and (3) data count. Next, we describe each section in detail.

Briefing list Data from different journals are presented on the platform in the briefing list that consists of the Geological Bulletin of China, Acta Geological Sinica, Geological Review and Mineral Deposits. The briefing list includes four columns: ID, briefing (in this research, this is generated based on the proposed model), data source and the original content with a link. Based on the briefing list, decision makers or users can obtain topical concerns or hotspots during different periods or stages in a timely and effective manner to formulate timely decisions.

Proportion of hotspots This section mainly shows the current geological hotspots and concerns from crawler data. A larger occupied area represents a higher level of attention or hotness. Only hotspots larger than 10% are displayed in the section. Based on this module, decision makers or users can understand the current concerns more clearly.

Data count The main goal of the statistics module is to be able to count the amount of data on the platform so that

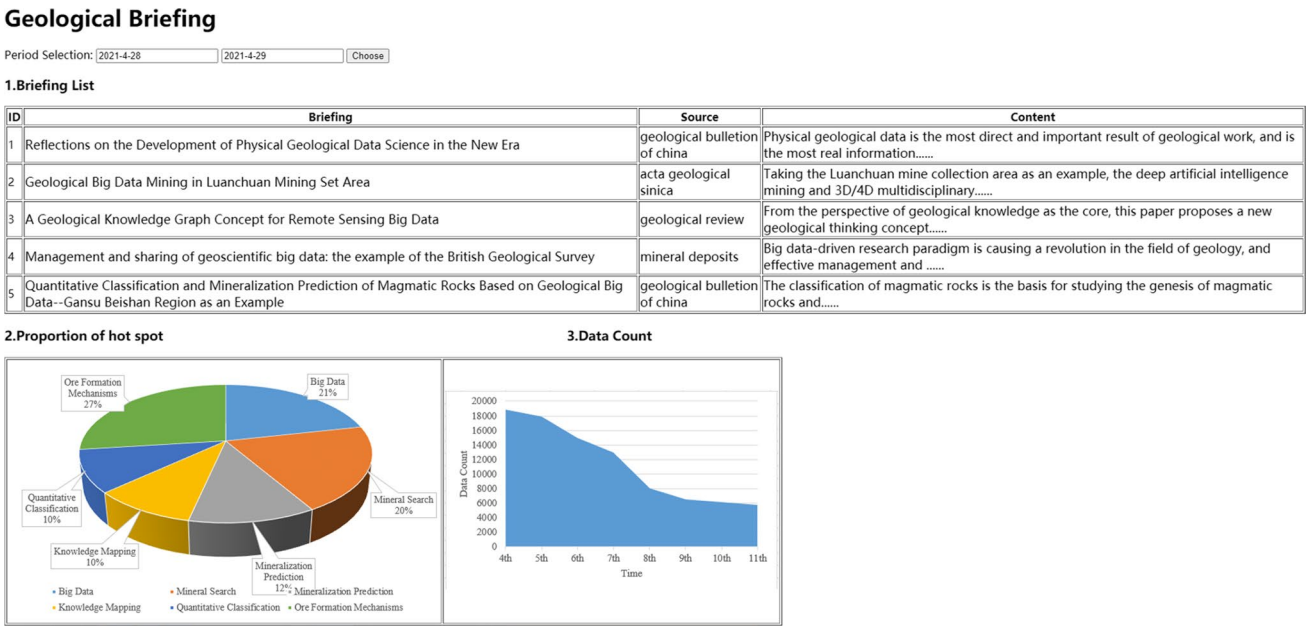

Fig. 9 Example of the Geology Journal Hotspot Mining Platform. The platform includes three sections: the briefing list, proportion of hot spots and data count

decision makers or users can visualize the geology trends over time.

Conclusions and future work

Automatic text summarization plays an important role in information extraction, briefing generation and knowledge discovery. In this paper, an automatic text summarization generation framework is proposed based on the fine-tuned BERT model. The framework is composed of an encoder module, decoder module and training module. The core stages of summary generation are the combination of encoder and decoder modules, after which the multi-stage function is used to connect modules, thus endowing the text summarization model with a multi-task learning architecture. We evaluate the performance of the proposed model on the developed dataset, and compared to a set of baseline models, our model achieves the best results. Additionally, we develop a briefing list platform to serve as an important part of the information extraction and knowledge discovery process in the domain of geoscience.

We conducted a set of primary experiments with the developed dataset to evaluate the summarization task for Chinese text. The experimental results confirm the following suggestions and findings. First, the fine-tuned BERT model achieves better results than other models, such as the word-2vec and GloVe models. Second, the text summarization based on neural network models trained on the developed datasets achieves a better performance than summarizations based on other models (i.e., RNN, RNN-context, Super-AE, CGU). Third, the results of our experiments show that deep learning with BERT representation significantly improves the results of the summarization system and outperforms the state-of-the-art approaches. As the BERT model learns much of the language information during unsupervised pre-training, it can be fine-tuned even with small datasets and hence performs better than CNN or RNN downstream-based models that must be trained from scratch.

Future work will focus on the following points: (1) more domain knowledge should be considered and added to the model to further enhance the representational power of the model. (2) Multi-source data are a very interesting exploration target, such as data retrieved from more literature journals and geological reports, which will help the model learn more valuable knowledge. (3) We intend to collect a vast amount of textual data, train BERT in the domain of geosciences for a variety of geoscience tasks in NLP, and further improve the word representation.

Acknowledgements 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 Natural Science Foundation of China (42050101, U1711267, 41871311, 41871305), National Key Research and Development Program (2018YFB0505500, 2018YFB0505504) and the Fundamental Research Funds for the Central Universities, China University of Geosciences (Wuhan) (No. CUG2106116)).

Declarations

Conflict of interest The authors declare no conflicts of interest.

References

Al-Abdallah RZ, Al-Taani AT (2017) Arabic single-document text summarization using particle swarm optimization algorithm[J]. Procedia Comput Sci 117:30–37

Bhat IK , Mohd M , Hashmy R (2018) SumItUp: a hybrid single-document text summarizer[M]

Carbonell J , Goldstein J (1998) The use of MMR, diversity-based reranking for reordering documents and producing summaries. ACM:335–336

Ceylan H , Mihalcea R , O'Zertem U, et al. (2010) Quantifying the limits and success of extractive summarization systems across domains[C]// human language technologies: the conference of the north American chapter of the Association for Computational Linguistics. Association for Computational Linguistics

Chen YC , Bansal M (2018) Fast abstractive summarization with reinforce-selected sentence rewriting[C]// proceedings of the 56th annual meeting of the Association for Computational Linguistics (volume 1: long papers)

Cheng J, Lapata M (2016) Neural summarization by extracting sentences and words. Proceedings of the 54th annual meeting of the Association for Computational Linguistics, Berlin, pp 484–494

Devlin J, Chang M W, Lee K, et al (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. arXiv preprint arXiv:1810.04805

El-Kassas WS, Salama CR, Rafea AA et al (2020) Automatic text summarization: a comprehensive survey[J]. Expert Syst Appl 113679

fxsjy (2018) https://github.com/fxsjy/jieba

Grusky M, Naaman M, Artzi Y (2018) NEWSROOM: A dataset of 1.3 million summarieswith diverse extractive strategies. Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans

Hou S, Lu R (2020) Knowledge-guided unsupervised rhetorical parsing for text summarization[J]. Inf Syst :101615

Hou L, Hu P, Bei C (2017) Abstractive document summarization via neural model with joint attention. Paper presented at the natural language processing and Chinese computing, Dalian

Howard J, Ruder S (2018) Universal language model fine-tuning for text classification. Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers):328–39

Hu B & Chen Q, Zhu F (2015) LCSTS: A Large Scale Chinese Short Text Summarization Dataset. https://doi.org/10.18653/v1/D15-1229

Hunter J, Freer Y, Gatt A et al (2012) Automatic generation of natural language nursing shift summaries in neonatal intensive care: BT-nurse[J]. Artif Intell Med 56(3):157–172

Joshi M , Hui W, Mcclean S (2018) Dense semantic graph and its application in single document summarisation. Emerging Ideas on Information Filtering and Retrieval

Jun Q (2019) Hybrid text summarization model based on reinforcement learning[J]. Information Technol Inform Technol 226(01):67–70
Kingma, D, Ba J (2014) Adam: a method for stochastic optimization. International Conference on Learning Representations
Liang Z , Du J , Li C (2020) Abstractive social media text summarization using selective reinforced Seq2Seq attention model[J]. Neurocomputing, 410
Lin C-Y (2004) ROUGE: a package for automatic evaluation of summaries. Proceedings of the ACL Workshop: Text Summarization Braches Out 10
Lin J, Sun X, Ma S, Su Q (2018) Global Encoding for Abstractive Summarization. 163–169. https://doi.org/10.18653/v1/P18-2027
Liu B (2012) Sentiment analysis and opinion mining[J]. Synthesis Lectures Human Language Technol 5(1):160–167
Liu Y (2019) Fine-tune BERT for extractive summarization[J]
Lu R, Wang T, Zeng BQ, Liu X (2020) TSPT: a three-stage composite text summarization model based on pre-training. Appl Res Comput 37(10):2917–2921
Ma X (2019) Geo-Data Science: Leveraging Geoscience Research with Geoinformatics, Semantics and Open Data. Acta Geologica Sinica 93:44–47. https://doi.org/10.1111/1755-6724.14240
Ma S, Sun X, Lin J, Wang H (2018) Autoencoder as Assistant Supervisor: Improving Text Representation for Chinese Social Media Text Summarization
Ma X, Ma C, Wang C (2020) A new structure for representing and tracking version information in a deep time knowledge graph. Comput Geosci 145:104620. https://doi.org/10.1016/j.cageo.2020.104620
Mao X, Yang H, Huang S et al (2019) Extractive Summarization Using Supervised and Unsupervised Learning[J]. Expert Syst Appl 133:173–181
Mohan MJ, Sunitha C, Ganesh A, Jaya A (2016) A study on ontology based abstractive summarization. Procedia Comput Sci 87:32–37
Nallapati R, Xiang B , Zhou B (2016a) Sequence-to-sequence RNNs for text summarization[J]
Nallapati R, Zhai F, Zhou B (2016b) SummaRuNNer: A Recurrent Neural Network based Sequence Model for Extractive Summarization of Documents
Narayan S , Cohen SB, Lapata M (2018a) Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization[J]
Narayan S , Cohen SB , Lapata M (2018b) Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization[J]
Narayan S, Cardenas R, Topoulos NP, Cohen SB, Lapata M, Yu JS, Chang Y (2018c) Document modeling with external attention for sentence extraction. Proceedings of the 56th annual meeting of the Association for Computational Linguistics, Melbourne
Nenkova A , Mckeown K (2012) A survey of text summarization techniques[J]. Springer US
Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. (2018) Deep contextualized word representations. Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long papers):2227–37
Qiu Q, Xie Z, Wu L, Wenjia L (2018a) DGeoSegmenter: A dictionary-based Chinese word segmenter for the geoscience domain. Comput Geosci 121. https://doi.org/10.1016/j.cageo.2018.08.006
Qiu Q, Xie Z, Wu L (2018b) A cyclic self-learning Chinese word segmentation for the geoscience domain. Geomatica. https://doi.org/10.1139/geomatica-2018-0007
Qiu Q, Xie Z, Wu L, Wenjia L (2019) Geoscience Keyphrase Extraction Algorithm Using Enhanced Word Embedding Expert Systems with Applications 125. https://doi.org/10.1016/j.eswa.2019.02.001
Qiu Q, Xie Z, Wu L, Tao L (2020) Automatic spatiotemporal and semantic information extraction from unstructured geoscience reports using text mining techniques Earth Sci Inform 13. https://doi.org/10.1007/s12145-020-00527-9
Radev DR (2004) LexRank: graph-based lexical centrality as salience in text summarization[J]. J Qiqihar Junior Teachers College 22:2004
Radford A, Narasimhan K, Salimans T, Sutskever I (2018) Improving language understanding by generative pre-training
Rush AM, Chopra S, Weston J (2015) A Neural Attention Model for Abstractive Sentence Summarization[J]. Computer Science. Sequence model for extractive summarization of documents. In Proceedings of the 31st AAAI conference on artificial intelligence, pages 3075–3081, San Francisco
Sandhaus E (2008) The New York Times Annotated Corpus. Linguistic Data Consortium, Philadelphia, 6(12)
Shuai W , Xiang Z , Bo L , et al. (2017) Integrating extractive and abstractive models for long text summarization[C]// 2017 IEEE international congress on big data (BigData congress). IEEE
Siddiqui T , Shamsi J A (2018) Generating abstractive summaries using sequence to sequence attention model[C]// Frontiers of information technology. IEEE Comput Soc
Sutskever I , Vinyals O , Le Q V (2014) Sequence to sequence learning with neural networks[J]. NIPS
Tan J , Wan X , Xiao J (2017) Abstractive document summarization with a graph-based attentional neural model[C]// meeting of the Association for Computational Linguistics
Wang C, Hazen R, Cheng Q, Stephenson M, Zhou C, Fox P, Shen S, Oberhänsli R, Hou Z, Ma X, Feng Z, Fan J, Ma C, Hu X, Luo B, Wang J (2021) The deep-time digital earth program: data-driven discovery in geosciences. Natl Sci Rev. https://doi.org/10.1093/nsr/nwab027
Wenjia L, Ma K, Qiu Q, Wu L, Xie Z, Li S, Chen S (2021) Chinese Word Segmentation Based on Self-Learning Model and Geological Knowledge for the Geoscience Domain Earth and Space Science 8. https://doi.org/10.1029/2021EA001673
Zhang H , Gong Y , Yan Y , et al. (2019) Pretraining-based natural language generation for text summarization[J]
Zhou Q , Yang N , Wei F , et al. (2018) Neural document summarization by jointly learning to score and select sentences