IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received October 22, 2017, accepted December 4, 2017, date of publication December 27, 2017, date of current version March 9, 2018.

Digital Object Identifier 10.1109/ACCESS.2017.2785229

Attention-Based Relation Extraction With Bidirectional Gated Recurrent Unit and Highway Network in the Analysis of Geological Data

XIONG LUO[1, 2, 3], (Member, IEEE), WENWEN ZHOU[1, 2, 3], WEIPING WANG[1, 2, 3], YUEQIN ZHU[3, 4], AND JING DENG[1, 2, 3]

[1] School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China
[2] Beijing Engineering Research Center of Industrial Spectrum Imaging, Beijing 100083, China
[3] Key Laboratory of Geological Information Technology, Ministry of Land and Resources, Beijing 100037, China
[4] Development and Research Center, China Geological Survey, Beijing 100037, China

Corresponding authors: Xiong Luo (xluo@ustb.edu.cn) and Weiping Wang (weipingwangjt@ustb.edu.cn)

This work was supported in part by the Key Laboratory of Geological Information Technology, Ministry of Land and Resources, under Grant 2017320, in part by the Fundamental Research Funds for the China Central Universities, University of Science and Technology Beijing under Grant FRF-BD-16-005A, in part by the National Key Technologies Research and Development Program of China under Grant 2015BAK38B01, in part by the National Natural Science Foundation of China under Grant 61174103 and Grant 61603032, in part by the National Key Research and Development Program of China under Grant 2016YFB0700502, Grant 2016YFB1001404, Grant 2016YFC0600510, and Grant 2017YFB0702300, and in part by the University of Science and Technology Beijing –National Taipei University of Technology Joint Research Program under Grant TW201705.

ABSTRACT Attention-based deep learning model as a human-centered smart technology has become the state-of-the-art method in addressing relation extraction, while implementing natural language processing. How to effectively improve the computational performance of that model has always been a research focus in both academic and industrial communities. Generally, the structures of model would greatly affect the final results of relation extraction. In this article, a deep learning model with a novel structure is proposed. In our model, after incorporating the highway network into a bidirectional gated recurrent unit, the attention mechanism is additionally utilized in an effort to assign weights of key issues in the network structure. Here, the introduction of highway network could enable the proposed model to capture much more semantic information. Experiments on a popular benchmark data set are conducted, and the results demonstrate that the proposed model outperforms some existing relation extraction methods. Furthermore, the performance of our method is also tested in the analysis of geological data, where the relation extraction in Chinese geological field is addressed and a satisfactory display result is achieved.

INDEX TERMS Relation extraction, bidirectional gated recurrent unit (BGRU), highway network, attention, geological data.

I. INTRODUCTION

Natural language processing (NLP) as an active research area in artificial intelligence fulfills computational tasks for large natural language data, including character, word, sentence, text, and many others. Then, relation extraction as an important sub task of NLP is critical to achieving natural language generation and understanding. It extracts semantic relationships between two entities from texts [1], [2]. The extracting process is coordinated through the use of some machine learning algorithms, e.g., kernel-based methods, in some applications [3], [4].

Recently, as one of the popular topics in machine learning, deep learning has achieved amazing performance in many fields, such as computer vision, speech recognition, and relation extraction, and many others [5]–[8]. Deep learning simulates human brain to construct models and extracts useful information from large-scale dataset automatically, without handcrafted features and lexical resources. Convolutional neural network (CNN) and recurrent neural network (RNN) are two popular approaches of deep learning. The main difference between them lies in the network architecture. The former usually consists of a convolution layer, pooling layer, and nonlinear layer. The latter mostly contains an unidirectional or bidirectional RNN layer which could capture memory of historical information through a gating mechanism. Deep learning methods used for relation extraction usually utilize word representation as the input of CNN or RNN model [9]. More complicated features are captured and the representation of relationships between two entities are generated by the model. Generally, deep learning methods require a large amount of manually labeled data, which spend much additional time and cost [10]. In response to this limitation, the idea of distant supervision was proposed [11]. Distant supervision assumes that a sentence including two entities implicates the relation of this entity pair in knowledge bases. Based on this hypothesis, unlabeled corpus could be aligned with knowledge bases. Distant supervision effectively solves the problem of manually labeling and enables deep learning model to become the state-of-the-art method in the task of relation extraction [12]–[14].

Furthermore, RNN models, such as long short-term memory (LSTM) [15] and gated recurrent unit (GRU) [16], could capture long short-term dependencies through the gating mechanism. GRU is a popular variant of LSTM with less gating units and higher efficiency [17]. Then, GRU has been widely used in many NLP tasks. For example, the GRU model was used in the realization of neural responding machine [18] and the implementation of language model [19]. Due to the introduction of gating mechanism, LSTM and GRU models have shown satisfactory performance in the sequence learning tasks, such as machine translation, speech recognition, and relation extraction especially [20].

In addition to the above works, some novel neural networks (NNs) are also developed in order to further improve the computational performance. Highway network is a special NN framework proposed recently [21]. It learns to control the flow of information through a network by the use of gating units as well. The proposal of highway network is with the purpose of designing an extremely deep and efficient network. Specifically, in NLP field, it has been verified that the addition of highway network could extract much more comprehensive semantics features when constructing a language model [22].

Motivated by it, in this article, we focus on the application of highway network in the task of distant supervised relation extraction. A highly effective relation extraction method is accordingly proposed by incorporating highway network into a bidirectional GRU (BGRU) model. Additionally, in the proposed method, the attention mechanism is also utilized to assign weights of words and sentences. We call the proposed model as attention-based BGRU and highway network (Att-BGRU-HN) model. The use of highway network between BGRU and attention mechanism could play a key role in extracting the most important features between words comprehensively. A popular benchmark dataset, i.e., New York Times corpus aligned with Freebase [23], is used to test the performance of our method. And the experimental results demonstrate that the proposed model achieves significant improvements compared with the single attention-based BGRU (Att-BGRU) model and some popular relation extraction models using distant supervision.

Finally, we employ the proposed model to the analysis of Chinese geological data. Considering that the proposed model could address relation extraction with a fixed number of specified relation classes, it is exactly suitable for the data analysis in geological field with limited types of relationship. It is noted that the word segmentation is necessary for Chinese relation extraction. On the basis of distant supervision hypothesis, we align segmented Chinese sentences acquired by Baidu encyclopedia crawler with three tuple set from Chinese geological thesaurus. These labeled sentences are fed to our model to capture semantic features and obtain the representation of relations. The experimental results also show that our model achieves satisfactory performance for relation extraction in the geological field.

Here, the contributions of this article would be summarized as follows:

(1) We introduces highway network into a BGRU model using attention mechanism in the task of relation extraction. The use of highway network helps to capture much more semantics features between words.

(2) Compared with the single Att-BGRU model and other popular relation extraction models using distant supervision, the proposed model could obtain more precise representation of relations and achieve better performance.

(3) To demonstrate the usability of the proposed model, we apply it in the analysis for geological data and obtain preferable display results for the task of relation extraction.

The remainder of this article is organized as follows. The next section analyzes the related work on the relation extraction methods. Section III presents the implementation of our method. Section IV illustrates the experiment results. Moreover, Section V discusses the application of our model in analyzing geological data. Finally, the conclusion of this article is drawn in Section VI.

II. RELATED WORK

As one of the most important task in NLP, relation extraction has drawn much attention over the years. Various methods have been proposed.

Traditional relation extraction methods require manual rules, and employ pattern matching technique to extract corresponding relational instances from texts. Then, supervised machine learning methods transform the relation extraction task into a classification problem on the basis of some annotation tools, e.g., part of speech (POS) tagger and parser, while acquiring effective features. Conventional machine learning classifiers used in the relation extraction task mainly include maximum entropy (ME) model, logistic regression (LR) model, support vector machine (SVM) model, and so on.

Since the distant supervision hypothesis was proposed, many approaches have been developed in the relation extraction based on distant supervision. For example, a multi-class logistic classifier optimized by Gaussian regularization was utilized to extract relationships based on a large-scale corpus that was automatically constructed by aligning a knowledge base with unlabeled sentences. Experiments demonstrated that this method could efficiently reduce the dependence of the model on manually labeling data and obtain satisfactory classification results [11]. Nevertheless, distant supervision may causes noise problem due to the introduction of wrong labels. In order to address this issue, some improvement strategies have been developed. For example, a multi-instance model was designed in the situation that an entity pair might have multiple relationships [12]. A multi-instance multi-label model with Bayesian network was proposed for relation extraction [13].

In recent years, a large number of relation extraction methods based on deep learning have been presented. For example, a CNN model was proposed for the relation extraction task [24]. As the input of CNN model, all words in corpus are embedded into low-dimensional vectors on the basis of word features and position features. An attention-based bidirectional LSTM model for relation extraction was proposed [20]. Through the combination of neural attention mechanism and LSTM, the most important semantic information in a sentence is captured. This model does not utilize any features derived from lexical resources or NLP systems and outperforms most of relation extraction methods. Moreover, deep learning method using distant supervision is a hot issue of handling relation extraction task and tremendously attracts researchers' attention. For instance, in order to minimize the error from distant supervision, a CNN relation extraction model combined with sentence level attention mechanism was proposed [14]. The high-quality sentences were assigned higher weights, while noisy sentences got smaller weights through this attention mechanism. The experimental result indicated that the combination of deep learning model with neural attention mechanism could effectively reduce the error and improve the performance of relation extraction.

Recently proposed highway network is a NN framework that learns how to control the flow of information in the network through gating units. Highway network is mostly suitable for image recognition field when building extremely deep networks. There are quite few attempts of highway network in NLP field. The combination of CNN and highway network was proven to be effective when building a character-aware neural language model. Then, the addition of highway network layer could extract much more comprehensive semantics features between characters [22].

Although there are many methods in the relation extraction task, there is still room on further improving the computational performance and enhancing the cross-domain adaptability of the relation extraction model. Then, the design of the model structure is one of the future research topics along this direction. To the best of our knowledge, this is the first effort to incorporate highway network into attention-based RNN model in distant supervised relation extraction.

III. THE PROPOSED MODEL

On the basis of the above works [14], [20], [22], we propose a novel hybrid relation extraction model through the use of neural attention mechanism. The architecture of our model is shown in Fig. 1, where those key modules are analyzed as follows.

A. INPUT WORD REPRESENTATION

In our method, the input layer is the bottom component of this NN model, whose output is sent to the BGRU layer. Here, each input word is converted into a low dimensional vector which implicates its semantic meaning. Considering the fact that those words which are closer to head or tail entity are more weighty in the task of relation extraction, we use position features (PF) to describe input words.

For a given sentence composed of n words, $S = \{c_1, c_2, \cdots, c_n\}$, we transform each word c_i into a low fixed dimensional vector $\mathbf{e}_i$, where $i = 1, 2, \cdots, n$. Specifically, all words in New York Times corpus used in our experiment are indexed first, and then for each word in S , we match the index of c_i to an embedding matrix $\mathbf{W} \in \mathbb{R}^{d^{\mathbf{W}} \times |V|}$, where $d^{\mathbf{W}}$ is the size of word embedding and V is a fixed vocabulary including all words of train and test datasets.

PF is used to describe the relative distances of the current word to head and tail entity in a sentence. It is similar to [24]. For example, for the sentence "John was born in Beijing where he completed his first degree in economics", firstly we can calculate the relative distances from "born" to head entity "John" and tail entity "Beijing". They are -2 and 2, respectively. Secondly, both relative distances are mapped to vectors $\mathbf{p}_1$, $\mathbf{p}_2$, respectively. And PF equals to the join of the two.

Finally, the input word c_i is represented as the join of $\mathbf{e}_i$ and PF, and we defined it as $\mathbf{x}_i$.

B. BIDIRECTIONAL GATED RECURRENT UNIT (BGRU) NETWORK

Generally, the thinking of human brain is persistent. For example, when reading articles, we can understand each word on the basis of understanding the previous words we read. RNN is a type of NN architecture that has recurrent structure to preserve previous information continually. However, the performance of RNN may be not theoretically perfect in practice. It would lose the ability to preserve and process information long ago, and suffer from gradient vanishing problem as time goes on [25].

In order to tackle this problem, some special alternatives of RNN were proposed, and the most representative methods are LSTM [15] and GRU [16]. Both of them could capture long short-term dependencies through a gating mechanism. However, GRU may be comparable to LSTM, since it is with a simpler structure and a lower computational complexity [17].

Here, GRU-based network is used to train and gain context semantic features for each input word of S. As a gated RNN, GRU controls the flow of information through reset gate and update gate. In this article, the input is $\mathbf{x}_t$, where t is the current time step. And $\mathbf{h}_t$ indicates the state of the hidden layer at time t .

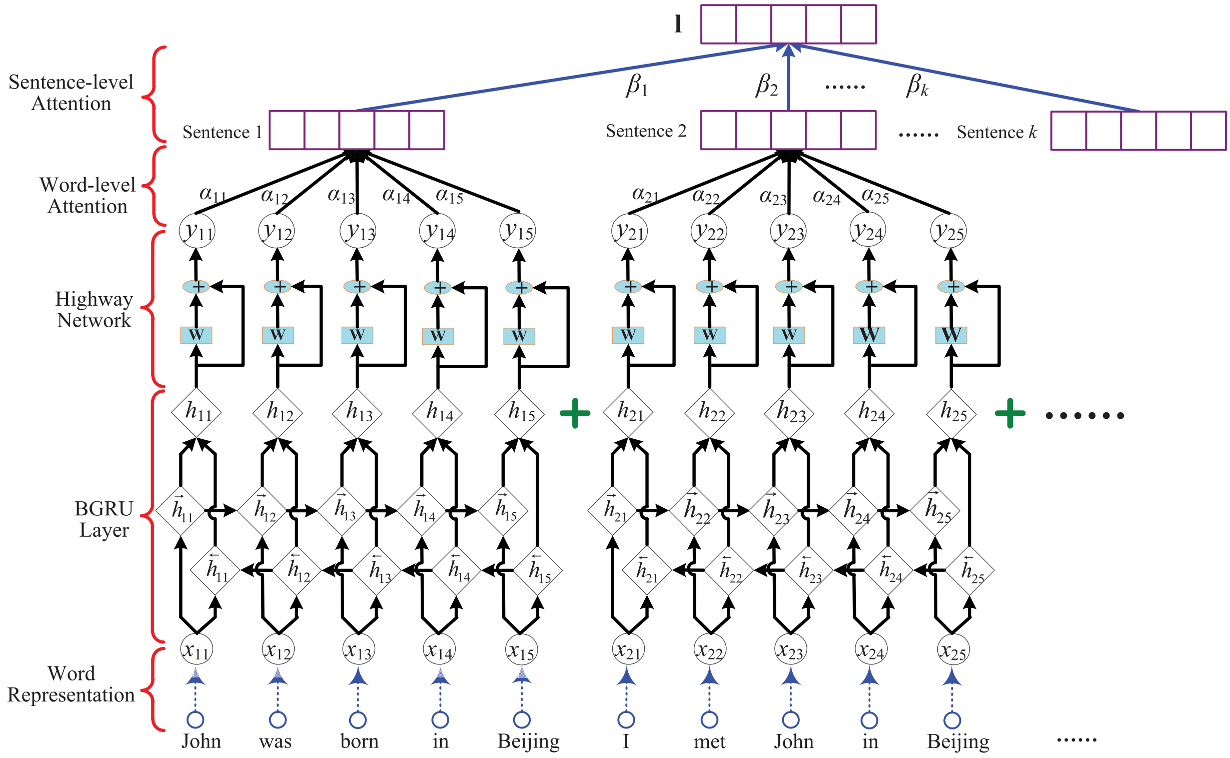

l
Sentence-level Attention
β_1
β_2
......
β_k
Sentence 1
Sentence 2
......
Sentence k
Word-level Attention
α_{11} α_{12} α_{13} α_{14} α_{15}
α_{21} α_{22} α_{23} α_{24} α_{25}
y_{11} y_{12} y_{13} y_{14} y_{15} y_{21} y_{22} y_{23} y_{24} y_{25}
Highway Network
W
h_{11} h_{12} h_{13} h_{14} h_{15} h_{21} h_{22} h_{23} h_{24} h_{25}
BGRU Layer
$\overrightarrow{h}_{11}$ ... $\overrightarrow{h}_{25}$
$\overleftarrow{h}_{11}$ ... $\overleftarrow{h}_{25}$
Word Representation
x_{11} x_{12} x_{13} x_{14} x_{15} x_{21} x_{22} x_{23} x_{24} x_{25}
John was born in Beijing
I met John in Beijing
......

FIGURE 1. **Architecture of our proposed model Att-BGRU-HN for the relation extraction task.**

Specifically, two key parts compose the GRU-based network. One is a reset gate r_t^j with corresponding weight matrices $\mathbf{W}_r$ and $\mathbf{U}_r$, for the j-th GRU hidden unit. When reset gate is closed ($r_t^j = 0$), the previous state $\mathbf{h}_{t-1}$ is ignored and the current state is defined by the current input $\mathbf{x}_t$ only, which means the possibility of dropping all information that is useless for current hidden state. The other is update gate z_t^j designed to regulate the degree of the information from previous state $\mathbf{h}_{t-1}$ transmitted to the current hidden state $\mathbf{h}_t$, with corresponding weight matrices $\mathbf{W}_z$ and $\mathbf{U}_z$. Moreover, $\widetilde{h}_t^j$ represents candidate activation based on previous state $\mathbf{h}_{t-1}$ and current input $\mathbf{x}_t$, with corresponding weight matrices $\mathbf{W}_h$ and $\mathbf{U}_h$. The final output of the hidden unit $\mathbf{h}_t$ includes the previous part computed by multiplying $1 - z_t^j$ and h_{t-1}^j, and the updated part computed by multiplying z_t^j and $\widetilde{h}_t^j$.

The whole data flows of the j-th GRU hidden unit are shown as follows:

$$r_t^j = \sigma(\mathbf{W}_r\mathbf{x}_t + \mathbf{U}_r\mathbf{h}_{t-1})^j, \tag{1}$$

$$z_t^j = \sigma(\mathbf{W}_z\mathbf{x}_t + \mathbf{U}_z\mathbf{h}_{t-1})^j, \tag{2}$$

$$\widetilde{h}_t^j = \tanh(\mathbf{W}_h\mathbf{x}_t + \mathbf{U}_h(r_t \odot \mathbf{h}_{t-1}))^j, \tag{3}$$

$$h_t^j = z_t^j\widetilde{h}_t^j + (1 - z_t^j)h_{t-1}^j. \tag{4}$$

where $\sigma(\cdot)$ is the logistic element-wise sigmoid function that can be represented as $\sigma(x) = \frac{1}{1+e^{-x}}$ $(x \in \mathbb{R})$ in this article, and $\tanh(\cdot)$ is the hyperbolic tangent function that can be represented as $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ $(x \in \mathbb{R})$ in our method. Here, $\odot$ represents the Hadamard product, also known as element-wise product, which means the product of the corresponding elements of two matrices.

When extracting semantic features of words, the above architecture based on positive-sequence of time could only consider the historical information well, even if future information is as important as historical information for the representation of words in S.

Then, BGRU network deals with this problem by introducing a future layer in which input sequence of data is in reverse direction. Therefore, this network uses two hidden layers to extract information from the past and the future. These two hidden layers are connected to the same output layer. This network makes full use of context information of input sequence. In consideration of it, BGRU network is adopted in this article. The left-GRU layer learns the historical information of $\mathbf{x}_t$ by feeding the positive-sequence of S, and the right-GRU obtains the future information of $\mathbf{x}_t$ through feeding the reverse of S. The final output of the i-th word of sentence S is represented by the following equation:

$$\mathbf{h}_i = [\overrightarrow{\mathbf{h}_i} \oplus \overleftarrow{\mathbf{h}_i}], \tag{5}$$

where $\overrightarrow{\mathbf{h}_i}$ and $\overleftarrow{\mathbf{h}_i}$ represent the output of the i-th word through left-GRU layer and right-GRU layer, respectively. And $\oplus$ is an element-wise sum, which means the sum of those corresponding elements of two matrices.

C. HIGHWAY NETWORK

It is verified that simply attention-based deep learning models could perform well in the task of relation extraction [20], [26], [27]. Nevertheless, there is still some additional room for improvement along this direction. Here, the highway network is employed to capture much more semantic features between words. In this article, the output $\mathbf{h}$ of BGRU is sent to highway network layer.

Highway network mainly learns how to control the flow of information in the network through gating units. Two nonlinear transforms T and C are defined in highway network. The former is called as the transform gate with corresponding weight matrix $\mathbf{W}_T$ and the latter is known as the carry gate with corresponding weight matrix $\mathbf{W}_C$. Hence, the output $\mathbf{y}$ is computed by the following affine transform:

$$\mathbf{y} = (H(\mathbf{h}, \mathbf{W}_H)) \odot (T(\mathbf{h}, \mathbf{W}_T)) + \mathbf{h} \odot (C(\mathbf{h}, \mathbf{W}_C)). \quad (6)$$

Normally, C is set to $1 - T$ for simplicity. Thus, the highway network layer could be also calculated by:

$$\mathbf{y} = (H(\mathbf{h}, \mathbf{W}_H)) \odot (T(\mathbf{h}, \mathbf{W}_T)) + \mathbf{h} \odot (1 - T(\mathbf{h}, \mathbf{W}_T)), \quad (7)$$

where the dimensions of $\mathbf{h}$ and $\mathbf{y}$ have to match, so that $\mathbf{W}_H$ and $\mathbf{W}_T$ should be square matrices here. The activation functions of H and T are rectified linear unit (ReLU) and sigmoid, respectively. Here, ReLU function is defined as $\mathrm{ReLU}(x) = \max\{0, x\}$, where $x \in \mathbb{R}$.

Particularly, from (7) we could observe that:

$$\mathbf{y} = \begin{cases} \mathbf{h}, & T(\mathbf{h}, \mathbf{W}_T) = 0, \\ H(\mathbf{h}, \mathbf{W}_H), & T(\mathbf{h}, \mathbf{W}_T) = 1. \end{cases} \quad (8)$$

By performing a Jacoby change on (8), we could obtain:

$$\frac{d\mathbf{y}}{d\mathbf{h}} = \begin{cases} \mathbf{I}, & T(\mathbf{h}, \mathbf{W}_T) = 0, \\ (H(\mathbf{h}, \mathbf{W}_H))', & T(\mathbf{h}, \mathbf{W}_T) = 1. \end{cases} \quad (9)$$

Here, because of the introduction of gating units, highway network has the ability to balance its behavior between a traditional plain layer and a layer just passing the inputs through. Highway network has an unique advantage for super deep NN training, whose optimization is not hampered even if the number of network depth increases to one hundred.

In addition, highway network could also be used in NLP tasks to deepen the intrinsic relevance of the features extracted by the upper NN. A typical example is the combination of CNN and highway network for constructing neural language models [22]. Compared with the single CNN model, the addition of highway network could be able to extract the intrinsic relevance between character features acquired by CNN, and experimental results demonstrated that the final acquired representations of characters are much more comprehensive and semantic.

Here, we use highway network to enhance the relevance between features acquired by BGRU layer, and get the vector representation of words with deeper semantic connotations. The bias $\mathbf{b}_T$ of T is initialized as negative number in general, so that the carry function C could carry large enough original information. We also adopt this way to initialize $\mathbf{b}_T$ in this article.

D. ATTENTION

Recently, neural attention mechanism originated from the human visual attention process has achieved great success in a wide range of applications, including machine translations [28], abstractive sentence summarization [29], speech recognition [30]–[32], and many others.

Here, we utilize neural attention mechanism in the relation extraction task. In Fig. 1, it is located on the fourth and fifth tiers of our model architecture, after word representation, BGRU, and highway network. It is obviously that each entity pair may correspond to more than one sentence in the corpus and each sentence consists of a number of words. In response to it, we employ two kinds of attention mechanisms, i.e., attention of words and attention of sentences.

1) ATTENTION OF WORDS

Generally, if all word vectors obtained by BGRU and highway network are treated equally for the representation of sentence S, there would be a waste of computational time on some unimportant words. Consequently, motivated by [20], we give a weight to each element of the input sequence through a attention of words and focus on the most important parts of the input sentence. Here, the weight represents the dependence of current output on each words, where "1" means totally dependent and "0" represents completely independent.

The key issue of attention here is to calculate the weight α_i of word c_i in a sentence S automatically. Here, α_i is computed by:

$$u_i = \mathbf{w}^{\mathrm{T}}\mathbf{y}_i, \quad (10)$$

$$\alpha_i = e^{u_i} \left(\sum_{i=1}^{n} e^{u_i} \right)^{-1}, \quad (11)$$

where $\mathbf{y}_i$ is generated by highway network, and $\mathbf{w}$ is a parameter vector to be trained. Note that $i \in \{1, 2, \cdots, n\}$ and n is the length of sentence S.

The vector representation $\mathbf{s}$ of sentence S is computed by a weighted sum of $\mathbf{y}_i$ as follows:

$$\mathbf{s} = \sum_{i=1}^{n} \alpha_i \mathbf{y}_i. \quad (12)$$

Finally, after applying a nonlinear function on $\mathbf{s}$, we obtain:

$$\mathbf{s}^{\star} = \tanh(\mathbf{s}). \quad (13)$$

2) ATTENTION OF SENTENCES

It is assumed in distant supervision that a sentence including two entities represents this entity pair's relation in the knowledge base. However, it may generate a large number of noise training samples, which seriously damages the final results [33]. To address this issue, the distant supervised relation extraction is treated as a multi-instance problem in some works [12]–[14].

Here, the attention of sentences is utilized to compute the weights of all sentences towards a specified entity pair. Thus, the noise caused by distant supervision could be greatly reduced.

Let L be a set containing all sentences for a specific entity pair. Suppose $L = \{S_1, S_2, \cdots, S_k\}$ and k is the number of sentences that belong to a same entity pair. Firstly, we calculate the weight β_j of sentence $S_j\,(j = 1, 2, \cdots, k)$ for the entity pair by:

$$v_j = (\mathbf{A}\mathbf{s}_j^{\star})^{\mathrm{T}}\mathbf{d}, \tag{14}$$

$$\beta_j = e^{v_j}\left(\sum_{j=1}^{k} e^{v_j}\right)^{-1}, \tag{15}$$

where $\mathbf{A}$ is a a weighted diagonal matrix and $\mathbf{d}$ is a parameter vector related to the relation class.

And then, the vector representation $\mathbf{l}$ of L can be obtained by a weighted sum of $\mathbf{s}_j^{\star}$:

$$\mathbf{l} = \sum_{j=1}^{k} \beta_j \mathbf{s}_j^{\star}. \tag{16}$$

E. CLASSIFYING AND REGULARIZATION

In our model, softmax classifier is used to obtain the conditional probability for each relation class and arg max function is used to pick the predicted relation with maximum probability:

$$\mathrm{P}(r|L,\theta) = \mathrm{softmax}(\mathbf{W}_l\mathbf{l} + \mathbf{b}_l), \tag{17}$$

$$\widehat{r} = \arg\max_{r} \mathrm{P}(r|L,\theta), \tag{18}$$

where θ means all parameters of the proposed model, and $\mathbf{W}_l$ and $\mathbf{b}_l$ represent the corresponding weight matrix and bias vector of L, respectively. Here, softmax(x_i) is a function calculated by $e^{x_i}\left(\sum_{i=1}^{n} e^{x_i}\right)^{-1}$, where vector $\mathbf{x} = [x_i]_{i=1}^{n} \in \mathbb{R}^n$.

We define the cost function through the cross entropy of real relation and predicted relation. Moreover, we need an optimized model that could explain the training data well with smallest complexity. Therefore, a regularizer of L_2-norm is added behind the empirical risk. Regularization is the function of selecting the model whose empirical risk and model complexity are both as small as possible and could avoid overfitting better. Then, the cost function is calculated by:

$$\mathcal{J}(\theta) = -\frac{1}{m}\sum_{i=1}^{m} t_i \log(r_i) + \frac{\lambda}{2}\|w\|^2, \tag{19}$$

where the first term is empirical risk and the second term is regularizer. Here, $\mathbf{t} \in \mathbb{R}^m$ represents the one-hot represented ground truth of L, r_i is the conditional probability of the i-th relation class through the computation of softmax, and m represents the number of relation classes. In (19), all weight parameters w are constrained through a regularizer of L_2-norm, where λ is the L_2-norm hyperparameter.

Additionally, through the use of dropout technique [34], we can effectively alleviate overfitting on the BGRU layer. The details of training would be found in Section IV.

IV. EXPERIMENT RESULTS AND DISCUSSION

In this section, our experiments are conducted on New York Times corpus with the purpose of verifying the effectiveness of adding highway network on the typical GRU-based model, while demonstrating the performance of our proposed model. Moreover, the hyperparameter settings and the impact of those parameters used in highway network layer are analyzed.

A. DATASET DESCRIPTION

The proposed model is evaluated on a real-world dataset which is originally released in [23]. This dataset was generated through aligning New York Times corpus with Freebase on the basis of distant supervision. Here, the entity mentions in text were found from those phrases which were tagged identically by the Stanford Named Entity Recognizer, and the phrases were matched to the entities of Freebase. This dataset as a benchmark one has been popular in many performance test tasks for the relation extraction [12], [13].

TABLE 1. Statistics of dataset.

Training set	The number of sentences	522,611
	The number of entity pairs	281,270
	The number of facts	18,252
Test set	The number of sentences	172,448
	The number of entity pairs	96,678
	The number of facts	1,950

In the experiments, 53 relationships are defined and NA (not any) relation is included in particular to represent the irrelevance between two entities. Hold-out validation is used for the model evaluation. The New York Times corpus is randomly divided into two mutually exclusive datasets. One is defined as train set. The other is defined as test set to evaluate the generalization performance of the proposed model. The detailed statistics of this dataset are showed in Table 1.

B. EXPERIMENTAL SETUP

Here, we train the proposed model and update model parameters through the back-propagation algorithm. We obtain word embedding by Word2Vec tool and the dimension of word vectors is set to 50. The whole hyperparameter settings of our model are summarized in Table 2. We use Adam [35] as an optimizer with the learning rate of 0.0001. The batch size b is the hyperparameter which means the number of entity pairs during one training or test, and it is fixed to 50. The hyperparameter l_S represents the fixed length of sentences, and it is set to 70. The regularization strength λ represents the decay rate of weight, and it is fixed as 10^{-4}. Meanwhile, the regularizer is unavailable when λ is set to 0.0. Dropout

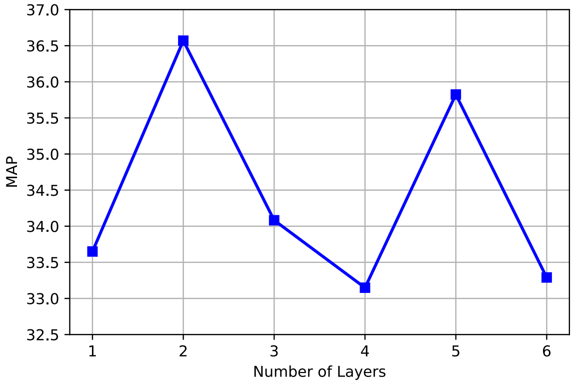


FIGURE 2. Impact of the number of highway network layers l_H.

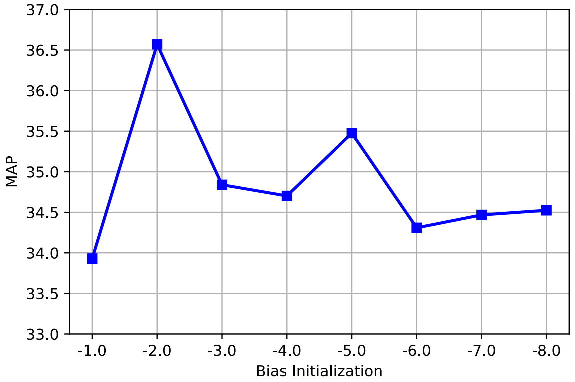


FIGURE 3. Impact of the initialization of $\mathbf{b}_T$.

TABLE 2. Hyperparameter settings.

Hyperparameter name	Hyperparameter setting
Sentence length l	70
Batch size b	50
Word dimension d_a	50
Position dimension (head entity) d_h	5
Position dimension (tail entity) d_t	5
Number of BGRU hidden units μ	230
Number of highway network layers l_H	2
Learning rate η	0.001
Regularization strength λ	0.0001
Dropout keep probability ρ	0.5

TABLE 3. AP comparison between Att-BGRU and Att-BGRU-HN models.

Model	AP (%)
Att-BGRU	35.09
Att-BGRU-HN	**36.57**

TABLE 4. Precision comparison of Att-BGRU and Att-BGRU-HN models.

Multi-Instance	Top N	Att-BGRU (%)	Att-BGRU-HN (%)
One	100	80.00	**80.00**
	200	68.50	**71.50**
	300	65.00	**65.00**
	Average	71.20	**72.20**
Two	100	82.00	**80.00**
	200	72.00	**72.00**
	300	65.00	**68.30**
	Average	73.00	**73.40**
All	100	83.00	**83.00**
	200	74.00	**77.00**
	300	69.00	**70.30**
	Average	75.30	**76.80**

hyperparameter ρ means the probability that each NN unit is kept and we set it as 0.5.

In term of highway network layer, the important hyperparameters in our model are the number of highway network layers l_H and the initialization of transform gate's bias $\mathbf{b}_T$. Here, the grid search is utilized to determine the optimum values of these two hyperparameters. We artificially define the range of l_H within $\{1, 2, \cdots, 6\}$. According to the fact that $\mathbf{b}_T$ is initialized to a negative value aiming to initially enlarge the carry behavior [21], we manually define the range of $\mathbf{b}_T$ within $\{-8.0, -7.0, \cdots, -1.0\}$. The impact of these two hyperparameters on model performance measured by average precision (AP) are shown in Figs. 2 and 3, respectively, where other hyperparameters of the model have been tuned to their optimum values.

C. EXPERIMENTAL RESULTS

Based on the above experimental corpus and settings, we conduct the experiments to evaluate the effectiveness of adding highway network and the performance of our model on the relation extraction task.

1) EFFECTIVENESS OF HIGHWAY NETWORK LAYER

To investigate the effectiveness of highway network layer, two models are trained on New York Times corpus in the same experimental environment. One is the Att-BGRU model, and the other is the proposed Att-BGRU-HN model. For all test data, the AP values of these two models are calculated as the metric, and the final results are shown in Table 3. It can be easily seen from the results that the addition of highway network layer could improve the performance of relation extraction commendably.

In addition, we compute the precisions for the top N ($N = 100, 200, 300$) of these two model on the entity pairs with more than one sentence in our corpus, so that the performance of sentence-level attention mechanism could be seen obviously. The comparison results are shown in Table 4, where the "Multi-Instance" column means the number of sentences we select and use to predict the relationship type for each entity pair to be tested. The result of Table 4 shows that the addition of highway network layer helps to improve the performance of sentence-level attention and extract much more comprehensive semantics features between words in a sentence.

2) COMPARISON WITH SOME EXISTING RELATION EXTRACTION MODEL BASED ON DISTANT SUPERVISION

To evaluate the performance of the proposed model in the relation extraction task, we additionally compare this model with four popular relation extraction methods based on distant supervision. They are Mintz's method [11], Hoffmann's method [12], Surdeanu's method [13], and Att-CNN [14].

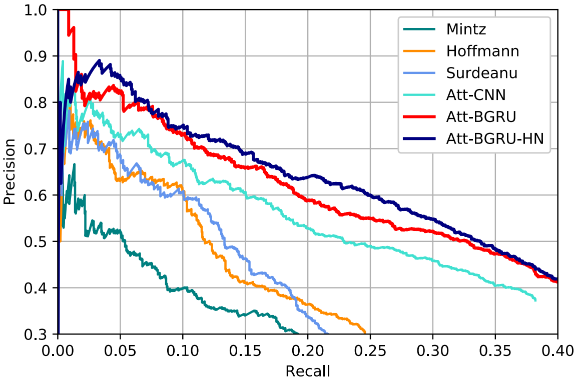


FIGURE 4. **Comparison of our model with some other popular models.**

Fig. 4 clearly shows the comparison results through precision with the increase of recall value, where computational performance of Att-BGRU and Att-BGRU-HN models is also compared. It is observed from the this figure that the precision value of our model is higher than any other model when the recall rate changes. Therefore, we could draw the conclusion that our Att-BGRU-HN model may outperform those popular relation extraction methods based on distant supervision. The additional highway network layer enables our method to effectively capture much more semantic features between words in the task of relation extraction.

V. APPLICATION OF ATT-BGRU-HN MODEL IN THE ANALYSIS FOR CHINESE GEOLOGICAL DATA

In addition to the above experimental comparison on a popular real-world dataset, we specifically provide an application case in the analysis for Chinese geological data, using our Att-BGRU-HN model.

Generally, the geological data is the veritable "big data" with huge quantity and complex types. In recent years, geological data service faces the dual demands of digitization and socialization for both institutions and the public [36]. Then, applying the idea of big data and machine learning techniques to the mining of geological data plays an important role to achieve satisfactory application performance [37].

The fragmented and unstructured data are a big part of the geological data [38], [39]. Meanwhile, it is obviously that text data is an important part of unstructured geological data, and the automatic extraction of relations between entities from text has always been an important research direction in unstructured data mining [40].

In this section, we apply the proposed relation extraction model to the field of Chinese geology. In particular, we select petrology area as the application object. In accordance with the characteristics of petrology, we explore the method of combining Chinese geological data with neural relation extraction based on deep learning, in an effort to further improve mining performance in dealing with geological data.

It is noted here that each sentence is originally a natural segmented word set in English corpus. However, the basic unit of a Chinese sentence is sinogram and therefore word segmentation is necessary. The popular segmentation methods include dictionary based methods and statistical machine learning methods, such as hidden Markov model (HMM), conditional random field (CRF), and deep learning model. Among those available approaches, Jieba [41], which utilizes HMM to identify out-of-vocabulary words and combines user-defined dictionary to adapt the word segmentation in a specific field, is a suitable word segmentation tool for the geological field.

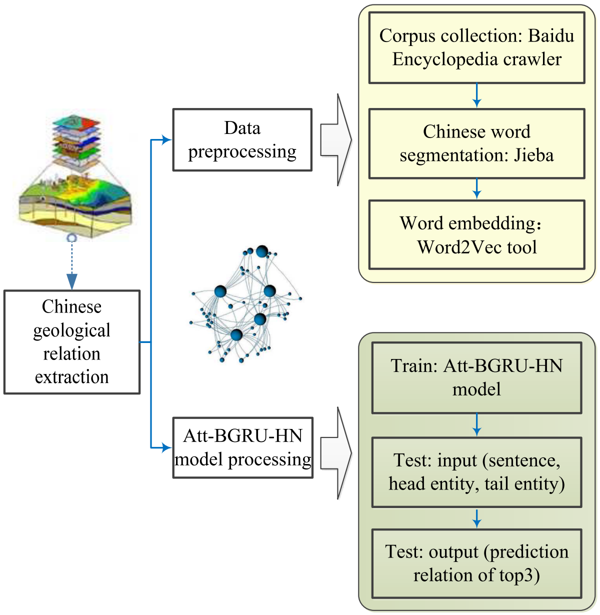

Chinese geological relation extraction
Data preprocessing
Att-BGRU-HN model processing
Corpus collection: Baidu Encyclopedia crawler
Chinese word segmentation: Jieba
Word embedding: Word2Vec tool
Train: Att-BGRU-HN model
Test: input (sentence, head entity, tail entity)
Test: output (prediction relation of top3)

FIGURE 5. **Schematic diagram of Chinese geological relation extraction.**

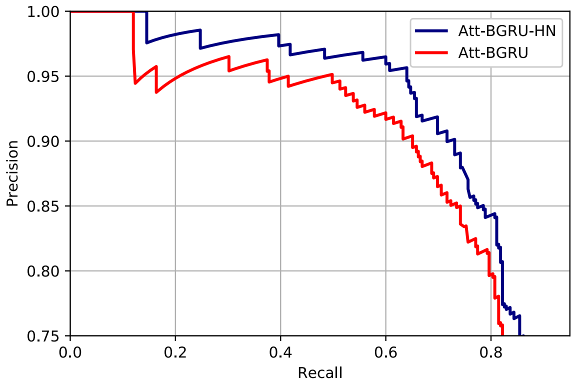


FIGURE 6. **Comparison of our model with Att-BGRU for Chinese geological data.**

Here, the schematic diagram of Chinese geological relation extraction with the proposed model is shown in Fig. 5. In the beginning, the web crawler technology is utilized to collect sentences of geological rocks from Baidu Encyclopedia, and Jieba is used to divide the sentences into words. Secondly, on the basis of distant supervision, we align the three tuples of Chinese geological thesaurus with the above segmented sentence set, which is called as the annotation of the corpus. Here, the final number of annotated sentences

TABLE 5. Some example results on chinese geological relation extraction task.

Head entity	Tail entity	Sentence	True relation	Test relation in top 3 (probability)
霞石	灰色	霞石无色或白色，有时也呈灰色、绿色或红色。	颜色	1 颜色 (0.999997) 2 结构 (1.68556e-6) 3 NA (4.69314e-7)
Nepheline	Gray	Nepheline is colorless or white, sometimes gray, green or red.	Color type	**1 Color type (0.999997)** **2 Structure (1.68556e-6)** **3 NA (4.69314e-7)**
钛铁矿矿山	伊尔门山	著名钛铁矿矿山有俄罗斯的伊尔门山、挪威的克拉格勒和美国怀俄明州的铁山、加拿大魁北克的埃拉德湖等。	产地	1 产地 (0.999983) 2 相关人物 (9.30787e-6) 3 NA (3.13827e-7)
Ilmenite mines	Il'menskiy Khrebet	Russian Il'menskiy Khrebet, Norway Kragero, Iron Mountain in Wyoming, USA and Ellard Lake in Quebec, Canada are famous ilmenite mines.	Producer	**1 Producer (0.999983)** **2 Related person (9.30787e-6)** **3 NA (3.13827e-7)**
月岩	月壤	自1969年美国“阿波罗”11 号登月以来，共采回380多千克月岩样品，按样品的结构和成因，月岩可分为3 类，即：结晶质火成岩、角砾岩和月尘。	包含	1 包含 (0.999154) 2 伴生 (0.000506) 3 相关人物 (0.000139)
Lunar rock	Lunar soil	Since 1969 the United States “Appollo” 11 lunar years, were collected more than 380 kilograms of lunar rock samples, according to the structure and origin of the sample, the rocks can be divided into 3 categories, namely: crystalline igneous rock, breccia and lunar soil.	Contain	**1 Contain (0.999154)** **2 Accompany (0.000506)** **3 Related person (0.00013)**
流纹岩	斑状	霞大多数流纹岩都具斑状结构，表明结晶作用在喷发作用以前就已开始。	结构	1 结构 (0.993710) 2 颜色 (0.003901) 3 构造 (0.001656)
Rhyolite	Porphyritic	Most rhyolite have porphyritic structure, which indicates that the crystallization began in the previous eruption.	Structure	**1 Structure (0.993710)** **2 Color type (0.003901)** **3 Construction (0.001656)**
袁奎荣	花岗岩	著名地质学家袁奎荣在花岗岩构造学、显微造与组构学以及隐伏花岗岩与立体找矿等领域较有研究。	相关名人	1 相关人物 (0.660765) 2 发现地 (0.132889) 3 产地 (0.132351)
Yuan Kuirong	Granite	The famous geologist Yuan Kuirong had done much research in the field of granite tectonics, microstructure, concealed granite and stereo prospecting.	Related person	**1 Related person (0.660765)** **2 Discovery site (0.132889)** **3 Producer (0.132351)**
地质学	学科	地质学是五大基础学科之一，其他四个分别是数学，物理，化学，生物。	是一种	1 是一种 (0.999694) 2 别称 (0.000137) 3 包含 (9.20257e-05)
Geology	Subjects	Geology is one of the basic subjects, and the other four are mathematics, physics, chemistry and biology.	Is a kind of	**1 Is a kind of (0.999694)** **2 Another name (0.000137)** **3 Contain (9.20257e-05)**

is 3050 and training set account for 80.00%. Thirdly, via the proposed relation extraction model, Chinese geological relation extraction task is transformed into a classification problem. We could obtain conditional probability of all relation types for test samples and select the relation type with maximum probability.

The number of relation types in our Chinese geological data is 22, and it includes the most common relation between rocks in the field of petrology. As mentioned before, NA relation is also included to represent the irrelevance between two entities. And the hyperparameter settings are same as those values in Table 2.

The final AP in this task using our proposed model is 88.10%. There may exist small errors in the results because of the limited experimental corpus. However, it could be found that when the number of relation classes decreases, the accuracy of relation extraction results is greatly increased, and the proposed model is especially suitable for the geological relation extraction. The results comparison between the proposed model with Att-BGRU model are seen in Fig. 6, which shows the effectiveness of introducing highway network in the task of Chinese geological relation extraction.

We randomly list the test results of six sentences, and they are shown in Table 5. The fourth column represents the true relationship of samples and the fifth column represents the prediction results of top 3 obtained through our model. The values appeared in parentheses represent their corresponding probability. We can easily see that our model Att-BGRU-HN achieves satisfactory performance in the analysis for Chinese geological data.

VI. CONCLUSION

In this article, we propose a neural relation extraction model, named Att-BGRU-HN, for the relation extraction task. By incorporating highway network into a typical attention-based BGRU network, a novel method is accordingly developed. Experimental results on New York Times corpus aligned with Freebase demonstrate that the proposed model achieves an improvement when highway network layer is added, and outperforms some existing relation extraction methods based on distant supervision. It can be concluded that the introduction of highway network could enable our approach to capture much more useful and semantic information between words. Furthermore, to demonstrate the effectiveness of our model in real data mining applications, we apply it in Chinese geological field and also achieve a satisfactory performance for the task of relation extraction. In the future, we will continue to explore the application of deep learning in relation extraction and analyze the convergence of the proposed model.

REFERENCES

[1] N. Bach and S. Badaskar, "A survey on relation extraction," Lang. Technol. Inst., Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep., 2007.

[2] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *J. Mach. Learn. Res.*, vol. 12, pp. 2493–2537, Nov. 2011.

[3] W. Zhao *et al.*, "A human-centered activity tracking service: Towards a healthier workplace," *IEEE Trans. Human Mach. Syst.*, vol. 47, no. 3, pp. 343–355, Jun. 2017.

[4] R. C. Bunescu and R. J. Mooney, "A shortest path dependency kernel for relation extraction," in *Proc. Conf. Human Lang. Technol. Empirical Methods Natural Lang. Process.*, Oct. 2005, pp. 724–731.

[5] I. Hendrickx *et al.*, "SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals," in *Proc. ACL Int. Workshop Semantic Eval.*, Jun. 2009, pp. 94–99.

[6] S. Zhang, D. Zheng, X. Hu, and M. Yang, "Bidirectional long short-term memory networks for relation classification," in *Proc. Pac. Asia Conf. Lang., Inf. Comput.*, Oct. 2015, pp. 73–78.

[7] X. Yan, L. Mou, G. Li, Y. Chen, H. Peng, and Z. Jin, "Classifying relations via long short term memory networks along shortest dependency path," in *Proc. Conf. Empirical Methods Nat. Lang. Process.*, Sep. 2015, pp. 1785–1794.

[8] X. Luo *et al.*, "Towards enhancing stacked extreme learning machine with sparse autoencoder by correntropy," *J. Franklin Inst.*, to be published, doi: 10.1016/j.jfranklin.2017.08.014.

[9] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *J. Mach. Learn. Res.*, vol. 3, pp. 1137–1155, Feb. 2003.

[10] T. Mikolov, M. Karafiát, L. Burget, C. Jan, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Sep. 2010, pp. 1045–1048.

[11] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, "Distant supervision for relation extraction without labeled data," in *Proc. 47th Annu. Meeting ACL 4th Int. Joint Conf. Natural Lang. Process. AFNLP*, Aug. 2009, pp. 1003–1011.

[12] R. Hoffmann, C. Zhang, X. Ling, L. Zettlemoyer, and D. Sweld, "Knowledge-based weak supervision for information extraction of overlapping relations," in *Proc. Annu. Meeting Assoc. Comput. Linguist, Human Lang. Technol.*, Jun. 2011, pp. 541–550.

[13] M. Surdeanu, J. Tibshirani, R. Nallapati, and C. D. Manning, "Multi-instance multi-label learning for relation extraction," in *Proc. Joint Conf. Empirical Methods Nat. Lang. Process. Comput. Natural Lang. Learn.*, Jul. 2012, pp. 455–465.

[14] Y. Lin, S. Shen, Z. Liu, H. Luan, and M. Sun, "Neural relation extraction with selective attention over instances," in *Proc. Annu. Meet. Assoc. Comput. Linguist*, Aug. 2016, pp. 2124–2133.

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[16] K. Cho, B. Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Oct. 2014, pp. 1724–1734.

[17] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. (Dec. 2014). "Empirical evaluation of gated recurrent neural networks on sequence modeling." [Online]. Available: https://arxiv.org/abs/1412.3555

[18] L. Shang, Z. Lu, and H. Li, "Neural responding machine for short-text conversion," in *Proc. Annu. Meet. Assoc. Comput. Linguist Int. Joint Conf. Natural Lang. Process. Asian Fed. Natural Lang. Process.*, Jul. 2015, pp. 1577–1586.

[19] R. Kiros *et al.*, "Skip-thought vectors," in *Proc. Conf. Neural Inf. Process. Syst.*, Dec. 2015, pp. 2394–3302.

[20] P. Zhou *et al.*, "Attention-based bidirectional long short-term memory networks for relation classification," in *Proc. Annu. Meeting Assoc. Comput. Linguist*, Aug. 2016, pp. 207–212.

[21] R. K. Srivastava, K. Greff, and J. Schmidhuber. (May 2015). "Highway networks." [Online]. Available: https://arxiv.org/abs/1505.00387

[22] Y. Kim, Y. Jernite, D. Sontag, and A. M. Rush, "Character-aware neural language models," in *Proc. AAAI Conf. Artif. Intell.*, Feb. 2016, pp. 2741–2749.

[23] S. Riedel, L. Yao, and A. McCallum, "Modeling relations and their mentions without labeled text," in *Proc. Eur. Conf. Mach. Learn. Prins Pract. Knowl. Discovery Databases*, Sep. 2010, pp. 148–163.

[24] D. Zeng, K. Liu, S. Lai, G. Zhou, and J. Zhao, "Relation classification via convolutional deep neural network," in *Proc. Int. Conf. Comput. Linguist*, Aug. 2014, pp. 2335–2344.

[25] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994.

[26] D. Zeng, K. Liu, Y. Chen, and J. Zhao, "Distant supervision for relation extraction via piecewise convolutional neural networks," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Sep. 2015, pp. 1753–1762.

[27] C. N. D. Santos, B. Xiang, and B. Zhou, "Classifying relations by ranking with convolutional neural networks," in *Proc. Annu. Meeting Assoc. Comput. Linguist Int. Joint. Conf. Natural Lang. Process. Asian Fed. Natural Lang. Process.*, Jul. 2015, pp. 626–634.

[28] M. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Sep. 2015, pp. 1412–1421.

[29] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Sep. 2015, pp. 379–389.

[30] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio. (Dec. 2014). "End-to-end continuous speech recognition using attention-based recurrent NN: First results." [Online]. Available: https://arxiv.org/abs/1412.1602

[31] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. Conf. Adv. Neural Inf. Process. Syst.*, Dec. 2015, pp. 577–585.

[32] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2013, pp. 6645–6649.

[33] S. Takamatsu, I. Sato, and H. Nakagawa, "Reducing wrong labels in distant supervision for relation extraction," in *Proc. Conf. Annu. Meet. Assoc. Comput. Linguist*, Jul. 2012, pp. 721–729.

[34] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. (Jul. 2012). "Improving neural networks by preventing coadaptation of feature detectors." [Online]. Available: https://arxiv.org/abs/1207.0580

[35] D. P. Kingma and J. L. Ba. (May 2015). "Adam: A method for stochastic optimization." [Online]. Available: https://arxiv.org/abs/1412.6980

[36] T. Zhang, Y. Du, T. Huang, and X. Li, "Stochastic simulation of geological data using isometric mapping and multiple-point geostatistics with data incorporation," *J. Appl. Geophys.*, vol. 125, pp. 14–25, Feb. 2016.

[37] M. G. Runge, M. S. Bebbington, S. J. Cronin, J. M. Lindsay, and M. R. Moufti, "Integrating geological and geophysical data to improve probabilistic hazard forecasting of Arabian shield volcanism," *J. Volcanol. Geothermal Res.*, vol. 311, pp. 41–59, Feb. 2016.

[38] Y. Zhu, W. Zhou, Y. Xu, J. Liu, and Y. Tan, "Intelligent learning for knowledge graph towards geological data," *Sci. Program.*, vol. 2017, Feb. 2017, Art. no. 5072427.

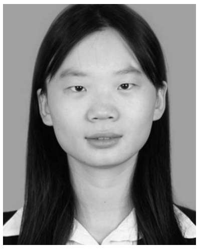
[39] L. Zhang, "Improvement of K-means algorithm and its applications in analysis of geological exploration seismic data," *Electron. J. Geotech. Eng.*, vol. 20, no. 12, pp. 4423–4434, 2015.

[40] Y. Zhu, Y. Tan, R. Li, and X. Luo, "Cyber-physical-social-thinking modeling and computing for geological information service system," *Int. J. Distrib. Sens. Netw.*, vol. 12, no. 11, pp. 1–9, Nov. 2016.

[41] (Jan. 2018). *Jieba: Chinese Text Segmentation*. [Online]. Available: https://github.com/fxsjy/jieba

XIONG LUO received the Ph.D. degree in computer applied technology from Central South University, Changsha, China, in 2004.

He is currently a Professor with the School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, China. He has published extensively in his areas of interest in several journals, such as IEEE Transactions on Industrial Informatics, IEEE Transactions on Human-Machine Systems, IEEE Access, *Future Generation Computer Systems*, and *Journal of The Franklin Institute*. His current research interests include machine learning, cloud computing, and computational intelligence.



WENWEN ZHOU is currently pursuing the master's degree with the University of Science and Technology Beijing, Beijing, China. Her current research interests include deep learning and knowledge engineering.

WEIPING WANG received the Ph.D. degree from the Beijing University of Posts and Telecommunications, China, in 2015. She is currently an Associate Professor with the University of Science and Technology Beijing. Her current research interests include neural networks and computational intelligence.

YUEQIN ZHU received the Ph.D. degree from Technical University of Munich, Germany, in 2012. She is currently a Senior Engineer with the Key Laboratory of Geological Information Technology, Ministry of Land and Resources, Development and Research Center, China Geological Survey, China. Her current research interests include cloud computing, computational intelligence, and cartograhpy.

JING DENG is currently pursuing the master's degree with the University of Science and Technology Beijing, Beijing, China. Her current research interests include machine learning and cyber-physical systems.