Chinese Abrupt Event Recognition Based on CBiGRU-Att Model

Haoran Yin, Jinxuan Cao*, Yanhui Du, Guodong Wang, Luzhe Cao, Xing Wang and Houlu Zhang
People's Public Security University of China, Beijing 100035, China

*Corresponding author e-mail: caojinxuan@163.com

Abstract. Aiming at the problem of poor portability of traditional event recognition methods, the need for a large number of learning features, and the poor interpretability of recurrent neural networks in different information features about degrees of importance, this paper proposes a Chinese abrupt event recognition method based CBiGRU-Att model. Firstly, the text corpus was preprocessed. Word2vec was used to generate word vectors and the local features of the word vectors were extracted by using the convolutional layer. Then the extracted features were used as the input of the BiGRU to obtain higher-order context features, and introduced the attention mechanism to weight feature. Finally, softmax function was used to activate the learned features and output the recognition results. Simulation results show that this method is superior to other methods in the precision and recall rate for Chinese abrupt event recognition.

1. Introduction

As a manifestation of information, event is defined as the objective fact that specific people and objects interact with each other at a specific time and place [1]. Event extraction mainly consists of two steps: (1) event recognition, (2) event analysis, and then extraction of event elements. Event recognition aims to obtain the structured event information that users are interested in from the unstructured text information and classify the corresponding events [2]. In event recognition, Event type classification is one of the core tasks. However, when classifying event types, two problems are often encountered: (1) same type may contain discriminating elements of other types of events; (2) The discriminant element of certain type event can appear in different events of type and in different contexts. This relies on the extracted features to solve these problems.

Traditional event extraction methods need to manually design features and annotate corpus, which consumes a lot of time and manpower. The selection of features directly affects the final effect of relational classifiers, and the use of NLP (natural language processing) annotation tools is likely to lead to the problem of error propagation [3]. Therefore, this paper applies the method of deep learning and proposes the CBiGRU-Att joint model to recognize Chinese abrupt event.

2. Related research

At present, research methods for event recognition in various fields are mainly divided into three categories: pattern matching, machine learning and deep learning.
2.1. Pattern Matching
Jiang jifa [4] proposed the model of GenPAM and conducted experiments in 100 corpus of flight accidents, with F value reaching 63%. Zheng jiaheng et al. [5] used clustering method to automatically generate information extraction mode for Chinese text, and conducted experiments in crop information text corpus in the agricultural field. The misclassification rate and misclassification rate were 0.21% and 1.07% respectively. Although the pattern-based matching method has a high recognition rate in the specific event domain, it needs the expert knowledge in this domain to build a large-scale knowledge base, which requires a lot of manpower, has poor portability, and is not applicable to the event recognition in the current open domain.

2.2. Machine Learning
Chieu et al. [6] introduced a maximum entropy classifier into event extraction to realize the extraction of personnel management events, and the F value reached 59.2%. Xu honglei et al. [7] used a classifier to filter non-event sentences and conducted experiments in ACE2005 Chinese corpus, showing that F value reached 76%. Zhao yanyan et al. [8] used the combination of trigger word expansion and classification to identify event categories, and conducted experiments on ACE2005 corpus, with F value reaching 64.64%. Based on the traditional machine learning method, although the recognition of event-triggered words is simple and efficient, and the recognition efficiency is improved compared with the pattern matching, to establish an accurate model, a large number of experiments and learning are needed to acquire learning features, and more linguistic knowledge is needed to design features.

2.3. Deep Learning
Wen chang et al. [9] used the LSTM network model with the attention mechanism to extract the evolution relationship of abrupt event, and carried out experiments in the built corpus. The F value reached 89.6%. He Xinyu [10], such as two-way LSTM and two stage method of trigger word, which can identify all the words will word vector and the corresponding training beforehand word after vector and fine-tuning the word vector difference and average sentence vector as input, to identify the trigger word, and the trigger in the MLEE corpus word recognition experiment, the F value were 73.62% and 77.13% respectively; Wang hongbin et al.[11] used neural network as classifier for event classification, word vector as the input of neural network to classify the semantic meaning of event sentences, used dependency analysis to mine the relationship between words, and conducted experiments in CEC corpus, and the F value reached 78.2%.

Based on the deep learning method, although can extract the characteristics of events, but a single neural network model, such as convolution neural network dependence of long-distance information, and circulation of neural network is good at capturing text context semantic characteristics, therefore, based on the existing research, this paper puts forward the CBiGRU - Att combined neural network model, using convolution neural network is good at extracting N - "gramm local semantic characteristics, the advantages of using cycle the learning ability of neural network for the context sequence information, use attention mechanism important degree of different characteristics can be interpreted, The text word vector features and context information features are extracted effectively, and the associated model is constructed to recognize the corresponding abrupt event.

3. Model
Considering that convolutional neural network is good at capturing local semantic features of text and it is relatively simple to train, but it is dependent on information from a long distance, and the size of convolution kernel cannot be too large, so it cannot effectively identify and represent temporal information such as context. Considering that the recurrent neural network is good at capturing the context semantic feature of the text, it can take advantage of the historical
information of the context and integrate the order of words into it, but the convergence is very slow in training. Therefore, the two neural networks are combined to integrate their advantages and introduce attention mechanism to build a CBiGRU-Att joint model for abrupt event recognition. The event recognition based on CBiGRU-Att joint model proposed in this paper includes five parts: input layer, convolution layer, recurrent neural network layer, attention layer and softmax layer. The whole event recognition process is shown in Fig.1.

Fig.1. CBiGRU-Att model

3.1. Input Layer
It is a common idea of English event detection to classify every word in the sentence through neural network, find out the existing trigger words and classify them [12]. However, for languages without natural segmentation, such as Chinese, we need to conduct word segmentation first, and then generate word embedding vector by word segmentation text. In this paper, stuttering participles [13] are adopted to conduct word segmentation on the text, and stop words are removed from the text after word segmentation. MIKOLOV [14] et al. proposed word2vec model for training the preprocessed text to obtain the word vector. The training mode used Skip gram model [15], and then the word vector was taken as the input of the convolution layer. By using the form of word vector instead of the traditional one-hot encoding method, the problem of dimension disaster can be avoided and the vector space can be transformed from sparse high-dimensional to dense low-dimensional.

3.2. Convolution Layer
In the task of event recognition, the operation of the convolution layer is to get vector features by sliding the convolution kernel on the word sequence. The main purpose of convolution operation is to obtain word vector features and then express them in the form of higher-order context information features. The process of extracting word vector features by convolution operation is shown in Fig.2.
Let inputted word vector (yunnan, ruili, earthquake...) of the sentence is \(V = \{v_1, v_2, \ldots, v_r\}\), where \(v_i \in \mathbb{R}^n\) (\(i = 1, 2, \ldots, r\)). In order to obtain the features of the word level, \(m\) convolution kernel \(w\) is selected, and their structure is \(k \times n\), namely the matrix of \(k\) rows and \(n\) columns. Therefore, the word vector can be expressed as:

\[
y'_i = f \left( w \ast V' + b \right)
\]

(1)

Notes: the "\(*\)" operator represents the multiplication of corresponding elements of the matrix; \(V'\) represents the \(k \times n\) matrix composed of the word sequence vector \([v_1, v_2, \ldots, v_k]\) in \(V\); \(b\) is the offset vector with \(n\) dimension, and \(f\) is the nonlinear activation function. Since the convolutional neural network needs to input sentences of fixed length, the input sentences are filled to a fixed length by means of filling, while the whole sentence length remains unchanged after the convolution operation. The convolution kernel is used to scan each word vector successively to obtain the word vector features of the whole sentence:

\[
F = \begin{bmatrix} y_1', y_2', \ldots, y_r' \end{bmatrix}
\]

(2)

Notes: \(F \in \mathbb{R}^{r \times m}\) is a matrix, and \(y_i' (i = 1, 2, \ldots)\) is represented as \(r \times m\), in \(F\). The eigenvector representing each word after it has been convolved and kernel and activated by an activation function.

In convolution neural networks, generally after the convolution convolution layer maximum pooling operation would be taken to reduce the dimension of the output, the local characteristics of capture significant, however on the sequence of the text with the context dependent information feature extraction, pooling the operation will be lost in a sequence of word order between
information, destroy the sequence characteristics, so in the convolution section of the model abandoned the pooling operation, the word level on convolution layer extract local features as circulation layer neural network input.

3.3. BiGRU

In the task of event recognition, the operation of recurrent neural network is to capture the context features of sentences by using the word vector features extracted by the convolution layer as input. From the theory of network structure, traditional RNN can process sequence information of any length, but in practice, RNN is not strong enough to capture remote features, and its robustness is not strong. Moreover, the propagation of error of RNN will decrease or expand with the increase of network depth in the training process, that is, there may be gradient disappearance or gradient explosion, which makes the training results of RNN difficult to get the expected effect. As a variant of RNN, Long Short Term Memory network (LSTM) can well deal with the problem of remote information dependence that RNN cannot solve, so as to learn long-term dependent information. The data set used in this paper is not large, so Bidirectional Gated Recurrent Unit is adopted. GRU is a variant of LSTM. It simplifies LSTM a lot, and there is one Gate less than LSTM. As a result, it is more efficient in calculation, occupies less memory and is easier to converge. The unit structure of GRU is shown in Fig.3.

![Fig.3. GRU unit structure](image)

1) Update gete: \( z_t = \sigma(W_z [h_{t-1}, x_t]) \)
2) Reset gate: \( r_t = \sigma(W_r [h_{t-1}, x_t]) \)
3) Candidate memory unit: \( \tilde{h}_t = \tanh(W [r_t \cdot h_{t-1}, x_t]) \)
4) Current memory unit: \( h_t = (1 - z_t) \cdot h_{t-1} + z_t \cdot \tilde{h}_t \)

Notes: "[ ]" means two vectors are connected, and the "*" operator means matrix element multiplication.

From the above, it can be known that the spread mechanism of GRU is a unidirectional, so model can not capture the below information feature. In order to extract higher-order features of context information, based on the forward GRU, model adds backward GRU helped to make up the disadvantage, so get a bidirectional GRU. Via the forward GRU, calculating above information feature \( \tilde{h}_t \) at time \( t \), and at the same time using backward GRU calculated below information feature \( \hat{h}_t \) at time \( t \), so get a bidirectional GRU.
feature $h_t$ at time $t$. Then result can get $h_t = [h_{t-1}, h_t]$, and using the $h_t$ as input of attention layer. The process of bidirectional GRU extract higher-order features as shown in Fig. 4.

![Fig. 4. BiGRU extract context feature](image)

In the task of event recognition, using feature vector $F$ as the input of RNN layer, result get the feature vector $H$ with context information:

$$H = [h_1, h_2, \ldots, h_r]$$

(3)

Notes: $h_t \in \mathbb{R}^{2m}(t=1, 2, \ldots, r)$, $H \in \mathbb{R}^{r \times 2m}$

### 3.4. Attention Layer

In the task of abrupt event recognition, the operation of the attention layer is to automatically weight the feature information of the previous layer. The main purpose is to make the contribution of important feature information to the result be reflected in the weighting process so that the model can focus on extracting the features of several words. Specific operation steps are as follows:

1) After multiplying the input eigenvector $h_t$ and the weight matrix $w_a$ and adding the bias $b_a$, $u_t$ was obtained through the activation function tanh:

$$u_t = \tanh(h_tw_a + b_a)$$

(4)

2) Let $e_t = u(w_A^T, e)$, $E = (e1, e2, \ldots, e_m)$, $A = (\alpha_1, \alpha_2, \ldots, \alpha_m)$, $A = \text{softmax}(E)$, and get the weight $\alpha_t$ automatically assigned by each eigenvector:

$$\alpha_t = \frac{\exp(u_t^w)}{\sum \exp(u_t^w)}$$

(5)
3) By multiplying the eigenvector $o_t$ at each moment and the weight $\alpha_t$ at the corresponding moment, the weighted summation result $S$ can be obtained:

$$ S = \sum \alpha_t o_t $$

Notes: $w_a, u$ present weight, and will be continuously optimized by model training after initialization; $o_t$ present feature vector of output matrix $O$; $b_a$ present offset; $w_a \in \mathbb{R}^{2m \times 2m}$, $b_a \in \mathbb{R}^{2m}$, $u \in \mathbb{R}^{2m}$, $S \in \mathbb{R}^{2m}$.

3.5. Output Layer

The operation of the output layer make the result $S$ of the previous layer activated through the final softmax layer. The layer uses softmax as the activation function to obtain the event category probability distribution $P$:

$$ P = \frac{\exp(Sw_{fc} + b_{fc})}{\sum \exp(Sw_{fc} + b_{fc})} $$

Notes: $w_{fc}$ present weight, $b_{fc}$ present offset; $w_{fc} \in \mathbb{R}^{2m \times c}$, $b_{fc} \in \mathbb{R}^c$, $P=[p_1, p_2, \ldots, p_i, \ldots, p_c]$ ($i \in [1, c]$), $c$ is number of abrupt event type. In the training stage, the cross entropy cost function is adopted as the loss function to obtain the loss value loss, and the learning rate is set as LR. The gradient descent is carried out by the optimizer Adam to minimize the loss, so that the training of the whole model can be iterated to convergence:

$$ loss = average(-\Sigma label*tf.log(p_i)) $$

AdamOptimizer(LR).minimize(loss)  

During validation and testing, arg_max function is used to obtain the index value of the maximum probability in the output vector $p_i$ to match the actual label of the event type:

$$ label = arg \_ max(p_i) $$

4. Experimental results and analysis

4.1. Abrupt Event Dataset

The Corpus of this paper adopts the CEC (Chinese Emergency Corpus) established by the semantic intelligence laboratory of Shanghai university, as well as the data obtained through the web-crawler. Corpus data are all derived from news reports of emergencies on the Internet and we-media data, which contain five types of emergencies: earthquake, fire, traffic accident, terrorist attack and food poisoning. The total number of CEC is 332; The data obtained by the crawler were 28,267, and 25,530 were counted after duplication eliminating. This paper conducted experiments on two kinds of data sets, randomly selected training sets, validation sets and test sets, and divided them into 7:2:1.

4.2. Training Stage

In the training stage, specific operations are as follows:
1) Use stuttering participles to process the training set text, obtain the participle text, and remove common stop words from the participle text. For example, punctuation, “的”, “了” et al;

2) The pre-processed training samples were trained by word2vec, and the training mode was skip-gram model. The word vector (200 dimension) was used as the feature vector to represent each word, so as to obtain the word vector \( v_i \);

3) If the processed word segmentation are counted and the text less than 300 words, the segmentation text will be filled up to 300 words. The shape of matrix formed by the input word vector is a 300×200.

4) Set 100 convolution kernel \( w \), their shape is 5×200 matrix. Input matrix \( V \) to get 300×100 matrix after convolution operation, and get feature matrix \( f \) after activation of nonlinear function \( f \).

5) Making \( F \) as the input of the recurrent neural network to obtain the context feature matrix \( H \) with the shape of 300×200. The \( H \) can make weighted summation through the attention layer to get the vector \( S \). The \( S \) is finally activated by softmax function to obtain category probability distribution \( P \).

6) The cross entropy cost function is adopted as the loss function to obtain loss value, and the learning rate is set as 1e-3 according to the empirical value. The Adam optimizer is called for gradient descent to minimize loss, so that the entire model can be trained to convergence.

4.3. Test Stage
In the test stage, the data preprocess is the same as in the training stage. The test sample was input into the trained model. Then the word feature extraction of the convolution layer, the context feature extraction of BiGRU and the weighted sum of the attention layer were processed, and the event type prediction probability was obtained through softmax function. Finally, the index value of the maximum probability in the output vector \( p_i \) was obtained through argmax function to match with the actual label of the event type, \( \text{label} = \text{arg\_max}(p_i) \), and the matching result was output.

4.4. Analysis
The recognition results of abrupt events were analyzed using a general method for performance evaluation, with the same standards as the document [16-20], including Precision, Recall and F-value. In this paper, a number of experiments were carried out, including the convolutional neural network, GRU and the joint model CNN-BiGRU, BiGRU-AM and CBiGRU-Att, and the results were compared with those of different models in other documents. The experimental contrast results in CEC dataset are shown in TABLE I.

| Model     | P/%  | R/%  | F/%  |
|-----------|------|------|------|
| SVM[16]   | 79.30| 59.90| 63.70|
| CNN       | 72.73| 64.00| 68.09|
| GRU       | 69.70| 66.67| 68.15|
| BiGRU[17] | 71.10| 69.00| 70.00|
| BiGRU-AM  | 75.76| 69.44| 72.46|
| CNN-BiGRU | 74.24| 71.01| 72.59|
| CBiGRU-Att| 76.81| 69.74| 73.10|

Since CEC corpus is treated with structured XML format language, after preprocess, the text feature is still obvious and easier to extract and training. That situation relatively weak convolution layer superiority in the local feature extraction aspect, so that it make loss of generality, and CEC fewer data may lead model lack of training. Therefore, the experiment used various models to conduct contrast experiments on unstructured information data obtained web-
crawler from Internet we-media, and the experimental comparison results were shown in TABLE II.

**Table 2. Experimental contrast results**

| Model       | P/%  | R/%  | F/%  |
|-------------|------|------|------|
| SVM         | 77.26| 53.43| 63.17|
| CNN         | 71.59| 62.69| 66.85|
| GRU         | 70.36| 64.96| 67.55|
| BiGRU       | 73.53| 67.56| 70.32|
| CNN-BiGRU  | 75.43| 72.29| 73.82|
| BiGRU-AM    | 76.32| 71.76| 73.96|
| CBiGRU-Att  | 77.53| 73.69| 75.56|

As can be seen from the experimental results, the traditional machine learning method SVM model was adopted in table I[16] and table II to extract trigger words, which achieved a high accuracy rate but a low recall rate. In terms of the effect of event recognition, the deep learning method is more effective than the traditional machine learning method, and the overall F-value is better than SVM, indicating that deep learning can extract more abstract features for learning. CNN uses convolutional neural network to recognize events, and GRU uses unidirectional recurrent neural network to recognize abrupt event. The F value is slightly higher than CNN, indicating that the context information captured by GRU is more conducive to the identification of abrupt event. In table I, the F-value of BiGRU model adopted in the document [17] reached 70.00%, which is nearly 2% higher than that of GRU, indicating that the bidirectional recurrent neural network is better than the unidirectional recurrent neural network in abrupt event recognition by capturing the features of context information. The CNN-BiGRU model output is as an input of bidirectional recurrent neural network. On the basis of bidirectional recurrent neural network model, BiGRU-AM introduce the attention mechanism. CBiGRU-Att combines the advantages of the above two models, improves the model ability in the local feature extraction, and strengthened the interpretability of different information feature, so that model has obtained the higher F value.

CBiGRU-Att joint model proposed in this paper achieve good effect compared with other methods in accuracy rate and recall rate and F value is superior to other methods. Indicating that combined with convolution, recurrent, and attention mechanism neural network has played a very good effect in the capture of local features, extraction of context semantic features, and important degree interpretability of different feature, and improves the performance of Chinese abrupt event recognition.

5. Conclusions

Through this research, this paper builds a CBiGRU-Att joint model. The model can effectively recognize the corresponding abrupt event. Compared with the traditional machine learning methods, this paper uses joint model based on the deep learning. The method has good generalization ability, and don’t need dependence for specific domain knowledge of experts. Relying on the neural network combination model of autonomous learning can avoid the large amount of manpower and not build expansion table of the trigger word. The portability is good; Compared with the single neural network model, CNN and RNN, the joint model can combine their advantages and integrate other functions, so as to extract local feature information of abrupt events, obtain higher-order context features, and distinguish the importance of different feature information.

The next step is to introduce attributes such as part of speech, named entity, dependency relationship and semantic role to generate word vector, so as to add more semantic features to
expand the dimension of word vector, and deepen the depth of convolution layer, capture more abstract higher-order features and further improve the effect of event recognition.

Acknowledgments
This work was supported by the National Key R&D Program of China (2016YFB0801100, 2018YFC0823200) and the Fundamental Research Funds for the Central Universities of PPSUC (2019JKF108). Correspondence should be addressed to Jinxuan Cao: caojinxuan@163.com.

References
[1] Qin Yanxia, Zhang Min and Zheng Dequan, “A survey on neural network-based methods for event extraction”, Intelligent Computer and Applications, 2018, 8(3): 1-10.
[2] Zhong Z, Jin L, Feng Z. “Multi-font printed Chinese character recognition using multi-pooling convolutional neural network”, International Conference on Document Analysis & Recognition. 2015.
[3] Wan Jing, Li Haoming, Yan Huanchun, et al. “Relation extraction based on recurrent convolutional neural network”, Application Research of Computers, 2018, 1-6[2019-09-27].
[4] Jiang Jifa, “Research on free text information extraction model acquisition”, Graduate school of Chinese Academy of Sciences (institute of computing technology), 2004.
[5] Zheng Jiaheng, Wang Xinyi, Li Fei, “Research on Automatic Generation of Extraction Patterns”, Journal of Chinese Information Processing, 2004, 18(1):48-54.
[6] Chieu H L, Ng H T, “A maximum entropy approach to information extraction from semi-structured and Free Text”. Proceedings of the 18th National Conference on Artificial Intelligence, USA:American Association for Artificial Intelligence, 2002:786-791.
[7] Xu Honglei, Chen Jinxiu, Zhou Changle, et al, “Research on Event Type Identification for Chinese Event Extraction”, Mind and Computation, 2010, 4(1): 34-44.
[8] Zhao Yanyan, “Research on Chinese event extraction technology”, Harbin Institute of Technology, 2007.
[9] WEN Chang, LIU Yu, GU Jinguang, “Evolution relationship extraction of emergency based on attention-based bidirectional long short-term memory network model”, Journal of Computer Applications, 2019, 39(06): 1646-1651. (in Chinese)
[10] He Xinyu, Li Lishuang, “Trigger Detection Based on Bidirectional LSTM and Two-stage Method”, Journal of Chinese Information Processing, 2017, 31(6): 147-154.
[11] WANG Hong-bin, GAO Hong-kui, “Fusion Trigger Word Extension, Neural Network and Dependency Analysis Event Recognition”, Software Guide, 2018, 17(1): 19-21.
[12] Hongyu Lin, Yaojie Lu, Xianpei Han, et al, “Nugget Proposal Networks for Chinese Event Detection”, Institute of Software Chinese Academy of Sciences, 2018.
[13] Tu Min, Liu Xiang, Liu Shuchun, “Natural Language Processing Core Technology and Algorithm with Python”, China Machine Press, 2018.
[14] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed Representations of Words and Phrases and their Compositionality[C]//The Neural Information Processing Systems (NIPS) Foundation. Proceedings of NIPS, December 5-10, 2013, Harrahs and Harveys, Lake Tahoe. USA:NIPS, 2013:3111-3119.
[15] Mikolov T, Chen K, Corrado G, et al, “Efficient Estimation of Word Representations in Vector Space”, Computer Science, 2013.
[16] XUAN Xiaoxing, LIAO Tao, GAO Beibei, “Automatic Extraction of Chinese Event Trigger Word”, Computer&Digital Engineering, 2015, (3): 457-461.
[17] MA Chenxi CHEN Xingshu, WANG Wenxian, WANG Haizhou, et al, “Chinese Event Detection Based on Recurrent Neural Network”, Netinfo Security, 2018, (5): 75-81.
[18] LI Hong, YU Long, TIAN Shengwei, Turgun Ibrahim, et al, “Uyghur Emergency Event Extracton Based on DCNNs-LSTM Model”, Journal of Chinese Information Processing, 2018, 32(6): 52-61.
[19] CHEN Zhenghao, FENG Ao, HE Jia, “Text sentiment classification based on 1D convolutional hybrid neural network”, Journal of Computer Applications, 2019, 39(07): 1936-1941.

[20] CHEN Bo, “Text classification method based on the cycle structured convolutional neural network”, Journal of Chongqing University of Posts and Telecommunications(Natural Science Edition), 2018, 30(5): 705-710.