Real-Time Extraction of News Events Based on BERT Model

Yuxin Jiao
School of Computer Science and Engineering
Xi'an Technological University
Xi 'an, China
E-mail: 2233980355@qq.com

Li Zhao
School of Computer Science and Engineering
Xi’an Technological University
Xi’an, China
E-mail: 332099732@qq.com;

***Abstract*— For the large number of news reports generated every day, in order to obtain effective information from these unstructured news text data more efficiently. In this paper, we study the real-time crawling of news data from news websites through crawling techniques and propose a BERT model-based approach to extract events from news long text. In this study, NetEase news website is selected as an example for real-time extraction to crawl the news data of this website. BERT model as a pre-trained model based on two-way encoded representation of transformer performs well on natural language understanding and natural language generation tasks. In this study, we will fine-tune the training based on BERT model on news corpus related dataset and perform sequence annotation through CRF layer to finally complete the event extraction task. In this paper, the DUEE dataset is chosen to train the model, and the experiments show that the overall performance of the BERT model is better than other network models. Finally, the model of this paper is further optimised, using the ALBERT and RoBERTa models improved on the basis of the BERT model, experiments were conducted, the results show that both models are improved compared to the BERT model, the ALBERT model algorithm performs the best, the model algorithm's F1 value is 1% higher than that of BERT. The results show that the performance is optimised.**

Keywords-Web News Events; BERT; Event Extraction

I. INTRODUCTION

As Internet technology has advanced, information resources have expanded and more unstructured text—such as news articles and brief videos—is now readily available. Information regarding a variety of events, including social, political, and commercial ones, is contained in this textual data. In the current international situation, thousands of domestic and international news are generated every day. People need to know the latest policies and situations through news. By integrating and analysing a large number of online news events, we can obtain the latest domestic and international information, including domestic and international economic, military and diplomatic situations, and provide people with reliable reference data.

As more and more people want to be able to quickly and accurately extract the most relevant information about the events they want to focus on, researchers have started to work on the development of systems that are able to quickly, automatically, and accurately identify structured knowledge about events from the huge amount of textual information available on the Internet, and the event extraction task has been born as a result. The goal of event extraction is to create a structured record of events from unstructured text that includes the who, what, where, when, why, and how of actual occurrences.

News events are numerous and complex and redundant, this paper focuses on the extraction of events from long news texts, and the extraction of unstructured news texts, specifically including the extraction of event types, event ontologies and ontological roles. A news event may contain more than one type and the corresponding role of argument, and the news event contains a variety of domains, and now many studies are single-domain event extraction. In addition, news is real-time, the information is updated very quickly, and many fields are closely related and interact with each other, so there is still a lack of research work on extracting news events from multiple fields.

DOI: 10.2478/ijanmc-2024-0023

This paper combines real-time crawling of news data with extraction of events, which can timely and effectively obtain multi-domain news events, and provides a basic prerequisite for subsequent downstream work such as graph building and correlation analysis tasks. Taking Netease News Network as the research object, this paper proposes a network algorithm based on the BERT model to extract events, firstly, the news data are crawled and processed, and the extraction process is determined. Secondly, the model's structure and sequence annotation are shown. The DUEE dataset is used to train the network model, which has been shown to be very effective through experimental comparison with conventional network approaches. Finally, the model is further optimised, and the pre-training models RoBERTa and ALBERT, which are enhanced based on the BERT model, are utilized for the trials, and the experimental findings also demonstrate that the model performance is improved.

II. Related Work

A difficult and sophisticated part of information extraction is event extraction. In essence, event extraction research started almost simultaneously with information extraction research, and as research continued to evolve, the subtask of event extraction (VDR) began to be explicitly mentioned in the ACE evaluation programme. Researchers studied related event extraction techniques after the term "event extraction" was established. These techniques evolved over three phases with qualitative leaps in extraction effects: from early template matching-based techniques to machine learning-based techniques, and ultimately to deep learning-based techniques. Currently, deep learning is being used by researchers to extract events from data, and they have discovered that deep learning produces better extraction outcomes for deep feature extraction.

Events are contained in an event description, which is usually a sentence or a cluster of sentences, and the elements that constitute an event include event trigger words, event elements, element roles, and event categories. Depending on how granular the events are, event extraction jobs can be categorized into sentence-level and document-level categories. Sentence-level event extraction task involves identifying and extracting events from individual sentences, the goal of which is to identify the event's trigger word or sentence and to extract the relevant paper elements and roles played by the paper. The paper by Yu et al. [1] A neural network model for sentence-level event extraction known as the "LSTM-based end-to-end biomedical event extraction framework" is put forth. It makes use of a distributed representation of words in conjunction with bi-directional long and short-term memory neural networks to extract contextual information from sentences. A document-level event extraction technique called Hang et al. [2] makes advantage of multi-level contextual embedding to extract events and their parameters while capturing intricate word associations. The model uses a hierarchical structure to capture the relationships between events and employs a multi-task learning approach to jointly predict event types, trigger words and parameter elements. The model produced state-of-the-art results on many event categories when tested on the ACE 2005 dataset. A graph-based method for document level event extraction that captures relationships and dependencies between events is proposed in the work "Document Level Event Extraction via Heterogeneous Graph Based Inter-Event Relationship Learning" by Xu et al. [3]. The model learns the links between events using a graph neural network and represents events and their corresponding parameters using a heterogeneous graph structure.

In subsequent research work, event extraction based on deep learning has become an important research area with significant progress and improvement in performance. It utilises deep neural networks to automatically identify and extract event information from text. Pre-trained language models are one of the major advancements in deep learning based event extraction techniques. It has been demonstrated that pre-trained language models enhance the performance of event extraction models by offering a pre-trained comprehension of the language and the capacity to produce representations that effectively capture the syntactic and semantic aspects of the text. In 2018, the BERT pre-trained language model proposed by

Google's Devlin [4] et al. was trained on a large-scale corpus by employing Transformer encoding and multi-head attention mechanism, which resulted in pre-trained word vectors with stronger representational power and made the application of pre-trained models in the field of NLP gained much attention. The BERT model has a very good representational power and can well solve the problem of multiple meanings of a word, so it is often used to generate the initial embedding matrix of a text. The BERT-Bi LSTM-CRF model was employed by Li Ni et al. [5] to enhance performance on the clinical named entity recognition datasets CCKS-2017 and CCKS-2018. The Chinese named entity recognition method of Li Ni et al. [6] allowed the network model to acquire 94.41% F1 value on the MSRA dataset, based on the BERT-IDCNN-CRF network structure. Jian Yuan et al. [7] used BERT and CNN models to extract character and glyph feature vectors from the text and fused word vectors to extract text characteristics from various dimensions. This approach outperformed other models in testing on the People's Daily dataset.

In summary, developments in deep learning-based event extraction are characterised by combining pre-trained language model attention mechanisms with multi-task learning techniques. These advancements have significantly increased the efficiency and accuracy of event extraction models, increasing their usefulness in a variety of information extraction and natural language processing applications.

III. Extraction Of Real-Time Event Information From Netease News

Real-time news extraction using the Bert model for the NetEase News Network requires the following four processes: data acquisition, data processing, model training, and news event extraction. As shown in the Fig.1, data acquisition uses crawler technology to crawl all kinds of news-related content from the NetEase News Network, pre-process the crawled content, i.e., check whether there is any error content, then train the Bert model, input the crawled news data into the Bert model for event extraction, get the events, and store them.

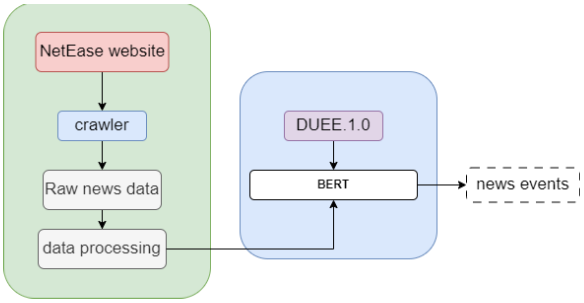

Figure 1. Event Extraction Process Map

A. Data Acquisition And Processing

This paper selects the real-time news website as Netease News Network. For the website, using crawler technology to obtain the website in 11 categories of news content, choose the Scrapy crawler framework, send requests to the website through regular expressions, and use beautiful After the extraction of news, the crawled data will be saved to a CSV file. Initially, we collected about 20,000 news stories. Due to the real-time capture, there may be news duplication or data is empty, the initial data first cleaning operation, that is, to remove the duplication and no content of the news data.

B. Model Introduction

The BERT model is a pre-trained model with a bi-directional Transformer's encoder as a feature extractor, whose internal structure consists of multiple Trans-former layers stacked on top of each other, and has been able to achieve good results in most of the NLP tasks by training on ultra-large scale datasets. The bidirectional coding ability of the BERT model is applied to obtain the correlation relationship between words and the contextual semantic information between sentences to achieve bidirectional feature extraction of news text data. At the same time, reasonable constraints on the tag sequence prediction results are achieved by combining the Conditional Random Field CRF.

In this study, news text statements are used as inputs to the BERT model, and the corpus is divided into sentences by space line breaks, and the sentence inputs are notated as Text={c1,c2,...,cn}(n≤max_len), where cn represents the individual words in a sentence, n is the number of words contained in a sentence, and max_len is the sentence maximum length. Every sentence undergoes preprocessing; if its length surpasses max_len, it is shortened; if its length falls below max_len, PAD tags are added to supplement the sentence length; CLS tags are added at the start of the sentence, and SEP tags are added at the conclusion to divide it from the following sentence. To obtain the outcome, the preprocessed input phrase sequences are computed using the BERT model's word embedding layer. As shown in equation (1).

$$E_{word} = E_{tok} + E_{seg} + E_{pos} \qquad (1)$$

E_{tok} is the symbol embedding; E_{seg} is the fragment embedding; and E_{pos} is the positional embedding. Summing the three based on the textual elements yields the final word vector representation of the phrase input.

The pre-training task for the BERT model consists of a Masked LM task at the text level and a Next Sentence Prediction task at the sentence level. The Masked LM task is set to mask 15% of the text of the input sequence, where 80% of the text is replaced by MASK symbols, 10% is replaced by other textual symbols, and 10% of the text is not replaced. The model predicts the masked original text based on the context of the unmasked text in the sequence, and thus learns the correlation relationship between the texts. During the execution of the Next Sentence Prediction task, the set of sentence sequences as input does not completely retain the order of the utterances in the original corpus, but instead, it consists of a randomly selected 50% of the sentences together with 50% of the sentences retaining the original order, and by learning the contextual relationship between sentences and obtaining the contextual information of the utterances. The BERT model is able to achieve both text-level and sentence-level rich feature extraction in the news event extraction task studied in this paper, and then undergoes the multi-head attention mechanism and fine-tuning training to obtain the prediction vectors of the corresponding tag sequences for each text position. The pre-training process and the fine-tuning process are shown in Fig. 2.

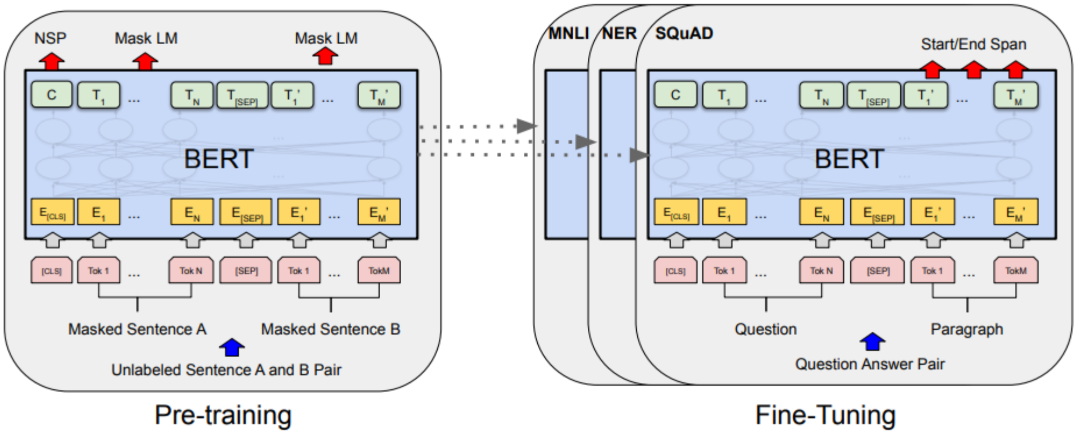

Figure 2. Pre-training and Fine-Tune process

Since the tag sequence prediction results output from the BERT model do not take into account the transformation rules between tags, the computed results are optimised using Conditional Random Field CRF after the BERT model. In the domains of segmentation, entity identification, and lexical annotation in natural language processing, CRF is a well-known sequence annotation technique. And in these annotation scenarios, the effect is significantly improved. Transfer property and state property are the two main categories of properties in CRF. The relationship between the current state and the input sequence is known as the state characteristic, and the transfer characteristic reflects the connection between the final output and the present output state. as seen in Fig.3.

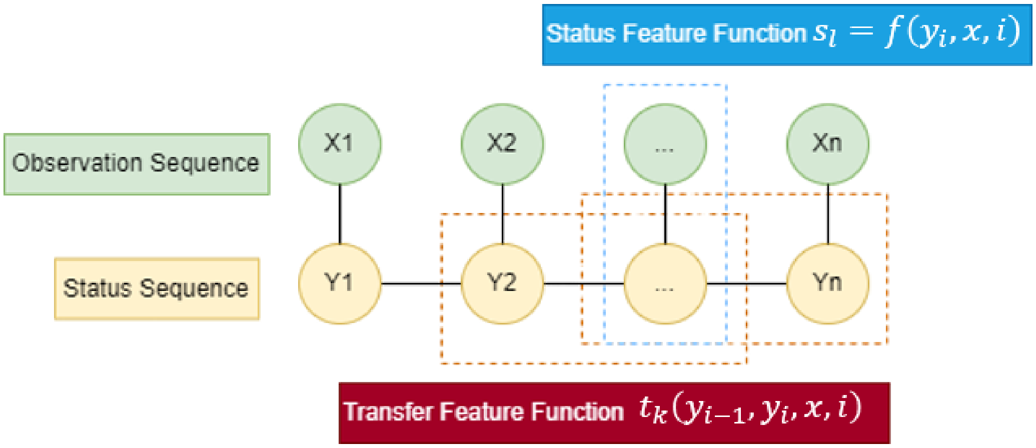
Status Feature Function $s_l = f(y_i, x, i)$
Observation Sequence
X1
X2
...
Xn
Status Sequence
Y1
Y2
...
Yn
Transfer Feature Function $t_k(y_{i-1}, y_i, x, i)$

Figure 3. Graph structure of CRFs for linear chain conditional random fields

Conditional Random Field (CRF) can fully consider the dependencies and constraints between neighbouring characters. Therefore, we adopt Conditional Random Field (CRF) in the last layer of the model to constrain the feature information output from the multi-attention layer to ensure the accuracy of the relationship between the labels obtained in the end. When given an input sequence X={x1,x2,x3,...,xn}, assume that the output sequence is y={y1,y2,y3,...,yn}. Then the score of the output sequence can be expressed by the following equation:

$$s(X,y)=\sum_{i=1}^{n}\left(W_{y_i,y_{i+1}}+P_{i+1,y_{i+1}}\right) \qquad (2)$$

W is the transfer matrix, $W_{y_i,y_{i+1}}$ is the number of scores of labels transferred from y_i to y_{i+1}, and $P_{i+1,y_{i+1}}$ is the number of scores of labels y_{i+1} corresponding to the i+1st word of the input sequence. The probability of the output sequence y is calculated and the sequence of labels when the conditional probability is maximum will be output as the result sequence. Where Y_X denotes the entire tag sequence of the input sequence. The formula can be expressed as:

$$P(y|X)=\frac{\exp(s(X,y))}{\sum \tilde{y}\in Y_X \exp(s(X,\tilde{y}))} \qquad (3)$$

C. Extracting events based on the Bert model

Two stages make up Bert's overall framework: pre-training and fine-tuning. This research focuses on the fine-tuning step, where the Bert model is first parameterized by the pre-training model and then all parameters are learned on labeled data [8]. The model in the pre-training stage is trained on unlabelled data. The CRF layer labels the sequences after the news text input is delivered into the Bert model for processing and training.

The specific process is to first read data from the DUEE dataset, build a thesaurus, feed the data into the Bert model for processing and training, and annotate the sequences with a CRF layer. The CRF layer makes each point annotated as a whole rather than individually, and the annotations of each point have a certain degree of relevance. In this way, the model understands the text in addition to the rules in the output sequence. Eventually event extraction is complete, predictions are presented, and the model is evaluated and saved. Data is processed using the trained Bert model; the captured news data is fed into the Bert model and processed to obtain the extracted event types. The structure of the extracted events is shown in Fig. 4; the news data is processed to extract the event type, thesis role, specific attribute value and specific category.

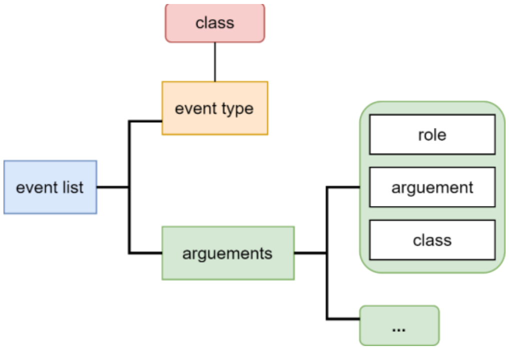

Figure 4. Extracted event output structure, including event types and argument roles

IV. EXPERIMENTAL ANALYSIS

In this paper, a web crawler based on the Scrapy framework is used to obtain news data from 11 categories in the NetEase news network, and initially, two news items are crawled as the basic processing data. The DUEE.1.0 dataset is selected to fine-tune the Bert model for training. And the crawled news data was input into the Bert model for extraction to realize the task of extracting real-time network news events

A. Event Information Extraction Experiment

In this paper, we first used a crawler program based on a scrapy framework applied to NetEaseNews.com to get 20000 data and store it in a CSV file.

The Bert model is trained using the DUEE.1.0 dataset, which consists of 17,000 words with event messages and 65 distinct events. This dataset has 20% designated as the validation set and 80% designated as the training set. The Keras framework and Tensorflow are used in the experimental environment. Using Adam as the optimizer, the model's learning rate parameter is set to e-5. The BERT layer uses the trained Chinese BERT model from Google, which uses a 12-layer Transformer encoder. The hidden layer has a dimensionality of 768 and a multi-head attention mechanism with 12 heads. During the training phase, the batch_size is set to 32 and the epochs are set to 10.

In this experiment, the precision, recall and F1 value of the machine learning method are used as the model measures for this experiment. The numbers TP, FP, and FN represent the number of positive samples and positive predictions, negative samples and positive predictions, and positive samples and negative predictions, respectively.The following is the formula:

$$\Pr ecision = \frac{TP}{TP + FP} \tag{4}$$

$$\text{Re}\,call = \frac{TP}{TP + FN} \tag{5}$$

$$F1 = \frac{\text{2xPrecisionxRecall}}{\Pr ecision + \text{Re}\,call} \tag{6}$$

In this study, in addition to using the BERT model, the traditional neural network models LSTM and BiLSTM were used for experimental comparison to prove the effectiveness of BERT. The line graph Fig.5 shows that the P, R, and F1 values of the BERT model are significantly higher than those of the other two models, and for the BiLSTM model, the overall level is lower because the effect of context and semantic environment on the classification of words and word labels is not taken into account, and the predicate labels of each element are independent of the other elements and not dependent on them.

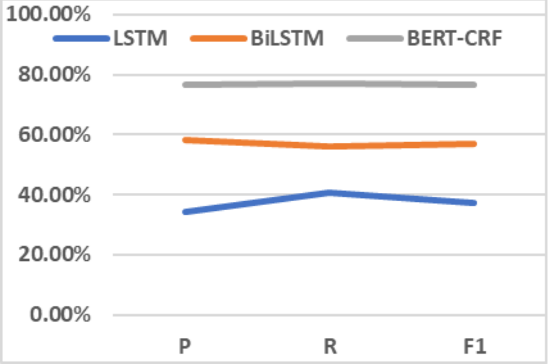

Figure 5. Comparison of P-value, R-value and F1-value of LSTM,BiLSTM and BERT-CRF models. P for Precision, R for Recall

In order to further optimize the model, the BERT pre-training model is replaced with ALBERT and RoBERTa for experiments, respectively. The RoBERTa model is an improved version of the Bert model, with more unlabeled data, longer training time, and larger batch sizes, which enhances the model's learning ability and generalization capability. Meanwhile, improvements are made in the training method by removing the next sentence prediction task to support longer word sequences; using dynamic masks to avoid repeated training of data; and using byte-level vocabulary to train the model to support processing of many common words. The ALBERT model is a lightweight model based on the BERT model, with a substantially lower number of parameters compared to the traditional BERT, and with a relative increase in the operation speed. By reducing the number of parameters and enhancing the resilience of the neural network parameters, ALBERT's technique of embedded layer factorization and cross-layer parameter sharing speeds up model training while compressing the overall number of parameters. The SOP (Sentence-Order Prediction) task, which focuses on inter-sentence order prediction independently of subject aspect, takes the role of the NSP task in ALBERT. When compared to NSP, SOP can achieve an approximate 2% increase in accuracy for the downstream job that requires numerous sentence inputs. The ALBERT model could increase semantic comprehension, speed up training, and have fewer parameters.

The BERT pre-trained model is replaced with ALBERT and RoBERTa model respectively on both the DUEE dataset taken for this experiment respectively and the data are shown in Fig. 6.

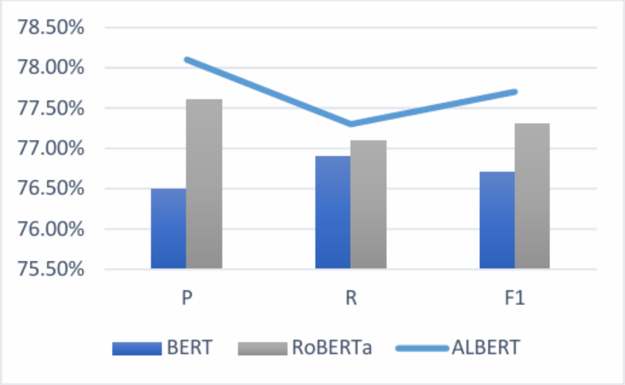

Figure 6. Comparison of P-value, R-value and F1-value of BERT, RoBERTa and ALBERT models Comparison of P, R and F1 values. P for Precision, R for Recall

B. Conclusion Of The Experiment

The use of a bi-directional architecture, which enables the model to better comprehend the context and meaning of the text, is one of the BERT model's significant advances. The core component of the BERT paradigm is the transformer encoder. It is composed of several layers, one feed-forward neural network and one multi-headed self-attention mechanism. While the feed-forward neural network analyzes the weighted input to build a contextual representation of the tokens, the self-attention mechanism enables the model to assess each token's relevance in the input text based on its relationship to other tokens in the sentence. Therefore, BERT is effective for event extraction. The specific experimental results are shown in Table I.

TABLE I. Experimental Results I

Module	P	R	F1
LSTM	34.2%	40.6%	37.1%
BiLSTM	58.1%	56.2%	57.1%
BERT-CRF	76.5%	76.9%	76.7%

Both ALBERT and RoBERTa models are based on BERT with different improvements, and the experimental results also show that both ALBERT and RoBERTa are better than the BERT model, and the specific experimental data are shown in Table II.

TABLE II. Experimental Results II

Module	P	R	F1
BERT	76.50%	76.90%	76.70%
RoBERTa	77.60%	77.10%	77.30%
ALBERT	78.10%	77.30%	77.70%

Input the organized news data into the three trained models for processing, and select the news body for event extraction. The news body is obtained from the organized news csv file for extraction, and the output results are then stored in the csv file to complete the extraction of news events. From the actual extraction results, there is not much difference between the three models, and ALBERT is slightly better than the other two models.

V. CONCLUSIONS

To address the issue of slow real-time news on websites, this research suggests an event extraction technique for real-time news. Crawler technology is utilized to crawl the news data, and after literature research for event extraction model determination, the BERT model is finally selected, and the public news dataset DUEE is utilized for training to extract event attributes and types, etc. When comparing BERT to the more established neural network models LSTM and BiLSTM, there is a clearer advantage. In this paper, the BERT model is also replaced with RoBERTa and ALBERT model, the model is further optimized, and the experiments show that the ALBERT model improves the F1 value of the original BERT model by 1%. The trained model is finally applied to the real-time crawled NetEase news network data to realize the real-time news event extraction.

The research on real-time news event extraction in this paper extracts complex unstructured news data and transforms it into structured content, which has important research significance and use value for its downstream tasks such as knowledge graph creation, event association analysis, etc., and has research and exploration significance.

REFERENCES

[1] Yu X, Rong W, Liu J, et al., Lstm-based end-to-end framework for biomedical event extraction [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2019, 17(6): 2029–2039.

[2] Yang H Chen Y., Liu K., et al. Multi-Turn and Multi-Granularity Reader for Document-Level Event Extraction [J]. ACM Transactions on Asian and Low-Resource Language Information Processing, 2022, 22(2):1–16.

[3] Xu R, Liu T, Li L, et al. [Document-level event extraction via heterogeneous graph-based interaction model with a tracker [J]. arXiv preprint arXiv:2105. 14924, 2021.

[4] Devlin J, Chang M W, Lee K, et al. Bert:Pre-training of deep bidirectional transformers for language understanding [J]. ArXiv Preprint ArXiv:1810.04805, 2018.

[5] LI Xiangyang, ZHANG Huan, ZHOU Xiaohua. Chinese clinical named entity recognition with variant neural structures based on BERT methods [J]. Journal of Biomedical Informatics, 2020(107):103422.

[6] LI Ni, GUAN Huanmei, YANG Piao, et al.BERT-IDCNN-CRF for named entity recognition in Chinese [J]. Journal of Shandong University (Natural Science), 2020, 55(01):102-109.

[7] YUAN Jian, ZHANG Haibo. Chinese Entity Recognition Model of Multi-granularity Fusion Embedded [J]. Journal of Chinese Computer Systems, 2022, 43(4):741-746.

[8] YANG Zhenyu, ZHANG Denghui. A complex long sentence intent classification method combining BERT and two-layer LSTM [J]. Computer Applications and Software, 2021, 38(12):207-212.