A Chinese Event Relation Extraction Model Based on BERT

Can Tian
Big Data Analysis Technology Lab
University of Chinese Academy of Sciences
Beijing, China
e-mail: tiancan@hust.edu.cn

Yawei Zhao*
Big Data Analysis Technology Lab
University of Chinese Academy of Sciences
Beijing, China
e-mail: zhaoyw@ucas.ac.cn

Liang Ren
Big Data Analysis Technology Lab
Beijing Knowlegene Data Technology Co., Ltd.
Beijing, China
e-mail: renliang@knowlegene.com

***Abstract*—Relation extraction and event extraction are important subtasks of information extraction. To identify the relations and events in Chinese text accurately can help to improve the performance of tasks such as graph construction and risk conduction. Different from the traditional methods, this paper proposes a joint model to extract entities and events in the text, and gives the concept of event relation, to discover the potential relations between the arguments of events and the relations between two or more events. We conduct experiments on a financial dataset, the results show that the new model is 4%-6% higher than the existing event extraction model in F1 score, and the proposed event relation is also meaningful and practical.**

Keywords-relation extraction; event extraction; event relation

I. INTRODUCTION

Relations and events are the blocks for building the knowledge graph. To identify relations and events in texts accurately can improve the performance of tasks such as graph completion and risk conduction. To recognize entities in the text is the pre-order task of relation and event extraction, but most of existing models to solve them are pipeline manner, to recognize the entities first, then analyze the relations between the entities or whether there are events existing in the sentence. It makes the tasks simple to deal with in this way, but ignores the connection among the three tasks, the performance of entity recognition may affect the results of relation or event extraction and lead to erroneous delivery.

It can be found that the occurrence of events is generally accompanied by new relations, such as the cooperation relation arising from the "new partner" event, investment relation arising from the "financial event", etc., and such relations can't be found in traditional ways. At the same time, an event may be caused by other events, that is, multiple events are related, such as the "company executives resign" event caused the "Stocks fell" event and the "enterprise loss" led to the "company layoffs" event. This article refers to the two new relations above as event relation.

In view of this, this paper first proposes a new model to extract events in the text, and then conduct event relation analysis on the extracted events, so that not only the connections between the arguments but also the relations between different events can be found. On this basis, an event relation graph can be constructed. It can be used as a supplement to relation graph or event graph of a specific field, or it can provide more diverse information for other tasks such as risk conduction and graph completion.

The major contributions of this paper are: (1)A novel tagging scheme is proposed to extract events. (2) The latest BERT model is used to settle the event extraction problem. (3) An improved loss function is proposed to describe the importance of labels. (4) Propose the concept of event relation to get more diverse relation types, and establish event relation graph.

II. RELATED WORKS

The premise of relation and event extraction needs to get the entities in the text. The most of existing methods to settle entity recognition research are based on the LSTM+CRF framework [1] and added different new features in different languages or fields, such as Chinese radicals [2], English character vectors [3], the model fusion [4] or new attention mechanism [5], as well as the latest BERT model [6] also is helpful to entity recognition.

The most used to solve the relation extraction task is the CNN model. By adding the word position feature [7], the position information of the entity in the text is considered, and then through the convolution layer and the pooling layer respectively to obtain the sentence representation for relation extraction. However, the same kernel size is used that the extract features are relatively simple. Therefore, Nguyen T H [8] added various kernel sizes and got better results. By introducing a new loss function which was named Ranking loss [9], the loss function can effectively improve the

978-1-7281-0831-5/19/$31.00 ©2019 IEEE

discriminability between different relation categories, and has a great improvement on the model. Some joint models [10] is also proposed to extract entities and relations in the text to avoid erroneous delivery.

Event extraction is separated into two parts in general, to recognize event trigger first, then to extract event arguments, traditional researches to settle them in a pipelined manner, the representative job is the Dynamic Multi-Pooling Convolutional Neural Networks [11]. There also have some researcher to combine the two parts together, Q Li [12] proposed a joint model based on structured prediction which extracts trigger and arguments. Some works are only need to identify whether an event in a sentence which named event detection [13]. There are few jobs about Chinese event extraction compared with English, but small researches [14].

This paper is mainly to use a joint model to extract the entities and events in the text and to discover the potential relations between the event arguments and the relations between two or more events.

III. THE JOINT ETRACTION MODEL

In this section, we first give the new tagging scheme which covert the event extraction task into a sequence labeling problem. Then we describe how to use the end-to-end model to deal with the labeling problem and discuss the improved loss function that we proposed in this paper.

A. The Tagging Scheme

For a given text, each word token is assigned a label that contributes to extract the results, we use "BIO" annotation to express the beginning, internal, and the other labels of the entities here. To recognize entities is mainly for the purpose of extracting event arguments better here. Therefore, for a text, in addition to labeling the event trigger and event arguments, it is necessary to mark the entities with the same entity type as the argument of the event. At the same time, to adding a pair of event type identifiers at the beginning and the tail of text, it is made to mix the corpus of multiple events to training become possible, the two identifiers will remind the model which event the text belongs to, and use the standard of the event to label it. The specific tagging scheme is shown in Fig 1.

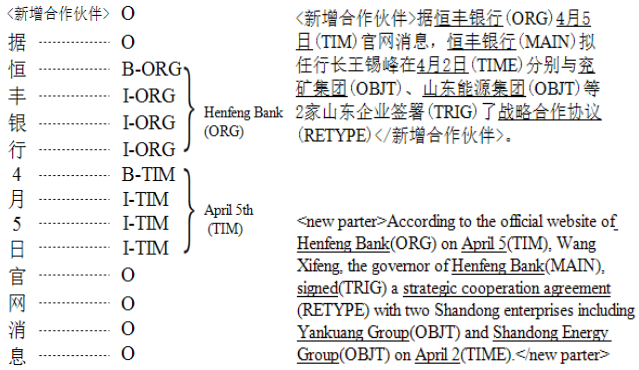

Figure 1. An example of tagging scheme.

The "new partner" indicates the event type, and "ORG" and "TIM" express irrelevant organization and time that are not related to the event. "TRIG" indicates the event trigger, "MAIN", "OBJT", "TIME", "RETYPE" represent the arguments of the event (subject, object organization, time of occurrence and the type of cooperation). It can be found that the token marked "O", "Wang Xifeng" and "Shandong" are entities of person and place respectively, it's unnecessary to label them because the entity type of the event arguments do not include these entities.

B. End-to-end Model

The end-to-end model based on neural network is widely used to solve the sequence labeling task. The LSTM+CRF framework is the most typical, and it has achieved better results in different sequence processing tasks. On this basis, the word2vec, attention mechanism and the advanced transformer are constantly improving the effects of different tasks. Google released the BERT model in 2018, which is considered to be a milestone in the NLP field. It is similar to word2vec, which uses a large number of unlabeled corpus training models to characterize the language grammar and provide the representation as a feature to downstream tasks. Different from the existing method, BERT has improved in the model framework, pre-trained loss function, and the training method, and achieves better results on many NLP tasks. Based on this, this paper uses the joint model of BERT and LSTM+CRF to extract the events in the text, as shown in Fig 2.

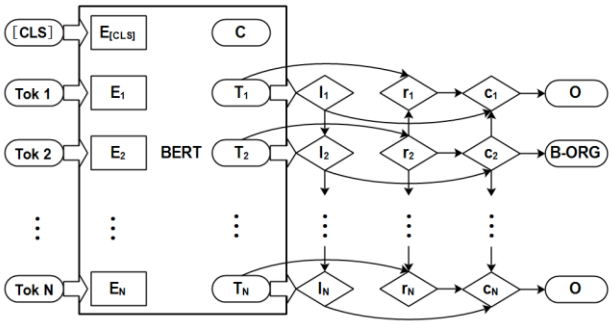

Figure 2. BERT+LSTM+CRF model.

C. Improved loss Function

It can be seen from the tagging scheme that the entity type of the irrelevant organization, the subject organization and the object organization which are marked as "ORG", "MAIN" and "OBJT" respectively are the same, but we regard them as different tags, the same to time, amount et al. In order to express the relations between them, we generate two sets of tag sequences during the data processing, the first sequence y_1 is the entity type of each token, the second y_2 is the tag of event role, and the loss is calculated in model by the two tag sequences, then optimize the two losses, and for each of the label i, the importance of it is different, in some events, we mainly focus on organization, regardless of the time or the amount involved, so it is possible to introduce different weights into each of the tags to calculate the loss. Combining the above strategies, the loss function is defined as follows.

$$L = \max \sum_{j=1}^{|D|} \sum_{t=1}^{L_j} (\beta_1 (\log(p_{1t}^{(j)} = y_{1t}^{(j)} \mid x_j, \Theta)) + \beta_2 \alpha_i \cdot \log(p_{2t}^{(j)} = y_{2t}^{(j)} \mid x_j, \Theta))$$

The $|D|$ is the size of training set, L_j is the length of the sentence x_j, $p_{1t}^{(j)}$, $y_{1t}^{(j)}$ refers to the prediction label and the event label of token t in sentence x_j respectively, and β_1, β_2 represent the weight of the losses which are calculated by y_1 and y_2, α_i indicating the weight of each tag, the value can be determined according to the importance of different event arguments.

IV. The Representation of Event Relation

There are two kinds of event relations mentioned in this paper, event relation within the arguments of an event and event relation between two or more events. It will be analyzed separately from the extracted events.

A. Event Relation within Event

The event relation within event is the connection of the arguments of an event, it can be expressed as a triple $(Role_i^{\Phi}, R_{\Phi}, Role_j^{\Phi})$, the same to traditional relations, $Role_i^{\Phi}$, $Role_j^{\Phi}$ are the i and j argument of event Φ, R_{Φ} represent the relation of argument i and j is produced following the event Φ. As shown in Fig 3, the gray part is the event type, and the arguments of event are connected with it. The thickness of the arrow indicates the importance of the event argument, such as the financing event of "XtalPi", "Sequoia" is the investment, "Google" and "Baidu" are the follow-up investments. In some cases, the amount of investment can also be expressed. The dotted line indicates the new relations that are accompanied by the events. The traditional relation definition or relation extraction methods cannot obtain such relations, so the relations can be used as the supplement to the traditional relations to provide more diverse relation information.

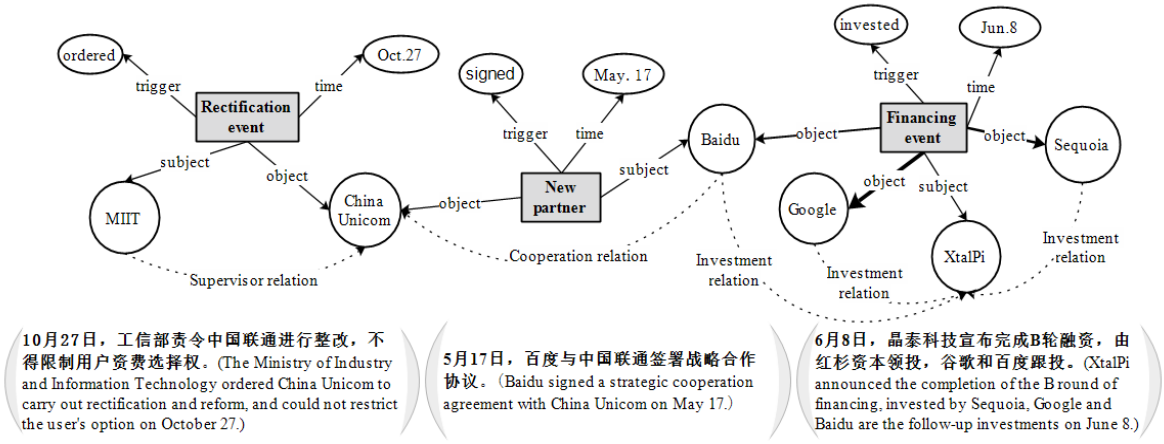

Figure 3. An example of event relation within event.

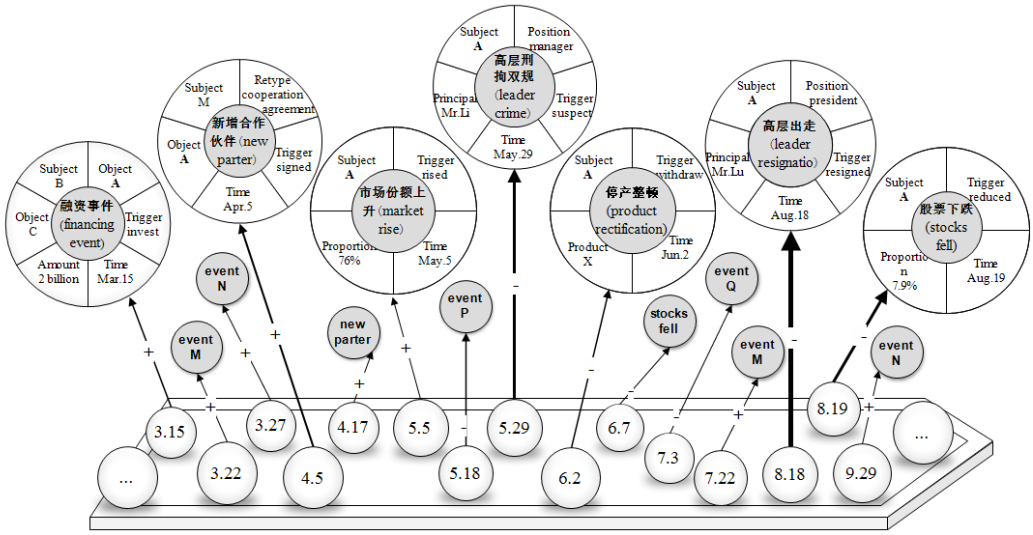

Figure 4. An example of event relation between events.

B. Event Relation between Events

The event relation between two or more events can express with $(\{\omega_1\Phi_1^+, \omega_2\Phi_2^-, ..., \omega_k\Phi_k^+\}, \Phi_t)$, Indicates that event Φ_t is affected by events $\Phi_1, \Phi_2, ..., \Phi_k$, $\omega = (\omega_1, \omega_2, ..., \omega_k)$ represent the weight parameters of events $\Phi_1, \Phi_2, ..., \Phi_k$, and "+", "-" indicate the positive and negative impact of the event. As shown in Fig 4, they are the related events that company A expands on the timeline. The gray part is the event type, and the event arguments are connected with it. The "+" and "-" on the arrow indicate the influence tendency of company A corresponds to positive and negative influence respectively. The thickness of the arrow corresponds to the degree of attention of the media to the event, which expresses as the number of news were reported about the event during the period. It can also be formalized as the extent to which the event affects the company, which may induce the creation of new events. Such connections have positive implications for the transmission of risks.

V. EXPERIMENTS

In this section, we first describe our experimental setting, including the dataset and the evaluation that we chose, the contrast models and the mainly hyperparameters of different models. Then we report the experimental results and analyze the error of different methods.

A. Experimental Setting

- Dataset: Most of the rearches to extract events are based on ACE2005 dataset. There are little jobs on Chinese corpus while the ACE2005 dataset only includes 633 Chinese documents with less than 2000 events, for lacking of training data, and the event type is sparse. We product experiments on the financial datasets to evaluate the performance of our methods, including 27 types of events, and a total of 32,480 sentences.
- Evaluation: We chose Precision(P), Recall(R) and F1 score(F1) to evaluate the experimental results.
- Baselines: We reproduced four benchmark methods, structured perceptron, DMCNN, word2vec+LSTM+CRF and BERT+softmax based on the dataset to compare with the proposed model.
- Hyperparameters: Five models are mentioned here, we set the beam search size equal to 4 in structured perceptron method, the dimension of word embeddings, event features, position features are 100, 5, 5 and the filter size is 3, the feature map size is 100 in DMCNN model, the dimension of word embeddings is 100 and the dimension of LSTM is 100 under the word2vec+LSTM+CRF framework, we choose the max sequence length equal to 128 in BERT+softmax, and $\beta_1 = 1$, $\beta_2 = 1$, $\alpha = (5, 2, 2, 1)$ where the parameter corresponding to the importance of event trigger and subject organization, other event arguments, other entities, "O" respectively in BERT+LSTM+CRF+L.

B. Experimental Results

We report the experiment results of different models as shown in TABLE I, it can be seen that the proposed end-to-end models in this paper outperform the contrast models in F1 score, it means that our new tagging scheme is good for the event extraction problem and it is always helpful no matter which end-to-end model we choose. We also can conclude that the improved loss function is good for extracting the event arguments from the last two experiments even though the performance of entity recognition will be reduced. We will analyze the results in details in the follow section.

TABLE I. EXPERIMENTAL RESULTS OF DIFFERENT MODELS

Methods	Entity			Trigger			Argument		
	P	R	F	P	R	F	P	R	F
Structured perceptron	NA	NA	NA	75.30	80.10	77.60	54.30	61.54	57.69
DMCNN	NA	NA	NA	85.53	87.39	86.43	63.55	64.19	63.87
Word2vec+LSTM+CRF	82.59	91.07	86.62	82.31	85.21	83.74	66.46	69.51	67.95
BERT+softmax	87.27	91.43	89.30	84.02	89.87	86.85	70.74	77.09	72.96
BERT+LSTM+CRF	**89.46**	**92.88**	**91.14**	**86.93**	91.24	89.03	**73.59**	77.58	75.53
BERT+LSTM+CRF+L	88.71	91.67	90.16	85.99	**93.51**	**89.59**	72.37	**82.12**	**76.94**

C. Error Analysis

It can be seen from the first two experiments that even though the structured perceptron is a joint model and DMCNN is the pipeline, while the results of CNN is better than the former, so it can be concluded that the deep neural network is the trend to solve information extraction tasks.

Compared with the DMCNN model, the third experiment transformed the event extraction into a sequence labeling problem, the LSTM+CRF framework was used to train the

model. Not only the entity and event but also combined the event trigger and argument to solve the error transmission problem. Get better results without the entity information, and increase the 4% in F1 value.

The last three models are based on the same data labeling scheme, using different end-to-end models, it can be found that the BERT model greatly improves the experimental results.

As can be seen from the last two experiments, the proposed improved loss function improves the event extraction results. In order to evaluate the influence of different parameters, we chose the parameter β in [0, 0.5, 1, 1.5, 2, 2.5] and chose several sets of α, we report the experimental results including trigger and event arguments as shown in Fig 5 and Fig 6.

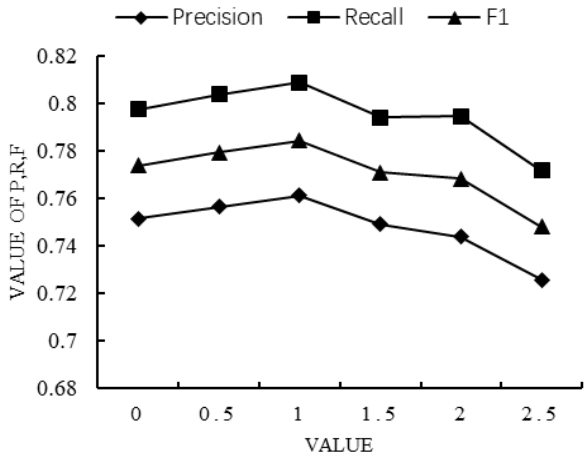

Figure 5. The results on different parameter β.

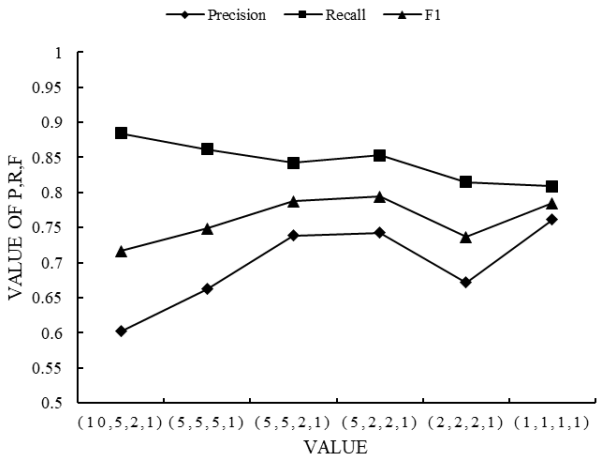

Figure 6. The results on different parameter α.

It can be seen from Fig 5 that the result of β=1 achieves a 1% improvement in F1 score comparing with β=0, and it is inversely proportional to β when the parameter become bigger. On the other hand, in order to describe the significance of event trigger and subject organization in an event, we set a larger weight for them, we can conclude from Fig 6 that when α is too large, it will improve the rate of recall but affect the accuracy of prediction. It can achieve a 1% improvement in F1 score when α = (5, 2, 2, 1).

VI. CONCLUSIONS

Entity recognition is the pre-order task of relation and event extraction, the results of entity recognition may affect the performance of relation or event extraction and lead to erroneous delivery, the same to recognize event trigger and extract event argument respectively. In this paper, a novel tagging scheme is proposed to covert the joint extraction task to a tagging problem, then we use the end-to-end model to extract entities and events together and we also propose a new loss function with different parameters to express the importance of different labels to extract events in Chinese text. We also proposed the concept of event relation to analyze the relation within the arguments of an event and between two or more events, it is important to reduce or control risks of a company and can be widely used to the tasks such as risk conduction and graph completion.

The experimental results show that the end-to-end models are better than other existing pipelined models, the new parameters are also improved the results and the proposed event relation is also meaningful and practical. However, sometimes this model cannot extract the event arguments completely, there are some problems to handle multiple different events in a sentence and how to use the event relation in other tasks in detail, which can be improved in the future works.

ACKNOWLEDGMENT

Give my heartfelt thanks to Beijing Knowlegene Data Technology Co., Ltd of providing the dataset details. This work is also supported by the 13th Five-Year Plan for Information Science of the Chinese Academy of Sciences (No. XXH13504-05) and the National Natural Science Foundation of China (No. 61872331).

Corresponding author: Yawei Zhao.

REFERENCES

[1] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF Models for Sequence Tagging[J]. Computer Science, 2015.

[2] Dong C, Zhang J, Zong C, Hattori M, Di H. Character-Based LSTM-CRF with Radical-Level Features for Chinese Named Entity Recognition[C]// International Conference on Computer Processing of Oriental Languages. Springer In-ternational Publishing, 2016:239-250.

[3] Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural Architectures for Named Entity Recognition[J]. 2016:260-270.

[4] Peters ME, Ammar W, Bhagavatula C, Power R. Semi-supervised sequence tagging with bidirectional language models[J]. 2017.

[5] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Aidan N. Attention Is All You Need[J]. 2017.

[6] Devlin J , Chang M W , Lee K , Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018.

[7] Turian J, Ratinov L, Bengio Y. Word representations: a simple and general method for semi-supervised learning[C]// Acl, Meeting of the Association for Computational Linguistics, July, Uppsala, Sweden. 2010.

[8] Nguyen T H , Grishman R. Relation Extraction: Perspective from Convolutional Neural Networks.[C]// Workshop on Vector Space Modeling for Natural Language Processing. 2015.

[9] Santos C N D , Xiang B , Zhou B. Classifying Relations by Ranking with Convolutional Neural Networks[J]. Computer Science, 2015, 86(86):132-137.

[10] Wang L , Cao Z , Melo G D. Relation Classification via Multi-Level Attention CNNs[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016.

[11] Chen Y, Xu L, Liu K. Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks[C]// The, Meeting of the Association for Computational Linguistics. 2015.

[12] Q Li, H Ji, L Huang. Joint Event Extraction via Structured Prediction with Global Features[C]// The Meeting of the Association for Computational Linguistics. 2013.

[13] Liu S, Chen Y, Liu K. Exploiting Argument Information to Improve Event Detection via Supervised Attention Mechanisms[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Lin-guistics (Volume 1: Long Papers). 2017.

[14] Li P, Zhou G. Joint Argument Inference in Chinese Event Extraction with Argument Consistency and Event Relevance.[J]. IEEE/ACM Transactions on Audio Speech & Language Processing, 2016, 24(4):612-622.

[15] Xia X, Peifeng L, Xin Z, Qiaoming Z. Event inference for semi-supervised Chinese event extraction[J]. Journal of Shandong University, 2014.

[16] Ma X, Hovy E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF[J]. 2016.

[17] Chiu JPC, Nichols E. Named Entity Recognition with Bidirectional LSTM-CNNs[J]. Computer Science, 2015.

[18] Yang B, Mitchell T. Joint Extraction of Events and Entities within a Document Context[J]. 2016.

[19] Nguyen T H, Cho K, Grishman R. Joint Event Extraction via Recurrent Neural Networks[C]// Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016.

[20] Zhou D, Zhang X, He Y. Event extraction from Twitter using non-parametric Bayesian mixture model with word embeddings[C]// Conference of the European Chapter of the Association for Computational Linguistics. 2017.

[21] Rao S, Marcu D, Knight K, Daume H. Biomedical Event Extraction using Abstract Meaning Representation[C]// Bi-oNLP 2017. 2017.

[22] Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space[J]. Computer Sci-ence, 2013.

[23] Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed Representations of Words and Phrases and their Composition-ality[J]. Advances in Neural Information Processing Systems, 2013, 26:3111-3119.