3.3 Model Training

We train the joint model and define L_T as the loss function of all binary classifiers that are responsible for detecting triggers, shown as follows:

$$L_T = \frac{1}{m \times n}(\sum_{l=0}^{m}\sum_{i=0}^{n} -logP_{Ts}^{l}(c_i) + \sum_{l=0}^{m}\sum_{i=0}^{n} -logP_{Te}^{l}(c_i)) \tag{8}$$

L_T denotes the average of cross entropy of output probabilities of all binary classifiers which detect starts and ends of triggers on each type label. In the same way, we define L_A as the loss function of all binary classifiers that are responsible for detecting event relation triples:

$$L_A = \frac{1}{m \times n}(\sum_{r=0}^{m}\sum_{i=0}^{n} -logP_{As}(c_k, r|t) + \sum_{r=0}^{m}\sum_{i=0}^{n} -logP_{Ae}(c_k, r|t)) \tag{9}$$

Where m denotes the sum of event label types and argument role types. L_A denotes the average of cross entropy of output probabilities of all binary classifiers which detect starts and ends of arguments on each role. The final loss function $L_E = L_T + L_A$. We minimize the final loss function to optimize the parameters of the model.

Algorithm 1 trigger and argument identification

Input: P_{Ts}^{l},P_{Te}^{l},P_{As}, P_{Ae},predicted trigger matrix TP,predicted argument matrix AP, sentence s, label list L
Output: predicted trigger list L_T, length of L_T l, predicted argument list L_A

Take out matrix S_t of ids and labels of starts that satisfy $P_{Ts}^{l} > \delta_s^l$ from TP and matrix E_t of ids and labels of ends that satisfy $P_{Te}^{l} > \delta_e^l$ from AP
for each (id_s,l_s) in S_t **do**
  for each (id_e,l_e) in E_t **do**
    if $id_s < id_e \& l_s == l_e$ **then**
      $trigger \leftarrow s[id_s - 1, id_e]$
      $label \leftarrow L[l_e]$
      $Append[trigger, label] to L_T$
      $break$
    end if
  end for
end for
return L_t
if L_T **then**
  for $i = 0 \rightarrow l$ **do**
    Take out matrix S_{ai} of ids and labels of starts that satisfy $P_{As} > \varepsilon_s^r$ from AP and matrix E_{ai} of ids and labels of ends that satisfy $P_{Ae} > \varepsilon_e^r$ for ith trigger from AP
    for each (id_{si},r_{si}) in S_{ai} **do**
      for each (id_{ei},r_{ei}) in E_{ai} **do**
        if $id_{si} < id_{ei} \& r_{si} == r_{ei}$ **then**
          $argument \leftarrow s[id_{si} - 1, id_{ei}]$
          $role \leftarrow L[r_{ei}]$
          $Append[argument, role] to L_A$
          $break$
        end if
      end for
    end for
  end for
end if
return L_A

4 Experiments

We evaluate JMCEE framework on the ACE 2005 dataset that contains 633 Chinese documents. We follow the same setup as (Chen and Ji, 2009; Lin et al., 2018; Zeng et al., 2016), in which 549/20/64 documents are used for training/development/test set. The proposed model is compared with the following state-of-the-art methods:

1) DMCNN (Chen et al., 2015) adopts dynamic multi-pooling CNN to extract sentence-level features automatically.

2) Rich-C (Chen and Ng, 2012) is a joint-learning, knowledge-rich approach including character-based features and discourse consistency features, which is the feature-based state-of-art system.

3) C-BiLSTM (Zeng et al., 2016) designs a convolutional Bi-LSTM model which conduct Chinese event extraction from perspective of a character-level sequential labeling paradigm.

4) NPNs (Lin et al., 2018) performs event extraction in a character-wise paradigm, where a hybrid representation for each character is learned to capture both structural and semantic information from both characters and words.

ACE 2005 dataset annotates 33 event subtypes and 35 role classes. The tasks of event trigger classification and argument classification in this paper are combined into a 70-category task along with "None" word and "Other" type. In order to evaluate the effectiveness of our proposed model, we evaluate models by micro-averaged Precision (P), Recall (R) and F1-score followed the computation measures of Chen and Ji (2009). The following criteria are utilized to evaluate the performance of predicted results:

1) A trigger prediction is correct only if its span and type match with the golden labels.

2) An argument prediction is correct only if its span, role, related trigger and trigger type match with the golden labels.

It is worth noting that all the predicted roles for an argument are required to match with the golden labels, instead of just one of them. We take a further step to see the impacts of pipelined model and joint model. The pipelined model called MCEE which identifies triggers and arguments in two separate stages based our classification algorithm. The highest F-score parameters on the development set are picked and listed in Table 1.

Hyper-parameter	Trigger classification	Argument classification
character embedding	768	768
maximum length	510	510
batch size	8	8
learning rate of Adam	0.0005	0.0005
classification thresholds	[0.5,0.5,0.5,0.5]	[0.5,0.4,0.5,0.4]

Table 1: Hyper-parameters for experiments.

4.1 Overall Results

Table 2 shows the results of trigger extraction on ACE 2005. As is seen, our JMCEE framework achieves the best F1 scores for trigger classification among all the compared methods.

Note that the results of Rich-C could obtain more accurate estimation of model performance since it performed 10-fold cross-validation experiments. However, our JMCEE gains at least 8% F1-score improvements on trigger classification task on ACE 2005, which steadily outperforms all baselines. The improvement on the trigger extraction is quite significant, with a sharp increase of near 10% on the F1 score compared with these conventional methods.

Model	Trigger identification			Trigger classification		
	P	R	F1	P	R	F1
DMCNN	66.6	63.6	65.1	61.6	58.8	60.2
Rich-C	62.2	71.9	66.7	58.9	68.1	63.2
C-BiLSTM	65.6	66.7	66.1	60.0	60.9	60.4
NPNs	75.9	61.2	67.8	73.8	59.6	65.9
MCEE(BERT-Pipeline)	82.5	78.0	80.2	72.6	68.2	70.3
JMCEE(BERT-Joint)	**84.3**	**80.4**	**82.3**	**76.4**	**71.7**	**74.0**

Table 2: Comparison of different methods on Chinese trigger extraction on ACE 2005 test set. Bold denotes the best result.

Table 3 shows results of argument extraction. Compared with these baselines, our JMCEE is at least 3% higher over other models on F1-score on argument classification task. While the improvement in argument extraction is not so obvious comparing to trigger extraction. This is probably due to the rigorous evaluation metric we have taken and the difficulty of argument extraction. Note that by our approach we identify 89% overlap roles in test set. Moreover, results show that our joint model substantially outperforms the pipelined model whether on trigger classification or argument classification. It is seen that joint model enables to capture the dependencies and interactions between the two subtasks and communicate deeper information between them, and thus improves the overall performance.

Model	Argument identification			Argument classification		
	P	R	F1	P	R	F1
Rich-C	43.6	**57.3**	49.5	39.2	**51.6**	44.6
C-BiLSTM	53.0	52.2	52.6	47.3	46.6	46.9
MCEE(BERT-Pipeline)	59.5	40.4	48.1	51.9	37.5	43.6
JMCEE(BERT-Joint)	**66.3**	45.2	**53.7**	**53.7**	46.7	**50.0**

Table 3: Comparison of different methods on Chinese argument extraction on ACE 2005 test set. Bold denotes the best result.

4.2 The Effect of Classification Thresholds

The effectiveness of thresholds settings for the trigger and argument classification is studies in this subsection. Table 4 lists the results of thresholds settings of the starts and ends of both two tasks. Specially, we tune two set of thresholds of starts and ends of trigger and arguments through setting δ^l to be 0.5, 0.5 and setting ε^r ranging from 0.5 to 0.4. Then, set δ^l to be 0.5, 0.4 and set ε^r ranging from 0.5 to 0.4. By analyzing the results, we find that the best performance of JMCEE on trigger extraction is achieved with parameters 0.5, 0.5, 0.5, 0.5, while the best performance of JMCEE on argument extraction is achieved with parameters 0.5, 0.4, 0.5, 0.4.

Proceedings of the 19th China National Conference on Computational Linguistics, pages 950-961,
Hainan, China, October 30 - Novermber 1, 2020. (c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

It suggests that when the ends of thresholds of both trigger and argument classification are set to be 0.4 could identify more candidate triggers and arguments. More candidate triggers could contribute to identifying arguments as we incorporate inter-dependencies between event triggers and argument roles in our joint extraction architecture, while the increased triggers could bring more noise to trigger classification with decreasing on the F1 score.

δ_l		ε_r		Trigger classification			Argument classification		
Start	End	Start	End	P	R	F1	P	R	F1
0.5	0.5	0.5	0.5	**76.4**	**71.7**	**74.0**	53.4	43.7	48.0
0.5	0.5	0.5	0.4	71.2	68.9	70.0	50.3	44.9	47.5
0.5	0.5	0.4	0.5	74.1	69.6	71.8	52.6	45.7	48.9
0.5	0.4	0.5	0.5	74.6	69.2	71.8	49.5	44.2	46.7
0.5	0.4	0.5	0.4	73.8	71.4	72.6	**53.7**	**46.7**	**50.0**
0.5	0.4	0.4	0.5	72.0	70.7	71.3	47.8	47.5	47.7

Table 4: Results of thresholds settings for the start and end of trigger and argument classification. Bold denotes the best result

Overall, the experimental results are remarkable facts given that our framework achieves better performance without any external and manually-generated features. We consider this as a strong promise toward our proposed joint framework which could be used as a good starting point.

5 Conclusions

In this paper, we propose a simple yet effective joint Chinese multiple events extraction framework which jointly extracts triggers and arguments by adopting a pre-trained BERT encoder without elaborate engineering features. Our contribution in this work is as follows:

1) Event relation triple is defined and incorporated into our framework to learn inter-dependencies among event triggers, arguments and arguments roles, which solves the roles overlap problem.

2) Our framework performs event extraction in a character-wise paradigm by utilizing multiple sets of binary classifiers to determine the spans, which allows to extract multiple events and relation triples and avoids Chinese language specific issues such as word-trigger mismatch and word boundary problem.

Experiments have shown that our method outperforms conventional methods. We believe our proposed framework could be applied to many other NLP tasks for exploiting inner composition structure during extraction, such as Entity Relation Extraction. Our future work will focus on data generation to enrich training data and try to extend our framework to the open domain.

Acknowledgements

This work has been supported by National Key Research and Development Program(No.2019YFB1406302), China Postdoctoral Science Foundation(NO.2020M670057) and Beijing Postdoctoral Research Foundation(No.ZZ2019-92).

References

Chen Chen and Vincent Ng. 2012. *Joint Modeling for Chinese Event Extraction with Rich Linguistic Features*. Proceedings of COLING 2012, 529–544.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng and Jun Zhao. 2015. *Event Extraction via Dynamic Multi-pooling Convolutional Neural Networks* . Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing,vol.1,167–176.

Zheng Chen and Heng Ji. 2009. *Language Specific Issue and Feature Exploration in Chinese Event Extraction*. Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 209–212.

Jacob Devlin, Ming-W. Chang, Kenton Lee and Kristina Toutanova. 1972. *Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv preprint arXiv:1810.04805

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel and Ralph Weischedel. 2004. *The Automatic Content Extraction (ACE) Program-tasks, Data, and Evaluation*. LREC,vol.2

Xiaocheng Feng, Bing Qin and Ting Liu. 2016. *A Language-independent Neural Network for Event Detection.*. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, vol.2, 66–71.

Ruifang He and Shaoyang Duan. 2019. Joint Chinese Event Extraction based Multi-task Learning. *Journal of Software*, 30(4):1015–1030.

Qi Li, Heng Ji and Liang Huang. 2013. *Joint Event Extraction via Structured Prediction with Global Features*. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, vol.1, 73–82.

Shasha Liao and Ralph Grishman. 2010. *Using Document Level Cross-event Inference to Improve Event Extraction*. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 789–797.

Hongyu Lin, Yaojie Lu, Xianpei Han and Le Sun. 2018. *Joint Chinese Event Extraction based Multi-task Learning*. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics,1565–1574.

Jian Liu, Yubo Chen, Kang Liu and Jun Zhao. 2018. *Event Detection via Gated Multilingual Attention Mechanism*. Proceedings of the 32nd AAAI Conference on Artificial Intelligence, 4865–4872.

Xiao Liu, Zhunchen Luo and Heyan Huang. 2018. *Jointly Multiple Events Extraction via Attention-based Graph Information Aggregation*. arXiv preprint arXiv:1809.09078.

Bryan McCann, James Bradbury, Caiming Xiong and Richard Socher. 2017. Learned in Translation: Contextualized Word Vectors. *In Advances in Neural Information Processing Systems*, 6294–6305.

Trung-M. Nguyen and Thien-H. Nguyen. 2019. *One for all: Neural Joint Modeling of Entities and Events*. Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 6851–6858.

Thien Huu Nguyen, Kyunghyun Cho and Ralph Grishman. 2016. *Joint Event Extraction via Recurrent Neural Networks*. Proceedings of NAACL-HLT 2016, 300–309.

Thien Huu Nguyen and Ralph Grishman. 2016. *Modeling Skip-grams for Event Detection with Convolutional Neural Networks*. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 886–891.

Matthew E. Peters, Mark Neumann, Mohit Iyyer and Matt Gardner. 2018. *Deep Contextualized Word Representations*. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, 2227–2237 .

Alec Radford, Karthik Narasimhan, Tim Salimans and Ilya Sutskever. 2018. *Improving Language Understanding by Generative Pre-training*. https://www.cs.ubc.ca/ amuham01/LING530/papers/radford2018improving.pdf.

Lei Sha, Feng Qian, Baobao Chang and Zhifang Sui. 2018. *Jointly Extracting Event Triggers and Arguments by Dependency-bridge RNN and Tensor-based Argument Interaction*. Proceedings of the 32nd AAAI Conference on Artificial Intelligence, 5916–5923.

Sen Yang, Dawei Feng, Linbo Qiao, Zhigang Kan and Dongsheng Li. 2019. *Exploring Pre-trained Language Models for Event Extraction and Generation*. Proceedings of the 57th Annual Meeting of the Association for Computa-tional Linguistics, 5284–5294.

Ying Zeng, Honghui Yang, Yansong Feng and Dongyan Zhao. 2016. *A Convolution Bilstm Neural Network Model for Chinese Event Extraction*. In: Lin CY., Xue N., Zhao D., Huang X., Feng Y. (eds) Natural Language Understanding and Intelligent Applications. ICCPOL 2016, NLPCC 2016. LNCS, vol. 10102, 275–287.

Proceedings of the 19th China National Conference on Computational Linguistics, pages 950-961,
Hainan, China, October 30 - Novermber 1, 2020. (c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China