MULTI-ROLE EVENT ARGUMENT EXTRACTION AS MACHINE READING COMPREHENSION WITH ARGUMENT MATCH OPTIMIZATION

Jingcong Tao[1,*], *Youcheng Pan*[1,*], *Xinyu Li*[2,*], *Baotian Hu*[1,†], *Weihua Peng*[2,†], *Cuiyun Han*[2], *Xiaolong Wang*[1]

[1]Harbin Institute of Technology, Shenzhen
[2]Baidu International Technology (Shenzhen) Co., Ltd
20s051009@stu.hit.edu.cn, panyoucheng4@gmail.com,
{hubaotian, xlwangsz}@hit.edu.cn, {lixinyu13, pengweihua, hancuiyun}@baidu.com

ABSTRACT

Extracting arguments for the pre-defined roles is a crucial step for event extraction. Recently, there are some insightful works that view it as a machine reading comprehension problem and achieve significant progress. However, most of them need multi-turns to extract the arguments of each role independently, which ignores the relationships among roles in the same event. To alleviate this problem, we propose a novel Multi-Role Argument Extraction method named MRAE which can exploit the relationship of event roles by extracting all arguments for an event simultaneously. To force MRAE to locate more arguments accurately, we propose an argument match optimization loss based on the minimum risk training to exploit sentence-level F1 score. We conduct experiments on the widely used ACE2005 dataset. The experimental results demonstrate that MRAE outperforms the competitor methods by at least +1.2% F1 score on argument extraction, and also shows superiority on data scarce scenarios.

Index Terms— multi-role event argument extraction, machine reading comprehension, minimum risk training, data scarce

1. INTRODUCTION

Event argument extraction (EAE) is a challenging subtask of event extraction (EE), whose purpose is to identify the corresponding arguments from the sentence according to the trigger and pre-defined roles of a specific event. Figure 1 illustrates an example, in which the sentence contains two events: *Personnel End-position* and *Personnel Start-position*. Figure 1(a) shows the *Personnel End-position* event with the trigger word '*steps down*', and three arguments: *The pro-reform director of Iran's biggest-selling daily newspaper* (role=**person**), *Tehran* (role=**place**) and *Iran's biggest-selling daily newspaper* (role=**entity**). Figure 1(b) illustrates the *Personnel Start-position* event with the trigger word '*appointment*', and two arguments: *a conservative* (role=**person**) and *the city* (role=**entity**).

Traditional EAE methods tend to identify entities firstly and then classify them to their roles [1, 2, 3, 4]. For example, in Figure 1, traditional methods firstly recognize an entity *Tehran*, and then assign the entity with the role *place*, which heavily relies on the entity recognition and may result in error propagation. In order to mitigate this issue, recent works view EAE as a machine reading comprehension (MRC) task. Li et al. [5] are the first to use multi-turn question answering on event argument extraction, in which they construct a question for each role to extract the arguments from the sentence. To explore a more proper question form, Liu et al. [6] pre-define some question templates like "where is the place". After that, to make questions more informative, Du et al. [7] propose to train a question generation model to generate questions. To obtain a closer relationship among roles and arguments, Zhou et al. [8] introduce a joint training task that asks the model what is the argument for the role and what is the role for the argument.

However, these MRC-based works only extract arguments for one role each time, whose inputs consisting of a question and a sentence for argument extraction model are in the form like:

[CLS] Where is the place [SEP] The pro-reform director ...

[CLS] Who is the person [SEP] The pro-reform director ...

But this kind of question form extracts arguments for each role independently, which ignores the connection among roles in the same event. As illustrated in Figure 1, both of *End-position* event and *Start-position* event have the roles *place* and *entity*. If we extract them independently, the model may extract the *person* in the *End-position* event for the *entity* role in the *Start-position* event.

In this work, we propose a novel Multi-Role Argument Extraction model (MRAE) to extract all the roles from the same event simultaneously. Our inputs are like:

[CLS] place [CLS] person ... [SEP] The pro-reform director ...

which combines all roles in the same event as the questions and uses [CLS] to separate each other. Then, we utilize [SEP] to connect the roles with the sentence. In this way, we can encode the sentence for all the roles in an event at the same time.

Moreover, current MRC frameworks tend to consider the token-level loss to determine whether a token is a start or an end position of an argument independently, which cannot guarantee that all the arguments are accurately extracted, especially for the situation where there are multiple arguments for a role. To address this issue, we propose a novel argument match optimization (AMO) algorithm based on the minimum risk training (MRT) [9, 10, 11] to optimize our model by using sentence-level F1 score to evaluate the contents of all the arguments instead of the start or end position respectively.

---

* equal contribution

† corresponding author

This work is jointly supported by grants: Natural Science Foundation of China (No.62006061), Stable Support Program for Higher Education Institutions of Shenzhen (No.GXWD20201230155427003-20200824155011001) and Strategic Emerging Industry Development Special Funds of Shenzhen (No.JCYJ20200109113441941).

The pro-reform director of Iran's biggest-selling daily newspaper and official organ of Tehran's municipality has stepped down following the appointment of a conservative as the city's new mayor, press reports said Sunday.

Event type	Personnel End-position	**Trigger**	stepped down
Arguments	Role=person	The pro-reform director of Iran's biggest-selling daily newspaper	
	Role=place	Tehran	
	Role=entity	Iran's biggest-selling daily newspaper	

(a) Personnel End-position

Event type	Personnel Start-position	**Trigger**	appointment
Arguments	Role=person	a conservative	
	Role=entity	the city	

(b) Personnel Start-position

Fig. 1: A sentence that contains two events: (a) Personnel End-position event & (b) Personnel Start-position event

We evaluate our proposed method on the widely used benchmark ACE2005, and the experimental result reaches 54.6% (+1.2%) F1 score on the argument classification, which outperforms the previous state-of-the-art demonstrates the effectiveness of extracting the multiple roles simultaneously with AMO. Our contributions can be summarized as follow:

- We propose a novel model MRAE that learns the internal connections among roles to extracts all the arguments from the same event simultaneously.
- We propose an MRT-based AMO algorithm on EAE task that uses the sentence-level F1 score to optimize our model.
- We conduct a series of experiments to prove the effectiveness of our proposed model, and improve the result of F1 score to 54.6% (+1.2) on argument classification.

2. METHOD

In this section, we will introduce the details of our MRAE, including (1) how MRAE extracts arguments for all roles in an event simultaneously; (2) how we apply AMO on the proposed MRAE to promote the extraction performance. The overall architecture of MRAE is illustrated in Figure 2.

2.1. Framework

As shown in Figure 2, we treat each role as a question and separate them with [CLS]. After questions, we use [SEP] to connect the

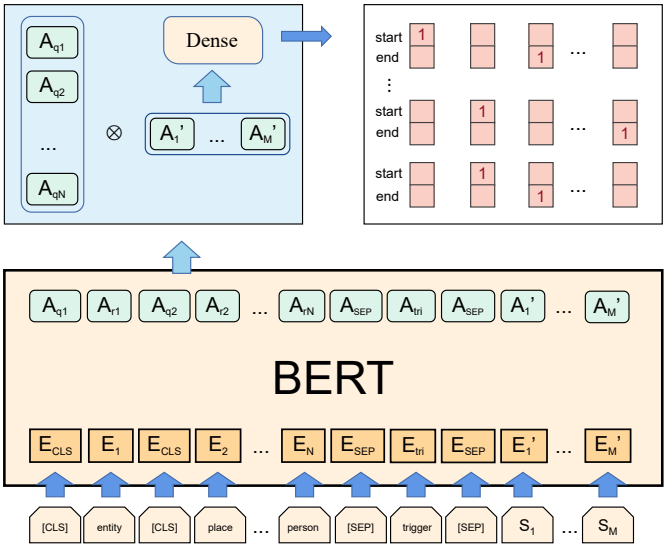

Fig. 2: Overview of our MRAE model. We use the representations of [CLS] before each role as the question representation, and then compute multiple role-specific sentence representations to extract the corresponding arguments.

questions with trigger words and the sentence S.

$$I = [\text{CLS}]\ Q_1\ [\text{CLS}]\ Q_2, ...[\text{SEP}]\ t_1, t_2, ...[\text{SEP}]\ s_1, s_2...$$

I is the input of our model; Q indicates the questions; and t is the trigger. Then we have representations of input sequence:

$$\boldsymbol{A_1}, \boldsymbol{A_2}, ..., \boldsymbol{A_l} = \text{BERT}(I_1, I_2, ..., I_l)$$

To separately extract arguments for each role, we choose the hidden state of the [CLS] before each role question as the question representation. Then, we obtain sentence representations for each role. We have:

$$\boldsymbol{H_s^{q_i}} = \boldsymbol{A_{q_i}} \otimes \boldsymbol{A_s}$$

where $\otimes$ indicates point multiplication; A_{q_i} is the representations of the i-th question; A_s is the sentence representations from BERT [12]; and $H_s^{q_i}$ is the sentence representations for the i-th role question. The possibility of each token to be selected as the start/end of the argument is:

$$p_i = sigmoid(\boldsymbol{H_s^{q_i}} W_i)$$

where $W_i \in \mathbb{R}^{d \times 2}$ is trainable parameter. p_i indicates the possibility of each token being a start or an end of an argument. Then, we can compute the prediction loss of the start/end positions as below:

$$L_{se} = \frac{1}{l} \sum \text{CE}(y_i, p_i)$$

where CE indicates the cross entropy loss function.

2.2. Argument Match Optimization

L_{se} just helps to optimize whether a token is a start or an end of an argument, which is a local loss and not directly related to the evaluation metric. To realize a global optimization, we propose an argument match optimization algorithm based on minimum risk training,

whose objective is to minimize the expected loss as below:

$$L_{amo}(\theta) = \sum_{(x,y)\in D} \sum_{y'\in\gamma(x)} p(y'|x,\theta)\Delta(y',y) \quad (1)$$

where D is the training set, x is the input sequence for an event and y is the golden arguments for the event roles. $\gamma(x)$ refers to the set of some possible output y' predicted by MRAE. $p(y'|x,\theta)$ indicates the possibility of y' being the arguments predicted by the model. $\Delta(y',y)$ measures the loss between y' and the golden arguments y.

Algorithm 1 The sampling algorithm

Input: our model θ, input x, sample number k, parameter μ
Output: A subset of all model outputs $\gamma(x)$

```
p ~ P(·|x, θ)
ȳ ← get_arguments(p)
γ(x) ← ȳ
for i = 1, 2, ..., k do
    p' ~ p + μN(0, 1)
    y' ← get_arguments(p')
    γ(x) ← γ(x) ∪ y'
end for
```

Algorithm 1 shows the sampling process to obtain $\gamma(x)$. In our framework, the original model output is a start-end possibility vector, which we denote as p. Then we deduce the model predicted arguments $\bar{y}$ according to p, and add $\bar{y}$ as the first member of $\gamma(x)$. To get other possible outputs in $\gamma(x)$, we add a normal distribution noise on p and get the corresponding arguments y'. We repeat the process for k times to get k noisy sample y', and add them to $\gamma(x)$.

Sun et al. [13] collects samples to obtain $\gamma(x)$ by randomly changing the label directly, e.g. from B to O, ignoring the possibilities. To address this problem, we add a normal distribution on the output possibilities. Thus tokens that possibility near the threshold are more likely to change the labels.

For all y' in $\gamma(x)$, we calculate the cross-entropy loss between them and p. This represents the similarity between noisy sample y' and the original model output p, which we denote as:

$$sim(y',p) = \mathrm{CE}(y',p) \quad (2)$$

Then we get the probability that y' output by the model:

$$p(y'|x,\theta) = \frac{e^{sim(y',p)}}{z} \quad (3)$$

$$z = \sum_{y'\in\gamma(x)} e^{sim(y',p)} \quad (4)$$

Then for $\Delta(y',y)$, which measures the loss between sample y' and golden arguments y, we use F1 score as the loss function. However, an event may find no corresponding argument in the sentence. Hence we calculate the F1 score for a batch of events each time.

$$\Delta(y',y) = 1 - \mathrm{F1}(y',y) \quad (5)$$

This equation enables the model to be optimized globally with F1 scores. By replace equation 1 with equation 2,3,4, we can get L_{amo}. Our total loss is:

$$L = \alpha L_{amo} + \beta L_{se}$$

where α, β are the weight parameters of each part.

3. EXPERIMENTS

3.1. Dataset

We train and test our model on widely used ACE2005 dataset, which contains 33 event types and 35 semantic roles. To ensure our results are comparable with previous works, we use the same data split and criteria with Li [14]. There are 529 documents for training set, 40 documents for the testing set, and 30 documents for the development set.

3.2. Experimental Settings

To make a fair comparison with previous works, we compare our model on four sub-tasks: (1) Trigger Identification (TI): a trigger is correctly identified only if the span matches the offset of a golden trigger. (2) Trigger Classification (TC): a trigger is correctly classified only if it is correctly identified and classified to the corresponding event type. (3) Argument Identification (AI): an argument is correctly identified only if the span matches a golden span and the event type is right. (4) Argument Classification (AC): an argument is correctly classified only if its offset, event type and role type match a reference argument.

Since we concentrate on the argument extraction task in this paper, we directly adopt the trigger identification and classification model by following BERT_QA_Arg [7]. F1 score is used to evaluate the performance on all the mentioned sub-tasks.

We use Bert-large [12] as our pretrained model. We set learning rate as 3e-5, hidden size as 1024, batch size as 16, epoch as 100, sequence length as 180, μ as 0.2 and k as 10.

3.3. Competitor Methods

We compare our proposed model with several competitor methods: **DMCNN** [3] a method that adopt convolution neural networks to capture sentence feature for EE. **dbRNN** [2] a method that employs sentence syntax information captured by RNNs and GCNs for EE. **JMEE** [4]: a joint framework that uses attention-based GCNs and RNNs to obtain syntax information to extract triggers and arguments for EE. **DYGIE++** [1]: a multi-task information extraction framework on extracting entities, events and relations using contextualized span representations. **Joint3EE** [15]: a joint framework to predict entities, triggers and arguments on the shared encoder representation. **MQAEE** [5]: a pipeline framework that takes EE as multi-turn question answering, which view each role as an independent question. **BERT_QA_Arg** [7]: a MRC-based model that tries different manually defined question templates.

3.4. Result and Analysis

According to the overall experimental results shown in Table 1, we have the following observations. Firstly, compared with the traditional extraction methods, the MRC-based methods are able to achieve better performance on event argument extraction, including that our proposed MRAE obtains the best F1 score among all the competitor methods. Secondly, compared with the other MRC-based methods, i.e. MQAEE and BERT_QA_Arg, MRAE outperforms them by at least +1.9% F1 score on argument identification and +1.2% F1 score on argument classification especially when the performance of trigger extraction is close, which demonstrates the superiority of MRAE.

In order to validate the effectiveness of the multi-role extraction and AMO, we also conduct an ablation study for analysis. As

Model	TI	TC	AI	AC
DMCNN [3]	–	69.1	–	48.0
dbRNN [2]	–	71.9	57.2	50.1
JMEE [4]	–	–	–	50.4
DYGIE++ [1]	–	68.9	54.1	51.4
Joint3EE [15]	72.5	69.8	–	52.1
MQAEE [5]	74.5	71.7	55.2	53.4
BERT_QA_Arg [7]	75.8	72.4	55.3	53.3
MRAE	75.8	72.4	**57.2**	**54.6**
w/o AMO	75.8	72.4	56.5	53.6
w/o Multi	75.8	72.4	56.5	53.8

Table 1: Experimental results by F1 score (%) of the four sub-tasks. '–' means that the corresponding results were not reported in the original paper.

shown in Table 1, when not employing AMO (w/o AMO) or multi-role extraction (w/o Multi), the performance of MRAE on argument extraction drops a bit but still outperforms all the competitor methods, which indicates that both the multi-role extraction and AMO are beneficial to improve the performance.

3.5. Results in Data Scarce Scenarios

Model	1%	5%	10%	20%
DMCNN [3]	–	8.7	16.6	23.7
dbRNN [2]	–	8.1	17.2	24.1
BERT_QA_Arg [7]	1.8	23.4	34.8	48.7
MRAE	**6.0**	**25.9**	**36.5**	**50.3**

Table 2: Experimental results by F1 score (%) of argument extraction on data scarce scenarios

Table 2 shows the experimental results of EAE on data-scarce scenarios. Golden triggers are provided in this experiment setting. We use 1%, 5%, 10%, 20% of the train data to fulfill the data-scarce scenarios. Compared with traditional neural networks DMCNN and dbRNN, we can observe that the MRC-based methods including BERT_QA_Arg and MRAE have outperformed them largely in all the scenarios. In comparison with the single-role MRC model BERT_QA_Arg, our proposed model MRAE achieves superior performance, with over +4.2%, +2.5%, +1.7%, +1.6% improvement on 1%, 5%, 10%, 20% settings, respectively. Noting that with less train data, MRAE can surpass other methods more obvious on the F1 score. We argue that MRAE can learn more information with multi-role extraction in the data-scarce scenarios. But as the amount of train data grows, a number of these information will be learned by the model from the new events.

3.6. Error Analysis

Although our proposed model achieves better performance, there are still many problems to solve. Some wrong predictions made by MRAE on ACE2005 test set are presented in Figure 3, which mostly consist of three types of errors: spurious errors (S-error), boundary error (B-error), and missing error (M-error).

S-error is the error that MRAE extracts a wrong argument for a role while the role has corresponding true argument. In this case, the wrong prediction may be caused by a complex context that confuses our model. More information is needed by the model to distinguish the correct argument for the role.

S-error	
Former senior banker Callum McCarthy begins what is one of the most important jobs in London's financial world in September, when incumbent Howard Davies steps down	
Event Type	personnel start-position
Role	position
Golden Arguments	one of the most important jobs in London's financial world
Predicted Arguments	senior banker
B-error	
Within weeks he was arrested and charged with sodomising an official driver several years previously and with abusing his powers to cover up the offence	
Event Type	justice charge-indict
Role	crime
Golden Arguments	(1) sodomising an official driver (2) abusing his powers to cover up the offence
Predicted Arguments	sodomising an official driver several years previously and with abusing his powers to cover up the offence
M-error	
As well as previously holding senior positions at Barclays Bank, BZW and Kleinwort Benson,McCarthy was formerly a top civil servant at the Department of Trade and Industry	
Event Type	personnel end-position
Role	entity
Golden Arguments	(1) the department of trade and industry (2) BZW (3) Kleinwort Benson (4) Barclays Bank
Predicted Arguments	(1) the department of trade and industry (2) BZW

Fig. 3: Examples of error analysis (S-error, B-error and M-error).

B-error is the most common error occurred in the extraction tasks, in which the model cannot identify the accurate boundaries of the arguments. There are two kinds of B-errors: The first is that the model cannot identify the core entity while always extract some other inessential modifiers around it, which may be caused by the annotated dataset; the second kind of B-errors is that model cannot separate two arguments connected by words like "and", as shown in Table 3. We make an attempt to use a span-level loss function [16] to solve this problem, but it doesn't work well. In this case, stronger constrains is needed for the model.

M-error indicates that the model predicts no arguments for the role while there actually exists golden arguments in the sentence. This kind of errors are caused by the lack of training data. Some arguments in the test data do not appear in the training data. With a larger scale of data, this kind of error may be reduced.

4. CONCLUSION

In this paper, we propose a novel EAE model MRAE to extract all the arguments simultaneously. MRAE can capture the inner connection among roles and optimize with a sentence-level loss AMO. We compared MRAE with the competitor methods and experiment results demonstrate that MRAE outperforms previous models on ACE2005 dataset.

5. REFERENCES

[1] David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi, "Entity, relation, and event extraction with contextualized span representations," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 5788–5793.

[2] Lei Sha, Feng Qian, Baobao Chang, and Zhifang Sui, "Jointly extracting event triggers and arguments by dependency-bridge rnn and tensor-based argument interaction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, vol. 32.

[3] Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao, "Event extraction via dynamic multi-pooling convolutional neural networks," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015, pp. 167–176.

[4] Xiao Liu, Zhunchen Luo, and He-Yan Huang, "Jointly multiple events extraction via attention-based graph information aggregation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 1247–1256.

[5] Fayuan Li, Weihua Peng, Yuguang Chen, Quan Wang, Lu Pan, Yajuan Lyu, and Yong Zhu, "Event extraction as multi-turn question answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 2020, pp. 829–838.

[6] Jian Liu, Yubo Chen, Kang Liu, Wei Bi, and Xiaojiang Liu, "Event extraction as machine reading comprehension," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 1641–1651.

[7] Xinya Du and Claire Cardie, "Event extraction by answering (almost) natural questions," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 671–683.

[8] Yang Zhou, Yubo Chen, Jun Zhao, Yin Wu, Jiexin Xu, and Jinlong Li, "What the role is vs. what plays the role: Semi-supervised event argument extraction via dual question answering," 2021.

[9] Franz Josef Och, "Minimum error rate training in statistical machine translation," in *Proceedings of the 41st annual meeting of the Association for Computational Linguistics*, 2003, pp. 160–167.

[10] David A Smith and Jason Eisner, "Minimum risk annealing for training log-linear models," in *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, 2006, pp. 787–794.

[11] Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu, "Minimum risk training for neural machine translation," in *ACL (1)*, 2016.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, June 2019, pp. 4171–4186, Association for Computational Linguistics.

[13] Changzhi Sun, Yuanbin Wu, Man Lan, Shiliang Sun, Wenting Wang, Kuang-Chih Lee, and Kewen Wu, "Extracting entities and relations with joint minimum risk training," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2256–2265.

[14] Qi Li, Heng Ji, and Liang Huang, "Joint event extraction via structured prediction with global features," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2013, pp. 73–82.

[15] Trung Minh Nguyen and Thien Huu Nguyen, "One for all: Neural joint modeling of entities and events," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 6851–6858.

[16] Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li, "A unified mrc framework for named entity recognition," in *ACL*, 2020.