Enhancing Dialogue-based Relation Extraction by Speaker and Trigger Words Prediction

Tianyang Zhao[◊§*], **Zhao Yan**[§], **Yunbo Cao**[§], **Zhoujun Li**[◊†]
[◊]State Key Lab of Software Development Environment, Beihang University, Beijing, China
[§]Tencent, Beijing, China
[◊]{tyzhao,lizj}@buaa.edu.cn; [§]{zhaoyan,yunbocao}@tencent.com

Abstract

Identifying relations from dialogues is more challenging than traditional sentence-level relation extraction (RE), since the difficulties of speaker information representation and the long-range semantic reasoning. Despite the successful efforts, existing methods do not fully consider the particularity of dialogues, making them difficult to truly understand the semantics between conversational arguments. In this paper, we propose two beneficial tasks, speaker prediction and trigger words prediction, to enhance the extraction of dialogue-based relations. Specifically, speaker prediction captures the characteristics of speaker-related entities, and the trigger words prediction provides supportive contexts for relations between arguments. Extensive experiments on the DialogRE dataset show noticeable improvements compared to the baseline models, which achieves a new state-of-the-art performance with a 65.5% of F1 score and a 60.5% of $F1_c$ score, respectively.

Dialogue 1

S1: Yeah you see umm, well, I'm an actor. Right? ……

Relation: (***S1***, per:title, actor)
Trigger Words: -

Dialogue 2

S1: Mom!
S2: Sweetie! So this is where you work? ……

Relation: (***S1***, per:parents, ***S2***)
Trigger Words: mom

Dialogue 3

S1: Hello, Mr. Bing.
S2: Loved your Stevie Wonder last night.
S3: Thanks. Listen, about the weekly numbers, I'm gonna need them on my desk by nine o'clock.
S1: Sure.
S2: No problem.

Relation: (***S1***, per:boss, ***S3***)
(***S2***, per:boss, ***S3***)
Trigger Words: -

Figure 1: Examples from the DialogRE dataset. s_n denotes the speaker of each utterance. The underlined text indicates the relation between the argument pairs.

1 Introduction

The task of relation extraction is to identify the relation facts between two arguments from plain text, which is the fundamental step of many natural language processing applications. Recent years have seen increasing efforts on sentence-level RE, e.g., relations only hold within a single sentence (Fu et al., 2019; Luan et al., 2019; Zhao et al., 2020; Wang and Lu, 2020; Wei et al., 2020). To adapt to complex scenarios, some current works have moved forward to the document-level RE, e.g., relations can exist across multiple sentences (Yao et al., 2019; Wang et al., 2020; Nan et al., 2020; Jain et al., 2020; Zhou et al., 2021).

A more challenging yet practical extension is the dialogue-based relation extraction. The dialogues contain multi-turn conversations among a group of speakers. The relations not only exist between the entities in the dialogue text but also the speakers of each utterance. Additionally, most of relations appear in multi-turn conversation, which require cross-sentence extraction. Considering the complexities, we divide the dialogue-based RE into three categories. In the first category, the relation can be directly inferred from the current utterance, as shown in the Dialogue 1 of Figure 1. In the second category, the relation involves utterances among multiple speakers and there is clear evidence in the dialogue that triggers the relation. Regarding the Dialogue 2 in Figure 1, "mom" is the trigger word of the relation `per:parents` be-

[*] Work done during an internship at Tencent Cloud Xiaowei.
[†] Corresponding Author.

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4580–4585

tween "S1" and "S2". Nevertheless, there are still cases where there is no clear context indicating the relationship. As shown in the Dialogue 3 of Figure 1, the relation between "S1" and "S3" as well as the relation between "S2" and "S3" can only be inferred from the tones and expression habits of speakers. Therefore, to identify relations from the complex dialogues, it is necessary to 1) discover highly supportive information about the arguments, and 2) capture the unique features of speakers.

Existing studies propose to solve this task through a speaker-aware BERT model (Yu et al., 2020) as well as a gaussian graph-based method (Xue et al., 2021). The former modifies the speaker arguments in dialogue text with special tokens to highlight the speaker-related information. The latter builds a latent multi-view graph to encode the long-distance dependency between arguments. However, these works regard dialogue as a plain text without considering the supporting information of relations and the characteristics of speakers. As been emphasized before (Xue et al., 2021), trigger words and speaker-related features play a critical role in dialogue-based relation extraction. In this case, it is difficult for them to detect the speaker-related relations from the complicated conversations.

To address the above limitations, we propose two beneficial tasks, speaker prediction and trigger words prediction, to enhance the dialogue-based relation extraction. Specifically, the speaker prediction task aims to capture the unique features of speakers. We randomly mask the speaker tokens and use the context to predict who said the utterance. The trigger words prediction task is to detect the supportive context of the current relation. We solve it with a sequence labeling method. Moreover, we design an integration module for the relation extraction task to combine both the global dialogue representation and the local arguments representation. Finally, the three tasks are jointly trained based on a multi-task learning framework.

The contributions of our work are summarized as follows. We propose two beneficial tasks, speaker prediction and trigger words prediction, to capture the unique features of speakers and detect the supportive information about arguments, both effectively enhance the dialogue-based relation extraction. We evaluate our method on the DialogRE dataset and achieve a new state-of-the-art performance with 65.5% of F1 score and 60.5% of $F1_c$ score, respectively. Our code is available at https://github.com/TanyaZhao/DialogRE-Trigger-Speaker-Prediction.

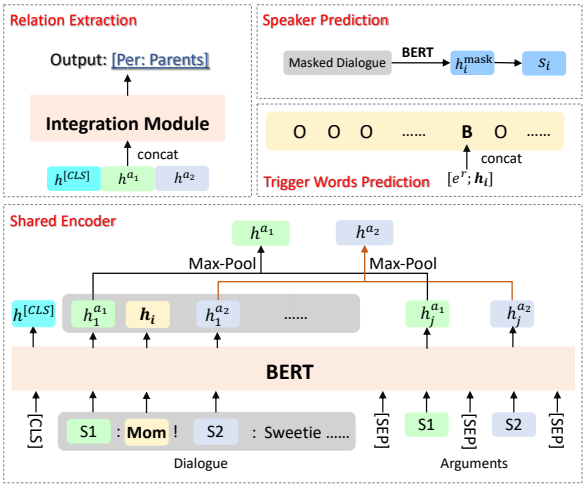

Figure 2: Overall structure of our method.

2 Problem Definition

Given a dialogue $d = s_1 : t_1, s_2 : t_2, ..., s_m : t_m$ and two arguments a_1, a_2, where s_i and t_i are the speaker and the utterance of the i-th turn, $t_i = x_{i1}, x_{i2}, ..., x_{in}$ is consisted of n words. The dialogue-based relation extraction aims to predict the relation type $r \in \mathcal{R}$ between a_1 and a_2, where $\mathcal{R}$ is the set of predefined relation categories.

3 Methodolgy

This section introduces the structure of our method, including three tasks, relation extraction, trigger words prediction and speaker prediction. Figure 2 shows the overall structure of our method.

3.1 Relation Extraction Task

The relation extraction task takes the dialogue d and the argument pair (a_1,a_2) as input and outputs a relation type r between the two arguments. For the dialogue d, we first modify it into $\hat{d} = \hat{s_1} : \hat{t_1}, \hat{s_2} : \hat{t_2}, ..., \hat{s_m} : \hat{t_m}$, where

$$\hat{s}_i = \begin{cases} \texttt{[B]}\ \texttt{[S}_1\texttt{]}\ \texttt{[E]} & \text{if } s_i = a_1 \\ \texttt{[B]}\ \texttt{[S}_2\texttt{]}\ \texttt{[E]} & \text{if } s_i = a_2 \\ s_i & \text{otherwise} \end{cases}, \quad (1)$$

$\hat{t_i} = x_{i1}..., \texttt{[B]}, a_k, \texttt{[E]}, ..., x_{in}, k \in \{1, 2\}$, if t_i contains a_k. Among them, $\texttt{[S}_1\texttt{]}$, $\texttt{[S}_2\texttt{]}$, [B], [E] are newly-defined tokens. [B] and [E] are used to mark the start and the end position of the argument. We further replace the argument a_k

with the pre-defined token [S$_k$], as $\hat{a_k}$ = [S$_k$] if $\exists i(s_i = a_k)$. Then, we concatenate $\hat{d}$, $\hat{a_1}$ and $\hat{a_2}$ as a sequence, and use special tokens [CLS] and [SEP] as sperators, formulated as

$$\texttt{[CLS]}\,\hat{d}\,\texttt{[SEP]}\,\hat{a_1}\,\texttt{[SEP]}\,\hat{a_2}\,\texttt{[SEP]}. \quad (2)$$

We feed the sequence into the pre-trained language model BERT (Devlin et al., 2019) and obtain the hidden semantic representation of each input tokens. Among them, $h^{[CLS]} \in \mathbb{R}^{d_h}$ is the hidden output of [CLS], where d_h is the hidden size of BERT. We use $h^{[CLS]}$ to represent the global relational feature between a_1 and a_2.

To better represent the semantic information of arguments, we distill all the hidden states of a_k's start marker in the sequence (Eq. 2), including those in the dialogue text, and formulate them as $h_1^{a_k}, h_2^{a_k}, ..., h_j^{a_k} \in \mathbb{R}^{d_h}$. Then, we apply a max pooling process to obtain a combined representation of a_k:

$$h^{a_k} = \text{max-pool}(h_1^{a_k}, h_2^{a_k}, ..., h_j^{a_k}). \quad (3)$$

Next, we concatenate $h^{[CLS]}$, h^{a_1} and h^{a_2} as $h = [h^{[CLS]}; h^{a_1}; h^{a_2}] \in \mathbb{R}^{3d_h}$. Note that, $h^{[CLS]}$ is the global relational information of the sequence, and h^{a_1}, h^{a_2} are the local features of the arguments.

Furthermore, we propose an integration module to enhance the correlation between the dialogue and arguments. Specifically, h is fed into a two-layer highway network (Srivastava et al., 2015) as

$$\begin{aligned} \hat{h} &= H(h) * T(h) + h * (1 - T(h)), \\ T(h) &= \sigma(W^t h + b^t), \end{aligned} \quad (4)$$

where $W^t \in \mathbb{R}^{3d_h \times d_t}$, $b^t \in \mathbb{R}^{d_t}$ are learned weights with d_t as the hidden size of the highway network. Finally, we conduct a multi-class classification to calculate the probability of the relation between a_1 and a_2 by

$$\Pr(\text{relation} = r|d, a_1, a_2) = \text{sigmoid}(W^r h + b^r), \quad (5)$$

where $W^r \in \mathbb{R}^{d_t \times |\mathcal{R}|}$, $b^r \in \mathbb{R}^{|\mathcal{R}|}$.

3.2 Trigger Words Prediction Task

Generally, to identify the relation between two arguments, it is necessary to detect the supportive context that triggers the relation. Yu et al. (2020) have verified that trigger words play an important role for relation extraction. However, their work directly append the ground truth trigger words to the input sequence, which is not feasible for scenarios where the golden triggers are not available. The intuitive idea is to predict the trigger words from the conversation. Therefore, we can still obtain supporting information without relying on the golden triggers to guide the relationship extraction.

We propose the trigger words prediction task, which applies a simple and effective way to improve the relation extraction task. Specifically, we perform sequence labeling over the hidden outputs of BERT. Considering the trigger words are closely related to the relation, we first map the predicted relation r (Eq. 5) into a distributed embedding $e^r \in \mathbb{R}^{d_r}$, and concatenate it with each hidden output of BERT as $z_i = [e^r; h_i]$[1]. Then, we predict the boundary label for every token. Formally, the probability of the token x_i with label l is calculated by

$$\Pr(\text{label} = l|x_i) = \text{softmax}(W^l z_i + b^l), \quad (6)$$

where $W^l \in \mathbb{R}^{(d_h+d_r)\times|\mathcal{B}|}$ and $b^l \in \mathbb{R}^{|\mathcal{B}|}$ with $\mathcal{B} = \{\texttt{B}, \texttt{I}, \texttt{O}\}$.

3.3 Speaker Prediction Task

Notably, a majority of relations in the dialogue-based RE are associated with the speakers. For example, the triplet (S1, per:parents, S2) in Figure 1. Different from the ordinary entities, speakers have distinctive personal features, including tone of voices, expression habits, etc., which are important indicators for relation extraction. Therefore, we further propose the speaker prediction task based on the discourse structure to capture the speaker-related features. The motivation behind it is that if the model can distinguish who said the utterance, it learned the speaker's unique information, which is helpful to the speaker-related relation prediction.

Concretely, we randomly select the speaker words s_i in d with a probability of 10% and replace them with a special token [MASK]. Next, the BERT model takes the modified sequence as input and obtain the hidden state of each masked speaker, denoted as h_i^{mask}. Then, we predict the speaker through a multi-type classification as

$$\Pr(s_i|s_i^{\text{mask}}) = \text{softmax}(W^s h_i^{\text{mask}} + b^s), \quad (7)$$

where $W^s \in \mathbb{R}^{d_h \times S}$, $b^r \in \mathbb{R}^S$ with S as the maximum number of speakers in dialogues.

[1]We use the golden relation during training.

Model	Dev		Test	
	F1(σ)	**F1**$_c$(σ)	**F1**(σ)	**F1**$_c$(σ)
BERT(Devlin et al., 2019)	60.6 (1.2)	55.4 (0.9)	58.5 (2.0)	53.2 (1.6)
BERTs(Yu et al., 2020)	63.0 (1.5)	57.3 (1.2)	61.2 (0.9)	55.4 (0.9)
GDPNet(Xue et al., 2021)	**67.1 (1.0)**	61.5 (0.8)	64.9 (1.1)	60.1 (0.9)
Ours	66.8 (0.9)	**61.5 (1.0)**	**65.5 (0.7)**	**60.5 (0.8)**

Table 1: Performance comparison of our method with the existing advanced models on DialogRE dataset. σ denotes the standard deviation of 5 runs with different initial random seeds.

3.4 Joint Training Objective

The above three tasks share the BERT encoder and are jointly trained based on the multi-task learning framework. During training, we minimize the following objective loss function as

$$\mathcal{L} = \mathcal{L}_{\text{RE}} + \mathcal{L}_{\text{TP}} + \mathcal{L}_{\text{SP}}, \tag{8}$$

where $\mathcal{L}_{\text{RE}}$ is the binary cross-entropy loss for relation extraction, $\mathcal{L}_{\text{TP}}$ and $\mathcal{L}_{\text{SP}}$ are the cross-entropy loss for trigger words prediction and speaker prediction, respectively. For inference, we directly use the relation predicted by Equation 5 as the final result.

4 Experiments

In this section, we compare the proposed method with the current state-of-the-art approaches to evaluate its effectiveness.

4.1 Experimental Setup

Dataset We conduct experiments on the dialogue-based RE benchmark dataset, **DialogRE** (Yu et al., 2020). It contains $1,788$ dialogues from the transcripts of *Friends* corpus, totally with 36 relation types. 49.6% of relation triples is annotated with trigger words.

Evaluation Metrics Following the previous work (Yu et al., 2020), we adopt F1 score and F1$_c$ score as the evaluation metrics. Among them, F1$_c$ is a supplement to the F1, which only considers the first $i \leq m$ turns of utterances, rather than the entire dialogue.

Baseline Models We compare our method with the existing advanced models, **BERT**(Devlin et al., 2019), **BERTs** (Yu et al., 2020) and **GDPNet** (Xue et al., 2021). **BERT** model for the dialogue-based RE directly applies BERT as the dialogue encoder, and uses the hidden state of `[CLS]` for relation prediction. **BERTs** is a speaker-awared BERT, with modifies the speaker tokens to special markers. **GDPNet** uses a gaussian graph-based network to capture the interaction within dialogues, and achieves the current state-of-the-art performance.

Model	Test	
	F1(σ)	**F1**$_c$(σ)
Ours	**65.5 (0.7)**	**60.5 (0.8)**
Ours w/o SP	65.4(0.6)	59.8 (0.6)
Ours w/o TP	63.5 (0.9)	58.8 (1.0)
Ours w/o SP and TP	63.0 (0.7)	58.0 (0.9)
BERTs	61.2 (0.8)	55.4 (0.9)

Table 2: Ablation study to investigate the influence of each proposed task. SP and TP denote speaker prediction and trigger words prediction, respectively.

4.2 Experimental Results

Main Results Table 1 presents the performance comparison of our method with the existing advanced models. The results show that our method obviously outperforms the previous models and achieve a new state-of-the-art on test set with a F1 score of 65.5% and a F1$_c$ score of 60.5%, demonstrating the effectiveness of the proposed method.

Ablation Study We conduct ablation study experiments to investigate the influence of each proposed task. Table 2 shows the results. We can observe that, 1) *Ours w/o SP*, which removes the speaker prediction task. This causes a performance drop on all metrics, expescially with a drop of 0.7% on F1$_c$ score. 2) *Ours w/o TP*, which eliminates the trigger prediction task. The performance in terms of F1 and F1$_c$ decreases by 2.0% and 1.7%, respectively, demonstrating the importance role of trigger words prediction. 3) *Ours w/o SP and TP*, which detaches both speaker and trigger words prediction tasks. In this case, the performance further drops 0.5% and 0.8% in terms of F1 and F1$_c$. Therefore, the results above indicate that both the two tasks are beneficial to the dialogue-based RE. 4) Note that,

Case 1	**S1:** Mom ! **S2:** Sweetie! So this is where you work? ...
BERTs	(**S1**, `unanswerable`, **S2**) ✗
Ours	(**S1**, `per:parents`, **S2**) ✓
Case 2	**S1:** Hello, Mr. Bing. **S2:** Loved your Stevie Wonder last night. **S3:** Thanks. Listen, about the weekly numbers, I'm gonna need them on my desk by nine o'clock. **S1:** Sure. **S2:** No problem.
BERTs	(**S1**, `unanswerable`, **S3**) ✗ (**S2**, `unanswerable`, **S3**) ✗
Ours	(**S1**, `per:boss`, **S3**) ✓ (**S2**, `per:boss`, **S3**) ✓

Table 3: Case study on the DialogRE test set. The highlighted text indicates the trigger words recognized by our method.

although *Ours w/o SP and TP* only retains the relation extraction task, it still outperforms BERTs by a large margin. The result shows that our improved method of relation extraction is also effective.

Analysis on Discourse Structure Modeling To show the necessity of considering the discourse structure in dialogue-based RE, we design a naive way to degenerate a dialogue into a plain document. We modify the colon after a speaker into text like "said", "responsed" or "continued". For example, the Dialogue 2 in Figure 1 is converted into "*S1 said Mom! S2 responsed Sweetie! So this is where you work? ...*". Then, we apply our method to the changed text. The performance on the test set significantly degrades with 58.0% for F1 and 56.8% for F1_c. The result indicates that dialogues contain important discourse structural information. Therefore, it is important to study the extraction strategies for dialogues rather than directly applying common sentence-level or document-level extraction methods.

Trigger Words Prediction To further evaluate the effect of trigger words prediction task, we calculate the prediction performance on the cases annotated with the ground truth trigger words. Note that, 49.6% of relational triplets have trigger words in DialogRE. The prediction accuracy is 75.6%. The result demonstrates that we can correctly recognize most of the the trigger words. And with the help of the supporting information, the performance of relation extraction is considerably improved, as shown in Table 2.

Case Study We give a case study to analyze the quality of the results produced by our approach and the baseline model. Cases in Table 3 show that our method is capable of capturing the trigger words information and the characteristic of speakers. In case 1, the base model fails to utilize the trigger words information and identifies the relation as `unanswerable`. However, our method correctly recognizes that the word "Mom" triggers the relation between "S1" and "S2", which promotes the right prediction result. Besides, in case 2, our method can capture the characteristics information of speakers and thus correctly predict that "S3" is the boss of "S1" and "S2". Contrarily, the baseline model has difficulties in handling such case.

5 Conclusion

In this paper, we propose to enhance the dialogue-base relation extraction with two benefical tasks, the speaker prediction and the trigger words prediction. Extensive experiments on the benchmark dataset DialogRE demonstrate the effectiveness of our method in achieving state-of-the-art performance in both F1 score and F1_c score.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant Nos.U1636211, 61672081, 61370126), the 2020 Tencet Wechat Rhino-Bird Focused Research Program, and the Fund of the State Key Laboratory of Software Development Environment (Grant No.SKLSDE2019ZX-17). We thank anonymous reviewers for their helpful comments.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186.

Tsu-Jui Fu, Peng-Hsuan Li, and Wei-Yun Ma. 2019. Graphrel: Modeling text as relational graphs for joint entity and relation extraction. In *ACL*, pages 1409–1418.

Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. Scirex: A challenge

dataset for document-level information extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7506–7516.

Yi Luan, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, and Hannaneh Hajishirzi. 2019. A general framework for information extraction using dynamic span graphs. In *NAACL*, pages 3036–3046.

Guoshun Nan, Zhijiang Guo, Ivan Sekulic, and Wei Lu. 2020. Reasoning with latent structure refinement for document-level relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1546–1557.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Highway networks. *arXiv preprint arXiv:1505.00387*.

Difeng Wang, Wei Hu, Ermei Cao, and Weijian Sun. 2020. Global-to-local neural networks for document-level relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3711–3721.

Jue Wang and Wei Lu. 2020. Two are better than one: Joint entity and relation extraction with table-sequence encoders. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1706–1721.

Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. 2020. A novel cascade binary tagging framework for relational triple extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1476–1488.

Fuzhao Xue, Aixin Sun, Hao Zhang, and Eng Siong Chng. 2021. Gdpnet: Refining latent multi-view graph for relation extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. Docred: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777.

Dian Yu, Kai Sun, Claire Cardie, and Dong Yu. 2020. Dialogue-based relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4927–4940.

Tianyang Zhao, Zhao Yan, Yunbo Cao, and Zhoujun Li. 2020. Asking effective and diverse questions: A machine reading comprehension based framework for joint entity-relation extraction. In *IJCAI*.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Proceedings of the AAAI Conference on Artificial Intelligence*.