TREND: Trigger-Enhanced Relation Extraction Network for Dialogues

Po-Wei Lin   Shang-Yu Su   Yun-Nung Chen

National Taiwan University, Taipei, Taiwan

{r09922a24, f05921117}@csie.ntu.edu.tw   y.v.chen@ieee.org

Abstract

The goal of dialogue relation extraction (DRE) is to identify the relation between two entities in a given dialogue. During conversations, speakers may expose their relations to certain entities by explicit or implicit clues, such evidences called "triggers". However, trigger annotations may not be always available for the target data, so it is challenging to leverage such information for enhancing the performance. Therefore, this paper proposes to learn how to identify triggers from the data with trigger annotations and then transfers the trigger-finding capability to other datasets for better performance. The experiments show that the proposed approach is capable of improving relation extraction performance of unseen relations and also demonstrate the transferability of our proposed trigger-finding model across different domains and datasets.[1]

1 Introduction

The goal of relation extraction (RE) is to identify the semantic relation type between two mentioned entities from a given text piece, which is one of basic and important natural language understanding (NLU) problems (Zhang et al., 2017; Zhou and Chen, 2021; Cohen et al., 2020). In this task setting, we are usually given a written sentence and a query pair containing two entities and asked to return the most possible relation type from a predefined set of relations. Dialogue relation extraction (DRE), on the other hand, aims to excavate underlying cross-sentence relation in natural human communications (Yu et al., 2020; Jia et al., 2021). The problem itself is well-motivated, because relations between entities in dialogues could potentially provide dialogue systems with additional features for better dialogue managing (Peng et al., 2018; Su et al., 2018a) or response generation (Su et al., 2018b).

Figure 1: An example of dialogue relation extraction; the dashed arrows connect subjects, triggers, and objects. Triggers are clues of relations annotated in DialogRE.

There are two popular datasets, DialogRE (Yu et al., 2020) and DDRel (Jia et al., 2021), focusing on relation extraction in dialogues illustrated in Figure 1. In DRE, given a conversation and a query pair, we aim to identify the interpersonal relationship between the given entities, where entities can be human or other types like locations. As shown in Figure 1, the evidences of relations within the conversation flow, called **Triggers**, provide informative cues for this task. A trigger can be a short phrase or even a single word with any possible part-of-speech. In the example, the clue for knowing the speaker 2 has a negative impression on the speaker 3 comes from the sentence "*You are arrogant.*" Such hint is intuitively useful for deciding the relations. However, Albalak et al. (2022) is the only prior work that tried to explicitly leverage such signal for improving DRE, because such explanation annotations may not be always available (Kung et al., 2020).

Prior work can be divided into two main lines, one of which is graph-based methods. DHGAT (Chen et al., 2020) presents an attention-based heterogeneous graph network to model multiple types of features; GDPNet (Xue et al., 2021) constructs latent multi-view graphs to model possible relationships among tokens in a long sequence, and then refines the graphs by iterative graph convo-

---

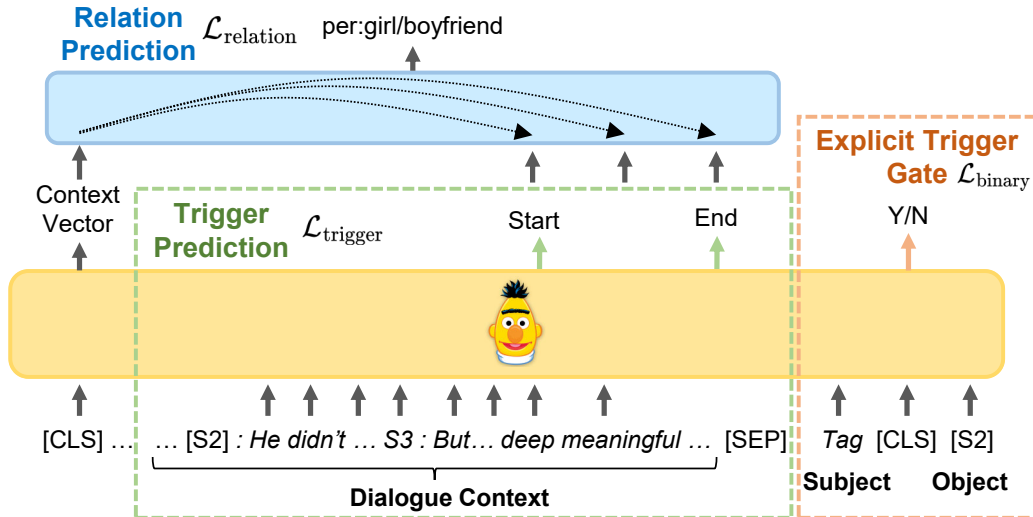[1] The source code is available at: http://github.com/MiuLab/TREND.

Figure 2: The proposed method contains two components: (1) a multi-tasking BERT with two fine-tuning tasks (explicit trigger classification and trigger prediction), and (2) a relation predictor with attentional feature fusion.

lution and pooling techniques. Another branch is BERT-based (Kenton and Toutanova, 2019) methods (Yu et al., 2020; Xue et al., 2022). SimepleRE (Xue et al., 2022) is a simple BERT model with an additional refinement gate for iteratively finding high-confidence prediction. LSR (Nan et al., 2020) is a latent structure refinement method for better reasoning in the document-level relation extraction task. Although it is known that using trigger information can significantly help the performance of relation extraction, only DialogRE has the annotated triggers. It is not guaranteed that utilizing the annotated triggers can generalize to other relations from other datasets, considering the discrepancy of their relation types.

Given the target data without trigger annotations, this paper proposes **TREND**, a simple multi-tasking model with an attentional relation predictor, where it learns the general capability of finding triggers and transfers it to the unseen relations for performance improvement. The experiments show that our proposed method can effectively identify the explicit triggers and generalize to unseen relations towards great flexibility and practicality.

2 Proposed Method

The core idea of this model is to identify trigger spans and accordingly leverage such signal to improve relation extraction. We hereby propose **T**rigger-enhanced **R**elation-**E**xtraction **N**etwork for **D**ialogues, **TREND**, illustrated in Figure 2.

2.1 Problem Formulation

Given a piece of dialogue context $\mathcal{D}$ composed of text tokens $\mathcal{D} = \{x_i\}$ and a query pair q containing a subject entity and an object entity $q = (s, o)$, the task aims at learning a function f that finds the most possible relations between the given entities from a predefined relation set $\mathcal{R}$, $f(\mathcal{D}, q) \to \mathcal{R}$. Note that a single query pair may contain multiple relations, and we duplicate the data samples when they have multiple relation labels by following the prior work.

2.2 TREND

The proposed model has two modules, (1) a multi-tasking BERT (Kenton and Toutanova, 2019) for encoding context and identifying triggers, and (2) a relation predictor with a feature fusion of the dialogue and the automatically identified trigger.

As illustrated in Figure 2, an input $(\mathcal{D}, q)$ will be first augmented into a BERT-style sequence. Specifically, the input format is "`[CLS]` $\mathcal{D}$ `[SEP]` s `[CLS]` o". We replace the target entity pair with their speaker tokens in $\mathcal{D}$ following Yu et al. (2020) illustrated in the figure. The first `[CLS]` encodes the dialogue contexts, and the second one is to predict whether the triggers are explicit via binary classification detailed below.

Explicit Trigger Gate Because triggers sometimes are implicit, it is difficult to identify the associated trigger spans of dialogue relations. We hereby propose to learn a binary classifier as a gate to identify if the explicit triggers exist, and empty trigger spans are inputted to relation prediction

when no explicit triggers. The binary cross entropy loss $\mathcal{L}_{\text{binary}}$ is used here.

Trigger Prediction The explicit triggers are identified by a extractive method with start-end pointer prediction (Kenton and Toutanova, 2019), which is prevalent in extractive question answering (Lee et al., 2016; Rajpurkar et al., 2016). This is a single-label classification problem of predicting the most possible positions; hence a cross entropy loss $\mathcal{L}_{\text{trigger}}$ is conducted.

Relation Prediction A learned context vector and a predicted trigger span are then fed into the relation predictor as depicted in the top part of Figure 2. The features are fused by a generic attention mechanism, where the query is the context vector $\mathbf{c}$, and the keys and the values are trigger words $\mathbf{x_i}$ encoded by BERT: $\sum \text{softmax}(\mathbf{c} \cdot \mathbf{x_i}) \cdot \mathbf{x_i}$. The merged feature is then fed into a 1-layer feed-forward network for final relation prediction using a cross entropy loss $\mathcal{L}_{\text{relation}}$.

Supervised Joint Learning Considering that only DialogRE contains the annotated trigger cues, we perform supervised joint learning for three above tasks. Three above losses are linearly combined as the learning objective for training the whole model in an end-to-end manner. The weights for adjusting the impact of each loss are tuned in the development set. We also apply schedule sampling (Bengio et al., 2015) on explicit trigger classification and trigger prediction when feeding into the relation predictor in order to mitigate the gap between the true triggers and the predicted ones.

Transfer Learning Because annotated triggers may not be available, this paper focuses on transferring the trigger-finding capability to another target dataset, DDRel, which does not contain trigger annotations and its relation types differ a lot from DialogRE. We replace the final feed-forward layer with a new one, since relation numbers may different in two datasets. Then we fine-tune the whole model using a single loss about relation prediction, $\mathcal{L}_{\text{relation}}$, where we assume the trigger-finding capability can be better transferred cross different datasets/relations.

3 Experiments

We focus on evaluating the performance of DRE on the dataset without trigger labels in order to investigate if the trigger-finding capability can be transferred across datasets/relations.

Model	F1
BERT	60.6
GDPNet	64.3
SimpleRE (single entity pair)	60.4
D-REX$_{\text{BERT}}$	59.2
TUCORE-GCN$_{\text{BERT}}$	65.5
TREND$_{\text{BERT-Base}}$	**66.8**
TREND$_{\text{BERT-Large}}$	**67.8**
SimpleRE (multiple entity pairs)	66.7
SocAoG (multiple entity pairs)	*69.1*
TREND$_{\text{BERT-Base}}$ (ground-truth triggers)	75.3

Table 1: The model performance on DialogRE.

3.1 Setting

The DRE datasets used in our experiments are DialogRE (v2) with trigger annotations (Yu et al., 2020) and DDRel (Jia et al., 2021) without trigger annotations. Text normalization like lemmatization and expanding contractions is applied to data pre-processing. In all experiments, we use mini-batch `adam` with a learning rate $3e-5$ as the optimizer on Nvidia Tesla V100. The ratio of teacher forcing and other hyper-parameters are selected by grid search in (0,1] with a step 0.1. The training takes 30 epochs without early stop. The detailed implementation can be found in Appendix A.

The following BERT-based methods are performed for fair comparison: 1) BERT, 2) GDP-Net (Xue et al., 2021), 3) SimpleRE (Xue et al., 2022), 4) D-REX$_{\text{BERT}}$ (Albalak et al., 2022), and 5) TUCORE-GCN$_{\text{BERT}}$ (Lee and Choi, 2021). Other approaches that take multiple entity pairs for global consideration cannot directly be compared with TREND but reported as reference.

3.2 Results of Supervised Joint Learning

The performance of our TREND model jointly trained on the trigger-available DialogRE dataset is presented in Table 1, where it is obvious that our TREND achieves the best performance in the fair setting. Unlike SimpleRE and GDPNet that need to iteratively refine the latent features or latent graphs, relation prediction in the proposed TREND is straight-forward, making training and inference efficient and robust. Furthermore, D-REX (Albalak et al., 2022) also leverages triggers for relation prediction but performs significantly worse than our simple TREND models in the same setting. Our trained binary gate has about 85% accuracy while the trigger prediction has no more than 50% of exact match. Although our model cannot per-

Model	4-class		6-class		13-class	
	Acc	**Macro-F**	**Acc**	**Macro-F**	**Acc**	**Macro-F**
BERT	47.1 / 58.1	44.5 / 52.0	41.9 / 42.3	39.4 / 38.0	39.4 / 39.7	20.4 / 24.1
TUCORE-GCN_BERT	43.8 / 60.3	41.9 / 56.6	36.9 / **52.6**	38.7 / 54.2	29.5 / 44.9	20.5 / **36.9**
TREND_BERT-Base	51.5 / **65.4**	**46.5 / 61.2**	40.3 / **52.6**	**43.0 / 55.0**	**40.5 / 46.2**	21.2 / 34.7
w/o binary gate	52.5 / 53.8	45.3 / 49.7	37.0 / 43.6	41.8 / 45.9	36.6 / 43.6	**26.4** / 36.3
TREND_BERT-Large	**51.6** / 60.3	**46.5** / 54.0	**42.5** / 46.2	**43.0** / 48.2	34.4 / 43.6	19.9 / 36.3
w/o binary gate	41.5 / 47.4	40.3 / 44.9	39.0 / 42.3	43.1 / 42.9	38.5 / 34.6	17.3 / 21.1

Table 2: The DDRel performance in session-level/pair-level settings and different granularity settings (4,6,13-class).

fectly extract the triggers, the predicted spans can still facilitate relation prediction in our proposed TREND. It demonstrates that our TREND model is capable of identifying potential triggers and utilizing such cues for predicting relations. Note that TREND_BERT-Large is for reference, indicating that a larger model has the potential of further improving the performance. The upper-bound of our proposed TREND_BERT-Base is 75.3 shown in the last row of Table 1, where the ground truth triggers are inputted in the relation predictor. This higher score suggests that our TREND model still has a room for improvement and the proposed model design is well validated.

3.3 Results of Transfer Learning

Due to the lack of trigger annotations in DDRel, our TREND model focuses on transferring the trigger-finding capability learned from DialogRE to DDRel. We compare our proposed TREND with two models, which are not designed for transferring across different relation extraction datasets, so they are directly trained on the DDRel data. Table 2 presents the performance achieved on DDRel evaluated in session-level and pair-level settings, where session-level relation extraction is given a *partical* dialogue the entity pair is involved in and pair-level is based on a *full* dialogue (Jia et al., 2021).[2] All scores are much lower than ones in DialogRE due to the higher difficulty of this dataset. The obtained improvement compared with the BERT baseline is larger when the longer dialogue contexts as the input; that is, pair-level improvement is more than session-level one. The probable reason is that extracting key evidences for predicting relations is more important to overcome information overload.

Furthermore, we report the performance of the current state-of-the-art (SOTA) relation extraction

model, TUCORE-GCN, on the DDRel dataset.[3] It can be found that our proposed method can effectively transfer the capability of capturing triggers from DialogRE to DDRel, and outperform TUCORE-GCN in most cases, achieving a new SOTA performance in DDRel.

Suprisingly, TREND_BERT-Large does not outperform TREND_BERT-Base, implying that TREND_BERT-Base already has enough good capability of capturing triggers and can generalize to another dataset (DDRel) and a new relation set.

3.4 Ablation Study

Because our trigger finding module contains a binary classifier deciding the existence of explicit triggers and a trigger predictor extracting trigger spans, we examine the effectiveness of the binary gate. By removing the binary gate, the performance is consistently degraded shown in Table 2, further demonstrating the effectiveness of the designed trigger-finding module in our TREND model.

3.5 Generalization of Unseen Relations

To further investigate if our trigger-finding capability can generalize to different relations, we categorize all relations into seen and unseen relations based on the relation similarity between the two datasets shown in Table 3, and show the session-level performance in Table 4. It can be seen that our proposed TREND is capable of transferring trigger-finding capability from DialogRE to DDRel, even DDRel does not contain trigger annotations. More importantly, our learned trigger-finding capability is demonstrated general to diverse relations, because TREND achieves better results for not only seen but also unseen relations whose triggers never appear in the DialogRE data. We qualitatively analyze the predicted triggers of unseen relations, where TREND extracts a dirty word ("fxxk") and a

---

[2]A session only contains multiple turns in a dialogue, so session-level results are worse than pair-level ones.

[3]The numbers are obtained based on the released code in Lee and Choi (2021).

DDRel Relation	DialogRE Relation
Workplace Superior-Subordinate	per:boss
Workplace Superior-Subordinate	per:subordinate
Friends	per:friends
Lovers	per:girl/boyfriend
Neighbors	per:neighbor
Roommates	per:roommate
Child-Parent	per:children
Child-Other Family Elder	per:other family
Siblings	per:siblings
Spouse	per:spouse
Colleague/Partners	per:works
Courtship	-
Opponents	-
Professional Contact	-

Table 3: Relation ontology mapping between DDRel and DialogRE datasets.

DDRel Relation	Seen	Unseen
BERT	23.77	9.94
TUCORE-GCN	23.39	10.81
TREND	28.30	13.13

Table 4: F1 results of DDRel seen and unseen relations.

word "client" as triggers for unseen relations "opponent" and "professional contact" in DDRel respectively. The full samples can be found in Table 6. It shows the effectiveness and generalizability of our proposed TREND model towards practical usage.

3.6 Qualitative Study

The predicted triggers and relation for DialogRE and DDRel datasets are presented in Table 5 and Table 6 respectively. Note that the triggers are not annotated in DDRel. It can be found that TREND can extract explicit cues as triggers not only for the seen relations, which are similar to relations in DialogRE, but also unseen ones.

4 Conclusion

This paper proposes TREND, a multi-tasking model with the generalizable trigger-finding capability, to improve dialogue relation extraction. TREND is a simple, flexible, end-to-end model based on BERT with three components: (1) an explicit trigger gate for trigger existence, (2) an extractive trigger predictor, and (3) a relation predictor with an attentional feature fusion. The experiments demonstrate that TREND can successfully transfer the learned trigger-finding capability across different datasets and diverse relations for better dialogue relation extraction performance, showing the great potential of improving explainability without rationale annotations.

S1: What's up?
S2: Monica and I are **engaged**.
S1: Oh my God. Congratulations.
S2: Thanks.

Argument	Relation	Trigger
(S2, Monica)	girl/boyfriend	engaged

Table 5: A predicted result of TREND on DialogRE.

S1: That's all.
S2: That's all?!
S1: You don't see it, do you, **father**?"
S2: No. Fellow wants to sell a house ...

Argument	Relation (Seen)	Trigger
(S1, S2)	Child-Parent	father

S1: **Fuck** me!
S2: Want a drink? Okay... I'm not good at this sort of thing, but we don't have a lot of time, so I'll just go ahead and get started.

Argument	Relation (Unseen)	Trigger
(S1, S2)	Opponent	fuck

S1: I'm Joe Galvin, I'm representing Deborah Ann Kaye, case against St.
S2: I told the guy I didn't want to talk to...
S1: I'll just take a minute. Deborah Ann Kaye. You know what I'm talking about.
S2: No.
S1: He's the Assistant Chief of Anesthesiology, Massachusetts Commonwealth. He says your doctors, Towler and Marx, put my girl in the hospital for life. And we can prove that. What we don't know is why. I want someone who was in the O.R.
S2: I've got nothing to say to you.
S1: You know what happened.
S2: Nothing happened.
S1: Then why aren't you testifying for their side? I can subpoena you, you know. I can get you up there on the stand.
S2: And ask me what?
S1: Who put my **client** in the hospital for life.
S2: I didn't do it, Mister.
S1: Who are you protecting, then?
S2: Who says that I'm protecting anyone?
S1: I do. Who is it? The Doctors. What do you owe them?
S2: I don't owe them a goddamn thing.
S1: Then why don't you testify?
S2: You know, you're pushy, fella...
S1: You think I'm pushy now, wait 'til I get you on the stand...
S2: Well, maybe you better do that, then.

Argument	Relation (Unseen)	Trigger
(S1, S2)	Professional Contact	client

Table 6: Predicted results of TREND on DDRel.

627

References

Alon Albalak, Varun Embar, Yi-Lin Tuan, Lise Getoor, and William Yang Wang. 2022. D-REX: Dialogue relation extraction with explanations. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 34–46.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 1171–1179.

Hui Chen, Pengfei Hong, Wei Han, Navonil Majumder, and Soujanya Poria. 2020. Dialogue relation extraction with document-level heterogeneous graph attention networks. *arXiv preprint arXiv:2009.05092*.

Amir DN Cohen, Shachar Rosenman, and Yoav Goldberg. 2020. Relation classification as two-way span-prediction. *arXiv preprint arXiv:2010.04829*.

Qi Jia, Hongru Huang, and Kenny Q Zhu. 2021. Ddrel: A new dataset for interpersonal relation classification in dyadic dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13125–13133.

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

Po-Nien Kung, Tse-Hsuan Yang, Yi-Cheng Chen, Sheng-Siang Yin, and Yun-Nung Chen. 2020. Zero-shot rationalization by multi-task transfer learning from question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2187–2197.

Bongseok Lee and Yong Suk Choi. 2021. Graph based network with contextualized representations of turns in dialogue. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 443–455.

Kenton Lee, Shimi Salant, Tom Kwiatkowski, Ankur Parikh, Dipanjan Das, and Jonathan Berant. 2016. Learning recurrent span representations for extractive question answering. *arXiv preprint arXiv:1611.01436*.

Guoshun Nan, Zhijiang Guo, Ivan Sekulić, 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.

Baolin Peng, Xiujun Li, Jianfeng Gao, Jingjing Liu, Kam-Fai Wong, and Shang-Yu Su. 2018. Deep dyna-q: Integrating planning for task-completion dialogue policy learning. *arXiv preprint arXiv:1801.06176*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Shang-Yu Su, Xiujun Li, Jianfeng Gao, Jingjing Liu, and Yun-Nung Chen. 2018a. Discriminative deep dyna-q: Robust planning for dialogue policy learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3813–3823.

Shang-Yu Su, Kai-Ling Lo, Yi-Ting Yeh, and Yun-Nung Chen. 2018b. Natural language generation by hierarchical decoding with linguistic patterns. In *Proceedings of The 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

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*, volume 35, pages 14194–14202.

Fuzhao Xue, Aixin Sun, Hao Zhang, Jinjie Ni, and Eng-Siong Chng. 2022. An embarrassingly simple model for dialogue relation extraction. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6707–6711. IEEE.

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.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45.

Wenxuan Zhou and Muhao Chen. 2021. An improved baseline for sentence-level relation extraction. *arXiv preprint arXiv:2102.01373*.

A Reproducibility

A.1 Hyperparameters

All the hyper-parameters were selected by grid search in (0,1] with step 0.1. The loss functions are linearly combined and each of them has an adjustable weight.

TREND$_{\text{BERT-Base}}$
- Loss: $0.3 \cdot \mathcal{L}_{\text{trigger}} + 1.0 \cdot \mathcal{L}_{\text{relation}} + 1.0 \cdot \mathcal{L}_{\text{binary}}$
- schedule sampling: 0.7 for trigger prediction, 0.7 for binary classification

TREND$_{\text{BERT-Large}}$
- Loss: $0.3 \cdot \mathcal{L}_{\text{trigger}} + 1.0 \cdot \mathcal{L}_{\text{relation}} + 1.0 \cdot \mathcal{L}_{\text{binary}}$
- schedule sampling: 0.5 for trigger prediction, 0.7 for binary classification

Data	Training	Inference
DialogRE	15 mins×30	5 mins
DDRel (session-level)	15 mins×30	5 mins
DDRel (pair-level)	1.5 mins×30	10 secs

Table 7: Time efficiency on three sets of experiments.

A.2 Time Efficiency

The training and inference cost in terms of time is reported in Table 7.