An Embarrassingly Simple Model for Dialogue Relation Extraction

Fuzhao Xue[1], **Aixin Sun**[1], **Hao Zhang**[1,2], **Eng Siong Chng**[1]
[1] School of Computer Science and Engineering, Nanyang Technological University, Singapore
[2] Institute of High Performance Computing, A*STAR, Singapore
fuzhao001@e.ntu.edu.sg, axsun@ntu.edu.sg,
hao007@e.ntu.edu.sg, aseschng@ntu.edu.sg

Abstract

Dialogue relation extraction (RE) is to predict the relation type of two entities mentioned in a dialogue. In this paper, we model Dialogue RE as a multi-label classification task and propose a simple yet effective model named SimpleRE. SimpleRE captures the interrelations among multiple relations in a dialogue through a novel input format, BERT Relation Token Sequence (BRS). In BRS, multiple [CLS] tokens are used to capture different relations between different pairs of entities. A Relation Refinement Gate (RRG) is designed to extract relation-specific semantic representation adaptively. Experiments on DialogRE show that SimpleRE achieves the best performance with much shorter training time. SimpleRE outperforms all direct baselines on sentence-level RE without using external resources.

1 Introduction

Relation Extraction (RE) is to identify the semantic relation type between two mentioned entities in a given piece of text, *e.g.,* a sentence or dialogue. As shown in Table 1, given a pair of entities (*i.e.,* an argument pair) "Monica" and "S2", the RE task is to predict their relation type, from a set of predefined relations. Many existing studies formulate Dialogue RE as a typical classification task. Researchers have tried to improve Dialogue RE by considering speaker information (Yu et al., 2020) and trigger tokens (Xue et al., 2020). There are also solutions based on graph attention network, where a graph is constructed to model speaker, entity, entity-type, and utterance nodes (Chen et al., 2020). However, Transformer-based models remain strong competitors and achieve the best $F1$ measure (Xue et al., 2020; Yu et al., 2020).

Multiple pairs of entities could be mentioned in a dialogue, and their relations are interrelated to some extent. For instance, based on the first few utterances in Table 1, "Richard" and 'Monica" have two possible relations, *i.e.,* "positive_impression" or "girl/boyfriend". However, the last utterance indicates that "Monica" is girlfriend of "S2"; hence "Richard" and 'Monica" can only be related by "positive_impression". We argue that such interrelationships could be helpful for relation extraction.

S1:	Where the hell have you been?!	
S2:	I was making a coconut phone with the professor.	
S1:	Richard told Monica he wants to marry her!	
S2:	What?!	
S1:	Yeah! Yeah, I've been trying to find ya to tell to stop messing with her and maybe I would have if these damn boat shoes wouldn't keep flying off!	
S2:	My—Oh my God!	
S1:	I know! They suck!!	
S2:	He's not supposed to ask my girlfriend to marry him! I'm supposed to do that!	
	Argument pair	**Relation type**
R1	(Monica, S2)	girl/boyfriend
R2	(Richard, Monica)	positive_impression

Table 1: An example from DialogRE dataset (Yu et al., 2020).

In this paper, we propose SimpleRE, an extremely simple model to reason and learn the interrelations among tokens and relations. Because of the strong semantic modeling capability, BERT becomes the natural choice to model such interrelationships. We formulate Dialogue RE as a multi-label classification task and design a BERT Relation Token Sequence (BRS). BRS contains multiple "[CLS]" tokens as the input to BERT-based model, with the aim to capture relations between multiple pairs of entities. We then propose a Relation Refinement Gate (RRG) to refine the semantic representation of each relation in an adaptive manner, for target relation prediction.

On DialogRE dataset, SimpleRE achieves best $F1$ by a large margin over two BERT-based methods, BERTs (Yu et al., 2020) and GDPNet (Xue et al., 2020). As a simple model, the training speed of SimpleRE is at least 5 times faster than these

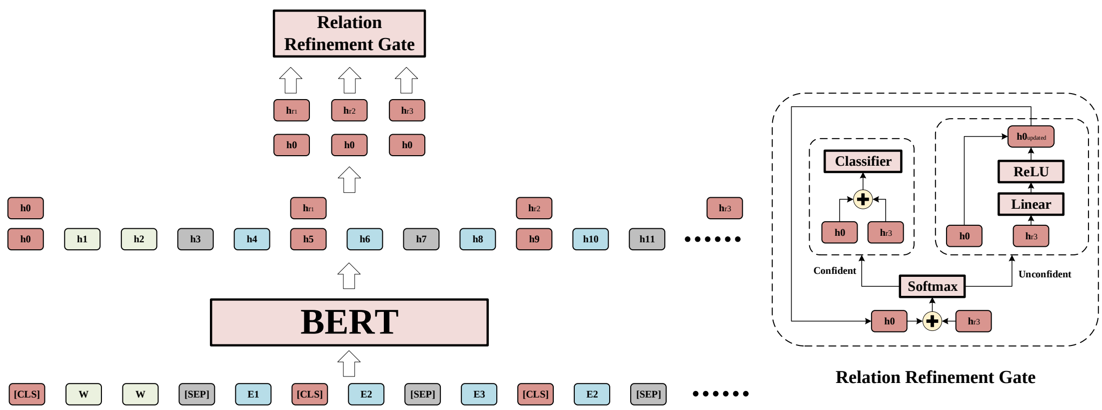

Figure 1: The overall architecture of SimpleRE.

two models. We also show that BRS is effective on sentence-level RE, and the adapted SimpleRE beats all direct baselines on TACRED dataset.

2 SimpleRE

The overall architecture of SimpleRE is shown in Figure 1. Its novelty is two-fold: the input format to BERT, *i.e.,* BERT Relation Token Sequence (BRS), and the way to utilize BERT encodings through Relation Refinement Gate (RRG).

2.1 Problem Formulation

Let $\mathcal{R}$ be the set of predefined relation types. Let $X = \{x_1, x_2, \ldots, x_T\}$ be a text sequence with T tokens, where x_t is the token at t^{th} position. X denotes an entire dialogue for Dialogue RE, and a single sentence for sentence-level RE. There could be multiple relations $R = \{r_1, r_2, \ldots, r_n\}$ between n pairs of entities mentioned in X. The i^{th} relation $r_i \in \mathcal{R}$ is predicted based on an argument pair: subject entity E_s^i and object entity E_o^i.

2.2 BERT Relation Token Sequence

BERT (Devlin et al., 2019) based models are powerful in modeling semantics in text sequences, and have achieved outstanding performance on various tasks (van Aken et al., 2019; Xu et al., 2020; Su et al., 2020). In SimpleRE, we adopt BERT to model the interrelations among all possible relations in a text sequence, through BRS.

Given a sequence X with T tokens, a set of subject entities $E_s = \{E_s^1, E_s^2, \ldots, E_s^n\}$ and a set of object entities $E_o = \{E_o^1, E_o^2, \ldots, E_o^n\}$, we form a BRS as input to BERT: $BRS = \big\langle$[CLS], X, [SEP], E_s^1, [CLS], E_o^1, [SEP], $\ldots$, [SEP], E_s^n, [CLS], E_o^n, [SEP]$\big\rangle$. [CLS] and [SEP] are classification and separator tokens, respectively. The [CLS] tokens at different positions in the BRS input may carry different meanings, due to the different contexts.

Multiple [CLS] tokens have been used to learn hierarchical representations of a document, where one [CLS] is put in front of a sentence (Liu, 2019; Chen et al., 2019). In BRS, multiple [CLS] tokens are for different relations between entity pairs, and also their interrelations, because these multiple [CLS] tokens are in the same input sequence.

2.3 Relation Refinement Gate

In BRS, representation of the first [CLS] token (denoted by h_0) encodes the semantic information of entire sequence. Representations of the subsequent [CLS] tokens capture the relations between each pair of entities. We denote the i^{th} relation representation as h_{ri}. To predict the relation type of r_i, in Relation Refinement Gate, we concatenate semantic representations of h_0 and h_{ri} as $c_i = [h_0; h_{ri}]$. We then use Shallow-Deep Networks (Kaya et al., 2019) to compute a confidence score:

$$s_c = \max\Big(\text{Softmax}\big(f(c_i)\big)\Big) \tag{1}$$

Here f denotes a single layer feed-forward neural network (FFN). If the confidence score s_c is larger than a predefined threshold τ, c_i is used to predict the target relation between E_s^i and E_o^i, by a classifier.[1] Otherwise, we refine h_0 to be more relation-specific to h_{ri}, since h_0 is weakly related to the target relation (Xue et al., 2020). To this end, we define a refinement mechanism to extract a task-specific semantic information by updating

[1]We use a linear layer as a classifier for Dialogue RE and a linear layer with softmax for sentence-level RE.

Model	English V1 ($F1 \pm \delta$)	English V2 ($F1 \pm \delta$)	Chinese ($F1 \pm \delta$)
CNN (Yu et al., 2020)	48.0±1.5	-	-
LSTM (Yu et al., 2020)	47.4±0.6	-	-
BiLSTM (Yu et al., 2020)	48.6±1.0	-	-
AGGCN (Guo et al., 2019)	46.2	-	-
LSR (Nan et al., 2020)	44.4	-	-
DHGAT (Chen et al., 2020)	56.1	-	-
BERT (Devlin et al., 2019)	58.5±2.0	60.6±0.5*	61.6±0.4*
BERTs (Yu et al., 2020)	61.2±0.9	61.8±0.6*	63.8±0.6*
GDPNet (Xue et al., 2020)	64.9±1.1	64.3±1.1*	62.2±0.9*
SimpleRE (Ours)	**66.3**±0.7	**66.7**±0.7	**65.2**±1.1

Table 2: Comparison of SoTA methods on DialogRE. We report 5-run averaged $F1$ and the standard deviation (δ). * denotes results produced by running author released codes.

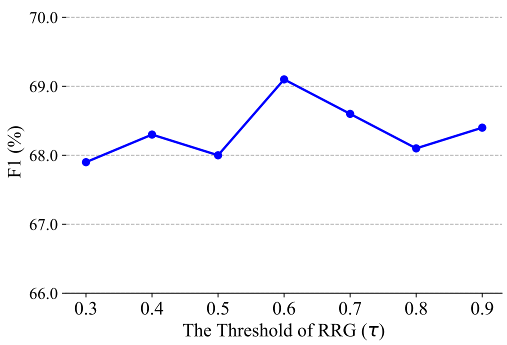

Figure 2: The performance of different thresholds τ.

h_0 for the prediction of r_i:

$$h_0' = \text{ReLU}\big(g(h_{ri})\big) + h_0 \qquad (2)$$

Here g is a single layer FFN and h_0' denotes the updated semantic representation. Then h_0' is used to predict the relation or updated further, depending on the recomputed s_c. Note h_0' updating is specific to each target relation prediction of r_i.

3 Experiments

We conduct experiments on Dialogue RE and sentence-level RE tasks to evaluate SimpleRE against baseline models.

3.1 Dataset

DialogRE is the first human-annotated Dialogue RE dataset (Yu et al., 2020) originated from the transcripts of American TV situation comedy, Friends. It contains $1,788$ dialogues and 36 predefined relation types. An example of this dataset is given in Table 1. Recently, Yu et al. (2020) releases a modified English version and a Chinese version of DialogRE. We evaluate SimpleRE on all three versions, to show its effectiveness and efficiency.

Model	Average Time (mins)
BERT (Devlin et al., 2019)	4.7
BERTs (Yu et al., 2020)	4.7
GDPNet (Xue et al., 2020)	12.6
SimpleRE (Ours)	0.9

Table 3: Efficiency comparison with existing models on Dialogue RE, by average training time per epoch (minutes).

TACRED is a widely-used sentence-level RE dataset. It contains more than 106K sentences drawn from the yearly TACKBP4 challenge, and has 42 different relations (including a special “no relation” type). We evaluate SimpleRE on both TACRED and TACRED-Revisit (TACREV) datasets; TACREV is a modified version of TACRED.

3.2 Experimental Settings

We compare SimpleRE with two recent BERT-based methods, BERTs (Yu et al., 2020) and GDPNet (Xue et al., 2020). We also include popular baselines AGGCN (Guo et al., 2019), LSR (Nan et al., 2020), and DHGAT (Chen et al., 2020) in our experiments. For a fair comparison with BERTs and GDPNet, we utilize the same hyperparameter settings, except for batch size. Specifically, we set batch size to 6 rather than 24 for SimpleRE because it predicts multiple relations (*i.e.,* all relations annotated in one dialogue) per forward process. To set threshold τ, we conduct a preliminary study on the development set with different τ values. Reported in Figure 2, SimpleRE achieves best performance when $\tau \approx 0.6$. Thus, we set $\tau = 0.6$ throughout the experiments, unless specified otherwise.

Model	$F1 \pm \sigma$
SimpleRE	**66.3**±0.7
SimpleRE w/o BRS	60.4±0.9
SimpleRE w/o RRG	65.5±0.7

Table 4: Ablation studies of SimpleRE.

3.3 Results on DialogRE

Performance by $F1$. Table 2 summarizes the results on DialogRE. Observed that BERT-based models significantly outperform non-BERT models. Among the three BERT-based models, SimpleRE surpasses GDPNet and BERTs by 1.4% and 5.1% respectively, on English V1 DialogRE, in terms of $F1$ measure. SimpleRE outperforms both models on English V2 and Chinese datasets as well.

Efficiency by Training Time. We now study the efficiency of SimpleRE by evaluating its average training time per epoch. The results are reported in Table 3. Observe that SimpleRE is 5 times faster than baselines despite its smaller batch size. Compared to the baselines, SimpleRE can predict multiple relations per step and its simple structure leads to better efficiency than baselines, *e.g.,* GDPNet with SoftDTW (Cuturi and Blondel, 2017).

Ablation Study We conduct ablation studies on DialogRE for the effectiveness of the proposed components, BERT Relation Token Sequence (BRS) and Relation Refinement Gate (RRG). To evaluate their impact on performance, we remove BRS and RRG from the model separately. To remove BRS, we modify the input format to predict one relation each time with a modified input format: $\langle$[CLS], X, [SEP], E_s, [CLS], E_o, [SEP] $\rangle$. To remove RRG, all relations are predicted based on the corresponding token representations $[h_0; h_{ri}]$, without updating h_0.

As reported in Table 4, the results show that removing BRS leads to large performance degradation, which indicates interrelations among relations have a significant impact on RE performance. Meanwhile, RRG module also contributes to the performance gains.

3.4 Results on TACRED

We now adapt SimpleRE to sentence-level RE. Becuase each sentence only contains a single relation in the sentence-level RE dataset, BRS becomes $\langle$[CLS], X, [SEP], E_s, [CLS], E_o, [SEP] $\rangle$. The representations of the two [CLS] tokens are concatenated for relation prediction. Compared to typical RE input sequence $\langle$[CLS], X, [SEP], E_s, [SEP], E_o, [SEP] $\rangle$, SimpleRE replaces the [SEP] token between two entities with a [CLS] token. RRG is not applicable here because there is only one relation in each sentence *i.e.,* it is not necessary to further refine h_0 to be target relation specific.

Model	TACRED	TACREV
LSTM (Zhang et al., 2017)	62.7	70.6
PA-LSTM (Zhang et al., 2017)	65.1	74.3
C-AGGCN (Guo et al., 2019)	68.2	75.5
LST-AGCN (Sun et al., 2020)	68.8	-
SpanBERT (Joshi et al., 2020)	70.8	78.0
GDPNet (Xue et al., 2020)	70.5	80.2
SimpleRE (Ours)	**71.7**	**80.7**
KnowBERT (Peters et al., 2019)	71.5	79.3

Table 5: $F1$ of all models on TACRED and TACRED-Revisit (TARREV).

Model	$F1$
SimpleRE	**71.7**
SimpleRE w/o 2nd [CLS] token	70.6
SimpleRE w/o relation representation h_r	70.0
SimpleRE w/o semantic representation h_0	70.6

Table 6: Ablation study of SimpleRE. For the model without 2nd [CLS] token, we use [SEP] token. Instead of using $[h_0 : h_r]$, we have evaluated SimpleRE with either h_0 or h_r alone for relation prediction.

For fair comparison, we refer Xue et al. (2020) to use SpanBERT as the backbone model (*i.e.,* BERT in Figure 1). The results on TACRED and TACREV are summarized in Table 5. Observed that SimpleRE outperforms all compared baselines including KnowBERT (Peters et al., 2019), and the latter incorporates an external knowledge base during training.

Table 6 summarizes the results of ablation studies on TACRED. We first replace the second [CLS] token with a [SEP] token, which leads to 1.1% performance degradation. This observation suggests that [CLS] is necessary to capture the relation between entities near it. Moreover, the performance of our model further drops without relation representation, *i.e.,* predicting relation purely based on h_0 instead of $[h_0 : h_r]$. Poorer performance is also observed when only h_r is used for prediction.

4 Conclusion

In this paper, we propose a simple yet effective model named SimpleRE for dialogue relation extraction. SimpleRE is designed to learn and reason the interrelations among multiple relations in a dialogue. The most important component of SimpleRE is the BERT Relation Token Sequence, where multiple [CLS] tokens are used to capture relations between entity pairs. The Relation Refinement Gate is designed to further improve the semantic representation in an adaptive manner. The SimpleRE can also be easily adapted to sentence-level relation extraction. On both datasets, DialogRE and TACRED, we show that our embarrassingly simple model is a strong competitor for relation extraction tasks. Expanding SimpleRE to other relation extraction tasks like document-level relation extraction and few-shot relation extraction is part of our future work.

References

Betty van Aken, Benjamin Winter, Alexander Löser, and Felix A. Gers. 2019. How does bert answer questions? a layer-wise analysis of transformer representations. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, page 1823–1832. ACM.

Deli Chen, Shuming Ma, Keiko Harimoto, Ruihan Bao, Qi Su, and Xu Sun. 2019. Group, extract and aggregate: Summarizing a large amount of finance news for forex movement prediction. In *Proceedings of the Second Workshop on Economics and Natural Language Processing*, pages 41–50, Hong Kong. ACL.

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*.

Marco Cuturi and Mathieu Blondel. 2017. Soft-DTW: a differentiable loss function for time-series. In *Proceedings of the 34th International Conference on Machine Learning*, pages 894–903. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. ACL.

Zhijiang Guo, Yan Zhang, and Wei Lu. 2019. Attention guided graph convolutional networks for relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 241–251, Florence, Italy. ACl.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Yigitcan Kaya, Sanghyun Hong, and Tudor Dumitras. 2019. Shallow-deep networks: Understanding and mitigating network overthinking. In *International Conference on Machine Learning*, pages 3301–3310. PMLR.

Yang Liu. 2019. Fine-tune bert for extractive summarization. *arXiv preprint arXiv:1903.10318*.

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, Online. ACL.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word 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)*, pages 43–54, Hong Kong, China. ACL.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

Kai Sun, Richong Zhang, Yongyi Mao, Samuel Mensah, and Xudong Liu. 2020. Relation extraction with convolutional network over learnable syntax-transport graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8928–8935.

Weidi Xu, Xingyi Cheng, Kunlong Chen, and Taifeng Wang. 2020. Symmetric regularization based bert for pair-wise semantic reasoning. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 1901–1904. ACM.

Fuzhao Xue, Aixin Sun, Hao Zhang, and Eng Siong Chng. 2020. Gdpnet: Refining latent multi-view graph for relation extraction. *arXiv preprint arXiv:2012.06780*.

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, Online. ACL.

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, Copenhagen, Denmark. ACL.