AN EMBARRASSINGLY SIMPLE MODEL FOR DIALOGUE RELATION EXTRACTION

Fuzhao Xue[1], *Aixin Sun*[1], *Hao Zhang*[1,2], *Jinjie Ni*[1], *Eng-Siong Chng*[1]

[1]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[2]Centre for Frontier AI Research, Agency for Science, Technology and Research, Singapore

ABSTRACT

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

***Index Terms*—** Dialogue Relation Extraction, Multi-Relations, BERT

1. INTRODUCTION

Relation extraction (RE) is to identify the semantic relation type between two entities mentioned in a piece of text, *e.g.,* a sentence or a dialogue. Table 1 shows an example dialogue. The RE task is to predict the relation type of a pair of entities like "Monica" and "S2" (*i.e.,* an argument pair) mentioned in the dialogue, from a set of predefined relations. Researchers have tried to improve Dialogue RE by considering speaker information [1] or trigger tokens [2]. There are also solutions based on graph attention network, where a graph models speaker, entity, entity-type, and utterance nodes [3]. However, Transformer-based models remain strong competitors [1, 2].

A dialogue may mention multiple pairs of entities, reflected by annotations in the DialogRE dataset [1] (see Table 1). Among multiple pairs of entities, the relations mentioned in the same dialog often interrelate with each other to some extent. An example is shown in Table 1, "Richard" and "Monica" in the first few utterances show two possible relations, *i.e.,* "positive_impression" or "girl/boyfriend". 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.

Table 1. An example from DialogRE dataset [1]. Relations of two pairs of entities are annotated.

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!

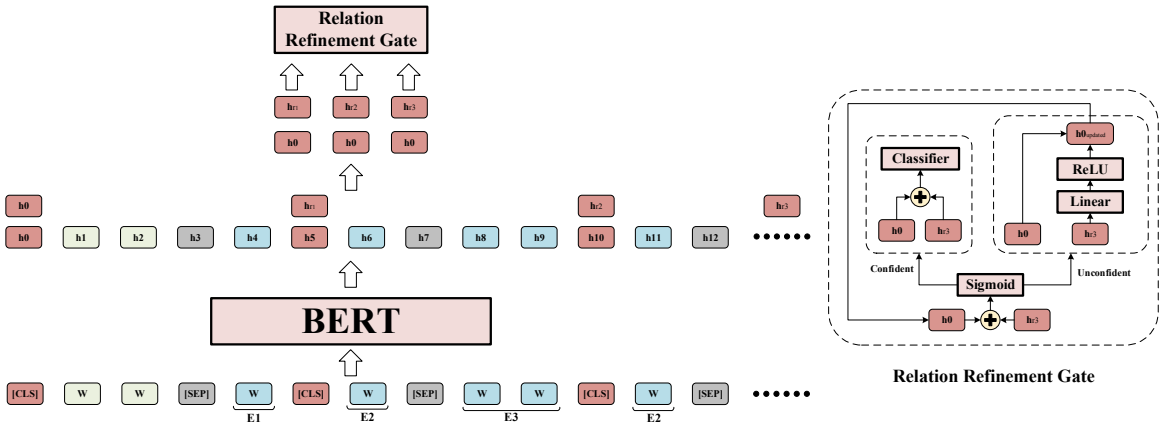
	Argument pair	**Relation type**
R1	(Monica, S2)	girl/boyfriend
R2	(Richard, Monica)	positive_impression

In this paper, we propose SimpleRE, an extremely simple model, to reason and learn interrelations among tokens and relations. SimpleRE is built on top of BERT. Due to its strong modeling capability, BERT is the natural choice to model such interrelationships. We first design a BERT Relation Token Sequence (BRS). BRS contains multiple "[CLS]" tokens in input sequence, 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 for target relation prediction in an adaptive manner.

On the DialogRE dataset, SimpleRE achieves best $F1$ over two BERT-based methods, BERTs [1] and GDPNet [2], by a large margin. As a simple model, the training of SimpleRE is at least 5 times faster than these two models. We also show that BRS is effective on sentence-level RE, and the adapted SimpleRE beats all direct baselines on the TACRED dataset.

This research is supported by the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funding Scheme (Project #A19E2b0098) and the National Research Foundation Singapore under its AI Singapore Programme (Award Number: AISG-100E-2018-006).

Fig. 1. The architecture of SimpleRE. An entity may contain one or more tokens as illustrated.



2. SimpleRE

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

2.1. Problem formulation

Let $\mathcal{R}$ be a 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, or a single sentence for sentence-level RE. Between n pairs of entities mentioned in X, there could be multiple relations $R = \{r_1, r_2, \ldots, r_n\}$. The i-th relation $r_i \in \mathcal{R}$ is predicted for an argument pair: *subject entity* E_s^i and *object entity* E_o^i. Note that, an entity may contain one or more tokens. In this problem setting, the pairs of entities whose relations are to be predicted are known.

2.2. BERT Relation Token Sequence

BERT [4] based models are powerful in modeling semantics in text sequences [5, 6, 7]. In SimpleRE, we adopt BERT to model the interrelations among all possible relations in a text sequence, through BRS.

Given a sequence X, which contains 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 the 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 [8, 9]. In BRS, multiple [CLS] tokens are for capturing different relations between entity pairs and 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 [10] to compute a confidence score:

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

Here f denotes a single layer feed-forward neural network (FFN). If 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 [2]. To this end, we define a refinement mechanism to extract task-specific semantic information by updating h_0 for the prediction of r_i:

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

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. To avoid possible endless refinement, we set an upper bound B to limit the maximum number of iterations for refining h_0'. $B = 3$ in our experiments.

3. EXPERIMENTS

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

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

Table 2. Comparison with baselines on DialogRE. Results are 5-run averaged $F1$ with standard deviation (δ).

Model	$F1 \pm \delta$
CNN [1]	48.0±1.5
LSTM [1]	47.4±0.6
BiLSTM [1]	48.6±1.0
AGGCN [11]	46.2
LSR [12]	44.4
DHGAT [3]	56.1
BERT [4]	58.5±2.0
BERTs [1]	61.2±0.9
GDPNet [2]	64.9±1.1
SimpleRE (Ours)	**66.3**±0.7

Table 3. Comparison with baselines on new versions of DialogRE. Results are 5-run averaged $F1$ with standard deviation.

Model	English V2 ($F1 \pm \delta$)	Chinese ($F1 \pm \delta$)
BERT [4]	60.6±0.5	61.6±0.4
BERTs [1]	61.8±0.6	63.8±0.6
GDPNet [2]	64.3±1.1	62.2±0.9
SimpleRE (Ours)	**66.7**±0.7	**65.2**±1.1

3.1. Dataset

DialogRE is the first human-annotated Dialogue RE dataset [1] originated from the transcripts of American comedy "Friends". It contains $1,788$ dialogues and 36 predefined relation types (see example in Table 1). Recently, [1] released a modified English version and a Chinese version of DialogRE. We evaluate SimpleRE on all three versions.

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 [1] and GDPNet [2]. We also include popular baselines AGGCN [11], LSR [12], and DHGAT [3] 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 Table 4, SimpleRE achieves best performance when $\tau \approx 0.6$. Thus, we set $\tau = 0.6$ throughout the experiments, unless specified otherwise. We set the maximal refinement iterations B as 3 for Dialogue RE.

Table 4. Performance with different threshold τ values.

τ	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$F1$	67.9	68.3	68.0	**69.1**	68.6	68.1	68.4

Table 5. Average training time (in minutes) per epoch on Dialogue RE

Model	Average Time (mins)
BERT [4]	4.7
BERTs [1]	4.7
GDPNet [2]	12.6
SimpleRE (Ours)	0.9

3.3. Results on DialogRE

3.3.1. Performance by $F1$.

Table 2 summarizes the results on DialogRE. Observe 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 DialogRE, by $F1$ measure.

Note that, the two modified DialogRE datasets, English V2 and Chinese version, are released recently. Since most existing models have not reported their performance, we obtain results by running author released codes of existing BERT-based models. Table 3 shows that SimpleRE achieves better performance than baselines on both versions of DialogRE.

3.3.2. Efficiency by training time.

The average training time per epoch is reported in Table 5, after training for 20 epochs. SimpleRE is about $5\times$ faster than baselines, despite its smaller batch size. Note a dialogue in Dialogue RE may contain multiple relations. Existing models only infer one relation per forward process. On the contrary, SimpleRE predicts multiple relations per forward process. Its simple structure leads to better efficiency than baselines, *e.g.,* GDPNet with SoftDTW [13].

3.3.3. Ablation study.

We conduct ablation studies on DialogRE, for the effectiveness of components in SimpleRE: BERT Relation Token Sequence (BRS) and Relation Refinement Gate (RRG). To evaluate their impacts, we remove BRS and RRG from our model separately. To remove BRS, we change 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

Table 6. Ablation study of SimpleRE on DialogRE.

Model	$F1 \pm \sigma$
SimpleRE	**66.3**±0.7
SimpleRE w/o BRS	60.4±0.9
SimpleRE w/ BRS-v2	62.8±1.1
SimpleRE w/ BRS-v3	63.5±0.8
SimpleRE w/o RRG	65.5±0.7

Table 7. $F1$ of all models on TACRED and TACRED-Revisit (TARREV), the sentence-level RE datasets.

Model	TACRED	TACREV
LSTM [14]	62.7	70.6
PA-LSTM [14]	65.1	74.3
C-AGGCN [11]	68.2	75.5
LST-AGCN [15]	68.8	-
SpanBERT [16]	70.8	78.0
GDPNet [2]	70.5	80.2
SimpleRE (Ours)	**71.7**	**80.7**
KnowBERT [17]	71.5	79.3

are predicted based on the corresponding token representations $[h_0; h_{ri}]$, without updating h_0. Besides, we also design two alternative BRS, *i.e.,* BRS-v2 and BRS-v3, for comparison. For BRS-v2, we exchange the [CLS] tokens between entity pairs with [SEP] tokens before subject entities. Similarly, we exchange the [CLS] tokens with [SEP] tokens after object entities in BRS-v3.

Reported in Table 6, the results show that removing BRS leads to large performance degradation, indicating interrelations among relations have a significant impact on RE performance. Meanwhile, RRG module also contributes to the performance gains. Another interesting finding is that BRS-v2 and -v3 cannot model the relations well although they both use multiple [CLS] tokens in the input sequence. The $F1$ score decreases from 66.3 to 62.8 and 63.5, respectively. This result further shows that [CLS] token in BRS is sensitive to its position in the sequence.

3.4. Results on TACRED

We now adapt SimpleRE to sentence-level RE. Note that SimpleRE was not evaluated on document-level RE due to the difference in problem settings. We leave the adaptation of SimpleRE to document-level RE as our future work.

Because each sentence only contains a single relation in 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. Hence, it is unnecessary to refine h_0 to be target relation specific.

Table 8. Ablation study of SimpleRE on TACRED. 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.

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

For fair comparison, we refer [2] to use SpanBERT as the backbone (*i.e.,* BERT in Figure 1). Results on TACRED and TACREV are summarized in Table 7. SimpleRE outperforms all baselines including KnowBERT [17] on both datasets. Note that KnowBERT incorporates external knowledge base during training.

Table 8 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 result suggests that [CLS] is necessary to capture the relation between entities near it. Moreover, the performance of our model further drops without relation representation h_r, *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. Hence, both h_0 and h_r contribute to the correct prediction of relations.

In short, through experiments on both dialogue and sentence-level RE tasks, we show SimpleRE is a strong competitor. Although simple, both components, BRS and RRG, are essential in SimpleRE's model design.

4. CONCLUSION

In this paper, we propose a simple yet effective model for dialogue relation extraction. Building on top of the powerful modeling capability of BERT, SimpleRE is designed to learn and reason the interrelations among multiple relations in a dialogue. The most important component in 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. Through experiments and ablation studies, we show that both components contribute to the success of SimpleRE. Due to its simple structure and fast training speed, we believe SimpleRE serves a good baseline in Dialogue RE task. The SimpleRE can also be easily adapted to sentence-level relation extraction. On datasets for both tasks, DialogRE and TACRED, we show that our simple model is a strong competitor for relation extraction tasks.

5. REFERENCES

[1] Dian Yu, Kai Sun, Claire Cardie, and Dong Yu, "Dialogue-based relation extraction," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, July 2020, pp. 4927–4940, ACL.

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

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

[4] 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, ACL.

[5] Betty van Aken, Benjamin Winter, Alexander Löser, and Felix A. Gers, "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*. 2019, p. 1823–1832, ACM.

[6] Weidi Xu, Xingyi Cheng, Kunlong Chen, and Taifeng Wang, "Symmetric regularization based bert for pairwise semantic reasoning," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2020, p. 1901–1904, ACM.

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

[8] Yang Liu, "Fine-tune bert for extractive summarization," *arXiv preprint arXiv:1903.10318*, 2019.

[9] Deli Chen, Shuming Ma, Keiko Harimoto, Ruihan Bao, Qi Su, and Xu Sun, "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*, Hong Kong, Nov. 2019, pp. 41–50, ACL.

[10] Yigitcan Kaya, Sanghyun Hong, and Tudor Dumitras, "Shallow-deep networks: Understanding and mitigating network overthinking," in *International Conference on Machine Learning*. PMLR, 2019, pp. 3301–3310.

[11] Zhijiang Guo, Yan Zhang, and Wei Lu, "Attention guided graph convolutional networks for relation extraction," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, July 2019, pp. 241–251, ACl.

[12] Guoshun Nan, Zhijiang Guo, Ivan Sekulic, and Wei Lu, "Reasoning with latent structure refinement for document-level relation extraction," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, July 2020, pp. 1546–1557, ACL.

[13] Marco Cuturi and Mathieu Blondel, "Soft-DTW: a differentiable loss function for time-series," in *Proceedings of the 34th International Conference on Machine Learning*, Doina Precup and Yee Whye Teh, Eds. 2017, pp. 894–903, PMLR.

[14] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning, "Position-aware attention and supervised data improve slot filling," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, Sept. 2017, pp. 35–45, ACL.

[15] Kai Sun, Richong Zhang, Yongyi Mao, Samuel Mensah, and Xudong Liu, "Relation extraction with convolutional network over learnable syntax-transport graph," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Apr. 2020, vol. 34, pp. 8928–8935.

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

[17] Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith, "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)*, Hong Kong, China, Nov. 2019, pp. 43–54, ACL.