EIDER: Empowering Document-level Relation Extraction with Efficient Evidence Extraction and Inference-stage Fusion

Yiqing Xie, Jiaming Shen, Sha Li, Yuning Mao, Jiawei Han
University of Illinois at Urbana-Champaign, IL, USA
{xyiqing2, js2, shal2, yuningm2, hanj}@illinois.edu

Abstract

Document-level relation extraction (DocRE) aims to extract semantic relations among entity pairs in a document. Typical DocRE methods blindly take the full document as input, while a subset of the sentences in the document, noted as the evidence, are often sufficient for humans to predict the relation of an entity pair. In this paper, we propose an evidence-enhanced framework, EIDER, that empowers DocRE by efficiently extracting evidence and effectively fusing the extracted evidence in inference.[1] We first jointly train an RE model with a lightweight evidence extraction model, which is efficient in both memory and runtime. Empirically, even training the evidence model on silver labels constructed by our heuristic rules can lead to better RE performance. We further design a simple yet effective inference process that makes RE predictions on both extracted evidence and the full document, then fuses the predictions through a blending layer. This allows EIDER to focus on important sentences while still having access to the complete information in the document. Extensive experiments show that EIDER outperforms state-of-the-art methods on three benchmark datasets (e.g., by 1.37/1.26 Ign F1/F1 on DocRED).

1 Introduction

Relation extraction (RE) is the task of extracting semantic relations among entities within a given text, which has abundant applications such as knowledge graph construction, question answering, and biomedical text analysis (Yu et al., 2017; Shi et al., 2019; Trisedya et al., 2019). Prior studies mostly focus on predicting the relation between two entity mentions in a single sentence. However, in reality, an entity may have multiple mentions throughout a document. It is also common that a relation can only be inferred given multiple sentences as the context. As a result, recent studies have been moving towards the more realistic setting of document-level relation extraction (DocRE) (Peng et al., 2017; Yao et al., 2019; Zeng et al., 2020).

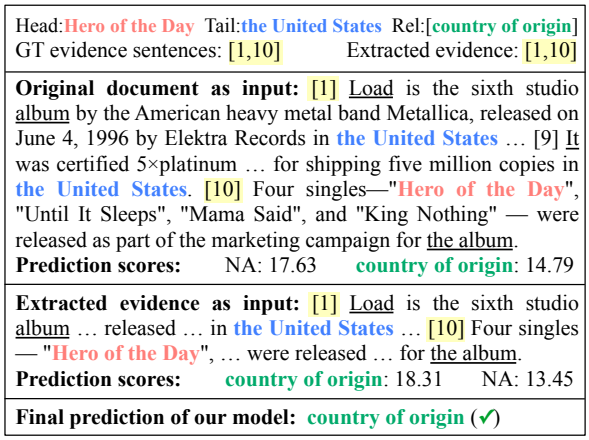

Figure 1: A test sample in the DocRED dataset (Yao et al., 2019), where the i^{th} sentence in the document is marked with [i] at the start. Our model correctly predicts [1,10] as evidence, and if we only use the extracted evidence as input, the model can predict the relation "country of origin" correctly.

Unlike typical DocRE models that blindly take the whole document as input, a human may only need a few sentences to infer the relation of an entity pair. For each entity pair, we define the minimal set of sentences required by human annotators to infer their relation as their *evidence sentences*. As shown in Figure 1, to predict the relation between *"Hero of the Day"* and *"the United States"*, it is sufficient to know that *Load (the album)* was released in *the United States* from the 1^{st} sentence, and *"Hero of the Day"* is a single of *Load* from the 10^{th} sentence. In other words, the 1^{st} and 10^{th} sentences serve as the evidence to infer this relation. Although the 9^{th} sentence also mentions *"the United States"*, it is irrelevant to this specific relation. Including such irrelevant sentences in input might sometimes introduce noise to the model and be more detrimental than beneficial.

[1]Our code is available at https://github.com/Veronicium/Eider

Findings of the Association for Computational Linguistics: ACL 2022, pages 257 - 268
May 22-27, 2022

Despite the usefulness of evidence, few prior studies leverage it in a proper way (Huang et al., 2021a,b). In particular, Huang et al. (2021a) extracts the evidence sentences together with RE but does not utilize them after extraction. Besides, it requires human-annotated evidence for training, and also suffers from massive memory usage and training time. Another work (Huang et al., 2021b) trains an RE model solely on evidence sentences, which misses important information in the original document and fails to show improvements when paired up with pre-trained language models.

In this paper, we propose an **ev**i**d**ence-**e**nhanced Doc**R**E framework EIDER, which efficiently extracts evidence and effectively leverages the extracted evidence to improve DocRE. During training, we enhance DocRE by jointly extracting relations and evidence using multi-task learning, which allows the two tasks to benefit from providing additional training signals for each other. There are two major challenges regarding evidence extraction. The first challenge is the memory and runtime overhead due to training an additional task. For example, a prior multi-task method (Huang et al., 2021a) needs over 14h and three consumer GPUs to train, while the individual RE model only takes around 90min on one GPU. In comparison, EIDER uses a simpler evidence extraction model, which can fit into a single GPU and only requires 95min runtime. The second challenge is that human-annotated evidence sentences are costly and heavily relying on them limits model applicability. Therefore, we design several heuristic rules to construct silver labels in case the evidence annotation is unavailable. We observe that EIDER still improves RE performance when trained with our silver labels, and sometimes even performs on par with using gold labels.

With the evidence extracted, either by our rules or evidence extraction model, we propose to further enhance DocRE by utilizing the evidence in inference. In the extreme case, if there is only one sentence related to the relation, one can make predictions solely based on this sentence and reduce the problem to sentence-level relation extraction. One naive approach is thus to directly replace the original document with the extracted evidence (Huang et al., 2021b). However, since no systems can extract evidence perfectly, solely relying on extracted sentences may miss important information and harm model performance in certain cases (see Table 5). To avoid information loss, we fuse the prediction results of the original document and extracted evidence through a blending layer (Wolpert, 1992). In this way, EIDER pays more attention to the extracted important sentences, while still having access to all the information in the document. Empirical analysis demonstrates that removing either source would lead to degenerate performance.

We conduct extensive experiments on three widely-adopted DocRE benchmarks: DocRED (Yao et al., 2019), CDR (Li et al., 2016) and GDA (Wu et al., 2019). Experiment results show that EIDER achieves state-of-the-art performance on all the datasets. Performance analysis further shows that the improvement of EIDER is most significant on inter-sentence entity pairs, suggesting that leveraging evidence is especially effective in reasoning over multiple sentences. In particular, EIDER significantly improves the performance on entity pairs that require co-reference/multi-hop reasoning by 1.98/2.08 F1 on DocRED, respectively.

Contributions. (1) We propose an efficient joint relation and evidence extraction model that allows the two tasks to mutually enhance each other without heavily relying on evidence annotation. (2) We design a simple and effective DocRE inference process enhanced by the extracted evidence, enabling more focus on the important sentences with no information loss. (3) We demonstrate that our evidence-enhanced framework outperforms state-of-the-art methods on three DocRE datasets.

2 Problem Formulation

Given a document d comprised of N sentences $\{s_n\}_{n=1}^{N}$, L tokens $\{h_l\}_{l=1}^{L}$, E named entities $\{e_i\}_{i=1}^{E}$ and all the proper-noun mentions of each entity, $\{m_j^i\}$, the task of document-level relation extraction (DocRE) is to predict the set of all possible relations between all entity pairs (e_h, e_t) from a pre-defined relation set $\mathcal{R} \bigcup \{\text{NA}\}$. We refer to e_h and e_t as the head entity and tail entity, respectively. A relation r belongs to the positive class $\mathcal{P}_{h,t}^{T}$ if it exists between (e_h, e_t) and otherwise the negative class $\mathcal{N}_{h,t}^{T}$. For each entity pair (e_h, e_t) that possesses a non-NA relation, we define its *evidence*[2] $V_{h,t} = \{s_{v_k}\}_{k=1}^{K}$ as the subset of sentences in the document that are sufficient for human annotators to infer the relation. Human annotation of evidence may or may not be given in training, depending on the datasets, but is not available in inference.

[2]We use "*evidence sentence*" and "*evidence*" interchangeably throughout the paper.

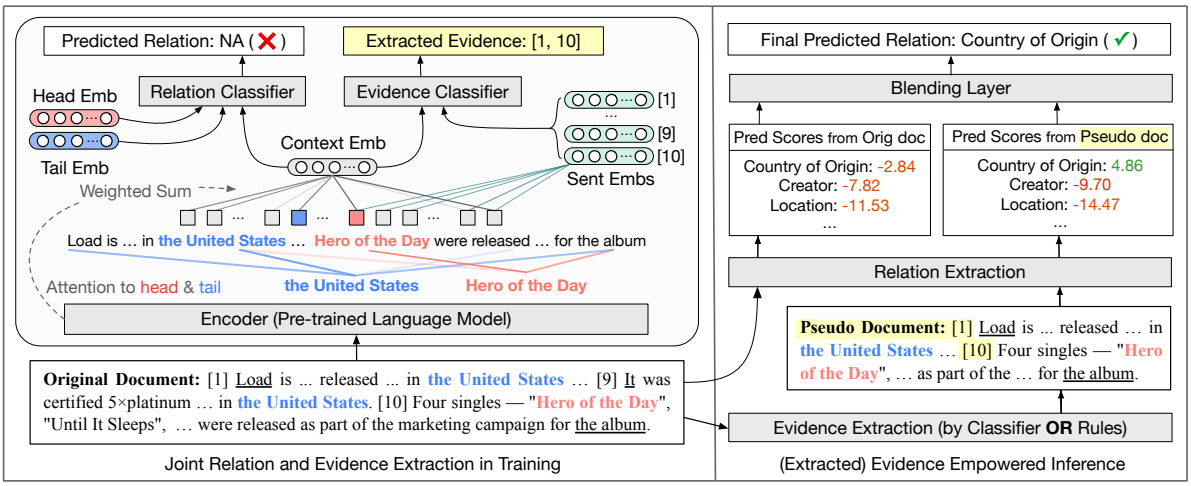

Figure 2: The overall architecture of EIDER. The left part illustrates the training stage and the right shows the inference stages of EIDER. We highlight **head entities**, **tail entities** and extracted evidences.

3 Methodology

An illustration of the framework of EIDER is shown in Figure 2. In training, we jointly extract relation and evidence using multi-task learning, where the two tasks have their own classifier and share the base encoder (Sec. 3.1). In inference, we fuse the predictions on the original document and the extracted evidence using a blending layer (Sec. 3.2). In case the evidence annotation is not available, we also provide several heuristic rules to construct silver evidence labels as an alternative (Sec. 3.3).

3.1 Joint Relation and Evidence Extraction

In our framework, we jointly train the relation extraction model with an evidence extraction model using multi-task learning. As shown in Figure 2, the two tasks have their own classifier but share the base encoder. Intuitively, tokens relevant to predicting the relation are essential in both models. By sharing the base encoder, the two tasks can provide additional training signals for each other and hence mutually enhance each other (Ruder, 2017).

Base Encoder. We leverage pre-trained language models (Devlin et al., 2019) to encode the semantic meanings of each token in the document. Specifically, given a document $d = [h_l]_{l=1}^{L}$, we insert a special token "*" before and after each entity mention $\{m_j^i\}$ and leverage the encoder to obtain the s-dim token embeddings $\boldsymbol{H} = [\mathbf{h}_1, ..., \mathbf{h}_L], \mathbf{h}_l \in \mathbb{R}^s$ and the cross token attention $\boldsymbol{A} \in \mathbb{R}^{L \times L}$:

$$\boldsymbol{H}, \boldsymbol{A} = \text{Encoder}([h_1, ..., h_L]), \tag{1}$$

where $\boldsymbol{A}$ is the average of the attention heads in the last transformer layer (Vaswani et al., 2017). For each mention of an entity e_i, we use the embedding of the start symbol "*" as its mention embedding $\mathbf{m_j^i}$. Then, we obtain the embedding of entity e_i by adopting LogSumExp pooling (Jia et al., 2019; Zhou et al., 2021) over the embeddings of all its mentions: $\mathbf{e}_i = \log \sum_j \exp(\mathbf{m_j^i})$.

To predict the relation of different entity pairs, a model may need to focus on different parts of the context. To capture the context relevant to each entity pair (e_h, e_t), we compute its context embedding $\mathbf{c}_{h,t} \in \mathbb{R}^s$ based on the attention matrix $\boldsymbol{A}$ from the pre-trained encoder (Zhou et al., 2021):

$$\mathbf{c}_{h,t} = \boldsymbol{H}^T \frac{\boldsymbol{A}_h \circ \boldsymbol{A}_t}{\boldsymbol{A}_h^T \boldsymbol{A}_t}, \tag{2}$$

where $\circ$ is the Hadamard product and $\boldsymbol{A}_h \in \mathbb{R}^L$ is e_h's attention to all the tokens in the document, obtained by averaging e_h's mention-level attention. Similarly for $\boldsymbol{A}_t$. The intuition is that tokens with high attention towards both e_h and e_t are important to both entities. Hence, these tokens are likely to be essential to the relation and should contribute more to the context embedding.

Relation Classifier. To predict the relation between an entity pair (e_h, e_t), we first compute their context-aware representations $(\mathbf{z_h}, \mathbf{z_t})$ by combining their entity embeddings $(\mathbf{e}_h, \mathbf{e}_t)$ with their context embedding $\mathbf{c}_{h,t}$ and then utilize a bilinear function to calculate the logit of how likely a relation $r \in \mathcal{R}$ exists between e_h and e_t:

$$\begin{aligned} \boldsymbol{z}_h &= \tanh\left(\boldsymbol{W}_h \mathbf{e}_h + \boldsymbol{W}_{c_h} \mathbf{c}_{h,t}\right), \\ \boldsymbol{z}_t &= \tanh\left(\boldsymbol{W}_t \mathbf{e}_t + \boldsymbol{W}_{c_t} \mathbf{c}_{h,t}\right), \\ \mathbf{y}_r &= \boldsymbol{z}_h \boldsymbol{W}_r \boldsymbol{z}_t + \boldsymbol{b}_r, \end{aligned} \tag{3}$$

where $\boldsymbol{W}_h, \boldsymbol{W}_t, \boldsymbol{W}_{c_h}, \boldsymbol{W}_{c_t}, \boldsymbol{W}_r$ and $\boldsymbol{b}_r$ are learnable parameters. As the model may have different confidence for different entity pairs, we apply the adaptive-thresholding loss (Zhou et al., 2021), which learns a dummy relation class TH that serves as the dynamic threshold for each entity pair:

$$\mathbf{y}_{\mathrm{TH}} = \boldsymbol{z}_h \boldsymbol{W}_{\mathrm{TH}} \boldsymbol{z}_t + \boldsymbol{b}_r. \tag{4}$$

During inference, for each tuple $(e_h, e_t, r), r \in \mathcal{R}$, we obtain the prediction score: $S^{(O)}_{h,t,r} = \mathbf{y}_r - \mathbf{y}_{TH}$. Finally, we define our training objective for relation extraction as follows:

$$\mathcal{L}_{RE} = -\sum_{h \neq t} \sum_{r \in \mathcal{P}^T_{h,t}} \log \left(\frac{\exp(\mathbf{y}_r)}{\sum_{r' \in \mathcal{P}^T_{h,t} \cup \{\mathrm{TH}\}} \exp(\mathbf{y}_{r'})} \right) - \log \left(\frac{\exp(\mathbf{y}_{\mathrm{TH}})}{\sum_{r' \in \mathcal{N}^T_{h,t} \cup \{\mathrm{TH}\}} \exp(\mathbf{y}_{r'})} \right). \tag{5}$$

Evidence Classifier. In addition to the relation, we also predict whether each sentence s_n is an evidence sentence of entity pair (e_h, e_t). Similar to entity embeddings, to obtain sentence embedding $\mathbf{s}_n$, we apply a LogSumExp pooling over all the tokens in s_n: $\mathbf{s}_n = \log \sum_{h_l \in s_n} \exp(\mathbf{h}_l)$. Intuitively, if s_n is an evidence sentence of (e_h, e_t), the tokens in s_n would be relevant to the relation prediction, and should contribute more to $\mathbf{c}_{h,t}$. Hence, we use a bilinear function between context embedding $\mathbf{c}_{h,t}$ and sentence embedding $\mathbf{s}_n$ to measure the importance of sentence s_n to entity pair (e_h, e_t):

$$\mathrm{P}(s_n | e_h, e_t) = \sigma(\mathbf{s}_n \boldsymbol{W}_v \mathbf{c}_{h,t} + \boldsymbol{b}_v), \tag{6}$$

where $\boldsymbol{W}_v$ and $\boldsymbol{b}_v$ are learnable parameters.

As an entity pair may have more than one evidence sentence, we use the binary cross entropy as the objective to train the evidence extraction model.

$$\mathcal{L}_{Evi} = -\sum_{h \neq t, \mathrm{NA} \notin \mathcal{P}^T_{h,t}} \sum_{s_n \in \mathcal{D}} y_n \cdot \mathrm{P}(s_n | e_h, e_t) + (1 - y_n) \cdot \log(1 - \mathrm{P}(s_n | e_h, e_t)), \tag{7}$$

where the evidence label y_n is 1 when $s_n \in V_{h,t}$ and otherwise 0. If golden labels are not provided, we use several heuristic rules to construct silver labels instead. Details are introduced in Sec 3.3.

Finally, we optimize our model by the combination of the relation extraction loss $\mathcal{L}_{RE}$ and evidence extraction loss $\mathcal{L}_{Evi}$:

$$\mathcal{L} = \mathcal{L}_{RE} + \mathcal{L}_{Evi}. \tag{8}$$

Efficiency Considerations. Compared to a previous method E2GRE (Huang et al., 2021a) that also extracts the evidence, EIDER is significantly more efficient in both memory and training time for two reasons. First, E2GRE learns $|\mathcal{R}|$ representations for each sentence. Namely, it makes evidence prediction for every (entity, entity, sentence, relation) tuple, which requires expensive computation especially when $|\mathcal{R}|$ is large (e.g., $|\mathcal{R}| = 96$ in DocRED). In contrast, we observe that most entity pairs only have one set of evidence across relations and thus predict only one set of evidence for each entity pair.

Second, E2GRE regards the evidence label of entity pairs with $r = \mathrm{NA}$ as an empty set. However, these entity pairs may still involve some relation beyond the pre-defined relation set $\mathcal{R}$, which also have their evidence sentences. Hence, we train the evidence extraction model only on entity pairs with at least one non-NA relation, which accounts for a small subset (e.g., 2.97% in DocRED) of all the entity pairs. Experiments show that EIDER achieves better performances than E2GRE in both RE and evidence extraction while requiring only 30% of its memory usage and 11% of its runtime.

Furthermore, E2GRE does not utilize the evidence after extraction and relies heavily on the human annotation of evidence, which we will address in the following sections.

3.2 Fusion of Evidence in Inference

Suppose the extracted evidence sentences already contain all the information relevant to the relation, then there is no need to use the whole document for relation extraction. However, no system can perfectly extract the evidence without missing any sentences. Solely relying on the extracted evidence may miss important information in the document and lead to sub-optimal performance. Therefore, we combine the prediction results on both the original document and the extracted evidence, which can either be learned by our evidence classifier (Sec. 3.1) or constructed by our heuristic rules (Sec. 3.3) if evidence annotation is unavailable. Even without joint training, one may directly improve general (trained) DocRE models by applying our proposed inference process (noted as EIDER (Rule)-Nojoint in Table 5).

Specifically, as shown in Figure 2, we first obtain a set of relation prediction scores $S^{(O)}_{h,t,r}$ from the original documents. Then we construct a pseudo document $d'_{h,t}$ for each entity pair by concatenating

the extracted evidence sentences $V'_{h,t}$ in the order they present in the original document. The prediction score of the RE model on the pseudo document is noted as $S^{(E)}_{h,t,r}$. Finally, we fuse the results by aggregating the two sets of prediction sores through a blending layer (Wolpert, 1992):

$$\mathrm{P}_{Fuse}(r|e_h, e_t) = \sigma(S^{(O)}_{h,t,r} + S^{(E)}_{h,t,r} - \tau). \quad (9)$$

We choose this design because it is simple and only includes one learnable parameter, τ, alleviating over-fitting in the development set. We optimize the parameter τ on the development set as follows:

$$\mathcal{L}_{Fuse} = -\sum_{d\in\mathcal{D}}\sum_{h\neq t}\sum_{r\in\mathcal{R}} y_r \cdot \mathrm{P}_{Fuse}(r|e_h, e_t) + (1 - y_r) \cdot \log(1 - \mathrm{P}_{Fuse}(r|e_h, e_t)), \quad (10)$$

where $y_r = 1$ if relation r holds between (e_h, e_t) and $y_r = 0$ otherwise. Empirically, using other loss functions does not affect the performance much.

3.3 Heuristic Evidence Label Construction

In case that human annotation of evidence is not available, we design a set of heuristic rules to automatically construct silver labels for evidence extraction. Then we train our joint model on the silver labels and directly use the silver labels as pseudo documents in inference. The percentage of test samples covered by each rule is shown in Table 6.

Co-occur. If the head and tail entities co-occur in the same sentence (e.g., "Load" and "the United States" co-occur in the 1^{st} sentence in Figure 2), we use all the sentences they co-occur as evidence.

Coref. If the proper-noun mentions of the head and tail entity do not co-occur, but their coreferential mentions co-occur (e.g., "Hero of the Day" and "the album", the co-reference of "Load" co-occur in the 10^{th} sentence in Figure 2), we use all the sentences where their coreferential mentions co-occur as evidence. In practice, we directly apply a pre-trained coreference resolution model, HOI (Xu and Choi, 2020), without fine-tuning on our dataset.

Bridge. If the first two conditions are not met, but there exists a third bridge entity whose coreferential mention co-occurs with both head and tail (e.g., "Load" or its coreferential mention "the album" co-occurs with both "the United States" and "Hero of the Day" in Figure 2), we take all the sentences where the bridge co-occurs with head or tail as the evidence. If there are more than one bridge entities, we choose the one with the highest frequency. While this rule can be easily extended to multiple bridges, we empirically observe that capturing one bridge already leads to satisfying results.

4 Experiments

4.1 Experiment Setup

Datasets. We evaluate the effectiveness of EIDER on three datasets: DocRED (Yao et al., 2019), CDR (Li et al., 2016) and GDA (Wu et al., 2019), where DocRED is the only dataset that provides evidence labels as part of the annotation. The details of the datasets are listed in Appendix A.1.

Implementation Details. Our model is implemented based on PyTorch and Huggingface's Transformers (Wolf et al., 2019). We use cased-$\text{BERT}_{\text{base}}$ (Devlin et al., 2019) and $\text{RoBERTa}_{\text{large}}$ as the base encoders and optimize our model using AdamW with learning rate 5e-5 for the encoder and $1e-4$ for other parameters. We adopt a linear warmup for the first 6% steps. The batch size (number of documents per batch) is set to 4 and the ratio between relation extraction and evidence extraction losses is set to 0.1. We perform early stopping based on the F1 score on the development set, with a maximum of 30 epochs. Our $\text{BERT}_{\text{base}}$ models are trained with one GTX 1080 Ti GPU and $\text{RoBERTa}_{\text{large}}$ models with one RTX A6000 GPU.

Evaluation Metrics. Following prior studies (Yao et al., 2019), we use **F1** and **Ign F1** as the main evaluation metrics for relation extraction, where **Ign F1** measures the F1 score excluding the relations shared by the training and development/test set. We also report **Intra F1** and **Inter F1**, where the former measures the performance on the co-occurred (intra-sentence) entity pairs and the latter evaluates the inter-sentence entity pairs where none of their proper-noun mentions co-occurs. For evidence extraction, we compute the F1 score (denoted as **Evi F1**) and further introduce **PosEvi F1**, which measures the F1 score of evidence only on positive entity pairs (i.e., those with non-NA relations).

4.2 Main Results

We compare our methods with both *Graph-based methods* and *transformer-based methods*. Graph-based methods explicitly perform inference on document-level graphs. Transformer-based methods, including EIDER, implicitly capture the long-distance token dependencies via transformers. Noted that EIDER is trained on gold labels and

Model	Dev				Test	
	Ign F1	F1	Intra F1	Inter F1	Ign F1	F1
LSR-BERT$_{base}$ (Nan et al., 2020)	52.43	59.00	65.26	52.05	56.97	59.05
GLRE-BERT$_{base}$ (Wang et al., 2020)	-	-	-	-	55.40	57.40
Reconstruct-BERT$_{base}$ (Xu et al., 2021)	58.13	60.18	-	-	57.12	59.45
GAIN-BERT$_{base}$ (Zeng et al., 2020)	59.14	61.22	67.10	53.90	59.00	61.24
BERT$_{base}$ (Wang et al., 2019)	-	54.16	61.61	47.15	-	53.20
BERT-Two-Step (Wang et al., 2019)	-	54.42	61.80	47.28	-	53.92
HIN-BERT$_{base}$ (Tang et al., 2020)	54.29	56.31	-	-	53.70	55.60
E2GRE-BERT$_{base}$ (Huang et al., 2021a)	55.22	58.72	-	-	-	-
CorefBERT$_{base}$ (Ye et al., 2020)	55.32	57.51	-	-	54.54	56.96
ATLOP-BERT$_{base}$ (Zhou et al., 2021)	59.11 $\pm$ 0.14†	61.01 $\pm$ 0.10†	67.26 $\pm$ 0.15†	53.20 $\pm$ 0.19†	59.31	61.30
EIDER (Rule)-BERT$_{base}$	60.36 $\pm$ 0.13	62.34 $\pm$ 0.08	68.40 $\pm$ 0.14	54.79 $\pm$ 0.13	60.23	62.21
EIDER-BERT$_{base}$	**60.51 $\pm$ 0.11**	**62.48 $\pm$ 0.13**	**68.47 $\pm$ 0.08**	**55.21 $\pm$ 0.21**	**60.42**	**62.47**
RoBERTa$_{large}$ (Ye et al., 2020)	57.14	59.22	-	-	57.51	59.62
CorefRoBERTa$_{large}$ (Ye et al., 2020)	57.35	59.43	-	-	57.90	60.25
E2GRE-RoBERTa$_{large}$ (Huang et al., 2021a)	59.55	62.91	-	-	60.29	62.51
GAIN-BERT$_{large}$ (Zeng et al., 2020)	60.87	63.09	-	-	60.31	62.76
ATLOP-RoBERTa$_{large}$ (Zhou et al., 2021)	61.30 $\pm$ 0.22†	63.15 $\pm$ 0.21†	69.61 $\pm$ 0.25†	55.01 $\pm$ 0.18†	61.39	63.40
EIDER (Rule)-RoBERTa$_{large}$	61.73 $\pm$ 0.07	63.91 $\pm$ 0.07	69.99 $\pm$ 0.09	56.27 $\pm$ 0.11	61.93	64.12
EIDER-RoBERTa$_{large}$	**62.34 $\pm$ 0.14**	**64.27 $\pm$ 0.10**	**70.36 $\pm$ 0.07**	**56.53 $\pm$ 0.15**	**62.85**	**64.79**

Table 1: Relation extraction results on DocRED. We report the mean and standard deviation on the development set by conducting 5 runs with different random seeds. We report the official test score of the best checkpoint on the development set. Results with † are based on our implementation. Others are reported in their original papers. We separate graph-based and transformer-based methods into two groups.

Model	CDR	GDA
LSR-BERT$_{base}$ (Nan et al., 2020)	64.8	82.2
SciBERT$_{base}$ (Zhou et al., 2021)	65.1 $\pm$ 0.6	82.5 $\pm$ 0.3
DHG-BERT$_{base}$ (Zhang et al., 2020b)	65.9	83.1
GLRE-SciBERT$_{base}$ (Wang et al., 2020)	68.5	-
ATLOP-SciBERT$_{base}$ (Zhou et al., 2021)	69.4 $\pm$ 1.1	83.9 $\pm$ 0.2
EIDER (Rule)-SciBERT$_{base}$	**70.63** $\pm$ 0.49	**84.54** $\pm$ 0.22

Table 2: Relation extraction results on CDR and GDA.

Model	Dev Evi F1	Test Evi F1
E2GRE-BERT$_{base}$	47.14	48.35
EIDER-BERT$_{base}$	**50.71**	**51.27**
E2GRE-RoBERTa$_{large}$	51.11	50.50
EIDER-RoBERTa$_{large}$	**52.54**	**53.01**

Table 3: Evidence extraction results on DocRED. We compare EIDER with E2GRE (Huang et al., 2021a).

leverages the evidence extracted by our model in inference. EIDER (Rule) is trained on silver evidence labels constructed by rules and also leverages them in inference.

Relation Extraction Results. Tables 1 and 2 show that EIDER outperforms the DocRE baseline methods in all datasets. Our improvement is especially large on Inter F1 (e.g., 1.21/2.01 Intra/Inter F1 compared to ATLOP-BERT$_{base}$). We hypothesize that the bottleneck of inter-sentence pairs is to locate the relevant context, which often spreads through the whole document. EIDER learns to capture important sentences in training and focuses more on these important sentences in inference.

Among the baselines, the Inter F1 of GAIN is 0.70 higher than ATLOP while the Intra F1 of ATLOP is 0.16 higher than GAIN, indicating that document-level graphs may be effective in multi-hop reasoning. Although EIDER does not involve explicit multi-hop reasoning modules, it still notably outperforms graph-based models in Inter F1.

Finally, EIDER (Rule) also outperforms all the baselines in both DocRED and the two biomedical datasets which do not have evidence annotation. The improvement on DocRED and CDR is much larger than that on GDA. We hypothesize that it is because more than 85% relations in GDA are intra-sentence ones, making it trivial even for the single RE model to focus on these sentences.

Evidence Extraction Results. To our knowledge, E2GRE is the only method that has reported their evidence extraction result. The results in Table 3 indicate that EIDER outperforms E2GRE significantly (e.g., by 3.57 Dev Evi F1 under BERT$_{base}$). The results show that it may be sufficient to train the evidence classifier only on pairs with $r \in \mathcal{R}$

	Rules (ours)	EIDER-BERT$_{base}$	NoJoint
PosEvi F1	77.43	**80.33**	51.13

Table 4: Ablation study for evidence extraction.

Ablation	Ign F1	F1	Intra F1	Inter F1
EIDER-BERT$_{base}$	**60.51**	**62.48**	**68.47**	**55.21**
NoJoint	59.98	62.03	68.51	54.10
NoPseudo	59.70	61.53	67.55	54.01
NoOrigDoc	58.47	60.44	66.24	53.23
NoBlending	58.93	61.46	67.33	54.37
FinetuneOnEvi	60.11	62.29	68.13	54.84
EIDER (Rule)-BERT$_{base}$	**60.36**	**62.34**	**68.40**	**54.79**
NoJoint	60.01	62.09	68.21	54.34

Table 5: Ablation study of EIDER on DocRED.

and over each (entity, entity, sentence) tuple instead of (entity, entity, sentence, relation) as in E2GRE.

Our ablation studies in Table 4 show that our three heuristic rules, denoted as **Rules (ours)**, already capture most of the evidence for positive entity pairs. The high quality of silver labels explains why our model can perform well using silver labels only. Furthermore, training the RE model and evidence extraction model separately (denoted as **NoJoint**) results in a sharp performance drop. As the relation and evidence classifiers share the same base encoder, discarding the relation classifier will result in insufficient training of the base encoder and harm the performance.

4.3 Performance Analysis

Ablation Study. Table 5 shows the ablation studies that analyzes the utility of each module in EIDER. We observe that **NoJoint** leads to sharp performance drop in DocRE. Besides, **EIDER (Rule)-Nojoint** achieves significant "free gains" (0.90/1.08 Ign F1/F1) by simply fusing the evidence constructed by rules in the inference of ATLOP. In principle, this inference process can be applied to general DocRE models.

We also remove the pseudo document (constructed from the extracted evidence) and the original document separately, denoted as **NoPseudo** and **NoOrigDoc**, respectively. We observe that removing either source will lead to performance drops. Also, the drop of Inter F1 is much larger than Intra F1 for **NoPseudo**, indicating that our inference process is effective for inter-sentence pairs where the evidence may not be consecutive.

As for **NoBlending**, we remove the blending layer and simply take the union of the two sets of

	Co-occur	Coref	Bridge	Total
Count	6711	984	3212	10,907
Percent	54.46%	7.99%	26.07%	88.52%

Table 6: Statistics of the 12,323 relations in the DocRED development set.

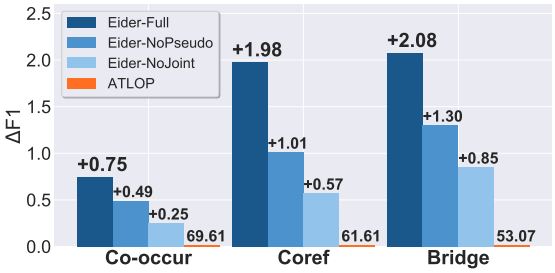

Figure 3: Performance gains in F1 by relation categories. The gains are relative to the second best baseline (ATLOP-RoBERTa$_{large}$).

results. The sharp drop of performance indicates the blending layer can successfully learn a dynamic threshold to combine the prediction results.

Finally, we further finetune the RE model on ground truth evidence before feeding it the extracted evidence (denoted as **FinetuneOnEvi**) but the performance is not improved, probably because the encoded entity representations in evidence and original documents are already highly similar.

Performance Breakdown. To further analyze the performance of EIDER on different types of entity pairs, we categorize the relations into three categories based on our three heuristic rules in Sec. 3.3: *Co-occur*, *Coref* and *Bridge*. The number and percentage of relations covered by each rule are listed in Table 6. We can see that the three categories cover over 88% of the relations in the development set. The results on each category are shown in Figure 3. We can see that our full model has the best performance in all three categories and our ablations also outperform ATLOP. For all our methods, the improvements over ATLOP is *Bridge* > *Coref* ≫ *Co-occur*. This reveals that both modules mainly improve the model's reasoning ability from multiple sentences, either by coreference reasoning or by multi-hop reasoning over a third entity.

Model	Memory	Training time
ATLOP-BERT$_{base}$	9,139 MB	5.19 it/s
E2GRE-BERT$_{base}$	36,182 MB	0.53 it/s
EIDER-BERT$_{base}$	10,933 MB	4.92 it/s

Table 7: Training time and memory usage on DocRED.

Ground Truth Relation: Located in **Ground Truth Evidence Sentence(s)**: [1, 2] **Extracted Evidence Sentence(s)**: [1, 2]
Document: [1] The Portland Golf Club is a private golf club in the northwest United States , in suburban Portland, Oregon. [2] It is located in the unincorporated Raleigh Hills area of eastern Washington County, southwest of downtown Portland and east of Beaverton. [3] The club was established in the winter of 1914, when a group of nine businessmen assembled to form a new club after leaving their respective clubs **...**
Final Prediction: Located in **Prediction on Orig. Doc**: Located in **Prediction on Extracted Evidences**: Located in

Ground Truth Relation: Characters **Ground Truth Evidence Sentence(s)**: [1, 3] **Extracted Evidence Sentence(s)**: [1, 3]
Document: [1] King Louie is a fictional character introduced in Walt Disney's 1967 animated musical film, The Jungle Book. [2] Unlike the majority of the adapted characters in the film, Louie was not featured in Rudyard Kipling's original works. [3] King Louie was portrayed as an orangutan who was the leader of the other jungle primates, and who attempted to gain knowledge of fire from Mowgli, **...**
Final Prediction: Characters **Prediction on Orig. Doc**: NA **Prediction on Extracted Evidences**: Characters

Ground Truth Relation: Inception **Ground Truth Evidence Sentence(s)**: [5, 6] **Extracted Evidence Sentence(s)**: [5]
Document: [1] Oleg Tinkov (born 25 December 1967) is a Russian entrepreneur and cycling sponsor. **...** [5] Tinkoff is the founder and chairman of the Tinkoff Bank board of directors (until 2015 it was called Tinkoff Credit Systems). [6] The bank was founded in 2007 and as of December 1, 2016, it is ranked 45 in terms of assets and 33 for equity among Russian banks. **...**
Final Prediction: Inception **Prediction on Orig. Doc**: Inception **Prediction on Extracted Evidences**: NA

Table 8: Case studies of our proposed framework EIDER. We use red, blue and green to color the **head entity**, **tail entity** and **relation**, respectively. The indices of extracted evidence sentences are highlighted with yellow.

Efficiency Comparison. We benchmark the time and memory usage of EIDER on an RTX A6000 GPU. Table 7 shows that our joint model incurs only ~5% training time and ~14% GPU memory overhead. Experiments also show that EIDER can be trained on a single consumer GPU (e.g., an 11GB GTX 1080 Ti) but E2GRE is not able to.

4.4 Case Studies

Table 8 shows a few examples of EIDER. Detailed statistics and error analysis are provided in Appendix A.2. In the first example, the head entity is mentioned in the first sentence and the tail entity appears in the second. We can see that EIDER correctly extracts these sentences as evidence. Since the evidence sentences are consecutive, the predictions on both the original document and the evidence sentences are correct. In the second example, the prediction using only the original document is incorrect, possibly because the "King Louie" in the 1^{st} and 3^{rd} sentences are so far away from each other that the model fails to recognize them as coreference. Hence, it fails to distinguish "King Louie" as a bridge entity and wrongly predicts "NA". Instead, these two sentences are consecutive in the extracted evidence, making it easier for the model to find the bridge. In the last example, the 6^{th} sentence is missing in the extracted evidence, so the extracted evidence does not contain enough information to predict the relation. However, the prediction on the original document is correct, leading to the correct final result.

5 Related Work

Relation Extraction. Previous research efforts on relation extraction mainly concentrate on predicting relations within a sentence (Cai et al., 2016; Zhang et al., 2018, 2019, 2020a). Despite their effectiveness, in the real world, certain relations can only be inferred from multiple sentences. Consequently, recent studies (Quirk and Poon, 2017; Peng et al., 2017; Yao et al., 2019) started to work on document-level relation extraction (DocRE).

Graph-based DocRE. Graph-based DocRE methods generally construct a graph with mentions, entities, sentences, or documents as the nodes, and infer the relations by reasoning on this graph. Zeng et al. (2020) performs multi-hop reasoning on both a mention-level graph and an entity-level graph. Xu et al. (2021) extracts a reasoning path for each relation and encourages the model to reconstruct the path during training. Zeng et al. (2021) separately deals with intra- and inter-sentential entity pairs and performs multi-hop reasoning on a mention-level graph for inter-sentential entity pairs. However, the extracted graph may omit some important information in the text. Complicated operations on the graphs may also hinder the model from capturing the text structure.

Transformer-based DocRE. Another line of studies model cross-sentence relations by implicitly capturing the long-distance token dependencies via the transformer (Vaswani et al., 2017). Zhou et al. (2021) uses attention in the transformers to extract useful context and adopts an adaptive threshold for

each entity pair. Zhang et al. (2021) views DocRE as a semantic segmentation task over the entity matrix and applies a U-Net to capture the correlations between relations. Huang et al. (2021a) guides DocRE by extracting evidence but does not leverage them after extraction. It also highly relies on evidence annotations and suffers from massive runtime and memory overhead. Huang et al. (2021b) predicts on only a few sentences selected by rules, which may miss important information and does not show consistent improvements. In comparison, we design a lightweight evidence extraction model that is significantly more efficient than Huang et al. (2021a) and can improve DocRE even trained on silver labels. EIDER also fuses the extracted evidence in inference, putting more attention to the important sentences without information loss.

6 Conclusion

In this work, we propose EIDER, an **evid**ence-**e**nhanced **R**E framework, which improves DocRE by joint relation and evidence extraction and fusion of extracted evidence in inference. In training, the RE and evidence extraction model provide additional training signals for each other and mutually enhance each other. The joint model is efficient in time and memory and does not rely heavily on the human annotation of evidence. During inference, the prediction results on both the original document and the extracted evidence are combined, which encourages the model to focus on the important sentences while reducing information loss. Experiment results demonstrate that EIDER significantly outperforms existing methods on three public datasets (DocRED, CDR, and GDA), especially on inter-sentence relations.

Acknowledgements

Research was supported in part by US DARPA KAIROS Program No. FA8750-19-2-1004, SocialSim Program No. W911NF-17-C-0099, and INCAS Program No. HR001121C0165, National Science Foundation IIS-19-56151, IIS-17-41317, and IIS 17-04532, and the Molecule Maker Lab Institute: An AI Research Institutes program supported by NSF under Award No. 2019897. Any opinions, findings, and conclusions or recommendations expressed herein are those of the authors and do not necessarily represent the views, either expressed or implied, of DARPA or the U.S. Government.

References

Rui Cai, Xiaodong Zhang, and Houfeng Wang. 2016. Bidirectional recurrent convolutional neural network for relation classification. In *ACL*.

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*.

Kevin Huang, Peng Qi, Guangtao Wang, Tengyu Ma, and Jing Huang. 2021a. Entity and evidence guided document-level relation extraction. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*.

Quzhe Huang, Shengqi Zhu, Yansong Feng, Yuan Ye, Yuxuan Lai, and Dongyan Zhao. 2021b. Three sentences are all you need: Local path enhanced document relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*.

Robin Jia, Cliff Wong, and Hoifung Poon. 2019. Document-level n-ary relation extraction with multiscale representation learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. Biocreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*.

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*.

Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-sentence n-ary relation extraction with graph LSTMs. *Transactions of the Association for Computational Linguistics*.

Chris Quirk and Hoifung Poon. 2017. Distant supervision for relation extraction beyond the sentence boundary. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *ArXiv*.

Y. Shi, Jiaming Shen, Yuchen Li, N. Zhang, Xinwei He, Zhengzhi Lou, Q. Zhu, M. Walker, Myung-Hwan Kim, and Jiawei Han. 2019. Discovering hypernymy in text-rich heterogeneous information network by exploiting context granularity. In *CIKM*.

Hengzhu Tang, Yanan Cao, Zhenyu Zhang, Jiangxia Cao, Fang Fang, Shi Wang, and Pengfei Yin. 2020. HIN: hierarchical inference network for document-level relation extraction. In *Advances in Knowledge Discovery and Data Mining - 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11-14, 2020, Proceedings, Part I*.

Bayu Distiawan Trisedya, Gerhard Weikum, Jianzhong Qi, and Rui Zhang. 2019. Neural relation extraction for knowledge base enrichment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*.

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)*.

Hong Wang, Christfried Focke, Rob Sylvester, Nilesh Mishra, and William Wang. 2019. Fine-tune bert for docred with two-step process. *Computing Research Repository*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*.

David H. Wolpert. 1992. Stacked generalization. *Neural Networks*.

Ye Wu, Ruibang Luo, Henry C. M. Leung, Hing-Fung Ting, and Tak Wah Lam. 2019. RENET: A deep learning approach for extracting gene-disease associations from literature. In *Research in Computational Molecular Biology - 23rd Annual International Conference, RECOMB 2019, Washington, DC, USA, May 5-8, 2019, Proceedings*.

Liyan Xu and Jinho D. Choi. 2020. Revealing the myth of higher-order inference in coreference resolution. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Wang Xu, Kehai Chen, and Tiejun Zhao. 2021. Document-level relation extraction with reconstruction. *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*.

Deming Ye, Yankai Lin, Jiaju Du, Zhenghao Liu, Peng Li, Maosong Sun, and Zhiyuan Liu. 2020. Coreferential Reasoning Learning for Language Representation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Mo Yu, Wenpeng Yin, Kazi Saidul Hasan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. 2017. Improved neural relation detection for knowledge base question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Shuang Zeng, Yuting Wu, and Baobao Chang. 2021. Sire: Separate intra- and inter-sentential reasoning for document-level relation extraction. In *The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*.

Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. 2020. Double graph based reasoning for document-level relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ningyu Zhang, Xiang Chen, Xin Xie, Shumin Deng, Chuanqi Tan, Mosha Chen, Fei Huang, Luo Si, and Huajun Chen. 2021. Document-level relation extraction as semantic segmentation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*.

Ningyu Zhang, Shumin Deng, Zhanlin Sun, Jiaoyan Chen, Wei Zhang, and Huajun Chen. 2020a. Relation adversarial network for low resource knowledge graph completion. In *Proceedings of The Web Conference 2020*.

Ningyu Zhang, Shumin Deng, Zhanlin Sun, Xi Chen, Wei Zhang, and Huajun Chen. 2018. Attention-based capsule networks with dynamic routing for relation extraction. In *EMNLP*.

Ningyu Zhang, Shumin Deng, Zhanlin Sun, Guanying Wang, Xi Chen, Wei Zhang, and Huajun Chen. 2019. Long-tail relation extraction via knowledge graph embeddings and graph convolution networks. In *NAACL-HLT*.

Zhenyu Zhang, Bowen Yu, Xiaobo Shu, Tingwen Liu, Hengzhu Tang, Wang Yubin, and Li Guo. 2020b. Document-level relation extraction with dual-tier heterogeneous graph. In *Proceedings of the 28th International Conference on Computational Linguistics*.

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*.

A Appendices

A.1 Dataset Statistics

Our model is evaluated on three benchmark datasets, where the statistics are shown in Table 9:

DocRED (Yao et al., 2019) is a large human-annotated document-level RE dataset constructed from Wikipedia. In the training set, around 97.03% entity pairs do not hold any explicit relations. In our experiments, the performance on the test set is validated through the Leader board[3].

CDR (Li et al., 2016) is a biomedical relation extraction dataset consisting of 1,500 PubMed abstracts. The only two entity types are chemicals and diseases and the only non-NA relation is the causal relation between chemicals and disease concepts.

GDA (Wu et al., 2019) contains 30,192 MEDLINE abstracts. It is also a biomedical dataset with two entity types only: diseases and genes, and one non-NA relation type only: the interactions between disease concepts and genes.

Statistics	DocRED	CDR	GDA
# Train	3053	500	23353
# Dev	1000	500	5839
# Test	1000	500	1000
# Relation types	97	2	2
# Avg.# entities per Doc	19.5	7.6	5.4
# Avg.# sentences per Doc	8.0	9.7	10.2
Percent of Intra Rel	54.2	75.7	84.7

Table 9: Statistics of the datasets in experiments. The percentage of intra-sentence relations is calculated from the development set of DocRED and calculated from the test set of CDR and GDA.

A.2 Error Analysis of EIDER

The detailed statistics of the predictions of our model are listed in Table 10. Among all the errors, the majority is because the model wrongly predicts the non-NA relations (i.e., $r \in \mathcal{R}$) as "NA" or predicts "NA" as some non-NA relations. Only $\frac{287}{287+4340+3613} = 3.48\%$ of the errors result from wrongly taking some non-NA relation as another.

To check the exact reason why our model makes these errors, we randomly select 50 cases from DocRED where our model predicts wrongly. We summarize the error types in Table 11 and provide one or two examples for each of the common error types in Table 12.

[3]Results can be found at https://competitions.codalab.org/competitions/20717.

		Ground Truth	
Prediction		$r \in \mathcal{R}$	NA
	$r \in \mathcal{R}$ (Correct)	7,696 (✓)	3,613 (✗)
	$r \in \mathcal{R}$ (Wrong)	287 (✗)	
	NA	4,340 (✗)	380,854 (✓)

Table 10: Statistics of one run of EIDER-RoBERTa$_{\text{large}}$. "$r \in \mathcal{R}$" means non-NA relations. We use "✓" and "✗" to denote correct and wrong predictions, respectively. For example, we have 4,340 wrong predictions where the ground truth is some $r \in \mathcal{R}$ but the prediction is NA.

Reason	Count
Labeling Mistakes	18
Fail in Commonsense Reasoning	8
Fail in Coreferential Reasoning	6
Fail in Multi-hop Reasoning	4
Fail in Surface-name Reasoning	3
Wrong Evidence Extraction	1
Others	10

Table 11: Error types of EIDER in 50 randomly sampled error cases in DocRED. Where "Labeling Mistakes" means our model predicts correctly but the annotation is wrong.

Our analysis shows that 18 out of 50 "error cases" are actually correct. It suggests that labeling mistakes are still prevalent in the DocRED dataset. We show an example under ***"Error Type 1"*** in Table 12. The annotator wrongly labels *"U.S. Route 20"*, a highway, as the country of *"Capital District"*.

Another common error type is ***"Error Type 2"***: *failing in commonsense reasoning*. These error examples normally require commonsense knowledge of the related entities that does not explicitly present in the document. In the first case, the document shows that the airport is located in *"Michigan"* and is near the *"Crooks Road"*. Then we still require the commonsense knowledge that a road (Crooks Road) is a rather small location compared to a state (Michigan). Finally, we can conclude that *"Crooks Road"* locates in *"Michigan"*.

The second case requires the commonsense knowledge about the church. Specifically, if a pope (Benedict XVI) can remove a priest (Maciel) from the ministry, they must be in the same church and hence share the same religion. From sentence [2] we know the priest, Maciel, is a Catholic, hence the pope, Benedict XVI, must also be a Catholic. Even though our prediction on extracted evidence is correct, the confidence is still not high, leading to the incorrect final prediction. As the logic chain of

***Error Type 1**: Labeling Mistakes* **Ground Truth Relation:** Country (✗) **Ground Truth Evidence Sentence(s)**: [1, 4, 5, 7] **Extracted Evidence Sentence(s)**: [5, 7] **Document**: [1] Westmere is a hamlet in the town of Guilderland, Albany County, New York. [4] It is a suburb of the neighboring city of Albany. [5] U.S. Route 20 (Western Avenue) bisects the community and is the major thoroughfare and main street. **...** [7] Crossgates Mall, the Capital District's largest shopping mall, is in Westmere's northeastern corner. **Final Prediction**: NA **Prediction on Orig. Doc**: NA **Prediction on Extracted Evidences**: NA
***Error Type 2**: Fail in Commonsense Reasoning* **Ground Truth Relation:** Located in **Ground Truth Evidence Sentence(s)**: [1, 5] **Extracted Evidence Sentence(s)**: [1, 5] **Document**: [1] Oakland / Troy Airport is a county-owned public-use airport located east of the central business district of Troy, a city in Oakland County, Michigan, United States. [2] It is included in the Federal Aviation Administration (FAA) National Plan of Integrated Airport Systems for 2017–2021, in which it is categorized as a regional reliever airport facility. **...** [5] It is located between Maple Road and 14 Mile Road and Coolidge Highway and Crooks Road. [6] **...** **Final Prediction**: NA (✗) **Prediction on Orig. Doc**: NA (✗) **Prediction on Extracted Evidences**: NA (✗) **Ground Truth Relation:** Religion **Ground Truth Evidence Sentence(s)**: [1, 6] **Extracted Evidence Sentence(s)**: [1, 2, 6] **Document**: [1] Marcial Maciel Degollado (March 10, 1920 – January 30, 2008) was a Mexican Catholic priest who founded the Legion of Christ and the Regnum Christi movement, serving as general director of the legion from 1941 to 2005. [2] Throughout most of his career, he was respected within the church as "the greatest fundraiser of the modern Roman Catholic church" and as a prolific recruiter of new seminarians. **...** [6] In 2006 Pope Benedict XVI removed Maciel from active ministry based on the results of an investigation that he had started while head of the Congregation for the Doctrine of the Faith, before his election as Pope in April 2005. **Final Prediction**: NA (✗) **Prediction on Orig. Doc**: NA (✗) **Prediction on Extracted Evidences**: Religion
***Error Type 3**: Fail in Coreferential Reasoning* **Ground Truth Relation**: NA **Ground Truth Evidence Sentence(s)**: [] **Extracted Evidence Sentence(s)**: [1] **Document**: [1] Manon Balletti (1740–1776) was the daughter of Italian actors performing in France and lover of the famous womanizer Giacomo Casanova. [2] She was ten years old when she first met him; she happened to be the daughter of Silvia Balletti, an actress of the Comédie Italienne company and younger sister of Casanova's closest friend. **...** **Final Prediction**: Child (✗) **Prediction on Orig. Doc**: Child (✗) **Prediction on Extracted Evidences**: Child (✗)
***Error Type 4**: Fail in Multi-hop Reasoning* **Ground Truth Relation:** Educated at **Ground Truth Evidence Sentence(s)**: [4] **Extracted Evidence Sentence(s)**: [4] **Document**: [1] Ronald Leonard is an American cellist. [2] He has had a distinguished career as a soloist, chamber musician, principal cellist and teacher. **...** [4] He was a winner of the Walter Naumburg Competition while a student at the Curtis Institute of Music, where he studied with Leonard Rose and Orlando Cole. **...** **Final Prediction**: NA (✗) **Prediction on Orig. Doc**: NA (✗) **Prediction on Extracted Evidences**: NA (✗)
***Error Type 5**: Fail in Surface-name Reasoning* **Ground Truth Relation:** Country **Ground Truth Evidence Sentence(s)**: [] **Extracted Evidence Sentence(s)**: [1, 4] **Document**: [1] A Route Army was a type of military organization during the Chinese Republic, and usually exercised command over two or more corps or a large number of divisions or independent brigades. [2] It was a common formation in China prior to the Second Sino-Japanese War but was discarded as a formation type by the National Revolutionary Army after 1938 (other than the 8th Route Army), in favor of the Group Army. [3] Some of the more famous of the Route Armies were: [4] 8th Route Army: Communist guerrilla force in North China. **...** **Final Prediction**: NA (✗) **Prediction on Orig. Doc**: NA (✗) **Prediction on Extracted Evidences**: NA (✗)

Table 12: Examples for the five most common error types. We use red, blue and green to color the **head entity**, **tail entity** and **relation**, respectively. The indices of extracted evidence sentences are highlighted with yellow.

commonsense reasoning is always complicated, it is not easy to find a very similar pattern in the training set, or even during pre-training, which makes the problem difficult for a model.

In most of the cases (5 out of 6) in "***Error Type 3**: Fail in Coreferential Reasoning*", human can still identify the correct relation based on the extracted evidence only. As shown in our example in Table 12, in the first sentence, the model wrongly predicts *"Giacomo Casanova"* as the father of *"Manon Balletti"*, but her real father should be an *"Italian actor performing in France"*. It shows that even the reasoning within a single sentence can be difficult.

Similarly, the example in "***Error Type 4***" also shows that the prediction can still be wrong even if we extract the correct evidence sentences and simplify the problem to sentence-level RE. This suggests that if the performance of sentence-level RE is improved, the performance of DocRE will also improve.

Finally, as described by "***Error Type 5***", some examples require direct reasoning from the surface names of the head and tail entities. As shown in the the last case in Table 12, humans can directly identify that *"China"* is the country of *North China* without reading the document, despite that there are no clue in the document indicates this relation. However, most DocRE models, including EIDER, learn to predict the relations only based on the given document and sometimes fail in such cases.