EIDER: Evidence-enhanced Document-level Relation Extraction

Anonymous ACL submission

Abstract

Document-level relation extraction (DocRE) aims to extract the semantic relations among entity pairs in a document. In DocRE, we observe that (1) a subset of the sentences in a document, noted as the evidence sentences, are often sufficient for predicting the relation between a specific entity pair; (2) these evidence sentences can be extracted in an effective and lightweight manner: by multi-task learning along with the RE model or by heuristic rules. In this paper, we propose a novel DocRE framework called EIDER that automatically extracts and makes use of evidence. EIDER enhances a DocRE model by combining the inference results from the evidence sentences and the original document through a blending layer. The performance can be further improved by jointly training an RE model with an evidence extraction model via multi-task learning. If human-annotated evidence is not available, we can use the evidence extracted by this joint model or by several heuristic rules. Extensive experiments show that EIDER achieves state-of-the-art performance on the DocRED, CDR, and GDA datasets. Remarkably, EIDER outperforms the runner-up by 1.37/1.26 Ign F1/F1 on DocRED. In particular, EIDER-RoBERTa$_{\text{large}}$ significantly improves the performance on entity pairs requiring co-reference/multi-hop reasoning by 1.98/2.08 F1, respectively.

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 entities in a single sentence. However, in the reality, it is 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) (Quirk and Poon, 2017; Peng et al., 2017; Gupta et al., 2019).

Head: Ontario Tail: Canada Relation: [Country, Located in]
Ground truth evidence: [1,4] Extracted evidence: [1,4]

Original document as input: [1] Paul Desmarais Jr. (born July 3, 1954) is a Canadian businessman in his hometown of Montreal. [2] He is the eldest son of Paul Desmarais Sr. and Jacqueline (Maranger) Desmarais [3] Currently he is the Chairman and Co-chief Executive Officer of … [4] Desmarais was born in Sudbury, Ontario.
Pred result (logits): NA: 16.46 **Country**: 15.41 **Located in**: 14.64

Extracted evidence as input: [1] Paul Desmarais Jr. (born July 3, 1954) is a Canadian businessman in his hometown of Montreal. [4] Desmarais was born in Sudbury, Ontario.
Pred result (logits): Country: 14.69 **Located in**: 13.63 NA: 10.93

Final prediction result of our model: Country, Located in

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 [1,4] as evidence, and if we only use the extracted evidence as input, the model can predict the relation "country" and "located in" correctly.

In each document, the sentences are not *equally important* for each entity pair and some sentences could be irrelevant for the relation prediction. For each entity pair, we refer to the minimal set of sentences required to infer their relation as *evidence sentences* (Yao et al., 2019). As shown in Figure 1, to predict the relation between *"Ontario"* and *"Canada"*, it is sufficient to know *Paul Desmarais* is a *Canadian* from the 1^{st} sentence, and *Paul Desmarais* was born in *Ontario* from the 4^{th} one. In other words, the 1^{st} and 4^{th} sentences serve as evidence sentences of this entity pair. Although the 2^{nd} and 3^{rd} sentences lie between these two sentences, they are irrelevant to this specific relation. Including such irrelevant sentences in input might sometimes introduce noise to the model and be more detrimental than beneficial.

In light of the observations above, we propose two approaches to make better use of evidence sentences. The first is to combine evidence in inference. One naive way is to directly make predictions on the evidence sentences, as in (Huang et al., 2021b). However, other sentences may also include relevant information, such as the information of the involved entities. Discarding all other sentences may result in loss of coherence and harm model performance in certain cases. Hence, we fuse the prediction results of the original document and evidence to highlighting the most important sentences while avoiding information loss. Notice that this method does not need additional training and can be applied to general DocRE models.

In case evidence sentences are not provided in inference, they can be extracted by training an evidence extracted model using multi-task learning. This also serves as our second approach to improve DocRE with evidence. Intuitively, both tasks should focus on the information relevant to the current entity pair, such as the underlined *"Paul Desmarais"* and *"Desmarais"* in the 4^{th} sentence of Figure 1. This suggests that the two tasks have certain commonalities and can provide additional training signals for each other. To avoid the massive training time and memory overhead due to training an additional task, our model adopts a simpler model structure and is trained on only part of the evidence annotation, which requires only 5% additional training time and 14% more memory. When human-annotated evidence sentences are not available even for training, we find that a simple set of heuristic rules can serve to construct silver labels with relatively high quality. Experiment results show that even the model trained with silver labels can outperform the baseline significantly.

In this paper, we propose an **evid**ence-**e**nhanced **R**E framework EIDER, which automatically extracts evidence and effectively leverages the extracted evidence to improve the performance of DocRE. We first train a relation extraction model and an evidence extraction model using multi-task learning. If the human-annotated evidence is not accessible even in training, we adopt several heuristic rules to construct silver labels instead. To reduce memory usage and training time, we use the same sentence representation across relations and only train the evidence extraction model on positive entity pairs with at least one relation. During inference, we construct a pseudo document by concatenating all the evidence (or predicted evidence). Finally, we fuse the predictions based on the original document and the pseudo document using a blending layer (Wolpert, 1992).

Extensive experiments show that EIDER outperforms the state-of-the-art methods on widely-adopted DocRE benchmarks DocRED (Yao et al., 2019), CDR (Li et al., 2016) and GDA (Wu et al., 2019). Further examination shows that the improvement of EIDER is especially large on inter-sentence entity pairs, where multiple sentences are involved.

Contributions. (1) We propose an evidence-empowered inference process of DocRE, which improves the performance without re-training the RE model. (2) We jointly learn relation and evidence extraction using multi-task learning, where the two tasks mutually enhance each other. (3) In the absence of human-annotated evidence labels, we design a set of rules to construct evidence labels and show that these silver labels can already improve DocRE performance. (4) We demonstrate that EIDER outperforms state-of-the-art methods on three DocRE datasets: DocRED, CDR, and GDA.

2 Problem Formulation

Given a document d comprised of N sentences $\{s_t\}_{t=1}^{N}$, L tokens $\{h_l\}_{l=1}^{L}$ and a set of entities $\{e_i\}$ appearing in d, 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. An entity e_i may appear multiple times in document d, where we denote its corresponding mentions as $\{m_j^i\}$. 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 sentences*[1] $V_{h,t} = \{s_{v_i}\}_{i=1}^{K}$ as the subset of sentences in the document that are sufficient for human annotators to infer the relation.

3 Methodology

In this section, we will first introduce our base relation extraction model (Sec. 3.1) and then propose two methods to improve DocRE by using evidence: evidence-empowered inference (Sec. 3.2) and evidence extraction as an auxiliary task (Sec. 3.3). We also provide several heuristic rules (Sec. 3.4) to construct evidence labels in case the evidence annotation is not available. An illustration of our framework is shown in Figure 2.

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

3.1 Base Relation Extraction Model

Base Encoder. Given a document $d = [h_l]_{l=1}^{L}$, we insert a special token "*" before and after each entity mention and encode the document with a pre-trained encoder (Devlin et al., 2019) to obtain the s-dim embedding of each token, aggregated as a matrix $\boldsymbol{H} \in \boldsymbol{R}^{L \times s}$:

$$\boldsymbol{H} = [\boldsymbol{h}_1, ..., \boldsymbol{h}_L] = \text{Encoder}([h_1, ..., h_L]). \quad (1)$$

For each mention of an entity e_i, we use the embedding of the start symbol "*" as its mention embedding. 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: $\mathbf{e}_i = \log \sum_j \exp(\mathbf{m_j^i})$.

Following Zhou et al. (2021), we capture the context for each entity pair (e_h, e_t) by computing a context embedding $\mathbf{c}_{h,t} \in \boldsymbol{R}^s$ based on the attention scores from the pre-trained encoder:

$$\begin{aligned} \mathbf{c}_{h,t} &= \boldsymbol{H}^T \mathbf{a}_{h,t} \\ \mathbf{a}_{h,t} &= \text{Normalize}(\sum_{k=1}^{K} \mathbf{A}_h^k \circ \mathbf{A}_t^k). \end{aligned} \quad (2)$$

where K is the number of attention heads, and $\mathbf{A}_h^k \in \mathbb{R}^L$ is the attention from e_h to each token under attention head k, computed by averaging the attention from each of its mentions m_j^h to each token. Similarly for $\mathbf{A}_t^k$. The intuition is that tokens with high attention towards both e_h and e_t are important to both entities. Hence, these tokens are essential to the relation and should contribute more to the context embedding.

Relation Prediction Head. We first map the embeddings of $(e_h,\ e_t)$ to context-aware representations $(\mathbf{z_h},\ \mathbf{z_t})$ by combining their entity embeddings with the context embedding $\mathbf{c}_{h,t}$, and then obtain the probability of relation $r \in \mathcal{R}$ between (e_h, e_t) via a bilinear function:

$$\begin{aligned} \boldsymbol{z}_h &= \tanh\left(\boldsymbol{W}_h \mathbf{e}_h + \boldsymbol{W}_{c_h} \mathbf{c}_{h,t}\right), \\ \boldsymbol{z}_t &= \tanh\left(\boldsymbol{W}_t \mathbf{e}_t + \boldsymbol{W}_{c_t} \mathbf{c}_{h,t}\right), \\ \mathbf{y}_r &= \left(\boldsymbol{z}_h \boldsymbol{W}_r \boldsymbol{z}_t + \boldsymbol{b}_r\right), \\ \mathrm{P}\left(r|e_h, e_t\right) &= \sigma \mathbf{y}_r \end{aligned} \quad (3)$$

where $\boldsymbol{W}_h, \boldsymbol{W}_t, \boldsymbol{W}_{c_h}, \boldsymbol{W}_{c_t}, \boldsymbol{W}_r, \boldsymbol{b}_r$ are learnable parameters. As the model may have different confidence for different entity pairs or classes, 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:

$$\mathbf{y}_{\text{TH}} = (\boldsymbol{z}_h \boldsymbol{W}_{\text{TH}} \boldsymbol{z}_t + \boldsymbol{b}_r) \quad (4)$$

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

$$\begin{aligned} \mathcal{L}_{RE} = &- \sum_{h \neq t} \sum_{r \in \mathcal{P}_{h,t}^T} \log \left(\frac{\exp\left(\mathbf{y}_r\right)}{\sum_{r' \in \mathcal{P}_{h,t}^T \cup \{\text{TH}\}} \exp\left(\mathbf{y}_{r'}\right)} \right) \\ &- \log \left(\frac{\exp\left(\mathbf{y}_{\text{TH}}\right)}{\sum_{r' \in \mathcal{N}_{h,t}^T \cup \{\text{TH}\}} \exp\left(\mathbf{y}_{r'}\right)} \right) \end{aligned} \quad (5)$$

3.2 Evidence-empowered Inference

Suppose we are given the ground truth evidence and it already contains all the information relevant to the relation, then there is no need to use the whole document for relation extraction. Instead, we can construct a pseudo document $d'_{h,t}$ for each entity pair by concatenating the evidence sentences $V_{h,t}$ in the order they are presented in the original document and feed the pseudo document to the trained model to obtain another set of prediction scores $S_{h,t,r}^{(E)}$. It may simplify the input, making it easier for the model to make the correct predictions.

However, the non-evidence sentences in the original document may also provide background information of the entities and possibly contributes to the prediction. Hence, solely relying on evidence sentences may result in information loss and lead to sub-optimal performance. As a result, we combine the prediction results on both the original documents and the extracted evidence.

After obtaining two sets of relation prediction results from the original documents and the pseudo documents, we fuse the results by aggregating the prediction scores from original documents and pseudo documents, denoting as $S^{(O)}$ and $S^{(E)}$, through a blending layer (Wolpert, 1992):

$$\mathrm{P}_{Fuse}\left(r|e_h, e_t\right) = \sigma(S_{h,t,r}^{(O)} + S_{h,t,r}^{(E)} - \tau), \quad (6)$$

where τ is a learnable parameter. We optimize the parameter τ on the development set as follows:

$$\begin{aligned} \mathcal{L}_{Fuse} = - \sum_{d \in \mathcal{D}} \sum_{h \neq t} \sum_{r \in \mathcal{R}} & y_r \cdot \mathrm{P}_{Fuse}\left(r|e_h, e_t\right) + \\ & (1 - y_r) \cdot \log(1 - \mathrm{P}_{Fuse}\left(r|e_h, e_t\right)), \end{aligned} \quad (7)$$

where $y_r = 1$ if the relation r holds between (e_h, e_t) and $y_r = 0$ otherwise.

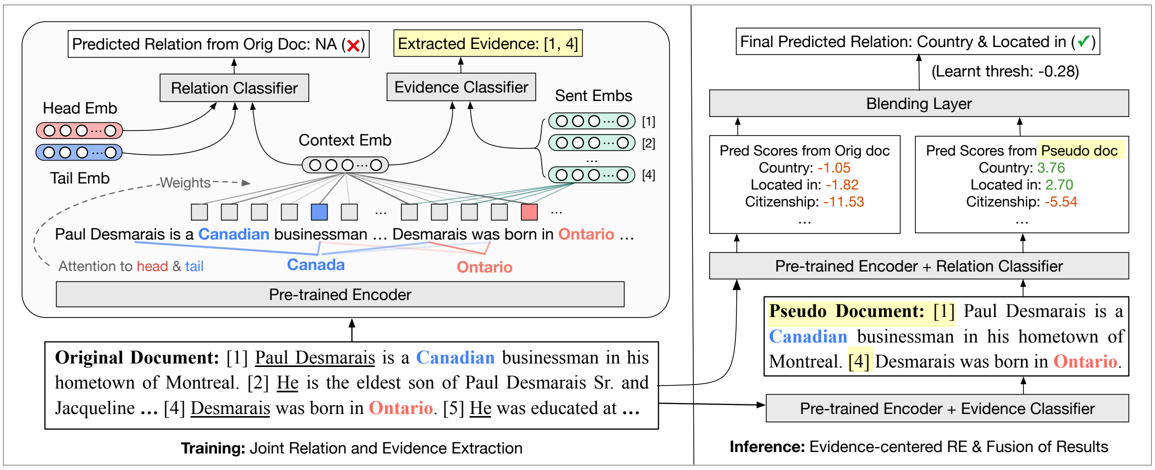
Predicted Relation from Orig Doc: NA (✗)
Extracted Evidence: [1, 4]
Relation Classifier
Evidence Classifier
Head Emb
Sent Embs
[1]
[2]
[4]
Context Emb
Tail Emb
Weights
Paul Desmarais is a Canadian businessman ... Desmarais was born in Ontario ...
Attention to head & tail
Canada
Ontario
Pre-trained Encoder
Original Document: [1] Paul Desmarais is a Canadian businessman in his hometown of Montreal. [2] He is the eldest son of Paul Desmarais Sr. and Jacqueline ... [4] Desmarais was born in Ontario. [5] He was educated at ...
Training: Joint Relation and Evidence Extraction
Final Predicted Relation: Country & Located in (✓)
(Learnt thresh: -0.28)
Blending Layer
Pred Scores from Orig doc
Country: -1.05
Located in: -1.82
Citizenship: -11.53
Pred Scores from Pseudo doc
Country: 3.76
Located in: 2.70
Citizenship: -5.54
Pre-trained Encoder + Relation Classifier
Pseudo Document: [1] Paul Desmarais is a Canadian businessman in his hometown of Montreal. [4] Desmarais was born in Ontario.
Pre-trained Encoder + Evidence Classifier
Inference: Evidence-centered RE & Fusion of Results

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

3.3 Evidence Extraction as Auxiliary Task

In reality, the annotation of evidence may be available only in training but not in inference. To automatically extract evidence, we jointly train the relation extraction model and evidence extraction model using multi-task learning. Intuitively, tokens relevant to the relation are essential in both models. By sharing the base encoder, the two models can provide additional training signals for each other and hence mutually enhance each other (Ruder, 2017; Liu et al., 2019).

The evidence extraction model predicts whether each sentence s_i is an evidence sentence of entity pair (e_h, e_t). Similar to entity embeddings, to obtain sentence embedding $\mathbf{s}_i$, we apply a LogSumExp pooling over all the tokens in s_i: $\mathbf{s}_i = \log \sum_{h_l \in s_i} \exp(\mathbf{h}_l)$. Intuitively, if s_i is an evidence sentence of (e_h, e_t), the tokens in s_i would be relevant to the relation prediction, and would contributing more to $\mathbf{c}_{h,t}$. Hence, we use a bilinear function between context embedding $\mathbf{c}_{h,t}$ and sentence embedding $\mathbf{s}_i$ to measure the importance of sentence s_i to entity pair (e_h, e_t):

$$\mathrm{P}(s_i|e_h, e_t) = \sigma(\mathbf{s}_i \boldsymbol{W}_v \mathbf{c}_{h,t} + \boldsymbol{b}_v), \tag{8}$$

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.

$$\mathcal{L}_{Evi} = -\sum_{h \neq t, NA \notin \mathcal{P}^T_{h,t}} \sum_{s_i \in \mathcal{D}} y_i \cdot \mathrm{P}(s_i|e_h, e_t) + (1 - y_i) \cdot \log(1 - \mathrm{P}(s_i|e_h, e_t)), \tag{9}$$

where y_i is 1 when $s_i \in V_{h,t}$ and $y_i = 0$ otherwise.

When training for evidence prediction, we only use the entity pairs with at least one relation $r \in \mathcal{R}$, which accounts for a small subset (2.97% in DocRED) of the total possible entity pairs. Note that only such pairs have human-annotated evidence sentences. While previous work (Huang et al., 2021a) treated the remaining entity pairs as negative examples for every sentence and predict on the (e_h, e_t, r, s_i) tuple level, we find this unnecessary and inefficient in terms of memory and time. We observe that most of the entity pairs only have one set of evidence across relations, and hence only predict for each (e_h, e_t, s_i) tuple. Then for this "relation-agnostic" evidence, a negative entity pair that does not have a relation $r \in \mathcal{R}$ does not necessarily imply that the entity pair does not have any relation. Taking such pairs as negative examples also makes the training set highly unbalanced.

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

$$\mathcal{L} = \mathcal{L}_{RE} + \mathcal{L}_{Evi}. \tag{10}$$

If we do not have access to evidence during inference, we can use the extracted evidence $V'_{h,t}$ instead of the ground truth evidence $V_{h,t}$ in the evidence-empowered inference introduced in Sec. 3.2.

3.4 Heuristic Evidence Label Construction

Both of our methods aim at facilitating relation extraction with evidence. However, human-annotated evidence is not always available. [2] In this case,

[2] If human-annotated evidence sentences are available, then we do not need to go through this step.

we design several heuristic rules to automatically construct silver labels for evidence extraction:

Intra. If the head and tail entities co-occur in the same sentence (e.g., "Desmarais" and "Ontario" co-occur in the 4^{th} sentence in Figure 2), we use all the sentences they co-occur as evidence.

Coref. If the entity mention pairs of the head and tail entity do not co-occur explicitly, but their coreferential mentions co-occur (e.g., "Paul Desmarais" and "Canadian", the co-reference of "Canada" co-occur in the 1^{st} sentence in Figure 2), we use all the sentences where their coreferential mentions co-occur. In practice, we may directly apply existing coreference resolution models such as HOI (Xu and Choi, 2020) without re-training.

Bridge. If the first two conditions are not satisfied, but there exists a third bridge entity whose coreferential mention co-occurs with both head and tail (e.g., "Paul Desmarais" co-occurs with both "Canadian" and "Ontario" in Figure 2), we take all the sentences where the bridge co-occurs with head or tail as the evidence. If there is more than one bridge entity, we choose the one with the highest frequency. This rule can be easily extend to multiple bridges. Empirically, we observe that capturing one bridge already leads to satisfying results.

4 Experiments

4.1 Experiment Setup

Dataset. 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). The details of the datasets are listed in Appendix A. DocRED is the only dataset that provides evidence sentences as part of the annotation, and the evidence annotation is not visible in inference.

Evaluation Metrics. Following prior studies (Yao et al., 2019; Zhou et al., 2021; Huang et al., 2021a), we use **F1** and **Ign F1** as the main evaluation metrics for relation extraction and use **Evi F1** and **PosEvi F1** as the metric for evidence extraction. Ign F1 measures the F1 score excluding the relations shared by the training and development/test set. PosEvi F1 measures the F1 score of evidence only on entity pairs with explicit relations (positive pairs). 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 relations where none of the entity mention pairs co-occurs.

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, model cross-sentence relations by implicitly capturing the long-distance token dependencies via the transformer. We also compare to two ablations of our method: EIDER (Rule), where we use rule-based evidence labels instead of golden labels, and ATLOP + Fuse, where we directly apply our evidence-enhance inference on the checkpoint of ATLOP without re-training. The evidence is also extracted by rules.

Relation Extraction Results. Table 1 presents the relation extraction results, where we observe that EIDER outperforms the baseline methods in all datasets. For instance, EIDER-BERT$_{\text{base}}$ significantly improves ATLOP (Zhou et al., 2021) by 1.47/1.40 F1/Ign F1, which uses the same base relation extraction model as our method.

The experiment results also show that our improvement on Inter F1 is much larger than that on Intra F1. For instance, EIDER outperforms ATLOP by 1.21/2.01 Intra/Inter F1 under BERT$_{\text{base}}$ (0.75/1.52 under RoBERTa$_{\text{large}}$). 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 during training and uses these important sentences during inference.

Among the baselines, we observe that the Inter F1 of GAIN (Zeng et al., 2020) is 0.70 higher while the Intra F1 of ATLOP is 0.16 higher. Such results indicate that graph-based methods may capture the long-distance dependency between entities by directly connecting them on the graph. Although EIDER does not involve an explicit multi-hop reasoning module, it still notably outperforms the graph-based models in terms of Inter F1, demonstrating that the evidence-empowered inference also relieves long-distance dependency challenge by directly concatenating important sentences.

Finally, in both DocRED and the two biomedical datasets which do not have evidence annotation, EIDER (Rule) also outperforms all baselines. This shows that EIDER still performs well without evidence annotation. The improvement of ATLOP + Fuse further shows that our inference approach can be applied to general DocRE models without retraining. The improvement on DocRED and CDR

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., 2020)	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 ± 0.14†	61.01 ± 0.10†	67.26 ± 0.15†	53.20 ± 0.19†	59.31	61.30
ATLOP-BERT$_{base}$ + Fuse	60.01 ± 0.14	62.09 ± 0.09	68.21 ± 0.10	54.34 ± 0.15	-	-
Eider (Rule)-BERT$_{base}$	60.36 ± 0.13	62.34 ± 0.08	68.40 ± 0.14	54.79 ± 0.13	-	-
Eider-BERT$_{base}$	**60.51 ± 0.11**	**62.48 ± 0.13**	**68.47 ± 0.08**	**55.21 ± 0.21**	**60.42**	**62.47**
BERT$_{large}$ (Ye et al., 2020)	56.67	58.83	-	-	56.47	58.69
CorefBERT$_{large}$ (Ye et al., 2020)	56.82	59.01	-	-	56.40	58.83
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
GAIN-BERT$_{large}$ (Zeng et al., 2020)	60.87	63.09	-	-	60.31	62.76
ATLOP-RoBERTa$_{large}$ (Zhou et al., 2021)	61.30 ± 0.22†	63.15 ± 0.21†	69.61 ± 0.25†	55.01 ± 0.18†	61.39	63.40
ATLOP-RoBERTa$_{large}$ + Fuse	61.48 ± 0.13	63.64 ± 0.14	69.61 ± 0.19	56.17 ± 0.22	-	-
Eider (Rule)-RoBERTa$_{large}$	61.73 ± 0.07	63.91 ± 0.07	69.99 ± 0.09	56.27 ± 0.11	-	-
Eider-RoBERTa$_{large}$	**62.34 ± 0.14**	**64.27 ± 0.10**	**70.36 ± 0.07**	**56.53 ± 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 ± 0.6	82.5 ± 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 ± 1.1	83.9 ± 0.2
Eider (Rule)-SciBERT$_{base}$	**70.63** ± 0.49	**84.54** ± 0.22

Table 2: Relation extraction results on CDR and GDA.

Model	Dev PosEvi F1	Dev Evi F1	Test Evi F1
Eider-rules	77.43	-	-
E2GRE-BERT$_{base}$	-	47.14	48.35
Eider-BERT$_{base}$	80.33	**50.71**	**51.27**
E2GRE-RoBERTa$_{large}$	-	51.11	50.50
Eider-RoBERTa$_{large}$	81.51	**52.54**	**53.01**

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

is much larger than that on GDA. We hypothesize that it is because more than 85% relations in GDA are intra-sentence relations, so the model might already learn to focus on the important sentences without the help of evidence.

Evidence Extraction Results. To our knowledge, E2GRE is the only method that has reported their evidence extraction result. The results of evidence prediction in Table 3 indicate that Eider outperforms E2GRE significantly (e.g., by 3.57 Dev Evi F1 under BERT$_{base}$). One possible reason is that the incorporation of context vector models the dependency between tokens, leading to better performance in evidence extraction. The results show that it may be sufficient to make predictions on positive pairs only and over each (entity, entity, sentence) tuple (instead of (sentence, relation, entity, entity) as in E2GRE). We also observe that our three heuristic rules already capture most of the evidence for the positive pairs (77.43 F1). This again demonstrates that our model can perform well even without relying on evidence annotations.

4.3 Performance Analysis

Ablation Study. We conduct ablation studies to further analyze the utility of each module in Eider. The results are shown in Table 4.

We first train the RE model and the evidence extraction model separately, denoted as **NoJoint**. We observe that the drop in Inter F1 is more significant (i.e., 0.50/1.04 Intra F1/Inter F1), which shows that the evidence and relation extraction model mutually enhance each other's ability to identify the related context of each entity pair.

Then, we remove the extracted evidence and

Ablation	Ign F1	F1	Intra F1	Inter F1
Eider-RoBERTa$_{\text{large}}$	**62.34**	**64.27**	**70.36**	**56.53**
NoJoint	61.56	63.40	69.86	55.49
NoEvi	61.94	63.81	70.10	55.94
NoOrigDoc	60.26	62.68	68.36	55.49
NoBlending	61.09	63.47	69.25	56.27
FinetuneOnEvi	61.84	63.92	69.86	56.40

Table 4: Ablation studies of Eider.

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

Table 5: The statistics of the 12,323 relations in the DocRED development set.

the original document separately, denoted as **NoEvi** 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 **NoEvi**, indicating that the extracted evidence is more effective for cross-sentence entity pairs where the important sentences may not be consecutive.

As for **NoBlending**, we remove the blending layer and simply take the union of the two sets of 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**). We observe that the performance is not improved, probably because the encoded entity representation in evidence and original documents are already similar to each other. In fact, when performing relation extraction on the training set using the ground truth evidence alone, the F1 is already over 95%.

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. 3.4: *Intra*, *Coref* and *Bridge*. The number and percentage of relations covered by each rule are listed in Table 5. 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. The differences between models vary by category. For all our methods, the improvements over ATLOP is *Bridge* $>$ *Coref* $\gg$ *Intra*. 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.

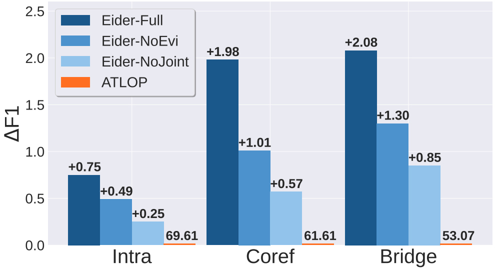

Figure 3: Performance gains in F1 by relation categories. The gains are relative to the second best baseline (ATLOP).

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

Table 6: Training time and memory usage.

Efficiency Comparison. We benchmark the time and memory usage of Eider on an RTX A6000 GPU. Table 6 show 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 7 shows a few examples of Eider. In the first example, the head entity is mentioned in the first sentence and the tail entity appears in the second sentence. We can see that the model correctly extracts these evidence sentences. Since the evidence sentences are consecutive, both the predictions on the original document and the evidence sentences are correct. In the second example, the 2^{nd} sentence is a distracting sentence as it does not contain any useful information involving the target entity pair. 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 link them. Hence, it also fails to distinguish "King Louie" as a bridge entity in this case. However, 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 for prediction. However, the prediction on the original document is correct, leading to the correct final result.

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 7: 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.

5 Related Work

Relation Extraction. Previous research efforts on relation extraction mainly concentrate on predicting relations within a sentence (Cai et al., 2016; Zeng et al., 2015; Feng et al., 2018; Zheng et al., 2021; Zhang et al., 2018, 2019, 2020a). While these approaches tackle the sentence-level RE task effectively, 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; Wang et al., 2019; Tang et al., 2020) have proposed to work on the 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. Specifically, Nan et al. (2020) constructs a document-level graph and iteratively updates the node representations, and refines the graph topological structure. Zeng et al. (2020) performs multi-hop reasoning on both a mention-level graph and an entity-level graph. Xu et al. (2020) extracts a reasoning path between each entity pair holding at least one relation and encourages the model to reconstruct the path during training. These methods simplify the input document by extracting a graph with entities and performing explicit graph reasoning. However, the complicated operations on the graphs lower the efficiency of the methods.

Transformer-based DocRE. Another line of studies solely relies on the transformer architecture (Devlin et al., 2019) to model cross-sentence relations since transformers can implicitly capture long-distance dependencies. Zhou et al. (2021) uses attention in the transformers to extract useful context and adopts an adaptive threshold for each entity pair. Huang et al. (2021b) makes predictions on the evidence sentences extracted by several hand-crafted rules, which may suffer from information loss. Instead, EIDER combines the predictions on both the evidence and the original document. Similar to our method, Huang et al. (2021a) jointly extracts relation and evidence. However, our method does not rely on human-annotated evidence and uses a much simpler model structure and hence reduces time and memory usage. We are also the first work to fuse the predictions based on extracted evidence sentences in inference.

6 Conclusion

In this work, we propose EIDER, an **evid**ence-**e**nhanced **RE** framework, which improves DocRE by joint relation and evidence extraction and fusion of extraction results in inference. We also provide an evidence label construction method so that our model does not rely heavily on the human annotation of evidence. In training, the relation extraction and evidence extraction model provide additional training signals for each other and mutually enhance each other. The joint model adopts a simple model structure and is efficient in time and memory. 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.

References

Rui Cai, Xiaodong Zhang, and Houfeng Wang. 2016. Bidirectional recurrent convolutional neural network for relation classification. In *ACL*, pages 756–765.

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*, pages 4171–4186.

Jun Feng, Minlie Huang, Li Zhao, Yang Yang, and Xiaoyan Zhu. 2018. Reinforcement learning for relation classification from noisy data. In *AAAI*, pages 5779–5786.

Pankaj Gupta, Subburam Rajaram, Hinrich Schütze, and Thomas A. Runkler. 2019. Neural relation extraction within and across sentence boundaries. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, pages 6513–6520.

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.

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*, pages 3693–3704.

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*, 2016.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *ACL*.

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.

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*, 5:101–115.

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*, pages 1171–1182.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *ArXiv*, abs/1706.05098.

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*, volume 12084 of *Lecture Notes in Computer Science*, pages 197–209.

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*, pages 229–240, Florence, Italy. Association for Computational Linguistics.

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

Hong Wang, Christfried Focke, Rob Sylvester, Nilesh Mishra, and William Wang. 2019. Fine-tune bert for docred with two-step process. *Computing Research Repository*, arXiv:1909.11898.

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*, pages arXiv–1910.

David H. Wolpert. 1992. Stacked generalization. *Neural Networks*, 5:241–259.

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*, volume 11467, pages 272–284.

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)*, pages 8527–8533. Association for Computational Linguistics.

Wang Xu, Kehai Chen, and Tiejun Zhao. 2020. Document-level relation extraction with reconstruction.

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

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)*, pages 7170–7186.

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)*, pages 571–581, Vancouver, Canada. Association for Computational Linguistics.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *EMNLP*, pages 1753–1762.

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)*, pages 1630–1640.

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*, pages 1630–1641, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hengyi Zheng, Rui Wen, Xi Chen, Yifan Yang, Yunnan Zhang, Ziheng Zhang, Ningyu Zhang, Bin Qin, Xu Ming, and Yefeng Zheng. 2021. Prgc: Potential relation and global correpondence based joint relational triple extraction. In *ACL*.

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 Dataset Statistics

Our model is evaluated on three benchmark datasets:

DocRED (Yao et al., 2019) is a large human-annotated document-level RE dataset, which consists of 3,053/1,000/1,000 documents for training/development/testing, respectively. DocRED is constructed from Wikipedia, involving 96 relation types, 132,275 entities, and 56,354 relations. In the training set, around 97.03% entity pairs do not hold any explicit relations. Around 54.2% of the relations are intra-sentence. The others can only be extracted by considering multiple sentences.

CDR (Li et al., 2016) is a relation extraction dataset in the biomedical domain, with 500/500/500 documents in the train/development/test set. The only two entity types are chemicals and diseases and the only explicit relation is the causal relation between chemicals and disease concepts. In the test set, around 75.7% of the relations are intra-sentence.

GDA (Wu et al., 2019) is also a biomedical dataset, which consists of 29,192/200/800 documents for training/development/testing. It also contains two entity types only: diseases and genes, and one relation type only: the interactions between disease concepts and genes. In the test set, around 84.7% of the relations are intra-sentence.

B Experimental Details

B.1 Implementation Details

Our model is implemented based on PyTorch and Huggingface's Transformers (Wolf et al., 2019). We use cased-BERT$_{\text{base}}$ (Devlin et al., 2019) and 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 BERT$_{\text{base}}$ models are trained with one GTX 1080 Ti GPU and RoBERTa$_{\text{large}}$ models with one RTX A6000 GPU.