EIDER: Evidence-enhanced Document-level Relation Extraction

Anonymous ACL submission

Abstract

Document-level relation extraction (DocRE) aims at extracting the semantic relations among entity pairs in a document. In DocRE, a subset of the sentences in a document, called the evidence sentences, might be sufficient for predicting the relation between a specific entity pair. To make better use of the evidence sentences, in this paper, we propose a three-stage evidence-enhanced DocRE framework called EIDER[1] consisting of joint relation and evidence extraction, evidence-centered relation extraction (RE), and fusion of extraction results. We first jointly train an RE model with a simple and memory-efficient evidence extraction model. Then, we construct pseudo documents based on the extracted evidence sentences and run the RE model again. Finally, we fuse the extraction results of the first two stages using a blending layer and make a final prediction. Extensive experiments show that our proposed framework achieves state-of-the-art performance on the DocRED dataset, outperforming the second-best method by 1.37/1.26 Ign F1/F1. In particular, EIDER-RoBERTa$_{\text{large}}$ significantly improves the performance on entity pairs requiring co-reference and multi-hop reasoning by 1.98/2.08 F1, respectively, which cover around 75% of the cross-sentence samples.

1 Introduction

Relation extraction (RE) is the task of extracting semantic relations among entities within a given text, which is a critical step in information extraction and has abundant applications (Yu et al., 2017; Shi et al., 2019; Trisedya et al., 2019). Prior studies mostly focus on sentence-level RE, where the two entities of interest co-occur in the same sentence and it is assumed that their relation can be derived from the sentence (Zeng et al., 2015; Cai et al., 2016). However, this assumption does not always hold and some relations between entities 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).

h: Hero of the Day t: the United States r: [country of origin]
Ground truth evidence: [1,10] Extracted evidence: [1,10]

Original document as input: [1] Load is the sixth studio album by the American heavy metal band Metallica, released on June 4, 1996 by Elektra Records in the United States and by Vertigo Records internationally. … [9] It was certified 5×platinum by the Recording Industry Association of America (RIAA) for shipping five million copies in the United States. [10] Four singles—"Until It Sleeps", "Hero of the Day", "Mama Said", and "King Nothing" — were released as part of the marketing campaign for the album.
Prediction result (logits): NA: 17.63 country of origin: 14.79

Extracted evidence as input: [1] Load is … in the United States and by Vertigo Records internationally. [10] Four singles —"Until It Sleeps", "Hero of the Day", … for the album.
Prediction result (logits): country of origin: 18.31 NA: 13.45

Final prediction result of our model: country of origin

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.

In each document, the sentences are not *equally important* for each entity pair and some sentences could be irrelevant for the relation prediction. We refer to the minimal set of sentences required to infer a relation as *evidence sentences* (Yao et al., 2019). In the example in Figure. 1, the 1^{st} and 10^{th} sentences serve as evidence sentences to the *"country of origin"* relation between *"Hero of the Day"* and *"the United States"*. The 1^{st} sentence indicates that *Load* is originated from *the United States*, and the 10^{th} indicates *Hero of the Day* is a song of *Load*. Although the 9^{th} sentence also mentions *"the United States"*, it is irrelevant to this specific relation. Including such irrelevant sentences in the input might sometimes introduce noise to the model and be more detrimental than beneficial.

[1]Our code will be released on GitHub for reproducibility.

In light of the observations above, we propose two approaches to make better use of evidence sentences. The first is to jointly extract relations and evidence. Intuitively, both tasks should focus on the information relevant to the current entity pair, such as the underlined *"Load"* and *"the album"* in the 10^{th} sentence of Figure 1. This suggests that the two tasks have certain commonalities and can provide additional training signals for each other. Huang et al. (2020) trains these two tasks in a multi-task learning manner. However, their model makes evidence prediction for every <sentence, relation, entity, entity> tuple, which requires massive GPU memory and long training time. Our method adopts a much simpler model structure, which only predicts for each <sentence, entity, entity> tuple and can be trained on a single consumer GPU.

The second approach is to conduct evidence-centered relation extraction with the evidence sentences as model input. In the extreme case, if there is only one sentence related to the relation prediction, we can make predictions solely based on this sentence and reduce the problem to sentence-level relation extraction. One concurrent work (Huang et al., 2021) shows the effectiveness of replacing the original documents with sentences extracted by hand-crafted rules. However, the sentences extracted by rules are not perfect. Solely relying on heuristically extracted sentences may result in information loss and harm model performance in certain cases. Instead, our evidence is obtained by a dedicated evidence extraction model. In addition, we fuse the prediction results of the original document and extracted evidence to avoid information loss, and demonstrate that the two sources of predictions are complementary to each other.

Specifically, in this paper, we propose an **evid**ence-**e**nhanced **RE** framework, named EIDER, which automatically extracts evidence and effectively leverages the extracted evidence to improve the performance of DocRE in three stages. In the first stage, we train a relation extraction model and an evidence extraction model in a multi-task learning manner. We adopt localized context pooling (Zhou et al., 2021) in both models, which enhances the entity embedding with additional context relevant to the current entity pair. To reduce memory usage and training time, we use the same sentence representation for each relation and only train the evidence extraction model on entity pairs with at least one relation. In the second stage, we regard the extracted evidence as a pseudo document and make another set of predictions based on the pseudo document. In the last stage, we fuse the predictions based on the original document and the pseudo document using a blending layer (Wolpert, 1992). In this way, EIDER puts more attention to the important sentences extracted in the first stage, while still having access to the whole document to avoid information loss.

Extensive experiments show that EIDER outperforms the state-of-the-art methods on the public DocRED (Yao et al., 2019) dataset. The performance analysis shows that the improvement of EIDER is especially large on inter-sentence entity pairs, which are more complicated than intra-sentence pairs. We also conduct a comparison among various ablations, which validates the benefits of the joint training and evidence-centered RE module.

Contributions. (1) We propose a memory-efficient multi-task learning DocRE framework for joint relation and evidence extraction, which only requires around 14% additional memory compared to a single RE model alone. (2) We refine the inference stage of DocRE by fusing the prediction results from the original document and the extracted evidence via a blending layer, which allows the model to focus more on the important sentences with no information loss. (3) We demonstrate that EIDER achieves new state-of-the-art results on the large-scale DocRED dataset.

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 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 $r \in \mathcal{R}$ between (e_h, e_t) exists if it is expressed by any pair of their mentions, and otherwise labeled as NA. For each entity pair (e_h, e_t) that possesses a non-NA relation, we define its *evidence sentences*[2] $V_{h,t} = \{s_{v_i}\}_{i=1}^{K}$ as the subset of sentences in the document that are sufficient for human annotators to infer the relation.

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

3 Methodology

Our EIDER framework consists of three stages: joint relation and evidence extraction (Sec. 3.1), evidence-centered relation extraction (Sec. 3.2) and fusion of extraction results (Sec. 3.3). An illustration of our framework is shown in Figure 2.

3.1 Joint Relation and Evidence Extraction

In our framework, the relation extraction model and evidence extraction model share a pre-trained encoder and have their own prediction heads. Intuitively, tokens relevant to the relation are essential in both models, such as "Paul Desmarais" and "Desmarais" in the 1^{st} and 4^{th} sentences of the example shown in Figure 2. By sharing the base encoder, the two models are able to provide additional training signals for each other and hence mutually enhance each other (Ruder, 2017; Liu et al., 2019).

Base Encoder. The base encoder inputs a document and outputs the embedding of each token in it. Given a document $d = [h_l]_{l=1}^{L}$, we insert a special token "*" before and after each entity mention. We then encode the document with a pre-trained encoder (Devlin et al., 2019) to obtain the embedding of each token:

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

For each mention of an entity e_i, we first use the embedding of the start symbol "*" as its mention embedding. Then, we adopt LogSumExp pooling to obtain the embedding of entity e_i, which is a smooth version of max pooling:

$$\mathbf{e}_i = \log \sum_{j=1}^{N_{e_i}} \exp\left(\mathbf{m}_\mathbf{j}^\mathbf{i}\right), \quad (2)$$

where N_{e_i} is the number of entity e_i's mentions in the document and $\mathbf{m}_\mathbf{j}^\mathbf{i}$ is the embedding of its j^{th} mention. We compute a context embedding for each entity pair (e_h, e_t) based on the attention matrix $\mathbf{A} \in \mathbb{R}^{K \times L \times L}$ in the pre-trained encoder following Zhou et al. (2021), where K is the number of attention heads. Intuitively, tokens with high attention towards both e_h and e_t are important to both entities and hence essential to the relation. For $i \in \{h, t\}$, we first compute the attention from each token to each mention m_j^i under the k^{th} head, noted as $\mathbf{A}_{j,k}^{M,i} \in \mathbb{R}^L$. Then, we compute the attention from each token to each entity e_i by averaging attention over mentions $m_j^i \in e_i$, denoted as $\mathbf{A}_{i,k}^{E} \in \mathbb{R}^L$. The context embedding of (e_h, e_t) is then obtained by:

$$\begin{aligned} \mathbf{c}_{h,t} &= \boldsymbol{H}\mathbf{a}^{(h,t)} \\ \mathbf{a}^{(h,t)} &= \text{softmax}(\sum_{i=1}^{K} \mathbf{A}_{h,k}^{E} \cdot \mathbf{A}_{t,k}^{E}). \end{aligned} \quad (3)$$

Relation Prediction Head. We first map the embeddings of e_h and 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 $r \in \mathcal{R}$ holds between (e_h, e_t) via a bilinear function:

$$\begin{aligned} \boldsymbol{z}_h &= \tanh\left(\boldsymbol{W}_h \mathbf{e}_h + \boldsymbol{W}_{c_h} \mathbf{c}_{h,t}\right), \\ \boldsymbol{z}_t &= \tanh\left(\boldsymbol{W}_t \mathbf{e}_t + \boldsymbol{W}_{c_t} \mathbf{c}_{h,t}\right), \\ \mathrm{P}\left(r | e_h, e_t\right) &= \sigma\left(\boldsymbol{z}_h \boldsymbol{W}_r \boldsymbol{z}_t + \boldsymbol{b}_r\right), \end{aligned} \quad (4)$$

where $\boldsymbol{W}_h, \boldsymbol{W}_t, \boldsymbol{W}_{c_h}, \boldsymbol{W}_{c_t}, \boldsymbol{W}_r, \boldsymbol{b}_r$ are learnable parameters. We adopt the adaptive-thresholding loss (Zhou et al., 2021) for our RE model. Specifically, we consider a relation belong to the positive class $\mathcal{P}_T$ if it exists between the entity pair, and otherwise the negative classes $\mathcal{N}_T$. Then, we introduce a dummy relation class TH, and encourage the logits of positive classes to be larger than that of TH, while the logits of negative classes smaller than TH:

$$\begin{aligned} \mathcal{L}_{RE} = &- \sum_{r \in \mathcal{P}_T} \log\left(\frac{\exp\left(\mathbf{y}_r\right)}{\sum_{r' \in \mathcal{P}_T \cup \{\mathrm{TH}\}} \exp\left(\mathbf{y}_{r'}\right)}\right) \\ &- \log\left(\frac{\exp\left(\mathbf{y}_{\mathrm{TH}}\right)}{\sum_{r' \in \mathcal{N}_T \cup \{\mathrm{TH}\}} \exp\left(\mathbf{y}_{r'}\right)}\right), \end{aligned} \quad (5)$$

where $\mathbf{y}$ is the logits, namely, the hidden representation in the last layer before Sigmoid.

Evidence Prediction Head. The evidence extraction model predicts whether each sentence s_i is an evidence sentence of entity pair (e_h, e_t)[3]. To obtain sentence embedding $\mathbf{s}_i$, we apply a mean pooling over all the tokens in s_i: $\mathbf{s}_i = \frac{1}{|s_i|} \sum_{h_l \in s_i} (\mathbf{h}_l)$, which shows better performance than LogSumExp pooling in preliminary experiments.

Intuitively, the tokens contributing more to $\mathbf{c}_{h,t}$ are more important to both e_h and e_t, and hence may be relevant to the relation prediction. Similarly, if s_i is an evidence sentence of (e_h, e_t), the tokens in s_i would also be relevant to the relation

[3]The evidence information is available during training but is not required during inference.

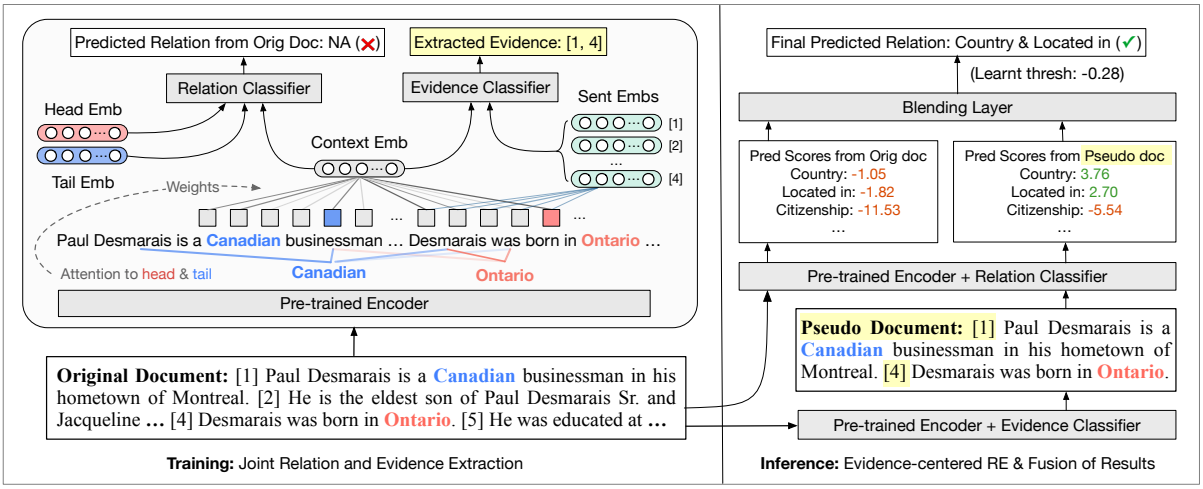

Figure 2: The overall architecture of EIDER. The left part illustrates the first stage (training) and the right shows the second and third stages (inference) of EIDER. We highlight **head entities**, **tail entities** and extracted evidences.

prediction. 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):

$$\mathrm{P}\left(s_i|e_h,e_t\right) = \sigma\left(\mathbf{s}_i \boldsymbol{W}_v \mathbf{c}_{h,t} + \boldsymbol{b}_v\right), \quad (6)$$

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.

$$\mathcal{L}_{Evi} = -\sum_{s_i \in \mathcal{D}} y_i \cdot \mathrm{P}\left(s_i|e_h,e_t\right) + (1-y_i)\cdot \log(1-\mathrm{P}\left(s_i|e_h,e_t\right)), \quad (7)$$

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

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

$$\mathcal{L} = \mathcal{L}_{RE} + \alpha \cdot \mathcal{L}_{Evi}. \quad (8)$$

α is a hyper-parameter that balances the two losses.

Inference. After the model is trained, we feed the original documents as input for relation extraction. For each entity pair (e_h, e_t), we obtain the prediction score of each relation $r \in \mathcal{R}$ by:

$$S_{h,t,r} = \begin{cases} \mathbf{y}_r - \mathbf{y}_{TH} & \text{if } \mathbf{y}_r \in \text{top\_k}(\mathbf{y}) \\ -\inf & \text{otherwise,} \end{cases} \quad (9)$$

where top_k($\mathbf{y}$) denotes the top k relations with the largest probability, which might also contain the dummy class TH.

We also extract the evidence from the joint model, noted as $V'_{h,t}$. For simplicity, we predict s_i as an evidence sentence if $\mathrm{P}\left(s_i|e_h,e_t\right) > 0.5$.

3.2 Evidence-centered Relation Extraction

Suppose we are given the ground truth evidence, that is, 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 (e_h, e_t) 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.

Since the evidence information is only available during training, we replace the evidence sentences in the construction of pseudo documents with the evidence extracted by our model, noted as $V'_{h,t}$, and obtain another set of prediction scores by Eq. 9. As the same RE model is used for both predictions without retraining and the inference speed is very fast, EIDER is still comparable to other methods with only one round of prediction.

3.3 Fusion of Extraction Results

Assuming the extracted evidence is completely accurate, directly using the extracted evidence for prediction may simplify the input, making it easier for the model to make the correct predictions. However, the quality of the extracted evidence is not perfect. Besides, the non-evidence sentences in the original document may also provide background information of the entities and is possible to contribute 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):

$$\mathrm{P}_{Fuse}(r|e_h, e_t) = \sigma(S^{(O)}_{h,t,r} + S^{(E)}_{h,t,r} - \tau), \quad (10)$$

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

$$\mathcal{L}_{Fuse} = -\sum_{d\in\mathcal{D}}\sum_{h\neq t}\sum_{r\in\mathcal{R}} y_r \cdot \mathrm{P}_{Fuse}(r|e_h, e_t) + (1 - y_r) \cdot \log(1 - \mathrm{P}_{Fuse}(r|e_h, e_t)), \quad (11)$$

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

4 Experiments

4.1 Experiment Setup

Dataset. We evaluate the effectiveness of EIDER on the DocRED benchmark (Yao et al., 2019), a large human-annotated document-level RE dataset, which consists of 3,053/1,000/1,000 documents for training/development/testing, respectively. The dataset provides evidence sentences as part of the annotation, which is not visible during inference.

Evaluation Metrics. Following prior studies (Yao et al., 2019; Zhou et al., 2021; Huang et al., 2020), we use **F1** and **Ign F1** as the main evaluation metrics for relation extraction and **Evi F1** as the metric for evidence extraction. 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 entity pairs (intra-sentence) and the latter measures the performance on inter-sentence relations where none of the entity mention pairs co-occur.

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 α 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.

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.

Relation Extraction Results. Table 1 presents the relation extraction results of EIDER and baseline models. First, we observe that EIDER outperforms the baseline methods in terms of all metrics on both the development and test sets. Furthermore, compared to ATLOP (Zhou et al., 2021), which uses the same base relation extraction model as our method, our BERT$_{\text{base}}$ model improves its performance significantly by 1.47/1.40 F1/Ign F1 and EIDER-RoBERTa$_{\text{large}}$ improves it by 1.12/1.04 F1/Ign F1 on the development set. The improvements on the test set over ATLOP are 1.17/1.11 and 1.39/1.46, respectively. Such results demonstrate the usefulness of joint extraction and integration of extracted evidence in both training and inference.

The experiment results also show that EIDER performs better than ATLOP by 1.21/2.01 (0.75/1.52) Intra/Inter F1 under BERT$_{\text{base}}$ (RoBERTa$_{\text{large}}$) on the development set. We notice that the improvement on Inter F1 is much larger than that on Intra F1, which indicates that using evidence extraction as an auxiliary task and utilizing the extracted evidence in the inference stage can largely improve the inter-sentence prediction ability of the model. GAIN-BERT$_{\text{base}}$ (Zeng et al., 2020) and ATLOP-BERT$_{\text{base}}$ have similar overall F1/Ign F1 scores, but the Inter F1 of GAIN is 0.70 higher and the Intra F1 of ATLOP is 0.16 higher. This indicates that these methods may capture the long-distance dependency between entities by directly connecting them on the graph. Although EIDER does not involve explicit multi-hop reasoning modules, it still significantly outperforms the graph-based models in terms of Inter F1, which demonstrates that the evidence-centered relation extraction also helps EIDER to capture long-distance dependencies between entities and further infer complicated relations from multiple sentences.

Evidence Extraction Results. We list the results

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., 2020)	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)	$59.11 \pm 0.14^{\dagger}$	$61.01 \pm 0.10^{\dagger}$	$67.26 \pm 0.15^{\dagger}$	$53.20 \pm 0.19^{\dagger}$	59.31	61.30
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)	$61.30 \pm 0.22^{\dagger}$	$63.15 \pm 0.21^{\dagger}$	$69.61 \pm 0.25^{\dagger}$	$55.01 \pm 0.18^{\dagger}$	61.39	63.40
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. 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	Dev F1	Test F1
E2GRE-BERT$_{base}$	47.14	-
EIDER-BERT$_{base}$	**50.71**	**51.27**
E2GRE-RoBERTa$_{large}$	-	50.50
EIDER-RoBERTa$_{large}$	**52.54**	**53.01**

Table 2: Evidence extraction results. We compare EIDER with E2GRE (Huang et al., 2020).

Ablation	Ign F1	F1	Intra F1	Inter F1
EIDER-RoBERTa$_{large}$	**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 3: Ablation studies of EIDER.

of evidence prediction in Table 2. To our knowledge, E2GRE is the only method that has reported their evidence extraction result. EIDER-BERT$_{base}$ outperforms E2GRE significantly by 3.57 on the development set and EIDER-RoBERTa$_{large}$ outperforms it by 2.51 on the test set. One possible reason is the incorporation of context vector models the dependency between tokens, leading to better performance in evidence extraction. Noted that the structure of EIDER is much simpler than E2GRE, which only makes evidence prediction on each <sentence, entity, entity> tuple. The results indicate that it is not necessary to make predictions for each <sentence, relation, entity, entity> tuple as in E2GRE.

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 3.

We first train the RE model and the evidence extraction model separately, denoted as **NoJoint**. The performance of Intra F1/Inter F1 drops by 0.50/1.04 compared to the full model. We observe that the drop in Inter F1 is more significant, which shows that the evidence and relation extraction model mutually enhance each other's ability of identifying the related context of each entity pair.

Then, we remove the extracted evidence and the original document during inference separately, denoted as **NoEvi** and **NoOrigDoc**, respectively. We observe that removing either source will lead to performance drops. The reason is probably because the original documents may contain irrelevant and noisy sentences, while using the extracted evidence sentences alone may fail to cover all of the important information in the original document. Also, when removing the extracted evidence, the drop of Inter F1 is much larger than Intra F1, Such results indicate that the extracted evidence is more effective for cross-sentence entity pairs where the

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

Table 4: The statistics of categories among the 12,323 relations in the DocRED development set.

important sentences may not be consecutive.

Additionally, we remove the blending layer and simply take the union of the two sets of results, noted as **NoBlending**. The performance drops sharply by 0.8/1.25 F1/Ign F1. It demonstrates that 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, the reason might be 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 train F1 is over 95%.

Performance Breakdown. To further analyze the performance of EIDER on different types of entity pairs, we categorize the relations into three categories: (1) *Intra*, where two entities co-occur in the same sentence, (2) *Coref*, where none of their explicit entity mention pairs co-occur, but their co-reference co-occurs, (3) *Bridge*, where the first two situations are not satisfied, but there exists a third entity whose co-reference co-occurs with both the head entity and the tail entity (e.g., "Load" in Figure 1). The statistics of each category are listed in Table 4, where the co-reference of each entity is extracted by HOI (Xu and Choi, 2020). From the statistics, 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* ≫ *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.

Memory Usage. We test the memory efficiency of EIDER. Experiments show that training EIDER-BERT$_{\text{base}}$ requires 10,916 MB on a single GTX

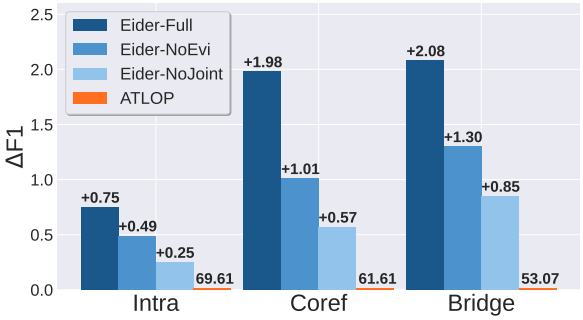

Figure 3: Performance gains in F1 by relation categories. The gains are relative to the second best baseline, ATLOP.

1080 Ti GPU, and the standalone relation extraction model consumes 9,579 MB GPU memory, indicating that our joint model incurs only ~14% GPU memory overhead. In comparison, when jointly trained with the same relation extraction model as ours, E2GRE-BERT$_{\text{base}}$ fails to run on the same GPU and requires 36,182 MB on an RTX A6000 GPU, which shows that EIDER is much more memory-efficient.

4.4 Case Studies

Table 5 shows a few example output cases of EIDER. In the first example, the extracted evidence contains the ground truth evidence, and the prediction on the pseudo document is correct. In the second example, the 6^{th} sentence is missing in the extracted evidence, but fortunately, the prediction on the original document is correct and the final result is correct. The last example shows a case where EIDER successfully predicts the evidence, but the prediction result on the pseudo document is "NA". The reason is probably because the non-evidence sentences in the original document may also provide background information of the entities and is possible to contribute to the prediction.

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, 2020). 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

Ground Truth Relation: Place of birth **Ground Truth Evidence Sentence(s)**: [3]
Document: [1] Kurt Tucholsky (9 January 1890 – 21 December 1935) was a German - Jewish journalist , satirist , and writer. [2] He also wrote under the pseudonyms Kaspar Hauser (after the historical figure), Peter Panter, Theobald Tiger and Ignaz Wrobel. [3] Born in Berlin - Moabit, he moved to Paris in 1924 and then to Sweden in 1929. [4] Tucholsky was one of the most important journalists of **...**
Extracted Evidence Sentence(s): [1, 3]
Prediction based on Orig. Document: NA **Prediction based on Extracted Evidences**: Place of Birth
Final Predicted Type: Place of Birth

Ground Truth Relation: Inception **Ground Truth Evidence Sentence(s)**: [5, 6]
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. **...**
Extracted Evidence Sentence(s): [5]
Prediction based on Orig. Document: Inception **Prediction based on Extracted Evidences**: NA
Final Predicted Type: Inception

Ground Truth Relation: Original network **Ground Truth Evidence Sentence(s)**: [1, 2]
Document: [1] "How to Save a Life" is the twenty-first episode of the eleventh season of the American television medical drama Grey's Anatomy, and is the 241st episode overall. [2] It aired on April 23, 2015 on ABC in the United States. [3] The episode was written by showrunner Shonda Rhimes and directed by Rob Hardy, making it the first episode Rhimes has written since the season eight finale "Flight". **...**
Extracted Evidence Sentence(s): [1, 2]
Prediction based on Orig. Document: Original network **Prediction based on Extracted Evidences**: NA
Final Predicted Type: Original network

Table 5: 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 ground truth evidence sentences are highlighted with yellow.

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. (2021) designs several hand-crafted rules to extract sentences that are important to the prediction. Similar to our method, Huang et al. (2020) learns a model to perform joint relation extraction and evidence extraction. However, our method uses a much simpler model structure for the evidence extraction model and hence reduces the memory usage and improves efficiency. 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 **evi**dence-**e**nhanced **R**E framework consisting of three stages: joint relation and evidence extraction, evidence-centered relation extraction, and fusion of extraction results. The joint training stage adopts a simple model structure and is memory-efficient. The relation extraction and evidence extraction model provide additional training signals for each other and mutually enhance each other. 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 the DocRED dataset, 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, Guangtao Wang, Tengyu Ma, and Jing Huang. 2020. Entity and evidence guided relation extraction for docred.

Quzhe Huang, Shengqi Zhu, Yansong Feng, Yuan Ye, Yuxuan Lai, and Dongyan Zhao. 2021. Three sentences are all you need: Local path enhanced document relation extraction.

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.

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. 2020. 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*.

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*.