DREEAM: Guiding Attention with Evidence for Improving Document-Level Relation Extraction

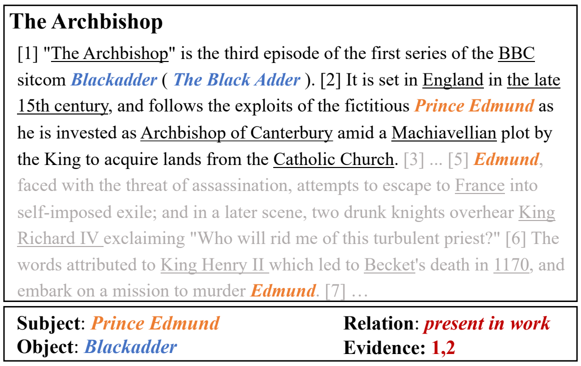
Youmi Ma and **An Wang** and **Naoaki Okazaki**
Tokyo Institute of Technology
{youmi.ma@nlp., an.wang@nlp., okazaki@}c.titech.ac.jp

Abstract

Document-level relation extraction (DocRE) is the task of identifying all relations between each entity pair in a document. Evidence, defined as sentences containing clues for the relationship between an entity pair, has been shown to help DocRE systems focus on relevant texts, thus improving relation extraction. However, evidence retrieval (ER) in DocRE faces two major issues: high memory consumption and limited availability of annotations. This work aims at addressing these issues to improve the usage of ER in DocRE. First, we propose DREEAM, a memory-efficient approach that adopts evidence information as the supervisory signal, thereby guiding the attention modules of the DocRE system to assign high weights to evidence. Second, we propose a self-training strategy for DREEAM to learn ER from automatically-generated evidence on massive data without evidence annotations. Experimental results reveal that our approach exhibits state-of-the-art performance on the DocRED benchmark for both DocRE and ER. To the best of our knowledge, DREEAM is the first approach to employ ER self-training[1].

1 Introduction

Document-level relation extraction (DocRE) has been recognized as a more realistic and challenging task compared with its sentence-level counterpart (Peng et al., 2017; Verga et al., 2018; Yao et al., 2019). In DocRE, an entity can have multiple mentions scattered throughout a document, and relationships can exist between entities in different sentences. Therefore, DocRE models are expected to apply information filtering to long texts by focusing more on sentences relevant to the current decision of relation extraction (RE) and less on irrelevant ones. To this end, existing studies retrieve *supporting evidence* (evidence hereafter, Yao et al., 2019), a set of sentences necessary for humans to identify the relation between an entity pair (Huang et al., 2021a,b; Xie et al., 2022; Xiao et al., 2022; Xu et al., 2022). As shown in Figure 1, to decide the *present in work* relation between *Prince Edmund* and *Blackadder*, reading sentences 1 and 2 should be sufficient. Although sentences 5 and 6 also mention the subject, they are irrelevant to the relation decision. Evidence of the relation triple (*Prince Edmund*, *present in work*, *Blackadder*) is thus defined as sentences 1 and 2.



The Archbishop

[1] "The Archbishop" is the third episode of the first series of the BBC sitcom ***Blackadder*** (***The Black Adder***). [2] It is set in England in the late 15th century, and follows the exploits of the fictitious ***Prince Edmund*** as he is invested as Archbishop of Canterbury amid a Machiavellian plot by the King to acquire lands from the Catholic Church. [3] ... [5] ***Edmund***, faced with the threat of assassination, attempts to escape to France into self-imposed exile; and in a later scene, two drunk knights overhear King Richard IV exclaiming "Who will rid me of this turbulent priest?" [6] The words attributed to King Henry II which led to Becket's death in 1170, and embark on a mission to murder ***Edmund***. [7] ...

Subject: ***Prince Edmund*** **Relation**: ***present in work***
Object: ***Blackadder*** **Evidence: 1,2**

Figure 1: Example document and one of the relation triples from DocRED, where the i-th sentence is marked with [i] in the beginning. Mentions in bold italics are those of subjects and objects, whereas entity mentions other than subject and object are underlined.

Despite the usefulness of evidence, automatic evidence retrieval (ER) faces two major issues. Firstly, the existing approaches for ER are memory-inefficient. Previous systems tackle ER and DocRE as separate tasks, introducing extra neural network layers to learn ER with DocRE jointly (Huang et al., 2021a; Xie et al., 2022; Xiao et al., 2022). The ER module typically involves a bilinear classifier that receives entity-pair-specific embeddings and sentence embeddings as the input. To compute the evidence score of each sentence for each entity pair, the module must walk through all (entity pair, sentence) combinations. The computations signifi-

[1] The source code is available at https://github.com/YoumiMa/dreeam

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 1971–1983
May 2-6, 2023 ©2023 Association for Computational Linguistics

cantly increase memory consumption, particularly in documents with numerous sentences and entities. Secondly, the availability of human annotations of evidence is limited. To make matters worse, gold training data for DocRE are more expensive to annotate than those for their sentence-level counterpart. Despite the difficulty of obtaining human annotations, acquiring evidence annotations at a low cost has been underexplored. Although automatically collecting silver training data for RE by distant supervision (Mintz et al., 2009; Yao et al., 2019), locating evidence for a sliver RE instance in the document is nontrivial.

This work aims at alleviating these issues to improve the usage of ER in DocRE. To reduce the memory consumption, we propose **D**ocument-level **R**elation **E**xtraction with **E**vidence-guided **A**ttention **M**echanism (DREEAM), a memory-efficient approach for incorporating DocRE with ER. We adopt ATLOP (Zhou et al., 2021), a Transformer-based DocRE system widely used in previous studies (Xie et al., 2022; Tan et al., 2022a; Xiao et al., 2022), as the backbone. Instead of introducing an external ER module, we directly guide the DocRE system to focus on evidence. Specifically, we supervise the computation of entity-pair-specific local context embeddings. The local context embedding, formed as a weighted sum among all token embeddings based on attention from the encoder, is trained to assign higher weights to evidence and lower weights otherwise.

To compensate for the shortage of evidence annotations, we propose performing ER under a weakly-supervised setting. Specifically, we design a strategy to perform self-training with DREEAM on massive, unlabeled data. The data is obtained from distant supervision (distantly-supervised data hereafter) and thus is automatically annotated with relation labels but not evidence labels. We expect the knowledge about ER learned from the human-annotated data to generate and grow on the distantly-supervised data. To enable self-training, we first adopt a teacher model trained on human-annotated data to retrieve silver evidence from distantly-supervised data. Next, we train a student model on the data for RE while learning ER from the silver evidence. The student model is further finetuned on the human-annotated data to refine its knowledge. Experiments on the DocRED benchmark (Yao et al., 2019) show that with the help of ER self-training, DREEAM exhibits state-of-the-art performance on both RE and ER.

In short, the contributions of this work are: (1) We propose DREEAM, a memory-efficient approach to incorporate evidence information into Transformer-based DocRE systems by directly guiding the attention. DREEAM does not introduce any extra trainable parameters for ER while achieving good performance on both RE and ER. (2) We propose incorporating distantly-supervised RE training with ER self-training, which improves the performance on both tasks. To the best of our knowledge, DREEAM is the first DocRE system that enables joint training of ER and RE under a weakly-supervised setting.

2 Preliminary

2.1 Problem Formulation

Given a document D containing sentences $\mathcal{X}_D = \{x_i\}_{i=1}^{|\mathcal{X}_D|}$ and entities $\mathcal{E}_D = \{e_i\}_{i=1}^{|\mathcal{E}_D|}$, DocRE aims to predict all possible relations between every entity pair. Each entity $e \in \mathcal{E}_D$ is mentioned at least once in D, with all its proper-noun mentions denoted as $\mathcal{M}_e = \{m_i\}_{i=1}^{|\mathcal{M}_e|}$. Each entity pair (e_s, e_o) can hold multiple relations, comprising a set $\mathcal{R}_{s,o} \subset \mathcal{R}$, where $\mathcal{R}$ is a pre-defined relation set. We let the set $\mathcal{R}$ include ϵ, which stands for *no-relation*. Additionally, if an entity pair (e_s, e_o) carries a valid relation $r \in \mathcal{R}\backslash\{\epsilon\}$, ER aims to retrieve the supporting evidence $\mathcal{V}_{s,r,o} \subseteq \mathcal{X}_D$ that are sufficient to predict the triplet (e_s, r, e_o).

2.2 ATLOP

This section reviews ATLOP, the backbone of our proposed method.

Text Encoding Before encoding, a special token $*$ is inserted at the beginning and the end of each entity mention. Then, tokens $\mathcal{T}_D = \{t_i\}_{i=1}^{|\mathcal{T}_D|}$ within document D are encoded with a Transformer-based pretrained language model (PLM, Vaswani et al., 2017) to obtain token embeddings and cross-token dependencies. Although the original ATLOP adopts only the last layer, this work takes the average of the last three layers[2]. Specifically, for a PLM with d hidden dimensions at each transformer layer, the token embeddings $\boldsymbol{H}$ and cross-token dependencies $\boldsymbol{A}$ are computed as:

$$\boldsymbol{H}, \boldsymbol{A} = \text{PLM}(\mathcal{T}_D), \tag{1}$$

[2] Pilot experiments showed that using the last 3 layers yields better performance than using only the last layer.

where $\boldsymbol{H} \in \mathbb{R}^{|\mathcal{T}_D| \times d}$ averages over hidden states of each token from the last three layers and $\boldsymbol{A} \in \mathbb{R}^{|\mathcal{T}_D| \times |\mathcal{T}_D|}$ averages over attention weights of all attention heads from the last three layers.

Entity Embedding The entity embedding $\boldsymbol{h}_e \in \mathbb{R}^d$ for each entity e with mentions $\mathcal{M}_e = \{m_i\}_{i=1}^{|\mathcal{M}_e|}$ is computed by collecting information from all its mentions. Specifically, `logsumexp` pooling, which has been empirically shown to be effective in previous studies (Jia et al., 2019), is adopted as: $\boldsymbol{h}_e = \log \sum_{i=1}^{|\mathcal{M}_e|} \exp(\boldsymbol{H}_{m_i})$, where $\boldsymbol{H}_{m_i}$ is the embedding of the special token $*$ at the starting position of mention m_i.

Localized Context Embedding To better utilize information from long texts, ATLOP introduces entity-pair specified localized context embeddings. Intuitively, for entity pair (e_s, e_o), tokens important to both e_s and e_o should contribute more to the embedding. The importance of each token is determined by the cross-token dependencies $\boldsymbol{A}$ obtained from Equation 1. For entity e_s, the importance of each token is computed using the cross-token dependencies of all its mentions $\mathcal{M}_{e_s}$. First, ATLOP collects and averages over the attention $\boldsymbol{A}_{m_i} \in \mathbb{R}^{|\mathcal{T}_D|}$ at the special token $*$ before each mention $m_i \in \mathcal{M}_{e_s}$ to get $\boldsymbol{a}_s \in \mathbb{R}^{|\mathcal{T}_D|}$ as the importance of each token for entity e_s. Then, the importance of each token for an entity pair (e_s, e_o), noted as $\boldsymbol{q}^{(s,o)} \in \mathbb{R}^{|\mathcal{T}_D|}$, is computed from $\boldsymbol{a}_s$ and $\boldsymbol{a}_o$ as:

$$\boldsymbol{q}^{(s,o)} = \frac{\boldsymbol{a}_s \circ \boldsymbol{a}_o}{\boldsymbol{a}_s^\top \boldsymbol{a}_o}, \tag{2}$$

where $\circ$ stands for the Hadamard product. $\boldsymbol{q}^{(s,o)}$ is thus a distribution that reveals the importance of each token for entity pair (e_s, e_o). Subsequently, ATLOP performs a localized context pooling,

$$\boldsymbol{c}^{(s,o)} = \boldsymbol{H}^\top \boldsymbol{q}^{(s,o)}, \tag{3}$$

where $\boldsymbol{c}^{(s,o)} \in \mathbb{R}^d$ is a weighted average over all token embeddings.

Relation Classification To predict the relation between entity pair (e_s, e_o), ATLOP first generates context-aware subject and object representations:

$$\boldsymbol{z}_s = \tanh(\boldsymbol{W}_s[\boldsymbol{h}_{e_s}; \boldsymbol{c}^{(s,o)}] + \boldsymbol{b}_s) \tag{4}$$

$$\boldsymbol{z}_o = \tanh(\boldsymbol{W}_o[\boldsymbol{h}_{e_o}; \boldsymbol{c}^{(s,o)}] + \boldsymbol{b}_o), \tag{5}$$

where $[\cdot;\cdot]$ represents the concatenation of two vectors and $\boldsymbol{W}_s, \boldsymbol{W}_o \in \mathbb{R}^{d \times 2d}$, $\boldsymbol{b}_s, \boldsymbol{b}_o \in \mathbb{R}^d$ are trainable parameters. Then, a bilinear classifier[3] is applied on the context-aware representations to compute the relation scores $\boldsymbol{y}^{(s,o)} \in \mathbb{R}^{|\mathcal{R}|}$:

$$\boldsymbol{y}^{(s,o)} = \boldsymbol{z}_s^\top \mathsf{W}_r \boldsymbol{z}_o + \boldsymbol{b}_r, \tag{6}$$

where $\mathsf{W}_r \in \mathbb{R}^{|\mathcal{R}| \times d \times d}$ and $\boldsymbol{b}_r \in \mathbb{R}^{|\mathcal{R}|}$ are trainable parameters. The probability that relation $r \in \mathcal{R}$ holds between entity e_s and e_o is thus $\mathrm{P}(r|s,o) = \sigma(y_r^{(s,o)})$, where σ is the sigmoid function.

Loss Function ATLOP proposes Adaptive Thresholding Loss (ATL) that learns a dummy threshold class TH during training, serving as a dynamic threshold for each relation class $r \in \mathcal{R}$. For each entity pair (e_s, e_o), ATL forces the model to yield scores above TH for positive relation classes $\mathcal{R}_P$ and scores below TH for negative relation classes $\mathcal{R}_N$, formulated as below:

$$\begin{aligned}\mathcal{L}_{\text{RE}} = &-\sum_{s \neq o} \sum_{r \in \mathcal{R}_P} \frac{\exp(y_r^{(s,o)})}{\sum_{r' \in \mathcal{R}_P \cup \{\text{TH}\}} \exp(y_{r'}^{(s,o)})} \\ &- \frac{\exp(y_{\text{TH}}^{(s,o)})}{\sum_{r' \in \mathcal{R}_N \cup \{\text{TH}\}} \exp(y_{r'}^{(s,o)})}.\end{aligned} \tag{7}$$

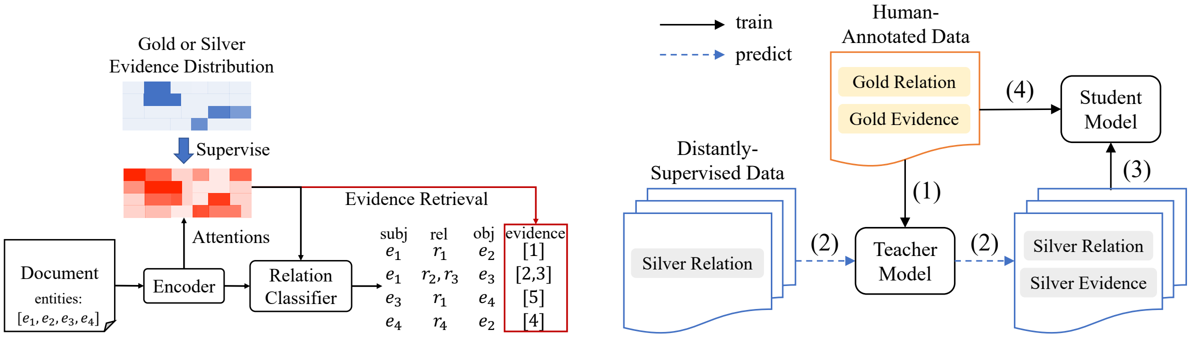
The idea of setting a threshold class is similar to the Flexible Threshold (Chen et al., 2020).

3 Proposed Method: DREEAM

To perform information filtering, ATLOP computes a localized context embedding based on attention weights from the Transformer-based encoder. The rationale is that cross-token dependencies are encoded as attention weights in Transformer layers. In this work, we propose DREEAM to enhance ATLOP with evidence. In addition to the automatically-learned cross-token dependencies, the attention modules are supervised to concentrate more on evidence sentences and less on others.

DREEAM can be employed for both supervised and self-training, sharing the same architecture with different supervisory signals, as shown in Figure 2 (a). Inspired by Tan et al. (2022a), we propose a pipeline to enable self-training of ER, with the data flow shown in Figure 2 (b). First, we train a teacher model on human-annotated data with gold relations and evidence labels. Next, we apply the trained teacher model to predict silver evidence for

[3] In practice, a grouped bilinear classifier (Zheng et al., 2019) is applied to save memory.



(a) Model architecture of DREEAM.

(b) Information flow of self-training using DREEAM.

Figure 2: Model architecture and the information flow during self-training. In (a), gold/silver evidence distributions come from human-annotations/the teacher model. In (b), arrows represent the direction of knowledge transfer.

the distantly-supervised data. Then, we train a student model on the distantly-supervised data, with ER supervised by the silver evidence. Finally, we finetune the student model on the human-annotated data to refine its knowledge. The rest of this section introduces the training processes of the teacher and student models, followed by the inference process.

3.1 Teacher Model

For each entity pair (s, o), we guide $\boldsymbol{q}^{(s,o)}$ with an evidence distribution to help generate an evidence-centered localized context embedding. While $\boldsymbol{q}^{(s,o)}$ yields token-level importance for e_s and e_o, we can obtain only sentence-level evidence from human annotations, as shown in Figure 1. To alleviate this gap, we sum the weight of each token within a sentence. Specifically, for a sentence $x_i \in \mathcal{X}_D$ consisting of tokens $t_{\text{START}(x_i)}, \dots, t_{\text{END}(x_i)}$, we obtain the sentence-level importance as:

$$p_i^{(s,o)} = \sum_{j=\text{START}(x_i)}^{\text{END}(x_i)} q_j^{(s,o)}. \tag{8}$$

Collecting the importance of all sentences yields a distribution $\boldsymbol{p}^{(s,o)} \in \mathbb{R}^{|\mathcal{X}_D|}$ that expresses the importance of each sentence within the document.

We further supervise $\boldsymbol{p}^{(s,o)}$ for each entity pair (e_s, e_o) using a human-annotated evidence distribution computed from gold evidence. First, we define a binary vector $\boldsymbol{v}^{(s,r,o)} \in \mathbb{R}^{|\mathcal{X}_D|}$ for each valid relation label $r \in \mathcal{R}_{s,o} \subset \mathcal{R}\backslash\{\epsilon\}$ that records whether each sentence $x_i \in \mathcal{X}_D$ is evidence of the relation triple (e_s, r, e_o) or not. For example, if x_i is evidence of (e_s, r, e_o), then $v_i^{(s,r,o)}$ is set to 1, and otherwise 0.

Next, we marginalize all valid relations and normalize the marginalized vector to obtain $\boldsymbol{v}^{(s,o)}$:

$$\boldsymbol{v}^{(s,o)} = \frac{\sum_{r \in \mathcal{R}_{s,o}} \boldsymbol{v}^{(s,r,o)}}{\sum_{r \in \mathcal{R}_{s,o}} \mathbf{1}^\top \boldsymbol{v}^{(s,r,o)}}, \tag{9}$$

where $\mathbf{1} = (1, 1, \dots, 1) \in \mathbb{R}^{|\mathcal{X}_D|}$ is an all-ones vector. The rationale behind Equation 9 is that modules before the relation classifier are not explicitly aware of specific relation types. We thus guide attention modules within the encoder to produce relation-agnostic token dependencies.

Loss Function Our purpose is to guide $\boldsymbol{p}^{(s,o)}$ with human evidence $\boldsymbol{v}^{(s,o)}$ to generate an evidence-focused localized context embedding $\boldsymbol{c}^{(s,o)}$. To achieve this, we train the model with Kullback-Leibler (KL) Divergence loss, minimizing the statistical distance between $\boldsymbol{p}^{(s,o)}$ and $\boldsymbol{v}^{(s,o)}$:

$$\mathcal{L}_{\text{ER}}^{\text{gold}} = -D_{\text{KL}}(\boldsymbol{v}^{(s,o)} || \boldsymbol{p}^{(s,o)}). \tag{10}$$

During training, we balance the effect of ER loss with RE loss using a hyper-parameter λ:

$$\mathcal{L}^{\text{gold}} = \mathcal{L}_{\text{RE}} + \lambda \mathcal{L}_{\text{ER}}^{\text{gold}}. \tag{11}$$

3.2 Student Model

We employ the system trained on human-annotated data as a teacher model to support ER self-training on massive data. The data, obtained from relation distant-supervision (Mintz et al., 2009), contains noisy labels for RE but no information for ER. We train a student model on the data. Supervision of the student model, similar to that of the teacher model, consists of two parts: an RE binary cross-entropy loss and an ER self-training loss.

In general, predictions from the teacher model are adopted as the supervisory signal for ER training. First, we let the teacher model infer on the distantly-supervised data, thereby yielding an evidence distribution over tokens $\hat{\boldsymbol{q}}^{(s,o)}$ for each entity pair (e_s, e_o). Next, we train the student model to reproduce $\hat{\boldsymbol{q}}^{(s,o)}$ for each entity pair (e_s, e_o).

Loss Function The objectives of self-training are identical to those of supervised training. We train ER of the student model using a KL-divergence loss similar to Equation 10:

$$\mathcal{L}_{\text{ER}}^{\text{silver}} = -D_{\text{KL}}(\hat{\boldsymbol{q}}^{(s,o)} || \boldsymbol{q}^{(s,o)}), \qquad (12)$$

where $\boldsymbol{q}^{(s,o)}$ is the student model's evidence distribution over tokens regarding entity pair (e_s, e_o), computed from Equation 2.

There are two notable differences between $\mathcal{L}_{\text{ER}}^{\text{silver}}$ and $\mathcal{L}_{\text{ER}}^{\text{gold}}$. Firstly, the supervisory signal of $\mathcal{L}_{\text{ER}}^{\text{gold}}$ is sentence-level, while that of $\mathcal{L}_{\text{ER}}^{\text{silver}}$ is token-level. The gap results from the availability of token-level evidence distributions. On human-annotated data, it is untrivial to obtain token-level evidence distributions from sentence-level annotations. On distantly-supervised data, however, the evidence distribution over tokens can be easily obtained from predictions of the teacher model. We thus adopt token-level evidence distributions to provide supervision from a micro perspective for ER self-training. Secondly, $\mathcal{L}_{\text{ER}}^{\text{gold}}$ is computed only on entity pairs with valid relation(s), while $\mathcal{L}_{\text{ER}}^{\text{silver}}$ is computed over all entity pairs within the document. The design choice is based on the low reliability of relation labels on distantly-supervised data. As these relation labels are collected automatically, it is possible that some of the annotated relations are irrelevant to the document. Therefore, it is hard to tell which relations are valid and which are not from the automatic annotations. For this reason, we compute the loss from all entity pairs to prevent missing important instances.

The overall loss is balanced by the same hyper-parameter λ in Equation 11:

$$\mathcal{L}^{\text{silver}} = \mathcal{L}_{\text{RE}} + \lambda \mathcal{L}_{\text{ER}}^{\text{silver}}. \qquad (13)$$

After training on the distantly-supervised data, the student model is further finetuned using the human-annotated data to refine its knowledge about DocRE and ER with reliable supervisory signals.

3.3 Inference

Following Zhou et al. (2021), we apply adaptive thresholding to obtain RE predictions, selecting relation classes with scores higher than the threshold class as predictions. For ER, we apply static thresholding and choose sentences with importance higher than a pre-defined threshold as evidence.

Statistics	Human	Distant
# of documents	3,053/998/1,000	101,873
# of relation types	97	97
Avg. # of ent. per doc.	19.5	19.3
Avg. # of sent. per doc.	8.0	8.1
Avg. # of ment. per ent.	1.3	1.3
Avg. # of rel. per doc.	12.5	14.8
Avg. # of evi. per rel.	1.6	-

Table 1: Data statistics of DocRED. *Human* stands for human-annotated data and *Distant* stands for distantly-supervised data. The abbreviations *doc.*, *ent.*, *sent.*, *ment.*, *rel.*, and *evi.* stand for document, entity, sentence, mention, relation, and evidence sentences, respectively.

We further incorporate the **inference-stage fusion** strategy proposed by Xie et al. (2022). Specifically, for each predicted relation triple (e_s, r, e_o) associated with evidence prediction $\mathcal{V}_{s,r,o}$, we create a pseudo-document $\hat{D}_{s,r,o}$ by collecting only evidence sentences $x_i \in \mathcal{V}_{s,r,o}$. Then, we feed pseudo-documents into the trained model to re-score the relation triples. To aggregate the predictions from the pseudo-documents and the entire document, we adopt a blending layer that contains only one parameter τ representing a threshold. Each triple (e_s, r, e_o) is chosen as the final prediction only if the summation of its scores on the entire document and pseudo-documents is higher than τ. We adjust τ to minimize the binary cross-entropy loss of RE on the development set. For more details, we refer the readers to the original paper (Xie et al., 2022).

4 Experiments

To evaluate DREEAM, we conduct experiments under supervised and weakly-supervised settings.

4.1 Setting

Dataset We conduct experiments on DocRED (Yao et al., 2019)[4], the largest dataset for DocRE with human annotations. As shown in Table 1, DocRED contains a small portion of human-annotated data and a large portion of distantly-supervised data made by aligning Wikipedia articles with the Wikidata knowledge base (Vrandečić and Krötzsch, 2014). In this work, we directly adopt the distantly-supervised data provided in DocRED.

Configuration We implement DREEAM based on Hugging Face's Transformers (Wolf et al., 2020).

[4] https://github.com/thunlp/DocRED

			Dev			Test	
Method	PLM	Ign F1	F1	Evi F1	Ign F1	F1	Evi F1
(a) without Distantly-Supervised Data							
SSAN (Xu et al., 2021a)	BERT$_{\text{base}}$	57.03	59.19	-	55.84	58.16	-
ATLOP (Zhou et al., 2021)	BERT$_{\text{base}}$	59.22	61.09	-	59.31	61.30	-
E2GRE (Huang et al., 2021a)	BERT$_{\text{base}}$	55.22	58.72	47.12	-	-	-
DocuNet (Zhang et al., 2021)	BERT$_{\text{base}}$	59.86	61.83	-	59.93	61.86	-
EIDER (Xie et al., 2022)	BERT$_{\text{base}}$	60.51	62.48	50.71	60.42	62.47	51.27
SAIS (Xiao et al., 2022)	BERT$_{\text{base}}$	59.98	62.96	53.70	60.96	62.77	52.88
DREEAM (teacher)	BERT$_{\text{base}}$	59.60$_{\pm 0.15}$	61.42$_{\pm 0.15}$	52.08$_{\pm 0.10}$	59.12	61.13	51.71
+ Inference-stage Fusion		60.51$_{\pm 0.06}$	62.55$_{\pm 0.06}$		60.03	62.49	
SSAN (Xu et al., 2021a)	RoBERTa$_{\text{large}}$	60.25	62.08	-	59.47	61.42	-
ATLOP (Zhou et al., 2021)	RoBERTa$_{\text{large}}$	61.32	63.18	-	61.39	63.40	-
DocuNet (Zhang et al., 2021)	RoBERTa$_{\text{large}}$	62.23	64.12	-	62.39	64.55	-
EIDER (Xie et al., 2022)	RoBERTa$_{\text{large}}$	62.34	64.27	52.54	62.85	64.79	53.01
SAIS (Xiao et al., 2022)	RoBERTa$_{\text{large}}$	62.23	65.17	55.84	63.44	65.11	55.67
DREEAM (teacher)	RoBERTa$_{\text{large}}$	61.71$_{\pm 0.09}$	63.49$_{\pm 0.10}$	54.15$_{\pm 0.11}$	61.62	63.55	54.01
+ Inference-stage Fusion		62.29$_{\pm 0.23}$	64.20$_{\pm 0.23}$		62.12	64.27	
(b) with Distantly-Supervised Data							
KD-DocRE (Tan et al., 2022a)	BERT$_{\text{base}}$	63.38	64.81	-	62.56	64.76	-
DREEAM (student)	BERT$_{\text{base}}$	63.47$_{\pm 0.02}$	65.30$_{\pm 0.03}$	**55.68**$_{\pm 0.04}$	63.31	65.30	**55.43**
+ Inference-Stage Fusion		**63.92**$_{\pm 0.02}$	**65.83**$_{\pm 0.04}$		**63.73**	**65.87**	
SSAN (Xu et al., 2021a)	RoBERTa$_{\text{large}}$	63.76	65.69	-	63.78	65.92	-
KD-DocRE (Tan et al., 2022a)	RoBERTa$_{\text{large}}$	65.27	67.12	-	65.24	67.28	-
DREEAM (student)	RoBERTa$_{\text{large}}$	65.24$_{\pm 0.07}$	67.09$_{\pm 0.07}$	**57.55**$_{\pm 0.07}$	65.20	67.22	**57.34**
+ Inference-Stage Fusion		**65.52**$_{\pm 0.07}$	**67.41**$_{\pm 0.04}$		**65.47**	**67.53**	

Table 2: Evaluation results on development and test sets of DocRED, with best scores **bolded**. The scores of existing methods are borrowed from corresponding papers. We group the methods first by whether they utilize the distantly-supervised data or not, then by the PLM encoder.

Following previous work, we evaluate the performance of DREEAM using BERT$_{\text{base}}$ (Devlin et al., 2019) and RoBERTa$_{\text{large}}$ (Liu et al., 2019) as the PLM encoder. The parameter for balancing ER loss with RE loss is set to 0.1 for BERT$_{\text{base}}$ and 0.05 for RoBERTa$_{\text{large}}$ when training both the teacher and the student model, chosen based on a grid search from $\lambda \in \{0.05, 0.1, 0.2, 0.3\}$. We train and evaluate DREEAM on a single Tesla V100 16GB GPU when utilizing BERT$_{\text{base}}$ and on a single NVIDIA A100 40GB GPU when utilizing RoBERTa$_{\text{large}}$. Details about hyper-parameters and running time are provided in Appendix A.

Evaluation During inference, sentences x_i with $p_i > 0.2$ computed from Equation 8 are retrieved as evidence. For the evaluation, we adopt official evaluation metrics of DocRED (Yao et al., 2019): Ign F1 and F1 for RE and Evi F1 for ER. Ign F1 is measured by removing relations present in the annotated training set from the development and test sets. We train our system five times, initialized with different random seeds, and report the average scores and standard error of these runs.

4.2 Main Results

Table 2 lists the performance of the proposed and existing methods. We select the best-performing model on the development set to make predictions on the test set and submit the predictions to CodaLab for evaluation[5].

Performance of the Student Model Table 2 shows that the student model outperforms existing systems on RE by utilizing the distantly-supervised data. In particular, when adopting BERT$_{\text{base}}$ as the PLM encoder, DREEAM performs better than KD-DocRE (Tan et al., 2022a), the previous state-of-the-art system, by 0.6/1.0 points on Ign F1/F1 for the development set. On the test set, the improvement reaches 1.1 F1 points on both Ign F1 and F1. Notably, DREEAM utilizing BERT$_{\text{base}}$ even performs comparably with SSAN utilizing RoBERTa$_{\text{large}}$ under the weakly-supervised setting (Xu et al., 2021a). When adopting RoBERTa$_{\text{large}}$ as the PLM encoder, the advantage of DREEAM remains on both development and test sets. These results support our hypothesis that ER self-training improves RE, which has not been demonstrated by any previous work.

Performance of the Teacher Model The upper half of Table 2 shows that the teacher model trained

[5] `https://codalab.lisn.upsaclay.fr/competitions/365`. Submissions under username `kgmr15` are from this work.

Setting	Ign F1	F1	Evi F1
(a) Teacher Model			
DREEAM	**59.60**$_{\pm 0.15}$	**61.42**$_{\pm 0.15}$	**52.08**$_{\pm 0.10}$
w/o ER training	59.21$_{\pm 0.19}$	61.01$_{\pm 0.20}$	42.79$_{\pm 1.65}$
(b) Student Model			
DREEAM	**63.47**$_{\pm 0.02}$	65.30$_{\pm 0.03}$	**55.68**$_{\pm 0.04}$
w/o ER self-training	61.96$_{\pm 0.39}$	63.77$_{\pm 0.44}$	53.72$_{\pm 0.43}$
w/o ER fine-tuning	63.34$_{\pm 0.02}$	**65.50**$_{\pm 0.02}$	55.27$_{\pm 0.05}$
w/o both	62.13$_{\pm 0.07}$	63.82$_{\pm 0.08}$	47.13$_{\pm 0.12}$

Table 3: Ablation studies evaluated on the DocRED development set.

on human-annotated data exhibits comparable performance to EIDER (Xie et al., 2022) on both RE and ER. Although there is a performance gap between DREEAM and SAIS, we attribute it to the difference in supervisory signals. While DREEAM incorporates RE with only relation-agnostic ER, SAIS is trained under three more tasks: coreference resolution, entity typing, and relation-specific ER (Xiao et al., 2022). These extra supervisory signals possibly contribute to the high performance of SAIS. Apart from the performance, our method has a critical advantage over previous ER-incorporated DocRE systems in memory efficiency. We provide a detailed discussion in Section 4.4.

Effectiveness of ER Self-Training Additionally, we observe that the student model leads the existing systems by a large margin on ER. As the first approach enabling weakly-supervised ER training, DREEAM utilizes considerable amounts of data without evidence annotation via self-training. The experimental results reveal that DREEAM improves over the state-of-the-art supervised approaches by approximately 2.0 points on Evi F1. Therefore, we conclude that our approach to ER self-training succeeds in acquiring evidence knowledge from the relation-distantly-supervised data with no evidence annotation.

4.3 Ablation Studies

This subsection investigates the effect of evidence-guided attention by ablation studies. All subsequent experiments adopt BERT$_{\text{base}}$ as the PLM encoder. We report scores without the inference-stage fusion strategy (Xie et al., 2022).

Teacher Model Firstly, we examine how guiding attention with evidence helps RE training on human-annotated data. We train a variant of our teacher model without ER training and evaluate its performance on the DocRED development set. In general, disabling ER training reduces the model to a baseline similar to ATLOP (Zhou et al., 2021)[6].

As presented in Table 3 (a), the RE performance of our system decreases without ER training. This observation supports the hypothesis that guiding attention with evidence is beneficial to improving RE. We further visualize the token importance $\boldsymbol{q}^{(s,o)}$ for some instances to investigate the effect of evidence-guided training and find that our method succeeds in guiding the attention to focus more on relevant contexts. The details can be found in Appendix B.

Additionally, we retrieve evidence from the ER-disabled model as sentences with importance higher than the pre-defined threshold. By doing so, we find that the Evi F1 is not far from its evidence-aware counterpart. This observation indicates that ER is a task highly coupled with RE.

Student Model Next, we investigate the student model trained on distantly-supervised data and fine-tuned on human-annotated data. The aim is to examine the effect of guiding attention with evidence at various stages of training. To this end, we remove ER supervisory signals from the student model during the training on distantly-supervised and human-annotated data. The baseline excludes ER supervision from both stages, pre-trained on distantly-supervised data and then finetuned on human-annotated data for only RE.

As shown in Table 3 (b), DREEAM without ER self-training performs comparably to the baseline, while DREEAM without ER fine-tuning performs comparably to the original model with no ablations. These results indicate that ER self-training is more essential than ER fine-tuning for the student model. On the one hand, we observe that disabling ER self-training on massive data causes a huge loss of evidence knowledge that cannot be recovered by finetuning on the much smaller evidence-annotated dataset. On the other hand, we can conclude that DREEAM succeeds in retrieving evidence knowledge from the data without any evidence annotation, demonstrating the effectiveness of our ER self-training strategy.

4.4 Memory Efficiency

This subsection discusses the memory inefficiency issue in previous ER approaches and shows how DREEAM solves it. Previous approaches regard

[6] The difference between ATLOP and our baseline is that our baseline utilizes the last three layers of PLM to obtain embeddings, whereas ATLOP adopts only the final layer.

Method	Memory (GiB)	Trainable Params. (M)
(a) without ER Module		
ATLOP (Zhou et al., 2021)	10.8	115.4
SSAN (Xu et al., 2021a)	6.9	113.5
KD-DocRE (Tan et al., 2022a)	15.2	200.1
(b) with ER Module		
EIDER (Xie et al., 2022)	43.1	120.2
SAIS (Xiao et al., 2022)	46.2	118.0
DREEAM (proposed)	11.8	115.4

Table 4: Memory consumption and the number of trainable parameters of DREEAM and existing methods.

ER as a separate task from RE that requires extra neural network layers to solve (Huang et al., 2021a; Xie et al., 2022; Xiao et al., 2022). To perform ER, all of them introduce a bilinear evidence classifier that receives an entity-pair-specific embedding and a sentence embedding as inputs. For example, EIDER computes an evidence score for sentence x_i with regard to entity pair (e_s, e_o) as below:

$$\mathrm{P}(x_i|e_s, e_o) = \sigma(\boldsymbol{x}_i \mathsf{W} \boldsymbol{c}^{(s,o)} + \boldsymbol{b}), \qquad (14)$$

where $\boldsymbol{x}_i$ is a sentence embedding, $\boldsymbol{c}^{(s,o)}$ is the localized context embedding computed from Equation 3, W and $\boldsymbol{b}$ are trainable parameters. EIDER and other existing systems thus need to compute over all combinations of (sentence, entity pair). Specifically, consider a document D with n sentences $\mathcal{X}_D = \{x_1, x_2, \dots, x_n\}$ and m entities $\mathcal{E}_D = \{e_1, e_2, \dots, e_m\}$, yielding $m \times (m-1)$ entity pairs. To obtain evidence scores, EIDER must perform bilinear classification $n \times m \times (m-1)$ times via Equation 14, resulting in huge memory consumption. In contrast, DREEAM takes the summations of attention weights over tokens as evidence scores, thus introducing neither new trainable parameters nor expensive matrix computations. Hence, we see that DREEAM is more memory-efficient than its competitors.

Table 4 summarizes the memory consumption and the number of trainable parameters when utilizing BERT$_{\text{base}}$ as the PLM encoder for existing and proposed methods. Values are measured when training the systems using the corresponding official repositories with a batch size of four[7]. We observe that the memory consumption of DREEAM is only 27.4% of EIDER and 25.5% of SAIS. Notably, DREEAM also consumes less memory than KD-DocRE, underscoring the memory efficiency of our proposed method.

Statistics	DocRED	Re-DocRED
# rel.	38,180	85,932
# rel. w/o evi.	1,421 (3.7%)	38,672 (45.0%)

Table 5: Statistics of relation triples in the training set of DocRED and Re-DocRED. *rel.* stands for relation triples and *rel. w/o evi.* stands for relation triples without evidence sentences.

Method	Ign F1	F1
(a) without Distantly-Supervised Data		
ATLOP (Zhou et al., 2021)	76.82	77.56
DocuNet (Zhang et al., 2021)	77.26	77.87
KD-DocRE (Tan et al., 2022a)	77.60	78.28
DREEAM	77.34$_{\pm 0.19}$	77.94$_{\pm 0.15}$
+ Inference-Stage Fusion	79.66$_{\pm 0.39}$	80.73$_{\pm 0.38}$
(b) with Distantly-Supervised Data		
ATLOP (Zhou et al., 2021)	78.52	79.46
DocuNet (Zhang et al., 2021)	78.52	79.46
KD-DocRE (Tan et al., 2022a)	80.32	81.04
DREEAM	78.67$_{\pm 0.17}$	79.35$_{\pm 0.18}$
+Inference-Stage Fusion	**80.39**$_{\pm 0.03}$	**81.44**$_{\pm 0.04}$

Table 6: Evaluation results on the test set of Re-DocRED, with best scores **bolded**. PLM encoder is aligned to RoBERTa-large. The scores of existing methods are borrowed from Tan et al. (2022b).

4.5 Performance on Re-DocRED

Although DocRED is a widely used benchmark, recent works have pointed out that annotations of the dataset are incomplete (Huang et al., 2022; Xie et al., 2022; Tan et al., 2022b). To paraphrase, many relation triples in DocRED are missing in human annotations, biasing the dataset with many false negatives. Tan et al. (2022b) thus proposed Re-DocRED, a more reliable benchmark for DocRE that revises DocRED to alleviate the false negative issue. In this subsection, we evaluate DREEAM on Re-DocRED to verify the soundness of our proposed method.

Similar to Section 4.2, we conducted experiments under two different settings: (a) a fully-supervised setting without distantly-supervised data and (b) a weakly-supervised setting utilizing distantly-supervised data. Notably, Re-DocRED introduces new relation triples without providing accurate evidence sentences. As shown in Table 5, compared with DocRED, the training set of Re-DocRED contains much more relation triples without evidence sentences. DREEAM trained on Re-DocRED could thus be inaccurate on ER, biased by the considerable amount of missing evidence. Therefore, during ER self-training of the student model, we adopt token evidence distributions predicted by a teacher model trained on DocRED as

[7] The value of EIDER is different from the original paper because we enable ER evaluations during training.

the supervisory signal. The student model is further finetuned on Re-DocRED to obtain more reliable knowledge about RE.

Table 6 compares the performance of DREEAM against existing methods. We observe that DREEAM outperforms existing methods under both the fully-supervised setting and the weakly-supervised setting. The observation indicates the soundness of our proposed method.

5 Related Work

DocRE Recent work has extended the scope of relation extraction task from sentence to document (Peng et al., 2017; Quirk and Poon, 2017; Yao et al., 2019). Compared with its sentence-level counterpart, DocRE is a more realistic and challenging setting, aiming at extracting both intra-sentence and inter-sentence relations. Although commonly-used benchmarks for DocRE include DocRED (Yao et al., 2021), CDR (Li et al., 2016) and GDA (Wu et al., 2019), only DocRED contains evidence annotation and massive pre-processed data obtained from relation distant supervision. Therefore, we adopt DocRED as our test bed.

Transformer-based DocRE Modeling DocRE with a Transformer-based system has been a popular and promising approach, outperforming its graph-based counterparts (Zeng et al., 2020, 2021; Xu et al., 2021b). One of the major topics of these systems is a better utilization of long-distance token dependencies captured by the PLM encoder. Zhang et al. (2021) formulate DocRE as a semantic segmentation task and introduce a U-Net (Ronneberger et al., 2015) on top of the PLM encoder to capture local and global dependencies between entities. Zhou et al. (2021) propose localized contextual pooling to focus on tokens relevant to each entity pair. Based on their work, Tan et al. (2022a) adopt an axial attention module to perform two-hop reasoning and capture the dependencies between relation triples. These designs provide no supervision on token dependencies, expecting the model to capture them implicitly during training. In contrast, we provide explicit supervision for token dependencies by utilizing evidence information.

ER in DocRE This study is not the first to incorporate evidence information into DocRE. Huang et al. (2021b) first report that heuristically selecting evidence sentences boosts the performance of DocRE systems. Huang et al. (2021a), Xie et al. (2022) and Xiao et al. (2022) train neural classifiers to automatically retrieve evidence together with RE. However, we perform ER with neither heuristic rules nor neural classifiers. Furthermore, our approach can be used for ER self-training on data without evidence annotations.

Distant Supervision Distant supervision has been widely adopted as a technique to generate automatically-labeled data for RE (Mintz et al., 2009; Quirk and Poon, 2017; Xiao et al., 2020). The method assumes that if a sentence contains an entity pair that participates in a known relation in a knowledge base (KB), the sentence probably expresses that relation. Thus unlabeled text can be aligned with a KB using entities as anchors, with each match distantly supervised by the relation described in the KB. Yao et al. (2019) apply the technique to annotate relations in documents automatically. In this work, we directly adopt those documents for ER self-training.

6 Conclusion

We have introduced methods to improve the usage of ER in DocRE. First, we propose DREEAM, a memory-efficient method to reduce the computation cost of ER. Unlike existing approaches that train an evidence classifier for ER, DREEAM directly supervises the attention to concentrate more on evidence than on others. Next, we propose to employ DREEAM in a weakly-supervised setting to compensate for the shortage of human annotations. Instead of gold evidence annotated by humans, we adopt evidence predictions from a teacher model trained on human-annotated data as the supervisory signal to realize ER self-training on unlabeled data. Experiments on the DocRED benchmark show that DREEAM exhibits state-of-the-art performance on both RE and ER, with the help of weakly-supervised training on data obtained from distant supervision of relations. Compared with existing approaches, DREEAM performs ER with zero trainable parameters introduced, thereby reducing the memory usage to 27% or less. The soundness of DREEAM is confirmed by conducting experiments on Re-DocRED, a revised version of DocRED.

In the future, we plan to transfer the evidence knowledge of DREEAM trained on DocRED to other DocRE datasets.

Limitations

A major limitation of this work is that our method can only retrieve relation-agnostic evidence. Unlike Xiao et al. (2022), DREEAM cannot specify evidence sentences for each relation label. Therefore, when an entity pair holds multiple relations, DREEAM retrieves the same evidence regardless of the relation type, even though the evidence may be correct for some of the relations but not for others.

Ethics Statement

In this work, we have proposed a method for incorporating ER into DocRE. Our approach directly supervises the weights of attention modules within a Transformer-based PLM encoder. Inside the research community, we hope our approach can provide a new viewpoint on the explainability of document-level relation extraction systems. Furthermore, a better DocRE system will benefit the research on other tasks, such as question answering and reading comprehension. In the real world, a DocRE system with good performance can help extract useful information from unstructured text, reducing human efforts and expenses. Furthermore, as our method is memory-efficient, it is also friendly to the environment.

We also have demonstrated a use case of our method in ER self-training, utilizing massive data obtained from relation distant-supervision. Although in this work, we directly adopt the data provided by Yao et al. (2019), it is possible to extend the scale of data by utilizing numerous unstructured texts. Utilizing a wide range of unstructured texts may expose our system to the risk of vulnerable data, potentially biasing our system in the wrong direction. To mitigate the problem, we encourage performing data pre-processing to detect and remove harmful contents before training.

Acknowledgements

This paper is based on results obtained from a project, JPNP18002, commissioned by the New Energy and Industrial Technology Development Organization (NEDO). We appreciate Marco Cognetta for the generous help and valuable discussions.

References

Tongfei Chen, Yunmo Chen, and Benjamin Van Durme. 2020. Hierarchical entity typing via multi-level learning to rank. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8465–8475, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kevin Huang, Peng Qi, Guangtao Wang, Tengyu Ma, and Jing Huang. 2021a. Entity and evidence guided document-level relation extraction. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 307–315, Online. Association for Computational Linguistics.

Quzhe Huang, Shibo Hao, Yuan Ye, Shengqi Zhu, Yansong Feng, and Dongyan Zhao. 2022. Does recommend-revise produce reliable annotations? an analysis on missing instances in DocRED. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6241–6252, Dublin, Ireland. Association for Computational Linguistics.

Quzhe Huang, Shengqi Zhu, Yansong Feng, Yuan Ye, Yuxuan Lai, and Dongyan Zhao. 2021b. Three sentences are all you need: Local path enhanced document relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 998–1004, Online. Association for Computational Linguistics.

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: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3693–3704, Minneapolis, Minnesota. Association for Computational Linguistics.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016. Baw068.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.

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

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham. Springer International Publishing.

Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. 2022a. Document-level relation extraction with adaptive focal loss and knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1672–1681, Dublin, Ireland. Association for Computational Linguistics.

Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. 2022b. Revisiting docred – addressing the false negative problem in relation extraction.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Patrick Verga, Emma Strubell, and Andrew McCallum. 2018. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 872–884, New Orleans, Louisiana. Association for Computational Linguistics.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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*, pages 272–284, Cham. Springer International Publishing.

Chaojun Xiao, Yuan Yao, Ruobing Xie, Xu Han, Zhiyuan Liu, Maosong Sun, Fen Lin, and Leyu Lin. 2020. Denoising relation extraction from document-level distant supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3683–3688, Online. Association for Computational Linguistics.

Yuxin Xiao, Zecheng Zhang, Yuning Mao, Carl Yang, and Jiawei Han. 2022. SAIS: Supervising and augmenting intermediate steps for document-level relation extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2395–2409, Seattle, United States. Association for Computational Linguistics.

Yiqing Xie, Jiaming Shen, Sha Li, Yuning Mao, and Jiawei Han. 2022. Eider: Empowering document-level relation extraction with efficient evidence extraction and inference-stage fusion. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 257–268, Dublin, Ireland. Association for Computational Linguistics.

Benfeng Xu, Quan Wang, Yajuan Lyu, Yong Zhu, and Zhendong Mao. 2021a. Entity structure within and throughout: Modeling mention dependencies for document-level relation extraction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14149–14157.

Wang Xu, Kehai Chen, Lili Mou, and Tiejun Zhao. 2022. Document-level relation extraction with sentences importance estimation and focusing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2920–2929, Seattle, United States. Association for Computational Linguistics.

Wang Xu, Kehai Chen, and Tiejun Zhao. 2021b. Document-level relation extraction with reconstruction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14167–14175.

Jie Yang and Yue Zhang. 2018. NCRF++: An open-source neural sequence labeling toolkit. In *Proceedings of ACL 2018, System Demonstrations*, pages

74–79, Melbourne, Australia. Association for Computational Linguistics.

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, Florence, Italy. Association for Computational Linguistics.

Yunzhi Yao, Shaohan Huang, Wenhui Wang, Li Dong, and Furu Wei. 2021. Adapt-and-distill: Developing small, fast and effective pretrained language models for domains. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 460–470, Online. Association for Computational Linguistics.

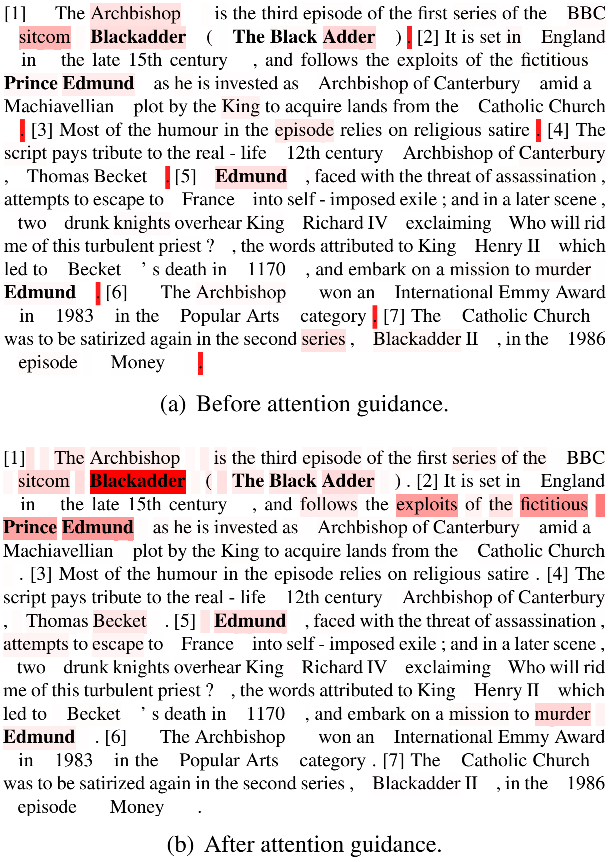
Shuang Zeng, Yuting Wu, and Baobao Chang. 2021. SIRE: Separate intra- and inter-sentential reasoning for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 524–534, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Ningyu Zhang, Xiang Chen, Xin Xie, Shumin Deng, Chuanqi Tan, Mosha Chen, Fei Huang, Luo Si, and Huajun Chen. 2021. Document-level relation extraction as semantic segmentation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3999–4006. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. 2019. Learning deep bilinear transformation for fine-grained image representation. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates Inc.

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*.



[1] The Archbishop is the third episode of the first series of the BBC sitcom **Blackadder** (**The Black Adder**) . [2] It is set in England in the late 15th century , and follows the exploits of the fictitious **Prince Edmund** as he is invested as Archbishop of Canterbury amid a Machiavellian plot by the King to acquire lands from the Catholic Church . [3] Most of the humour in the episode relies on religious satire . [4] The script pays tribute to the real - life 12th century Archbishop of Canterbury , Thomas Becket . [5] **Edmund** , faced with the threat of assassination , attempts to escape to France into self - imposed exile ; and in a later scene , two drunk knights overhear King Richard IV exclaiming Who will rid me of this turbulent priest ? , the words attributed to King Henry II which led to Becket ' s death in 1170 , and embark on a mission to murder **Edmund** . [6] The Archbishop won an International Emmy Award in 1983 in the Popular Arts category . [7] The Catholic Church was to be satirized again in the second series , Blackadder II , in the 1986 episode Money .

(a) Before attention guidance.

[1] The Archbishop is the third episode of the first series of the BBC sitcom **Blackadder** (**The Black Adder**) . [2] It is set in England in the late 15th century , and follows the exploits of the fictitious **Prince Edmund** as he is invested as Archbishop of Canterbury amid a Machiavellian plot by the King to acquire lands from the Catholic Church . [3] Most of the humour in the episode relies on religious satire . [4] The script pays tribute to the real - life 12th century Archbishop of Canterbury , Thomas Becket . [5] **Edmund** , faced with the threat of assassination , attempts to escape to France into self - imposed exile ; and in a later scene , two drunk knights overhear King Richard IV exclaiming Who will rid me of this turbulent priest ? , the words attributed to King Henry II which led to Becket ' s death in 1170 , and embark on a mission to murder **Edmund** . [6] The Archbishop won an International Emmy Award in 1983 in the Popular Arts category . [7] The Catholic Church was to be satirized again in the second series , Blackadder II , in the 1986 episode Money .

(b) After attention guidance.

Figure 3: Heatmaps of token importance for localized context pooling before and after guiding the attention with evidence when deciding the relation for entity pair (*Prince Edmund*, *The Black Adder*). The gold relation is *present in work* with evidence sentences 1 and 2. Deeper the color, the larger the value.

A Hyper-Parameters and Runtime

We adopt AdamW as the optimizer (Loshchilov and Hutter, 2019) and apply a linear warmup for the learning rate at the first 6% steps. Important hyper-parameters are shown in Table 7, which are mainly borrowed from existing works. Specifically, we borrow hyper-parameters from Zhou et al. (2021) to train the teacher model and borrow those from Tan et al. (2022a) to train and finetune the student model. The only exception is the number of epochs for training the student model, which is determined by a grid search from $\{2, 5, 8, 10\}$.

The average running time spent for our system at each training stage is shown in Table 8. Note that we employ a single Tesla V100 16GB GPU when utilizing BERT$_{\text{base}}$ and a single NVIDIA A100 40GB GPU when utilizing RoBERTa$_{\text{large}}$.

B Visualization: Evidence-Guided Attention

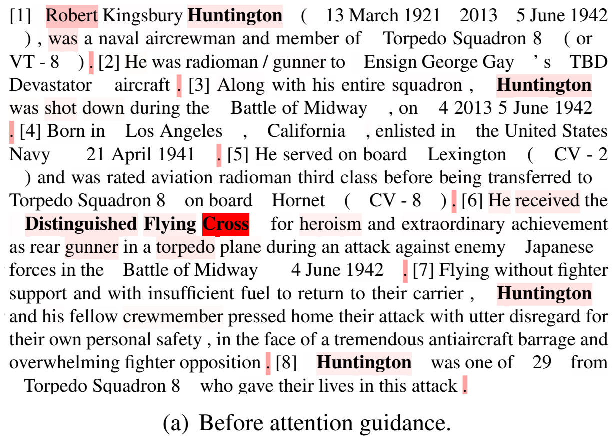
As introduced in Section 3.1, evidence knowledge of DREEAM originates from sentence-level supervision. We hypothesize that sentence-level supervision, from a more macro perspective, should improve its micro counterpart of token-level focusing. To test the hypothesis, we examine the token-level evidence distribution for localized context pool-

Hyperparam.	Train (teacher)		Train (student)		Finetune (student)	
	BERT$_{\text{base}}$	RoBERTa$_{\text{large}}$	BERT$_{\text{base}}$	RoBERTa$_{\text{large}}$	BERT$_{\text{base}}$	RoBERTa$_{\text{large}}$
# Epoch	30	30	2	5	10	10
lr for encoder	5e-5	3e-5	3e-5	1e-5	1e-6	1e-6
lr for classifier	1e-4	1e-4	1e-4	5e-5	3e-6	3e-6
max gradient norm	1.0	1.0	5.0	5.0	2.0	2.0

Table 7: Hyper-parameters in training.

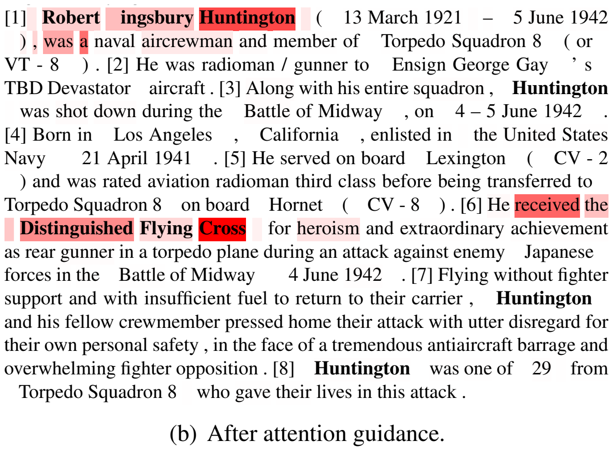
Phase	BERT$_{\text{base}}$	RoBERTa$_{\text{large}}$
Train (teacher)	1h18min	1h18min
Train (student)	2h55min	6h12min
Finetune (student)	26min	29min

Table 8: Runtime for each training stage.



[1] Robert Kingsbury **Huntington** (13 March 1921 2013 5 June 1942) , was a naval aircrewman and member of Torpedo Squadron 8 (or VT - 8) . [2] He was radioman / gunner to Ensign George Gay ’ s TBD Devastator aircraft . [3] Along with his entire squadron , **Huntington** was shot down during the Battle of Midway , on 4 2013 5 June 1942 . [4] Born in Los Angeles , California , enlisted in the United States Navy 21 April 1941 . [5] He served on board Lexington (CV - 2) and was rated aviation radioman third class before being transferred to Torpedo Squadron 8 on board Hornet (CV - 8) . [6] He received the **Distinguished Flying Cross** for heroism and extraordinary achievement as rear gunner in a torpedo plane during an attack against enemy Japanese forces in the Battle of Midway 4 June 1942 . [7] Flying without fighter support and with insufficient fuel to return to their carrier , **Huntington** and his fellow crewmember pressed home their attack with utter disregard for their own personal safety , in the face of a tremendous antiaircraft barrage and overwhelming fighter opposition . [8] **Huntington** was one of 29 from Torpedo Squadron 8 who gave their lives in this attack .

(a) Before attention guidance.



[1] **Robert** **ingsbury** **Huntington** (13 March 1921 – 5 June 1942) , was a naval aircrewman and member of Torpedo Squadron 8 (or VT - 8) . [2] He was radioman / gunner to Ensign George Gay ’ s TBD Devastator aircraft . [3] Along with his entire squadron , **Huntington** was shot down during the Battle of Midway , on 4 – 5 June 1942 . [4] Born in Los Angeles , California , enlisted in the United States Navy 21 April 1941 . [5] He served on board Lexington (CV - 2) and was rated aviation radioman third class before being transferred to Torpedo Squadron 8 on board Hornet (CV - 8) . [6] He received the **Distinguished Flying Cross** for heroism and extraordinary achievement as rear gunner in a torpedo plane during an attack against enemy Japanese forces in the Battle of Midway 4 June 1942 . [7] Flying without fighter support and with insufficient fuel to return to their carrier , **Huntington** and his fellow crewmember pressed home their attack with utter disregard for their own personal safety , in the face of a tremendous antiaircraft barrage and overwhelming fighter opposition . [8] **Huntington** was one of 29 from Torpedo Squadron 8 who gave their lives in this attack .

(b) After attention guidance.

Figure 4: Heatmaps of token importance for localized context pooling before and after guiding the attention with evidence when deciding the relation for entity pair (*Robert Kingsbury Huntington*, *Distinguished Flying Cross*). The gold relation is *award received* with evidence sentences 1 and 6. Deeper the color, the larger the value.

ing. Specifically, we utilize heatmaps to visualize $\boldsymbol{q}^{(s,o)}$ and observe the differences before and after evidence-guided training.

Results are shown in Figure 3 and 4. We adopt the toolkit developed by Yang and Zhang (2018). It is obvious that the distribution is more focused on sentences 1 and 2 in Figure 3(b) than in Figure 3(a). Before training the evidence-guided attention, the model tends to focus on the period of each sentence. Guiding the attention with evidence helps the model to focus more on sentences 1 and 2, as well as the critical tokens providing a clue for relation classification, such as *fictitious* in Figure 3 and *received* in Figure 4.