SAIS: Supervising and Augmenting Intermediate Steps for Document-Level Relation Extraction

Yuxin Xiao[1], **Zecheng Zhang**[2], **Yuning Mao**[3], **Carl Yang**[4], **Jiawei Han**[3]
[1]Carnegie Mellon University, [2]Stanford University,
[3]University of Illinois at Urbana-Champaign, [4]Emory University
[1]yuxinxia@cs.cmu.edu [2]zecheng@cs.stanford.edu
[3]{yuningm2,hanj}@illinois.edu [4]j.carlyang@emory.edu

Abstract

Stepping from sentence-level to document-level, the research on relation extraction (RE) confronts increasing text length and more complicated entity interactions. Consequently, it is more challenging to encode the key information sources—relevant contexts and entity types. However, existing methods only *implicitly* learn to model these critical information sources while being trained for RE. As a result, they suffer the problems of ineffective supervision and uninterpretable model predictions. In contrast, we propose to *explicitly* teach the model to capture relevant contexts and entity types by Supervising and Augmenting Intermediate Steps (SAIS) for RE. Based on a broad spectrum of carefully designed tasks, our proposed SAIS method not only extracts relations of better quality due to more effective supervision, but also retrieves the corresponding supporting evidence more accurately so as to enhance interpretability. By assessing model uncertainty, SAIS further boosts the performance via evidence-based data augmentation and ensemble inference while reducing the computational cost. Eventually, SAIS delivers state-of-the-art RE results on three benchmarks (DocRED, CDR, and GDA) and outperforms the runner-up by 5.04% relatively in F1 score in evidence retrieval on DocRED.[1]

1 Introduction

Playing a crucial role in the continuing effort of transforming unstructured text into structured knowledge, RE (Bach and Badaskar, 2007) seeks to identify relations between an entity pair based on a given piece of text. Earlier studies mostly pay attention to sentence-level RE (Zhang et al., 2017; Hendrickx et al., 2019) (i.e., the targeting entity pair co-occur within a sentence) and achieve promising results (Zhang et al., 2019; Zhou et al., 2020). Based on an extensive empirical analysis, Peng et al. (2020) reveals that textual contexts and entity types are the major information sources that lead to the success of prior approaches.

Given that more complicated relations are often expressed by multiple sentences, recent focus of RE has been largely shifted to the document level (Yao et al., 2019; Cheng et al., 2021). Existing document-level RE methods (Zeng et al., 2020; Zhou et al., 2021) utilize advanced neural architectures such as heterogeneous graph neural networks (Yang et al., 2020) and pre-trained language models (Xu et al., 2021b). However, although documents typically include longer contexts and more intricate entity interactions, most prior methods only *implicitly* learn to encode contexts and entity types while being trained for RE. As a result, they deliver inferior and uninterpretable results.

On the other hand, it has been a trend that many recent datasets support the training of more powerful language models by providing multi-task annotations such as coreference and evidence (Yao et al., 2019; Li et al., 2016; Wu et al., 2019). Therefore, in contrast to existing methods, we advocate for *explicitly* guiding the model to capture textual contexts and entity type information by Supervising and Augmenting Intermediate Steps (SAIS) for RE. More specifically, we argue that, from the input document with annotated entity mentions to the ultimate output of RE, there are four intermediate steps involved in the reasoning process. Consider the motivating example in Figure 1:

(1) **Coreference Resolution (CR):** Although Sentence 0 describes the "*citizenship*" of "*Carl Linnaeus the Younger*" and Sentence 1 discusses the "*father*" of "*Linnaeus filius*", the two names essentially refer to the same person. Hence, given a document, we need to first resolve various contextual roles represented by different mentions of the same entity via CR.

(2) **Entity Typing (ET):** After gathering contextual information from entity mentions, ET reg-

[1]Our code is available at https://github.com/xiaoyuxin1002/SAIS.

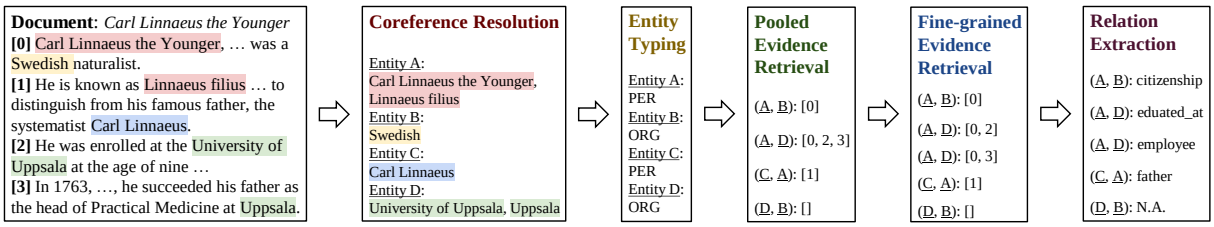

Figure 1: Motivating example adapted from DocRED. From the input document with annotated entity mentions to the RE output, there are four intermediate steps involved in the reasoning process. These steps are complementary to RE, in the sense that CR, PER, and FER capture textual contexts while ET preserves entity type information.

ularizes entity representations with the corresponding type information (e.g., Entity A, "*Linnaeus filius*", is of type "*PER*" (person)). Within an entity pair, the type information of the head and tail entities can be used to filter out impossible relations, as the relation "*year_of_birth*" can never appear between two entities of type "*PER*", for instance.

(3) **Pooled** and (4) **Fine-grained Evidence Retrieval (PER** and **FER):** A unique task for locating the relevant contexts within a document for an entity pair with any valid relation is to retrieve the evidence sentences supporting the relation. Nonetheless, some entity pairs may not express valid relations within the given document (e.g., Entities D and B in the example). Meanwhile some entity pairs possess multiple relations (e.g., Entity A is both "*educated_at*" and an "*employee*" of Entity D), each with a different evidence set. Therefore, we use PER to distinguish entity pairs with and without valid supporting sentences and FER to output more interpretable evidence unique to each valid relation of an entity pair.

In this way, the four intermediate steps are *complementary* to RE, in the sense that CR, PER, and FER capture textual contexts while ET preserves entity type information. Consequently, by explicitly supervising the model's outputs in these intermediate steps via carefully designed tasks, we extract relations of improved quality.

In addition, based on the predicted evidence, we filtrate relevant contexts by augmenting specific intermediate steps with pseudo documents or attention masks. By assessing model confidence, we apply these two kinds of evidence-based data augmentation together with ensemble inference, only when the model is *uncertain* about its original predictions. Eventually, we further boost the performance with negligible computational cost.

Altogether, our SAIS method achieves state-of-the-art RE performance on three benchmarks (DocRED (Yao et al., 2019), CDR (Li et al., 2016), and GDA (Wu et al., 2019)) due to more effective supervision and enhances interpretability by improving the evidence retrieval (ER) F1 score on DocRED by 5.04% relatively compared to the runner-up.

2 Background

2.1 Problem Formulation

Consider a document d containing sentences $\mathcal{S}_d = \{s_i\}_{i=1}^{|\mathcal{S}_d|}$ and entities $\mathcal{E}_d = \{e_i\}_{i=1}^{|\mathcal{E}_d|}$ where each entity e is assigned an entity type $c \in \mathcal{C}$ and appears at least once in d by its mentions $\mathcal{M}_e = \{m_i\}_{i=1}^{|\mathcal{M}_e|}$. For a pair of head and tail entities (e_h, e_t), document-level RE aims to predict if any relation $r \in \mathcal{R}$ exists between them, based on whether r is expressed by some pair of e_h's and e_t's mentions in d. Here, $\mathcal{C}$ and $\mathcal{R}$ are pre-defined sets of entity and relation types, respectively. Moreover, for (e_h, e_t) and each of their valid relations $r \in \mathcal{R}_{h,t}$, ER aims to identify the subset $\mathcal{V}_{h,t,r}$ of $\mathcal{S}_d$ that is sufficient to express the triplet (e_h, e_t, r).

2.2 Related Work

Early research efforts on RE (Bach and Badaskar, 2007; Pawar et al., 2017) center around predicting relations for entity pairs at the sentence level (Zhang et al., 2017; Hendrickx et al., 2019). Many pattern-based (Califf and Mooney, 1999; Qu et al., 2018; Zhou et al., 2020) and neural network-based (Cai et al., 2016; Feng et al., 2018; Zhang et al., 2019) models have shown impressive results. A recent study (Peng et al., 2020) attributes the success of these models to their ability to capture textual contexts and entity type information.

Nevertheless, since more complicated relations can only be expressed by multiple sentences, there has been a shift of focus lately towards document-level RE (Yao et al., 2019; Li et al., 2016; Cheng

et al., 2021; Wu et al., 2019). According to how an approach models contexts, there are two general trends within the domain. Graph-based approaches (Nan et al., 2020; Wang et al., 2020; Zeng et al., 2020; Li et al., 2020; Zeng et al., 2021; Xu et al., 2021c,d; Sahu et al., 2019; Guo et al., 2019) typically infuse contexts into heuristic-based document graphs and perform multi-hop reasoning via advanced neural techniques. Transformer-based approaches (Wang et al., 2019; Tang et al., 2020; Huang et al., 2020; Xu et al., 2021a; Zhou et al., 2021; Zhang et al., 2021; Xie et al., 2022; Ye et al., 2020) leverage the strength of pre-trained language models (Devlin et al., 2019; Liu et al., 2019) to encode long-range contextual dependencies. However, most prior methods only implicitly learn to capture contexts while being trained for RE. Consequently, they experience ineffective supervision and uninterpretable model predictions.

On the contrary, we propose to explicitly teach the model to capture textual contexts and entity type information via a broad spectrum of carefully designed tasks. Furthermore, we boost the RE performance by ensembling the results of evidence-augmented inputs. Compared to EIDER (Xie et al., 2022), we leverage the more precise and interpretable FER for retrieving evidence and present two different kinds of evidence-based data augmentation. We also save the computational cost by applying ensemble learning only to the uncertain subset of relation triplets. As a result, our SAIS method not only enhances the RE performance due to more effective supervision, but also retrieves more accurate evidence for better interpretability.

3 Supervising Intermediate Steps

This section describes the tasks that explicitly supervise the model's outputs in the four intermediate steps. Together they complement the quality of RE.

3.1 Document Encoding

Given the promising performance of pre-trained language models (PLM) in various downstream tasks, we resort to PLM for encoding the document. More specifically, for a document d, we insert a classifier token "[CLS]" and a separator token "[SEP]" at the start and end of each sentence $s \in \mathcal{S}_d$, respectively. Each mention $m \in \mathcal{M}_d$ is wrapped with a pair of entity markers "*" (Zhang et al., 2017) to indicate the position of entity mentions. Then we feed the document, with alternating segment token indices for each sentence (Liu and Lapata, 2019), into a PLM:

$$\mathbf{H}, \mathbf{A} = \text{PLM}(d), \tag{1}$$

to obtain the token embeddings $\mathbf{H} \in \mathbb{R}^{N_d \times H}$ and the cross-token attention $\mathbf{A} \in \mathbb{R}^{N_d \times N_d}$. $\mathbf{A}$ is the average of the attention heads in the last transformer layer (Vaswani et al., 2017) of the PLM. N_d is the number of tokens in d, and H is the embedding dimension of the PLM. We take the embedding of "*" or "[CLS]" before each mention or sentence as the corresponding mention or sentence embedding, respectively.

3.2 Coreference Resolution (CR)

As a case study, it is reported by Yao et al. (2019) that around 17.6% of relation instances in DocRED require coreference reasoning. Hence, after encoding the document, we resolve the repeated contextual mentions to an entity via CR. In particular, consider a pair of mentions (m_i, m_j), we determine the probability of whether m_i and m_j refer to the same entity by passing their corresponding embeddings $\mathbf{m}_i$ and $\mathbf{m}_j$ through a group bilinear layer (Zheng et al., 2019). The layer splits the embeddings into K equal-sized groups ($[\mathbf{m}_i^1, \dots, \mathbf{m}_i^K] = \mathbf{m}_i$, similar for $\mathbf{m}_j$) and applies bilinear with parameter $\mathbf{W}_m^k \in \mathbb{R}^{H/K \times H/K}$ within each group:

$$\mathbb{P}_{i,j}^{\text{CR}} = \sigma\left(\sum_{k=1}^{K} \mathbf{m}_i^{k\top} \mathbf{W}_m^k \mathbf{m}_j^k + b_m\right), \tag{2}$$

where $b_m \in \mathbb{R}$ and σ is the sigmoid function.

Since most mention pairs refer to distinct entities (each entity has only 1.34 mentions on average in DocRED), we adopt the focal loss (Lin et al., 2017) on top of the binary cross-entropy to mitigate this extreme class imbalance:

$$\ell_d^{\text{CR}} = -\sum_{m_i \in \mathcal{M}_d} \sum_{m_j \in \mathcal{M}_d} \Big(y_{i,j}^{\text{CR}} (1 - \mathbb{P}_{i,j}^{\text{CR}})^{\gamma^{\text{CR}}} \log \mathbb{P}_{i,j}^{\text{CR}} + (1 - y_{i,j}^{\text{CR}}) (\mathbb{P}_{i,j}^{\text{CR}})^{\gamma^{\text{CR}}} \log(1 - \mathbb{P}_{i,j}^{\text{CR}}) \Big) w_{i,j}^{\text{CR}}, \tag{3}$$

where $y_{i,j}^{\text{CR}} = 1$ if m_i and m_j refer to the same entity, and 0 otherwise. Class weight $w_{i,j}^{\text{CR}}$ is inversely proportional to the frequency of $y_{i,j}^{\text{CR}}$, and γ^{CR} is a hyperparameter.

3.3 Entity Typing (ET)

In a pair of entities, the type information can be used to filter out impossible relations. Therefore,

we regularize entity embeddings via ET. More specifically, we first derive the embedding of an entity e by integrating the embeddings of its mentions $\mathcal{M}_e$ via logsumexp pooling (Jia et al., 2019): $\mathbf{e} = \log \sum_{m \in \mathcal{M}_e} \exp(\mathbf{m})$. Since entity e could appear either at the head or tail in an entity pair, we distinguish between the head entity embedding $\mathbf{e}'_h$ and the tail entity embedding $\mathbf{e}'_t$ via two separate linear layers:

$$\mathbf{e}'_h = \mathbf{W}_{e_h}\mathbf{e} + \mathbf{b}_{e_h}, \;\; \mathbf{e}'_t = \mathbf{W}_{e_t}\mathbf{e} + \mathbf{b}_{e_t}, \quad (4)$$

where $\mathbf{W}_{e_h}, \mathbf{W}_{e_t} \in \mathbb{R}^{H \times H}$ and $\mathbf{b}_{e_h}, \mathbf{b}_{e_t} \in \mathbb{R}^H$.

However, no matter where e appears in an entity pair, its head and tail embeddings should always preserve e's type information. Hence, we calculate the probability of which entity type e belongs to by passing $\mathbf{e}'_\nu$ for $\nu \in \{h, t\}$ through a linear layer

$$\mathbb{P}^{\text{ET}}_e = \varsigma(\mathbf{W}_c \tanh(\mathbf{e}'_\nu) + \mathbf{b}_c), \quad (5)$$

followed by the multi-class cross-entropy loss:

$$\ell^{\text{ET}}_d = -\sum_{e \in \mathcal{E}_d} \sum_{c \in \mathcal{C}} y^{\text{ET}}_{e,c} \log \mathbb{P}^{\text{ET}}_{e,c}, \quad (6)$$

where $\mathbf{W}_c \in \mathbb{R}^{|\mathcal{C}| \times H}$, $\mathbf{b}_c \in \mathbb{R}^{|\mathcal{C}|}$, and ς is the softmax function. $y^{\text{ET}}_{e,c} = 1$ if e is of entity type c, and 0 otherwise.

3.4 Pooled Evidence Retrieval (PER)

To further capture textual contexts, we explicitly guide the attention in the PLM to the supporting sentences of each entity pair via PER. That is, we want to identify the pooled evidence set $\mathcal{V}_{h,t} = \cup_{r \in \mathcal{R}_{h,t}} \mathcal{V}_{h,t,r}$ in d that is important to an entity pair (e_h, e_t), regardless of the specific relation expressed by a particular sentence $s \in \mathcal{V}_{h,t}$. In this case, given (e_h, e_t), we first compute a unique context embedding $\mathbf{c}_{h,t}$ based on the cross-token attention from Equation 1:

$$\mathbf{c}_{h,t} = \mathbf{H}^\top \frac{\mathbf{A}_h \otimes \mathbf{A}_t}{\mathbf{1}^\top (\mathbf{A}_h \otimes \mathbf{A}_t)}. \quad (7)$$

Here, $\otimes$ is the element-wise product. $\mathbf{A}_h$ is e_h's attention to all the tokens in the document (i.e., the average of e_h's mention-level attention). Similar for $\mathbf{A}_t$. Then we measure the probability of whether a sentence $s \in \mathcal{S}_d$ is part of the pooled supporting evidence $\mathcal{V}_{h,t}$ by passing (e_h, e_t)'s context embedding $\mathbf{c}_{h,t}$ and sentence s' embedding $\mathbf{s}$ through a group bilinear layer:

$$\mathbb{P}^{\text{PER}}_{h,t,s} = \sigma\left(\sum_{k=1}^{K} \mathbf{c}^{k\top}_{h,t} \mathbf{W}^k_p \mathbf{s}^k + b_p\right), \quad (8)$$

where $\mathbf{W}^k_p \in \mathbb{R}^{H/K \times H/K}$ and $b_p \in \mathbb{R}$.

Again, we face a severe class imbalance here, since most entity pairs (97.1% in DocRED) do not have valid relations or supporting evidence. As a result, similar to Section 3.2, we also use the focal loss with the binary cross-entropy:

$$\begin{aligned}\ell^{\text{PER}}_d = -\sum_{e_h \in \mathcal{E}_d} \sum_{e_t \in \mathcal{E}_d} \sum_{s \in \mathcal{S}_d} \Big(& y^{\text{PER}}_{h,t,s} (1 - \mathbb{P}^{\text{PER}}_{h,t,s})^{\gamma^{\text{PER}}} \\ & \log \mathbb{P}^{\text{PER}}_{h,t,s} + (1 - y^{\text{PER}}_{h,t,s}) (\mathbb{P}^{\text{PER}}_{h,t,s})^{\gamma^{\text{PER}}} \\ & \log(1 - \mathbb{P}^{\text{PER}}_{h,t,s}) \Big) w^{\text{PER}}_{h,t,s}, \end{aligned} \quad (9)$$

where $y^{\text{PER}}_{h,t,s} = \mathbb{1}\{s \in \mathcal{V}_{h,t}\}$, class weight $w^{\text{PER}}_{h,t,s}$ is inversely proportional to the frequency of $y^{\text{PER}}_{h,t,s}$, and γ^{PER} is a hyperparameter.

3.5 Fine-grained Evidence Retrieval (FER)

In addition to PER, we would like to further refine $\mathcal{V}_{h,t}$, since an entity pair could have multiple valid relations and, correspondingly, multiple sets of evidence. As a result, we explicitly train the model to recover contextual evidence unique to a triplet (e_h, e_t, r) via FER for better interpretability. More specifically, given (e_h, e_t, r), we first generate a triplet embedding $\mathbf{l}_{h,t,r}$ by merging $\mathbf{e}_h$, $\mathbf{e}_t$, $\mathbf{c}_{h,t}$, and r's relation embedding $\mathbf{r}$ via a linear layer:

$$\mathbf{l}_{h,t,r} = \tanh(\mathbf{W}_l[\mathbf{e}_h \| \mathbf{e}_t \| \mathbf{c}_{h,t} \| \mathbf{r}] + \mathbf{b}_l), \quad (10)$$

where $\mathbf{W}_l \in \mathbb{R}^{H \times 4H}$, $\mathbf{b}_l \in \mathbb{R}^H$, $\|$ represents concatenation, and $\mathbf{r}$ is initialized from the embedding matrix of the PLM.

Similarly, we use a group bilinear layer to assess the probability of whether a sentence $s \in \mathcal{S}_d$ is included in the fine-grained evidence set $\mathcal{V}_{h,t,r}$:

$$\mathbb{P}^{\text{FER}}_{h,t,r,s} = \sigma\left(\sum_{k=1}^{K} \mathbf{l}^{k\top}_{h,t,r} \mathbf{W}^k_f \mathbf{s}^k + b_f\right), \quad (11)$$

where $\mathbf{W}^k_f \in \mathbb{R}^{H/K \times H/K}$ and $b_f \in \mathbb{R}$.

Since FER only involves entity pairs with valid relations, the class imbalance is milder here than in PER. Hence, let $y^{\text{FER}}_{h,t,r,s} = \mathbb{1}\{s \in \mathcal{V}_{h,t,r}\}$, we deploy the standard binary cross-entropy loss:

$$\begin{aligned}\ell^{\text{FER}}_d = -\sum_{e_i \in \mathcal{E}_d} \sum_{e_j \in \mathcal{E}_d} \sum_{r \in \mathcal{R}_{h,t}} \sum_{s \in \mathcal{S}_d} \big(& y^{\text{FER}}_{h,t,r,s} \log \mathbb{P}^{\text{FER}}_{h,t,r,s} \\ & + (1 - y^{\text{FER}}_{h,t,r,s}) \log(1 - \mathbb{P}^{\text{FER}}_{h,t,r,s}) \big). \end{aligned} \quad (12)$$

3.6 Relation Extraction (RE)

Based on the four complementary tasks introduced above, for an entity pair (e_h, e_t), we encode relevant contexts in $\mathbf{c}_{h,t}$ and preserve entity type information in $\mathbf{e}'_h$ and $\mathbf{e}'_t$. Ultimately, we acquire the contexts needed by the head and tail entities from $\mathbf{c}_{h,t}$ via two separate linear layers:

$$\mathbf{c}'_h = \mathbf{W}_{c_h}\mathbf{c}_{h,t} + \mathbf{b}_{c_h}, \mathbf{c}'_t = \mathbf{W}_{c_t}\mathbf{c}_{h,t} + \mathbf{b}_{c_t}, \quad (13)$$

where $\mathbf{W}_{c_h}, \mathbf{W}_{c_t} \in \mathbb{R}^{H\times H}$ and $\mathbf{b}_{c_h}, \mathbf{b}_{c_t} \in \mathbb{R}^H$, and then combine them with the type information to generate the head and tail entity representations:

$$\mathbf{e}''_h = \tanh(\mathbf{e}'_h + \mathbf{c}'_h), \ \mathbf{e}''_t = \tanh(\mathbf{e}'_t + \mathbf{c}'_t). \quad (14)$$

Next, a group bilinear layer is utilized to calculate the logit of how likely a relation $r \in \mathcal{R}$ exists between e_h and e_t:

$$\mathbb{L}^{\text{RE}}_{h,t,r} = \sum_{k=1}^{K} \mathbf{e}''^{k\top}_h \mathbf{W}^k_r \mathbf{e}''^k_t + b_r\,, \quad (15)$$

where $\mathbf{W}^k_r \in \mathbb{R}^{H/K\times H/K}$ and $b_r \in \mathbb{R}$.

As discussed earlier, only a small portion of entity pairs have valid relations, among which multiple relations could co-exist between a pair. Therefore, to deal with the problem of multi-label imbalanced classification, we follow Zhou et al. (2021) by introducing a threshold relation class TH and adopting an adaptive threshold loss:

$$\ell^{\text{RE}}_d = -\sum_{e_h\in\mathcal{E}_d}\sum_{e_t\in\mathcal{E}_d} \left[\sum_{r\in\mathcal{P}_{h,t}} \log\left(\frac{\exp \mathbb{L}^{\text{RE}}_{h,t,r}}{\sum_{r'\in\mathcal{P}_{h,t}\cup\{\text{TH}\}} \mathbb{L}^{\text{RE}}_{h,t,r'}}\right) + \log\left(\frac{\exp \mathbb{L}^{\text{RE}}_{h,t,\text{TH}}}{\sum_{r'\in\mathcal{N}_{h,t}\cup\{\text{TH}\}} \mathbb{L}^{\text{RE}}_{h,t,r'}}\right)\right]. \quad (16)$$

In essence, we aim to increase the logits of valid relations $\mathcal{P}_{h,t}$ and decrease the logits of invalid relations $\mathcal{N}_{h,t}$, both relative to TH.

Overall, with the goal of improving the model's RE performance by better capturing entity type information and textual contexts, we have designed four tasks to explicitly supervise the model's outputs in the corresponding intermediate steps. To this end, we visualize the entire pipeline $\text{SAIS}^{\text{O}}_{\text{All}}$ in Appendix A and integrate all the tasks by minimizing the multi-task learning objective

$$\ell = \sum_{d\in\mathcal{D}_{\text{train}}} \left(\ell^{\text{RE}}_d + \sum_{\text{Task}} \eta^{\text{Task}}\ell^{\text{Task}}_d\right), \quad (17)$$

where Task $\in$ {CR, ET, PER, FER}. η^{Task}'s are hyperparameters balancing the relative task weight.

During inference with the current pipeline $\text{SAIS}^{\text{O}}_{\text{All}}$, we predict if a triplet (e_h, e_t, r) is valid (i.e., if relation r exists between entity pair (e_h, e_t)) by checking if its logit is larger than the corresponding threshold logit (i.e., $\mathbb{L}^{\text{RE}}_{h,t,r} > \mathbb{L}^{\text{RE}}_{h,t,\text{TH}}$). For each predicted triplet (e_h, e_t, r), we assess if a sentence s belongs to the evidence set $\mathcal{V}_{h,t,r}$ by checking if $\mathbb{P}^{\text{FER}}_{h,t,r,s} > \alpha^{\text{FER}}$ where α^{FER} is a threshold.

4 Augmenting Intermediate Steps

We further improve RE after training the pipeline $\text{SAIS}^{\text{O}}_{\text{All}}$ by augmenting the intermediate steps in $\text{SAIS}^{\text{O}}_{\text{All}}$ with the retrieved evidence from FER.

4.1 When to Augment Intermediate Steps

The evidence predicted by FER is unique to each triplet (e_h, e_t, r). However, consider the total number of all possible triplets (around 40 million in the develop set of DocRED), it is computationally prohibitive to augment the inference result of each triplet with individually predicted evidence. Instead, following the idea of selective prediction (El-Yaniv et al., 2010), we identify the triplet subset $\mathcal{U}$ for which the model is *uncertain* about its relation predictions with the original pipeline $\text{SAIS}^{\text{O}}_{\text{All}}$. More specifically, we set the model's confidence for (e_h, e_t, r) as $\mathbb{L}^{\text{O}}_{h,t,r} = \mathbb{L}^{\text{RE}}_{h,t,r} - \mathbb{L}^{\text{RE}}_{h,t,\text{TH}}$. Then, the uncertain set $\mathcal{U}$ consists of triplets with the lowest $\theta\%$ absolute confidence $|\mathbb{L}^{\text{O}}_{h,t,r}|$. Consequently, we reject the original relation predictions for $(e_h, e_t, r) \in \mathcal{U}$ and apply evidence-based data augmentation to enhance the performance (more details in Section 4.2).

To determine the rejection rate $\theta\%$ (note that $\theta\%$ is NOT a hyperparameter), we first sort all the triplets in the develop set based on their absolute confidence $|\mathbb{L}^{\text{O}}_{h,t,r}|$. When $\theta\%$ increases, the risk (i.e., inaccuracy rate) of the remaining triplets that are not in $\mathcal{U}$ is expected to decrease, and vice versa. On the one hand, we wish to reduce the risk for more accurate relation predictions; on the other hand, we want a low rejection rate so that data augmentation on a small rejected set incurs little computational cost. To balance this trade-off, we set $\theta\%$ as the rate that achieves the minimum of $\text{risk}^2 + \text{rejection rate}^2$. As shown in Figure 2, we find $\theta\% \approx 4.6\%$ in the develop set of DocRED. In practice, we can further limit the maximum number of rejected triplets per entity pair. By setting it as

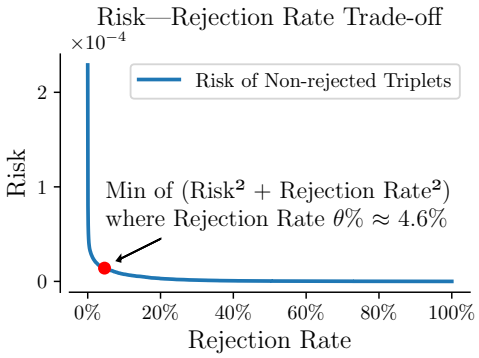

Figure 2: Trade-off between risk and rejection rate on the develop set of DocRED.

10 in experiments, we reduce the size of $\mathcal{U}$ to only 1.5% of all the triplets in the DocRED develop set.

4.2 How to Augment Intermediate Steps

Consider a triplet $(e_h, e_t, r) \in \mathcal{U}$. We first assume its validity and calculate the probability $\mathbb{P}^{\text{FER}}_{h,t,r,s}$ of a sentence s being part of $\mathcal{V}_{h,t,r}$ based on Section 3.5. Then in a similar way to how $\mathbb{L}^{\text{O}}_{h,t,r}$ is generated with $\text{SAIS}^{\text{O}}_{\text{All}}$, we design two types of evidence-based data augmentation as follows:

Pseudo Document-based ($\text{SAIS}^{\text{D}}_{\text{All}}$): Construct a pseudo document using sentences with $\mathbb{P}^{\text{FER}}_{h,t,r,s} > \alpha^{\text{FER}}$ and feed it into the original pipeline to get the confidence $\mathbb{L}^{\text{D}}_{h,t,r}$.

Attention Mask-based ($\text{SAIS}^{\text{M}}_{\text{All}}$): Formulate a mask $\mathbf{P}^{\text{FER}}_{h,t,r} \in \mathbb{R}^{N_d}$ based on $\mathbb{P}^{\text{FER}}_{h,t,r,s}$ and modify the context embedding to $\mathbf{c}_{h,t} = \mathbf{H}^\top \frac{\mathbf{A}_h \otimes \mathbf{A}_t \otimes \mathbf{P}^{\text{FER}}_{h,t,r}}{\mathbf{1}^\top (\mathbf{A}_h \otimes \mathbf{A}_t \otimes \mathbf{P}^{\text{FER}}_{h,t,r})}$. Maintain the rest of the pipeline and get the confidence $\mathbb{L}^{\text{M}}_{h,t,r}$.

Following Xie et al. (2022), we ensemble $\mathbb{L}^{\text{D}}_{h,t,r}$, $\mathbb{L}^{\text{M}}_{h,t,r}$, and the original confidence $\mathbb{L}^{\text{O}}_{h,t,r}$ with a blending parameter $\tau_r \in \mathbb{R}$ (Wolpert, 1992) for each relation $r \in \mathcal{R}$ as

$$\begin{aligned}\mathbb{P}^{\text{B}}_{h,t,r} &= \sigma(\mathbb{L}^{\text{B}}_{h,t,r}) \\ &= \sigma(\mathbb{L}^{\text{O}}_{h,t,r} + \mathbb{L}^{\text{D}}_{h,t,r} + \mathbb{L}^{\text{M}}_{h,t,r} - \tau_r). \quad (18)\end{aligned}$$

These parameters are trained by minimizing the binary cross-entropy loss on $\mathcal{U}$ of the develop set:

$$\begin{aligned}\ell^B = -\sum_{(e_h,e_t,r)\in\mathcal{U}} \big(&y^{\text{RE}}_{h,t,r} \log \mathbb{P}^{\text{B}}_{h,t,r} \\ &+ (1 - y^{\text{RE}}_{h,t,r}) \log(1 - \mathbb{P}^{\text{B}}_{h,t,r})\big), \quad (19)\end{aligned}$$

where $y^{\text{RE}}_{h,t,r} = 1$ if (e_h, e_t, r) is valid, and 0 otherwise. When making relation predictions for $(e_h, e_t, r) \in \mathcal{U}$, we check whether its blended confidence is positive (i.e., $\mathbb{L}^{\text{B}}_{h,t,r} > 0$).

In this way, we improve the RE performance when the model is uncertain about its original predictions and save the computational cost when the model is confident. The overall steps for evidence-based data augmentation and ensemble inference $\text{SAIS}^{\text{B}}_{\text{All}}$ are summarized in Appendix B. These steps are executed only after the training of $\text{SAIS}^{\text{O}}_{\text{All}}$ and, therefore, adds negligible computational cost.

5 Experiments

5.1 Experiment Setup

We evaluate the proposed SAIS method on the following three document-level RE benchmarks. DocRED (Yao et al., 2019) is a large-scale crowd-sourced dataset based on Wikipedia articles. It consists of 97 relation types, seven entity types, and 5,053 documents in total, where each document has 19.5 entities on average. CDR (Li et al., 2016) and GDA (Wu et al., 2019) are two biomedical datasets where CDR studies the binary interactions between disease and chemical concepts with 1,500 documents and GDA studies the binary relationships between gene and disease with 30,192 documents. We follow Christopoulou et al. (2019) for splitting the train and develop sets.

We run our experiments on one Tesla A6000 GPU and carry out five trials with different seeds to report the mean and one standard error. Based on Huggingface (Wolf et al., 2019), we apply cased BERT-base (Devlin et al., 2019) and RoBERTa-large (Liu et al., 2019) for DocRED and cased SciBERT (Beltagy et al., 2019) for CDR and GDA. The embedding dimension H of BERT or SciBERT is 768, and that of RoBERTa is 1,024. The number of groups K in all group bilinear layers is 64.

For the general hyperparameters of language models, we follow the setting in (Zhou et al., 2021). The learning rate for fine-tuning BERT is 5e−5, that for fine-tuning RoBERTa or SciBERT is 2e−5, and that for training the other parameters is 1e−4. All the trials are optimized by AdamW (Loshchilov and Hutter, 2019) for 20 epochs with early stopping and a linearly decaying scheduler (Goyal et al., 2017) whose warm-up ratio = 6%. Each batch contains 4 documents and the gradients of model parameters are clipped to a maximum norm of 1.

For the unique hyperparameters of our method, we choose 2 from $\{1, 1.5, 2\}$ for the focal hyperparameters γ^{CR} and γ^{PER} based on the develop

Model	DocRED Dev			DocRED Test		
	Relation		Evidence	Relation		Evidence
	Ign F1	F1	F1	Ign F1	F1	F1
HeterGSAN-BERT$_{base}$ (Xu et al., 2021d)	58.13	60.18	-	57.12	59.45	-
GAIN-BERT$_{base}$ (Zeng et al., 2020)	59.14	61.22	-	59.00	61.24	-
DRN-BERT$_{base}$ (Xu et al., 2021c)	59.33	61.39	-	59.15	61.37	-
SIRE-BERT$_{base}$ (Zeng et al., 2021)	59.82	61.60	-	60.18	62.05	-
BERT$_{base}$ (Wang et al., 2019)	-	54.16	-	-	53.20	-
E2GRE-BERT$_{base}$ (Huang et al., 2020)	55.22	58.72	47.14	-	-	-
SSAN-BERT$_{base}$ (Xu et al., 2021a)	57.03	59.19	-	56.06	58.41	-
ATLOP-BERT$_{base}$ (Zhou et al., 2021)	59.22	61.09	-	59.31	61.30	-
DocuNet-BERT$_{base}$ (Zhang et al., 2021)	59.86	61.83	-	59.93	61.86	-
Eider-BERT$_{base}$ (Xie et al., 2022)	**60.51**	62.48	50.71	60.42	62.47	51.27
SAIS$^{B}_{All}$-BERT$_{base}$ (Ours)	59.98 ± 0.13	**62.96 ± 0.11**	**53.70 ± 0.21**	**60.96**	**62.77**	**52.88**
RoBERTa$_{large}$ (Ye et al., 2020)	57.19	59.40	-	57.74	60.06	-
SSAN-RoBERTa$_{large}$ (Xu et al., 2021a)	60.25	62.08	-	59.47	61.42	-
E2GRE-RoBERTa$_{large}$ (Huang et al., 2020)	-	-	-	60.30	62.50	50.50
ATLOP-RoBERTa$_{large}$ (Zhou et al., 2021)	61.32	63.18	-	61.39	63.40	-
DocuNet-RoBERTa$_{large}$ (Zhang et al., 2021)	62.23	64.12	-	62.39	64.55	-
Eider-RoBERTa$_{large}$ (Xie et al., 2022)	**62.34**	64.27	52.54	62.85	64.79	53.01
SAIS$^{B}_{All}$-RoBERTa$_{large}$ (Ours)	62.23 ± 0.15	**65.17 ± 0.08**	**55.84 ± 0.23**	**63.44**	**65.11**	**55.67**

Table 1: RE and ER results (%) on DocRED. Ign F1 refers to the F1 score excluding the relation instances mentioned in the train set. Baselines using BERT$_{base}$ are separated into the graph-based (upper) and transformer-based (lower) groups. We report the test scores from the official scoreboard and the baseline scores from the corresponding papers. SAIS$^{B}_{All}$ achieves state-of-the-art performance on both RE and ER. Full details in Appendix C.

set. We also follow Xie et al. (2022) for setting the FER prediction threshold α^{FER} as 0.5 and all the relative task weights η^{Task} for Task $\in \{\text{CR}, \text{ET}, \text{PER}, \text{FER}\}$ as 0.1.

5.2 Quantitative Evaluation

Besides RE, DocRED also suggests to predict the supporting evidence for each relation instance. Therefore, we apply SAIS$^{B}_{All}$ to both RE and ER. We report the results of SAIS$^{B}_{All}$ as well as existing graph-based and transformer-based baselines in Table 1[2] (full details in Appendix C). Generally, thanks to PLMs' strength in modeling long-range dependencies, transformer-based methods perform better on RE than graph-based methods. Moreover, most earlier approaches are not capable of ER despite the interpretability ER adds to the predictions. In contrast, our SAIS$^{B}_{All}$ method not only establishes a new state-of-the-art result on RE, but also outperforms the runner-up significantly on ER.

Since neither CDR nor GDA annotates evidence sentences, we apply SAIS$^{O}_{RE+CR+ET}$ here. It is trained with RE, CR, and ET and infers without data augmentation. As shown in Table 2 (full details in Appendix C), our method improves the prior best RE F1 scores by 2.7% and 1.8% absolutely on CDR and GDA, respectively. It indicates that our proposed method can still improve upon the baselines even if only part of the four complementary tasks are annotated and operational.

5.3 Ablation Study

To investigate the effectiveness of each of the four complementary tasks proposed in Section 3, we carry out an extensive ablation study on the DocRED develop set by training SAIS with all possible combinations of those tasks. As shown in Table 3, without any complementary tasks, the RE performance of SAIS is comparable to ATLOP (Zhou et al., 2021) due to similar neural architectures. When only one complementary task is allowed, PER is the most effective single task, followed by ET. Although FER is functionally analogous to PER, since FER only involves the small subset of entity pairs with valid relations, the performance gain brought by FER alone is limited. When

[2]For a fair comparison, we report the scores of SSAN (Xu et al., 2021a) without being pretrained on an extra dataset.

Model	CDR	GDA
LSR (Nan et al., 2020)	64.8	82.2
SciBERT (Beltagy et al., 2019)	65.1	82.5
DHG (Zhang et al., 2020)	65.9	83.1
SSAN-SciBERT (Xu et al., 2021a)	68.7	83.7
ATLOP-SciBERT (Zhou et al., 2021)	69.4	83.9
SIRE-BioBERT (Zeng et al., 2021)	70.8	84.7
DocuNet-SciBERT (Zhang et al., 2021)	76.3	85.3
$\text{SAIS}^{\text{O}}_{\text{RE+CR+ET}}$-SciBERT (Ours)	**79.0** $\pm$ **0.8**	**87.1** $\pm$ **0.3**
$\text{SAIS}^{\text{O}}_{\text{RE+ET}}$-SciBERT	75.9 ± 0.9	86.1 ± 0.5
$\text{SAIS}^{\text{O}}_{\text{RE+CR}}$-SciBERT	74.5 ± 0.4	85.4 ± 0.2
$\text{SAIS}^{\text{O}}_{\text{RE}}$-SciBERT	72.8 ± 0.6	84.5 ± 0.3

Table 2: RE F1 results (%) on the CDR and GDA test sets. The baseline scores are from the corresponding papers. $\text{SAIS}^{\text{O}}_{\text{RE+CR+ET}}$ scores the highest on both datasets. Full details in Appendix C.

two tasks are used jointly, the pair of PER and ET, which combines textual contexts and entity type information, delivers the most significant improvement. The pair of PER and FER also performs well, which reflects the finding in (Peng et al., 2020) that context is the most important source of information. The version with all tasks except CR sees the least drop in F1, indicating that CR's supervision signals on capturing contexts can be covered in part by PER and FER. Last but not least, the SAIS pipeline with all four complementary tasks achieves the highest F1 score. Similar trends are also recognized on CDR and GDA in Table 2, where SAIS trained with both CR and ET (besides RE) scores higher than its single-task counterpart.

Moreover, as compared to the original pipeline $\text{SAIS}^{\text{O}}_{\text{All}}$, pseudo document-based data augmentation $\text{SAIS}^{\text{D}}_{\text{All}}$ acts as a hard filter by directly removing predicted non-evidence sentences, while attention mask-based data augmentation $\text{SAIS}^{\text{M}}_{\text{All}}$ distills the context more softly. Therefore, we observe in Table 4 that $\text{SAIS}^{\text{D}}_{\text{All}}$ earns a higher precision, whereas $\text{SAIS}^{\text{M}}_{\text{All}}$ attains a higher recall. By ensembling $\text{SAIS}^{\text{O}}_{\text{All}}$, $\text{SAIS}^{\text{D}}_{\text{All}}$, and $\text{SAIS}^{\text{M}}_{\text{All}}$, we improve the RE F1 score by 0.57% absolutely on the DocRED develop set.

5.4 Qualitative Analysis

To obtain a more insightful understanding of how textual contexts and entity type information help with RE, we present a case study in Figure 3 (a). Here, $\text{SAIS}^{\text{O}}_{\text{RE+ET}}$ is trained with the task (i.e., ET) related to entity type information while $\text{SAIS}^{\text{O}}_{\text{RE+CR+PER+FER}}$ is trained with the tasks (i.e., CR, PER, and FER) related to textual contexts.

CR	ET	PER	FER	RE	F1
				✓	61.18 ± 0.09
✓				✓	61.41 ± 0.11
	✓			✓	61.52 ± 0.10
		✓		✓	61.68 ± 0.04
			✓	✓	61.44 ± 0.07
✓	✓			✓	61.65 ± 0.12
✓		✓		✓	61.79 ± 0.08
✓			✓	✓	61.64 ± 0.10
	✓	✓		✓	61.88 ± 0.05
	✓		✓	✓	61.81 ± 0.04
		✓	✓	✓	61.85 ± 0.10
	✓	✓	✓	✓	62.13 ± 0.04
✓		✓	✓	✓	62.06 ± 0.09
✓	✓		✓	✓	61.91 ± 0.06
✓	✓	✓		✓	61.98 ± 0.05
✓	✓	✓	✓	✓	62.39 ± 0.08

Table 3: Ablation study (%) using SAIS^{O}-$\text{BERT}_{\text{base}}$ to assess the effectiveness of the four complementary tasks (i.e., CR, ET, PER, and FER) for RE based on the DocRED develop set.

$\text{SAIS}^{\text{O}}_{\text{All}}$	$\text{SAIS}^{\text{D}}_{\text{All}}$	$\text{SAIS}^{\text{M}}_{\text{All}}$	Precision	Recall	F1
✓			66.58	58.70	62.39
	✓		73.21	45.59	56.19
		✓	53.14	67.49	59.46
✓	✓		71.14	54.35	61.62
✓		✓	61.61	62.90	62.25
✓	✓	✓	67.76	58.79	62.96

Table 4: Ablation study (%) using $\text{BERT}_{\text{base}}$ to assess the effectiveness of data augmentation (i.e., original ($\text{SAIS}^{\text{O}}_{\text{All}}$), pseudo document-based ($\text{SAIS}^{\text{D}}_{\text{All}}$), and attention mask-based ($\text{SAIS}^{\text{M}}_{\text{All}}$)) for RE based on the DocRED develop set.

Compared to $\text{SAIS}^{\text{O}}_{\text{All}}$, which is trained with all four complementary tasks, they both exhibit drawbacks qualitatively. In particular, $\text{SAIS}^{\text{O}}_{\text{RE+ET}}$ can easily infer the relation "*country*" between Entities E and C based on their respective types "*ORG*" and "*LOC*", whereas $\text{SAIS}^{\text{O}}_{\text{RE+CR+PER+FER}}$ may misinterpret Entity E as of type "*PER*" and infer the relation "*citizenship*" wrongly. On the other hand, $\text{SAIS}^{\text{O}}_{\text{RE+CR+PER+FER}}$ can directly predict the relation "*place_of_birth*" between Entities A and B by pattern matching, while overemphasizing the type "*LOC*" of Entity B may cause $\text{SAIS}^{\text{O}}_{\text{RE+ET}}$ to deliver the wrong relation prediction "*location*". Last but

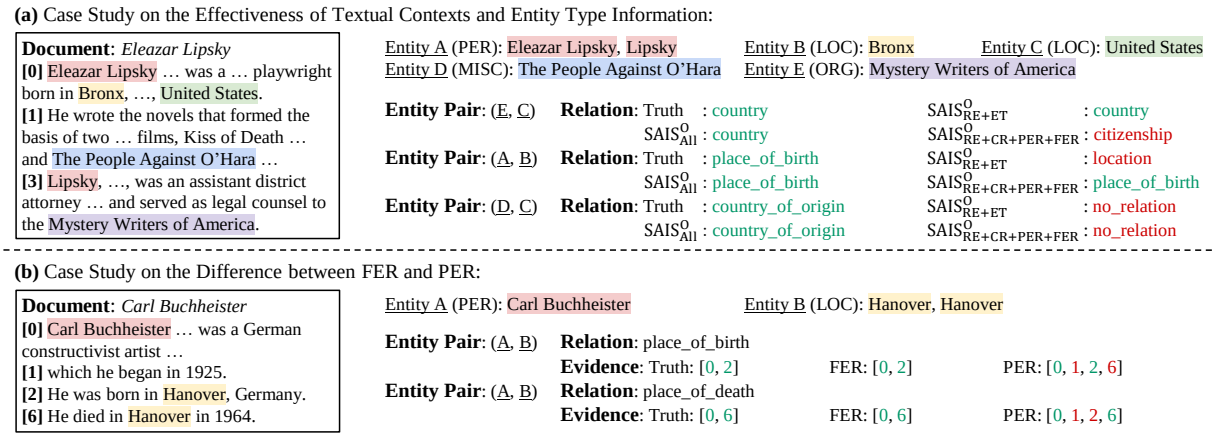

Figure 3: **(a)** Case study on the effectiveness of textual contexts and entity type information based on models' extracted relations from the DocRED develop set. By capturing contexts across sentences and regularizing them with entity type information, $\text{SAIS}^{\text{O}}_{\text{All}}$ extracts relations of better quality. **(b)** Case study on the difference between FER and PER based on retrieved evidence from the DocRED develop set. FER considers evidence unique to each relation for better interpretability. Irrelevant sentences are omitted here.

not least, $\text{SAIS}^{\text{O}}_{\text{All}}$ effectively models contexts spanning multiple sentences and regularizes them with entity type information. As a result, it is the only SAIS variant that correctly predicts the relation "*country_of_origin*" between Entities D and C.

Furthermore, to examine why SAIS (which uses FER for retrieving evidence) outperforms Eider (Xie et al., 2022) (which uses PER) significantly on ER in Table 1, we compare the performance of FER and PER based on a case study in Figure 3 (b). More specifically, PER identifies the same set of evidence for both relations between Entities A and B, among which Sentence 2 describes "*place_of_birth*" while Sentence 6 discusses "*place_of_death*". In contrast, FER considers an evidence set unique to each relation and outputs more interpretable results.

6 Conclusion

In this paper, we propose to explicitly teach the model to capture the major information sources of RE—textual contexts and entity types by Supervising and Augmenting Intermediate Steps (SAIS). Based on a broad spectrum of carefully designed tasks, SAIS extracts relations of enhanced quality due to more effective supervision and retrieves more accurate evidence for improved interpretability. SAIS further boosts the performance with evidence-based data augmentation and ensemble inference while preserving the computational cost by assessing model uncertainty. Experiments on three benchmarks demonstrate the state-of-the-art performance of SAIS on both RE and ER.

If given a plain document, we shall utilize existing tools (e.g., spaCy) to get noisy annotations and apply our method afterward. It is also interesting to investigate how other tasks (e.g., named entity recognition) could be incorporated into the multi-task learning pipeline of our SAIS method. We plan to explore these extensions in future works.

References

Nguyen Bach and Sameer Badaskar. 2007. A review of relation extraction. *Literature review for Language and Statistics II*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *EMNLP*.

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

Mary Elaine Califf and Raymond J. Mooney. 1999. Relational learning of pattern-match rules for information extraction. In *AAAI*.

Qiao Cheng, Juntao Liu, Xiaoye Qu, Jin Zhao, Jiaqing Liang, Zhefeng Wang, Baoxing Huai, Nicholas Jing Yuan, and Yanghua Xiao. 2021. Hacred: A large-scale relation extraction dataset toward hard cases in practical applications. In *ACL*.

Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. In *EMNLP*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep

bidirectional transformers for language understanding. In *NAACL*.

Ran El-Yaniv et al. 2010. On the foundations of noise-free selective classification. *JMLR*.

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

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. 2017. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*.

Zhijiang Guo, Yan Zhang, and Wei Lu. 2019. Attention guided graph convolutional networks for relation extraction. In *ACL*.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid O Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2019. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. *arXiv preprint arXiv:1911.10422*.

Kevin Huang, Guangtao Wang, Tengyu Ma, and Jing Huang. 2020. Entity and evidence guided relation extraction for docred. *arXiv preprint arXiv:2008.12283*.

Robin Jia, Cliff Wong, and Hoifung Poon. 2019. Document-level n-ary relation extraction with multiscale representation learning. In *NAACL*.

Bo Li, Wei Ye, Zhonghao Sheng, Rui Xie, Xiangyu Xi, and Shikun Zhang. 2020. Graph enhanced dual attention network for document-level relation extraction. In *COLING*.

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*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *ICCV*.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *EMNLP*.

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 preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

Guoshun Nan, Zhijiang Guo, Ivan Sekulić, and Wei Lu. 2020. Reasoning with latent structure refinement for document-level relation extraction. In *ACL*.

Dat Quoc Nguyen and Karin Verspoor. 2018. Convolutional neural networks for chemical-disease relation extraction are improved with character-based word embeddings. *BioNLP Workshop*.

Sachin Pawar, Girish K Palshikar, and Pushpak Bhattacharyya. 2017. Relation extraction: A survey. *arXiv preprint arXiv:1712.05191*.

Hao Peng, Tianyu Gao, Xu Han, Yankai Lin, Peng Li, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2020. Learning from context or names? an empirical study on neural relation extraction. In *EMNLP*.

Meng Qu, Xiang Ren, Yu Zhang, and Jiawei Han. 2018. Weakly-supervised relation extraction by pattern-enhanced embedding learning. In *WWW*.

Sunil Kumar Sahu, Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Inter-sentence relation extraction with document-level graph convolutional neural network. In *ACL*.

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 *PAKDD*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2018. Graph attention networks. In *ICLR*.

Patrick Verga, Emma Strubell, and Andrew McCallum. 2018. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. In *NAACL*.

Difeng Wang, Wei Hu, Ermei Cao, and Weijian Sun. 2020. Global-to-local neural networks for document-level relation extraction. In *EMNLP*.

Hong Wang, Christfried Focke, Rob Sylvester, Nilesh Mishra, and William Wang. 2019. Fine-tune bert for docred with two-step process. *arXiv preprint 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 preprint arXiv:1910.03771*.

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

Ye Wu, Ruibang Luo, Henry CM Leung, Hing-Fung Ting, and Tak-Wah Lam. 2019. Renet: A deep learning approach for extracting gene-disease associations from literature. In *RECOMB*.

Yiqing Xie, Jiaming Shen, Sha Li, Yuning Mao, and Jiawei Han. 2022. Eider: Evidence-enhanced document-level relation extraction. In *ACL (Findings)*.

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. In *AAAI*.

Han Xu, Zhang Zhengyan, Ding Ning, Gu Yuxian, Liu Xiao, Huo Yuqi, Qiu Jiezhong, Zhang Liang, Han Wentao, Huang Minlie, et al. 2021b. Pre-trained models: Past, present and future. *arXiv preprint arXiv:2106.07139*.

Wang Xu, Kehai Chen, and Tiejun Zhao. 2021c. Discriminative reasoning for document-level relation extraction. In *ACL*.

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

Carl Yang, Yuxin Xiao, Yu Zhang, Yizhou Sun, and Jiawei Han. 2020. Heterogeneous network representation learning: A unified framework with survey and benchmark. *IEEE TKDE*.

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 *ACL*.

Deming Ye, Yankai Lin, Jiaju Du, Zhenghao Liu, Peng Li, Maosong Sun, and Zhiyuan Liu. 2020. Coreferential reasoning learning for language representation. In *EMNLP*.

Shuang Zeng, Yuting Wu, and Baobao Chang. 2021. Sire: Separate intra-and inter-sentential reasoning for document-level relation extraction. In *ACL*.

Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. 2020. Double graph based reasoning for document-level relation extraction. In *EMNLP*.

Ningyu Zhang, Xiang Chen, Xin Xie, Shumin Deng, Chuanqi Tan, Mosha Chen, Fei Huang, Luo Si, and Huajun Chen. 2021. Document-level relation extraction as semantic segmentation. In *IJCAI*.

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*.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. 2017. Position-aware attention and supervised data improve slot filling. In *EMNLP*.

Zhenyu Zhang, Bowen Yu, Xiaobo Shu, Tingwen Liu, Hengzhu Tang, Wang Yubin, and Li Guo. 2020. Document-level relation extraction with dual-tier heterogeneous graph. In *COLING*.

Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. 2019. Learning deep bilinear transformation for fine-grained image representation. In *NeurIPS*.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. In *AAAI*.

Wenxuan Zhou, Hongtao Lin, Bill Yuchen Lin, Ziqi Wang, Junyi Du, Leonardo Neves, and Xiang Ren. 2020. Nero: A neural rule grounding framework for label-efficient relation extraction. In *WWW*.

A Multi-Task Learning Pipeline by Supervising Intermediate Steps ($\text{SAIS}^{\text{O}}_{\text{All}}$)

To explicitly teach the model to capture relevant contexts and entity type information for RE, we design four tasks to supervise the model's outputs in the corresponding intermediate steps. We illustrate the overall multi-task pipeline $\text{SAIS}^{\text{O}}_{\text{All}}$ in Figure 4.

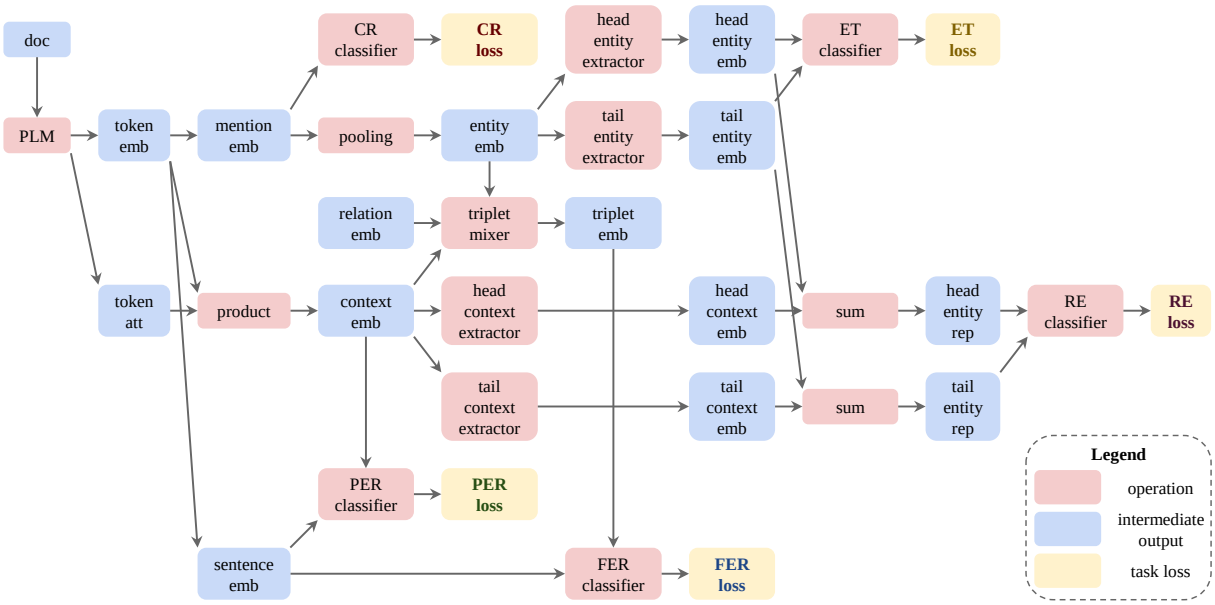

Figure 4: The overall multi-task learning pipeline of the proposed SAIS method ($\text{SAIS}^{\text{O}}_{\text{All}}$). By explicitly supervising the model's outputs in the intermediate steps via carefully designed tasks, we improve the RE performance.

B Ensemble Inference Algorithm with Evidence-based Data Augmentation ($\text{SAIS}^{\text{B}}_{\text{All}}$)

After training the multi-task pipeline $\text{SAIS}^{\text{O}}_{\text{All}}$ proposed in Section 3, we further boost the model performance by evidence-based data augmentation and ensemble inference as discussed in Section 4. The detailed steps are explained in Algorithm 1 below.

Algorithm 1: Evidence-based Data Augmentation and Ensemble Inference ($\text{SAIS}^{\text{B}}_{\text{All}}$)

input: trained pipeline $\text{SAIS}^{\text{O}}_{\text{All}}$ from Section 3, FER threshold α^{FER}, develop set $\mathcal{D}_{\text{dev}}$, test set $\mathcal{D}_{\text{test}}$

for $\mathcal{D} \in \{\mathcal{D}_{dev}, \mathcal{D}_{test}\}$ **do**

Original RE Prediction with $\text{SAIS}^{\text{O}}_{\text{All}}$ (Section 3.6)**:**

For $(e_h, e_t, r) \in \mathcal{D}$, get $\mathbb{L}^{\text{O}}_{h,t,r}$ from $\text{SAIS}^{\text{O}}_{\text{All}}$.

Identify the Uncertain Set $\mathcal{U}$ (Section 4.1)**:**

If $\mathcal{D}$ is $\mathcal{D}_{\text{dev}}$, calculate $\theta\%$ by minimizing $(\text{risk}^2 + \text{rejection rate}^2)$.

$\mathcal{U}$ contains triplets with the lowest $\theta\%$ absolute confidence $|\mathbb{L}^{\text{O}}_{h,t,r}|$.

Predict Evidence Probability for $(e_h, e_t, r) \in \mathcal{U}$ with $\text{SAIS}^{\text{O}}_{\text{All}}$ (Section 3.5)**:**

For $(e_h, e_t, r) \in \mathcal{U}$ and $s \in \mathcal{S}_d$ in the corresponding document d, get $\mathbb{P}^{\text{FER}}_{h,t,r,s}$ from $\text{SAIS}^{\text{O}}_{\text{All}}$.

Pseudo Document-based Data Augmentation $\text{SAIS}^{\text{D}}_{\text{All}}$ (Section 4.2)**:**

For $(e_h, e_t, r) \in \mathcal{U}$, get $\mathbb{L}^{\text{D}}_{h,t,r}$ by feeding the corresponding pseudo document into $\text{SAIS}^{\text{O}}_{\text{All}}$.

Attention Mask-based Data Augmentation $\text{SAIS}^{\text{M}}_{\text{All}}$ (Section 4.2)**:**

For $(e_h, e_t, r) \in \mathcal{U}$, get $\mathbb{L}^{\text{M}}_{h,t,r}$ by applying the corresponding attention mask to $\text{SAIS}^{\text{O}}_{\text{All}}$.

Ensemble Inference $\text{SAIS}^{\text{B}}_{\text{All}}$ (Section 4.2)**:**

If $\mathcal{D}$ is $\mathcal{D}_{\text{dev}}$, train τ_r for $r \in \mathcal{R}$ based on $\mathbb{L}^{\text{O}}_{h,t,r}$, $\mathbb{L}^{\text{D}}_{h,t,r}$, and $\mathbb{L}^{\text{M}}_{h,t,r}$ for $(e_h, e_t, r) \in \mathcal{U}$.

For $(e_h, e_t, r) \in \mathcal{U}$, get $\mathbb{L}^{\text{B}}_{h,t,r} = \mathbb{L}^{\text{O}}_{h,t,r} + \mathbb{L}^{\text{D}}_{h,t,r} + \mathbb{L}^{\text{M}}_{h,t,r} - \tau_r$.

Ultimate RE Prediction with $\text{SAIS}^{\text{B}}_{\text{All}}$ and $\text{SAIS}^{\text{O}}_{\text{All}}$ (Section 4.2 and 3.6)**:**

For $(e_h, e_t, r) \in \mathcal{U}$, extract relation r for entity pair (e_h, e_t) if $\mathbb{L}^{\text{B}}_{h,t,r} > 0$.

For $(e_h, e_t, r) \notin \mathcal{U}$, extract relation r for entity pair (e_h, e_t) if $\mathbb{L}^{\text{O}}_{h,t,r} > 0$.

Ultimate ER Prediction with $\text{SAIS}^{\text{O}}_{\text{All}}$ (Section 3.5)**:**

For predicted (e_h, e_t, r), retrieve $s \in \mathcal{S}_d$ in the corresponding document d if $\mathbb{P}^{\text{FER}}_{h,t,r,s} > \alpha^{\text{FER}}$.

output: sets of predicted triplet (e_h, e_t, r) and corresponding evidence $\mathcal{V}_{h,t,r}$ for $\mathcal{D}_{\text{dev}}$ and $\mathcal{D}_{\text{test}}$

C Experiment Details

We compare the proposed SAIS method against existing baselines based on three benchmarks: CDR (Li et al., 2016) and GDA (Wu et al., 2019) in Table 5, and DocRED (Yao et al., 2019) in Table 6. The details are explained in Section 5.

In particular, DocRED uses the MIT License, CDR is freely available for the research community, and GDA uses the GNU Affero General Public License. DocRED is constructed from Wikipedia and Wikidata and, therefore, contains information that names people. However, since our research focuses on identifying relations among real-world entities (including public figures) based on a given document, it is impossible to fully anonymize the dataset. We ensure that we only use publicly available information in our experiments. Our use of these datasets is consistent with their intended use. Although our method achieves state-of-the-art performance for RE and ER, using the predicted relations and evidence directly for downstream tasks without manual validation may increase the risk of errors carried forward due to the incorrect predictions. The experiments in this paper focus on English documents from biomedical and general domains, but our proposed framework can be easily extended to documents of other languages.

Model	**CDR**	**GDA**
BRAN (Verga et al., 2018)	62.1	-
CNN (Nguyen and Verspoor, 2018)	62.3	-
EoG (Christopoulou et al., 2019)	63.6	81.5
LSR (Nan et al., 2020)	64.8	82.2
SciBERT (Beltagy et al., 2019)	65.1	82.5
DHG (Zhang et al., 2020)	65.9	83.1
GLRE (Wang et al., 2020)	68.5	-
SSAN-SciBERT (Xu et al., 2021a)	68.7	83.7
ATLOP-SciBERT (Zhou et al., 2021)	69.4	83.9
SIRE-BioBERT (Zeng et al., 2021)	70.8	84.7
DocuNet-SciBERT (Zhang et al., 2021)	76.3	85.3
$\text{SAIS}^{\text{O}}_{\text{RE+CR+ET}}$-SciBERT (Ours)	**79.0** $\pm$ **0.8**	**87.1** $\pm$ **0.3**
$\text{SAIS}^{\text{O}}_{\text{RE+ET}}$-SciBERT	75.9 ± 0.9	86.1 ± 0.5
$\text{SAIS}^{\text{O}}_{\text{RE+CR}}$-SciBERT	74.5 ± 0.4	85.4 ± 0.2
$\text{SAIS}^{\text{O}}_{\text{RE}}$-SciBERT	72.8 ± 0.6	84.5 ± 0.3

Table 5: RE F1 results (%) on the CDR and GDA test sets. We report the baseline performances from the corresponding papers. $\text{SAIS}^{\text{O}}_{\text{RE+CR+ET}}$ using three training tasks (i.e., RE, CR, and ET) scores the highest on both datasets and better than its variants with fewer training tasks.

Model	DocRED Dev			DocRED Test		
	Relation		Evidence	Relation		Evidence
	Ign F1	F1	F1	Ign F1	F1	F1
CNN (Yao et al., 2019)	41.58	43.45	-	40.33	42.26	-
GAT (Veličković et al., 2018)	45.17	51.44	-	47.36	49.51	-
BiLSTM (Yao et al., 2019)	48.87	50.94	44.07	48.78	51.06	43.83
GCNN (Sahu et al., 2019)	46.22	51.52	-	49.59	51.62	-
EoG (Christopoulou et al., 2019)	45.94	52.15	-	49.48	51.82	-
AGGCN (Guo et al., 2019)	46.29	52.47	-	48.89	51.45	-
GEDA-BERT$_{base}$ (Li et al., 2020)	54.52	56.16	-	53.71	55.74	-
GLRE-BERT$_{base}$ (Wang et al., 2020)	-	-	-	55.40	57.40	-
LSR-BERT$_{base}$ (Nan et al., 2020)	52.43	59.00	-	56.97	59.05	-
HeterGSAN-BERT$_{base}$ (Xu et al., 2021d)	58.13	60.18	-	57.12	59.45	-
GAIN-BERT$_{base}$ (Zeng et al., 2020)	59.14	61.22	-	59.00	61.24	-
DRN-BERT$_{base}$ (Xu et al., 2021c)	59.33	61.39	-	59.15	61.37	-
SIRE-BERT$_{base}$ (Zeng et al., 2021)	59.82	61.60	-	60.18	62.05	-
BERT$_{base}$ (Wang et al., 2019)	-	54.16	-	-	53.20	-
BERT-TS$_{base}$ (Wang et al., 2019)	-	54.42	-	-	53.92	-
HIN-BERT$_{base}$ (Tang et al., 2020)	54.29	56.31	-	53.70	55.60	-
CorefBERT$_{base}$ (Ye et al., 2020)	55.32	57.51	-	54.54	56.96	-
E2GRE-BERT$_{base}$ (Huang et al., 2020)	55.22	58.72	47.14	-	-	-
SSAN-BERT$_{base}$ (Xu et al., 2021a)	57.03	59.19	-	56.06	58.41	-
ATLOP-BERT$_{base}$ (Zhou et al., 2021)	59.22	61.09	-	59.31	61.30	-
DocuNet-BERT$_{base}$ (Zhang et al., 2021)	59.86	61.83	-	59.93	61.86	-
Eider-BERT$_{base}$ (Xie et al., 2022)	**60.51**	62.48	50.71	60.42	62.47	51.27
SAIS$^{B}_{All}$-BERT$_{base}$ (Ours)	59.98 ± 0.13	**62.96 ± 0.11**	**53.70 ± 0.21**	**60.96**	**62.77**	**52.88**
BERT$_{large}$ (Ye et al., 2020)	56.51	58.70	-	56.01	58.31	-
CorefBERT$_{large}$ (Ye et al., 2020)	56.82	59.01	-	56.40	58.83	-
RoBERTa$_{large}$ (Ye et al., 2020)	57.19	59.40	-	57.74	60.06	-
CorefRoBERTa$_{large}$ (Ye et al., 2020)	57.35	59.43	-	57.90	60.25	-
SSAN-RoBERTa$_{large}$ (Xu et al., 2021a)	60.25	62.08	-	59.47	61.42	-
E2GRE-RoBERTa$_{large}$ (Huang et al., 2020)	-	-	-	60.30	62.50	50.50
ATLOP-RoBERTa$_{large}$ (Zhou et al., 2021)	61.32	63.18	-	61.39	63.40	-
DocuNet-RoBERTa$_{large}$ (Zhang et al., 2021)	62.23	64.12	-	62.39	64.55	-
Eider-RoBERTa$_{large}$ (Xie et al., 2022)	**62.34**	64.27	52.54	62.85	64.79	53.01
SAIS$^{B}_{All}$-RoBERTa$_{large}$ (Ours)	62.23 ± 0.15	**65.17 ± 0.08**	**55.84 ± 0.23**	**63.44**	**65.11**	**55.67**

Table 6: RE and ER results (%) on the develop and test sets of DocRED. Ign F1 refers to the F1 score excluding the relation instances mentioned in the train set. Baselines using BERT$_{base}$ are separated into the graph-based (upper) and transformer-based (lower) groups. We report the test set scores from the official scoreboard and the baseline scores from the corresponding papers. SAIS$^{B}_{All}$ achieves state-of-the-art performance on both RE and ER.