MD-RE: A Multi-Discrimination Framework for Document-Level Relation Extraction with Adaptive Threshold Shifted Loss

Anonymous ACL submission

Abstract

Document-level relation extraction (DocRE) aims to identify relations for an entity pair within a document. Existing methods can be broadly classified into two categories: direct encoding of the entire document or enhancement using extracted evidence sentences. However, the former often introduces noise unrelated to relations, while the latter is heavily dependent on the quality of evidence extraction. Moreover, these DocRE models typically use an adaptive threshold to predict all potential relations for an entity pair. As a result, class imbalance in DocRE often leads the model to learn a high throshold for an entity pair, which in turn causes the model to frequently predict that the entity pair has no relation. To address these issues, we propose a Multi-Discrimination framework (MD-RE) that does not rely on evidence sentences. MD-RE employs three discriminators with dynamically adjusted thresholds to independently predict relations, and aggregates their outputs via a weighted fusion strategy. Furthermore, we propose an Adaptive Threshold Shifted Loss (ATSL), which encourages lower threshold to alleviate the high false negative rate resulting from class imbalance. Experiments on three datasets demonstrate that MD-RE achieves new state-of-the-art results. In addition, ATSL significantly improves the performance of various existing DocRE models. Moreover, combining other losses with MD-RE also yields competitive results[1].

1 Introduction

Document-level relation extraction (DocRE) aims to extract relations for an entity pair from multiple sentences within a document. Since entities may span multiple sentences, the model must reason over more complex contexts and handle more relation types, making DocRE more challenging than sentence-level relation extraction. DocRE supports tasks such as knowledge graph construction (Mondal et al., 2021), information retrieval (Zeng et al., 2024), and question answering (Liu et al., 2024).

Most existing DocRE models are based on Transformer or graph-based architectures. Representative methods, such as ATLOP (Zhou et al., 2021), employ localized context pooling to guide attention toward relation-relevant regions, while KD-DocRE (Tan et al., 2022a) enhances multi-hop relation modeling via axial attention. In addition, some graph-based methods construct graphs and use graph neural networks to reason about relations between entities (Peng et al., 2022; Sun et al., 2023). Most of these models use the entire document as input context, but studies (Huang et al., 2021a,b) suggest that this may introduce noise irrelevant to relations. To mitigate this issue, recent works (Xie et al., 2022; Ma et al., 2023; Lu et al., 2023) propose extracting evidence sentences relevant to a given entity pair. However, these methods depend on the quality of evidence extraction and often exhibit limited effectiveness in low-resource scenarios (e.g., without evidence annotations).

To address these challenges, we propose a Multi-Discrimination framework (MD-RE), which does not rely on evidence sentences and reduces noise from the entire document through multi-discrimination perspectives. Specifically, MD-RE employs three discriminators with varying recall rates, each adopting a different threshold to determine the existence of relations for an entity pair. Higher-recall discriminators apply lower thresholds to retain more candidate relations, whereas lower-recall ones use higher thresholds to filter them more strictly. To enable each discriminator to use a different threshold for an entity pair, we propose a **L**oss-aware **N**egative **S**election (LNS) method: for each batch, we retain all positive examples[2] and

[1] Code: `https://anonymous.4open.science/r/MD-RE`

[2] If an entity pair has at least one relation, then the entity pair is a *positive example*, otherwise it is a *negative example*.

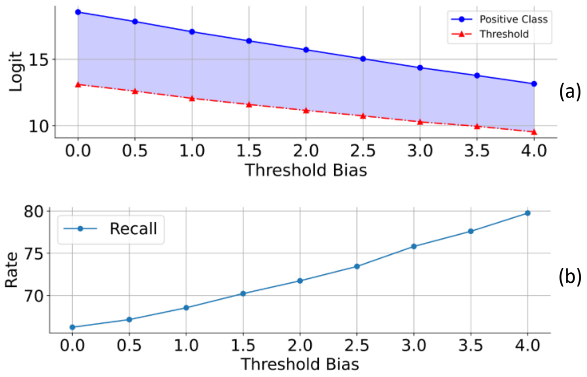


Figure 1: (a) Impact of ATSL threshold bias on the logits of the threshold and the positive class. (b) Recall changes across different threshold bias of ATSL.

select the top-k negative examples based on their loss. By reducing the number of negative examples, we can initially adjust the threshold and recall rate of each discriminator.

Furthermore, in order to more flexibly adjust the threshold and recall rate of each discriminator by introducing different threshold biases, we propose a novel **A**daptive **T**hreshold **S**hifted **L**oss (ATSL). The existing DocRE models usually employ an adaptive threshold loss (ATL) (Zhou et al., 2021) to predict relations for an entity pair, where a relation exists if its predicted logit exceeds the threshold. In the meantime, real-world datasets often face class imbalance[3]. This imbalance drives the model to learn a higher threshold for an entity pair, which in turn causes the model to frequently predict that the entity pair has no relation. An intuitive method is to lower the threshold at the initial stage of prediction, allowing more entity pairs to be classified as having relations. *Motivated by this idea*, we propose the ATSL, which introduces threshold biases based on the ATL loss, and improves the recall rate by lowering the threshold, thereby reducing the high false negative rate caused by class imbalance. In addition, we further observe that ATSL loss has *two key capabilities* that enable the discriminators to flexibly control both thresholds and recall rates: 1) ATSL can flexibly adjust the boundary between the positive class and the threshold, giving the model a more fine-grained logit judgment ability. In **Fig. 1(a)**, as the threshold bias increases, both the threshold and the positive class logit decrease, while the boundary between the positive class and the threshold shrinks; 2) ATSL also enables flexible control of the recall rate. In **Fig. 1(b)**, increasing the threshold bias leads to an increase in recall. To summarize, the contributions of our work are as follows:

- We propose a Multi-Discrimination framework (MD-RE) for DocRE, consisting of three discriminators with different discrimination criteria. Unlike previous methods, MD-RE does not rely on evidence sentences and effectively reduces document-level noise by incorporating multiple discrimination perspectives.
- We propose a Loss-aware Negative Selection (LNS) method to initially adjust the threshold and recall rate of each discriminator, and design a weighted fusion strategy to aggregate their outputs, aiming to achieve better prediction performance.
- We propose ATSL, a novel loss that introduces threshold biases to more flexibly adjust the threshold and recall rate of each discriminator and effectively mitigate class imbalance.
- Experiments on three datasets show that MD-RE achieves state-of-the-art (SOTA) results. In addition, ATSL consistently enhances performance and generalizes well across different baseline models. Moreover, combining other losses with MD-RE yields competitive results.

2 Related Work

Document-Level Relation Extraction. DocRE methods can be broadly categorized into: (1) directly encoding the entire document, such as GAIN (Zeng et al., 2020), ATLOP (Zhou et al., 2021), DocuNet (Zhang et al., 2021), KMGRE (Jiang et al., 2022), KD-DocRE (Tan et al., 2022a), TTM-RE (Gao et al., 2024), ABRE (Xu et al., 2024), and VaeDiff-DocRE (Tran et al., 2025); (2) introducing evidence sentences, including Eider (Xie et al., 2022), SAIS (Xiao et al., 2022), DREEAM (Ma et al., 2023), and AA (Lu et al., 2023).

Loss for DocRE. DocRE typically employs the ATL (Zhou et al., 2021) loss, which adaptively assigns a threshold to each entity pair, considering a relation to exist only when its predicted logit surpasses the threshold. Based on this, Tan et al. (2022a); Zhou and Lee (2022) find that the class imbalance is prevalent in DocRE. To address this, some subsequent studies have enhanced ATL, including Balanced-Softmax (Zhang et al., 2021),

[3] Class imbalance: negative examples significantly outnumber positive ones. For instance, in the Re-DocRED (Tan et al., 2022b) dataset, about 94% of entity pairs have no relation.

AML (Wei and Li, 2022), AFL (Tan et al., 2022a), SSR-PU (Wang et al., 2022), NCRL (Zhou and Lee, 2022), PEMSCL (Guo et al., 2023), and HingeABL (Wang et al., 2023). AML (Wei and Li, 2022) and HingeABL (Wang et al., 2023) aim to maximize the margin between positive and negative classes.

The above losses mitigate the class imbalance to some extent by optimizing the decision boundary between positive and negative classes. However, they lack the flexibility to adjust thresholds and recall rates, and their effectiveness in addressing class imbalance remains limited. To this end, we extend ATL by proposing the Adaptive Threshold Shifted Loss (ATSL).

Incomplete Labeling in DocRE. Due to prevalent false negatives in the DocRED (Yao et al., 2019) dataset, Tan et al. (2022b) introduces the revised version, Re-DocRED. To evaluate model robustness under incomplete annotations, a new task is proposed: train on DocRED, test on Re-DocRED. Some of the effective methods for this task include SSR-PU (Wang et al., 2022), CAST (Tan et al., 2023), and P^{3}M (Wang et al., 2024).

3 Methodology

Our MD-RE framework in **Fig. 2** consists of four main parts: Document Encoding module, Discrimination and Loss-aware Negative Selection module, Fusion module, and our loss ATSL.

3.1 Problem Formulation

The goal of DocRE is to predict relations $R \cup \{\textsc{na}\}$ for entity pairs $(e_h, e_t)_{h,t=1}^{n}, h \neq t$ within a document D. Here, $\{e_i\}_{i=1}^{n}$ denotes the set of entities in the document, and e_h and e_t refer to the head and tail entities, respectively. R is a predefined set of relations, while NA indicates the absence of any relation. For a given entity pair T=(e_h, e_t), the positive classes $\mathcal{P}_T \subseteq R$ correspond to relations expressed by any entity mention pair of e_h and e_t, whereas the negative classes $\mathcal{N}_T \subseteq R$ represent relations not expressed between them. If T expresses no relations, $\mathcal{P}_T$ is empty, and $\mathcal{N}_T = R$.

3.2 Document Encoding Module

The token sequence of a document D is denoted as $\mathrm{T}_D = \{t_i\}_{i=1}^{|\mathrm{T}_D|}$, where a special token "*" is inserted at the beginning and end of each entity mention. Following ATLOP (Zhou et al., 2021) and DREEAM (Ma et al., 2023), we obtain token-level hidden states $\mathrm{H} \in \mathbb{R}^{|\mathrm{T}_D| \times d}$ and attention weights $\mathrm{A} \in \mathbb{R}^{|\mathrm{T}_D| \times |\mathrm{T}_D|}$ by averaging outputs and last-head attentions from the last three encoder layers, respectively, where d is the hidden size:

$$\mathrm{H, A} = \mathrm{PLM}(\mathrm{T}_D) \tag{1}$$

The embedding h_e for each entity e is obtained by aggregating information from all its mentions $\mathrm{M}_e = \{m_i\}_{i=1}^{|\mathrm{M}_e|}$, where H_{m_i} denotes the embedding of the special token "*" that marks the starting position of the i-th mention:

$$h_e = \log \sum_{i=1}^{|\mathrm{M}_e|} \exp(\mathrm{H}_{m_i}) \tag{2}$$

Then we use the localized context pooling method to compute $c_{h,t}$ from token embeddings $\mathbf{H}$ and cross-token attention $\mathbf{A}$, where $\mathbf{A}_h$ and $\mathbf{A}_t$ represent attention for entities e_h and e_t, and $\otimes$ denotes element-wise product.

$$\mathrm{c}_{h,t} = \mathrm{H}^\top \frac{\mathrm{A}_h \otimes \mathrm{A}_t}{\mathrm{A}_h^\top \mathrm{A}_t} \tag{3}$$

The localized context $c_{h,t}$ is concatenated with the head h_{e_h} and tail entity h_{e_t} individually. Here, $\|$ denotes concatenation, W_h, W_t are trainable weights, and b_h, b_t are the biases for head and tail entities. Finally, z_h and z_t are fed into a bilinear classifier to compute the relation scores $\mathrm{logit}_{h,t}$ for the entity pair (e_h, e_t):

$$\begin{aligned} z_h &= \tanh(\mathrm{W}_h[h_{e_h} \| c_{h,t}] + b_h) \\ z_t &= \tanh(\mathrm{W}_t[h_{e_t} \| c_{h,t}] + b_t) \end{aligned} \tag{4}$$

$$\mathrm{logit}_{h,t} = \mathrm{z}_h^\top W_r z_t + b_r \tag{5}$$

3.3 Discrimination and Loss-aware Negative Selection Module

The core idea of our MD-RE framework is to progressively refine candidate relations through multiple stages. To achieve this, we employ three discriminators with distinct criteria, each applying a differently adjustable threshold to determine whether relations exist between an entity pair. The *motivation for designing the three discriminators and the details of each one* are described in the corresponding sections below.

Recall Discriminator. This discriminator is designed to achieve a high recall rate by applying lower thresholds to retain more candidate relations. Specifically, given an entity pair, we obtain its $\mathrm{logit}_{h,t}$ (see **Eq. (5)**), and then evaluate it using our ATSL loss (see **Section 3.5**). By adjusting the hyperparameter λ of ATSL, we can flexibly control the threshold and recall rate of the discriminator.

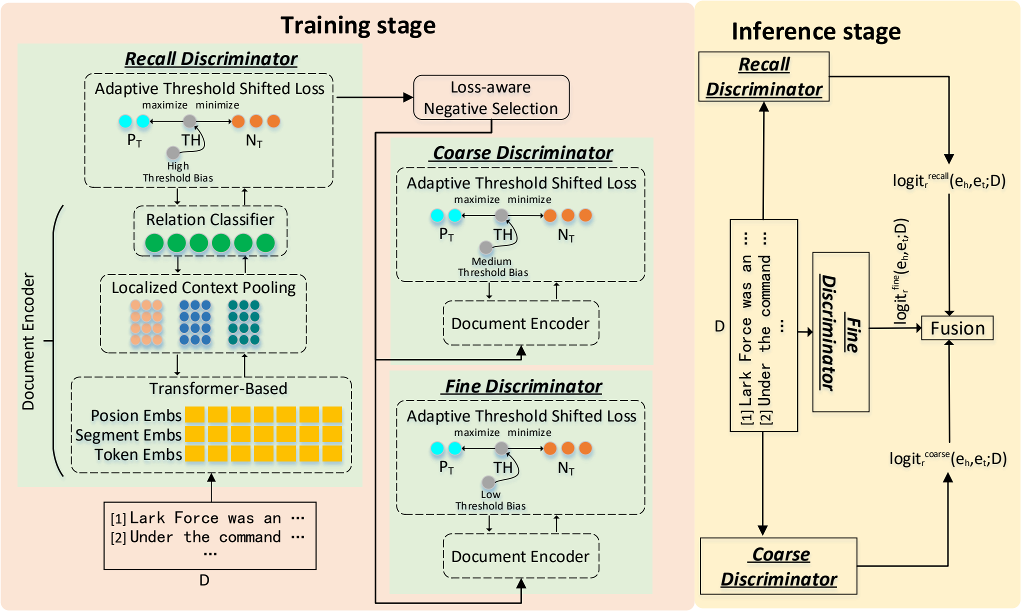


Figure 2: Overview of our MD-RE framework. In training, three discriminators with distinct decision criteria apply different thresholds to determine whether a relation exists for an entity pair. The Loss-aware Negative Selection (LNS) and Adaptive Threshold Shifted Loss (ATSL) jointly enable dynamic adjustment of threshold and recall for each discriminator. In inference, we adopt a weighted fusion strategy to integrate the outputs of three discriminators.

Loss-aware Negative Selection. After the recall discriminator, there are still a large number of negative samples, most of which are easy for the model to classify. This causes the model to focus on trivial instances while overlooking harder, more informative ones. To address this, we introduce the Loss-aware Negative Selection (LNS) method. By applying LNS, we effectively reduce the impact of too many negative samples, which helps to improve the performance of subsequent coarse and fine discriminators. In addition, LNS can also cooperate with ATSL loss to further dynamically adjust the threshold and recall of each discriminator. Specifically, for each batch, we retain all positive examples and select the Top-k negative examples based on their loss, defined as:

$$\begin{aligned} k &= \min(\rho \cdot |\mathcal{S}_{\text{pos}}|, |\mathcal{S}_{\text{neg}}|) \\ \mathcal{S}_{\text{neg-hard}} &= \text{Top-}k\left(\mathcal{S}_{\text{neg}}\right) \end{aligned} \tag{6}$$

Here, $\mathcal{S}_{\text{pos}}$ and $\mathcal{S}_{\text{neg}}$ denote the sets of positive and negative examples within the batch, respectively. ρ controls the ratio of selected negatives relative to the number of positives.

Coarse Discriminator. Subsequently, we feed all positive samples along with the negative samples selected by the LNS method into the coarse discriminator, and incorporate the ATSL loss to dynamically set a moderate threshold for each entity pair. This discriminator aims to further filter candidate relations at a coarse granularity.

Fine Discriminator. The fine discriminator is trained similarly to the coarse discriminator but with a higher threshold. It applies stricter criteria than the coarse discriminator to further refine candidate relations and improve prediction reliability.

3.4 Fusion Module

Since the three discriminators use different decision criteria, we design a weighted fusion strategy to effectively integrate their outputs and make the final decision, fully leveraging their respective strengths. Specifically, we directly accept the prediction for a triplet (h, r, t) if all three discriminators predict its existence; otherwise, we compute a fused score:

$$\text{logit}^{\text{final}}_{h,r,t} = \text{logit}^{\text{recall}}_{h,r,t} + \text{logit}^{\text{coarse}}_{h,r,t} + \text{logit}^{\text{fine}}_{h,r,t} \tag{7}$$

We further define an adaptive threshold based on the threshold of each discriminator:

$$\text{logit}^{\text{final}}_{h,\text{TH},t} = \alpha \cdot \text{logit}^{\text{recall}}_{h,\text{TH},t} + \text{logit}^{\text{coarse}}_{h,\text{TH},t} + \text{logit}^{\text{fine}}_{h,\text{TH},t} \tag{8}$$

A relation r exists if $\text{logit}^{\text{final}}_{h,r,t} > \text{logit}^{\text{final}}_{h,\text{TH},t}$. The method leverages complementary decision patterns of discriminators to improve performance. Since the recall discriminator plays a more dominant role in controlling recall, we apply the weighting factor α only to its threshold.

3.5 Loss Design

An Empirical Analysis of ATL. As shown in **Eq. (9)**, the Adaptive Threshold Loss (ATL) (Zhou et al., 2021) divides the set R of predefined relations into two subsets: the positive classes $\mathcal{P}_T$ and the negative classes $\mathcal{N}_T$, with an external threshold class TH used to distinguish between them. The objective is to encourage the logits of $\mathcal{P}_T$ to be higher than that of the TH class, and the logits of $\mathcal{N}_T$ to be lower than that of the TH class.

$$\begin{gathered}\mathcal{L}_1 = -\sum_{r \in \mathcal{P}_T} \log \left(\frac{\exp(\mathrm{logit}_r)}{\sum_{r' \in \mathcal{P}_T \cup \{\mathrm{TH}\}} \exp(\mathrm{logit}_{r'})} \right) \\ \mathcal{L}_2 = -\log \left(\frac{\exp(\mathrm{logit}_{\mathrm{TH}})}{\sum_{r' \in \mathcal{N}_T \cup \{\mathrm{TH}\}} \exp(\mathrm{logit}_{r'})} \right) \\ \mathcal{L}_{ATL} = \mathcal{L}_1 + \mathcal{L}_2\end{gathered} \tag{9}$$

Wang et al. (2023)'s analysis finds a significant difference in the number of relations between $\mathcal{P}_T$ and $\mathcal{N}_T$, with $\mathcal{N}_T$ being larger, causing $\mathcal{L}_2$ to dominate the loss calculation. Building on this analysis, we find that the dominance of $\mathcal{L}_2$ stems not only from the imbalance between $\mathcal{P}_T$ and $\mathcal{N}_T$, but more fundamentally from the overwhelming proportion of negative examples. When no relation exists between an entity pair, $\mathcal{P}_T$ is empty, resulting in no contribution from $\mathcal{L}_1$, and the loss is solely determined by $\mathcal{L}_2$. Since most entity pairs have no relation, $\mathcal{L}_2$ dominates. Building on this and Wang et al. (2023), we reformulate $\mathcal{L}_2$ as in **Eq. (10)**. When $\mathrm{logit}_{r'} - \mathrm{logit}_{\mathrm{TH}} \to -\infty$, $\mathcal{L}_2 \to 0$, indicating that $\mathrm{logit}_{\mathrm{TH}} \gg \mathrm{logit}_{r'}$. This suggests that ATL learns a relatively high threshold for an entity pair.

$$\begin{aligned}\mathcal{L}_2 &= -\log \left(\frac{\exp(\mathrm{logit}_{\mathrm{TH}})}{\sum_{r' \in \mathcal{N}_T \cup \{\mathrm{TH}\}} \exp(\mathrm{logit}_{r'})} \right) \\ &= -\log \left(\frac{1}{1 + \sum_{r' \in \mathcal{N}_T} \exp(\mathrm{logit}_{r'} - \mathrm{logit}_{\mathrm{TH}})} \right)\end{aligned} \tag{10}$$

Adaptive Threshold Shifted Loss. As analyzed above, we extend the findings of Wang et al. (2023) and further reveal the limitations: when the number of entity pairs with no relations significantly exceeds the number of entity pairs with relations (class imbalance), the logit of threshold class TH increases and eventually surpasses the logits of many candidate relations, leading to a large number of false negative predictions.

To address this issue, we introduce a threshold bias $\lambda > 0$ in the TH class to ensure that:

$$\begin{aligned}\mathcal{L}_2' &= -\log \left(\frac{\exp(\mathrm{logit}_{\mathrm{TH}} + \lambda)}{\exp(\mathrm{logit}_{\mathrm{TH}} + \lambda) + \sum_{r' \in \mathcal{N}_T} \exp(\mathrm{logit}_{r'})} \right) \\ &= -\log \left(\frac{1}{1 + \sum_{r' \in \mathcal{N}_T} \exp(\mathrm{logit}_{r'} - (\mathrm{logit}_{\mathrm{TH}} + \lambda))} \right)\end{aligned} \tag{11}$$

Minimizing $\mathcal{L}_2'$ requires that:

$$\mathrm{logit}_{r'} - (\mathrm{logit}_{\mathrm{TH}} + \lambda) \to -\infty \tag{12}$$

Consequently, we have:

$$\begin{gathered}\mathrm{logit}_{\mathrm{TH}} + \lambda \gg \mathrm{logit}_{r'} \\ \mathrm{logit}_{\mathrm{TH}} \gg \mathrm{logit}_{r'} - \lambda\end{gathered} \tag{13}$$

This shows that λ effectively reduces the $\mathrm{logit}_{\mathrm{TH}}$ in the optimization process. From **Eq.** (13), adding λ to the target logit boosts its value in the softmax calculation, allowing the model to achieve the same margin with a smaller $\mathrm{logit}_{\mathrm{TH}}$.

Similarly, we add a threshold bias β to the other part of the loss $\mathcal{L}_1'$, as shown in **Eq. (14)**. When $\mathcal{L}_1' \to 0$, it implies that $\mathrm{logit}_{\mathrm{TH}} + \beta - \mathrm{logit}_r \to -\infty$. From this, we can derive that $\mathrm{logit}_{\mathrm{TH}} + \beta \ll \mathrm{logit}_r$.

$$\begin{aligned}\mathcal{L}_1' &= -\sum_{r \in \mathcal{P}_T} \log \left(\frac{\exp(\mathrm{logit}_r)}{\exp(\mathrm{logit}_{\mathrm{TH}} + \beta) + \sum_{r' \in \mathcal{P}_T} \exp(\mathrm{logit}_{r'})} \right) \\ &= -\sum_{r \in \mathcal{P}_T} \log \left(\frac{1}{1 + \exp(\mathrm{logit}_{\mathrm{TH}} + \beta - \mathrm{logit}_r) + \sum_{r' \in \mathcal{P}_T, r' \neq r} \exp(\mathrm{logit}_{r'} - \mathrm{logit}_r)} \right)\end{aligned} \tag{14}$$

Finally, we obtain the **A**daptive **T**hreshold **S**hifted **L**oss (ATSL), as shown in **Eq. (15)**.

$$\begin{gathered}\mathcal{L}_1' = -\sum_{r \in \mathcal{P}_T} \log \left(\frac{\exp(\mathrm{logit}_r)}{\exp(\mathrm{logit}_{\mathrm{TH}} + \beta) + \sum_{r' \in \mathcal{P}_T} \exp(\mathrm{logit}_{r'})} \right) \\ \mathcal{L}_2' = -\log \left(\frac{\exp(\mathrm{logit}_{\mathrm{TH}} + \lambda)}{\exp(\mathrm{logit}_{\mathrm{TH}} + \lambda) + \sum_{r' \in \mathcal{N}_T} \exp(\mathrm{logit}_{r'})} \right) \\ \mathcal{L}_{ATSL} = \mathcal{L}_1' + \mathcal{L}_2'\end{gathered} \tag{15}$$

4 Experimental Setup

Implementation Details. Our experiments are implemented using PyTorch (Paszke, 2019) and Transformers (Wolf et al., 2020), using BERT$_{\mathrm{base}}$ (Devlin et al., 2019) and RoBERTa$_{\mathrm{large}}$ (Liu et al., 2019) as encoders. See **Appendix A.1** for details.

Datasets and Metrics. We experiment on the DocRED (Yao et al., 2019), DWIE (Zaporojets et al., 2021), and Re-DocRED (Tan et al., 2022b) datasets, which are detailed in **Appendix A.2**. We follow Zhou et al. (2021) and evaluate using F1 and Ign-F1, where **F1 represents** the standard F1, while **Ign-F1 is** computed by excluding relational facts shared between the train and dev/test sets.

5 Main Results and Analysis

We conduct experiments to answer questions about our main contributions: MD-RE and ATSL.

- **Q1:** How does our MD-RE framework perform? (Section 5.1)

Model	Dev		Test	
	F1	Ign-F1	F1	Ign-F1
with BERT$_{\text{base}}$				
ATLOP (Zhou et al., 2021)	74.22 $^{\diamond}$	73.35 $^{\diamond}$	74.02 $^{\diamond}$	73.22 $^{\diamond}$
DocuNET (Zhang et al., 2021)	74.65 $^{\diamond}$	73.68 $^{\diamond}$	74.49 $^{\diamond}$	73.60 $^{\diamond}$
KD-DocRE (Tan et al., 2022a)	74.69 $^{\diamond}$	73.76 $^{\diamond}$	74.55 $^{\diamond}$	73.67 $^{\diamond}$
DREEAM (Ma et al., 2023)	74.58 $^{\triangle}$	73.74 $^{\triangle}$	74.23 $^{\triangle}$	73.42 $^{\triangle}$
CAST (Tan et al., 2023)	-	-	74.67 ‡	73.32 ‡
SA-KD (Zhang et al., 2023)	75.85 $^{\diamond}$	75.03 $^{\diamond}$	75.77 $^{\diamond}$	74.85 $^{\diamond}$
ABRE (Xu et al., 2024)	$\underline{76.26}$ *	$\underline{75.54}$ *	$\underline{76.30}$ *	$\underline{75.70}$ *
VaeDiff-DocRE (Tran et al., 2025)	75.89 ‡	74.96 ‡	75.07 ‡	74.13 ‡
MD-RE (ours)	**77.70±0.10**	**76.46±0.07**	**77.80±0.04**	**76.63±0.02**
with RoBERTa$_{\text{large}}$				
ATLOP (Zhou et al., 2021)	77.63 †	76.88 †	77.73 †	76.94 †
DocuNET (Zhang et al., 2021)	78.16 †	77.53 †	77.92 †	77.27 †
KD-DocRE (Tan et al., 2022a)	78.65 †	77.92 †	78.35 †	77.63 †
PEMSCL (Guo et al., 2023)	79.89 †	79.02 †	79.86 †	79.01 †
AA (Lu et al., 2023)	$\underline{81.15}$ †	$\underline{80.04}$ †	$\underline{81.20}$ †	$\underline{80.12}$ †
TTM-RE (Gao et al., 2024)	78.13 *	78.05 *	79.95 *	78.20 *
VaeDiff-DocRE (Tran et al., 2025)	79.19 ‡	78.35 ‡	79.03 ‡	78.22 ‡
MD-RE (ours)	**81.44±0.12**	**80.38±0.12**	**81.49±0.05**	**80.45±0.06**

Table 1: Results on Re-DocRED. The underlined values indicate the results of the previous SOTA. † from Lu et al. (2023), * from original paper, ◇ from Zhang et al. (2023), ‡ from Tran et al. (2025), and △ our reproduced results.

Model	Dev		Test	
	F1	Ign-F1	F1	Ign-F1
GAIN *	62.55	58.63	67.57	62.37
ATLOP *	69.96	63.57	74.36	67.56
KMGRE *	71.40	65.56	76.71	69.94
MILR $^{\diamond}$	$\underline{72.05}$	$\underline{67.18}$	$\underline{76.51}$	$\underline{69.84}$
DREEAM ‡	72.40	65.93	74.66	67.27
TTM-RE ‡	64.51	56.62	65.01	54.71
MD-RE (ours)	**73.81±0.32**	**68.37±0.36**	**78.32±0.12**	**71.28±0.10**

Table 2: Results on DWIE with BERT$_{\text{base}}$. * from Jiang et al. (2022), $^{\diamond}$ from original paper, and ‡ from ours.

Model	Test	
	F1	Ign-F1
ATLOP (Zhou et al., 2021) *	45.19	45.09
DocuNET (Zhang et al., 2021) †	45.99	45.88
KD-DOcRE (Tan et al., 2022a) †	47.57	47.32
SSR-PU (Wang et al., 2022) *	59.50	58.68
CAST (Tan et al., 2023) †	$\underline{65.32}$	$\underline{64.25}$
P^{3}M (Wang et al., 2024) *	64.34	63.16
MD-RE (ours)	**65.93±0.04**	**64.96±0.03**

Table 3: Results on DocRED using RoBERTa$_{\text{large}}$. * Results from Wang et al. (2024); † from Tan et al. (2023).

- **Q2:** How effective is our ATSL loss when applied to different models? (Section 5.2)
- **Q3:** How does the performance of our ATSL loss compare to other losses? (Section 5.2)
- **Q4:** An ablation study. (Section 5.3)

5.1 Main Results

Results on Re-DocRED. As shown in **Table 1**, our MD-RE framework consistently outperforms all strong baselines and the previous SOTA models. Specifically, with the BERT$_{\text{base}}$ encoder, MD-RE achieves F1 of 77.70 and 77.80 on the dev and test sets, respectively, outperforming the previous SOTA model ABRE by 1.44 and 1.50. Similarly, with the RoBERTa$_{\text{large}}$ encoder, MD-RE achieves F1 of 81.44 and 81.49 on the dev and test sets, respectively, outperforming the SOTA model AA by 0.29 points on both splits.

Results on DWIE. As shown in **Table 2**, MD-RE consistently outperforms baseline models on the DWIE dataset, reaching 73.81 F1 and 68.37 Ign-F1 on the dev set, and 78.32 F1 and 71.28 Ign-F1 on the test set. Compared with the strong baseline MILR, MD-RE improves the F1 and Ign-F1 on the test set by 1.81 and 1.44, respectively. Notably, MD-RE also surpasses the recent DREEAM model, further demonstrating its effectiveness across challenging DocRE benchmarks.

Results on DocRED. To evaluate the weakly supervised generalization ability of MD-RE, we train on the incomplete dataset DocRED and test on the fully annotated dataset Re-DocRED. **Table 3** shows that MD-RE framework achieves the best results among all compared methods, with an F1 of 65.93 and an Ign-F1 of 64.96 on the test set. Compared to the previous competitive model CAST, MD-RE obtains gains of 0.61 F1 and 0.71 Ign-F1. When compared with other strong baselines such as P^{3}M and SSR-PU, MD-RE's improvement ranges from a minimum of 1.59 to a maximum of 6.43.

Model	Dev				Test			
	F1	F1 with ATSL	Ign-F1	Ign-F1 with ATSL	F1	F1 with ATSL	Ign-F1	Ign-F1 with ATSL
Re-DocRED with $\text{BERT}_{\text{base}}$								
ATLOP (Zhou et al., 2021)	74.22 △	**76.23** (+2.01)	73.35 △	**74.83** (+1.48)	74.02 △	**76.48** (+2.46)	73.22 △	**75.12** (+1.90)
DocuNet (Zhang et al., 2021)	74.65 △	**76.26** (+1.61)	73.68 △	**74.81** (+1.13)	74.49 △	**76.45** (+1.96)	73.60 △	**75.07** (+1.47)
KD-DocRE (Tan et al., 2022a)	74.69 △	**76.70** (+2.01)	73.76 △	**75.52** (+1.76)	74.55 △	**76.65** (+2.10)	73.67 △	**75.50** (+1.83)
DREEAM (Ma et al., 2023)	74.58 †	**76.08** (+1.50)	73.74 †	**74.81** (+1.07)	74.23 †	**76.14** (+1.91)	73.42 †	**74.92** (+1.50)
TTM-RE (Gao et al., 2024)	76.21 †	**80.16** (+3.95)	74.74 †	**79.05** (+4.31)	76.33 †	**80.51** (+4.18)	74.89 †	**79.48** (+4.59)
Re-DocRED with $\text{RoBERTa}_{\text{large}}$								
ATLOP (Zhou et al., 2021)	77.63 *	**80.35** (+2.72)	76.88 *	**79.22** (+2.34)	77.73 *	**80.40** (+2.67)	76.94 *	**79.29** (+2.35)
DocuNet (Zhang et al., 2021)	78.16 *	**79.76** (+1.60)	77.53 *	**78.78** (+1.25)	77.92 *	**79.85** (+1.93)	77.27 *	**78.91** (+1.64)
KD-DocRE (Tan et al., 2022a)	78.65 *	**79.06** (+0.41)	77.92 *	**78.07** (+0.15)	78.35 *	**78.76** (+0.41)	77.63 *	**77.78** (+0.15)
DREEAM (Ma et al., 2023)	77.60 †	**79.56** (+1.96)	77.20 †	**78.62** (+1.42)	77.94 ◇	**79.86** (+1.92)	77.34 ◇	**78.96** (+1.62)
TTM-RE (Gao et al., 2024)	78.13 ◇	**82.57** (+4.44)	78.05 ◇	**81.70** (+3.65)	79.95 ◇	**82.36** (+2.41)	78.20 ◇	**81.53** (+3.33)

Table 4: Performance of different DocRE models using ATSL loss. * from Lu et al. (2023), △ from Zhang et al. (2023), ◇ from original paper, and † our reproduced results.

Loss Function	F1	Ign-F1
ATL (Zhou et al., 2021) *	73.29	72.46
Balanced-Softmax (Zhang et al., 2021) *	73.68	72.85
AML (Wei and Li, 2022) *	72.60	71.78
AFL (Tan et al., 2022a) *	74.15	73.20
$\text{HingeABL}_{\text{SAT}}$ (Wang et al., 2023) *	73.46	72.61
$\text{HingeABL}_{\text{MeanSAT}}$ (Wang et al., 2023) *	74.68	72.90
HingeABL (Wang et al., 2023) *	75.15	73.84
ATSL (Our Loss)	**76.48** (1.33↑)	**75.12** (1.28↑)

Table 5: Results of different losses on Re-DocRED test set. * from Wang et al. (2023). All results use ATLOP (Zhou et al., 2021) and $\text{BERT}_{\text{base}}$ for encoding.

Model	F1	Ign-F1
MD-RE (ours)	**77.70**	**76.46**
w/o LNS	77.52	76.27
w/o Fusion w Pipeline	72.83	72.27
w/o Fusion w Add	77.02	75.53
w/o Coarse Discriminator	77.18	75.67
w/o Fine Discriminator	76.38	74.58

Table 6: An ablation study on the Re-DocRED dev set using $\text{BERT}_{\text{base}}$ as the encoder, where w/o denotes removal and w indicates inclusion.

5.2 Results of ATSL

Different DocRE Models with ATSL. To evaluate the generality of our loss, we *apply ATSL to different models by replacing their original losses.* Specifically, ATLOP and DREEAM use ATL loss, DocuNet uses Balanced-Softmax loss, KD-DocRE uses AFL loss, and TTM-RE uses S-PU loss.

Table 4 shows that the ATSL loss significantly improves the performance of all baseline models. Specifically, the TTM-RE model with $\text{BERT}_{\text{base}}$ achieves improvements of **4.18** in F1 and **4.59** in Ign-F1 on the test set. Similarly, the ATLOP model with $\text{RoBERTa}_{\text{large}}$ achieves gains of **2.67** in F1 and **2.35** in Ign-F1. Moreover, our loss achieves an average improvement of 2.52 in F1 on the BERT test set, and 2.23 in F1 on the RoBERTa dev set. These results demonstrate the generality of ATSL in enhancing different DocRE models.

ATSL vs. Other Loss. To verify the effectiveness of ATSL compared to other losses, we evaluate them using the ATLOP and $\text{BERT}_{\text{base}}$ encoder. **Table 5** shows that ATSL achieves 76.48 in F1 and 75.12 in Ign-F1, outperforming other losses. Specifically, compared to the current SOTA HingeABL loss, ATSL improves by 1.33 in F1 and 1.28 in Ign-F1, significantly boosting model performance and demonstrating its effectiveness in DocRE task.

5.3 Ablation Study

We perform an ablation study to assess each component's impact. See **Table 6** for results.

w/o LNS. After removing LNS, the F1 drops by 0.18, indicating that removing it does not affect MD-RE's ability to set different thresholds for individual discriminators, thanks to the ATSL loss. In addition, the negative samples selected by LNS also bring a slight performance gain.

w/o Fusion w Pipeline. Replacing the fusion module with a pipeline method severely reduces performance, as it only retains triplets unanimously judged as true by all three discriminators, discarding many potential relations and lowering recall.

w/o Fusion w Add. Replacing the fusion module with a union method results in a performance drop. Results indicate that the fusion module effectively integrates information from different discriminators and exploits their complementary capabilities.

w/o Coarse Discriminator. Removing the coarse discriminator results in a slight decrease of 0.52 in the F1. This suggests that the coarse discriminator

Loss + Model	F1	Ign-F1
ATL loss (Zhou et al., 2021)		
-ATLOP	74.02	73.22
-MD-RE (ours)	**77.06** (3.04↑)	**76.02** (2.80↑)
AFL loss (Tan et al., 2022a)		
-KD-DocRE	74.55	73.67
-MD-RE (ours)	**77.41** (2.86↑)	**76.19** (2.52↑)

Table 7: MD-RE vs. baselines under same losses, all using $\text{BERT}_{\text{base}}$ on the Re-DocRED test set.

brings a slight improvement.

w/o Fine Discriminator. Removing the fine discriminator results in a slightly larger performance drop than removing the coarse one, with the F1 decreasing by 1.32. This suggests that the fine discriminator contributes more to refining the process.

6 Further Analysis

To further investigate our method's performance, we answer the following research questions:

- **Q5:** How does our MD-RE compare to other models under the same loss? (Section 6.1)
- **Q6:** Does the ATSL loss alleviate the class imbalance? (Section 6.2)
- **Q7:** How does our MD-RE compare to other models in resource efficiency? (Section 6.3)
- **Q8:** How does the hyperparameter λ influence the performance of ATSL? (Appendix B.1)
- **Q9:** A case study. (Appendix B.2)

Due to space limitations, we provide detailed analysis for Q8-Q9 questions in Appendix B.

6.1 MD-RE vs. Baselines under Same Losses

To verify the effectiveness and generalizability of the MD-RE framework, **Table 7** compares MD-RE with strong baselines under the same losses. Under the ATL loss, MD-RE significantly outperforms ATLOP, achieving a 3.04 and 2.80 improvement in F1 and Ign-F1 scores, respectively. Similarly, when trained with the AFL loss, MD-RE surpasses KD-DocRE by 2.86 in F1 and 2.52 in Ign-F1. These consistent improvements across losses demonstrate MD-RE's effectiveness and generalizability.

6.2 Analyzing the Class Imbalance

Model	FN ↓	FP ↓	FN/(FN+FP) ↓
ATLOP (Zhou et al., 2021)	5833	1942	0.75
ATLOP with ATSL (ours)	4181	4029	0.51

Table 8: FN (False Negative): Predicts a positive example as negative. FP (False Positive): Predicts a negative example as positive. $\text{BERT}_{\text{base}}$ on Re-DocRED dev set.

Method	Memory (GiB)	Training Time (Min)	F1
without evidence			
ATLOP	10.37	55.80	74.02
KD-DocRE	14.71	169.59	74.55
TTM-RE	20.30	208.37	-
with evidence			
Eider	43.10 *	-	-
SAIS	46.20 *	-	-
DREEAM	13.46	59.02	74.23
MD-RE (ours)	17.63	81.80	**77.80**

Table 9: Resource usage comparison. * from Ma et al. (2023). $\text{BERT}_{\text{base}}$ on Re-DocRED with batch size 4.

To illustrate the effectiveness of our ATSL in alleviating the class imbalance, we apply ATSL to the ATLOP and examine the number of two prediction patterns: FN and FP. **Table 8** shows that applying ATSL leads to a substantial reduction in false negatives, decreasing from 5833 to 4181. The ratio FN/(FN+FP) also drops from 0.75 to 0.51, indicating that the false negative issue caused by class imbalance is effectively mitigated. However, this improvement comes at the cost of a noticeable increase in false positives, suggesting that while ATSL helps capture more true positives, it may also introduce more incorrect predictions.

6.3 Resource Efficiency

To assess the resource efficiency of MD-RE, **Table 9** presents a comparison of memory usage and training time. MD-RE demonstrates favorable efficiency, requiring only 17.63 GiB of memory, which is lower than that of TTM-RE (without evidence), as well as Eider and SAIS (both of which use evidence). Its training time of 81.80 minutes is also significantly shorter than that of KD-DocRE and TTM-RE, both of which do not use evidence.

7 Conclusion

We propose a novel MD-RE framework that incorporates three discriminators with different decision criteria. By leveraging the unique characteristics of each discriminator, we integrate their outputs using a weighted fusion method. This design alleviates the issue of introducing noise unrelated to the relation from the entire document and does not rely on evidence sentences. In addition, we propose a novel multi-label classification loss, ATSL, which effectively mitigates the class imbalance. Extensive experiments show the effectiveness and general applicability of both MD-RE and ATSL.

Limitations

Although our proposed MD-RE framework and ATSL loss show promising results, there are still some limitations. *First*, while MD-RE achieves state-of-the-art performance with favorable efficiency as shown in Section 6.3, the use of multiple discriminators increases memory consumption and training time, which may potentially limit its applicability in resource-constrained environments. *In addition*, the weighted fusion method that combines three discriminators requires manual setting of the weight value α, which reduces its flexibility and applicability. This sensitivity to parameter selection may affect the method's generalization ability across different datasets. Future work may explore adaptive weighting strategies or parameter-free fusion mechanisms to enhance the scalability and generalization of the framework.

References

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) (NAACL-HLT)*, pages 4171–4186.

Chufan Gao, Xuan Wang, and Jimeng Sun. 2024. TTM-RE: memory-augmented document-level relation extraction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jia Guo, Stanley Kok, and Lidong Bing. 2023. Towards integration of discriminability and robustness for document-level relation extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 2598–2609.

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.

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.

Feng Jiang, Jianwei Niu, Shasha Mo, and Shengda Fan. 2022. Key mention pairs guided document-level relation extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1904–1914, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Bulou Liu, Zhenhao Zhu, Qingyao Ai, Yiqun Liu, and Yueyue Wu. 2024. Ledqa: A chinese legal case document-based question answering dataset. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 5385–5389.

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*.

Chonggang Lu, Richong Zhang, Kai Sun, Jaein Kim, Cunwang Zhang, and Yongyi Mao. 2023. Anaphor assisted document-level relation extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 15453–15464.

Youmi Ma, An Wang, and Naoaki Okazaki. 2023. Dreeam: Guiding attention with evidence for improving document-level relation extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 1971–1983.

Ishani Mondal, Yufang Hou, and Charles Jochim. 2021. End-to-end construction of nlp knowledge graph. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1885–1895.

A Paszke. 2019. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*.

Xingyu Peng, Chong Zhang, and Ke Xu. 2022. Document-level relation extraction via subgraph reasoning. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4331–4337.

Qi Sun, Kun Huang, Xiaocui Yang, Pengfei Hong, Kun Zhang, and Soujanya Poria. 2023. Uncertainty guided label denoising for document-level distant relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL)*, pages 15960–15973.

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 Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1672–1681.

Qingyu Tan, Lu Xu, Lidong Bing, and Hwee Tou Ng. 2023. Class-adaptive self-training for relation extraction with incompletely annotated training data. In

Findings of the Association for Computational Linguistics: ACL 2023, pages 8630–8643.

Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. 2022b. Revisiting docred-addressing the false negative problem in relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8472–8487.

Khai Phan Tran, Wen Hua, and Xue Li. 2025. Vaediff-docre: End-to-end data augmentation framework for document-level relation extraction. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7307–7320.

Jize Wang, Xinyi Le, Xiaodi Peng, and Cailian Chen. 2023. Adaptive hinge balance loss for document-level relation extraction. In *Findings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3872–3878.

Ye Wang, Xinxin Liu, Wenxin Hu, and Tao Zhang. 2022. A unified positive-unlabeled learning framework for document-level relation extraction with different levels of labeling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4123–4135.

Ye Wang, Huazheng Pan, Tao Zhang, Wen Wu, and Wenxin Hu. 2024. A positive-unlabeled metric learning framework for document-level relation extraction with incomplete labeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19197–19205.

Ying Wei and Qi Li. 2022. Sagdre: Sequence-aware graph-based document-level relation extraction with adaptive margin loss. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 2000–2008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP)*, pages 38–45.

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 (NAACL-HLT)*, pages 2395–2409.

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 ACL*, pages 257–268.

Xiaolong Xu, Chenbin Li, Haolong Xiang, Lianyong Qi, Xuyun Zhang, and Wanchun Dou. 2024. Attention based document-level relation extraction with none class ranking loss. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*.

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

Klim Zaporojets, Johannes Deleu, Chris Develder, and Thomas Demeester. 2021. DWIE: an entity-centric dataset for multi-task document-level information extraction. *Inf. Process. Manag.*, 58(4):102563.

Hansi Zeng, Chen Luo, Bowen Jin, Sheikh Muhammad Sarwar, Tianxin Wei, and Hamed Zamani. 2024. Scalable and effective generative information retrieval. In *Proceedings of the ACM Web Conference 2024*, pages 1441–1452.

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.

Liang Zhang, Zijun Min, Jinsong Su, Pei Yu, Ante Wang, and Yidong Chen. 2023. Exploring effective inter-encoder semantic interaction for document-level relation extraction. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 5278–5286.

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)*, pages 3999–4006.

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 (AAAI)*, volume 35, pages 14612–14620.

Yang Zhou and Wee Sun Lee. 2022. None class ranking loss for document-level relation extraction. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4538–4544.

A Datasets and Hyperparameters

A.1 Hyperparameter Settings

We summarize the hyperparameters used for training on three datasets in **Table 10**. To ensure robust

Dataset	Re-DocRED		DocRED	DWIE
	BERT	RoBERTa	RoBERTa	BERT
epoch	20	20	8	30
lr_encoder	4e-5	2e-5	1e-5	5e-5
lr_classifier	1e-4	1e-4	1e-4	1e-4
batch size	4	4	4	4
warmup_ratio	0.06	0.06	0.06	0.06
ρ	4	4	4	4
α	0.65	0.60	1.0	1.10
λ_{Recall}	3.0	3.0	4.5	3.0
λ_{Coarse}	1.5	1.5	4.0	1.5
λ_{Fine}	0.0	0.0	-	0.0
β_{Recall}	0.0	0.0	0.0	0.0
β_{Coarse}	0.0	0.0	0.0	0.0
β_{Fine}	0.0	0.0	-	0.0

Table 10: Hyperparameters.

Dataset	Split	#Docs.	#Entities.	#Rels.
DocRED	train	3,053	59,493	96
	dev†	500	9,684	96
	test†	500	9,779	96
DWIE	train	602	16,494	65
	dev	98	2,785	65
	test	99	2,623	65
Re-DocRED	train	3,053	59,359	96
	dev†	500	9,684	96
	test†	500	9,779	96

Table 11: Statistics of datasets.

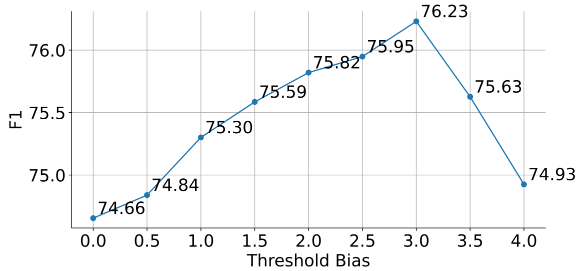


Figure 3: Performance under different λ values of the ATSL loss on the Re-DocRED dev set, using ATLOP and BERT$_{\text{base}}$.

evaluation, we conduct experiments using three different random seeds and report the averaged results. The batch size is set to 4, and learning rates for the encoder and classifier are tuned separately: a lower rate (1e-5 to 5e-5) for the encoder and a higher rate (1e-4) for the classifier. A warm-up ratio of 0.06 is applied across all settings to stabilize the early stages of training. The ATSL-related hyperparameters include $\boldsymbol{\lambda} = \{\lambda_{\text{Recall}}, \lambda_{\text{Coarse}}, \lambda_{\text{Fine}}\}$ and $\boldsymbol{\beta} = \{\beta_{\text{Recall}}, \beta_{\text{Coarse}}, \beta_{\text{Fine}}\}$, which control the weights for the recall, coarse, and fine discriminators, respectively. Since the hyperparameters λ and β in the ATSL loss serve the same purpose, we adjust only λ and set β to 0 to streamline the process. The number of training epochs is determined based on the dataset, ranging from 8 epochs on DocRED (Yao et al., 2019) to 30 epochs on DWIE (Zaporojets et al., 2021).

A.2 Datasets

The statistics of the three datasets are shown in **Table 11**, with detailed descriptions provided below.

DocRED (Yao et al., 2019) is a large-scale, human-annotated dataset constructed from Wikipedia, specifically designed for DocRE task. Although it contains 3,053 documents for training and 1,000 documents each for development and testing, it suffers from a substantial number of missing annotations. To address this issue, we adopt the Re-DocRED (Tan et al., 2022b) dataset for evaluation on the dev and test sets.

DWIE (Zaporojets et al., 2021) is a multi-task dataset focused on entity-centric tasks, with 602 documents in the train set, 98 in the dev set, and 99 in the test set.

Re-DocRED (Tan et al., 2022b) is a revised version of the DocRED (Yao et al., 2019) dataset, which has been reprocessed and manually validated to address the numerous false negative issues present in the original DocRED. Additionally, the validation and test sets of Re-DocRED are derived by splitting the dev set of DocRED, with each containing 500 documents. The number of documents in the train set of Re-DocRED remains the same as in DocRED, with a total of 3,053 documents.

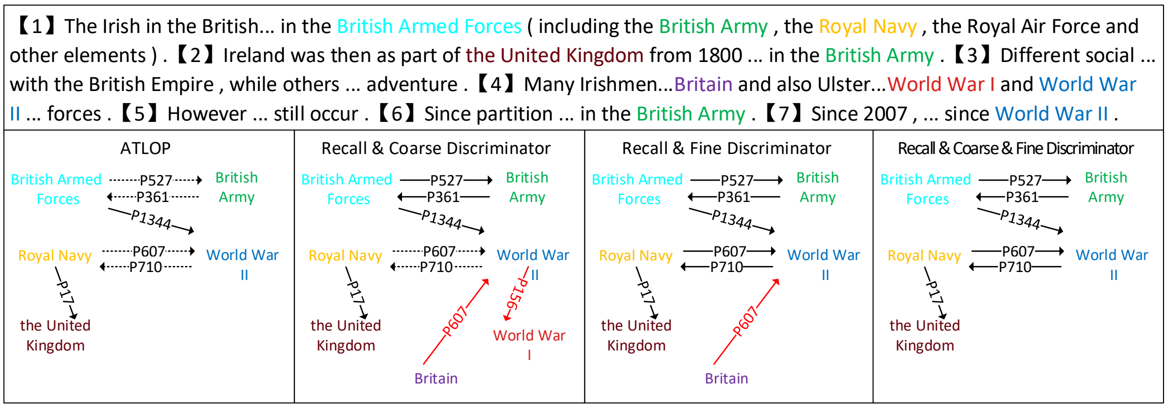
B Further Analysis

B.1 Effect of Hyperparameter λ in ATSL

To evaluate the impact of the hyperparameter λ in ATSL, we vary its value from 0.0 to 4.0 and systematically analyze its effect on performance. As shown in **Fig. 3**, the F1 consistently improves with increasing λ, achieving the highest value of 76.23 at $\lambda = 3.0$. A marginal decline in performance is observed beyond this point. These observations suggest that while ATSL is generally robust to the choice of λ, careful tuning around $[2.0, 3.5]$ is beneficial for maximizing performance.

B.2 Case Study

To better illustrate the effectiveness and interpretability of our MD-RE framework, we present a representative example in **Fig. 4**, where we provide a detailed comparison between the results of our MD-RE framework and those of the baseline model ATLOP. In the predictions of the ATLOP model, only two triples were correctly identified, with four triples missing. In the predictions from the recall discriminator combined with the coarse discriminator, compared to ATLOP, two additional triples were correctly predicted. However, the triples (**Royal Navy**, P607, **World War II**) and (**World War II**, P710, **Royal Navy**) were still missed. Moreover, the incorrect predictions (**Britain**, P607, **World War II**) and (**World War II**, P156, **World War I**) were introduced. This is because the recall discriminator focuses on improving recall, leading to some triples being over-recalled, while the coarse discriminator emphasizes initial filtering and reducing excessive predictions, which introduces incorrect predictions. In the predictions from the recall discriminator combined with the fine discriminator, the previously missed triples (**Royal Navy**, P607, **World War II**) and (**World War II**, P710, **Royal Navy**) were successfully recovered, and the incorrect prediction (**World War II**, P156, **World War I**) was corrected. This indicates that the fine discriminator is more effective at filtering out irrelevant or incorrect triples, thus improving prediction accuracy. Finally, by integrating the recall, coarse, and fine discriminators, the full MD-RE framework successfully predicted all the triples. Furthermore, the incorrect prediction (**Britain**, P607, **World War II**) from the combined recall and fine discriminator was eliminated. These results demonstrate that the integration of all three discriminators substantially enhances the overall performance of the framework.



【1】The Irish in the British... in the British Armed Forces (including the British Army , the Royal Navy , the Royal Air Force and other elements) . 【2】Ireland was then as part of the United Kingdom from 1800 ... in the British Army . 【3】Different social ... with the British Empire , while others ... adventure . 【4】Many Irishmen...Britain and also Ulster...World War I and World War II ... forces . 【5】However ... still occur . 【6】Since partition ... in the British Army . 【7】Since 2007 , ... since World War II .

Figure 4: Case study of the baseline model ATLOP and our proposed MD-RE framework on Re-DocRED dev set. Black solid lines indicate correct predictions, red solid lines represent incorrect predictions, and dashed lines denote missing predictions.