Uncertainty Guided Label Denoising for Document-level Distant Relation Extraction

Qi Sun[1,2], **Kun Huang**[1], **Xiaocui Yang**[2,3], **Pengfei Hong**[2], **Kun Zhang**[1*], **and Soujanya Poria**[2*]
[1]Nanjing University of Science and Technology
[2]Singapore University of Technology and Design, [3]Northeastern University
{319106003718, huangkun, zhangkun}@njust.edu.cn,
{pengfei_hong, sporia}@sutd.edu.sg,
yangxiaocui@stumail.neu.edu.cn

Abstract

Document-level relation extraction (DocRE) aims to infer complex semantic relations among entities in a document. Distant supervision (DS) is able to generate massive auto-labeled data, which can improve DocRE performance. Recent works leverage pseudo labels generated by the pre-denoising model to reduce noise in DS data. However, unreliable pseudo labels bring new noise, e.g., adding false pseudo labels and losing correct DS labels. Therefore, how to select effective pseudo labels to denoise DS data is still a challenge in document-level distant relation extraction. To tackle this issue, we introduce uncertainty estimation technology to determine whether pseudo labels can be trusted. In this work, we propose a **D**ocument-level distant **R**elation **E**xtraction framework with **U**ncertainty **G**uided label denoising, UGDRE. Specifically, we propose a novel instance-level uncertainty estimation method, which measures the reliability of the pseudo labels with overlapping relations. By further considering the long-tail problem, we design dynamic uncertainty thresholds for different types of relations to filter high-uncertainty pseudo labels. We conduct experiments on two public datasets. Our framework outperforms strong baselines by 1.91 F_1 and 2.28 Ign F_1 on the RE-DocRED dataset. [1]

1 Introduction

Document-level Relation Extraction (DocRE) aims to extract relations among entities in a document. In contrast to the conventional RE task that mainly focuses on sentence-level (Zhou et al., 2016; Guo et al., 2019; Tian et al., 2021), DocRE is more challenging due to the complex semantic scenarios, discourse structure of the document, and long-distant interactions between entities.

To understand complex inter-sentence entity relations, most existing methods employ transformer-based (Huang et al., 2021a; Zhou et al., 2021; Li et al., 2021) or graph-based models (Nan et al., 2020; Zeng et al., 2020, 2021) that aggregate effective entity representations. Although these methods achieve reasonable performance, they heavily rely on the large-scale human-annotated corpus, which is time-consuming and labor-intensive. Distant supervision mechanism (Mintz et al., 2009) provides large-scale distantly supervised (DS) data auto-labeled by existing relational triples from knowledge bases (Xie et al., 2021). Recent works observe that leveraging DS data to pretrain DocRE models can improve performance by a great margin (Xu et al., 2021; Yang Zhou, 2022; Wang et al., 2022).

Despite a great quantity of training data auto-labeled by distant supervision can enhance the performance of the model, noisy labels in DS data are non-negligible. Yao et al. (2019) show that there are 61.8% noisy inter-sentence instances in

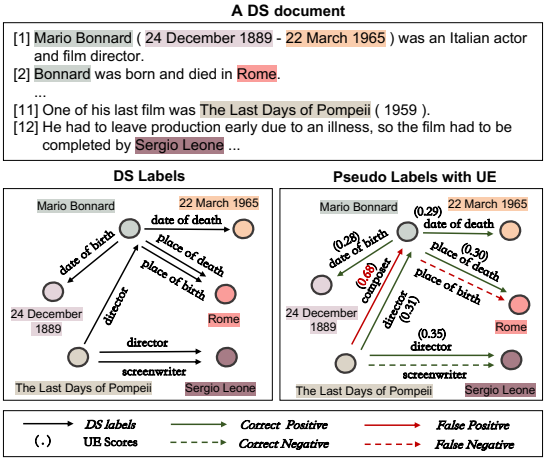


Figure 1: An example of the DS document. We present two types of noise caused by pseudo labels. One is adding new false labels as shown by the solid red line. Another is losing the correct DS label as shown by the red dotted line. We also show the proposed instance-level uncertainty estimation (UE) scores of pseudo labels. We present partly entities that are marked with different colors.

*Corresponding author
[1]https://github.com/QiSun123/UGDRE

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 15960–15973
July 9-14, 2023

their provided document-level distant relation extraction dataset. Current efforts (Xiao et al., 2020; Tan et al., 2022a) to alleviate the noise problem mainly employ a pre-denoising model. They train a DocRE model on human-annotated data first and then re-label DS data by the trained model.

However, the above methods still persist the risk of noise induction in the DS data due to false positive re-labeling. Besides, false negative pseudo labels also lead to the loss of effective labels in DS data. As shown in Figure 1, we obtain an extra false instance *(The Last Days of Pompeii, Mario Bonnard, composer)* and lose the correct DS instance *(Mario Bonnard , Rome, place of birth)*, when merely relying on pseudo labels. Thus, how to mitigate noise caused by pseudo labels and take full advantage of DS data is still a challenge in document-level distant RE.

In this work, we propose a **D**ocument-level distant **R**elation **E**xtraction framework with **U**ncertainty **G**uided label denoising, UGDRE. We first train a pre-denoising DocRE model with both DS and human-annotated data to generate pseudo labels. Since false pseudo labels predicted by the pre-denoising model are inevitable, we introduce Uncertainty Estimation (UE) to determine whether model predictions can be trusted or not. As shown in Figure 1, we can remove the false positive pseudo instance *(The Last Days of Pompeii, Mario Bonnard, composer)* according to its high uncertainty score. In this way, we can abstain from unreliable decisions of the pre-denoising model, which can mitigate the risk of false pseudo labels. Considering there might be multiple relations between an entity pair, we propose an instance-level UE method to capture uncertainty scores for overlapping relations. Moreover, we design a re-labeling strategy with dynamic class uncertainty thresholds by taking the DocRE long-tail problem into account to obtain high-quality DS data. With the proposed uncertainty guided label denoising mechanism, we design a multi-phase training strategy to further boost the performance of our final DocRE model.

The main contributions of our work are summarized as follows:

- We propose a document-level relation distant extraction framework with uncertainty guided label denoising, which greatly improves the label quality of DS data.
- We propose a novel instance-level uncertainty estimation method for overlapping relations to measure the reliability of instance-level pseudo labels.
- We design an iterative re-label strategy with dynamic class uncertainty thresholds for the problem of long-tail in DocRE to filter high uncertainty pseudo labels.
- The proposed framework achieves significant performance improvements over existing competitive baselines on two public datasets. Extensive experiments illustrate that the performance of baselines trained on our denoised DS (DDS) data is obviously improved.

2 Related Work

Sentence-level Relation Extraction. Conventional works on RE mainly focus on sentence-level supervised relation extraction (Zhou et al., 2016; Guo et al., 2019; Sun et al., 2022). Although these models achieve great success in RE, they primarily rely on the large-scale human-annotated corpus that needs time-consuming labels. Therefore, early works prefer to use extra data that are auto-labeled by distant supervision (DS) (Zeng et al., 2015; Huang et al., 2021b; Peng et al., 2022; Qin et al., 2018). However, the noisy labels caused by distant supervision will influence the performance of these models. Thus, various works are proposed to select effective instances, separate noisy data, and enhance the robustness of models. Most of them tend to perform attention mechanism(Li et al., 2020; Yuan et al., 2019; Han et al., 2018), negative training (Ma et al., 2021), reinforcement learning(Feng et al., 2018), and soft-label strategies (Liu et al., 2017). However, the above DS methods mainly focus on sentence-level RE, which can not be transferred to DocRE directly.

Document-level Relation Extraction. DocRE aims to extract relations between each entity pair expressed by multiple mentions across the sentences in a document. Different from the conventional sentence-level RE, DocRE needs the ability to reason relations in a more complex semantic scene. Existing approaches employ transformer-based models to extract contextual information for aggregating entity representations (Yao et al., 2019; Huang et al., 2021a; Zhou et al., 2021; Li et al., 2021). To further capture non-local syntactic and semantic structural information, some works construct document-level graphs and aggregate graph representations by Graph Neural Net-

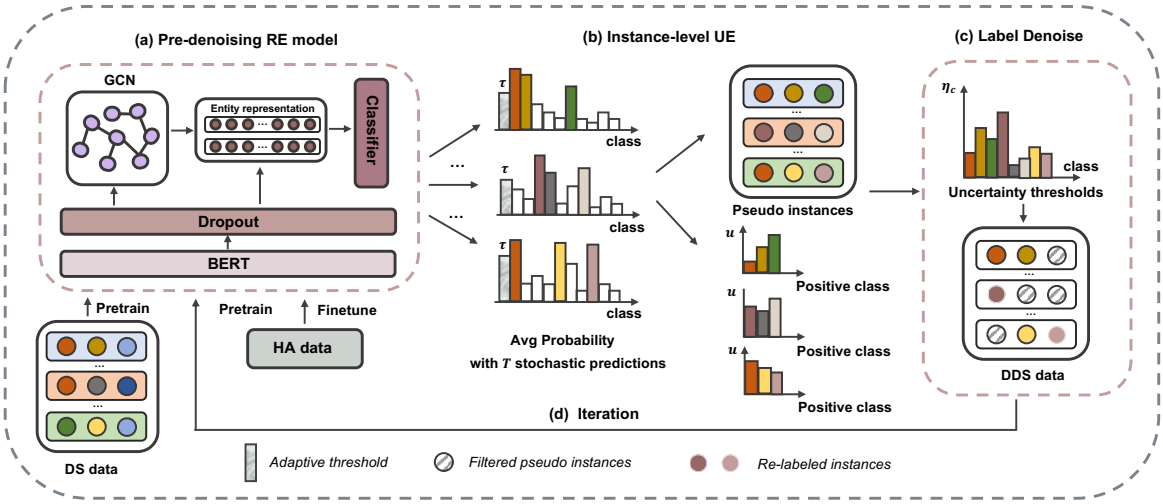


Figure 2: The overview of our UGDRE framework. It contains four key parts as follows: (a) Pre-denoising DocRE model; (b) instance-level UE of pseudo instances generated by Pre-denoising RE model; (c) Label denoising strategy to re-label with pseudo instances that contain low uncertainty scores; (d) Iterative training strategy for further re-label and improve the performance.

works (GNN) (Sahu et al., 2019; Wang et al., 2020; Eberts and Ulges, 2021; Christopoulou et al., 2019; Nan et al., 2020; Zeng et al., 2020; Wang Xu and Zhao, 2021; Zeng et al., 2021; Sun et al., 2023). Recent works observe that utilizing large-scale auto-labeled data generated by distant supervision (Mintz et al., 2009) to pretrain the DocRE model can attain great performance improvements (Xu et al., 2021; Yang Zhou, 2022; Wang et al., 2022; Hogan et al., 2022). Most of them directly utilize the DS data and ignore the accuracy of DS labels. To obtain high-quality DS data, several methods introduce re-label strategies based on the pre-denoising RE model trained on human-annotated data (Xiao et al., 2020; Tan et al., 2022a). However, these methods ignore the noise caused by pseudo labels. In this work, we introduce uncertainty estimation to determine the reliability of pseudo labels, which can reduce the noisy pseudo labels to further improve the quality of DS data.

3 Methodology

In this section, we introduce our proposed framework in detail. As shown in Figure 2, our UGDRE contains four key components: (1) Training a document-level pre-denoising model by the original DS and human-annotated training data; (2) Estimating uncertainty scores of pseudo labels generated by the pre-denoising model; (3) Denoising the DS data by pseudo labels and uncertainty scores; (4) Leveraging a multi-phase training strategy to iteratively train the DocRE model by denoised DS (DDS) data and human-annotated data.

3.1 Problem Formulation

Given a document $D = \{s_i\}_{i=1}^{t}$, which is composed of t sentences. Each document contains a set of entities $E = \{e_i\}_{i=1}^{q}$, where q is the number of entities. An entity might be mentioned multiple times in a document, formulated as $e_i = \{m_j^i\}_{j=1}^{p_i}$, where p_i is the number of times the entity e_i is mentioned. The aim of the document-level relation extraction is to predict relation types between entities, formulated as $\{(e_i, e_j, r_k)|e_i, e_j \in E, r_k \in R\}$, where R is the set of pre-defined relation types. In addition, there can be more than one relation type between a specific entity pair in a document. Thus, the DocRE task can be regarded as a multi-label classification task. In the document-level distant relation extraction setting, we have a clean human-annotated dataset and a distantly supervised dataset, while the quantity of DS data is significantly larger than the human-annotated data.

3.2 Document-level Pre-denoising Model

In order to alleviate the noisy label problem in the DS data, we construct a pre-denoising DocRE model to generate pseudo labels. As shown in Figure 2(a), we adopt BERT (Devlin et al., 2019) to capture the contextual representation $\{z_i\}_{i=1}^{n}$, where n is the number of tokens in a document. We also adopt a dropout layer to enhance the generalization ability of our DocRE model.

To capture non-local dependency among entities,

we construct a graph for each document. Specifically, we take all tokens in a document as nodes and connect them using the task-specific rules: (1) To capture the semantic information of mentions, tokens belonging to the same mention are connected. (2) To capture interactions between mentions, tokens of mentions belonging to the same entities are connected. (3) To capture the interactions of entities, tokens of entities that co-occur in a single sentence are connected.

We construct the adjacency matrix according to the document graph and apply Graph Convolutional Networks (GCN) to capture graph representations $\{g_i\}_{i=1}^{n}$, which is formulated as follows:

$$g_i = \rho\left(\sum_{j=1}^{n} A_{ij} W z_j + b\right), \tag{1}$$

where $W \in \mathbb{R}^{d\times d}$ and $b \in \mathbb{R}^{d}$ are trainable parameters. z_j is the contextual representation of j-th token, which is introduced above. A_{ij} is the adjacency matrix of the document graph. ρ is the activation function. To obtain the global representations $\{h_i\}_{i=1}^{n}$, we concatenate the contextual representations $\{z_i\}_{i=1}^{n}$ and graph representations $\{g_i\}_{i=1}^{n}$ as follows:

$$h_i = [z_i, g_i]. \tag{2}$$

Following Zhou et al. (2021), we also apply log-sumexp pooling (Jia et al., 2019) to obtain entity representations $\{e_i\}_{i=1}^{q}$. Finally, group bilinear (Van Amersfoort et al., 2020) is utilized to obtain the probabilities $\{p_{ij}^{c}\}_{c=1}^{N_c}$ of each class c for the entity pair (e_i, e_j) to predict relation types.

3.3 Instance-level Uncertainty Estimation

Uncertainty Estimation (UE) is a vital technology for misclassification detection (Vazhentsev et al., 2022), out-of-distribution instances detection (Van Amersfoort et al., 2020), and active learning (Burkhardt et al., 2018). In order to model the uncertainty in pre-denoising DocRE model, we introduce the Monte Carlo (MC) dropout (Gal and Ghahramani, 2016) technology into the DocRE task. As a popular UE technology, MC dropout is formally equivalent to approximate Bayesian inference in deep Gaussian processes (Gal and Ghahramani, 2016). This method requires multiple stochastic forward-pass predictions with activated dropout to capture the model uncertainty.

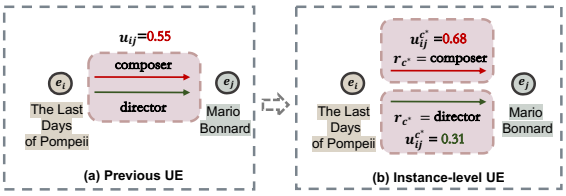


Figure 3: An example of our instance-level UE score for each predicted positive relation between an entity pair. We present two overlapping relations predicted by the pre-denoising model between an entity pair *(The Last Days of Pompeii, Mario Bonnard)*.

Previous works based on MC dropout (Gal et al., 2017; Vazhentsev et al., 2022) calculate the uncertainty score of the model prediction as follows:

$$u_s = \frac{1}{N_c}\sum_{c=1}^{N_c}\left(\frac{1}{N_t}\sum_{t=1}^{N_t}(p_t^c - \overline{p^c})^2\right), \tag{3}$$

where N_c is the number of the class number. N_t is the number of stochastic forward passes. p_t^c is the probability of the c-th class at t-th stochastic forward passes. $\overline{p^c} = \frac{1}{N_t}\sum_{t=1}^{N_t} p_t^c$ is the average probability of the c-th class.

The above uncertainty estimation method provides one uncertainty score for each prediction. However, there exist multiple relations for one entity pair, which can be called overlapping relations. Intuitively, different overlapping relations should have their own uncertainty scores. As shown in Figure 3(a), there are two different types of relations between an entity pair *(The Last Days of Pompeii, Mario Bonnard)*. It is hard to separate the false positive pseudo label *composer* and correct positive pseudo label *director* by previous UE methods (Gal et al., 2017; Vazhentsev et al., 2022).

To solve this issue, we modify the estimation process to obtain the instance-level UE score for each positive pseudo label between an entity pair, which can be seen in Figure 3(b). Inspired by ATLOP (Zhou et al., 2021) that introduces a threshold class $\tilde{c}$ to separate positive and negative relation classes. We calculate the adaptive threshold score τ_{ij} for entity pair (e_i, e_j) as follows:

$$\tau_{ij} = \frac{1}{N_t}\sum_{t=1}^{N_t} p_{ijt}^{\tilde{c}}, \tag{4}$$

where $p_{ijt}^{\tilde{c}}$ is the probability of the threshold class for entity pair (e_i, e_j) at t-th stochastic forward passes. N_t is the number of stochastic forward passes. Then, we regard classes of which average

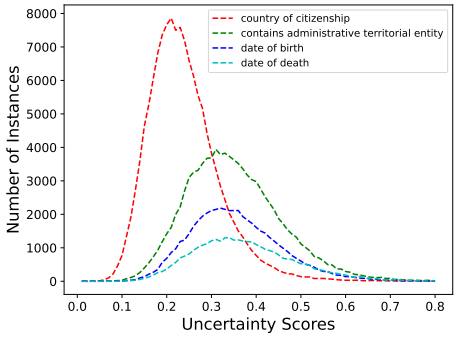


Figure 4: The distribution of UE scores of several relation types.

probability $\overline{p_{ij}^c} = \frac{1}{N_t}\sum_{t=1}^{N_t} p_{ijt}^c$ are higher than the threshold τ_{ij} as positive classes. If all the class probabilities are lower than the probability of the threshold class, we regard "NA (no relationship)" as the relationship type for the entity pair. Then, we calculate the uncertainty score of each positive class for entity pair (e_i, e_j) as follows:

$$u_{ij}^{c^*} = \frac{1}{N_t}\sum_{t=1}^{N_t}(p_{ijt}^{c^*} - \overline{p_{ij}^{c^*}})^2, c^* \in \{\overline{p_{ij}^c} > \tau_{ij}\}, \tag{5}$$

where $p_{ijt}^{c^*}$ is the probability of the positive class c^* at t-th stochastic forward passes. $\overline{p_{ij}^{c^*}} = \frac{1}{N_t}\sum_{t=1}^{N_t} p_{ij}^{c^*}$ is the average probability of the positive class c^*.

In this way, we can obtain each positive pseudo label with its uncertainty score between an entity pair, which is shown in Figure 2(b). Each pseudo instance is formulated as $(e_i, e_j, r_{c^*}, u_{ij}^{c^*})$.

3.4 Uncertainty Guided Label Denoising

After obtaining instance-level pseudo labels and their corresponding uncertainty scores, we re-label DS data by the proposed uncertainty-guided denoising strategy (Figure 2(c)). We observe that the distribution of uncertainty scores for each relation class is obviously different, which is shown in Figure 4. Moreover, it can be observed that frequent classes usually contain lower average uncertainty than long-tail classes. Therefore, considering the long-tail problem in the DocRE task, we propose dynamic class uncertainty thresholds to filter pseudo labels with high uncertainty. For each class of relation, the corresponding uncertainty threshold is calculated as follows:

$$\eta_{c^*} = \overline{u^{c^*}} + \sqrt{\frac{1}{N_{c^*}^{\eta} - 1}\sum_{l=1}^{N_{c^*}^{\eta}}(u_l^{c^*} - \overline{u^{c^*}})^2}, \tag{6}$$

where $u_l^{c^*}$ is the uncertainty score of the l-th pseudo instance of class c^*. $\overline{u^{c^*}} = \frac{1}{N_{c^*}^{\eta}}\sum_{l=1}^{N_{c^*}^{\eta}} u_i^{c^*}$ is the average of uncertainty scores for class c^* in all pseudo instances. $N_{c^*}^{\eta}$ is the quantity of pseudo instances that belong to class c^*.

In our re-label strategy (Figure 2(c)), for each entity pair (e_i, e_j), we replace the original DS label with the pseudo label r_{c^*} that contain the lower uncertainty score $u_{ij}^{c^*}$ than its class uncertainty thresholds η_{c^*}. In this way, we are able to reduce false positive pseudo labels and keep correct DS labels with high-quality pseudo labels. Besides, to reduce false negative pseudo labels, we keep the original DS positive labels where do not exist positive pseudo labels between an entity pair.

Algorithm 1 Multi-phase Training Strategy

Define: Human-annotated training and test data: HA and DT, DS data: DS, Iteration: K, DocRE Model: M, Pseudo labels with UE: PU, Denoised DS data: DDS, Relations: TR.

Require: DS, HA, K, M.

Ensure: $K > 0$.

$M_{pretrain} \leftarrow Train\,(M, DS)$
$M_{finetune} \leftarrow Train\,(M_{pretrain}, HA)$
for $i = 1$; $i <= K$; $i++$ **do**
1. $PU \leftarrow Predict\,(M_{finetune}, DS)$
2. $DDS \leftarrow Denoise\,(DS, PU)$
3. $M \leftarrow Reinitialized\,(M_{finetune})$
4. $M_{pretrain} \leftarrow Train\,(M, DDS)$
5. $M_{finetune} \leftarrow Train\,(M_{pretrain}, HA)$
6. $DS \leftarrow DDS$

end for
$TR \leftarrow Predict\,(M_{finetune}, DT)$
return TR

3.5 Multi-phase Training Strategy

In order to take full advantage of the DS data for further boosting the performance of the DocRE model, we design a multi-phase training strategy to iteratively re-label the DS data, which is shown in Algorithm 1. We introduce the overall process as follows. (1) We train the initial pre-denoising RE model with the original DS data and human-annotated data. (2) We leverage the pre-denoising

Model	DocRED				Re-DocRED			
	Dev		Test		Dev		Test	
	F_1	Ign F_1	F_1	Ign F_1	F_1	Ign F_1	F_1	Ign F_1
ATLOP (Zhou et al., 2021)	63.42	61.57	63.48	61.43	74.34	73.62	74.23	73.53
DocuNet (Zhang et al., 2021)	64.35	62.66	64.00	61.93	76.22	75.50	75.35	74.61
NCRL (Yang Zhou, 2022)	63.87	61.65	63.45	60.98	75.85	74.91	75.90	75.00
SSR-PU (Wang et al., 2022)	63.00	60.43	62.66	59.80	76.83	75.57	76.23	74.96
KD-NA* (Tan et al., 2022a)	64.17	62.18	64.12	61.77	76.14	75.25	76.00	75.12
KD-DocRE* (Tan et al., 2022a)	64.81	62.62	64.76	62.56	76.47	75.30	76.14	74.97
UGDRE (Ours)	**65.71**	**63.62**	**65.58**	**63.26**	**78.28**	**77.32**	**78.14**	**77.24**

Table 1: Experimental results on public datasets: DocRED and Re-DocRED dataset. The baselines are trained with both DS data and human-annotated training data. The test results of DocRED are obtained from the leaderboard submission. Results of DocRED with * are from Tan et al. (2022a)

RE model to generate pseudo instances with uncertainty scores on DS data. (3) We perform a re-label strategy to obtain denoised distantly supervised (DDS) data. (4) We re-initialize and train the pre-denoising DocRE with DDS data and human-annotated data to boost performance. We iterate the above (2), (3), and (4) phases until we obtain the best DocRE model.

4 Experiments

4.1 Dataset and Settings

Dataset. DocRED (Yao et al., 2019) is a popular DocRE dataset with 96 pre-defined relation types, which is constructed from Wikipedia and Wikidata. It provides a distant-supervised dataset with 101873 documents and a large-scale human-annotated dataset with 5053 documents. Re-DocRED is a high-quality revised version of human-annotated documents of DocRED, which is provided by Tan et al. (2022b) recently. Re-DocRED contains 3053, 500, and 500 documents for training, development, and testing. See Appendix A.1 for more details.

Settings. Following previous works (Zhou et al., 2021; Tan et al., 2022a), we adopt BERT$_{base}$ (Devlin et al., 2019) as the context encoder. We use AdamW (Loshchilov and Hutter, 2019) as the optimizer. We set the learning rate to 3e-5. We apply warmup for the initial 6% steps. We set the batch size to 8 for both the training and test process. The rate of the dropout is 0.25. All hyper-parameters are tuned on the development set. The experiments are conducted on a single NVIDIA RTX A6000-48G GPU. DocRED and RE-DocRED both contain 3053 human-annotated training documents for fine-tuning and 101873 distantly supervised training documents for pretraining. Thus, for each dataset, our framework takes about 55 hours and consumes about 23G GPU memory for training. Following Yao et al. (2019), we use F_1 and $IgnF_1$ as the evaluation metrics. The $IgnF_1$ represents F_1 scores, which excludes the relational facts shared by the human-annotated training set.

4.2 Compared Methods

We compare our UGDRE with several strong baselines that are trained on both DS and human-annotated data. **ATLOP** (Zhou et al., 2021) utilizes an adaptive thresholding loss to solve the overlapping relation problem, and adopts a localized context pooling to aggregate entity representations. **DocuNet** (Zhang et al., 2021) regards the DocRE task as a semantic segmentation task that provides a new view to extract document-level relations. **NCRL** (Yang Zhou, 2022) uses a multi-label loss that prefers large label margins between the "NA" relation class and each predefined class. **SSR-PU** (Wang et al., 2022) is a positive-unlabeled learning framework, which adapts DocRE with incomplete labeling. **KD-DocRE** (Tan et al., 2022a) attempts to overcome the differences between human-annotated and DS data by knowledge distillation. They also provide the **KD-NA** (Tan et al., 2022a), which is pretrained by DS data first and then fine-tuned by human-annotated data.

4.3 Experimental Results

We compare our UGDRE framework with the above baselines, which are also based on BERT$_{base}$ (Devlin et al., 2019) and trained on both DS data and human-annotated data. As shown in Table 1, our framework UGDRE outperforms the previous baselines on both DocRED and RE-DocRED datasets. Specifically, our UGDRE achieves 65.71

Model	Origin				After Denoising				Improvement	
	Dev		Test		Dev		Test			
	F_1	Ign F_1	F_1	Ign F_1	F_1	Ign F_1	F_1	Ign F_1	ΔF_1	ΔIgn F_1
+DocRED										
ATLOP	54.38	51.62	53.10	50.01	59.00	56.35	58.34	55.35	+5.24	+5.34
DocuNet	53.79	50.91	52.96	49.69	59.03	56.17	58.05	54.85	+5.09	+5.16
NCRL	54.53	51.66	53.26	50.03	59.39	56.71	58.50	55.46	+5.24	+5.43
KD-NA	54.02	50.94	54.10	50.65	58.39	55.31	58.20	54.79	+4.10	+4.14
UGDRE (Ours)	54.74	51.91	54.47	51.27	59.75	56.84	58.92	55.67	+4.45	+4.40
+RE-DocRED										
ATLOP	43.48	42.69	42.59	41.77	75.99	74.86	75.29	74.16	+32.70	+32.39
DocuNet	44.22	43.38	43.89	43.02	76.38	75.18	75.64	74.44	+31.75	+31.42
NCRL	44.71	43.87	44.09	43.23	76.39	75.19	75.69	74.50	+31.60	+31.27
KD-NA	45.55	44.58	45.38	44.41	76.11	74.78	75.37	74.00	+29.99	+29.59
UGDRE (Ours)	45.56	44.71	44.76	43.94	76.47	75.24	75.57	74.32	+30.81	+30.38

Table 2: Experimental results of DocRE baselines trained on original DS data and our denoised DS (DDS) data.

F_1 and 78.14 F_1 on the test set of DocRED and RE-DocRED datasets, respectively. Our UGDRE outperforms the KD-DocRE (Tan et al., 2022a) that leverages knowledge distillation to denoise by 2.00 F_1 and 0.82 F_1 on the test set of RE-DocRE and DocRED datasets. Moreover, our UGDRE significantly outperforms the latest strong baseline SSR-PU (Wang et al., 2022) by 1.91 F_1 and 2.28 Ign F_1 on the Re-DocRED dataset. This suggests the effectiveness of our uncertainty-guided denoise strategy.

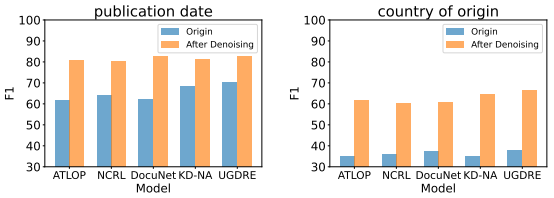
Besides, we observe that improvements on the RE-DocRED dataset are obviously higher than DocRED dataset, which can be caused by the following: 1) The RE-DocRED dataset is a revised version of the DocRED dataset by adding more positive instances. It alleviates the imbalance problem of positive and negative instances. 2) The pre-denoising model trained on RE-DocRED achieves a higher ability to discover relations, which will enhance the denoise process.

5 Analysis and Discussion

In this section, we conduct extensive experiments to further analyze the effectiveness of our proposed denoising strategy and instance-level UE method. We also conduct the ablation study to discuss the contribution of each component of the framework.

5.1 Effectiveness of the Denoising Strategy

In order to intuitively demonstrate the effectiveness of our uncertainty-guided denoising strategy. We present experimental results of several DocRE baselines only trained on original DS data and our denoised DS (DDS) data. As shown in Table 2, we



(a) Frequent relation. (b) Long-tail relation.

Figure 5: Experiment results for the frequent relation type *publication date* and long-tail relation type *country of origin* on the RE-DocRED dataset.

can observe that all baselines trained on our DDS data obtain significant performance improvements on both DocRED and RE-DocRED. In contrast to the original DS data, the performance of baselines trained on our DDS data increases more than 4 F_1 and 29 F_1 on the test set of the DocRED and RE-DocRED datasets. This further demonstrates the effectiveness of our uncertainty guided denoising strategy.

We observe that when training on original DS data, the performance of baselines on the RE-DocRED dataset is obviously lower than the DocRED dataset. This is because there are more positive instances in the RE-DocRED dataset than in the DocRED dataset, which makes the noise problem more obvious. Thus, the performance improvement of models trained on our denoised data will also be more obvious. The performance improvements of baselines that are fine-tuned on human-annotated data can be seen in Appendix A.2.

In addition, we also evaluate the performance of the above models on each relation type. As shown in Figure 5, the performance improvement

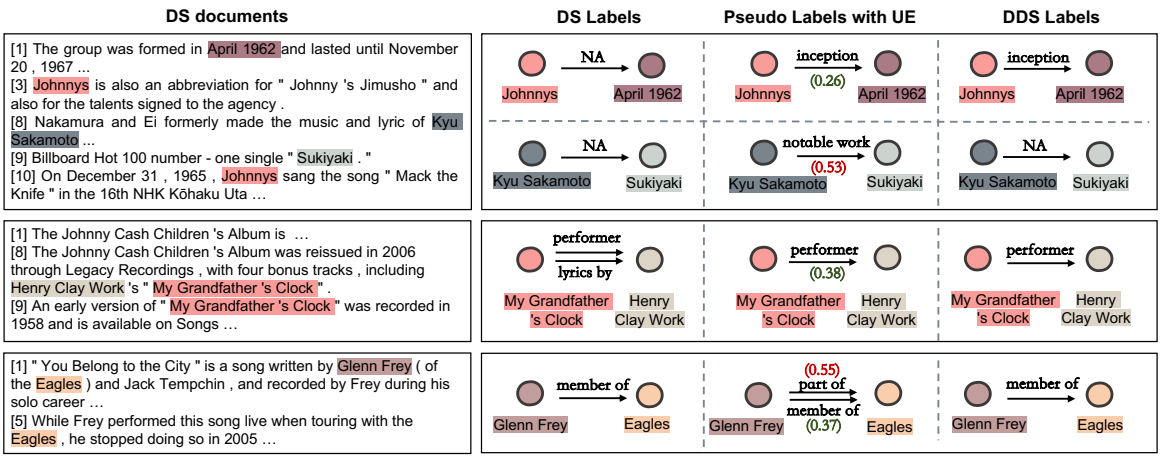


Figure 6: Case Study. We present several samples from the DS data, which contain original DS labels, pseudo labels with our proposed instance-level UE scores, and denoised distantly supervised (DDS) labels. We mark the UE scores that exceed their class uncertainty thresholds with red color.

Model	Dev		Test	
	F_1	Ign F_1	F_1	Ign F_1
SR	73.41	72.43	72.40	71.39
Entropy	74.42	73.27	73.49	72.31
PV Dropout	75.54	74.41	74.87	73.71
UGDRE	**76.47**	**75.24**	**75.57**	**74.32**

Table 3: Comparison of uncertainty estimation methods on Re-DocRED dataset.

Model	Dev		Test	
	F_1	Ign F_1	F_1	Ign F_1
UGDRE	78.28	77.32	78.14	77.24
w/o Pretrain	74.25	73.36	74.10	73.21
w/o DDS	76.91	76.00	76.16	75.23
w/o UE	77.66	76.80	76.84	75.99

Table 4: Ablation study on the RE-DocRED dataset.

of the long-tail relation type *country of origin* is obviously higher than the frequent relation type *publication date* after training on our denoised data. This indicates the effectiveness of our dynamic class uncertainty thresholds designed for the long-tail problem in DocRE task.

5.2 Effectiveness of Instance-level Uncertainty Estimation

We compare our proposed instance-level UE with existing popular UE technologies (Vazhentsev et al., 2022) as follows: 1) Softmax Response (SR); 2) Entropy; 3) Probability Variance (PV) with MC dropout. The performance of the DocRE model trained on denoised DS data that performed different UE technology is shown in Table 3. It can be observed that the DocRE model based on our instance-level UE outperforms SR, entropy, and PV dropout based methods on the test set of the RE-DocRED dataset. This is because our instance-level UE provides specific UE scores for different overlapping relations, which enables our downstream uncertainty guided relabel strategy to separate the false pseudo label from the overlapping relations. The experimental results also demonstrate the effectiveness of our proposed instance-level UE method.

5.3 Case Study

We present several samples of DS data that are denoised by our UGDRE framework in Figure 6. It can be observed that our framework denoises the DS data by 1) adding the extra correct positive instance, such as *(Johnnys, April 1962, inception)*; 2) Removing false DS instances, such as *(My Grandfather's Clock, Henry Clay Work, lyrics by)*. Moreover, we also present pseudo labels with our instance-level UE scores to show the process of re-relabel strategy. As shown in the second and fourth samples of Figure 6, our framework is able to reduce the false pseudo labels by their high uncertainty scores, such as *(Kyu Sakamoto, Sukiyaki, notable work)* and *(Glenn Frey, Eagles, member of)*.

5.4 Ablation Study

To analyze the effectiveness of each component in our UGDRE framework, we conduct the ablation study by removing different components. As shown in Table 3, the performance decreases as

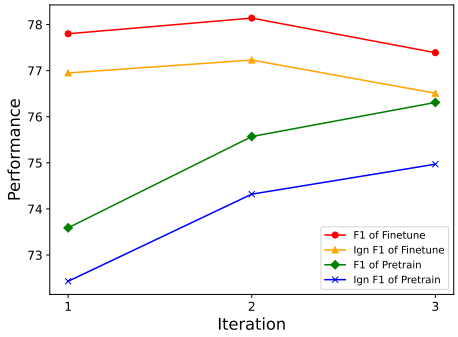


Figure 7: Performance of the model under different iterations on the test set of RE-DocRED.

removing each component, which demonstrates the effectiveness of our framework. When we remove the pretrain mechanism with DS data, the DocRE model trained by merely human-annotated data achieves 74.10 F_1 and 73.21 Ign F_1 on the test set of RE-DocRED dataset. This drop demonstrates that leveraging DS data can enhance the performance of the DocRE model. Removing the denoised distantly supervised (DDS) data leads to a 1.98 and 2.01 drop in terms of F_1 and Ign F_1 on the test set of RE-DocRED dataset. This indicates the significant effect of our uncertainty guided label denoising strategy. Our UGDRE framework is also effective on sentence-level distant RE, which can be seen in Appendix A.3.

As shown in Figure 7, we also present the performance of each iteration of the model that is pretrained on DDS and fine-tuned on human-annotated data. We can observe that the final model performance achieves the best by the second iteration of Algorithm 1, which proves the effectiveness of our multi-phase training strategy. Moreover, the removal of our instance-level uncertainty estimation also causes an obvious drop, which illustrates the importance of estimating uncertainty in our framework.

6 Conclusion

In this paper, we propose a **D**ocument-level distant **R**elation **E**xtraction framework with **U**ncertainty **G**uided label denoising, UGDRE. Specifically, we propose instance-level uncertainty estimation to measure the reliability of pseudo labels. Considering the long-tail problem, we design dynamic class uncertainty thresholds to filter high-uncertainty pseudo labels. Our proposed uncertainty guided denoising strategy can improve the quality of DS data. Experimental results demonstrate that our UGDRE outperforms competitive baselines. Moreover, extensive experiments verify the effectiveness of our label denoising. There are various challenges in DocRE worth exploring, one is to research the low-resource relation extraction.

Limitations

In this section, we discuss the limitations of our proposed framework. Our UGDRE can reduce the false positive pseudo label by estimating the uncertainty of the model prediction. However, it is difficult to reduce the false negative pseudo labels by uncertainty estimation. Our framework also relies on human-annotated data to train the pre-denoising model, which causes the sensitivity of our framework to the quality of human-annotated data. Thus, the improvements of models that continue training on the DocRED dataset are not as well as on the RE-DocRED dataset. Moreover, iterative training introduces additional computing overhead, which makes the training process time-consuming.

Acknowledgements

Thanks to all co-authors for their hard work. The work is supported by the Chinese Scholarship Council, the National Program on Key Research Project of China (Project no. 2020XXXXXX6404), the Ministry of Education, Singapore, under its AcRF Tier-2 grant (Project no. T2MOE2008, and Grantor reference no. MOE-T2EP20220-0017), and A*STAR under its RIE 2020 AME programmatic grant (project reference no. RGAST2003). Any opinions, findings, and conclusions, or recommendations expressed in this material are those of the author(s) and do not reflect the views of the Ministry of Education, Singapore.

References

Sophie Burkhardt, Julia Siekiera, and Stefan Kramer. 2018. Semisupervised bayesian active learning for text classification. In *Bayesian Deep Learning Workshop at NeurIPS*.

Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. In *Proceedings of EMNLP*, pages 4927–4938.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186.

Markus Eberts and Adrian Ulges. 2021. An end-to-end model for entity-level relation extraction using multi-instance learning. In *Proceedings of EACL*, pages 3650–3660.

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

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of ICML*, pages 1050–1059.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep bayesian active learning with image data. In *Proceedings of ICML*, pages 1183–1192.

Zhijiang Guo, Yan Zhang, and Wei Lu. 2019. Attention guided graph convolutional networks for relation extraction. In *Proceedings of ACL*, pages 241–251.

Xu Han, Pengfei Yu, Zhiyuan Liu, Maosong Sun, and Peng Li. 2018. Hierarchical relation extraction with coarse-to-fine grained attention. In *Proceedings of EMNLP*, pages 2236–2245.

William Hogan, Jiacheng Li, and Jingbo Shang. 2022. Fine-grained contrastive learning for relation extraction. In *Proceedings of EMNLP*, pages 1083–1095.

Quzhe Huang, Shengqi Zhu, Yansong Feng, Yuan Ye, Yuxuan Lai, and Dongyan Zhao. 2021a. Three sentences are all you need: Local path enhanced document relation extraction. In *Proceedings of ACL*, pages 998–1004.

Wenti Huang, Yiyu Mao, Liu Yang, Zhan Yang, and Jun Long. 2021b. Local-to-global gcn with knowledge-aware representation for distantly supervised relation extraction. *Knowledge-Based Systems*, page 107565.

Robin Jia, Cliff Wong, and Hoifung Poon. 2019. Document-level n-ary relation extraction with multiscale representation learning. In *Proceedings of NAACL*, pages 3693–3704.

Jingye Li, Kang Xu, Fei Li, Hao Fei, Yafeng Ren, and Donghong Ji. 2021. Mrn: A locally and globally mention-based reasoning network for document-level relation extraction. In *Findings of ACL*, pages 1359–1370.

Yang Li, Guodong Long, Tao Shen, Tianyi Zhou, Lina Yao, Huan Huo, and Jing Jiang. 2020. Self-attention enhanced selective gate with entity-aware embedding for distantly supervised relation extraction. In *Proceedings of AAAI*, 05, pages 8269–8276.

Tianyu Liu, Kexiang Wang, Baobao Chang, and Zhifang Sui. 2017. A soft-label method for noise-tolerant distantly supervised relation extraction. In *Proceedings of EMNLP*, pages 1790–1795.

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

Ruotian Ma, Tao Gui, Linyang Li, Qi Zhang, Xuan-Jing Huang, and Yaqian Zhou. 2021. Sent: Sentence-level distant relation extraction via negative training. In *Proceedings of ACL*, pages 6201–6213.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of ACL*, pages 1003–1011.

Guoshun Nan, Zhijiang Guo, Ivan Sekulic, and Wei Lu. 2020. Reasoning with latent structure refinement for document-level relation extraction. In *Proceedings of ACL*, pages 1546–1557.

Tao Peng, Ridong Han, Hai Cui, Lin Yue, Jiayu Han, and Lu Liu. 2022. Distantly supervised relation extraction using global hierarchy embeddings and local probability constraints. *Knowledge-Based Systems*, page 107637.

Pengda Qin, Weiran Xu, and William Yang Wang. 2018. Dsgan: Generative adversarial training for distant supervision relation extraction. In *Proceedings of ACL*, pages 496–505.

Sunil Kumar Sahu, Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Inter-sentence relation extraction with document-level graph convolutional neural network. In *Proceedings of ACL*, pages 4309–4316.

Qi Sun, Kun Zhang, Kun Huang, Tiancheng Xu, Xun Li, and Yaodi Liu. 2023. Document-level relation extraction with two-stage dynamic graph attention networks. *Knowledge-Based Systems*, 267:110428.

Qi Sun, Kun Zhang, Laishui Lv, Xun Li, Kun Huang, and Ting Zhang. 2022. Joint extraction of entities and overlapping relations by improved graph convolutional networks. *Applied Intelligence*, 52(5):5212–5224.

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 ACL*, pages 1672–1681.

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 EMNLP*.

Yuanhe Tian, Guimin Chen, Yan Song, and Xiang Wan. 2021. Dependency-driven relation extraction with attentive graph convolutional networks. In *Proceedings of ACL*, pages 4458–4471.

Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. 2020. Uncertainty estimation using a single deep deterministic neural network. In *Proceedings of ICML*, pages 9690–9700.

Artem Vazhentsev, Gleb Kuzmin, Artem Shelmanov, Akim Tsvigun, Evgenii Tsymbalov, Kirill Fedyanin, Maxim Panov, Alexander Panchenko, Gleb Gusev,

Mikhail Burtsev, Manvel Avetisian, and Leonid Zhukov. 2022. Uncertainty estimation of transformer predictions for misclassification detection. In *Proceedings of ACL*, pages 8237–8252.

Difeng Wang, Wei Hu, Ermei Cao, and Weijian Sun. 2020. Global-to-local neural networks for document-level relation extraction. In *Proceedings of EMNLP*, pages 3711–3721.

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 EMNLP*, page 4123–4135.

Kehai Chen Wang Xu and Tiejun Zhao. 2021. Discriminative reasoning for document-level relation extraction. In *Findings of ACL*, pages 1653–1663.

Chaojun Xiao, Yuan Yao, Ruobing Xie, Xu Han, Zhiyuan Liu, Maosong Sun, Fen Lin, and Leyu Lin. 2020. Denoising relation extraction from document-level distant supervision. In *Proceedings of EMNLP*, pages 3683–3688.

Chenhao Xie, Jiaqing Liang, Jingping Liu, Chengsong Huang, Wenhao Huang, and Yanghua Xiao. 2021. Revisiting the negative data of distantly supervised relation extraction. In *Proceedings of ACL*.

Benfeng Xu, Quan Wang, Yajuan Lyu, Yong Zhu, and Zhendong Mao. 2021. Entity structure within and throughout: Modeling mention dependencies for document-level relation extraction. In *Proceedings of AAAI*, 16, pages 14149–14157.

Wee Sun Lee Yang Zhou. 2022. None class ranking loss for document-level relation extraction. In *Proceedings of IJCAI*, pages 4538–4544.

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 ACL*, pages 764–777.

Yujin Yuan, Liyuan Liu, Siliang Tang, Zhongfei Zhang, Yueting Zhuang, Shiliang Pu, Fei Wu, and Xiang Ren. 2019. Cross-relation cross-bag attention for distantly-supervised relation extraction. In *Proceedings of AAAI*, pages 419–426.

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

Shuang Zeng, Yuting Wu, and Baobao Chang. 2021. Sire: Separate intra- and inter-sentential reasoning for document-level relation extraction. In *Findings of EMNLP*, pages 524–534.

Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. 2020. Double graph based reasoning for document-level relation extraction. In *Proceedings of EMNLP*, pages 1630–1640.

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 IJCAI*, pages 3999–4006.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of ACL*, pages 207–212.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Proceedings of AAAI*, 16, pages 14612–14620.

A Appendix

A.1 Statistics of Datasets

We present statistics of the public DocRE datasets, including DocRED (Yao et al., 2019), RE-DocRED (Tan et al., 2022b), and the distantly supervised data provided by Yao et al. (2019).

Dataset	# Document	Avg. # Instance
DocRED-Train	3,053	12.5
DocRED-Dev	1,000	12.3
DocRED-Test	1,000	12.8
Re-DocRED-Train	3,053	28.1
Re-DocRED-Dev	500	34.6
Re-DocRED-Test	500	34.9
Distantly Supervised	101,873	14.8

Table 5: Statistics of the Re-DocRED, DocRED, and distantly supervised dataset.

A.2 Results of Baselines with Fine-tuning

We present the results of baseline models that are pretrained on our denoised data and fine-tuned on the human-annotated data of the RE-DocRED dataset. As shown in Table 6, the final performance of most baseline models that are pretrained on our denoised data is significantly improved.

A.3 Sentence-level Relation Extraction

Our framework can also be applied to the sentence-level relation extraction task. We reconstruct a sentence-level relation extraction dataset from the distantly supervised training data and RE-DocRED datasets. The sentence-level RE dataset contains 231,107 DS training data, 10,033 human-annotated training data, 1,862 human-annotated development data, and 1,794 human-annotated test data. We perform our UGDRE on the sentence-level RE task in the same training way. As shown in Table

Model	DS				After Denoising				Improvement	
	Dev		Test		Dev		Test			
	F_1	Ign F_1	F_1	Ign F_1	F_1	Ign F_1	F_1	Ign F_1	ΔF_1	ΔIgn F_1
ATLOP	74.34	73.62	74.23	73.53	77.30	76.63	76.95	76.28	2.72	2.75
DocuNet	76.22	75.50	75.35	74.61	77.69	76.90	77.72	76.97	2.37	2.36
NCRL	75.85	74.91	75.90	75.00	77.71	76.84	76.78	75.92	0.88	0.92
KD-NA	76.14	75.25	76.00	75.12	78.16	77.23	77.73	76.86	1.73	1.74
UGDRE (Ours)	76.91	76.00	76.16	75.23	78.28	77.32	78.14	77.24	1.98	2.01

Table 6: Experimental results of baselines fine-tuned on human-annotated training data of RE-DocRED dataset, which are pretrained on original DS data and our denoised DS data.

Model	Dev		Test	
	F_1	Ign F_1	F_1	Ign F_1
UGDRE-SRE	79.00	78.38	78.52	77.94
w/o Pretrain	76.88	76.28	76.41	75.86
w/o DDS	78.08	77.47	77.68	77.14

Table 7: Experimental results of our framework on sentence-level relation extraction task.

7, the performance of the final sentence-level RE model UGDRE-SRE that pretrained on the DDS data is also improved.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations

☑ A2. Did you discuss any potential risks of your work?
Limitations

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Introduction (Sec.1)

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Not applicable. Left blank.

C ☑ Did you run computational experiments?

Experiments(Sec.4), Analysis and Discussion(Sec.5)

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Dataset and Settings (Sec.4.1)

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Dataset and Settings (Sec.4.1)

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Experiments(Sec.4), Analysis and Discussion(Sec.5)

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Not applicable. Left blank.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.