A Unified Positive-Unlabeled Learning Framework for Document-Level Relation Extraction with Different Levels of Labeling

Ye Wang[1], Xinxin Liu[1], Wenxin Hu[1]*, Tao Zhang[2]
[1]East China Normal University, Shanghai, China
[2]Tsinghua University, Beijing, China
{yewang, xxliu}@stu.ecnu.edu.cn, wxhu@cc.ecnu.edu.cn
tao-zhan20@mails.tsinghua.edu.cn

Abstract

Document-level relation extraction (RE) aims to identify relations between entities across multiple sentences. Most previous methods focused on document-level RE under full supervision. However, in real-world scenario, it is expensive and difficult to completely label all relations in a document because the number of entity pairs in document-level RE grows quadratically with the number of entities. To solve the common incomplete labeling problem, we propose a unified positive-unlabeled learning framework – shift and squared ranking loss positive-unlabeled (SSR-PU) learning. We use positive-unlabeled (PU) learning on document-level RE for the first time. Considering that labeled data of a dataset may lead to prior shift of unlabeled data, we introduce a PU learning under prior shift of training data. Also, using none-class score as an adaptive threshold, we propose squared ranking loss and prove its Bayesian consistency with multi-label ranking metrics. Extensive experiments demonstrate that our method achieves an improvement of about 14 F1 points relative to the previous baseline with incomplete labeling. In addition, it outperforms previous state-of-the-art results under both fully supervised and extremely unlabeled settings as well.[1]

1 Introduction

Relation extraction (RE) aims to identify the relations between two entities in a given text. It has rich applications in knowledge graph construction, question answering, and biomedical text understanding. Most of the previous work was to extract relations between entities in a single sentence (Miwa and Bansal, 2016; Zhang et al., 2018). Recently, document-level RE aiming to identify the relations among various entity pairs expressed in multiple sentences has received increasing research attention (Yao et al., 2019; Zhou et al., 2021; Xu et al., 2022).

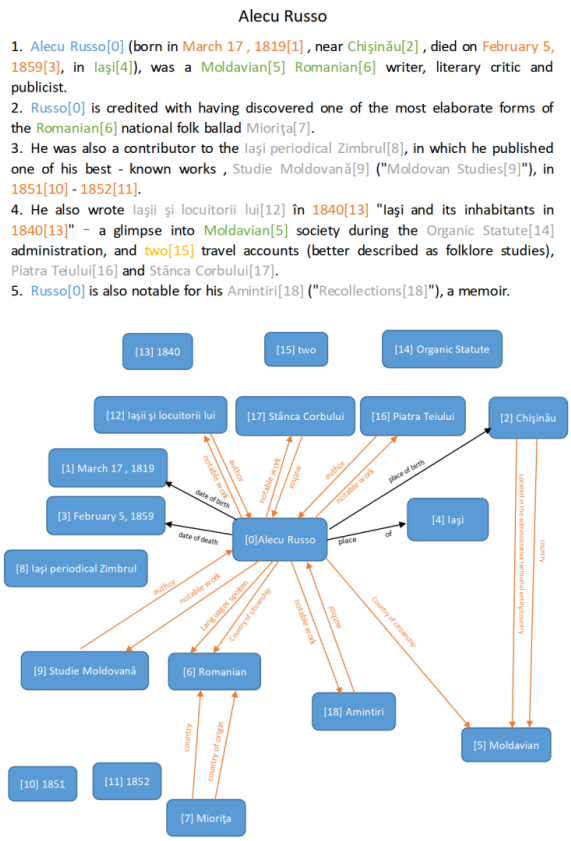

Figure 1: A case from DocRED. Entities are highlighted in different colors depending on their type. Black arrows indicate relations annotated with the original dataset, orange arrows indicate relations that are re-annotated by (Tan et al., 2022b).

Previous document-level RE methods mainly deal with fully supervised scenarios. However, in real-world scenarios, incomplete labeling is a common problem in document-level RE because the number of entity pairs grows quadratically with the number of entities. DocRED (Yao et al., 2019) is a popular dataset for document-level RE. Recent studies (Huang et al., 2022; Tan et al., 2022b) found that DocRED, which annotates data with a

*Corresponding author.
[1]Code and data are available at https://github.com/www-Ye/SSR-PU.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 4123–4135
December 7-11, 2022 ©2022 Association for Computational Linguistics

recommend-revise scheme, contains a large number of false negative samples, i.e., many positive samples being unlabeled. As shown in Figure 1, document *Alecu Russo* contains a large number of unlabeled positive relations. Consequently, the models trained on this dataset tend to overfit in real scenarios and get lower recall. As a result, document-level RE with incomplete labeling has become an emergency need.

To solve this problem, we propose a unified positive-unlabeled learning framework − shift and squared ranking loss positive-unlabeled (SSR-PU) learning, which can be adapted to labeling under different levels. We use positive-unlabeled (PU) learning for the first time on the document-level RE task. Since document-level RE is a multi-label classification task, we apply a binary PU learning method for each class (one-vs-all), converting it to multi-label PU learning. In addition, according to our observations, a considerable portion of the relations in DocRED, a dataset annotated by *recommend-revise* scheme, have already been annotated. This leads to the deviation between the prior distribution of the unlabeled data and the overall prior distribution. To address this problem, we introduce an adaptive PU learning under prior shift of training data that adjusts the model based on the estimated overall prior distribution and the labeled positive sample distribution to be similar to ordinary PN learning or ordinary PU learning. Here positive-negative (PN) learning means treating all unlabeled samples as negative samples.

Also, to distinguish between none-class and pre-defined classes, we propose a squared ranking loss for none-class ranking such that positive pre-defined labels are ranked higher than none-class label and negative pre-defined labels are ranked lower. This is an ideal multi-label surrogate loss metric, and we theoretically prove its Bayesian consistency with the multi-label ranking metric proposed by (Zhou and Lee, 2022). This loss function can be well adapted to PU learning.

We conduct extensive experiments on two multi-label document-level RE datasets with incomplete labeling, DocRED (Yao et al., 2019) and ChemDisGene (Zhang et al., 2022), a newly proposed multi-labeled biomedical document-level RE dataset. Experimental results show that our method SSR-PU outperforms previous baseline that did not consider the labeling incompleteness phenomenon by about 14 F1 points. In addition, we perform fully supervised experiments, as well as experiments on an extremely unlabeled data that is newly constructed, in which the number of each relation type labeled in each document is limited to 1. Experiments under two complementary settings demonstrate the effectiveness of our method with different levels of labeling. The contributions of this paper are summarized as follows:

- We propose a unified positive-unlabeled learning framework, SSR-PU, to adapt document-level RE with different levels of incomplete labeling.
- We apply PU learning for the first time to the document-level RE task and introduce a PU learning under prior shift of training data that can reach a balance between ordinary PN learning and ordinary PU learning based on the estimated prior and labeling distribution.
- We propose squared ranking loss, which effectively improves performance relative to other loss functions, and prove its Bayesian consistency with multi-label ranking metrics.
- Our method achieves state-of-the-art results in a variety of settings and provides a robust baseline for document-level RE with incomplete labels.

2 Related Work

Document-level relation extraction. Previous generally effective methods for document-level RE are mainly graph-based models and transformer-based models. Graph-based models (Nan et al., 2020; Li et al., 2020; Zeng et al., 2020, 2021; Xu et al., 2021b) gather entity information for relational inference with graph neural networks, and transformer-based methods (Zhou et al., 2021; Xu et al., 2021a; Zhang et al., 2021; Tan et al., 2022a) implicitly capture long-range dependencies. Recently, (Huang et al., 2022; Tan et al., 2022b) found that a large number of positive relations remain unlabeled in document-level RE datasets, especially unpopular relations. However, the previous methods did not consider unlabeled data separately. They simply treated them all as negative samples, which led to a lower recall and a significant drop in performance in realistic scenarios.

PU learning. Positive-unlabeled (PU) learning (Elkan and Noto, 2008; du Plessis et al., 2014; Plessis et al., 2015; Kiryo et al., 2017; Garg et al.,

2021) aims to learn a classifier from positive and unlabeled data. PU learning is a kind of semi-supervised learning but there is a fundamental difference between them: while semi-supervised learning requires labeled negative data, PU learning requires only labeled positive data. Many current PU learning methods rely on an overall prior estimate, while some recent studies (Charoenphakdee and Sugiyama, 2019; Nakajima and Sugiyama, 2021) have noticed a prior shift between the training set and the test set. On the other hand, PU learning has been used in many NLP applications, e.g., text classification (Li and Liu, 2003), sentence embedding (Cao et al., 2021), named entity recognition (Peng et al., 2019; Zhou et al., 2022), knowledge graph completion (Tang et al., 2022) and sentence-level RE (He et al., 2020). However, this method is rarely applied to the document-level RE task.

Multi-label classification. Multi-label classification is a widely investigated problem, and here we focus on the loss function. Binary cross entropy (BCE) is the most popular multi-label loss, reducing the multi-label problem to a number of independent binary (one-vs-all) classification tasks. Recently, (Hui and Belkin, 2020) have found that squared loss can also achieve better results in classification tasks. Another common multi-label loss function is pairwise ranking loss, which transforms multi-label learning into a ranking problem via pairwise (one-vs-one) comparison (Fürnkranz et al., 2008; Li et al., 2017). For multi-label PU learning, (Kanehira and Harada, 2016) treated it as a multi-label PU ranking problem, and (Aota et al., 2021) applied PU learning to multi-label common vulnerabilities and exposure classification by using one-vs-all strategy. For document-level RE task, (Zhou and Lee, 2022) proposed a none-class ranking multi-label metric. This multi-label metric has not yet been applied to PU learning.

3 Methodology

In this section, we introduce the details of our method shift and squared ranking loss positive-unlabeled (SSR-PU) learning for document-level RE with incomplete labeling. Firstly, we introduce the definition of positive-unlabeled learning for document-level RE. Next, we present the PU learning under prior shift of training data. Finally, squared ranking loss using the none-class score as an adaptive threshold is proposed.

3.1 Positive-unlabeled learning for document-level RE

Document-level RE can be viewed as a multi-label classification task, where each entity pair is an instance and the associated relations are label samples. Previous supervised learning methods only treated unlabeled relations as negative samples, which may lead to low recall in the presence of a large number of false negatives. To address this problem, we adopt PU learning (du Plessis et al., 2014; Plessis et al., 2015) for each class.

Let $\mathcal{X}$ be an instance space and $\mathcal{Y} = \{-1, +1\}^K$ be a label space, where K is the number of predefined classes. An instance $\boldsymbol{x} \in \mathcal{X}$ is associated with a subset of labels, identified by a binary vector $\boldsymbol{y} \in \mathcal{Y} = (y_1, \ldots, y_K)$, where $y_i = +1$ if the i-th label is positive for $\boldsymbol{x}$, and $y_i = -1$ otherwise. A score function is defined as $\boldsymbol{f}(\boldsymbol{x}) = (f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), ..., f_K(\boldsymbol{x}))$. In the following we use f_i instead, to omit the dependency on x.

For i-th class, assume that the data follow an unknown probability distribution with density $p(\boldsymbol{x}, y_i)$, $p_{\mathrm{P}_i} = p(\boldsymbol{x} \mid y_i = +1)$ as the positive marginal, $p_{\mathrm{N}_i} = p(\boldsymbol{x} \mid y_i = -1)$ as the negative marginal, and $p_i(\boldsymbol{x})$ as the marginal. In positive-negative (PN) learning, the goal is to minimize the expected classification risk:

$$R_{\mathrm{PN}}(f) = \sum_{i=1}^{K} \mathbb{E}_{\boldsymbol{x}, y_i \sim p(\boldsymbol{x}, y_i)}[\ell(f_i, y_i)], \quad (1)$$

Here, Eq.1 can be calculated by equivalently using the sum of the errors of positive and negative samples:

$$\begin{aligned} R_{\mathrm{PN}}(f) = \sum_{i=1}^{K} (\pi_i \mathbb{E}_{\mathrm{P}_i}[\ell(f_i, +1)] \\ + (1 - \pi_i)\mathbb{E}_{\mathrm{N}_i}[\ell(f_i, -1)]), \end{aligned} \quad (2)$$

where $\pi_i = p(y_i = +1)$ and $(1-\pi_i) = (1-p(y_i = +1)) = p(y_i = -1)$ is the positive and negative prior of the i-th class. $\mathbb{E}_{\mathrm{P}_i}[\cdot] = \mathbb{E}_{\boldsymbol{x} \sim p(\boldsymbol{x}|y_i=+1)}[\cdot]$, $\mathbb{E}_{\mathrm{N}_i}[\cdot] = \mathbb{E}_{\boldsymbol{x} \sim p(\boldsymbol{x}|y_i=-1)}[\cdot]$ and the loss function is represented by ℓ. Rewriting Eq.2 into a form that uses the data for approximation, we get:

$$\begin{aligned} \widehat{R}_{\mathrm{PN}}(f) = \sum_{i=1}^{K} (\frac{\pi_i}{n_{\mathrm{P}_i}} \sum_{j=1}^{n_{\mathrm{P}_i}} \ell(f_i(\boldsymbol{x}_j^{\mathrm{P}_i}), +1) \\ + \frac{(1 - \pi_i)}{n_{\mathrm{N}_i}} \sum_{j=1}^{n_{\mathrm{N}_i}} \ell(f_i(\boldsymbol{x}_j^{\mathrm{N}_i}), -1)), \end{aligned} \quad (3)$$

where $\boldsymbol{x}_j^{\mathrm{P}_i}$ and $\boldsymbol{x}_j^{\mathrm{N}_i}$ denote cases that the j-th sample of class i is positive or negative. n_{P_i} and n_{N_i} are the number of positive and negative samples of class i, respectively.

In positive-unlabeled (PU) learning, due to the absence of negative samples, we cannot estimate $\mathbb{E}_{\mathrm{N}_i}[\cdot]$ from the data. Following (du Plessis et al., 2014), PU learning assumes that unlabeled data can reflect the true overall distribution, that is, $p_{\mathrm{U}_i}(\boldsymbol{x}) = p_i(\boldsymbol{x})$. The expected classification risk formulation can be defined as:

$$R_{\mathrm{PU}}(f) = \sum_{i=1}^{K}(\pi_i \mathbb{E}_{\mathrm{P}_i}[\ell(f_i, +1)] + \mathbb{E}_{\mathrm{U}_i}[\ell(f_i, -1)] - \pi_i \mathbb{E}_{\mathrm{P}_i}[\ell(f_i, -1)]), \tag{4}$$

Here $\mathbb{E}_{\mathrm{U}_i}[\cdot] = \mathbb{E}_{\boldsymbol{x}\sim p_i(\boldsymbol{x})}[\cdot]$ and $\mathbb{E}_{\mathrm{U}_i}[\ell(f_i, -1)] - \pi_i \mathbb{E}_{\mathrm{P}_i}[\ell(f_i, -1)]$ can alternatively represent $(1 - \pi_i)\mathbb{E}_{\mathrm{N}_i}[\ell(f_i, -1)]$ because $p_i(\boldsymbol{x}) = \pi_i p_{\mathrm{P}_i}(\boldsymbol{x}) + (1 - \pi_i) p_{\mathrm{N}_i}(\boldsymbol{x})$.

By rewriting Eq.4 in a form that can be approximated using the data, we get the following:

$$\begin{aligned}\widehat{R}_{\mathrm{PU}}(f) = &\sum_{i=1}^{K}(\frac{\pi_i}{n_{\mathrm{P}_i}} \sum_{j=1}^{n_{\mathrm{P}_i}} \ell(f_i(\boldsymbol{x}_j^{\mathrm{P}_i}), +1) \\ &+ [\frac{1}{n_{\mathrm{U}_i}} \sum_{j=1}^{n_{\mathrm{U}_i}} \ell(f_i(\boldsymbol{x}_j^{\mathrm{U}_i}), -1) \\ &- \frac{\pi_i}{n_{\mathrm{P}_i}} \sum_{j=1}^{n_{\mathrm{P}_i}} \ell(f_i(\boldsymbol{x}_j^{\mathrm{P}_i}), -1)]),\end{aligned} \tag{5}$$

where $\boldsymbol{x}_j^{\mathrm{U}_i}$ denote cases that the j-th sample is unlabeled as class i and n_{U_i} is the number of samples unlabeled as class i.

However, the second term in Eq.5 can be negative and can be prone to overfitting when using a highly flexible model. Thus, a non-negative risk estimator (Kiryo et al., 2017) is proposed to alleviate the overfitting problem:

$$\begin{aligned}\widehat{R}_{\mathrm{PU}}(f) = &\sum_{i=1}^{K}(\frac{\pi_i}{n_{\mathrm{P}_i}} \sum_{j=1}^{n_{\mathrm{P}_i}} \ell(f_i(\boldsymbol{x}_j^{\mathrm{P}_i}), +1) + \\ &\max(0, [\frac{1}{n_{\mathrm{U}_i}} \sum_{j=1}^{n_{\mathrm{U}_i}} \ell(f_i(\boldsymbol{x}_j^{\mathrm{U}_i}), -1) \\ &- \frac{\pi_i}{n_{\mathrm{P}_i}} \sum_{j=1}^{n_{\mathrm{P}_i}} \ell(f_i(\boldsymbol{x}_j^{\mathrm{P}_i}), -1)])).\end{aligned} \tag{6}$$

For ℓ, we use the convex function squared loss:

$$\ell(f_i, y_i) = \frac{1}{4}(y_i f_i - 1)^2, \tag{7}$$

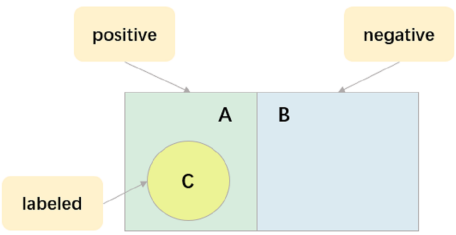

Figure 2: Positive sample distribution shift after labeled, i.e., $p(A \mid \overline{C}) \neq p(A)$

and we compare the performance of using squared loss and log-sigmoid loss functions in Section 4.4. The latter is a convex loss function commonly used in classification.

In addition, to solve the heavy class imbalance problem, we multiply $\gamma_i = (\frac{1-\pi_i}{\pi_i})^{0.5}$ before positive risk estimations as the class weight.

3.2 Class prior shift of training data

Ordinary PU learning requires an assumption that the overall distribution needs to be the same as the distribution of the unlabeled data. In contrast, with the document-level RE dataset constructed by a *recommend-revise* scheme, many relations are probably already annotated, especially the common ones. This leads to a prior shift in the unlabeled data of the training set. When this assumption is broken, ordinary PU learning will yield a biased result. To address this problem, inspired by the method (Charoenphakdee and Sugiyama, 2019) for handling a prior shift between the test set and the training set, we introduce the PU learning under prior shift of training data.

For each class, assume that the original prior $\pi_i = p(y_i = +1)$. We set $\pi_{labeled,i} = p(s_i = +1)$ and $(1 - \pi_{labeled,i}) = (1 - p(s_i = +1)) = p(s_i = -1)$ where $s_i = +1$ or $s_i = -1$ mean that the i-th class is labeled or unlabeled, respectively. As shown in Figure 2, the conditional probability of a positive sample under unlabeled data is different from the probability of an overall positive sample. The conditional probability of a positive sample under unlabeled data is:

$$p(y_i = 1 \mid s_i = -1) = \frac{p(y_i = 1, s_i = -1)}{p(s_i = -1)}, \tag{8}$$

where $p(y_i = 1, s_i = -1) = \pi_i - \pi_{labeled,i}$, we can obtain the prior of positive samples in the new unlabeled data after labeling as $\pi_{u,i} = p(y_i = 1 \mid s_i = -1) = \frac{\pi_i - \pi_{labeled,i}}{1 - \pi_{labeled,i}}$.

For document-level RE, the goal is to minimize the following misclassification risk for the original distribution of the training data:

$$R_{\mathrm{ori}}(f) = \sum_{i=1}^{K}(\pi_i \mathbb{E}_{\mathrm{P}_i}[\ell(f_i, +1)] + (1-\pi_i)\mathbb{E}_{\mathrm{N}_i}[\ell(f_i, -1)]). \tag{9}$$

We can express $R_{\mathrm{ori}}(f)$ using the expectation of positive and unlabeled data by the following theorem.

Theorem 1. *The misclassification risk $R_{\mathrm{ori}}(f)$ can be equivalently expressed as*

$$R_{\mathrm{S-PU}}(f) = \sum_{i=1}^{K}(\pi_i \mathbb{E}_{\mathrm{P}_i}[\ell(f_i, +1)] + \frac{1-\pi_i}{1-\pi_{u,i}}\mathbb{E}_{\mathrm{U}_i}[\ell(f_i, -1)] - \frac{\pi_{u,i} - \pi_{u,i}\pi_i}{1-\pi_{u,i}}\mathbb{E}_{\mathrm{P}_i}[\ell(f_i, -1)]). \tag{10}$$

Proof. Proof appears in Appendix A.1. □

As a result, we can obtain the non-negative risk estimator (Kiryo et al., 2017) under class prior shift of training data as follows:

$$\widehat{R}_{\mathrm{S-PU}}(f) = \sum_{i=1}^{K}(\frac{1}{n_{\mathrm{P}_i}}\pi_i \sum_{j=1}^{n_{\mathrm{P}_i}} \ell(f_i(\boldsymbol{x}_j^{\mathrm{P}_i}), +1) + \max(0, [\frac{1}{n_{\mathrm{U}_i}}\frac{1-\pi_i}{1-\pi_{u,i}}\sum_{j=1}^{n_{\mathrm{U}_i}} \ell(f_i(\boldsymbol{x}_j^{\mathrm{U}_i}), -1) - \frac{1}{n_{\mathrm{P}_i}}\frac{\pi_{u,i} - \pi_{u,i}\pi_i}{1-\pi_{u,i}}\sum_{j=1}^{n_{\mathrm{P}_i}} \ell(f_i(\boldsymbol{x}_j^{\mathrm{P}_i}), -1)])). \tag{11}$$

We can observe that PN learning and PU learning are special cases of this function. When $\pi_{u,i} = 0$, this equation reduces to the form of ordinary PN learning, and when $\pi_{u,i} = \pi_i$, this equation reduces to the form of ordinary PU learning.

3.3 Squared ranking loss

To better measure the performance of document-level RE, (Zhou and Lee, 2022) proposed a new multi-label performance measure:

$$L_{\mathrm{NA}}(\boldsymbol{f}, \boldsymbol{y}) = \sum_{i=1}^{K}([\![y_i > 0]\!][\![f_i < f_0]\!] + [\![y_i \leq 0]\!][\![f_i > f_0]\!] + \frac{1}{2}[\![f_i = f_0]\!]), \tag{12}$$

Dataset	DocRED		ChemDisGene	
	train	test	train	test
# docs	3,053	500	76,942	523
# rels	96		14	
Avg # ents	19.5	19.6	7.5	10.0
Avg # rels	12.5	34.9	2.1	7.2

Table 1: Statistics of Document-level RE Datasets

where positive pre-defined labels should be ranked higher than the none-class label and negative ones should be ranked below. $[\![\cdot]\!]$ is an indicator function that takes the value of 1 when the conditions in the parentheses are met, otherwise 0.

However, it is difficult to optimize the above equation directly. Thus, we propose the squared ranking surrogate loss by rewriting Eq.7 as:

$$\ell_{\mathrm{SR}}(f_i, y_i) = \frac{1}{4}(y_i(f_i - f_0) - margin)^2, \tag{13}$$

where $margin$ is a hyper-parameter and f_0 is the none-class score, when f_i is greater than f_0 the label exists, and otherwise not.

Next we prove the Bayesian consistency of ℓ_{SR} with the multi-label ranking metric L_{NA} when $margin \neq 0$. Given an instance x, let $\Delta_i = \mathrm{P}(y_i = 1 \mid x)$ be the marginal probability when the i-th label is positive, the Bayes optimal score function $\boldsymbol{f}^*_{\mathrm{NA}}$ that minimizes the multi-label risk $\mathbb{E}[L_{\mathrm{NA}}(\mathrm{P}, \boldsymbol{f}) \mid \boldsymbol{x}]$ is given by:

$$\boldsymbol{f}^*_{\mathrm{NA}} \in \{\boldsymbol{f} : f_i > f_0 \ if \ \Delta_i > \frac{1}{2}, \ and \ f_i < f_0 \ if \ \Delta_i < \frac{1}{2}\}. \tag{14}$$

The next theory guarantees that the classifier obtained by minimizing the surrogate loss ℓ_{SR} converges to the classifier with the lowest multi-label risk, thus making it possible to achieve a better classification performance w.r.t. corresponding to the multi-label performance metric.

Theorem 2. *ℓ_{SR} (Eq.13) is Bayes consistent w.r.t. L_{NA} (Eq.12) when $margin \neq 0$.*

Proof. Proof appears in Appendix A.2. □

As a supplement, we likewise compare the log-sigmoid ranking loss performance in Section 4.4.

4 Experiments

In this section, we evaluate our method on two multi-label document-level RE datasets with in-

Model	Ign F1	F1	P	R
BiLSTM*	32.57 ± 0.22	32.86 ± 0.22	77.04 ± 1.01	20.89 ± 0.17
GAIN+BERT$^*_{Base}$	45.57 ± 1.36	45.82 ± 1.38	88.11 ± 1.07	30.98 ± 1.36
DocuNET+RoBERTa$^*_{Large}$	45.88 ± 0.33	45.99 ± 0.33	94.16 ± 0.32	30.42 ± 0.29
ATLOP+BERT$^*_{Base}$	43.12 ± 0.24	43.25 ± 0.25	$\mathbf{92.49 \pm 0.33}$	28.23 ± 0.23
PN+ATLOP+BERT$_{Base}$	51.11 ± 0.49	51.68 ± 0.40	77.55 ± 3.10	38.79 ± 0.49
SR-PN+ATLOP+BERT$_{Base}$	52.70 ± 0.28	53.10 ± 0.26	83.76 ± 0.49	38.87 ± 0.23
PU+ATLOP+BERT$_{Base}$	51.80 ± 1.11	53.14 ± 1.01	58.81 ± 2.41	48.15 ± 0.14
SR-PU+ATLOP+BERT$_{Base}$	53.87 ± 0.27	55.06 ± 0.25	63.42 ± 0.64	$\mathbf{48.66 \pm 0.11}$
S-PU+ATLOP+BERT$_{Base}$	53.36 ± 1.22	54.44 ± 1.12	65.95 ± 2.84	46.38 ± 0.22
SSR-PU+ATLOP+BERT$_{Base}$	$\mathbf{55.21 \pm 0.12}$	$\mathbf{56.14 \pm 0.12}$	70.42 ± 0.18	46.67 ± 0.14
ATLOP+RoBERTa$^*_{Large}$	45.09 ± 0.26	45.19 ± 0.27	$\mathbf{94.75 \pm 0.25}$	29.67 ± 0.24
PN+ATLOP+RoBERTa$_{Large}$	54.21 ± 0.34	54.47 ± 0.35	89.22 ± 0.36	39.20 ± 0.41
SR-PN+ATLOP+RoBERTa$_{Large}$	56.06 ± 0.21	56.39 ± 0.23	87.47 ± 0.60	41.61 ± 0.39
PU+ATLOP+RoBERTa$_{Large}$	56.97 ± 0.47	58.04 ± 0.43	67.39 ± 1.22	50.98 ± 0.39
SR-PU+ATLOP+RoBERTa$_{Large}$	57.64 ± 0.25	58.77 ± 0.26	66.39 ± 0.47	$\mathbf{52.72 \pm 0.44}$
S-PU+ATLOP+RoBERTa$_{Large}$	58.19 ± 0.24	58.95 ± 0.25	75.68 ± 0.36	48.29 ± 0.40
SSR-PU+ATLOP+RoBERTa$_{Large}$	$\mathbf{58.68 \pm 0.43}$	$\mathbf{59.50 \pm 0.45}$	74.21 ± 0.53	49.67 ± 0.77

Table 2: Results on Re-DocRED revised test set. Results with $*$ are based on our implementation.

complete labeling. We also demonstrate the effectiveness of our method with different levels of labeling.

4.1 Experimental Setups

Datasets. *DocRED* (Yao et al., 2019) is a large-scale document-level RE dataset with 96 predefined relations constructed by a *recommend-revise* scheme from Wikipedia. (Tan et al., 2022b) observed a large number of false negatives in the annotation of DocRED and provided a high-quality revised version, Re-DocRED. In our experiments, we use the incompletely labeled DocRED original training set for training and the revised test set for testing. *ChemDisGene* (Zhang et al., 2022) is a newly proposed biomedical multi-label document-level RE dataset. This corpus is automatically derived from CTD database (Davis et al., 2021) by distantly supervised method and has 523 abstracts labeled by domain experts as an additional *All relationships* test set. We use the distantly supervised training set for training and the *All relationships* test set for testing. The average number of relations per document in the test set on both two datasets is much larger than the average number of relations in the training set, which indicates the incomplete labeling phenomenon in the training set, with a large number of false negatives present. The statistics of the two datasets are listed in Table 1.

Implementation details. For each dataset, we use ATLOP (Zhou et al., 2021) as the encoding model for the representation learning of relations. Further, we apply cased BERT$_{Base}$ (Devlin et al., 2019) and RoBERTa$_{Large}$ (Liu et al., 2019) for DocRED and PubmedBert (Gu et al., 2021) for ChemDisGene. We use Huggingface's Transformers (Wolf et al., 2020) to implement all the models and AdamW (Loshchilov and Hutter, 2019) as the optimizer, and apply a linear warmup (Goyal et al., 2017) at the first 6% steps followed by a linear decay to 0. For DocRED, we set the learning rates for BERT$_{Base}$ and RoBERTa$_{Large}$ settings to 5e-5 and 3e-5, respectively, in the same way as ATLOP. For ChemDisGene, the learning rate is set to 2e-5. The batch size (number of documents per batch) is set to 4 and 8 for two datasets, respectively. During our experiment, we set $\pi_i = 3\pi_{labeled,i}$ and $margin = 0.25$. To evaluate the efficacy of our methods in realistic settings, we do not use any fully labeled validation or test sets in any stage of the training process. The training stopping criteria are set as follows: 30 epochs for both two dataset. We report the performance of the final model instead of the best checkpoint. All experiments are conducted with 1 Tesla A100-40G GPU.

Baseline. We re-implemented the existing fully supervised methods BiLSTM (Yao et al., 2019), GAIN (Zeng et al., 2020), DocuNET (Zhang et al., 2021) and ATLOP (Zhou et al., 2021) as the baseline models for DocRED in this new setup, where

Model	F1	P	R
BRAN†	32.5	41.8	26.6
PubmedBert†	42.1	64.3	31.3
BRAN+PubmedBert†	43.8	70.9	31.6
ATLOP+PubmedBert*	42.73 ± 0.36	$\mathbf{76.17 \pm 0.54}$	29.70 ± 0.36
PN+ATLOP+PubmedBert	44.25 ± 0.24	73.46 ± 0.95	31.67 ± 0.16
SR-PN+ATLOP+PubmedBert	46.56 ± 0.35	69.84 ± 0.54	34.93 ± 0.40
PU+ATLOP+PubmedBert	44.60 ± 0.70	46.56 ± 1.17	42.80 ± 0.35
SR-PU+ATLOP+PubmedBert	45.86 ± 0.38	46.91 ± 0.79	$\mathbf{44.86 \pm 0.37}$
S-PU+ATLOP+PubmedBert	46.73 ± 0.49	53.95 ± 1.14	41.23 ± 0.36
SSR-PU+ATLOP+PubmedBert	$\mathbf{48.56 \pm 0.23}$	54.27 ± 0.40	43.93 ± 0.32

Table 3: Results on ChemDisGene *All relationships* test set. Results with † are reported from (Zhang et al., 2022). Results with $*$ are based on our implementation.

for GAIN and BiLSTM we use a fixed threshold of 0.5 and all methods take the final result of the model instead of the best checkpoint. For ChemDisGene, we used BRAN (Verga et al., 2018), PubmedBert (Gu et al., 2021) and PubmedBert + BRAN mentioned in (Zhang et al., 2022) as the baseline models, and ATLOP is re-implemented as a supplementary baseline.

Evaluation metric. For DocRED, we use the micro F1 (F1), micro ignore F1 (Ign F1), precision (P) and recall (R) as the evaluation metrics to evaluate the overall performance of a model. Ign F1 measures the F1 score excluding the relations shared by the training and test set. For ChemDisGene, we use micro F1 (F1), precision (P) and recall (R) as the evaluation metrics.

4.2 Main Results

In this subsection, we present the results of comparison of PN learning (PN), squared ranking loss PN learning (SR-PN), PU learning (PU), squared ranking loss PU learning (SR-PU), PU learning under prior shift of training data (S-PU) and SSR-PU. All methods use the same encoder and different loss functions. For each method, we use the same hyper-parameter settings and report the mean and standard deviation on the test set by conducting 5 runs with different random seeds (62, 63, 64, 65, 66).

Results on DocRED. As shown in Table 2, our SSR-PU method achieves a state-of-the-art F1 and Ign F1 in both BERT_{Base} and $\mathrm{RoBERTa}_{Large}$ settings and outperforms the original ATLOP by 13.58 and 14.52 F1 points, respectively. Meanwhile, consistent with the observation in the paper (Huang et al., 2022), existing document-level RE methods under full supervision have a significant performance degradation in the incompletely labeled scenario.

The original ATLOP method has the highest precision (P) but low recall (R), which implies that supervised learning methods that simply treat unlabeled data as negative samples lack the generalization ability to extract instances of relations that are systematically missed in the dataset. PN learning uses an estimated prior, but will yield a biased result because there are still positive samples in the unlabeled data. While PU learning uses both unlabeled and labeled data to better estimate the expectation of negative samples, which results in a higher recall rate. In addition, ordinary PU methods without prior shift overestimate the content of positive samples in unlabeled data, which means that the model will tend to identify more samples as positive, i.e., higher recall, but also leads to more false-positive prediction results, i.e., lower precision. In contrast, the S-PU method with prior shift effectively mitigates this phenomenon by bringing the positive samples estimated by the model in the unlabeled data closer to their true distribution. For example, in experiments under the BERT_{Base} setting, there is a small decrease in recall of less than 2 percentage points, while the precision improves by about 7 percentage points, leading to an improvement in the final results. And this phenomenon is more evident in common relations as analyzed in Section 4.4. Finally, applying squared ranking loss in PN learning, PU learning and S-PU learning can further improve the performance of the model, demonstrating the effectiveness of the method with none-class score as an adaptive threshold for document-level RE.

Model	Ign F1	F1
ATLOP+BERT$^*_{Base}$	72.70	73.47
SSR-PU+BERT$_{Base}$	**72.91**	**74.33**
ATLOP+RoBERTa$^*_{Large}$	76.92	77.58
DocuNET+RoBERTa$^\dagger_{Large}$	77.27	77.92
KD-DocRE+RoBERTa$^\dagger_{Large}$	77.63	78.35
SSR-PU+RoBERTa$_{Large}$	**77.67**	**78.86**

Table 4: Results on Re-DocRED revised test set under the fully supervised setting. Results with † are reported from (Tan et al., 2022b). Results with $*$ are based on our implementation.

Model	Ign F1	F1
ATLOP+BERT$^*_{Base}$	16.99	17.01
SSR-PU+BERT$_{Base}$	**46.47**	**47.24**
ATLOP+RoBERTa$^*_{Large}$	17.29	17.31
SSR-PU+RoBERTa$_{Large}$	**48.98**	**49.74**

Table 5: Results on Re-DocRED revised test set under the extremely unlabeled setting. Results with $*$ are based on our implementation.

Results on ChemDisGene. As shown in Table 3, the improvement of our method agrees with the results on DocRED, reaching the state-of-the-art F1, which is 5.83 F1 points higher than the original ATLOP. Notice that the improvement on ChemDisGene is not as dramatic as that on DocRED. We argue that this may be due to the fact that some of the documents in the extra annotated *All relationships* test set are from another corpus DrugProt (Miranda et al., 2021), and that the annotation by human experts has a large deviation from the original training set distribution. This suggests that it is a challenging direction to make the document-level RE model more generalizable when it is difficult to estimate the true distribution of the test set.

4.3 Different Levels of Labeling

Fully supervised setting. In this setting, we set $\pi_i = \pi_{labeled,i}$ and other hyper-parameters identically. As shown in Table 4, we use the (Tan et al., 2022b) revised Re-DocRED dataset in the same fully supervised setting to compare with the current state-of-the-art baseline models ATLOP (Zhou et al., 2021), DocuNET (Zhang et al., 2021) and KD-DocRE (Tan et al., 2022a). Our method achieves the same state-of-the-art results, demonstrating the effectiveness of our method with full labeling. The result with this setting can be seen as an upper bound for document-level RE with incomplete labeling. More details of the experiment are shown in Appendix A.3.

Model	Freq. F1	Freq. P	Freq. R
SR-PN	60.79	87.83	46.49
SR-PU	62.43	60.28	64.74
SSR-PU	64.88	68.36	61.74

Table 6: Results for the 10 most common relation types on Re-DocRED test set under the BERT$_{Base}$ setting.

Model	Freq. F1	Freq. P	Freq. R
SR-PN	47.62	71.76	35.64
SR-PU	47.65	44.35	51.48
SSR-PU	50.91	52.09	49.78

Table 7: Results for the 5 most common relation types on ChemDisGene *All relationships* test set.

Extremely unlabeled setting. In this setting, we use the original training set of DocRED to construct an extremely unlabeled training set, i.e., the number of labels for each relation type in the document being limited to 1. The average number of relations in the processed documents is reduced to 5.4. We consider this a more difficult and challenging scenario. We set $\pi_i = 12\pi_{labeled,i}$ and other hyper-parameters identically. As shown in Table 5, traditional supervised learning methods fail, while our proposed SSR-PU method still yields a robust result. It is worth noting that since the labeled sample is only a fraction of the true positive sample, i.e., the biased distribution, which means $p(x \mid y_i = 1)$ is not equal to $p(x \mid s_i = 1)$, the first term in Eq.11 is actually a biased approximation to the first term in Eq.10. We consider this bias as one of the bottlenecks of the current method and the main reason why the method degrades a lot in extremely unlabeled scenarios, i.e., the bias is widened in extremely unlabeled scenarios. This is a good direction for future research, where possible solutions might involve adding some data augmentation or bootstrapping methods for labeling to alleviate this bias. More details of the experiment are shown in Appendix A.4.

4.4 Additional Analysis

Analysis of common relations. As shown in table 6 and table 7, we show the results for common relations on DocRED and ChemDisGene, these frequent relation types account for about 60% of the relation triples (Tan et al., 2022b; Zhang et al., 2022). It can be seen that the SR-PU method has

Model	Ign F1	F1
S-PU$_{log-sigmoid}$	52.23	53.43
S-PU$_{squared}$	54.00	55.01
S-PU$_{log-sigmoid\ ranking}$	52.42	53.66
SSR-PU	55.43	56.36

Table 8: Results on Re-DocRED test set under the BERT$_{Base}$ setting with different loss functions.

a slightly higher recall and much lower precision, which corresponds to an overestimation of the positive sample size in the unlabeled data. The SSR-PU method, on the other hand, can alleviate this problem well, contributing to a better balance among precision and recall and better performance. This indicates a large amount of prior shift in common relations, which is consistent with (Huang et al., 2022) observation that common relations are more likely to be labeled in the dataset.

Comparison with other loss functions. We compare the squared loss with the log-sigmoid loss, which is commonly used in multi-label classification at the document-level RE. And again, this loss function is rewritten into a none-class ranking form for further comparison with squared ranking loss. The details of the loss function are listed in Appendix A.5. As shown in Table 8, both the squared loss function and the squared ranking loss function are significantly improved compared to the other loss functions, which demonstrates the effectiveness of our proposed loss function in the multi-label document-level RE task.

5 Conclusion and Future Work

In this paper, we propose a unified positive-unlabeled learning framework, SSR-PU, which can effectively solve the incomplete labeling of document-level RE. We use PU learning on document-level RE for the first time and introduce a PU learning under prior shift of training data to adapt to different levels of labeling. Also, we propose squared ranking loss, using none-class score as an adaptive threshold. Experiments demonstrate that our method achieves state-of-the-art results with different levels of labeling and provides a robust new baseline for incompletely labeled document-level RE. In the future, we will consider methods that do not require estimation of priors, allowing generalization to unknown distributions more accurately, as well as addressing the problem of biased distributions with incomplete labeled positive samples and further improving the extraction performance of long-tail relations.

Limitations

Regarding the limitations of our proposed method, our method requires an estimation of an overall prior that will affect the final result. In a realistic scenario, a very accurate prior estimation may be difficult to obtain. In addition, the biased distribution caused by the incomplete labeling of positive samples is one of the bottlenecks of the current method, and there is still much left to be improved for extremely unlabeled scenarios and scenarios where the gap between the test set and the training set distribution is too large, which can be a direction for further research. However, for now, we believe that our task is a valuable contribution to advancing the application of document-level RE in more realistic scenarios and provides a robust baseline for this direction.

Acknowledgements

We sincerely thank all anonymous reviewers for their valuable comments to improve our work. This research is funded by the Basic Research Project of Shanghai Science and Technology Commission (No.19JC1410101). The computation is supported by ECNU Multifunctional Platform for Innovation (001).

References

Masaki Aota, Tao Ban, Takeshi Takahashi, and Noboru Murata. 2021. Multi-label positive and unlabeled learning and its application to common vulnerabilities and exposure categorization. In *2021 IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pages 988–996.

Lele Cao, Emil Larsson, Vilhelm von Ehrenheim, Dhiana Deva Cavalcanti Rocha, Anna Martin, and Sonja Horn. 2021. PAUSE: Positive and annealed unlabeled sentence embedding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10096–10107, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nontawat Charoenphakdee and Masashi Sugiyama. 2019. Positive-unlabeled classification under class prior shift and asymmetric error. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pages 271–279. SIAM.

Allan Peter Davis, Cynthia J Grondin, Robin J Johnson, Daniela Sciaky, Jolene Wiegers, Thomas C Wiegers, and Carolyn J Mattingly. 2021. Comparative toxicogenomics database (ctd): update 2021. *Nucleic acids research*, 49(D1):D1138–D1143.

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

Marthinus C du Plessis, Gang Niu, and Masashi Sugiyama. 2014. Analysis of learning from positive and unlabeled data. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Charles Elkan and Keith Noto. 2008. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 213–220, New York, NY, USA. Association for Computing Machinery.

Johannes Fürnkranz, Eyke Hüllermeier, Eneldo Loza Mencía, and Klaus Brinker. 2008. Multilabel classification via calibrated label ranking. *Machine learning*, 73(2):133–153.

Saurabh Garg, Yifan Wu, Alexander J Smola, Sivaraman Balakrishnan, and Zachary Lipton. 2021. Mixture proportion estimation and pu learning:a modern approach. In *Advances in Neural Information Processing Systems*, volume 34, pages 8532–8544. Curran Associates, Inc.

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

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Zhengqiu He, Wenliang Chen, Yuyi Wang, Wei Zhang, Guanchun Wang, and Min Zhang. 2020. Improving neural relation extraction with positive and unlabeled learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7927–7934.

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

Like Hui and Mikhail Belkin. 2020. Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks. In *International Conference on Learning Representations*.

Atsushi Kanehira and Tatsuya Harada. 2016. Multi-label ranking from positive and unlabeled data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ryuichi Kiryo, Gang Niu, Marthinus C du Plessis, and Masashi Sugiyama. 2017. Positive-unlabeled learning with non-negative risk estimator. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Bo Li, Wei Ye, Zhonghao Sheng, Rui Xie, Xiangyu Xi, and Shikun Zhang. 2020. Graph enhanced dual attention network for document-level relation extraction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1551–1560, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Xiaoli Li and Bing Liu. 2003. Learning to classify texts using positive and unlabeled data. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, IJCAI'03, page 587–592, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Yuncheng Li, Yale Song, and Jiebo Luo. 2017. Improving pairwise ranking for multi-label image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

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

Antonio Miranda, Farrokh Mehryary, Jouni Luoma, Sampo Pyysalo, Alfonso Valencia, and Martin Krallinger. 2021. Overview of drugprot biocreative vii track: quality evaluation and large scale text mining of drug-gene/protein relations. In *Proceedings of the seventh BioCreative challenge evaluation workshop*.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using LSTMs on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1116, Berlin, Germany. Association for Computational Linguistics.

Shota Nakajima and Masashi Sugiyama. 2021. Positive-unlabeled classification under class-prior shift: A prior-invariant approach based on density ratio estimation. *arXiv preprint arXiv:2107.05045*.

Guoshun Nan, Zhijiang Guo, Ivan Sekulic, and Wei Lu. 2020. Reasoning with latent structure refinement for document-level relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1546–1557, Online. Association for Computational Linguistics.

Minlong Peng, Xiaoyu Xing, Qi Zhang, Jinlan Fu, and Xuanjing Huang. 2019. Distantly supervised named entity recognition using positive-unlabeled learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2409–2419, Florence, Italy. Association for Computational Linguistics.

Marthinus Du Plessis, Gang Niu, and Masashi Sugiyama. 2015. Convex formulation for learning from positive and unlabeled data. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1386–1394, Lille, France. PMLR.

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

Qingyu Tan, Lu Xu, Lidong Bing, and Hwee Tou Ng. 2022b. Revisiting docred–addressing the overlooked false negative problem in relation extraction. *arXiv preprint arXiv:2205.12696*.

Zhenwei Tang, Shichao Pei, Zhao Zhang, Yongchun Zhu, Fuzhen Zhuang, Robert Hoehndorf, and Xiangliang Zhang. 2022. Positive-unlabeled learning with adversarial data augmentation for knowledge graph completion. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 2248–2254. International Joint Conferences on Artificial Intelligence Organization. Main Track.

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

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

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

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

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

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

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

Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. 2020. Double graph based reasoning for document-level relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1630–1640, Online. Association for Computational Linguistics.

Dongxu Zhang, Sunil Mohan, Michaela Torkar, and Andrew McCallum. 2022. A distant supervision corpus for extracting biomedical relationships between chemicals, diseases and genes. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1073–1082, Marseille, France. European Language Resources Association.

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

Yuhao Zhang, Peng Qi, and Christopher D. Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2205–2215, Brussels, Belgium. Association for Computational Linguistics.

Kang Zhou, Yuepei Li, and Qi Li. 2022. Distantly supervised named entity recognition via confidence-based multi-class positive and unlabeled learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7198–7211, Dublin, Ireland. Association for Computational Linguistics.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):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-22*, pages 4538–4544. International Joint Conferences on Artificial Intelligence Organization. Main Track.

A Appendix

A.1 Proof of Theorem 1

Proof. Based on the fact that $p_{\mathrm{U}_i}(\boldsymbol{x}) = \pi_{u,i} p_{\mathrm{P}_i}(\boldsymbol{x}) + (1-\pi_{u,i}) p_{\mathrm{N}_i}(\boldsymbol{x})$, $(1-\pi_{u,i})\mathbb{E}_{\mathrm{N}_i}[\ell(f_i,-1)]$ can be alternatively expressed as $\mathbb{E}_{\mathrm{U}_i}[\ell(f_i,-1)] - \pi_{u,i}\mathbb{E}_{\mathrm{P}_i}[\ell(f_i,-1)]$. We can rewrite $R_{\mathrm{ori}}(f)$ as follows:

$$\begin{aligned} R_{\mathrm{ori}}(f) =& \sum_{i=1}^{K}(\pi_i \mathbb{E}_{\mathrm{P}_i}[\ell(f_i,+1)] \\ &+ (1-\pi_i)\mathbb{E}_{\mathrm{N}_i}[\ell(f_i,-1)]) \\ =& \sum_{i=1}^{K}(\pi_i \mathbb{E}_{\mathrm{P}_i}[\ell(f_i,+1)] \\ &+ \frac{1-\pi_i}{1-\pi_{u,i}}(\mathbb{E}_{\mathrm{U}_i}[\ell(f_i,-1)] \\ &- \pi_{u,i}\mathbb{E}_{\mathrm{P}_i}[\ell(f_i,-1)])) \\ =& R_{\mathrm{S-PU}}(f). \end{aligned} \tag{15}$$

We conclude that $R_{\mathrm{ori}}(f) = R_{\mathrm{S-PU}}(f)$. □

A.2 Proof of Theorem 2

Proof. Let $\Delta_i = \mathrm{P}(y_i = 1 \mid x)$ be the marginal probability when the i-th label is positive. The conditional risk of ℓ_{SR} is:

$$\begin{aligned} R_{\ell_{\mathrm{SR}}}(\mathrm{P}, \boldsymbol{f}) =& \sum_{i=1}^{K}(\Delta_i \frac{1}{4}((f_i - f_0) - margin)^2 \\ &+ (1-\Delta_i)\frac{1}{4}(-(f_i - f_0) - margin)^2). \end{aligned} \tag{16}$$

For $i = 1, ..., K$, the partial derivative can be computed by

$$\begin{aligned} &\frac{\partial}{f_i}\mathbb{E}[\ell_{\mathrm{SR}}(\mathrm{P}, \boldsymbol{f}) \mid \boldsymbol{x}] = \\ &\quad \sum_{i=1}^{K}(\Delta_i \frac{1}{2}((f_i - f_0) - margin) + \\ &\quad (1-\Delta_i)\frac{1}{2}((f_0 - f_i) - margin)), \end{aligned} \tag{17}$$

since ℓ_{SR} is convex and differentiable, we can obtain the optimal $\boldsymbol{f}^*$ by setting the partial derivatives to zero, which leads to

$$f_i^* - f_0^* = 2\Delta_i margin - margin, i = 1, ..., K. \tag{18}$$

When $margin \neq 0$, for the optimal score function $\boldsymbol{f}^*$, $f_i^* > f_0^*$ if and only if $\Delta_i > \frac{1}{2}$, which minimizes the ℓ_{SR} risk according to Eq.14. Therefore, ℓ_{SR} is Bayes consistent w.r.t. L_{NA}. □

A.3 Results under the Fully Supervised Setting

The detailed results under the fully supervised setting are shown in Table 9. We report the mean and standard deviation on the validation and test set by conducting 5 runs with different random seeds (62, 63, 64, 65, 66).

A.4 Results under the Extremely Unlabeled Setting

The detailed results under the extremely unlabeled setting are shown in Table 10. We report the mean and standard deviation on the test set by conducting 5 runs with different random seeds (62, 63, 64, 65, 66).

A.5 Details of Other Loss Functions

We first show the convex loss function log-sigmoid loss, which is commonly used in classification task:

$$\ell_{LS}(f_i, y_i) = -log(\sigma(y_i f_i)), \tag{19}$$

where $\sigma(x)$ is the sigmoid function.

Since log-sigmoid loss is convex and differentiable, we can obtain its none-class ranking form.

Model	Dev		Test	
	Ign F1	F1	Ign F1	F1
ATLOP+BERT$^*_{Base}$	73.12 ± 0.35	73.93 ± 0.38	72.70 ± 0.23	73.47 ± 0.25
SSR-PU+BERT$_{Base}$	$\mathbf{73.27 \pm 0.19}$	$\mathbf{74.69 \pm 0.20}$	$\mathbf{72.91 \pm 0.23}$	$\mathbf{74.33 \pm 0.20}$
ATLOP+RoBERTa$^*_{Large}$	76.98 ± 0.20	77.68 ± 0.21	76.92 ± 0.15	77.58 ± 0.16
DocuNET+RoBERTa$^\dagger_{Large}$	77.53	78.16	77.27	77.92
KD-DocRE+RoBERTa$^\dagger_{Large}$	**77.92**	78.65	77.63	78.35
SSR-PU+RoBERTa$_{Large}$	77.44 ± 0.25	$\mathbf{78.66 \pm 0.23}$	$\mathbf{77.67 \pm 0.25}$	$\mathbf{78.86 \pm 0.23}$

Table 9: Results on revised Re-DocRED under the fully supervised setting. Results with † are reported from (Tan et al., 2022b). Results with $*$ are based on our implementation.

Model	Ign F1	F1	P	R
ATLOP+BERT$^*_{Base}$	16.99 ± 0.24	17.01 ± 0.24	$\mathbf{93.17 \pm 0.48}$	9.36 ± 0.14
SSR-PU+BERT$_{Base}$	$\mathbf{46.47 \pm 0.21}$	$\mathbf{47.24 \pm 0.23}$	59.52 ± 0.87	$\mathbf{39.18 \pm 0.61}$
ATLOP+RoBERTa$^*_{Large}$	17.29 ± 0.28	17.31 ± 0.28	$\mathbf{94.85 \pm 0.19}$	9.52 ± 0.17
SSR-PU+RoBERTa$_{Large}$	$\mathbf{48.98 \pm 0.30}$	$\mathbf{49.74 \pm 0.30}$	61.57 ± 1.34	$\mathbf{41.75 \pm 0.42}$

Table 10: Results on Re-DocRED revised test set under the extremely unlabeled setting. Results with $*$ are based on our implementation.

Model	Ign F1	F1
SSR-PU$_{margin=0}$	0.18	0.20
SSR-PU$_{margin=0.1}$	55.76	56.81
SSR-PU$_{margin=0.25}$	55.43	56.36
SSR-PU$_{margin=0.5}$	55.27	56.19
SSR-PU$_{margin=1.0}$	54.25	55.24

Table 11: Results on Re-DocRED revised test set under the BERT$_{Base}$ setting with different $margin$.

Model	F1	P	R
SSR-PU$_{\pi_i=2\pi_{labeled,i}}$	55.44	78.98	42.71
SSR-PU$_{\pi_i=3\pi_{labeled,i}}$	56.36	70.53	46.93
SSR-PU$_{\pi_i=4\pi_{labeled,i}}$	54.74	61.45	49.35

Table 12: Results on Re-DocRED revised test set under the BERT$_{Base}$ setting with different π_i estimation.

Log-sigmoid ranking loss:

$$\ell_{LSR}(f_i, y_i) = -log(\sigma(y_i(f_i - f_0))). \quad (20)$$

This ranking loss function remain Bayesian consistent with L_{NA} (Eq.12).

A.6 Sensitivity to Hyper-Parameter *margin*

As shown in Table 11, the model fail to train when $margin = 0$, and the model is insensitive to $margin$ when $margin \neq 0$. This is consistent with our proof.

A.7 Influence of Prior Estimation

As shown in Table 12, the experimental results with different π_i show that our method is insensitive to the estimation of π_i. Smaller estimates of π_i lead to higher precision rates as well as lower recall rates, while the opposite is true for higher estimates of π_i.