

A Positive-Unlabeled Metric Learning Framework for Document-Level Relation Extraction with Incomplete Labeling

Ye Wang¹, Huazheng Pan¹, Tao Zhang², Wen Wu¹, Wenxin Hu^{1*}

¹East China Normal University, Shanghai, China

²Tsinghua University, Beijing, China

{yewang, hzpan}@stu.ecnu.edu.cn, tao-zhan20@mails.tsinghua.edu.cn

wwu@cs.ecnu.edu.cn, wxhu@cc.ecnu.edu.cn

Abstract

The goal of document-level relation extraction (RE) is to identify relations between entities that span multiple sentences. Recently, incomplete labeling in document-level RE has received increasing attention, and some studies have used methods such as positive-unlabeled learning to tackle this issue, but there is still a lot of room for improvement. Motivated by this, we propose a positive-augmentation and positive-mixup positive-unlabeled metric learning framework (P³M). Specifically, we formulate document-level RE as a metric learning problem. We aim to pull the distance closer between entity pair embedding and their corresponding relation embedding, while pushing it farther away from the none-class relation embedding. Additionally, we adapt the positive-unlabeled learning to this loss objective. In order to improve the generalizability of the model, we use dropout to augment positive samples and propose a positive-none-class mixup method. Extensive experiments show that P³M improves the F1 score by approximately 4-10 points in document-level RE with incomplete labeling, and achieves state-of-the-art results in fully labeled scenarios. Furthermore, P³M has also demonstrated robustness to prior estimation bias in incomplete labeled scenarios.

Introduction

Relation extraction (RE) involves identifying the relations between two entities in a given text, which is a fundamental task in information extraction. In the past, most RE research focused on extracting relations within a single sentence (Miwa and Bansal 2016; Zhang, Qi, and Manning 2018). However, more recent work has begun to examine document-level RE, which involves identifying relations between entities across multiple sentences in a document (Yao et al. 2019; Zhou et al. 2021; Xu et al. 2022; Yu, Yang, and Tian 2022; Zhou and Lee 2022).

Previously, document-level RE focused on fully supervised scenarios. However, due to the fact that the number of entity pairs is related to the number of entities in a quadratic way, it is very difficult to fully annotate all the relations in a document. This has made the problem of incomplete labeling a common problem in document-level RE and has attracted increasing attention from researchers. (Huang et al.

2022) noticed that the popular document-level RE dataset DocRED (Yao et al. 2019) annotated using the *recommend-revise* scheme contains a large number of unlabeled positive relations, i.e. false negatives. (Tan et al. 2022b) obtained a high-quality Re-DocRED dataset by supplementing the large number of missing relations in DocRED. (Wang et al. 2022) was the first to use positive-unlabeled (PU) learning, a method of learning risk estimators from positive and unlabeled data, to solve the document-level RE task with incomplete labeling and provided a powerful baseline. Despite this, they still suffer greatly from distribution bias caused by incomplete annotation of positive samples, and lack of generalization in the model.

Inspired by these studies, we propose a positive-augmentation and positive-mixup positive-unlabeled metric learning framework (P³M). Firstly, for metric learning in document-level RE, we initialize an embedding for each relation and an embedding for the none-class relation. During training, we pull the entity pair embedding closer to the corresponding relation embedding and push it away from the none-class relation embedding. We adapt this goal to the positive-unlabeled learning paradigm. Then, due to the fact that the labeled positive samples are a subset of the overall positive samples, the distribution of the labeled positive samples cannot approximate the true positive sample distribution, especially in extreme cases of incomplete labeling. To alleviate this problem, inspired by (Gao, Yao, and Chen 2021), we use the dropout noise (Srivastava et al. 2014) inherent in the model to augment the positive samples and experimentally verify the effectiveness of this augmentation.

Finally, to further enhance the model’s generalization, we use mixup to interpolate between the embeddings of positive and negative samples. Under the positive-unlabeled setting, it is not possible to obtain true negative samples, i.e. unlabeled entity pairs may still have some relations. Directly interpolating between the two would introduce noise. Thanks to the metric learning framework, which puts the embeddings of none-class relation and none-class entity pairs in the same feature space, we can use the embedding of none-class relation as pseudo-negative entity pair embedding.

We conduct experiments on the DocRED (Yao et al. 2019) dataset under incomplete labeling and extreme incomplete labeling settings, as well as the ChemDisGene (Zhang et al. 2022) dataset in the biomedical domain. We improve

*Corresponding author.

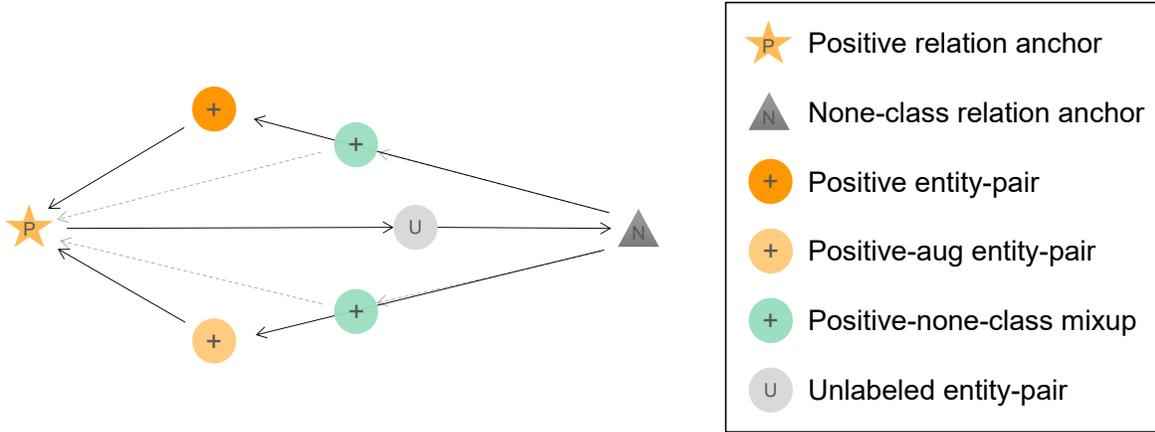


Figure 1: In the dense representation space for a specific positive relation, the P³M framework brings the positive sample (orange circle) and its augmented embedding (light orange circle) closer to the positive relation embedding (yellow pentagram), while distancing them from the none-class relation embedding (grey triangle). The unlabeled sample (light grey circle) is distanced from the positive relation and brought closer to the none-class relation. To address scarcity of positive samples, extra positive sample embeddings (light green circles) are obtained using mixup, partially aligning them with the positive relation and distancing them from the none-class relation.

the F1 score by about 4-10 points compared to the baseline, demonstrating the effectiveness of our proposed P³M method. We also conduct experiments in the fully labeled scenario and achieved the best results. Finally, experiments under different estimated priors demonstrate the robustness of our method to prior estimation bias, which is greatly beneficial for the application of P³M in real-world scenarios. The contributions of this paper can be summarized as follows:¹

- We propose a positive-unlabeled metric learning framework that adapts the metric learning objective to the positive-unlabeled learning paradigm in document-level RE.
- We use the dropout noise inherent in the model to augment the positive samples, expanding the distribution of the positive samples.
- We use mixup to interpolate between the embeddings of positive entity pairs and none-class relation, further enhancing the model’s generalization.
- Experiments show that our method achieves the state-of-the-art results in various incomplete labeling settings and in fully labeled scenario, as well as robustness to prior estimation bias.

Methodology

In this section, we introduce the details of P³M. Firstly, we propose positive-unlabeled metric learning for document-level RE. Then, we introduce an augmentation method for positive samples based on dropout. Finally, we propose positive-none-class mixup to further enhance the model’s generalization. The overall architecture of P³M is shown in Figure 1.

¹Code is available at <https://github.com/www-Ye/P3M>

Positive-Unlabeled Metric Learning for Document-Level RE

Document-level RE can be viewed as a multi-label classification task, and there are a large number of entity pairs with no relation. Previous work (Zhou and Lee 2022; Wang et al. 2022) has shown that setting an additional none-class relation can be very helpful for performance. Therefore, in our method, we transform document-level RE with none-class relation into a proxy-based metric learning task, setting an anchor for each positive relation and none-class relation, respectively.

Let \mathcal{X} be an instance space and $\mathcal{Y} = \{-1, +1\}^K$ be a label space, where K is the number of pre-defined classes. An instance $x \in \mathcal{X}$ is associated with a subset of labels, identified by a binary vector $y \in \mathcal{Y} = (y_1, \dots, y_K)$, where $y_i = +1$ if the i -th label is positive for x , and $y_i = -1$ otherwise. We define each relation embedding as $c \in \mathcal{C} = (c_0, c_1, \dots, c_K)$, where c_0 is the none-class relation embedding, and the rest are predefined relation embeddings. The goal is to learn an embedding $f : \mathcal{X} \rightarrow \mathbb{R}^d$ that brings it closer to its corresponding relation embedding c_i and push it further away from the none-class relation embedding c_0 .

For simplicity, we use the SoftMax_{norm} proposed by (Qian et al. 2019) as the metric learning loss function, which can be seen as a smoothed version of triplet loss. For a given entity pair and a given relation, the loss function can be expressed as:

$$\ell_{SoftMax_{norm}}(f(x), c_i, c_0) = -\log \frac{\exp(\lambda c_i^\top f(x))}{\exp(\lambda c_i^\top f(x)) + \exp(\lambda c_0^\top f(x))}, \tag{1}$$

where $f(x)$, c_i , c_0 need to be normalized, and λ is a scaling factor.

In the inference stage, for any entity pair, the relation i exists if $\mathbf{c}_i^\top f(\mathbf{x}) > \mathbf{c}_0^\top f(\mathbf{x})$ and vice versa. In the following part we use ℓ instead of $\ell_{SoftMaxNorm}$ as an abbreviation.

For i -th class, assume that the data follow an unknown probability distribution with density $p(\mathbf{x}, y_i)$, $p_{P_i} = p(\mathbf{x} | y_i = +1)$ as the positive marginal, $p_{N_i} = p(\mathbf{x} | y_i = -1)$ as the negative marginal, and $p_i(\mathbf{x})$ as the marginal. In positive-negative metric learning (PNM), the ideal loss to optimize would be:

$$L_{PNM} = \sum_{i=1}^K (\pi_i \mathbb{E}_{P_i} [\ell(f(\mathbf{x}), \mathbf{c}_i, \mathbf{c}_0)] + (1 - \pi_i) \mathbb{E}_{N_i} [\ell(f(\mathbf{x}), \mathbf{c}_0, \mathbf{c}_i)]), \quad (2)$$

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. $\mathbb{E}_{P_i}[\cdot] = \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x}|y_i=+1)}[\cdot]$, $\mathbb{E}_{N_i}[\cdot] = \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x}|y_i=-1)}[\cdot]$.

In positive-unlabeled metric learning (PUM), due to the absence of negative samples, we cannot estimate $\mathbb{E}_{N_i}[\cdot]$ from the data. Following (du Plessis, Niu, and Sugiyama 2014), PU learning assumes that unlabeled data can reflect the true overall distribution, that is, $p_{U_i}(\mathbf{x}) = p_i(\mathbf{x})$. The expected loss formulation can be defined as:

$$L_{PUM} = \sum_{i=1}^K (\pi_i \mathbb{E}_{P_i} [\ell(f(\mathbf{x}), \mathbf{c}_i, \mathbf{c}_0)] + \mathbb{E}_{U_i} [\ell(f(\mathbf{x}), \mathbf{c}_0, \mathbf{c}_i)] - \pi_i \mathbb{E}_{P_i} [\ell(f(\mathbf{x}), \mathbf{c}_0, \mathbf{c}_i)]), \quad (3)$$

here $\mathbb{E}_{U_i}[\cdot] = \mathbb{E}_{\mathbf{x} \sim p_i(\mathbf{x})}[\cdot]$ and $\mathbb{E}_{U_i}[\ell(f(\mathbf{x}), \mathbf{c}_0, \mathbf{c}_i)] - \pi_i \mathbb{E}_{P_i}[\ell(f(\mathbf{x}), \mathbf{c}_0, \mathbf{c}_i)]$ can alternatively represent $(1 - \pi_i) \mathbb{E}_{N_i}[\ell(f(\mathbf{x}), \mathbf{c}_0, \mathbf{c}_i)]$ because $p_i(\mathbf{x}) = \pi_i p_{P_i}(\mathbf{x}) + (1 - \pi_i) p_{N_i}(\mathbf{x})$.

Since there are already some labeled relations in the document-level RE dataset, this leads to prior shift in the unlabeled data. We also use the method of prior shift in the training data to obtain the final positive-unlabeled metric learning (PM) expected loss:

$$L_{PM} = \sum_{i=1}^K (\pi_i \mathbb{E}_{P_i} [\ell(f(\mathbf{x}), \mathbf{c}_i, \mathbf{c}_0)] + \frac{1 - \pi_i}{1 - \pi_{u,i}} \mathbb{E}_{U_i} [\ell(f(\mathbf{x}), \mathbf{c}_0, \mathbf{c}_i)] - \frac{\pi_{u,i} - \pi_{u,i} \pi_i}{1 - \pi_{u,i}} \mathbb{E}_{P_i} [\ell(f(\mathbf{x}), \mathbf{c}_0, \mathbf{c}_i)]), \quad (4)$$

where $\pi_{u,i} = p(y_i = 1 | s_i = -1) = \frac{\pi_i - \pi_{labeled,i}}{1 - \pi_{labeled,i}}$, $\pi_{labeled,i} = p(s_i = +1)$ and $(1 - \pi_{labeled,i}) = (1 - p(s_i = +1)) = p(s_i = -1)$. $s_i = +1$ or $s_i = -1$ mean that the i -th class is labeled or unlabeled, respectively. For details on prior shift in document-level RE, please refer to (Wang et al. 2022).

As a result, by rewriting Eq.4 in the form of data approximation and applying non-negative risk estimation (Kiryo et al. 2017) to the PM framework to address the overfitting problem caused by the complexity of the model, we can ob-

tain:

$$\begin{aligned} \widehat{L}_{PM} = & \sum_{i=1}^K \left(\frac{1}{n_{P_i}} \pi_i \sum_{j=1}^{n_{P_i}} \ell(f(\mathbf{x}_j^{P_i}), \mathbf{c}_i, \mathbf{c}_0) \right. \\ & + \max(0, \left[\frac{1}{n_{U_i}} \frac{1 - \pi_i}{1 - \pi_{u,i}} \sum_{j=1}^{n_{U_i}} \ell(f(\mathbf{x}_j^{U_i}), \mathbf{c}_0, \mathbf{c}_i) \right. \\ & \left. \left. - \frac{1}{n_{P_i}} \frac{\pi_{u,i} - \pi_{u,i} \pi_i}{1 - \pi_{u,i}} \sum_{j=1}^{n_{P_i}} \ell(f(\mathbf{x}_j^{P_i}), \mathbf{c}_0, \mathbf{c}_i) \right] \right), \end{aligned} \quad (5)$$

where $\mathbf{x}_j^{P_i}$ and $\mathbf{x}_j^{U_i}$ denote cases that the j -th sample of class i is positive or unlabeled. n_{P_i} and n_{U_i} are the number of positive and unlabeled samples of class i , respectively. In addition, in order to address the severe class imbalance problem, we multiply the first term in Eq.5 by $\gamma_i = (\frac{1 - \pi_i}{\pi_i})^{0.5}$ as the class weight.

Positive Augmentation Based on Dropout Noise

It is important to note that since the labeled sample is only a portion of the true positive sample, meaning the distribution is biased, $p(\mathbf{x} | y_i = 1)$ is not equal to $p(\mathbf{x} | s_i = 1)$. This means that the first term in Eq.5 is a biased approximation of the first term in Eq.4. To alleviate this issue, we can use data augmentation to expand the distribution of positive samples. Here, inspired by (Gao, Yao, and Chen 2021), we use the model's own dropout perturbation to augment the positive samples to get \mathbf{x}' , and the perturbed entity pair embedding is $f(\mathbf{x}')$. However, since we have only augmented the positive samples, the prior $\pi_{u,i}$ in the unlabeled data does not change, and we can obtain the positive-unlabeled metric learning (P²M) objective loss:

$$\begin{aligned} L_{P^2M} = & \sum_{i=1}^K (\pi_i \mathbb{E}_{P_{new,i}} [\ell(f(\mathbf{x}), \mathbf{c}_i, \mathbf{c}_0)] \\ & + \frac{1 - \pi_i}{1 - \pi_{u,i}} \mathbb{E}_{U_i} [\ell(f(\mathbf{x}), \mathbf{c}_0, \mathbf{c}_i)] \\ & - \frac{\pi_{u,i} - \pi_{u,i} \pi_i}{1 - \pi_{u,i}} \mathbb{E}_{P_{new,i}} [\ell(f(\mathbf{x}), \mathbf{c}_0, \mathbf{c}_i)]), \end{aligned} \quad (6)$$

here $\mathbb{E}_{P_{new,i}}[\cdot] = \mathbb{E}_{\mathbf{x}, \mathbf{x}' \sim p_i(\mathbf{x}, \mathbf{x}' | y_i = +1)}[\cdot]$. It can be written in the non-negative form of data approximation as:

$$\begin{aligned} \widehat{L}_{P^2M} = & \sum_{i=1}^K \left(\frac{1}{2n_{P_i}} \pi_i \left(\sum_{j=1}^{n_{P_i}} \ell(f(\mathbf{x}_j^{P_i}), \mathbf{c}_i, \mathbf{c}_0) \right. \right. \\ & \left. \left. + \sum_{j=1}^{n_{P_i}} \ell(f(\mathbf{x}'_j^{P_i}), \mathbf{c}_i, \mathbf{c}_0) \right) \right. \\ & + \max(0, \left[\frac{1}{n_{U_i}} \frac{1 - \pi_i}{1 - \pi_{u,i}} \sum_{j=1}^{n_{U_i}} \ell(f(\mathbf{x}_j^{U_i}), \mathbf{c}_0, \mathbf{c}_i) \right. \\ & \left. - \frac{1}{2n_{P_i}} \frac{\pi_{u,i} - \pi_{u,i} \pi_i}{1 - \pi_{u,i}} \left(\sum_{j=1}^{n_{P_i}} \ell(f(\mathbf{x}_j^{P_i}), \mathbf{c}_0, \mathbf{c}_i) \right. \right. \\ & \left. \left. + \sum_{j=1}^{n_{P_i}} \ell(f(\mathbf{x}'_j^{P_i}), \mathbf{c}_0, \mathbf{c}_i) \right) \right] \right). \end{aligned} \quad (7)$$

We will compare the difference between augmenting positive samples and augmenting all samples in the main results subsection of the experiments.

Positive-None-Class Mixup

In order to further enhance the generalization of the model, our goal is to obtain more diverse entity pair embeddings. To achieve this, we can interpolate between positive and negative samples of each class to obtain mixed entity pair representations:

$$f_{mix(ori)}(\mathbf{x}) = \mu f(\mathbf{x}) + (1 - \mu)f(\mathbf{x}_-), \quad (8)$$

here $\mathbf{x} \sim p(\mathbf{x} | y_i = 1)$ and $\mathbf{x}_- \sim p(\mathbf{x} | y_i = -1)$ are the positive and negative samples of class i , respectively. μ is sampled from a Beta(α, α) distribution ($\mu \in [0, 1]$ and $\alpha > 0$). However, in PU learning, we cannot obtain true negative samples, which means that there are some positive entity pairs in the unlabeled samples, resulting in bias when interpolating with unlabeled entity pairs. Thanks to the metric learning framework, which places the relation embedding and the entity pair embedding in the same feature space, we can use the none-class relation embedding \mathbf{c}_0 to stand in for pseudo-negative entity pairs. Therefore, Eq.8 can be rewritten as:

$$f_{mix}(\mathbf{x}) = \mu f(\mathbf{x}) + (1 - \mu)\mathbf{c}_0. \quad (9)$$

We will compare the difference between using mixup with none-class relation embedding and the original method in the main results subsection of the experiments. According to this formulation, the mixup loss function can be reformulated as:

$$L_{p-mix} = \sum_{i=1}^K (\mu \mathbb{E}_{P_{new,i}} [\ell(f_{mix}(\mathbf{x}), \mathbf{c}_i, \mathbf{c}_0)] + (1 - \mu) \mathbb{E}_{P_{new,i}} [\ell(f_{mix}(\mathbf{x}), \mathbf{c}_0, \mathbf{c}_i)]), \quad (10)$$

here we perform mixup on both the original positive samples and the augmented positive samples. Finally, we rewrite this equation in the form of data approximation:

$$\begin{aligned} \hat{L}_{p-mix} &= \sum_{i=1}^K \left(\frac{\mu}{2n_{P_i}} \left(\sum_{j=1}^{n_{P_i}} \ell(f_{mix}(\mathbf{x}_j^{P_i}), \mathbf{c}_i, \mathbf{c}_0) \right. \right. \\ &\quad \left. \left. + \sum_{j=1}^{n_{P_i}} \ell(f_{mix}(\mathbf{x}'_j^{P_i}), \mathbf{c}_i, \mathbf{c}_0) \right) \right. \\ &\quad \left. + \frac{(1 - \mu)}{2n_{P_i}} \left(\sum_{j=1}^{n_{P_i}} \ell(f_{mix}(\mathbf{x}_j^{P_i}), \mathbf{c}_0, \mathbf{c}_i) \right) \right. \\ &\quad \left. + \sum_{j=1}^{n_{P_i}} \ell(f_{mix}(\mathbf{x}'_j^{P_i}), \mathbf{c}_0, \mathbf{c}_i) \right). \end{aligned} \quad (11)$$

Therefore, the final loss of our positive-augmentation and positive-mixup positive-unlabeled metric learning (P^3M) framework is:

$$\hat{L}_{P^3M} = \hat{L}_{P^2M} + \nu \hat{L}_{p-mix}, \quad (12)$$

where ν is a hyperparameter that controls the strength of positive-mixup.

Experiments

In this section, we evaluate the performance of P^3M in various incompletely labeled document-level RE datasets and settings as well as in the fully labeled scenario. We also analyze the effectiveness of different components of the method.

Experimental Setups

Datasets. *DocRED* (Yao et al. 2019) is a large-scale document-level RE dataset constructed from Wikipedia, containing 96 predefined relations. However, the original dataset contains a large amount of incomplete labeling phenomena, (Tan et al. 2022b) proposed a high-quality revised version Re-DocRED. In our experiments, we use the incompletely labeled DocRED original training set and the fully labeled Re-DocRED test set. In order to further analyze the performance of the method in incompletely labeled scenarios, we also use the extreme incompletely labeled training set DocRED_ext constructed by (Wang et al. 2022) for experiments. *ChemDisGene* (Zhang et al. 2022) is a multi-label document-level RE dataset in the biomedical field. We use the incompletely labeled training set constructed by distantly supervised of CTD database (Davis et al. 2021) and the fully labeled *All relationships* test set constructed by additional annotation by domain experts for experiments. The datasets are in English and used for their intended purpose. The detailed statistics of the datasets are shown in Table 1. The average number of relations in the incompletely labeled training sets, especially the extremely incompletely labeled sets, is far less than that in the test sets, indicating the large number of false negatives in the training sets.

Implementation Details. In our experiments, we use AT-LOP (Zhou et al. 2021) as the encoding model for relation representation learning. Further, we apply cased BERT_{Base} (Devlin et al. 2019) and RoBERTa_{Large} (Liu et al. 2019) for DocRED and PubmedBert (Gu et al. 2022) 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 to 3e-5. 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. In our experiments, we fixed $\lambda = 10$, $\alpha = 1.0$, $\nu = 0.05$, and the dropout rate to 0.2. In order to make a fair comparison, we use the same prior estimation as in (Wang et al. 2022), setting $\pi_i = 3\pi_{labeled,i}$ and for extreme incomplete labeling, setting $\pi_i = 12\pi_{labeled,i}$. The training stopping criteria are set as follows: 10 epochs for both two dataset. We do not use any fully labeled validation or test sets in any stage of the training process and report the results of the final model by running five times with different random seeds (62, 63, 64, 65, 66). All experiments are conducted with 1 Tesla A100 or 1 Tesla V100 GPU.

Baseline. For DocRED, we use fully supervised models BiLSTM (Yao et al. 2019), GAIN (Zeng et al. 2020),

Dataset	DocRED	DocRED_ext	Re-DocRED		ChemDisGene	
	train	train	train	test	train	test
# docs	3,053	3,053	3,053	500	76,942	523
# rels	96	96	96		14	
Avg # ents	19.5	19.5	19.4	19.6	7.5	10.0
Avg # rels	12.5	5.4	28.1	34.9	2.1	7.2

Table 1: Statistics of document-level RE datasets.

Model	DocRED				DocRED_ext			
	Ign F1	F1	P	R	Ign F1	F1	P	R
BiLSTM [†]	32.57	32.86	77.04	20.89	—	—	—	—
GAIN+BERT [†] _{Base}	45.57	45.82	88.11	30.98	—	—	—	—
DocuNET+RoBERTa [†] _{Large}	45.88	45.99	94.16	30.42	—	—	—	—
ATLOP+BERT [†] _{Base}	43.12	43.25	92.49	28.23	16.99	17.01	93.17	9.36
SSR-PU+ATLOP+BERT [†] _{Base}	55.21	56.14	70.42	46.67	46.47	47.24	59.52	39.18
PM+ATLOP+BERT [†] _{Base}	57.97	59.34	60.76	58.01	53.84	54.85	54.91	54.81
P ² M(all)+ATLOP+BERT [†] _{Base}	58.27	59.54	63.31	56.19	53.77	54.71	56.81	52.78
P ² M+ATLOP+BERT [†] _{Base}	58.85	60.08	64.30	56.40	54.64	55.57	58.10	53.26
P ³ M(ori)+ATLOP+BERT [†] _{Base}	59.48	60.79	62.53	59.14	55.91	56.82	59.13	54.70
P ³ M+ATLOP+BERT [†] _{Base}	59.81	61.03	64.57	57.87	56.17	57.02	61.12	53.44
ATLOP+RoBERTa [†] _{Large}	45.09	45.19	94.75	29.67	17.29	17.31	94.85	9.52
SSR-PU+ATLOP+RoBERTa [†] _{Large}	58.68	59.50	74.21	49.67	48.98	49.74	61.57	41.75
PM+ATLOP+RoBERTa [†] _{Large}	60.72	62.13	61.15	63.16	56.67	57.72	54.60	61.27
P ² M(all)+ATLOP+RoBERTa [†] _{Large}	61.17	62.46	64.19	60.83	56.49	57.46	57.42	57.51
P ² M+ATLOP+RoBERTa [†] _{Large}	61.55	62.82	65.19	60.62	57.27	58.21	58.91	57.55
P ³ M(ori)+ATLOP+RoBERTa [†] _{Large}	62.64	63.96	64.15	63.80	58.48	59.42	59.53	59.33
P ³ M+ATLOP+RoBERTa [†] _{Large}	63.16	64.34	67.43	61.52	59.02	59.86	63.04	57.01

Table 2: Results on Re-DocRED revised test set. Results with † are reported from (Wang et al. 2022).

DocuNET (Zhang et al. 2021), and ATLOP (Zhou et al. 2021), as well as the positive-unlabeled learning method SSR-PU (Wang et al. 2022) as the baseline models. For ChemDisGene, we use BRAN (Verga, Strubell, and McCallum 2018), PubmedBert (Gu et al. 2022), PubmedBert+BRAN (Zhang et al. 2022), ATLOP, and SSR-PU as the baselines.

Evaluation Metric. We use micro F1 (F1), micro ignore F1 (Ign F1), precision (P), and recall (R) to evaluate the overall performance of models on DocRED. Ign F1 denotes the F1 score excluding the relations shared by the training and test set. We use micro F1 (F1), precision (P), and recall (R) to evaluate the models on ChemDisGene.

Main Results

In this subsection, we compare the results between PM, P²M(all), P²M, P³M(ori), and P³M. P²M(all) refers to augmenting all samples, while P³M(ori) refers to the original mixup method that uses unlabeled samples for mixing.

Results on DocRED. As shown in Table 2, traditional supervised learning methods such as BiLSTM, GAIN, DocuNET, and ATLOP have a dramatic decline in performance,

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	76.17	29.70
SSR-PU+PubmedBert [†]	48.56	54.27	43.93
PM+PubmedBert	52.02	58.26	47.00
P ² M(all)+PubmedBert	51.29	57.54	46.27
P ² M+PubmedBert	52.19	59.02	46.78
P ³ M(ori)+PubmedBert	53.58	59.44	48.78
P ³ M+PubmedBert	53.62	60.20	48.34

Table 3: Results on ChemDisGene *All relationships* test set. Results with † are reported from (Wang et al. 2022). Results with ‡ are reported from (Zhang et al. 2022).

especially in recall, in the incompletely labeled scenario. The SSR-PU method, which uses PU learning, effectively alleviates this problem and achieves a huge improvement on the basis of the ATLOP encoder model. Our framework

Model	Ign F1	F1
ATLOP+BERT [†] _{Base}	72.70	73.47
SSR-PU+BERT [†] _{Base}	72.91	74.33
P ³ M+BERT _{Base}	74.10	75.60
ATLOP+RoBERTa [†] _{Large}	76.92	77.58
DocuNET+RoBERTa [‡] _{Large}	77.26	77.87
KD-DocRE+RoBERTa [‡] _{Large}	77.60	78.28
SSR-PU+RoBERTa [†] _{Large}	77.67	78.86
P ³ M+RoBERTa _{Large}	78.82	80.02

Table 4: Results on Re-DocRED revised test set under the fully supervised setting. Results with [†] are reported from (Wang et al. 2022). Results with [‡] are reported from (Tan et al. 2022b).

P³M, on the other hand, improves the F1 score by 4.89 and 4.84, respectively, for the BERT_{Base} and RoBERTa_{Large} settings, compared to SSR-PU in the incompletely labeled scenario. And when using extremely incompletely labeled training sets, the two settings respectively improve the F1 score by 9.78 and 10.12. The outstanding improvement shows the effectiveness of our proposed framework.

In the DocRED experiment, P³M shows a slight precision drop compared to SSR-PU, possibly due to expanded positive sample distribution causing errors in ambiguous case classification. However, recall increases of 11.20 and 11.85 under the BERT_{Base} and RoBERTa_{Large} settings justify this trade-off. This classification error, likely stemming from the base model’s limitations, can be mitigated by enhancing the base model. In DocRED_ext, our method not only improves recall by 14.26 and 15.26 over SSR-PU under the same settings but also raises precision by 1.60 and 1.47, respectively, highlighting its value in label-scarce scenarios.

We compare different variations of our method. P²M(all) and P²M use dropout to augment samples. P²M(all) augments all samples, and there is a slight improvement when using the DocRED training set compared to the basic PM framework. However, in extreme scenarios of incomplete labeling, performance deteriorates. We believe this is caused by the fact that since the data itself is positive and unlabeled, augmenting all samples instead introduces some additional noise. P²M, which only augments positive samples, does not have this problem. The increase in the distribution of positive samples further improves the performance of the model and to some extent relieves the distribution bias caused by incompletely labeled positive samples. The regular P³M(ori) method has a considerable improvement over P²M because in document-level RE, the number of negative samples is far greater than that of positive samples. Therefore, direct sampling of unlabeled samples will only introduce a small number of false negatives, but there will still be bias. P³M has more performance improvement compared to P³M(ori), which shows that using none-class relation embedding as pseudo-negative samples effectively mitigates the bias of directly using unlabeled samples for mixup.

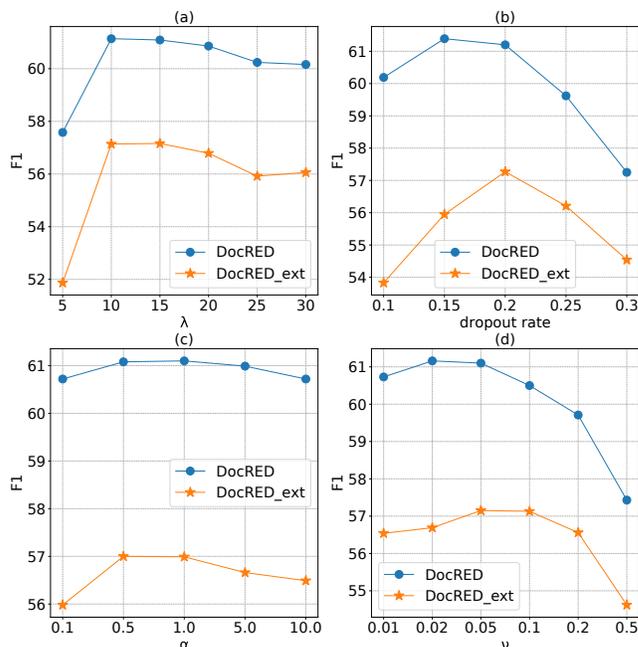


Figure 2: Effect of hyperparameters on DocRED

Results on ChemDisGene. As shown in Table 3, similar to the results on DocRED, the performance of supervised learning methods has a large decline, while SSR-PU has a large improvement compared to them. Our proposed P³M improved by 5.06 F1 score compared to SSR-PU and achieved a new best result. Notably, the *All relationships* test set of ChemDisGene is sourced from another corpus DrugProt (Miranda et al. 2021) and is additionally annotated by human experts, making the test set have a larger deviation from the training set. However, our proposed framework has better robustness under this deviation.

For different variations of the method, P²M(all) has an obvious performance decrease compared to PM, which may be caused by the larger deviation between the training set and the test set, and this deviation is further amplified by the augmentation of all samples. P²M shows further improvement compared to PM, indicating the help of dropout augmentation in improving the diversity of positive sample distribution. P³M(ori) and P³M, which added positive-mixup method, both have larger improvements, verifying the help of mixup in improving the generalization of positive-unlabeled metric learning framework. Due to the scarcity of positive samples, sampling n_{P_i} as negative samples from the unlabeled samples only causes a small bias, making P³M(ori) still able to achieve a good result and the performance gap between P³M and P³M(ori) is smaller.

Additional Analysis

Fully Supervised Setting. We conduct experiments on Re-DocRED (Tan et al. 2022b) under a fully supervised setting. In the experiments, we set $\pi_i = \pi_{labeled,i}$, $\nu = 0.01$, with a dropout rate of 0.1, and other hyperparameters remained unchanged. As shown in Table 4, we compared our

Model	Ign F1	F1	P	R
$P^3M_{\pi_i=\pi_{labeled,i}}$	59.84	60.77	70.99	53.12
$P^3M_{\pi_i=2\pi_{labeled,i}}$	60.00	61.06	68.06	55.38
$P^3M_{\pi_i=3\pi_{labeled,i}}$	59.92	61.13	65.01	57.69
$P^3M_{\pi_i=4\pi_{labeled,i}}$	59.06	60.43	61.08	59.78
$P^3M_{\pi_i=5\pi_{labeled,i}}$	57.98	59.48	57.75	61.31

Table 5: Results on Re-DocRED revised test set under the BERT_{Base} setting with different π_i estimation.

framework with the existing state-of-the-art methods AT-LOP (Zhou et al. 2021), DocuNET (Zhang et al. 2021), KD-DocRE (Tan et al. 2022a), and SSR-PU (Wang et al. 2022). Our framework achieves the best results as well.

Effect of Hyperparameters. Figure 2 shows the effect of hyperparameters on the model under the DocRED and DocRED_ext incomplete labeling settings. (a) shows the effect of the scaling factor λ , with similar trends under both settings, and $\lambda = 10$ being the best choice. (b) shows the effect of the dropout rate, indicating that the greater the degree of incompleteness in labeling, the greater the dropout rate needed to enhance diversity, but too large a dropout rate will also introduce more noise. (c) shows the effect of α , indicating that the model is not sensitive to the choice of α , and $\alpha = 1.0$ can be seen as a uniform mixup interpolation between distributions. (d) shows the effect of ν , with similar trends under both settings, and more severe incomplete labeling requires slightly larger mixup strength.

Effect of Prior Estimation. Table 5 shows the effect of different prior estimates on the model. It can be seen that our framework is not sensitive to errors in prior estimates, especially in cases where the prior estimate is too small. Even when $\pi_i = \pi_{labeled,i}$, the model still performs well, demonstrating the robustness of our method under errors in prior estimates, which is very helpful for real-world applications.

Related Work

Document-Level Relation Extraction. Previously, effective methods for document-level relation extraction (RE) have mainly been graph-based models and transformer-based models. Graph-based models (Nan et al. 2020; Li et al. 2020; Zeng et al. 2020; Zeng, Wu, and Chang 2021; Xu, Chen, and Zhao 2021) use graph neural networks to gather entity information for relational inference, while transformer-based methods (Zhou et al. 2021; Xu et al. 2021; Zhang et al. 2021; Tan et al. 2022a) capture long-range dependencies implicitly. Recently, it has been found that there are a large number of false negatives in document-level RE datasets, i.e. incomplete labels (Huang et al. 2022; Tan et al. 2022b). (Wang et al. 2022) proposed using positive-unlabeled learning to address this problem.

Positive-Unlabeled Learning. Positive-unlabeled (PU) learning (Elkan and Noto 2008; du Plessis, Niu, and Sugiyama 2014, 2015; Kiryo et al. 2017; Garg et al. 2021), as an emerging weakly supervised learning paradigm, aims

to learn classifiers from positive and unlabeled data, and has gained continuous attention from researchers. PU learning has been widely applied in various tasks, such as text classification (Li and Liu 2003), sentence embedding (Cao et al. 2021), named entity recognition (Peng et al. 2019; Zhou, Li, and Li 2022), knowledge graph completion (Tang et al. 2022), and sentence-level RE (He et al. 2020) in the NLP field. (Chuang et al. 2020) used PU learning to address the issue of negative samples potentially carrying the same label in contrastive learning.

Deep Metric Learning. Our work is inspired by metric learning and mainly falls into two categories: pair-based losses and proxy-based losses. Pair-based methods (Hadsell, Chopra, and LeCun 2006; Schroff, Kalenichenko, and Philbin 2015; Sohn 2016; Wang et al. 2019) focus on the relationships between individual samples, and contrastive learning can be considered a subset of this approach. Proxy-based methods like Proxy-NCA (Movshovitz-Attias et al. 2017) and NormFace (Wang et al. 2017) consider the relationships between proxies and samples, and (Qian et al. 2019) unified the relationship between SoftMax loss and triplet loss, and proposed a new SoftTriplet loss. Proxy-based methods are a type of approach that focuses on improving generalization while keeping training complexity low, although they may not fully utilize the relationships between individual samples.

Data Augmentation. Data augmentation is a key factor in deep learning performance and is widely used in many fields (Shorten and Khoshgoftaar 2019; Hedderich et al. 2021). (Wei and Zou 2019; Ma 2019) proposed to augment words by randomly inserting and replacing them, while (Lee et al. 2021) augmented the word embeddings directly. (Gao, Yao, and Chen 2021) used simple dropout to augment sentence embeddings for unsupervised contrastive learning. Mixup (Zhang et al. 2018; Verma et al. 2019) can be considered as another common data augmentation method, where interpolation is used to improve the generalization performance of the model between two samples. It is increasingly used and researched in the NLP (Chen, Yang, and Yang 2020; Yin et al. 2021; Wu et al. 2022) and the PU learning (Chen et al. 2020; Wei et al. 2020; Li et al. 2022; Zhao et al. 2022) fields. (Jeong et al. 2022) proposed a document augmentation dense retrieval framework that uses both methods.

Conclusion and Future Work

To address document-level RE with incomplete labeling, we propose a positive-unlabeled metric learning framework P^3M . First, we combine positive-unlabeled learning with metric learning to learn better representations. Then, we use dropout augmentation to expand the distribution of labeled positive samples. Finally, we use none-class relation embedding as pseudo-negative samples and propose a positive-none-class mixup method to further improve the model’s generalization performance. Experiments demonstrate that our method achieve state-of-the-art results in both incomplete and complete labeling scenarios, as well as robustness to prior estimation bias. In the future, we will explore various metric learning losses and data augmentation methods.

Acknowledgments

This work is funded by National Natural Science Foundation of China (under project No. 62377013). The computation is supported by the ECNU Multifunctional Platform for Innovation (001).

References

- Cao, L.; Larsson, E.; von Ehrenheim, V.; Rocha, D. D. C.; Martin, A.; and Horn, S. 2021. PAUSE: Positive and Annealed Unlabeled Sentence Embedding. In *EMNLP*, 10096–10107.
- Chen, H.; Liu, F.; Wang, Y.; Zhao, L.; and Wu, H. 2020. A Variational Approach for Learning from Positive and Unlabeled Data. In *NeurIPS*.
- Chen, J.; Yang, Z.; and Yang, D. 2020. MixText: Linguistically-Informed Interpolation of Hidden Space for Semi-Supervised Text Classification. In *ACL*, 2147–2157.
- Chuang, C.; Robinson, J.; Lin, Y.; Torralba, A.; and Jegelka, S. 2020. Debiased Contrastive Learning. In *NeurIPS*.
- Davis, A. P.; Grondin, C. J.; Johnson, R. J.; Sciaky, D.; Wieggers, J.; Wieggers, T. C.; and Mattingly, C. J. 2021. Comparative Toxicogenomics Database (CTD): update 2021. *Nucleic Acids Res.*, D1138–D1143.
- Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, 4171–4186.
- du Plessis, M. C.; Niu, G.; and Sugiyama, M. 2014. Analysis of Learning from Positive and Unlabeled Data. In *NeurIPS*, 703–711.
- du Plessis, M. C.; Niu, G.; and Sugiyama, M. 2015. Convex Formulation for Learning from Positive and Unlabeled Data. In *ICML*, volume 37, 1386–1394.
- Elkan, C.; and Noto, K. 2008. Learning classifiers from only positive and unlabeled data. In *SIGKDD*, 213–220.
- Gao, T.; Yao, X.; and Chen, D. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *EMNLP*, 6894–6910.
- Garg, S.; Wu, Y.; Smola, A. J.; Balakrishnan, S.; and Lipton, Z. C. 2021. Mixture Proportion Estimation and PU Learning: A Modern Approach. In *NeurIPS*, 8532–8544.
- Goyal, P.; Dollár, P.; Girshick, R. B.; Noordhuis, P.; Wesolowski, L.; Kyrola, A.; Tulloch, A.; Jia, Y.; and He, K. 2017. Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour. *CoRR*, abs/1706.02677.
- Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; and Poon, H. 2022. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Heal.*, 2:1–2:23.
- Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality Reduction by Learning an Invariant Mapping. In *CVPR*, 1735–1742.
- He, Z.; Chen, W.; Wang, Y.; Zhang, W.; Wang, G.; and Zhang, M. 2020. Improving Neural Relation Extraction with Positive and Unlabeled Learning. In *AAAI*, 7927–7934.
- Hedderich, M. A.; Lange, L.; Adel, H.; Strötgen, J.; and Klakow, D. 2021. A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. In *NAACL-HLT*, 2545–2568.
- Huang, Q.; Hao, S.; Ye, Y.; Zhu, S.; Feng, Y.; and Zhao, D. 2022. Does Recommend-Revise Produce Reliable Annotations? An Analysis on Missing Instances in DocRED. In *ACL*, 6241–6252.
- Jeong, S.; Baek, J.; Cho, S.; Hwang, S. J.; and Park, J. C. 2022. Augmenting Document Representations for Dense Retrieval with Interpolation and Perturbation. In *ACL*, 442–452.
- Kiryo, R.; Niu, G.; du Plessis, M. C.; and Sugiyama, M. 2017. Positive-Unlabeled Learning with Non-Negative Risk Estimator. In *NeurIPS*, 1675–1685.
- Lee, S.; Kang, M.; Lee, J.; and Hwang, S. J. 2021. Learning to Perturb Word Embeddings for Out-of-distribution QA. In *ACL-IJCNLP*, 5583–5595.
- Li, B.; Ye, W.; Sheng, Z.; Xie, R.; Xi, X.; and Zhang, S. 2020. Graph Enhanced Dual Attention Network for Document-Level Relation Extraction. In *COLING*, 1551–1560.
- Li, C.; Li, X.; Feng, L.; and Ouyang, J. 2022. Who Is Your Right Mixup Partner in Positive and Unlabeled Learning. In *ICLR*.
- Li, X.; and Liu, B. 2003. Learning to Classify Texts Using Positive and Unlabeled Data. In *IJCAI*, 587–594.
- Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.
- Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *ICLR*.
- Ma, E. 2019. NLP Augmentation. <https://github.com/makcedward/nlpaug>. Accessed: 2023-12-27.
- Miranda, A.; Mehryary, F.; Luoma, J.; Pysalo, S.; Valencia, A.; and Krallinger, M. 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*.
- Miwa, M.; and Bansal, M. 2016. End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures. In *ACL*.
- Movshovitz-Attias, Y.; Toshev, A.; Leung, T. K.; Ioffe, S.; and Singh, S. 2017. No Fuss Distance Metric Learning Using Proxies. In *ICCV*, 360–368.
- Nan, G.; Guo, Z.; Sekulic, I.; and Lu, W. 2020. Reasoning with Latent Structure Refinement for Document-Level Relation Extraction. In *ACL*, 1546–1557.
- Peng, M.; Xing, X.; Zhang, Q.; Fu, J.; and Huang, X. 2019. Distantly Supervised Named Entity Recognition using Positive-Unlabeled Learning. In *ACL*, 2409–2419.
- Qian, Q.; Shang, L.; Sun, B.; Hu, J.; Tacoma, T.; Li, H.; and Jin, R. 2019. SoftTriple Loss: Deep Metric Learning Without Triplet Sampling. In *ICCV*, 6449–6457.

- Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. FaceNet: A unified embedding for face recognition and clustering. In *CVPR*, 815–823.
- Shorten, C.; and Khoshgoftaar, T. M. 2019. A survey on Image Data Augmentation for Deep Learning. *J. Big Data*, 60.
- Sohn, K. 2016. Improved Deep Metric Learning with Multi-class N-pair Loss Objective. In *NeurIPS*, 1849–1857.
- Srivastava, N.; Hinton, G. E.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 1929–1958.
- Tan, Q.; He, R.; Bing, L.; and Ng, H. T. 2022a. Document-Level Relation Extraction with Adaptive Focal Loss and Knowledge Distillation. In *ACL (Findings)*, 1672–1681.
- Tan, Q.; Xu, L.; Bing, L.; Ng, H. T.; and Aljunied, S. M. 2022b. Revisiting DocRED - Addressing the False Negative Problem in Relation Extraction. In *EMNLP*, 8472–8487.
- Tang, Z.; Pei, S.; Zhang, Z.; Zhu, Y.; Zhuang, F.; Hoehndorf, R.; and Zhang, X. 2022. Positive-Unlabeled Learning with Adversarial Data Augmentation for Knowledge Graph Completion. In *IJCAI*, 2248–2254.
- Verga, P.; Strubell, E.; and McCallum, A. 2018. Simultaneously Self-Attending to All Mentions for Full-Abstract Biological Relation Extraction. In *NAACL-HLT*, 872–884.
- Verma, V.; Lamb, A.; Beckham, C.; Najafi, A.; Mitliagkas, I.; Lopez-Paz, D.; and Bengio, Y. 2019. Manifold Mixup: Better Representations by Interpolating Hidden States. In *ICML*, 6438–6447.
- Wang, F.; Xiang, X.; Cheng, J.; and Yuille, A. L. 2017. NormFace: L_2 Hypersphere Embedding for Face Verification. In *ACM MM*, 1041–1049.
- Wang, X.; Han, X.; Huang, W.; Dong, D.; and Scott, M. R. 2019. Multi-Similarity Loss With General Pair Weighting for Deep Metric Learning. In *CVPR*, 5022–5030.
- Wang, Y.; Liu, X.; Hu, W.; and Zhang, T. 2022. A Unified Positive-Unlabeled Learning Framework for Document-Level Relation Extraction with Different Levels of Labeling. In *EMNLP*, 4123–4135.
- Wei, J. W.; and Zou, K. 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In *EMNLP-IJCNLP*, 6381–6387.
- Wei, T.; Shi, F.; Wang, H.; Tu, W.; and Li, Y. 2020. MixPUL: Consistency-based Augmentation for Positive and Unlabeled Learning. *CoRR*, abs/2004.09388.
- Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Scao, T. L.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. M. 2020. Transformers: State-of-the-Art Natural Language Processing. In *EMNLP-Demos*, 38–45.
- Wu, L.; Xie, P.; Zhou, J.; Zhang, M.; Ma, C.; Xu, G.; and Zhang, M. 2022. Robust Self-Augmentation for Named Entity Recognition with Meta Reweighting. In *NAACL-HLT*, 4049–4060.
- Xu, B.; Wang, Q.; Lyu, Y.; Zhu, Y.; and Mao, Z. 2021. Entity Structure Within and Throughout: Modeling Mention Dependencies for Document-Level Relation Extraction. In *AAAI*, 14149–14157.
- Xu, W.; Chen, K.; Mou, L.; and Zhao, T. 2022. Document-Level Relation Extraction with Sentences Importance Estimation and Focusing. In *NAACL-HLT*, 2920–2929.
- Xu, W.; Chen, K.; and Zhao, T. 2021. Document-Level Relation Extraction with Reconstruction. In *AAAI*, 14167–14175.
- Yao, Y.; Ye, D.; Li, P.; Han, X.; Lin, Y.; Liu, Z.; Liu, Z.; Huang, L.; Zhou, J.; and Sun, M. 2019. DocRED: A Large-Scale Document-Level Relation Extraction Dataset. In *ACL*, 764–777.
- Yin, W.; Wang, H.; Qu, J.; and Xiong, C. 2021. BatchMixup: Improving Training by Interpolating Hidden States of the Entire Mini-batch. In *ACL-IJCNLP (Findings)*, 4908–4912.
- Yu, J.; Yang, D.; and Tian, S. 2022. Relation-Specific Attentions over Entity Mentions for Enhanced Document-Level Relation Extraction. In *NAACL-HLT*, 1523–1529.
- Zeng, S.; Wu, Y.; and Chang, B. 2021. SIRE: Separate Intra- and Inter-sentential Reasoning for Document-level Relation Extraction. In *ACL-IJCNLP (Findings)*, 524–534.
- Zeng, S.; Xu, R.; Chang, B.; and Li, L. 2020. Double Graph Based Reasoning for Document-level Relation Extraction. In *EMNLP*, 1630–1640.
- Zhang, D.; Mohan, S.; Torkar, M.; and McCallum, A. 2022. A Distant Supervision Corpus for Extracting Biomedical Relationships Between Chemicals, Diseases and Genes. In *LREC*, 1073–1082.
- Zhang, H.; Cissé, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond Empirical Risk Minimization. In *ICLR*.
- Zhang, N.; Chen, X.; Xie, X.; Deng, S.; Tan, C.; Chen, M.; Huang, F.; Si, L.; and Chen, H. 2021. Document-level Relation Extraction as Semantic Segmentation. In *IJCAI*, 3999–4006.
- Zhang, Y.; Qi, P.; and Manning, C. D. 2018. Graph Convolution over Pruned Dependency Trees Improves Relation Extraction. In *EMNLP*, 2205–2215.
- Zhao, Y.; Xu, Q.; Jiang, Y.; Wen, P.; and Huang, Q. 2022. Dist-PU: Positive-Unlabeled Learning from a Label Distribution Perspective. In *CVPR*, 14441–14450.
- Zhou, K.; Li, Y.; and Li, Q. 2022. Distantly Supervised Named Entity Recognition via Confidence-Based Multi-Class Positive and Unlabeled Learning. In *ACL*, 7198–7211.
- Zhou, W.; Huang, K.; Ma, T.; and Huang, J. 2021. Document-Level Relation Extraction with Adaptive Thresholding and Localized Context Pooling. In *AAAI*, 14612–14620.
- Zhou, Y.; and Lee, W. S. 2022. None Class Ranking Loss for Document-Level Relation Extraction. In *IJCAI*, 4538–4544.