Class-Adaptive Self-Training for Relation Extraction with Incompletely Annotated Training Data

Qingyu Tan[1, 2] **Lu Xu**[1, 3] **Lidong Bing**[† 1] **Hwee Tou Ng**[2]
[1]DAMO Academy, Alibaba Group
[2]Department of Computer Science, National University of Singapore
[3]Singapore University of Technology and Design
{qingyu.tan,lu.x,l.bing}@alibaba-inc.com
{qtan6,nght}@comp.nus.edu.sg

Abstract

Relation extraction (RE) aims to extract relations from sentences and documents. Existing relation extraction models typically rely on supervised machine learning. However, recent studies showed that many RE datasets are incompletely annotated. This is known as the false negative problem in which valid relations are falsely annotated as *no_relation*. Models trained with such data inevitably make similar mistakes during the inference stage. Self-training has been proven effective in alleviating the false negative problem. However, traditional self-training is vulnerable to confirmation bias and exhibits poor performance in minority classes. To overcome this limitation, we proposed a novel class-adaptive re-sampling self-training framework. Specifically, we re-sampled the pseudo-labels for each class by precision and recall scores. Our re-sampling strategy favored the pseudo-labels of classes with high precision and low recall, which improved the overall recall without significantly compromising precision. We conducted experiments on document-level and biomedical relation extraction datasets, and the results showed that our proposed self-training framework consistently outperforms existing competitive methods on the Re-DocRED and ChemDisgene datasets when the training data are incompletely annotated[1].

1 Introduction

Relation extraction (RE) (Wang et al., 2019; Chia et al., 2022a) is an important yet highly challenging task in the field of information extraction (IE). Compared with other IE tasks, such as named entity recognition (NER) (Xu et al., 2021), semantic role labeling (SRL) (Li et al., 2021), and aspect-based sentiment analysis (ABSA) (Li et al., 2018; Zhang et al., 2021b), RE typically has a significantly larger label space and requires graphical reasoning (Christopoulou et al., 2019). The complexity of the RE task inevitably increases the difficulty and cost of producing high-quality benchmark datasets for this task.

In recent years, several works that specifically focus on revising the annotation strategy and quality of existing RE datasets were conducted (Stoica et al., 2021; Alt et al., 2020; Tan et al., 2022b). For example, the DocRED (Yao et al., 2019) dataset is one of the most popular benchmarks for document-level relation extraction. This dataset is produced by the recommend-revise scheme with machine recommendation and human annotation. However, Huang et al. (2022) and Tan et al. (2022b) pointed out the false negative problem in the DocRED dataset, indicating that over 60% of the relation triples are not annotated. To provide a more reliable evaluation dataset for document-level relation extraction tasks, Huang et al. (2022) re-annotated 96 documents that are selected from the original development set of DocRED. In addition, Tan et al. (2022b) developed the Re-DocRED dataset to provide a high-quality revised version of the development set of DocRED. The Re-DocRED dataset consists of a development set that contains 1,000 documents and a silver-quality training set that contains 3,053 documents. Nevertheless, both works on DocRED revision did not provide gold-quality datasets due to the high cost of annotating the relation triples for long documents. Learning from incompletely annotated training data is crucial and practical for relation extraction. Hence, in this work, we focused on improving the training process with incompletely annotated training data.

To tackle the problem of training with incompletely annotated datasets, prior works leveraged the self-training method to alleviate the detrimental effects of false negative examples (Feng et al., 2018; Hu et al., 2021; Chen et al., 2021; Zhou

Qingyu Tan and Lu Xu are under the Joint PhD Program between Alibaba and NUS/SUTD.

[†] Corresponding author.

[1] Our code is released at https://github.com/DAMO-NLP-SG/CAST

July 9-14, 2023 ©2023 Association for Computational Linguistics

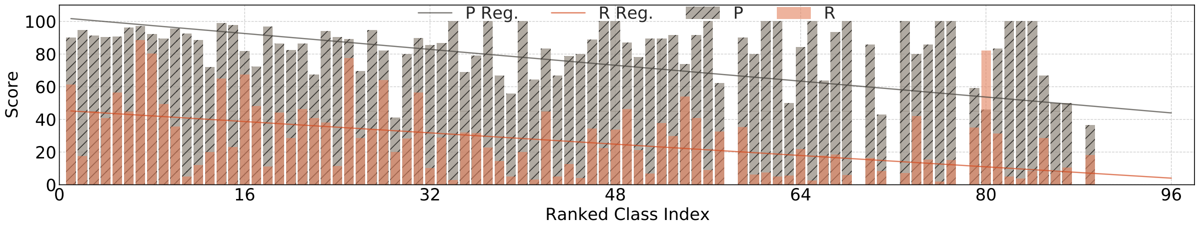


Figure 1: Precision and recall scores of each class (ranked by class frequency: left [high] →right [low]) on the development set of Re-DocRED when the model is trained on DocRED. P Reg. and R Reg. stand for the regression lines of the scores.

et al., 2023). However, self-training-based methods are highly susceptible to confirmation bias, that is, the erroneously predicted pseudo-labels are likely to deteriorate the model's performance in subsequent rounds of training (Arazo et al., 2020; Tarvainen and Valpola, 2017; Li et al., 2020a). Furthermore, the label distribution of relation extraction task is highly imbalanced. Therefore, the predictions made by prior self-training methods are likely to be of the majority classes. Wei et al. (2021) proposed a re-sampling strategy based on class frequencies to alleviate this problem in image classification. In this way, not all generated pseudo-labels will be used to update the training datasets. The pseudo labels of the minority classes have higher probabilities to be preserved than those of the frequent classes. However, such a sampling strategy does not specifically address the problems caused by the erroneously generated pseudo labels. When a model is trained on incompletely annotated datasets, minority classes exhibit bad performance and frequent classes may have low recall scores, as shown in Figure 1. Merging pseudo labels with original labels of the training dataset without considering the correctness of the former potentially deteriorates performance in subsequent iterations.

In order to overcome confirmation bias in self-training, we proposed a class-adaptive self-training (**CAST**) approach that considers the correctness of the pseudo labels. Instead of sampling the pseudo labels based on class frequencies, we introduced a class-adaptive sampling strategy to determine how the generated pseudo labels should be preserved. Specifically, we calculated the precision and recall scores of each class on the development set and used the calculated scores to compute the sampling probability of each class. Through such an approach, CAST can alleviate confirmation bias caused by erroneous pseudo labels. Our proposed approach preserves the pseudo labels from classes that have high precision and low recall scores and penalizes the sampling probability for the pseudo labels that belong to classes with high recall but low precision scores.

Our contributions are summarized as follows. (1) We proposed CAST, an approach that considers the correctness of generated pseudo labels to alleviate confirmation bias in the self-training framework. (2) Our approach was evaluated with training datasets of different quality, and the experimental results demonstrated the effectiveness of our approach. (3) Although our approach is not specifically designed for favoring the minority classes, the minority classes showed more significant performance improvements than the frequent classes, which is a nice property as the problem of long-tail performance is a common bottleneck for real applications.

2 Related Work

Neural Relation Extraction Deep neural models are successful in sentence-level and document-level relation extraction. Zhang et al. (2017) proposed position-aware attention to improve sentence-level RE and published TACRED, which became a widely used RE dataset. Yamada et al. (2020) developed LUKE, which further improved the SOTA performance with entity pre-training and entity-aware attention. Chia et al. (2022b) proposed a data generation framework for zero-shot relation extraction. However, most relations in real-world data can only be extracted based on inter-sentence information. To extract relations across sentence boundaries, recent studies began to explore document-level RE. As previously mentioned, Yao et al. (2019) proposed the popular benchmark dataset DocRED for document-level RE. Zeng et al. (2020) leveraged a double-graph network to model the entities and relations within a document. To address the multi-label problem of DocRE, Zhou et al. (2021) proposed using adaptive thresholds to extract all relations of a given entity pair. Zhang et al. (2021a) developed the DocUNET model to reformulate document-level RE as a semantic segmentation task and used a U-shaped network architecture to improve the performance of DocRE. Tan et al. (2022a) proposed using knowledge distillation and focal loss to denoise the distantly supervised data for DocRE and achieved great performance on the DocRED leaderboard. However, all preceding methods are based on a closed-world assumption (i.e., the entity pairs without relation annotation are negative instances). This assumption ignores the presence of false negative examples. Hence, even the above-mentioned state-of-the-art methods may not perform well when the training data are incompletely annotated.

Denoising for Relation Extraction RE is susceptible to noise in the training data. Noisy data can be categorized into two types: false positives (FPs) and false negatives (FNs). False positive examples are mainly caused by misalignment of knowledge bases. Xiao et al. (2020) proposed a denoising algorithm that filters FP examples in distantly supervised data. Wang et al. (2019) tackled the class-imbalance problem of RE and NER by meta-learning. The false negative problem is also common in information extraction. Li et al. (2020b); Xu et al. (2023) used simple negative sampling strategies to alleviate the detrimental effects of FN examples on NER. Most recently, Guo et al. (2023) tackled the multi-label problem in RE by entropy minimization and supervised contrastive learning. Given that the FN problem is related to incomplete annotation, supplementing the annotation by self-training is a viable way to tackle this problem (Erkan et al., 2007; Sun et al., 2011; Chen et al., 2021; Hu et al., 2021). However, self-training is susceptible to confirmation bias; conventional self-training suffers from the problem of error propagation and makes overwhelming predictions for frequent classes. Prior research on semi-supervised image classification (Wei et al., 2021; He et al., 2021) indicated that re-sampling of pseudo-labels can be beneficial to class-imbalanced self-training. However, existing re-sampling strategies are dependent only on the frequencies of the classes and do not consider the actual performance of each class. Our method alleviates confirmation bias by employing a novel re-sampling strategy that considers the precision and recall of each class on the development set. In this way, we can downsample the predictions for popular classes and maintain high-quality predictions for long-tail classes.

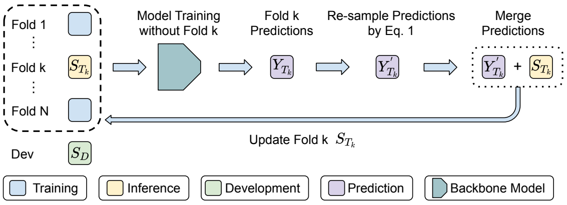


Figure 2: Illustration of training dataset update of CAST, and Algorithm 1 describes its full details.

3 Methodology

3.1 Problem Definition

Document-level relation extraction (DocRE) is defined as follows: given a text T and a set of n entities $\{e_1, ..., e_n\}$ appearing in the text, the objective of the document-level RE is to identify the relation type $r \in C \cup \{no\_relation\}$ for each entity pair (e_i, e_j). Note that e_i and e_j denote two different entities, and C is a predefined set of relation classes. The complexity of this task is quadratic in the number of entities, and the ratio of the NA instances (*no_relation*) is very high compared with sentence-level RE. Therefore, the resulting annotated datasets are often incomplete. The setting of this work is to train a document-level RE model with an incompletely labeled training set, and then the model is evaluated on a clean evaluation dataset, such as Re-DocRED (Tan et al., 2022b).

We denote the training set as S_T and the development set as S_D. Two types of training data are used in this work, each representing a different annotation quality. The first type is the training split of the original DocRED data (Yao et al., 2019), which we refer to as bronze-level training data. This data is obtained by a recommend-revise scheme. Even though the annotation of this bronze level is precise, there are a significant number of missing triples in this dataset. On the other hand, the training set of the Re-DocRED dataset has added a considerable number of triples to the bronze dataset, though a small number of triples might still be missed. We refer to this Re-DocRED dataset as silver-quality training data.

3.2 Our Approach

3.2.1 Overview

The main objective of our approach is to tackle the RE problem when the training data S_T is incompletely annotated. We propose a class-adaptive self-training (CAST) framework, as shown in Figure 2, to pseudo-label the potential false negative examples within the training set. First, we split the training set into N folds and train an RE model with $N-1$ folds. The remaining fold S_{T_k} is used for inference. Next, we use a small development set S_D to evaluate the models and calculate the sampling probability for each relation class (Eq. 1). The predicted label set Y_{T_k} is obtained by conducting inference on S_{T_k}. Then, we re-sample the predicted labels based on the computed probability, which is calculated based on the performance of each class. The re-sampled label set is denoted as Y'_{T_k}. Lastly, Y'_{T_k} will be merged with the initial labels of S_{T_k}. The details of the proposed framework are discussed in the following subsections.

3.2.2 Self-Training

In traditional self-training, models are trained on a small amount of well-annotated data and pseudo-labels are generated on unlabeled instances (Zhu and Goldberg, 2009). However, we do not have access to well-annotated training data, and our training data contains false negative examples. Therefore, we need to construct an N-fold cross-validation self-training system. Given a set of training documents S_T with relation triplet annotation, these documents are divided into N folds. The first $N-1$ folds will be used for training an RE model. Then, the trained model will be used to generate pseudo-labels for the held out N-th fold. The pseudo-labels will be merged with the original labels, and the merged data will be used to train a new model. The N-fold pseudo labeling process will be repeated for multiple rounds until no performance improvement is observed on the final RE system. However, because the class distribution of the document-level RE task is highly imbalanced, pseudo-labeling may favor the popular classes during prediction. This inevitably introduces large confirmation bias to popular classes, which is similar to the "rich-get-richer" phenomenon (Cho and Roy, 2004).

3.2.3 Intuition

When the annotation of the training set is incomplete, the model trained on such data typically shows high precision and low recall scores for most of the classes. Figure 1 shows the precision and recall of each class of the model that is trained on the DocRED dataset and evaluated on the development set of Re-DocRED. Among the 96 classes, most of the classes obtain higher precision scores than recall scores. Only one class that has a higher recall score than precision score; some classes have 0 precision and recall scores. Given this empirical observation, boosting self-training performance by sampling more pseudo-labeled examples from the classes that have high precision and low recall is a good strategy because (1) the pseudo labels of such classes tend to have better quality and (2) the recall performance of these classes can be improved by adding true positive examples. For extreme cases in which a class has predictions that are all wrong (i.e. its precision and recall are both 0), the logical action is to discard the corresponding pseudo-labels.

3.2.4 Class-Adaptive Self-Training (CAST)

As previously mentioned, traditional self-training suffers from confirmation bias, especially for RE task that has a highly imbalanced class distribution. The pseudo-labels that are generated by such an approach tend to be biased toward the majority classes. To alleviate this problem, we propose a class-adaptive self-training framework that filters the pseudo-labels by the per-class performance. Unlike existing self-training re-sampling techniques (Wei et al., 2021; He et al., 2021) that take only the class frequencies into account, our framework samples pseudo-labels based on their performance on the development sets.

First, we evaluate the model for pseudo-labeling on the development set S_D and calculate the precision P and recall R for each class. Then, we define our sampling probability μ_i for each relation class i as:

$$\mu_i = [P_i * (1 - R_i)]^{\beta} \tag{1}$$

where P_i and R_i are the precision and recall scores of class i, respectively, and β is a hyper-parameter that controls the smoothness of the sampling rates.

Note that all pseudo labels will be used when the sampling probability equals to 1. Conversely, all the pseudo labels will be discarded when the sampling probability equals to 0. If the recall of a specific class is very small and its precision is close to 1, the sampling rate of the class will be closer to 1. On the contrary, if the recall for a certain class is high, the sampling rate of the class will

Algorithm 1 Class-Adaptive Self-Training

```
Input:
$M$: Number of rounds
$N$: Number of folds
$S_T$: An incompletely annotated training set
$S_D$: A task-specific development set
$\theta$: A backbone model with parameters
$\beta$: Smoothness coefficient

for $j \in \{1, .., M\}$ do
    $S_T = \{S_{T_1}, ..., S_{T_N}\}$                                   ▷ Split $S_T$ into $N$ folds
    for $k \in \{1, .., N\}$ do
        $S_{T-} \leftarrow S_T - S_{T_k}$
        $\theta_k^* \leftarrow$ Optimize $\theta_k$ with $(S_{T-}, S_D)$              ▷ Training w/o $S_{T_k}$
        $Y_{T_k} \leftarrow$ Inference on $S_{T_k}$ by $\theta_k^*$                ▷ Predict labels of $S_{T_k}$
        Compute $P_i$, $R_i$ with $\theta_k^*$ and $S_D, \forall i \in C$
        $\mu_i \leftarrow [P_i * (1 - R_i)]^\beta, \forall i \in C$                       ▷ Eq. 1
        $Y_{T_k}^{'} \leftarrow$ Re-sample $Y_{T_k}$ with rates $\{\mu_i\}, \forall i \in C$
        $S_{T_k}^{'} \leftarrow$ Merge $Y_{T_k}^{'}$ with annotation of $S_{T_k}$
    $S_T \leftarrow S_{T_1}^{'} \cup ... \cup S_{T_N}^{'}$                            ▷ Update training set
    $\theta_j^* \leftarrow$Optimize $\theta$ with $(S_T, S_D)$                   ▷ Save model for round $j$
$\theta^* \leftarrow$ evaluate $\{\theta_1^*, ..., \theta_M^*\}$ on $S_D$
return $\theta^*$
```

DocRE	DocRED	Re-DocRED	Re-DocRED	
	Train	Train	Dev	Test
# Documents	3,053	3,053	500	500
Avg. # Entities per Doc	19.4	19.4	19.4	19.6
Avg. # Triples per Doc	12.5	28.1	34.6	34.9
Avg. # Sentences per Doc	7.9	7.9	8.2	7.9
# NA rate	97.0%	94.3%	93.1%	93.1%

BioRE	ChemDisGene		
	Train	Dev	Test
# Documents	76,544	1,480	523
Avg. # Words	196.6	237.3	235.6
Avg. # Entities per Doc	7.6	9.0	10.0
Avg. # Triples per Doc	2.2	2.2	7.2
Avg. # Sentences per Doc	12.6	14.0	13.2
# NA rate	96.8%	97.7%	93.8%

Table 1: Dataset statistics of our experiments for DocRE and BioRE.

be low. In this way, our method is able to alleviate confirmation bias toward the popular classes, which typically have higher recall. The pseudocode of our proposed CAST framework is provided in Algorithm 1.

4 Experiments

4.1 Experimental Setup

Our proposed CAST framework can be applied with any backbone RE model. For the experiment on DocRED, we adopted the ATLOP (Zhou et al., 2021) model as the backbone model, which is a well-established baseline for the DocRE task. We used BERT-Base (Devlin et al., 2019) and RoBERTa-Large (Liu et al., 2019) as the encoders. In addition to DocRED, we conduct experiments on ChemDisGene (Zhang et al., 2022), a DocRE dataset for biomedical relation extraction (BioRE). We used the PubMedBERT (Gu et al., 2021) encoder for the BioRE experiments. We use the development set of Re-DocRED in the document-level RE experiments because the Re-DocRED dataset has a high quality. Moreover, we use the distantly-supervised development set of ChemDisGene for the BioRE experiments. Our final models are evaluated on the test sets of Re-DocRED and ChemDisGene. Both of the test sets are human-annotated and have high quality, the statistics of the datasets can be found in Table 1.

For the hyper-parameters, we set $M = 5$ (i.e., the iteration round in Algorithm 1) and $N = 5$ for the self-training-based methods because these methods typically reach the highest performance before the fifth round and five-fold training is the conventional practice for cross validation. For β, we grid searched $\beta \in \{0.0, 0.25, 0.5, 0.75, 1\}$. For evaluation, we used micro-averaged F1 score as the evaluation metric. We also evaluate the F1 score for frequent classes and long-tail classes, denoted as Freq_F1 and LT_F1, respectively. For the DocRED dataset, the frequent classes include the top 10 most popular relation types[2] in the label space; the rest of the classes are categorized as the long-tail classes. Following Yao et al. (2019), we use an additional metric Ign_F1 on the DocRE task. This metric calculates the F1 score for the triples that do not appear in the training data.

4.2 Baselines

Vanilla Baselines This approach trains existing state-of-the-art RE models on incompletely annotated data and serves as our baseline method. As stated earlier, we use **ATLOP** as the backbone model for the DocRE experiments. In addition to ATLOP, we compare **GAIN** (Zeng et al., 2020), **DocuNET** (Zhang et al., 2021a), and **KD-DocRE** (Tan et al., 2022a) as our vanilla baselines. These methods are top-performing methods on the Re-DocRED dataset. However, similar to ATLOP, the performances of these models deteriorate significantly under the incomplete annotation setting.

Negative Sampling (NS) (Li et al., 2020b) This method tackles the incomplete annotation problem through negative sampling. To alleviate the effects of false negatives, this method randomly selects partial negative samples for training. Such an approach can help to alleviate the detrimental effect of the false negative problem.

[2] They cover 59.4% of the positive instances.

	Model	P	R	F1	Ign_F1	Freq_F1	LT_F1
BERT	GAIN†	88.11	30.98	45.82	45.57	-	-
	ATLOP	**88.39** ±0.39	28.87 ±0.34	43.52 ±0.25	43.28 ±0.24	45.49 ±0.24	40.46 ±0.28
	SSR-PU-ATLOP†	65.10 ±0.90	50.53 ±0.89	56.84 ±0.72	55.45 ±0.59	60.21 ±0.64	51.84 ±0.82
	NS-ATLOP	74.79 ±0.31	46.33 ±0.34	57.22±0.25	56.28 ±0.21	59.23 ±0.23	54.13 ±0.24
	VST-ATLOP	63.53 ±1.17	**56.41** ±0.86	59.56 ±0.16	58.03 ±0.25	63.17 ±0.46	55.61 ±0.25
	CREST-ATLOP	69.34 ±1.55	50.58 ±1.35	58.48 ±0.30	57.33 ±0.21	60.31 ±0.64	56.33 ±0.15
	CAST-ATLOP (Ours)	70.49 ±1.12	54.34 ±1.07	**61.36** ±0.67	**60.16** ±0.79	**63.66** ±0.44	**58.12** ±0.36
RoBERTa	DocuNET†	**94.16**	30.42	45.99	45.88	-	-
	KD-DocRE*	92.08	32.07	47.57	47.32	-	-
	ATLOP	92.62 ±0.35	33.61 ±0.48	49.32 ±0.29	49.16 ±0.27	51.49 ±0.51	45.36 ±0.43
	SSR-PU-ATLOP†	65.71 ± 0.28	57.01 ±0.47	61.05 ±0.21	59.48 ±0.18	62.85 ±0.10	58.19 ±0.54
	NS-ATLOP	68.39 ±2.23	56.05 ±0.98	61.58 ±0.48	60.43 ±0.55	65.35 ±0.12	57.16 ±0.44
	VST-ATLOP	62.85 ±0.48	**63.58** ±0.62	63.21 ±0.39	61.83 ±0.41	65.68 ±0.43	60.09 ±0.45
	CREST-ATLOP	73.09 ±0.79	55.06 ±0.86	62.81 ±0.35	61.90 ±0.33	63.71 ±0.41	61.75 ±0.49
	CAST-ATLOP (Ours)	72.83 ±0.50	59.22 ±0.61	**65.32** ±0.22	**64.25** ±0.15	**66.99** ±0.29	**63.05** ±0.11

Table 2: Experimental results on the test set of Re-DocRED when trained with DocRED. Model selection is based on the dev set of Re-DocRED. The reported results are the average of five runs. †: The results are reproduced from Wang et al. (2022) with the same development set S_D. *: The results are retrieved from Tan et al. (2022b).

Vanilla Self-Training (VST) (Peng et al., 2019; Jie et al., 2019) VST is a variant of simple self-training. In this approach, models are trained with N folds, and all pseudo-labels are directly combined with the original labels. Then, a new model is trained on the datasets with combined labels.

Class Re-balancing Self-Training (CREST) (Wei et al., 2021) This algorithm is the most advanced baseline of class-imbalanced semi-supervised training, re-samples the pseudo-labels generated by models. However, this sampling strategy only considers the frequencies of the training samples, whereas our CAST considers the per-class performance on the development set.

SSR Positive Unlabeled Learning (SSR-PU) (Wang et al., 2022) This method applies a positive unlabeled learning algorithm for DocRE under the incomplete annotation scenario. SSR-PU utilizes a shift-and-squared ranking (SSR) loss to accommodate the distribution shifts for the unlabeled examples.

BioRE Baselines For the BioRE experiments, we compare our methods with Biaffine Relation Attention Network **BRAN** (Verga et al., 2018) and **PubmedBERT** (Gu et al., 2021), which is a pretrained language model in the biomedical domain.

4.3 Experimental Results

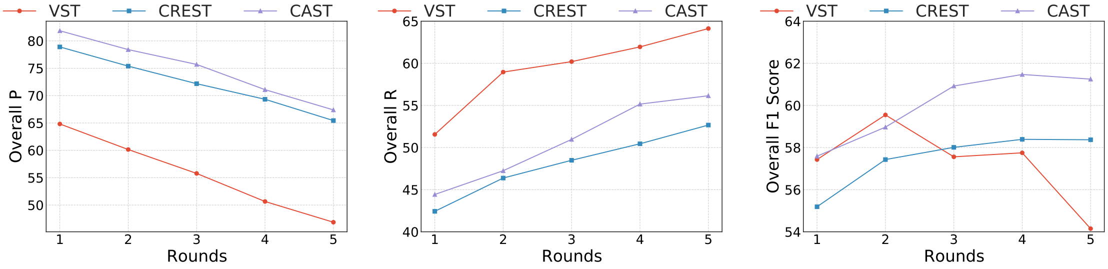
Table 2 presents the experimental results for the document-level RE. The experimental results on the original DocRED dataset show that the F1 score of the ATLOP-RoBERTa model is only 49.32. This finding can be ascribed to the low recall score of this method, as shown in Figure 1. NS significantly improves the performance compared with the baseline. After comparing vanilla self-training with the baseline, we observe that although the recall score is the highest for this method, its precision is significantly reduced. We observe similar trends for all self-training based methods (i.e., VST, CREST, and CAST), the recall improved at the expense of precision. Notably, the performance of the simple NS baseline exceeds the performance of SSR-PU when

	Model	P	R	F1	Ign_F1	Freq_F1	LT_F1
	Re-DocRED Training Data						
BERT	ATLOP	**86.70**	62.46	72.61	71.86	75.92	67.46
	NS-ATLOP	77.63	69.17	73.16	72.92	77.28	67.59
	VST-ATLOP	72.77	**75.55**	74.14	72.48	78.47	68.13
	CREST-ATLOP	75.94	72.47	74.17	72.77	77.93	68.68
	CAST-ATLOP (Ours)	76.59	72.84	**74.67**	**73.32**	**78.53**	**69.34**

Table 3: Experimental results on the test set of Re-DocRED when trained on silver quality data.

	Model	P	R	F1
	BRAN†	41.8	26.6	32.5
PubMedBERT	PubMedBERT†	64.3	31.3	42.1
	BRAN†	70.9	31.6	43.8
	ATLOP*	**76.17** ± 0.36	29.70 ±0.54	42.73 ±0.36
	SSR-PU-ATLOP*	54.27 ±0.23	43.93 ±0.40	48.56 ±0.32
	NS-ATLOP	71.54 ±0.50	35.52 ±0.29	47.47 ±0.37
	VST-ATLOP	54.92 ±0.42	**48.39** ±0.58	51.24 ±0.30
	CREST-ATLOP	59.42 ±1.63	42.12 ±0.65	49.28 ±0.21
	CAST-ATLOP (Ours)	66.68 ±2.22	45.48 ±1.27	**54.03** ±0.17

Table 4: Experimental results on ChemDisGene. The results with numeric superscripts are taken from the respective papers. †: The results are retrieved from Zhang et al. (2022). *: The results are retrieved from Wang et al. (2022).

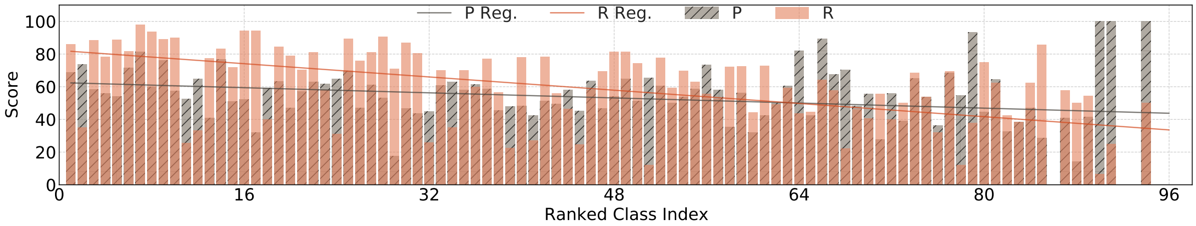


(a) Precision

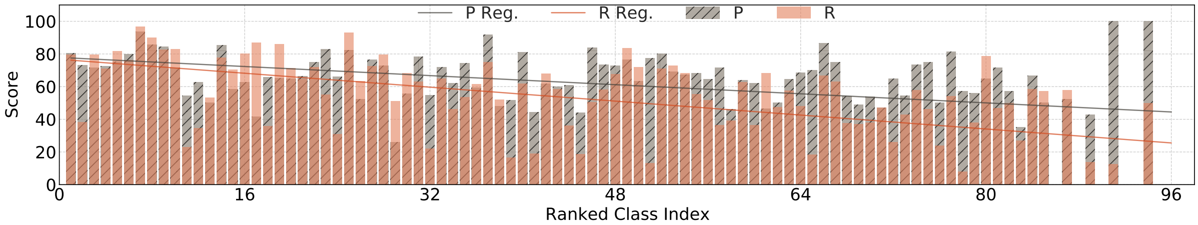
(b) Recall

(c) Overall F1

Figure 3: Comparison of different self-training strategies when training on DocRED.



(a) VST



(b) CAST

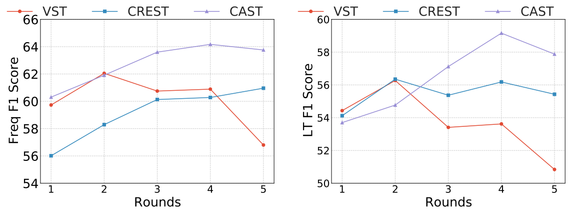
Figure 4: Precision and recall scores for each class (ranked by class frequency: left (high)→right (low)) on the dev set of Re-DocRED of VST and CAST's best models, which are trained on DocRED. Better view in color.

trained on the DocRED data. Our proposed CAST framework consistently outperforms the competitive baselines and achieves the highest performance for both BERT and RoBERTa encoders. Our best-performing model outperforms the baseline by 16.0 F1 (49.32 vs. 65.32). Moreover, the CAST obtains the highest precision score among the three self-training methods, thereby showing that the examples added by our class-adaptive sampling strategy have better quality.

The experimental results on the test set of Re-DocRED (Table 3) depict that the baseline F1 score is significantly improved due to the large gain in the recall score when the training data are switched from bronze-quality to silver-quality. Compared with baseline approaches, our CAST achieves consistent performance improvements in terms of F1 score. The F1 difference between the baseline and our CAST is 2.06 (72.61 vs. 74.67). However, the performance gap between our approach and the baseline is smaller than the corresponding gap when both are trained with DocRED. This indicates that the performance of existing state-of-the-art models for document-level RE is decent when high-quality training data is provided but declines when the training data are incompletely annotated. This finding verifies the necessity of developing better self-training techniques because preparing high-quality training data is costly.

Table 4 presents the experiments on biomedical RE. Our CAST model consistently outperforms strong baselines, exceeding the performance of SSR-PU by 5.47 F1 (54.03 vs. 48.56).

On the basis of the results of DocRE and BioRE experiments, self-training-based methods aim to improve recall and consistently improve overall performance when the training data is incompletely annotated. However, our CAST maintains a better balance between increasing recall and maintaining precision.



(a) Frequent F1

(b) Long-tail F1

Figure 5: F1 scores of frequent and long-tail classes with respect to rounds when trained on DocRED.

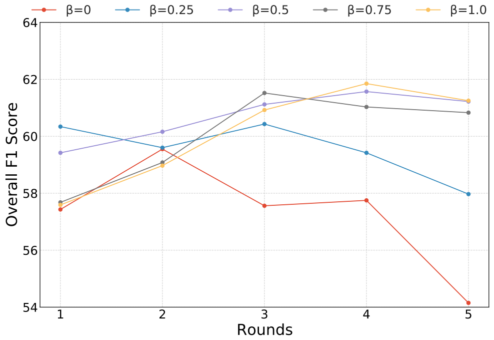


Figure 6: Effect of different β values.

5 Analysis

5.1 Comparisons of Self-Training Strategies

To further compare different self-training strategies, we illustrate the detailed performance with respect to the self-training rounds in Figure 3. The reported scores are on the development set of Re-DocRED and the training data is from DocRED. Figure 3b shows that all self-training-based methods generally have improving recall scores as the number of self-training rounds increases. On the contrary, the precision scores decline. From Figure 3c, we observe that VST outperforms CREST and CAST in the first two rounds. This is mainly because VST does not perform re-sampling on the pseudo-labels and it utilizes all pseudo-labels. At the beginning stage, these labels are of relatively good quality. However, the performance of VST drops after the second round of pseudo-labeling because as the number of rounds increases, the increase in the number of false positive examples in the pseudo-labels outweighs the benefit. Meanwhile, the performance gains of CREST and CAST are relatively stable, and both methods produce their best-performing models at round 4. Compared with CREST, our CAST maintains higher precision scores as the number of rounds increases (Figure 3a).

We also assess the F1 performance of the frequent and long-tail classes with respect to the number of rounds, and the comparison is shown in Figure 5. The results reveal that VST suffers greatly from confirmation bias on both frequent and LT classes, i.e., Figure 5a and Figure 5b, and its performance becomes very poor in round 5. In Figure 5b, we can see that the performance gains of CAST is stable across the training rounds and achieved the best LT performance.

5.2 Detailed Analysis of CAST

In this section, we analyze the performance of our CAST framework in detail. We first plot the precision and recall scores of VST and CAST for all the classes in Figure 4, where the experimental results are obtained by training with the DocRED dataset. The formulation of Figure 4 is the same as Figure 1. Figure 4a demonstrates that VST significantly improves the recall scores of many classes compared with the baseline in Figure 1. However, the improvements in recall scores are accompanied by a large decline in precision scores. This observation shows that the pseudo-labels in VST contain a considerable amount of erroneous predictions. By contrast, our CAST framework is able to better maintain the precision scores for most of the classes. The recall scores for most of the classes are significantly higher compared with those of the baseline. This observation justifies the improvements of the overall F1 scores in Table 2 despite the lower recall of CAST model than VST.

5.3 Effect of β

We further analyze the effect of the sampling coefficient β on our CAST framework in Figure 6, the experiments are conducted by training with the DocRED dataset. When β value is small, CAST behaves like the VST model, exhibits some F1 improvements in the first few rounds, and demonstrates diminishing positive effects in the later rounds. Larger β leads to better overall improvements and smaller fluctuations across different rounds. However, because the term $[P_i * (1 - R_i)]$ in Eq. 1 is smaller than 1, higher β may lead to lower sampling rates for all the classes. As a result, the convergence time of self-training may be longer. The interpretation for other values of β is provided in the Appendix C.

6 Conclusions and Future Work

In this work, we study the under-explored problem of learning from incomplete annotation in relation extraction. This problem is highly important in real-world applications. We show that existing state-of-the-art models suffer in this scenario. To tackle this problem, we proposed a novel CAST framework. We conducted experiments on DocRE and BioRE tasks, and experimental results show that our method consistently outperforms competitive baselines on both tasks. For future work, we plan to extend our framework to the distant supervision scenario. From the domain perspective, we plan to apply our framework to image classification tasks.

7 Limitations

The proposed CAST framework carries the same limitation of self-training-based methods, which is the requirement for multiple rounds and multiple splits of training. As a result, the GPU computing hours of CAST are longer than those of vanilla baselines and NS.

References

Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. 2020. TACRED revisited: A thorough evaluation of the TACRED relation extraction task. In *Proceedings of ACL*.

Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. 2020. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *Proceedings of IJCNN*.

Jhih-wei Chen, Tsu-Jui Fu, Chen-Kang Lee, and Wei-Yun Ma. 2021. H-FND: hierarchical false-negative denoising for distant supervision relation extraction. In *Findings of ACL*.

Yew Ken Chia, Lidong Bing, Sharifah Mahani Aljunied, Luo Si, and Soujanya Poria. 2022a. A dataset for hyper-relational extraction and a cube-filling approach. In *Proceedings of EMNLP*.

Yew Ken Chia, Lidong Bing, Soujanya Poria, and Luo Si. 2022b. RelationPrompt: Leveraging prompts to generate synthetic data for zero-shot relation triplet extraction. In *Findings of ACL*.

Junghoo Cho and Sourashis Roy. 2004. Impact of search engines on page popularity. In *Proceedings of WWW*, page 20–29.

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

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

Güneş Erkan, Arzucan Özgür, and Dragomir R. Radev. 2007. Semi-supervised classification for extracting protein interaction sentences using dependency parsing. In *Proceedings of EMNLP-CoNLL*.

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

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 Transactions on Computing for Healthcare (HEALTH)*.

Jia Guo, Stanley Kok, and Lidong Bing. 2023. Towards integration of discriminability and robustness for document-level relation extraction. In *Proceedings of EACL*.

Ju He, Adam Kortylewski, Shaokang Yang, Shuai Liu, Cheng Yang, Changhu Wang, and Alan Yuille. 2021. Rethinking re-sampling in imbalanced semi-supervised learning. *arXiv preprint arXiv:2106.00209*.

Xuming Hu, Chenwei Zhang, Fukun Ma, Chenyao Liu, Lijie Wen, and Philip S. Yu. 2021. Semi-supervised relation extraction via incremental meta self-training. In *Findings of EMNLP*.

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 ACL*.

Zhanming Jie, Pengjun Xie, Wei Lu, Ruixue Ding, and Linlin Li. 2019. Better modeling of incomplete annotations for named entity recognition. In *Proceedings of NAACL*.

Junnan Li, Richard Socher, and Steven C.H. Hoi. 2020a. Dividemix: Learning with noisy labels as semi-supervised learning. In *Proceedings of ICLR*.

Xin Li, Lidong Bing, Piji Li, Wai Lam, and Zhimou Yang. 2018. Aspect term extraction with history attention and selective transformation. In *Proceedings of IJCAI*.

Yangming Li, Shuming Shi, et al. 2020b. Empirical analysis of unlabeled entity problem in named entity recognition. In *Proceedings of ICLR*.

Zuchao Li, Hai Zhao, Shexia He, and Jiaxun Cai. 2021. Syntax role for neural semantic role labeling. *Computational Linguistics*, 47(3).

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*.

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*.

George Stoica, Emmanouil Antonios Platanios, and Barnabás Póczos. 2021. Re-tacred: addressing shortcomings of the tacred dataset. In *Proceedings of AAAI*.

Ang Sun, Ralph Grishman, and Satoshi Sekine. 2011. Semi-supervised relation extraction with large-scale word clustering. In *Proceedings of ACL*.

Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. 2022a. Document-level relation extraction with adaptive focal loss and knowledge distillation. In *Findings of ACL*.

Qingyu Tan, Lu Xu, Lidong Bing, and Hwee Tou Ng. 2022b. Revisiting docred–addressing the overlooked false negative problem in relation extraction. In *Proceedings of EMNLP*.

Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proceedings of NIPS*.

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

Ye Wang, Xinxin Liu, Wenxin Hu, and Tao Zhang. 2022. A unified positive-unlabeled learning framework for document-level relation extraction with different levels of labeling. In *Proceedings of EMNLP*.

Zihao Wang, Kwunping Lai, Piji Li, Lidong Bing, and Wai Lam. 2019. Tackling long-tailed relations and uncommon entities in knowledge graph completion. In *Proceedings of EMNLP*.

Chen Wei, Kihyuk Sohn, Clayton Mellina, Alan Yuille, and Fan Yang. 2021. Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In *Proceedings of CVPR*.

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

Lu Xu, Lidong Bing, and Wei Lu. 2023. Better sampling of negatives for distantly supervised named entity recognition. In *Findings of ACL*.

Lu Xu, Zhanming Jie, Wei Lu, and Lidong Bing. 2021. Better feature integration for named entity recognition. In *Proceedings of NAACL*.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of EMNLP*.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: a large-scale document-level relation extraction dataset. In *Proceedings of ACL*.

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

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 LREC*.

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

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021b. Aspect sentiment quad prediction as paraphrase generation. In *Proceedings of EMNLP*.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of EMNLP*, pages 35–45.

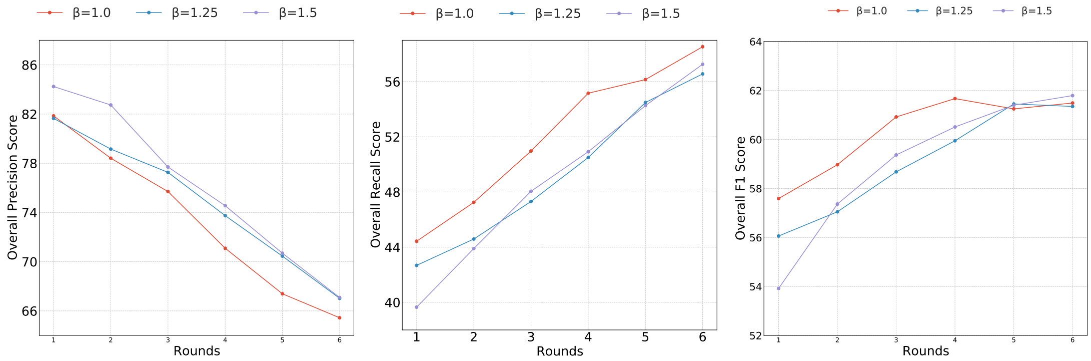
Ran Zhou, Xin Li, Lidong Bing, Erik Cambria, and Chunyan Miao. 2023. Improving self-training for cross-lingual named entity recognition with contrastive and prototype learning. In *Proceedings of ACL*.

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

Xiaojin Zhu and Andrew B. Goldberg. 2009. Introduction to semi-supervised learning. In *Introduction to Semi-Supervised Learning*.

A Sentence-Level Relation Extraction

Besides document-level RE, we also examined our method for sentence-level relation extraction (SentRE), the task is a simplified version of its document-level counterpart. Compared to the DocRE setting, there are two main differences for



(a) Precision

(b) Recall

(c) Overall F1

Figure 7: Effect of larger β when training on DocRED.

sentence-level RE. First, there are exactly $n = 2$ entities for each SentRE example. Second, there is only one relation type for an entity pair in SentRE, whereas there can be multiple relation types for DocRE. Again, we used two types of training data for the SentRE task. The first set of training data is from the original TACRED dataset, and the second set of training data is from Re-TACRED. Compared to the revision of Re-DocRED, which only resolved the false negative problem[3], the revision of Re-TACRED not only resolved the false negative problem but also relabeled the false positive instances.

The experimental results on SentRE are shown in Table 5. For the TACRED dataset, the top 5 classes[4] are included in the frequent classes. We can also see that when training with bronze-quality data (i.e., the upper section), our proposed CAST still achieves the best performance in terms of F1 score. This observation shows that our method is effective across different relation extraction scenarios and backbone models. On the other hand, we can observe that the baseline model achieves the highest F1 score when training with the Re-TACRED dataset (i.e., the lower section). As mentioned in the section of problem definition, the Re-TACRED training set has resolved the false negative and false positive problems of TACRED. Therefore, by simply using all training samples of Re-TACRED, the baseline approach achieves the best F1. It is worth noting that our CAST is very robust and does not hurt the performance, i.e., achieving slightly worse F1 but slightly better recall compared with the baseline.

Model	P	R	F1	Freq_F1	LT_F1
TACRED Training Data					
Baseline	**80.64**	40.44	53.87	70.62	36.43
NS	62.37	**53.96**	57.86	74.49	40.57
VST	67.24	52.83	59.17	80.64	47.25
CREST	67.92	52.48	59.21	80.36	47.64
CAST (Ours)	73.33	51.03	**60.18**	**81.12**	**48.75**
Re-TACRED Training Data					
Baseline	**88.01**	87.82	**87.91**	**89.21**	**87.37**
NS	85.44	88.56	86.97	88.75	86.24
VST	83.71	**89.82**	86.65	87.94	85.99
CREST	86.46	88.45	87.45	88.64	86.97
CAST (Ours)	87.62	87.96	87.78	89.14	87.31

Table 5: Experimental results on the test set of Re-TACRED when trained on TACRED and Re-TACRED, respectively. Model selection is based on the dev set of Re-TACRED.

B Hyper-Parameters of the Baselines

In this section, we report the hyper-parameters of the baseline experiments. For the negative sampling experiments, we used sampling rate $\gamma = 0.1$ for the DocRED experiment, $\gamma = 0.5$ for TACRED experiment and $\gamma = 0.7$ for the Re-TACRED and Re-DocRED experiments. γ is searched from $\gamma \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$.

From CREST (Wei et al., 2021), the classes are first ranked by their frequencies, and the sampling rate for class i is calculated as:

$$\mu_i = (\frac{X_{|C|+1-i}}{X_1})^{\alpha} \tag{2}$$

where X_1 is the count of the most frequent class among the positive classes. We set the power $\alpha = 0.33$ as reported in their paper. For all the self-training-based experiments (VST, CREST, and CAST), we trained with 10 epochs per fold. All our experiments were run on a NVIDIA-V100 GPU.

[3] The problem of false positive is minor in DocRED.

[4] They cover 57.5% of the positive instances.

	Model	P	R	F1	Ign_F1	Freq_F1	LT_F1
DocRED Training Data with Incomplete S_D							
BERT	ATLOP	**88.39**	28.87	43.52	43.28	45.49	40.46
	SSR-PU	70.42	46.67	56.14	55.21	59.38	49.24
	NS-ATLOP	55.98	55.63	55.78	53.90	58.73	51.92
	VST-ATLOP	63.03	**51.60**	56.71	55.26	60.75	51.52
	CREST-ATLOP	72.83	47.81	57.72	56.71	59.05	54.82
	CAST-ATLOP (Ours)	70.97	50.70	**59.14**	**58.03**	**61.20**	**56.22**

Table 6: Experimental results on the test set of Re-DocRED when trained on DocRED training data and using the development set of DocRED for model selection.

C Experiments on larger β

In Figure 7, we show the experimental results when β is larger than 1.0. Increasing β inevitably reduces the sampling probability for all the classes, which is more conservative. Therefore, larger β tends to have higher precision scores and lower recall scores. From Figure 7, we see that the optimal round for F1 scores is 4 for $\beta = 1.0$ and 5 for $\beta = 1.25$. When $\beta > 1.5$, the F1 score may not reach the optimal point before the 6th round. Since CAST would require training MN times, larger β may lead to significantly longer computation time to reach the optimal F1 score.

D Experiments with Incomplete S_D

In this section, we conducted experiments on DocRED with a development set of lower quality. Specifically, we used S_D from the DocRED dataset instead of Re-DocRED. The experiment results are shown in Table 6. We can see that the over performances of most methods were decreased. This observation showed the importance of a high-quality development set when training with incomplete data. Nevertheless, our CAST model still achieves the best overall performance among the compared methods.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
7

□ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

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

B ☑ Did you use or create scientific artifacts?

section 4

☑ B1. Did you cite the creators of artifacts you used?
section 4

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We only used open-source scientific data in this paper.

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

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We only used open-source scientific data in this paper.

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

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

C ☑ Did you run computational experiments?

4

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

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4 and Appendix B

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

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4

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

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.