Distantly-Supervised Named Entity Recognition with Uncertainty-Aware Teacher Learning and Student-Student Collaborative Learning

Anonymous ACL submission

Abstract

Distantly-Supervised Named Entity Recognition (DS-NER) effectively alleviates the burden of annotation, but meanwhile suffers from the label noise. Recent works attempt to adopt the teacher-student framework to gradually refine the training labels and improve the overall robustness. However, we argue that these teacher-student methods achieve limited performance because the poor calibration of the teacher network produces incorrectly pseudo-labeled samples, leading to error propagation. Therefore, we attempt to mitigate this issue by proposing: (1) Uncertainty-Aware Teacher Learning that leverages the prediction uncertainty to reduce the number of incorrect pseudo labels in the self-training stage; (2) Student-Student Collaborative Learning that allows the transfer of reliable labels between two student networks instead of indiscriminately relying on all pseudo labels from its teacher, and further enables a full exploration of mislabeled samples rather than simply filtering unreliable pseudo-labeled samples. We evaluate our proposed method on five DS-NER datasets, demonstrating that our method is superior to the state-of-the-art DS-NER denoising methods.

1 Introduction

Named Entity Recognition (NER) aims to detect entity spans in text and then classify them into predefined categories, which plays an important role in many applications such as dialogue systems (Li and Zhao, 2023; Liu et al., 2023). However, deep learning-based NER methods usually require a substantial quantity of high-quality annotation for training models, which is not only exceedingly costly but also time-consuming.

To alleviate the burden of annotation, Distantly-Supervised Named Entity Recognition (DS-NER)

[1]Our code will be open-sourced after peer review.

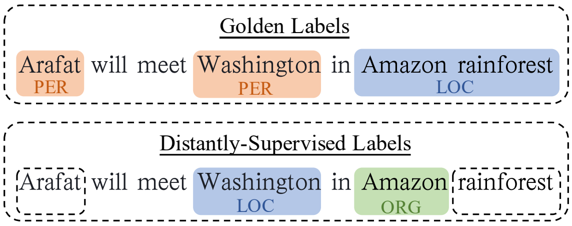

Figure 1: A sample generated by DS-NER. "Amazon" and "Washington" are inaccurate annotations. "Arafat" and "rainforest" are the incomplete annotations.

is widely used in real-world scenarios. It can automatically generate massive labeled training data by matching entities in existing knowledge bases with snippets in text. However, DS-NER suffers from two inherent issues: (1) **Inaccurate Annotation**: due to the context-free matching, the entity with multiple types in the knowledge bases may be labeled as an inaccurate type, and (2) **Incomplete Annotation**: due to the limited coverage of knowledge bases, many entity mentions in text cannot be matched and are wrongly labeled as non-entity. As shown in Figure 1, the entity types of "Washington" and "Amazon" are wrongly labeled owing to context-free matching, and "Arafat" is not recognized due to the limited coverage of resources.

Therefore, many works attempt to address these issues (Peng et al., 2019; Zhou et al., 2022; Li et al., 2021; Si et al., 2022, 2023). Recently, the self-training teacher-student framework in DS-NER has attracted increasing attention (Liang et al., 2020; Zhang et al., 2021a; Qu et al., 2023), as it can handle inaccurate and incomplete labels simultaneously, and use generated pseudo labels to make full use of the mislabeled samples from DS-NER dataset. This self-training framework firstly uses generated reliable pseudo labels from the teacher network to train the student network, and then updates a new teacher by shifting the weights of the trained student. Through this self-training loop, the training labels are gradually refined and model generalization can be improved. Specifically, BOND (Liang et al., 2020) designs a teacher-student network and selects high-confidence pseudo labels as reliable labels to get a more robust model. SCDL (Zhang et al., 2021b) further improves the performance by jointly training two teacher-student networks, then selects consistent and high-confidence pseudo labels between two teachers as reliable labels. ATSEN (Qu et al., 2023) designs two teacher-student networks by considering both consistent and inconsistent high-confidence pseudo labels between two teachers and also proposes fine-grained teacher updating to achieve advanced performance.

The above teacher-student methods highly rely on using the high-confidence pseudo labels (e.g., pseudo labels with confidence values greater than 0.7) as reliable labels, as they assume that the teacher model's predictions with high confidence tend to be correct. However, this assumption may be far from reality. Neural networks are usually poorly calibrated (Guo et al., 2017; Rizve et al., 2021), i.e., the probability associated with the predicted label usually reflects the bias of the teacher network and does not reflect the likelihood of its ground truth correctness. Therefore, a poorly calibrated teacher network can easily generate incorrect pseudo labels with high confidence. We argue that previous teacher-student methods achieve limited performance because poor network calibration produces incorrect pseudo-labeled samples, leading to error propagation.

We aim to reduce the effect of incorrect pseudo labels within the teacher-student framework by un**C**ertainty-aware t**E**acher a**N**d **S**tudent-Student c**O**llaborative lea**R**ning (**CENSOR**). Specifically, we apply two teacher-student networks to provide multi-view predictions on training samples. We propose Uncertainty-aware Teacher Learning that leverages the prediction uncertainty to guide the selection procedure of pseudo labels. Then, we use both uncertainty and confidence as indicators to select pseudo labels, reducing the number of incorrect pseudo labels selected by confidence scores from poorly calibrated teacher networks. We only select the pseudo labels with high confidence and low uncertainty as reliable labels, since these selected labels are more likely to contain less noise. Subsequently, to further reduce the risk of learning incorrect pseudo labels and make a full exploration of mislabeled samples, we introduce Student-Student Collaborative Learning that allows the transfer of reliable labels between two student networks. In each batch of data, each student network views its small-loss pseudo labels (e.g., pseudo labels of 10% samples with the smallest loss) as reliable labels and then teaches such reliable labels to the other student network for updating the parameters. In this way, a student network does not completely rely on all the pseudo labels from its poorly calibrated teacher network. Meanwhile, different from just filtering unreliable pseudo-labeled samples, this component provides the opportunity for the incorrect pseudo-labeled samples to be correctly labeled by the other teacher-student network, allowing the full exploration of training data. Experiments demonstrate that our method significantly outperforms previous methods, e.g., improving the F1 score by an average of 1.87% on five DS-NER datasets.

2 Related Work

To alleviate the burden of annotation, previous studies attempted to annotate NER datasets via distant supervision, which suffers from noisy annotation.

DS-NER Methods To address these issues, various methods have been proposed. Several studies (Shang et al., 2018; Yang et al., 2018; Jie et al., 2019) modify CRF to get better performance under the noise. Peng et al. (2019); Zhou et al. (2022) try to employ PU learning to obtain the unbiased estimation of loss value. Li et al. (2021, 2022) introduce negative sampling to mitigate the misguidance from unlabeled entities. Liang et al. (2020); Zhang et al. (2021b); Qu et al. (2023) adopt the teacher-student framework to handle both inaccurate and incomplete labels simultaneously. In this paper, we attempt to reduce the effect of incorrect pseudo labels and error propagation in the teacher-student framework to achieve better performance.

Teacher-Student Framework Teacher-student framework is a popular architecture in many semi-supervised tasks (Huo et al., 2021). Recently, the teacher-student framework has attracted increasing attention in DS-NER task. BOND (Liang et al., 2020) firstly attempts to apply self-training with a teacher-student network in DS-NER. SCDL (Zhang et al., 2021b) further improves the performance by jointly training two teacher-student networks. ATSEN (Qu et al., 2023) considers both consistent and inconsistent predictions between two teachers and proposes fine-grained teacher updating to achieve more robustness. We improve the teacher-student

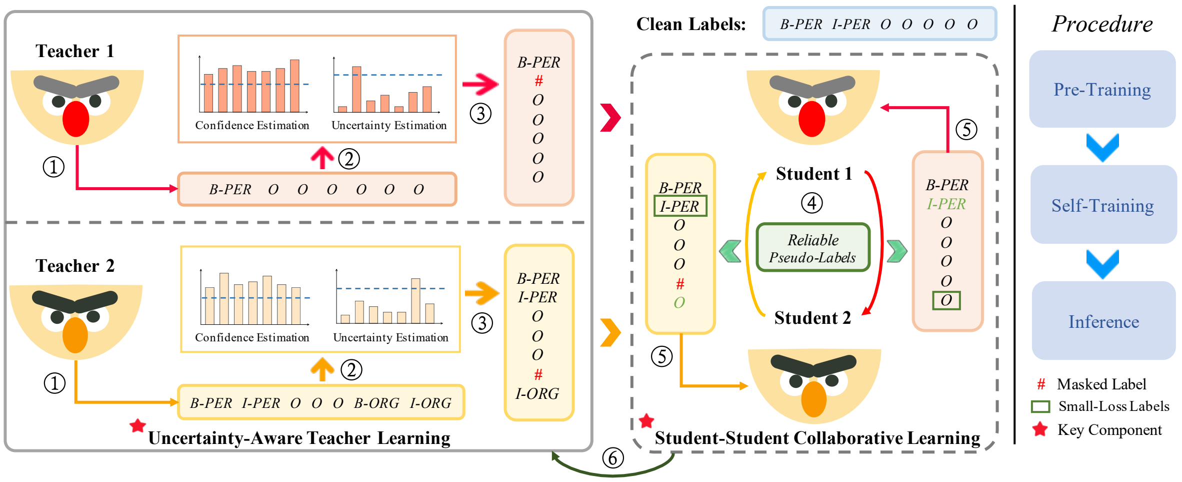

Figure 2: General architecture of CENSOR, which consists of two teacher-student networks. [①] means the teacher network first generates pseudo labels. [②] means estimating the confidence and uncertainty of generated pseudo labels. [③] means selecting reliable pseudo labels according to confidence and uncertainty, where masked pseudo labels will not be used to update the student network. [④] means using Student-Student Collaborative Learning to transfer the reliable pseudo labels. [⑤] means using selected reliable pseudo labels to update the corresponding student network. [⑥] means updating a new teacher by shifting the weights of the trained student.

framework by Uncertainty-Aware Teacher Learning and Student-Student Collaborative Learning, jointly reducing the effect of incorrect pseudo labels. In this way, our method can avoid error propagation and achieve better overall performance.

3 Task Definition

Given the training corpus D_{ds} where each sample (x_i, y_i), x_i represents i-th token, and y_i is the label. Each entity is a span of the text, associated with an entity type. We use the BIO scheme for sequence labeling. The beginning token of an entity is labeled as *B-type*, and others are *I-type*. The non-entity tokens are labeled as *O*. Traditional NER is a supervised learning task based on a clean dataset. We focus on the practical scenario where the training labels are noisy due to distant supervision, i.e., the revealed tag y_i may not correspond to the underlying correct one. Thus, the challenge of DS-NER is to reduce the negative effect of noisy annotations.

4 Methodology

As shown in Figure 2, CENSOR consists of two teacher-student networks to handle the noisy label. To avoid overfitting the incorrect pseudo labels generated by poorly calibrated teacher networks, we introduce Uncertainty-Aware Teacher Learning that leverages the prediction uncertainty to guide the label selection. We also propose Student-Student Collaborative Learning that allows reliable label transfer between two student networks, further reducing the risk of learning incorrect pseudo labels and making a full use of mislabeled samples.

4.1 Teacher-student Framework

Neural networks excel at memorization (Arpit et al., 2017). However, when noisy labels become prominent, deep-learning-based NER models inevitably overfit noisy labeled data, resulting in poor performance. The purpose of the teacher-student methods is to select reliable labels (i.e., pseudo labels that are more likely to be labeled correctly), to reduce the negative effect of label noise. Self-training involves the teacher-student network, where the teacher network first generates pseudo labels to participate in label selection. Then the student is optimized via back-propagation based on selected reliable labels, and the teacher is updated by gradually shifting the weights of the student with an exponential moving average (EMA). Following Qu et al. (2023), we train two sets of teacher-student networks using two different NER models to provide multi-view predictions on training samples.

4.2 Uncertainty-Aware Teacher Learning

In the DS-NER task, one of the main challenges of the teacher-student framework is to evaluate the correctness of the generated pseudo labels of the teacher model. Previous methods (Liang et al., 2020; Zhang et al., 2021a; Qu et al., 2023) generally assume that high-confidence predictions tend to be correct. Therefore, they select the samples with high-confidence pseudo labels (e.g., pseudo labels with confidence values greater than 0.7) as training data. However, the teacher network is prone to generating high-confidence yet incorrect pseudo labels due to the poor calibration (Guo et al., 2017). This overconfidence is indicative of model bias rather than the true likelihood of correctness. Therefore, relying solely on the teacher network's confidence as the indicator may not efficiently evaluate the correctness of the pseudo labels.

Meanwhile, we observe that when the NER model performs supervised learning on a mislabeled token, it receives two types of supervision from the incorrect label of the mislabeled token and the labels of semantically similar but correctly labeled tokens. For example, "Washington" in Figure 1 is mislabeled as "LOC" (location), and the model trained with it tends to predict "Washington" as "LOC" instead of "PER" (person). The model is also exposed to semantically similar but correctly labeled tokens, such as the token "James" labeled as "PER" in the training sentence "U.S. President will meet James at the White House", thus the model may also learn to generalize "Washington" as a "PER". The knowledge in both types of supervision is eventually learned and saved to the network neurons. However, as the training continues, the deep-learning-based model inevitably overfits the noisy labels due to its memorization capability (Arpit et al., 2017), rather than utilizing the correct knowledge learned from the labels of semantically similar but correctly labeled tokens.

Uncertainty Estimation Based on our observation, we find that randomly deactivating neurons introduces variability in predicted confidence of the incorrect pseudo label, which can be attributed to varying subsets of active neurons influencing each prediction. Specifically, the randomness of deactivation of the network neurons makes the remaining network neurons sometimes retain more knowledge learned from the incorrect label of the mislabeled token, and sometimes retain more knowledge learned from the labels of semantically similar but correctly labeled tokens. Consequently, such discrepancies can lead to inconsistencies in multiple predictions. For the correctly labeled tokens, since their labels are the same as those of semantically similar tokens, the two types of knowledge stored in the network neurons are more consistent, so the predictions from the different subsets of active neurons tend to be more consistent. Thus, we define the inconsistency of predictions from sampled teacher network neurons as uncertainty and evaluate the correctness of the generated pseudo labels.

Specifically, given the new input token x^* and the pseudo label $\hat{y}^*$ generated by the teacher network W, we perform K forward passes with Dropouts (Krizhevsky et al., 2012) through our teacher networks at inference time. In each pass, pre-defined parts of network neurons are randomly deactivated. Then, we could yield K subsets of active neurons $\{\hat{W}_1, \hat{W}_2, ..., \hat{W}_K\}$. To estimate the uncertainty for each token in the sequence labeling task, we leverage the variance of the model outputs for each token from multiple forward passes:

$$s_{un}(y^* = \hat{y}^*|W, x^*) = Var[p(y^* = \hat{y}^*|\hat{W}_k, x^*)]_{k=1}^{K}, \quad (1)$$

where $Var[.]$ is the variance of distribution over the K passes through the teacher network. The lower uncertainty indicates the predictions from sampled teacher network neurons and the learned knowledge are more consistent, thus the pseudo label is more likely to be correct.

Uncertainty-Aware Label Selection Different from previous teacher-student methods only using confidence as the indicator to select reliable pseudo labels, we jointly consider the confidence and uncertainty in label selection. For the confidence of the pseudo label $\hat{y}^*$, as follows:

$$\begin{gathered}\hat{y}^* = argmax(p(y^*|W, x^*)) \\ s_{co}(y^* = \hat{y}^*|W, x^*) = p(y^* = \hat{y}^*|W, x^*)\end{gathered} \quad (2)$$

A higher confidence value s_{co} means the model is more confident for the pseudo label $\hat{y}^*$. However, many of these selected pseudo labels with high confidence are also incorrect due to the poorly calibrated teacher network (Guo et al., 2017), leading to error propagation in the self-training. To reduce the effect of incorrect pseudo labels, we additionally use uncertainty score s_{un} as the indicator. Specifically, we select a subset of pseudo labels which are both high-confidence and low-uncertainty as reliable labels, since jointly considering confidence and uncertainty can further filter the incorrect pseudo labels with high confidence. Thus, we define a masked matrix, i.e.,

$$M_{x^*} = \begin{cases} 1 & s_{un} < \sigma_{ua} \quad and \quad s_{co} > \sigma_{co}; \\ 0 & Otherwise; \end{cases} \quad (3)$$

When $M = 0$, it means the pseudo-label may be incorrect and the sample should be masked in the self-training. σ_{co} and σ_{ua} are hyperparameters.

4.3 Student-Student Collaborative Learning

Based on Uncertainty-Aware Teacher Learning, the teacher network can utilize the correctly pseudo-labeled samples to alleviate the negative effect of label noise. However, simply masking unreliable pseudo-labeled samples can lead to underutilization of the training set, as there is no chance for the incorrect pseudo-labeled samples to be corrected and further learned. Intuitively, if we can correct the incorrect pseudo label with the correct one, it will become a useful training sample. Therefore, to address these shortcomings and incorporate Uncertainty-Aware Teacher Learning to make the teacher-student network more effective, we propose Student-Student Collaborative Learning.

The idea of Student-Student Collaborative Learning is to utilize two different student networks and let them learn from each other. We regard small-loss samples as clean samples for training, in each batch of data, each student network views its small-loss pseudo labels (e.g., pseudo labels of 10% samples with the smallest loss) as the reliable labels, and transfers such reliable labels to another student network for updating the parameters. These small-loss samples are far from the decision boundaries of the two models and thus are more likely to be true positives and true negatives (Feng et al., 2019). In this way, a student network is able to not completely rely on all pseudo labels from the teacher network, further reducing the risk of learning incorrect pseudo labels generated by the poorly calibrated teacher network. Moreover, the two different student networks may have different decision boundaries and thus are good at recognizing different patterns in data. Different from simply masking unreliable pseudo-labeled samples, this component also provides the opportunity for the incorrect pseudo-labeled samples to be correctly labeled by the other teacher-student network to make full use of the training data.

Specifically, for two student networks s_1, s_2 and their parameters W_{s_1}, W_{s_2}, we first let s_1 (resp. s_2) select a small ratio of samples in this batch of data $\hat{D}$ that have small training loss. For these selected samples $\hat{D}_{s_1}$ (resp. $\hat{D}_{s_2}$) from s_1 (resp. s_2), we use the corresponding generated pseudo labels $\hat{Y}_{s_1}$ (resp. $\hat{Y}_{s_2}$) as reliable labels and transfer such reliable labels to the other student network s_2 (resp. s_1) for updating the parameters W_2 (resp. W_1). The ratio of transferred labels is controlled by hyperparameter δ. In this way, two student networks can learn from each other's reliable labels, reducing the risk of learning from incorrect pseudo labels and making full use of the training data.

4.4 Training and Inference

Algorithm 1 in Appendix A.3 gives the pseudocode. The process can be divided into three stages: the pre-training, the self-training, and the inference.

Pre-Training Stage We warm up two different NER models W_A and W_B on the noisy DS-NER dataset to obtain a better initialization, and then duplicate the parameters W for both the teacher W_t and the student W_s (i.e., $W_{t_1}= W_{s_1}= W_A$, $W_{t_2}= W_{s_2}= W_B$). The training objective function is the cross entropy loss with the following form:

$$\mathcal{L} = -\frac{1}{N}\sum_{D_{ds}} y_i \log(p(y_i|W_s, x_i)) \tag{4}$$

where y_i means the i-th token label of the i-th token x_i in the DS-NER corpus D_{ds} and $p(y_i|W_s, x_i)$ denotes its probability produced by student network W_s. N is the size of the training corpus.

Self-Training Stage In this stage, we select reliable pseudo-labeled tokens to train the two teacher-student networks respectively. Specifically, we select reliable labels generated by teachers W_t and supervise the students W_s with cross-entropy loss. During the label selection, we use the proposed Uncertainty-Aware Label Selection to jointly consider the confidence and uncertainty as shown in Eq. 3 to reduce the effect of incorrect pseudo-labeled samples. Meanwhile, we use Student-Student Collaborative Learning to allow student networks can learn from each other's reliable labels by selecting the pseudo labels from small-loss samples. Therefore, the training objective function of student networks W_s in this stage is the cross entropy loss with the following form:

$$\mathcal{L} = -\frac{1}{N}\sum_{D_{ds}} M_i \hat{y}_i \log(p(\hat{y}_i|W_s, x_i)) \tag{5}$$

where $\hat{y}_i$ means the i-th pseudo-label generated by Student-Student Collaborative Learning and its teacher W_t. $p(\hat{y}_i|W_s, x_i)$ denotes its probability produced by student network W_s on generated pseudo-label. M_i is indicator where the i-th token x_i should be masked according to Eq. 3. Meanwhile, if $\hat{y}_i$ is the transferred pseudo-label from the other student, M_i will be automatically set to 1 (unmasked). That is, we are more inclined to

Method	CoNLL03			OntoNotes5.0			Webpage			Wikigold			Twitter		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
KB-Matching	81.13	63.75	71.40	63.86	55.71	59.51	62.59	45.14	52.45	47.90	47.63	47.76	40.34	32.22	35.83
BiLSTM-CRF	75.50	49.10	59.50	**68.44**	64.50	66.41	58.05	34.59	43.34	47.55	39.11	42.92	46.91	14.18	21.77
DistilRoBERTa	77.87	69.91	73.68	66.83	68.81	67.80	56.05	59.46	57.70	48.85	52.05	50.40	45.72	43.85	44.77
RoBERTa	82.29	70.47	75.93	66.99	69.51	68.23	59.24	62.84	60.98	47.67	58.59	52.57	50.97	42.66	46.45
AutoNER	75.21	60.40	67.00	64.63	69.95	67.18	48.82	54.23	51.39	43.54	52.35	47.54	43.26	18.69	26.10
LRNT	79.91	61.87	69.74	67.36	68.02	67.69	46.70	48.83	47.74	45.60	46.84	46.21	46.94	15.98	23.84
Co-teaching+	86.04	68.74	76.42	66.63	69.32	67.95	61.65	55.41	58.36	55.23	49.26	52.08	51.67	42.66	46.73
JoCoR	83.65	69.69	76.04	66.74	68.74	67.73	62.14	58.78	60.42	51.48	51.23	51.35	49.40	45.59	47.42
NegSampling	80.17	77.72	78.93	64.59	**72.39**	68.26	70.16	58.78	63.97	49.49	55.35	52.26	50.25	44.95	47.45
BOND	82.05	80.92	81.48	67.14	69.61	68.35	67.37	64.19	65.74	53.44	**68.58**	60.07	53.16	43.76	48.01
SCDL	**87.96**	79.82	83.69	67.49	69.77	68.61	68.71	68.24	68.47	62.25	66.12	64.13	**59.87**	44.57	51.09
ATSEN	85.75	83.86	84.79	65.69	70.71	68.11	71.08	70.03	70.55	57.67	54.71	56.15	59.31	45.83	51.71
CENSOR	87.33	**85.90**	**86.61**	67.11	71.01	**69.01**	**75.89**	**72.30**	**74.05**	**66.01**	68.10	**67.05**	58.63	**47.38**	**52.41**

Table 1: Main results on five DS-NER datasets. We report the baseline results from Liang et al. (2020); Zhang et al. (2021a) and our experimental results with their official implementation in our devices.

trust judgments from the student model because the student network is updated earlier and more frequently than the teacher network, and therefore better able to capture the changes of pseudo labels. N is the size of the training corpus.

Different from the optimization of the student network, we apply EMA as Zhang et al. (2021a) to gradually update the parameters of the teacher:

$$W_t \leftarrow \alpha W_t + (1-\alpha) W_s \tag{6}$$

where α denotes the smoothing coefficient. With the conservative and ensemble properties, the usage of EMA has largely mitigated the bias. As a result, the teacher tends to generate more reliable pseudo labels, which can be used as new supervision signals in the denoising self-training stage.

Inference Stage In the inference stage, only the best model $W_{best} \in \{W_{t_1}, W_{s_1}, W_{t_2}, W_{s_2}\}$ on the dev set is adopted for predicting the test data.

5 Experiment

5.1 Dataset

We conduct experiments on five DS-NER datasets, including CoNLL03 (Tjong Kim Sang and De Meulder, 2003), Webpage (Ratinov and Roth, 2009), Wikigold (Balasuriya et al., 2009), Twitter (Godin et al., 2015) and OntoNotes5.0 (Weischedel et al., 2013). For the fair comparison, we follow the same knowledge bases and settings as Liang et al. (2020), re-annotate the training set by distant supervision, and use the original dev and test set. Statistics of datasets are shown in Appendix A.1.

5.2 Evaluation Metrics and Baselines

We use Precision (P), Recall (R), and F1 score as our evaluation metrics. We compare CENSOR with various baseline methods, including supervised methods and DS-NER methods. We also present the results of **KB-Matching**, which directly uses knowledge bases to annotate the test sets.

Supervised Methods We select **BiLSTM-CRF** (Ma and Hovy, 2016), **RoBERTa** (Liu et al., 2019) and **DistilRoBERTa** (Sanh et al., 2019) as original supervised methods. As trained on noisy DS-NER datasets, these methods achieve poor performance.

DS-NER Methods We compare several DS-NER baselines. **AutoNER** (Shang et al., 2018) modifies the standard CRF to get better performance under the noise. **LRNT** (Cao et al., 2019) leaves training data unexplored fully to reduce the negative effect of noisy labels. **Co-teaching+** (Yu et al., 2019) and **JoCoR** (Wei et al., 2020) are two classical collaborative learning methods to handle noisy labels in computer vision area. **NegSampling** (Li et al., 2021) uses down-sampling in non-entities to relief the misleading from incomplete annotation.

Teacher-Student Methods for DS-NER Specifically, **BOND** (Liang et al., 2020) designs a teacher-student network and selects high-confidence predictions as pseudo labels to get a robust model. **SCDL** (Zhang et al., 2021b) improves the performance by training two teacher-student networks and selecting consistent high-confidence predictions between two teachers as pseudo labels. **ATSEN** (Qu

Method	P	R	F1
CENSOR	**87.33**	**85.90**	**86.61**
-w/o UTL	86.56 (-0.77)	84.37 (-1.53)	85.45 (-1.16)
-w/o SCL	86.44 (-0.89)	83.98 (-1.92)	85.19 (-1.42)

Table 2: Ablation study on CoNLL03. UTL means Uncertainty-Aware Teacher Learning and SCL means Student-Student Collaborative Learning.

et al., 2023) considers both consistent and inconsistent predictions with high confidence between two teachers and further proposes a fine-grained teacher updating method. We report the results of ATSEN with official implementation in our devices.

5.3 Experimental Settings

Following Qu et al. (2023), we adopt RoBERTa-base and DistilRoBERTa-base as two NER models for two teacher-student networks. We use Adam (Kingma and Ba, 2015) as our optimizer. We list detailed hyperparameters in the Appendix A.2.

5.4 Main Results

Table 1 presents the performance of different methods measured by precision, recall, and F1 score. Specifically, (1) CENSOR achieves new SOTA performance, showing superiority in the DS-NER task; (2) Compared to original supervised methods, including BiLSTM-CRF, RoBERTa, and DistilRoBERTa, CENSOR improves the F1 score with an average increase of 23.04%, 10.96%, and 8.99%, respectively, which demonstrates the necessity of DS-NER models and the effectiveness; (3) Compared to classical de-noising methods in the computer vision area (e.g., Co-teaching+), simply using these methods can not achieve strong performance, since these methods were not initially designed for sequence labeling tasks and ignore the characteristics of the DS-NER task. (4) Compared with teacher-student methods such as BOND, SCDL, and ATSEN, CENSOR achieves advanced performance, confirming that these teacher-student methods achieve limited performance because of the incorrect pseudo-labeled samples.

5.5 Analysis

Ablation Study Shown in Table 2, it is clear that Uncertainty-Aware Teacher Learning and Student-Student Collaborative Learning are both important to the model performance. Removing each component can lead to a simultaneous decrease in precision and recall at the same time, showing that proposed components indeed improve performance.

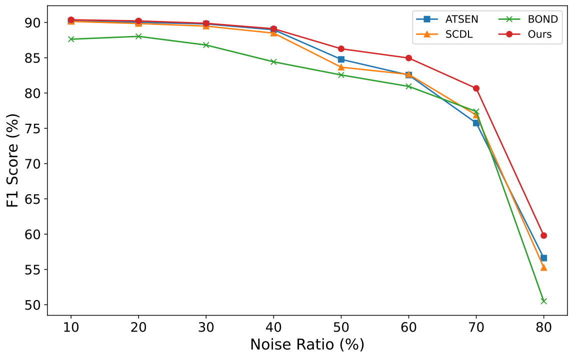

Figure 3: F1 on CoNLL03 with different noise ratios.

Method	P	R	F1
BOND	80.87 (-13.49)	78.04 (- 7.09)	79.43 (-10.08)
SCDL	94.18 (- 0.18)	77.11 (- 8.02)	84.80 (- 4.71)
ATSEN	93.01 (- 1.35)	82.96 (- 2.17)	87.70 (- 1.87)
CENSOR	**94.36**	**85.13**	**89.51**

Table 3: Comparison of the effectiveness of reducing label noise on CoNLL03.

Robustness to Different Noise Ratios To investigate the robustness of the CENSOR in different noise ratios, we randomly replace $k\%$ entity labels in the clean version (instead of the distantly-supervised version) of CoNLL03 training set with other entity types or non-entity. In this way, we can construct different noise ratios of label noise and we further report the test F1 score on CoNLL03. As shown in Figure 3, CENSOR achieves consistent advanced performance in different noise ratios, showing its satisfactory de-noising ability and strong robustness. Meanwhile, when the noise ratio is above 50%, CENSOR achieves more significant robustness, since CENSOR can select and generate more reliable labels due to the Uncertainty-Aware Teacher Learning and Student-Student Collaborative Learning from highly noisy data. More detailed data can be found in Table 7 in the Appendix.

Effectiveness of Reducing Learned Noise To confirm previous teacher-student methods achieve limited performance because of incorrectly pseudo-labeled samples, we try to explore the effectiveness of reducing label noise from different teacher-student methods, including CENSOR, BOND, SCDL, ATSEN. Specifically, we report the average F1 score of all selected (unmasked) pseudo labels for training during the self-training stage, using the labels from the clean version of the CoNLL03 train-

Method	P	R	F1
BOND	80.42 (-9.44)	76.46 (-8.69)	78.39 (-9.05)
SCDL	87.42 (-2.44)	75.85 (-9.30)	81.22 (-6.22)
ATSEN	87.84 (-2.02)	82.83 (-2.32)	85.26 (-2.18)
CENSOR	**89.86**	**85.15**	**87.44**

Table 4: Comparison of teacher pseudo-labeling ability of different teacher-student methods on CoNLL03.

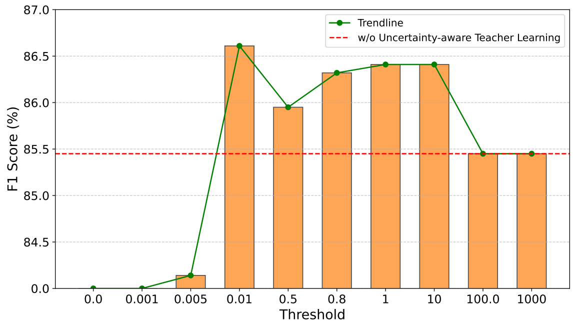

Figure 4: F1 on CoNLL03 with different threshold σ_{ua} in Uncertainty-Aware Label Selection.

ing set as ground truth labels. As shown in Table 3, CENSOR achieves a consistent advanced F1 score, which indicates CENSOR can select more correct labels based on Uncertainty-Aware Label Selection and Student-Student Collaborative Learning. Thus, CENSOR can use more correct pseudo labels to update the parameters of student networks and further avoid error propagation, leading to outstanding overall performance on the test set.

Effectiveness of Teacher Pseudo-labeling After confirming the effectiveness of reducing label noise, we attempt to further explore whether the teacher network could use more reliable labels to avoid error propagation, thus generating more correct pseudo labels. As shown in Table 4, we report the best F1 score of teacher networks from different teacher-student methods on the clean version of CoNLL03 training set. In detail, the teacher network from CENSOR correctly labels 87.44% samples, achieving the most advanced precision, recall, and F1 score. Compared to other teacher-student methods, including BOND, SCDL, and ATSEN, CENSOR improves the F1 score with an average increase of 9.05%, 6.22%, and 2.18%, respectively, which demonstrates using more correct labels can avoid error propagation and make the teacher network generate more reliable labels. In this way, the teacher network can make full use of the noisy samples in the DS-NER training set and help the teacher-student framework achieve outstanding performance on the test set.

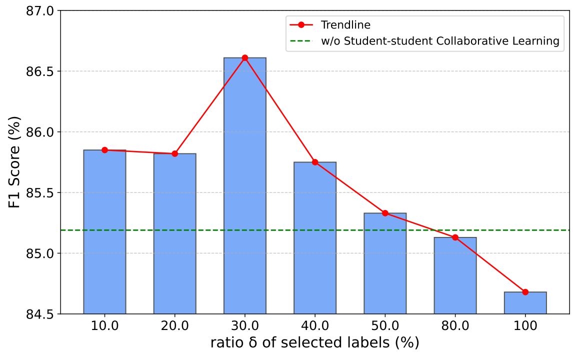

Figure 5: F1 on CoNLL03 with different ratio δ of selected labels in Student-student Collaborative Learning.

Parameter Study As shown in Figure 4 and Figure 5, we conduct experiments to explore the impact of important hyperparameters to further understand Uncertainty-Aware Label Selection and Student-Student Collaborative Learning. Overall, although the choice of different hyperparameters will have some impact on the model performance, as long as the hyperparameters are chosen wisely rather than at extreme values (e.g., wrongly setting the threshold σ_{ua} in Uncertainty-Aware Label Selection to 0), the performance of the model will always be improved over what it would have been without using the components. More detailed analysis are shown in the Appendix A.5.

Case Study We also conduct a case study to show the advantages of the proposed CENSOR, which can be found in Appendix A.6. We show the prediction of several teacher-student methods for DS-NER, including BOND, SCDL, and ATSEN.

6 Conclusion

We introduce CENSOR, a novel teacher-student framework designed for DS-NER task. CENSOR incorporates Uncertainty-Aware Teacher Learning, utilizing prediction uncertainty to guide the pseudo-label selection. It mitigates the usage of incorrect pseudo labels by avoiding reliance on confidence scores from poorly calibrated teacher networks. We also introduce Student-Student Collaborative Learning to enable a student network not to completely rely on pseudo labels from its teacher network, minimizing the risk of learning incorrect ones. Meanwhile, this component allows the training set can be fully explored. Our experimental results demonstrate CENSOR's superior performance compared to previous methods.

Limitations

Our proposed CENSOR has two tiny limitations, specifically: (1) CENSOR focuses on addressing the label noise in the DS-NER task, and all our analyses are specific to this task. As a result, our model may not be robust enough compared to other models if it is not specific to the DS-NER task. (2) Due to introducing the proposed Uncertainty-Aware Teacher Learning, our model will perform multiple forward passes in the uncertainty estimation phase, increasing the self-training time. Compared to ATSEN, the self-training of our model takes about 4 times as long as that of ATSEN.

References

Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron C. Courville, Yoshua Bengio, and Simon Lacoste-Julien. 2017. A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 233–242. PMLR.

Dominic Balasuriya, Nicky Ringland, Joel Nothman, Tara Murphy, and James R. Curran. 2009. Named entity recognition in wikipedia. In *Proceedings of the 1st 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources@IJCNLP 2009, Suntec, Singapore, August 7, 2009*, pages 10–18. Association for Computational Linguistics.

Yixin Cao, Zikun Hu, Tat-Seng Chua, Zhiyuan Liu, and Heng Ji. 2019. Low-resource name tagging learned with weakly labeled data. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 261–270. Association for Computational Linguistics.

Jiazhan Feng, Chongyang Tao, Wei Wu, Yansong Feng, Dongyan Zhao, and Rui Yan. 2019. Learning a matching model with co-teaching for multi-turn response selection in retrieval-based dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3805–3815, Florence, Italy. Association for Computational Linguistics.

Fréderic Godin, Baptist Vandersmissen, Wesley De Neve, and Rik Van de Walle. 2015. Multimedia lab \$@\$ ACL WNUT NER shared task: Named entity recognition for twitter microposts using distributed word representations. In *Proceedings of the Workshop on Noisy User-generated Text, NUT@IJCNLP 2015, Beijing, China, July 31, 2015*, pages 146–153. Association for Computational Linguistics.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 8536–8546.

Xinyue Huo, Lingxi Xie, Jianzhong He, Zijie Yang, Wengang Zhou, Houqiang Li, and Qi Tian. 2021. ATSO: asynchronous teacher-student optimization for semi-supervised image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 1235–1244. Computer Vision Foundation / IEEE.

Zhanming Jie, Pengjun Xie, Wei Lu, Ruixue Ding, and Linlin Li. 2019. Better modeling of incomplete annotations for named entity recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 729–734, Minneapolis, Minnesota. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 1106–1114.

Yangming Li, Lemao Liu, and Shuming Shi. 2021. Empirical analysis of unlabeled entity problem in named entity recognition. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Yangming Li, Lemao Liu, and Shuming Shi. 2022. Rethinking negative sampling for handling missing entity annotations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7188–7197, Dublin, Ireland. Association for Computational Linguistics.

Yiyang Li and Hai Zhao. 2023. EM pre-training for multi-party dialogue response generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 92–103, Toronto, Canada. Association for Computational Linguistics.

Chen Liang, Yue Yu, Haoming Jiang, Siawpeng Er, Ruijia Wang, Tuo Zhao, and Chao Zhang. 2020. BOND: bert-assisted open-domain named entity recognition with distant supervision. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 1054–1064. ACM.

Yajiao Liu, Xin Jiang, Yichun Yin, Yasheng Wang, Fei Mi, Qun Liu, Xiang Wan, and Benyou Wang. 2023. One cannot stand for everyone! leveraging multiple user simulators to train task-oriented dialogue systems. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–21, Toronto, Canada. Association for Computational Linguistics.

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*.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Minlong Peng, Xiaoyu Xing, Qi Zhang, Jinlan Fu, and Xuanjing Huang. 2019. Distantly supervised named entity recognition using positive-unlabeled learning. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2409–2419. Association for Computational Linguistics.

Xiaoye Qu, Jun Zeng, Daizong Liu, Zhefeng Wang, Baoxing Huai, and Pan Zhou. 2023. Distantly-supervised named entity recognition with adaptive teacher learning and fine-grained student ensemble. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 13501–13509. AAAI Press.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 147–155, Boulder, Colorado. Association for Computational Linguistics.

Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S. Rawat, and Mubarak Shah. 2021. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Jingbo Shang, Liyuan Liu, Xiaotao Gu, Xiang Ren, Teng Ren, and Jiawei Han. 2018. Learning named entity tagger using domain-specific dictionary. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2054–2064, Brussels, Belgium. Association for Computational Linguistics.

Shuzheng Si, Zefan Cai, Shuang Zeng, Guoqiang Feng, Jiaxing Lin, and Baobao Chang. 2023. SANTA: Separate strategies for inaccurate and incomplete annotation noise in distantly-supervised named entity recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3883–3896, Toronto, Canada. Association for Computational Linguistics.

Shuzheng Si, Shuang Zeng, Jiaxing Lin, and Baobao Chang. 2022. SCL-RAI: Span-based contrastive learning with retrieval augmented inference for unlabeled entity problem in NER. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2313–2318, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. 2020. Combating noisy labels by agreement: A joint training method with co-regularization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 13723–13732. Computer Vision Foundation / IEEE.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 23.

YaoSheng Yang, Wenliang Chen, Zhenghua Li, Zhengqiu He, and Min Zhang. 2018. Distantly supervised NER with partial annotation learning and reinforcement learning. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 2159–2169. Association for Computational Linguistics.

Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor W. Tsang, and Masashi Sugiyama. 2019. How does disagreement help generalization against label corruption? In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 7164–7173. PMLR.

Xinghua Zhang, Bowen Yu, Tingwen Liu, Zhenyu Zhang, Jiawei Sheng, Xue Mengge, and Hongbo Xu. 2021a. Improving distantly-supervised named entity recognition with self-collaborative denoising learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1518–1529, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhenyu Zhang, Bowen Yu, Xiaobo Shu, Xue Mengge, Tingwen Liu, and Li Guo. 2021b. From what to why: Improving relation extraction with rationale graph. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 86–95, Online. Association for Computational Linguistics.

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

A Appendix

A.1 DS-NER Datasets

Statistics of five datasets are shown in Table 5.

Dataset		Train	Dev	Test	Types
CoNLL03	**Sentence**	14041	3250	3453	4
	Token	203621	51362	46435	
OntoNotes5.0	**Sentence**	115812	15680	12217	18
	Token	2200865	304701	230118	
Webpage	**Sentence**	385	99	135	4
	Token	5293	1121	1131	
Wikigold	**Sentence**	1142	280	274	4
	Token	25819	6650	6538	
Twitter	**Sentence**	2393	999	3844	10
	Token	44076	15262	58064	

Table 5: The statistics of five DS-NER datasets.

A.2 Hyperparameters

Detailed hyperparameters are shown in Table 6. Experiments are run on a single NVIDIA A40.

Algorithm 1 Training Procedure of CENSOR.

Input: DS-NER dataset $D_{ds} = \{(X_i, Y_i)\}_{i=1}^{N}$
Parameter: Two teacher-student network parameters, including W_{t_1}, W_{s_1}, W_{t_2}, and W_{s_2}
Output: The best model

```
Pre-training two models W_A, W_B with D_ds.                ▷Pre-Training.
Initialize two teacher-student networks: W_t1 ← W_A, W_s1 ← W_A,
    W_t2 ← W_B, W_s2 ← W_B.
Initialize training step: step ← 0.
Initialize noisy labels: Y_I ← Y, Y_II ← Y.
while not reach max training epochs do
    Get a batch D̂ = (X^(b), Y_I^(b), Y_II^(b)) from D_ds,
    step ← step + 1.                                       ▷Self-Training.
    Get pseudo labels via the teacher W_t1, W_t2:
        Ỹ_I^(b) ← f(X^(b); W_t1),
        Ỹ_II^(b) ← f(X^(b); W_t2).
    Select reliable labels via Uncertainty-Aware Teacher Learning:
        Estimate Confidence and Uncertainty by Eq.3 and Eq.4, separately
        T_I^(b) ← Uncertainty-Aware Label Selection(Y_I^(b), Ỹ_I^(b)),
        T_II^(b) ← Uncertainty-Aware Label Selection(Y_II^(b), Ỹ_II^(b)).
    Select reliable labels via Student-Student Collaborative Learning:
        D̂*_s1 = arg min_{D̂:|D̂|≥δ%|D̂|} Loss(s1, D̂),
                                  //sample δ% small-loss instances
        D̂*_s2 = arg min_{D̂:|D̂|≥δ%|D̂|} Loss(s2, D̂).
                                  //sample δ% small-loss instances
        Transfer the pseudo labels between D̂*_s1 and D̂*_s2 .
    Update the student W_s1 and W_s2 by Eq. 7.
    Update the teacher W_t1 and W_t2 by Eq. 8.
end while
Evaluate models W_t1, W_s1, W_t2, W_s2 on Dev set.
return The best model W ∈ {W_t1, W_s1, W_t2, W_s2}
```

A.3 Pseudocode

Algorithm 1 gives the pseudocode of our method.

A.4 Robustness to Different Noise Ratios

Detailed data in Figure 3 can be found in Table 7.

A.5 Parameter Study

In Figure 4 and Table 8, we analyze the impact of σ_{ua} in Eq.3 within Uncertainty-Aware Label Selection. Notably, for minimal values of σ_{ua}, such as 0 and 0.001, the Uncertainty-Aware Label Selection phase filters and masks all samples. Consequently, the student network becomes incapable of parameter updates, rendering the entire teacher-student framework non-trainable. When the parameter σ_{ua} is in a reasonable interval, the effectiveness of the model is always improved due to the inclusion of filtered reliable labels in the self-training stage. Ultimately, when σ_{ua} reaches an excessive magnitude, the filtering capacity of the Uncertainty-Aware Label Selection stage is nullified, rendering the outcome akin to Uncertainty-Aware Teacher Learning omission. Therefore, while using different values of σ_{ua} tends to improve the performance, choosing σ_{ua} wisely and rationally is crucial for optimizing Uncertainty-Aware Teacher Learning. In Figure 5 and Table 9, we also explore the impact of the ratio δ of selected labels in Student-Student Col-

Name	CoNLL03	Ont5.0	Webpage	Wikigold	Twitter
Learning Rate	1e-5	2e-5	1e-5	1e-5	2e-5
Batch Size	8	16	16	16	8
EMA α	0.995	0.995	0.99	0.99	0.995
Sche. Warmup	200	500	100	200	200
Total Epoch	50	50	50	50	50
Pre-training Epoch	1	2	12	5	6
σ_{co} in Eq.5 of UTL	0.9	0.9	0.9	0.9	0.9
σ_{ua} in Eq.5 of UTL	0.01	0.05	0.1	0.2	0.2
K in Eq.2 of UTL	8	8	8	8	8
Dropout Rate	0.5	0.5	0.5	0.5	0.5
ratio δ of SCL	0.3	0.4	0.3	0.1	0.1
Update Cycle (iterations)	6000	7240	300	2000	3200

Table 6: Hyperparameters on five DS-NER datasets. UTL means Uncertainty-Aware Teacher Learning and SCL means Student-Student Collaborative Learning.

Ratio	ATSEN	SCDL	BOND	Ours
10%	90.19	90.15	87.63	**90.38**
20%	90.03	89.85	88.03	**90.22**
30%	89.79	89.48	86.80	**89.88**
40%	88.97	88.49	84.42	**89.11**
50%	84.77	83.66	82.56	**86.27**
60%	82.55	82.64	80.94	**84.96**
70%	75.75	76.88	77.38	**80.66**
80%	56.61	55.26	50.49	**59.80**
90%	19.59	17.09	14.85	**22.26**

Table 7: F1 on CoNLL03 with different noise ratios.

θ_{ua}	P	R	F1
-w/o UTL	86.56	84.37	85.45
0.000	00.00	00.00	00.00
0.001	00.00	00.00	00.00
0.005	85.65	82.68	84.14
0.010	**87.33**	**85.90**	**86.61**
0.500	87.22	84.71	85.95
0.800	87.60	85.06	86.32
1.000	87.27	85.56	86.41
10.00	87.27	85.56	86.41
100.0	86.56	84.37	85.45
1,000	86.56	84.37	85.45

Table 8: F1 on CoNLL03 with different threshold σ_{ua} in Uncertainty-Aware Label Selection. UTL means Uncertainty-Aware Teacher Learning.

K	P	R	F1
-w/o SCL	86.44	83.98	85.19
0.1	86.81	84.92	85.85
0.2	**87.35**	84.33	85.82
0.3	87.33	**85.90**	**86.61**
0.4	86.95	84.58	85.75
0.5	86.28	84.41	85.33
0.8	86.27	84.01	85.13
1.0	85.70	83.68	84.68

Table 9: F1 on CoNLL03 with different ratio δ of selected labels in Student-Student Collaborative Learning. SCL means Student-Student Collaborative Learning.

laborative Learning. A small δ enables the student network to partially leverage reliable labels from its counterpart, resulting in improved outcomes compared to scenarios without such collaborative learning. As δ increases, the transfer of these reliable labels diminishes the likelihood of learning incorrect labels from teacher-generated pseudo labels, thereby enhancing overall performance. Conversely, an excessively large δ adversely affects performance. This is attributed to the pseudo labels of selected samples, which, with a high transfer proportion (e.g., $\delta = 0.8$), cease to qualify as small-loss samples and are more prone to containing noise. Hence, proportion selection of δ proves critical for optimizing the efficacy of Student-Student Collaborative Learning.

A.6 Case Study

Finally, we perform case study to understand the advantage **CENSOR** with two examples in Table 10 and Table 11. We show the prediction of BOND, SCDL, ATSEN and CENSOR on a training sequence with label noise and a test sequence with ground truth. As shown in Table 10, BOND and SCDL can slightly generalize to unseen mentions and relieve partial incomplete annotation, e.g., they can successfully recognize the “John McNamara" and “New York”. However, these methods still suffer from label noise. For comparison, for hard labels “California Angels", CENSOR and ATSEN are able to detect them with advanced teacher-student design (e.g., Adaptive Teacher Learning in ATSEN and Student-Student Collaborative Learning in CENSOR) instead of relying purely on distant labels. However, as shown in Table 11, ATSEN still struggles to distinguish between easily confused samples and achieves inadequate generalization. In contrast, as CENSOR can use fewer incorrect pseudo-labeled samples due to Uncertainty-Aware Teacher Learning and Student-Student Collaborative Learning, a higher degree of robustness and generalization can be achieved.

A.7 Difference between Previous Methods

We will carefully compare previous methods to explain our motivation and the differences between previous methods and our proposed components.

Distant Match: [Johnson]$_{\text{PER}}$ is the second manager to be hospitalized after California [Angels]$_{\text{PER}}$ skipper [John]$_{\text{PER}}$ McNamara was admitted to New [York]$_{\text{PER}}$'s [Columbia]$_{\text{PER}}$ Presby Hospital .
Ground Truth: [Johnson]$_{\text{PER}}$ is the second manager to be hospitalized after [California Angels]$_{\text{ORG}}$ skipper [John McNamara]$_{\text{PER}}$ was admitted to [New York]$_{\text{LOC}}$'s [Columbia Presby Hospital]$_{\text{ORG}}$.
BOND: [Johnson]$_{\text{PER}}$ is the second manager to be hospitalized after [California]$_{\text{LOC}}$ [Angels]$_{\text{PER}}$ skipper [John McNamara]$_{\text{PER}}$ was admitted to [New York]$_{\text{LOC}}$'s [Columbia]$_{\text{PER}}$ Presby Hospital.
SCDL: [Johnson]$_{\text{PER}}$ is the second manager to be hospitalized after [California]$_{\text{LOC}}$ [Angels]$_{\text{PER}}$ skipper [John McNamara]$_{\text{PER}}$ was admitted to [New York]$_{\text{LOC}}$'s [Columbia Presby Hospital]$_{\text{ORG}}$.
ATSEN: [Johnson]$_{\text{PER}}$ is the second manager to be hospitalized after [California Angels]$_{\text{ORG}}$ skipper [John McNamara]$_{\text{PER}}$ was admitted to [New York]$_{\text{LOC}}$'s [Columbia Presby Hospital]$_{\text{ORG}}$.
CENSOR: [Johnson]$_{\text{PER}}$ is the second manager to be hospitalized after [California Angels]$_{\text{ORG}}$ skipper [John McNamara]$_{\text{PER}}$ was admitted to [New York]$_{\text{LOC}}$'s [Columbia Presby Hospital]$_{\text{ORG}}$.

Table 10: Case study with CENSOR and previous teacher-student methods for DS-NER. The sentence is from CoNLL03 training set.

Ground Truth: All-conquering [Juventus]$_{\text{ORG}}$ field their most recent signing, [Portuguese]$_{\text{MISC}}$ defender [Dimas]$_{\text{PER}}$, while [Alessandro Del Piero]$_{\text{PER}}$ and [Croat]$_{\text{MISC}}$ [Alen Boksic]$_{\text{PER}}$ lead the attack.
SCDL: All-conquering [Juventus]$_{\text{ORG}}$ field their most recent signing, [Portuguese]$_{\text{MISC}}$ defender [Dimas]$_{\text{PER}}$, while [Alessandro Del Piero]$_{\text{PER}}$ and [Croat Alen Boksic]$_{\text{PER}}$ lead the attack.
SCDL: All-conquering [Juventus]$_{\text{ORG}}$ field their most recent signing, [Portuguese]$_{\text{MISC}}$ defender [Dimas]$_{\text{PER}}$, while [Alessandro Del Piero]$_{\text{PER}}$ and [Croat Alen Boksic]$_{\text{PER}}$ lead the attack.
ATSEN: All-conquering [Juventus]$_{\text{ORG}}$ field their most recent signing, [Portuguese]$_{\text{MISC}}$ defender [Dimas]$_{\text{PER}}$, while [Alessandro Del Piero]$_{\text{PER}}$ and [Croat]$_{\text{ORG}}$ [Alen Boksic]$_{\text{PER}}$ lead the attack.
CENSOR: All-conquering [Juventus]$_{\text{ORG}}$ field their most recent signing, [Portuguese]$_{\text{MISC}}$ defender [Dimas]$_{\text{PER}}$, while [Alessandro Del Piero]$_{\text{PER}}$ and [Croat]$_{\text{MISC}}$ [Alen Boksic]$_{\text{PER}}$ lead the attack.

Table 11: Case study with CENSOR and previous teacher-student methods for DS-NER. Then sentence is from CoNLL03 test set.

Uncertainty-Aware Teacher Learning Most research on uncertainty estimation focuses on computer vision because it provides visual validation on uncertainty quality. For example, Rizve et al. (2021) first introduces uncertainty to filter the low-quality labels in the semi-supervised image classification task. However, very little research about uncertainty has been presented in the natural language process domain. As far as we know, we are the first to introduce the uncertainty in the DS-NER task. Meanwhile, different from the instance-level image classification task, the DS-NER task is based on token-level classification, which requires the model to capture the inherent token-wise label dependency. So different from estimating uncertainty at the instance level, we analyze the unique characteristics of the DS-NER task in the paper and design Uncertainty-Aware Teacher Learning to measure uncertainty at the token level. On the other hand, we are the first to find that previous teacher-student methods achieved limited performance because poor network calibration produces incorrect pseudo-labeled samples in the DS-NER task. Thus, we attempt to use uncertainty as the indicator to reduce the effect of incorrect pseudo labels within the teacher-student framework.

Student-Student Collaborative Learning Collaborative Learning (Han et al., 2018; Yu et al., 2019; Wei et al., 2020) is a popular method to handle label noise, which attempts to use two different networks to provide multi-view knowledge and let them learn from each other. **Co-teaching** (Han et al., 2018) first attempts to completely exchange reliable samples of two different networks and then update the networks by the exchanged multi-view information. **Co-teaching+** (Yu et al., 2019) further proposes to use disagreement strategy to update two networks, i.e., only using prediction disagreement data from two networks to update two networks. **JoCoR** (Wei et al., 2020) aims to use a designed joint loss to reduce the diversity of two networks during training and further improve the robustness of two networks. However, these methods are designed for tasks in the computer vision area (especially image classification), and as shown in Table 1, these methods often achieve limited performance in the DS-NER task. **SCDL** designs the teacher-student framework and adopts collaborative learning in the DS-NER task. Similar to Co-teaching, all of the pseudo labels predicted by the teacher are applied to update the noisy labels of the peer teacher-student network periodically since two teacher-student networks have different learning abilities based on different network structures. Different from SCDL, we aim to utilize two different student networks and let them learn from each other to reduce the negative effect of incorrect pseudo labels. Specifically, instead of completely exchanging pseudo labels between two teachers, we allow students to transfer reliable pseudo labels and at the same time allow students to learn on their own pseudo labels generated by their teacher network. In this way, we not only ensure that the transferred pseudo labels contain multi-view information but also ensure that the pseudo labels we transfer are high-quality by selective transfer. Meanwhile, as the student network is updated earlier and more frequently than the teacher network, the student network is better able to capture the changes of pseudo labels than the teacher network.

Relation between Two Components Designs on Uncertainty-Aware Teacher Learning and Student-Student Collaborative Learning are not independent. The two components can collaborate and achieve better results. Specifically, (1) Uncertainty-Aware Teacher Learning can help the teacher network to generate more reliable pseudo labels and further reduce the risk of the student network updating parameters on the incorrect pseudo label. At the same time, a more efficient student network can be achieved by learning to pseudo-label with fewer errors, which will further improve the efficiency of the Student-Student Collaborative Learning component; (2) Based on Uncertainty-Aware Teacher Learning, the teacher network can utilize the correctly pseudo-labeled samples to alleviate the negative effect of label noise. However, simply masking unreliable pseudo-labeled samples can lead to underutilization of the training set, as there is no chance for the incorrect pseudo-labeled samples to be corrected and further learned. Student-Student Collaborative Learning can allow the student network to learn from transferred reliable labels from the other student network. Therefore, this component further enables a full exploration of mislabeled samples rather than simply filtering unreliable pseudo-labeled samples. Through the collaboration of the two components, as shown in Table 1, CENSOR achieves the best performance among 12 baselines.