Improving Distantly-Supervised Named Entity Recognition with Self-Collaborative Denoising Learning

Xinghua Zhang, Bowen Yu, Tingwen Liu*, Zhenyu Zhang,
Jiawei Sheng, Mengge Xue and **Hongbo Xu**
Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
{zhangxinghua,yubowen,liutingwen,zhangzhenyu1996}@iie.ac.cn
{shengjiawei,xuemengge,hbxu}@iie.ac.cn

Abstract

Distantly supervised named entity recognition (DS-NER) efficiently reduces labor costs but meanwhile intrinsically suffers from the label noise due to the strong assumption of distant supervision. Typically, the wrongly labeled instances comprise numbers of incomplete and inaccurate annotation noise, while most prior denoising works are only concerned with one kind of noise and fail to fully explore useful information in the whole training set. To address this issue, we propose a robust learning paradigm named Self-Collaborative Denoising Learning (SCDL), which jointly trains two teacher-student networks in a mutually-beneficial manner to iteratively perform noisy label refinery. Each network is designed to exploit reliable labels via self denoising, and two networks communicate with each other to explore unreliable annotations by collaborative denoising. Extensive experimental results on five real-world datasets demonstrate that SCDL is superior to state-of-the-art DS-NER denoising methods[1].

1 Introduction

Named Entity Recognition (NER) is the task of detecting entity spans and then classifying them into predefined categories, such as person, location and organization. Due to the capability of extracting entity information and benefiting many NLP applications (e.g., relation extraction (Lin et al., 2017), question answering (Li et al., 2019)), NER appeals to many researchers. Traditional supervised methods for NER require a large amount of high-quality corpus for model training, which is extremely expensive and time-consuming as NER requires token-level labels.

Therefore, in recent years, distantly supervised named entity recognition (DS-NER) has been proposed to automatically generate labeled training set by aligning entities in knowledge bases (e.g., Freebase) or gazetteers to corresponding entity mentions in sentences. This labeling procedure is based on a strong assumption that each entity mention in a sentence is a positive instance of the corresponding type according to the extra resources. However, this assumption is far from reality. Due to the limited coverage of existing resources, many entity mentions in the text cannot be matched and are wrongly annotated as non-entity, resulting in incomplete annotations. Moreover, two entity mentions with the same surface name can belong to different entity types, thus simple matching rules may fall into the dilemma of labeling ambiguity and produce inaccurate annotations. As illustrated in Figure 1, the entity mention "*Jack Lucas*" is not recognized due to the limited coverage of extra resources and "*Amazon*" is wrongly labeled with organization type owing to the labeling ambiguity.

	Jack	Lucas	was	born	in	the	Amazon	region	.
Golden Labels:	*B-PER*	*I-PER*	*O*	*O*	*O*	*O*	*B-LOC*	*O*	*O*
Noisy Labels:	*O*	*O*	*O*	*O*	*O*	*O*	*B-ORG*	*O*	*O*

Figure 1: A noisy sample generated by distantly-supervised methods, where *Jack Lucas* is the incomplete annotation and *Amazon* is inaccurate.

Recently, many denoising methods (Shang et al., 2018b; Yang et al., 2018; Cao et al., 2019; Peng et al., 2019; Li et al., 2021) have been developed to handle noisy labels in DS-NER. For example, Shang et al. (2018b) obtained high-quality phrases through *AutoPhrase* (Shang et al., 2018a) and designed AutoNER to model these phrases that may be potential entities. Peng et al. (2019) proposed a positive-unlabeled learning algorithm to unbiasedly and consistently estimate the NER task loss, and Li et al. (2021) used negative sampling to eliminate the misguidance brought by unlabeled entities. Though achieving good performance, most studies mainly focus on solving incomplete annotations

*Corresponding author

[1] The source code and data can be found at https://github.com/AIRobotZhang/SCDL.

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 1518–1529
November 7–11, 2021. ©2021 Association for Computational Linguistics

with a strong assumption of no inaccurate ones existing in DS-NER. Meanwhile, these methods aim to reduce the negative effect of noisy labels by weakening or abandoning the wrongly labeled instances. Hence, they can at most alleviate the noisy supervision and fail to fully mine useful information from the mislabeled data. Intuitively, if we can rectify those unreliable annotations into positive instances for model training, a higher data utilization and better performance will be achieved. We argue that an ideal DS-NER denoising system should be capable of solving two kinds of label noise (i.e., incomplete and inaccurate annotations) and making full use of the whole training set.

In this work, we strive to reconcile this gap and propose a robust learning framework named SCDL (Self-Collaborative Denoising Learning). SCDL co-trains two teacher-student networks to form inner and outer loops for coping with label noise without any assumption, as well as making full exploration of mislabeled data. The inner loop inside each teacher-student network is a self denoising scheme to select reliable annotations from two kinds of noisy labels, and the outer loop between two networks is a collaborative denoising procedure to rectify unreliable instances into useful ones. Specifically, in the inner loop, each teacher-student network selects consistent and high-confidence labeled tokens generated by the teacher to train the student, and then updates the teacher gradually via exponential moving average (EMA)[2] based on the re-trained student. And as for the outer loop, the high-quality pseudo labels generated by one network's teacher are used to update the noisy labels of the other network thanks to the stability of EMA and different noise sensitivities between two networks. Moreover, the inner and outer loop procedures will be performed alternately. Obviously, a successful self denoising process (inner loop) can generate high-quality pseudo labels which benefit the collaborative learning procedure (outer loop) a lot and a promising outer loop will promote the inner loop by refining noisy labels, thus handling the label noise in DS-NER effectively.

We evaluate our method on five DS-NER datasets. Experimental results indicate that SCDL consistently achieves superior performance over previous competing approaches. Extensive validation studies demonstrate the rationality and robustness of our self-collaborative denoising framework.

2 Related Work

Many studies have obtained reliable performance in NER. For example, BiLSTM-CRF (Lample et al., 2016) and BERT (Devlin et al., 2019) based methods become the paradigm in NER due to their promising performances. However, most of these works rely on high-quality labels, which are quite expensive. To address this issue, several studies attempted to annotate tokens via distant supervision (Liang et al., 2020). They matched unlabeled sentences with external gazetteers or knowledge Graphs (KGs). Despite the success of distant supervision, it still suffers from noisy labels (i.e., incomplete and inaccurate annotations in NER).

DS-NER Denoising. Many studies (Shang et al., 2018b; Cao et al., 2019; Jie et al., 2019) tried to modify the standard CRF for adapting to the scenario of label noise, e.g., Fuzzy CRF. Ni et al. (2017) selected high-confidence labeled data from noisy data to train NER models. And many new training paradigms were proposed to resist label noise in DS-NER, such as AutoNER (Shang et al., 2018b), Reinforcement Learning (Yang et al., 2018; Nooralahzadeh et al., 2019), AdaPU (Peng et al., 2019) and Negative Sampling (Li et al., 2021). In addition, some studies (Mayhew et al., 2019; Liang et al., 2020) performed iterative training procedures to mitigate noisy labels in DS-NER. However, most studies mainly focus on incomplete annotations regardless of inaccurate ones or depending on manually labeled data. What's more, most prior methods are insufficient since they can at most alleviate the negative effect caused by label noise and fail to mine useful information from the whole training set. Different from previous studies, we propose two denoising learning procedures which can be enhanced each other mutually with the devised teacher-student network and co-training paradigm, mitigating two kinds of label noise and making full use of the whole training set.

Teacher-Student Network. The teacher-student network is well known in knowledge distillation (Hinton et al., 2014). A teacher is generally a complicated model and the light weight student imitates its output. Recently, there are many variations of teacher-student network. For example, self-training copies the student as a new teacher to gen-

[2] A momentum technique that has been explored in several studies, e.g., Adam (Kingma and Ba, 2015), semi-supervised (Tarvainen and Valpola, 2017) and self-supervised (Grill et al., 2020) learning.

erate pseudo labels (Xie et al., 2020; Wang et al., 2020). Liang et al. (2020) applied self-training with teacher-student network to handle label noise in DS-NER. However, for the teacher-student network in our framework, the teacher selects reliable annotations with devised strategies for training student and then we use EMA to update the teacher based on re-trained student. With this loop, our method can learn entity knowledge effectively.

Co-Training. The co-training paradigm which jointly trains two models is used to improve the robustness of models (Blum and Mitchell, 1998; Nigam and Ghani, 2000; Kiritchenko and Matwin, 2011). Many previous frameworks (Han et al., 2018; Yu et al., 2019; Wei et al., 2020; Li et al., 2020) have adopted co-training to denoise, but they mainly use the diversity of two single models and the single one doesn't have the denoising ability. But supervision signals from the peer model are not always clean. Instead, we train two groups of teacher-student networks and each group can also perform label denoising effectively which further improves the co-training paradigm.

3 Task Definition

Given the training corpus $\mathcal{D}$ where each sample is a form of (X_i, Y_i), $X_i = x_1, x_2, ..., x_N$ represents a sentence with *N* tokens and $Y_i = y_1, y_2, ..., y_N$ is the corresponding tag sequence. Each entity mention $e = x_i, ..., x_j (0 \leq i \leq j \leq N)$ is a span of the text , associated with an entity type, e.g., person, location. In this paper, we use the BIO scheme following (Liang et al., 2020). In detail, the begin token of an entity mention is labeled as *B-type* and others are *I-type*. The non-entity tokens are annotated as *O*.

The traditional NER problem is a supervised learning task by fitting a sequence labeling model based on the training dataset. However, we mainly explore the practical scenario when the labels of training data are contaminated due to the distant supervision. In other words, the revealed tag y_i may not correspond to the underlying correct one. The challenge posed in this setting is to reduce the negative influence of noisy annotations and generate high-confidence labels for them to make full use of the training data.

4 Methodology

In this section, we give a detailed description of our self-collaborative denoising learning framework, which consists of two interactive teacher-student networks to address both the incomplete and inaccurate annotation issues. As illustrated in Figure 2, each teacher-student network contributes to an inner loop for self denoising and the outer loop between two networks is a collaborative denoising scheme. These two procedures can be optimized in a mutually-beneficial manner, thus improving the performance of the NER system.

4.1 Self Denoising Learning

It is widely known that deep neural networks have high capacity for memorization (Arpit et al., 2017). When noisy labels become prominent, deep neural NER models inevitably overfit noisy labeled data, resulting in poor performance. The purpose of self denoising learning is to select reliable labels to reduce the negative influence of noisy annotations. To achieve this end, self denoising learning involves a teacher-student network, where the teacher first generates pseudo labels to participate in labeled token selection, then the student is optimized via back-propagation based on selected tokens, and finally the teacher is updated by gradually shifting the weights of the student in continuous training with exponential moving average (EMA). We take two neural NER models with the same architecture as the teacher and student respectively.

4.1.1 Labeled Token Selection

This subsection illustrates our labeled token selection strategy based on the consistency and high confidence predictions.

Consistency Predictions. It has been observed that the model's predictions of wrongly labeled instances fluctuate drastically in previous studies (Huang et al., 2019). A mislabeled instance will be supervised by both its wrong label and similar instances. For example, *Amazon* is wrongly annotated as *organization* in Figure 1. The wrong label *organization* pushes the model to fit this supervision signal while other clean tokens with similar context will encourage the model to predict it as *location*. Therefore, we can take advantage of this property to separate clean tokens from noisy ones.

Based on above analysis, how to quantify the fluctuation becomes a key issue. One straightforward solution is to integrate predictions from different training iterations but with more time-space complexity. Thanks to the widespread concern of EMA, we use it to update the teacher's parameters.

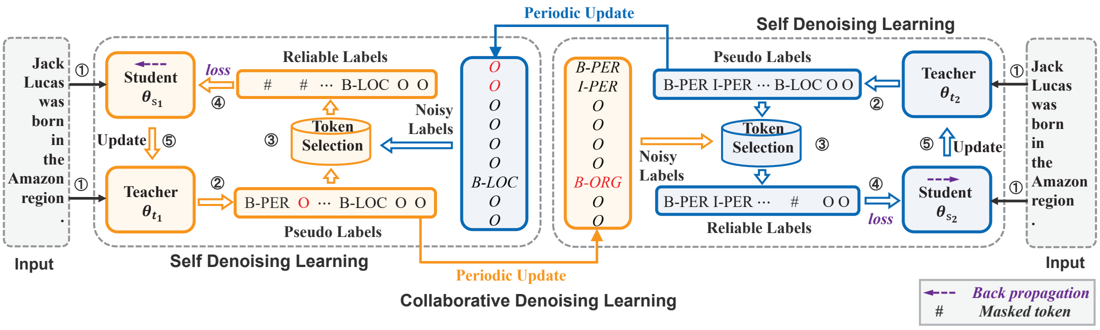

Figure 2: Overview of SCDL with two procedures performed iteratively. (1) Each teacher-student network contributes to an **inner loop** (i.e., **self denoising**): [②] the teacher first generates pseudo labels to [③] select tokens along with noisy labels, then [④] the student is optimized based on selected tokens, and finally [⑤] the teacher is updated by the student. (2) The interplay between two teacher-student networks is an **outer loop** (i.e., **collaborative denoising**): the pseudo labels are applied to update the noisy labels of the peer network periodically.

In this way, the teacher can be viewed as the temporal ensembling of the student models in different training steps and then its prediction will be the ensemble of predictions from past iterations. Therefore, the pseudo labels predicted by the teacher can quantify the fluctuation of noisy labels naturally. Subsequently, we devise the first token selection strategy based on the fluctuation of noisy labels to identify the correctly labeled tokens $(\bar{X}_i, \bar{Y}_i)$ via the consistency between noisy labels and predicted pseudo labels, denoted as:

$$(\bar{X}_i, \bar{Y}_i)_{\text{CP}} = \{(x_j, y_j) \mid y_j = \tilde{y}_j, \tilde{y}_j \in f(X_i; \theta_t)\} \tag{1}$$

where $y_j \in Y_i$ is the noisy label of the j-th token in the i-th sentence and $\tilde{y}_j$ is the pseudo label predicted by the teacher θ_t.

High Confidence Predictions. As studied in previous works (Bengio et al., 2009; Arpit et al., 2017), hard samples can not be learnt effectively at first, thus predictions of those mislabeled hard samples may not fluctuate and then they are mistakenly believed to be reliable. To alleviate this issue, we propose the second selection strategy to pick tokens with high confidence predictions, as formulated in Equation 2, where $\tilde{p}_j$ is the label distribution of the j-th token predicted by the teacher, δ denotes the confidence threshold.

$$(\bar{X}_i, \bar{Y}_i)_{\text{HCP}} = \{(x_j, y_j) \mid \max(\tilde{p}_j) \geq \delta\} \tag{2}$$

4.1.2 Optimization

Loss Function of the Student. Standard supervised NER methods are fitting the outputs of a model to hard labels (i.e, one-hot vectors) to optimize the parameters. However, when the model is trained with tokens and mismatched hard labels, wrong information is being provided to the model. Compared with hard labels, the supervision with soft labels is more robust to the noise because it carries the uncertainty of the predicted results. Therefore, we modify the standard cross entropy loss into a soft label form defined as:

$$\mathcal{L}(\theta_s) = -\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\sum_{c=1}^{C}\mathbb{I}_{i,j}\tilde{p}^{i}_{j,c}\log(p^{i}_{j,c}) \tag{3}$$

$$\mathcal{T}_i = (\bar{X}_i, \bar{Y}_i)_{\text{CP}} \cap (\bar{X}_i, \bar{Y}_i)_{\text{HCP}} \tag{4}$$

where $p^i_{j,c}$ is the probability of the j-th token with the c-th class in the i-th sentence predicted by the student and $\tilde{p}^i_{j,c}$ is from the teacher. $\mathcal{T}_i$ includes the tokens in the i-th sentence meeting the consistency and high confidence selection strategies simultaneously. $\mathbb{I}$ is the indicator function, $\mathbb{I}_{i,j} = 1$ when the j-th token is in $\mathcal{T}_i$, otherwise $\mathbb{I}_{i,j}$ is 0.

Then the parameters of the student model can be updated via back-propagation as follows:

$$\theta_s \leftarrow \theta_s - \gamma\frac{\partial\mathcal{L}}{\partial\theta_s} \tag{5}$$

Update of the Teacher. Different from the optimization of the student model, we apply EMA to gradually update the parameters of the teacher, as shown in Equation 6, where α denotes the smoothing coefficient.

$$\theta_t \leftarrow \alpha\theta_t + (1-\alpha)\theta_s \tag{6}$$

Although the clean token selection strategies indeed alleviate noisy annotations, they also suffer from unreliable token choice which misguides the model into generating biased predictions. As formulated in Equation 7, the update of the teacher θ_t^i in *i*-th iteration can be converted into the form of back-propagation (derivations in Appendix A.1):

$$\theta_t^i = \theta_t^{i-1} - \gamma(1-\alpha)\sum_{j=0}^{i-1}\alpha^{i-1-j}\frac{\partial \mathcal{L}}{\partial \theta_s^j} \quad (7)$$

where γ is the learning rate and $(1-\alpha)$ is a small number because α is generally assigned a value close to 1 (e.g., 0.995), equivalent to multiplying a small coefficient on the weighted sum of student's past gradients. Therefore, with the conservative and ensemble property, the application 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 collaborative denoising phase.

4.2 Collaborative Denoising Learning

Based on the devised clean token selection strategy in self denoising learning, the teacher-student network can utilize the correctly labeled tokens in an ideal situation to alleviate the negative effect of label noise. However, just filtering unreliable labeled tokens will inevitably lose useful information in training set since there is no opportunity for the wrongly labeled tokens to be corrected and explored. Intuitively, if we can change the wrong label to the correct one, it will be transformed into a useful training instance.

Inspired by some co-training paradigms (Han et al., 2018; Yu et al., 2019; Wei et al., 2020), we propose the collaborative denoising learning to update noisy labels mutually for mining more useful information from dataset by deploying two teacher-student networks with different architecture. As stated in (Bengio, 2014), a human brain can learn more effectively if guided by the signals produced by other humans. Similarly, 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 initial conditions and network structures. With this outer loop, the noisy labels can be improved continuously and the training set can be fully explored.

Algorithm 1 Training Procedure of SCDL

```
Input: Training corpus $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^M$ with noisy labels
Parameter: Two network parameters $\theta_{t_1}, \theta_{s_1}, \theta_{t_2}$, and $\theta_{s_2}$
Output: The best model
Pre-training two models $\theta_1, \theta_2$ with $\mathcal{D}$.    ▷Pre-Training.
$\theta_{t_1} \leftarrow \theta_1, \theta_{s_1} \leftarrow \theta_1, \theta_{t_2} \leftarrow \theta_2, \theta_{s_2} \leftarrow \theta_2$, step $\leftarrow 0$.
Initialize noisy labels: $Y_I \leftarrow Y, Y_{II} \leftarrow Y$.
while not reach max training epochs do
    Get a batch $(X^{(b)}, Y_I^{(b)}, Y_{II}^{(b)})$ from $\mathcal{D}$,
    step $\leftarrow$ step $+ 1$.    ▷Self Denoising Learning.
    Get pseudo-labels via the teacher $\theta_{t_1}, \theta_{t_2}$:
    $\tilde{Y}_I^{(b)} \leftarrow f(X^{(b)}; \theta_{t_1})$,
    $\tilde{Y}_{II}^{(b)} \leftarrow f(X^{(b)}; \theta_{t_2})$.
    Get clean tokens:
    $\mathcal{T}_I^{(b)} \leftarrow \text{TokenSelection}(Y_I^{(b)}, \tilde{Y}_I^{(b)})$,
    $\mathcal{T}_{II}^{(b)} \leftarrow \text{TokenSelection}(Y_{II}^{(b)}, \tilde{Y}_{II}^{(b)})$.
    Update the student $\theta_{s_1}$ and $\theta_{s_2}$ by Eq. 3 and Eq. 5.
    Update the teacher $\theta_{t_1}$ and $\theta_{t_2}$ by Eq. 6.
    if step mod $Update\_Cycle = 0$ then
        Update noisy labels mutually:    ▷Collaborative Denoising Learning.
        $Y_I = \{Y_i \leftarrow f(X_i; \theta_{t_2})\}_{i=1}^M$,
        $Y_{II} = \{Y_i \leftarrow f(X_i; \theta_{t_1})\}_{i=1}^M$.
    end if
end while
Evaluate models $\theta_{t_1}, \theta_{s_1}, \theta_{t_2}, \theta_{s_2}$ on Dev set.
return The best model $\theta \in \{\theta_{t_1}, \theta_{s_1}, \theta_{t_2}, \theta_{s_2}\}$
```

4.3 Algorithm Workflow

In this subsection, we introduce the overall procedure of our SCDL framework. Algorithm 1 gives the pseudocode. To summarize, the training process of SCDL can be divided into three procedures: (1) **Pre-Training with Noisy Labels.** We warm up two NER models θ_1 and θ_2 on the noisy labels to obtain a better initialization, and then duplicate the parameters θ for both the teacher θ_t and the student θ_s (i.e., θ_{t_1}= θ_{s_1}= θ_1, θ_{t_2}= θ_{s_2}= θ_2). The training objective function in this stage is the cross entropy loss with the following form:

$$\mathcal{L}(\theta) = -\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N} y_j^i \log(p(y_j^i|X_i;\theta)) \quad (8)$$

where y_j^i means the *j*-th token label of the *i*-th sentence in the noisy training corpus and $p(y_j^i|X_i;\theta)$ denotes its probability produced by model θ. *M* and *N* are the size of training corpus and the length of sentence respectively. (2) **Self Denoising Learning.** In this stage, we can select correctly labeled tokens to train the two teacher-student networks respectively. (3) **Collaborative Denoising Learning.** Self denoising can only utilize correct annotations and this phase will update noisy labels mutually to relabel tokens for two teacher-student networks. The initial noisy labels of two networks comes from distant supervision. The second and third phase are conducted alternately, which will promote each

	Method	CoNLL03			OntoNotes5.0			Webpage			Wikigold			Twitter		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
(i)	**BiLSTM-CRF**♣	91.35	91.06	91.21	85.99	86.36	86.17	50.07	54.76	52.34	55.40	54.30	54.90	60.01	46.16	52.18
	RoBERTa♣	89.14	91.10	90.11	84.59	87.88	86.20	66.29	79.73	72.39	85.33	87.56	86.43	51.76	52.63	52.19
(ii)	**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
(iii)	**AutoNER**‡ (Shang et al., 2018b)	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‡ (Cao et al., 2019)	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+‡* (Yu et al., 2019)	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‡* (Wei et al., 2020)	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‡* (Li et al., 2021)	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‡ (Liang et al., 2020)	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 (Ours)	**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**

Table 1: Main results on five benchmark datasets. (i) ♣ marks the model trained on the fully clean dataset. (ii) † marks the model trained on noisy dataset without label denoising. (iii) ‡ marks the prior label denoising framework. ⋆ marks produced with official implementation.

other to perform label denoising. It's worth noting that only the best model $\theta \in \{\theta_{t_1}, \theta_{s_1}, \theta_{t_2}, \theta_{s_2}\}$ is adopted for predicting.

5 Experiments

In this section, we evaluate the performance of SCDL, compared with several comparable baselines. Additionally, we conduct lots of auxiliary experiments and provide comprehensive analyses to justify the effectiveness of SCDL.

5.1 Experimental Settings

Datasets. We conduct experiments on five publicly available NER datasets: **CoNLL03** (Tjong Kim Sang, 2002), **OntoNotes5.0** (Weischedel et al., 2013), **Webpage** (Ratinov and Roth, 2009), **Wikigold** (Balasuriya et al., 2009) and **Twitter** (Godin et al., 2015). Liang et al. (2020) re-annotated the training set by distant supervision, and left the development and test set unchanged. The statistics of datasets are in Appendix A.2.

Baselines and Evaluation Metrics. We compare our method with several competitive baselines from three aspects. (i) *Fully-Clean.* **BiLSTM-CRF** (Ma and Hovy, 2016) and **RoBERTa** (Liu et al., 2019) are fully trained on clean dataset (without noisy labels) for NER, as the upper bound of denoising. (ii) *Fully-Noisy.* **KB-Matching** uses distant supervision to annotate test set. **BiLSTM-CRF**, **DistilRoBERTa** and **RoBERTa** are trained on noisy dataset without label denoising, as the lower bound of denoising. (iii) *Label-Denoising.* We compare several DS-NER denoising baselines which propose to solve noisy labels. **AutoNER** (Shang et al., 2018b) and **LRNT** (Cao et al., 2019) try to reduce the negative effect of noisy labels, leaving training dataset unexplored fully. **Co-teaching+** (Yu et al., 2019) and **JoCoR** (Wei et al., 2020) are two classical label denoising methods, developed in computer vision. **NegSampling** (Li et al., 2021) only handles incomplete annotations by negative sampling. **BOND** (Liang et al., 2020) adapts self-training directly to DS-NER, suffering from confirmation bias (a problem from self-training itself). We use Precision (**P**), Recall (**R**) and **F1** score as the evaluation metrics.

Implementation Details. For fair comparison, we adopt RoBERTa (θ_1) and DistilRoBERTa (θ_2) as the basic models. The max training epochs is 50, and the confidence threshold δ is 0.9. The batch size is set to 16 or 32, the learning rate is 1e-5 or 2e-5 according to different datasets. We tune EMA parameter α from {0.9,0.99,0.995,0.998}, tune update cycle according to the size of dataset (e.g., 6000 iterations (about 7 epochs) for CoNLL03) on development set. We implement our code with Pytorch based on huggingface Transformers[3]. Detailed hyperparameter settings for each dataset and tuning procedures are listed in Appendix A.3.

5.2 Experimental Results

Table 1 shows the results of our proposed method compared with baselines and highlights the best overall performance in bold. Obviously, SCDL achieves the best performance, and improves the precision as well as F1 score significantly, compared with previous state-of-the-art models.

[3] https://huggingface.co/transformers/

	P	R	F1
SCDL	89.42	80.74	**84.86**
w/o consistency prediction	87.01	**81.11**	83.96
w/o high confidence prediction	88.14	80.94	84.38
w/o θ_{t_2}, θ_{s_2} (co-training paradigm)	88.45	78.32	83.08
w/o θ_{t_1}, θ_{t_2} (teacher-student network)	87.90	77.22	82.22
w/o soft labels	**89.86**	79.12	84.15

Table 2: Ablation study on CoNLL03 dev set.

Compared to our basic models (i.e., DistilRoBERTa and RoBERTa), SCDL improves the F1 score with an average increase of 8.33% and 6.37% respectively, which demonstrates the necessity of label denoising in the distantly-supervised NER task and the effectiveness of the proposed method.

In addition, SCDL performs much better than previous studies which consider the noisy labels in NER, including AutoNER, LRNT, NegSampling and BOND. The reason is that they mainly focus on one kind of label noise in DS-NER or fail to make full use of the mislabeled data with their strategies. On contrast, our method can not only exploit correctly labeled tokens but also explore valuable information in wrongly labeled ones by correction. Compared to the popular denoising methods in computer vision: Co-teaching+ and JoCoR, SCDL gains of up to 12.05% absolute percentage points in F1 score. We guess this is beacause most computer vision denoising studies focus on instance-level classification, while NER is a token-level task where non-entity category accounts for the majority, and this case is not fully considered. Thus corruption occurs easily in DS-NER denoising task for these methods as the training goes.

5.3 Analysis

Ablation Study. To evaluate the influence of each component in our method, we conduct the ablation study for further exploration (see Table 2). Overall, although SCDL is not optimal on precision or recall, it achieves the best in F1 score, which indicates that our method can balance well when taking two kinds of annotation noise into account and exploring full training set. Based on these ablations, we observe that: (1) Token selection strategy with the consistency and high confidence predictions indeed promote the overall performance (F1 score) by improving the precision and marginally lowering the recall. The recall value doesn't decrease sharply in our framework because of the unbiased predictions generated by teacher model and alternate optimization. (2) When we keep only one teacher-student network (i.e., w/o θ_{t_2}, θ_{s_2}), both recall and F1 decrease visibly, which validates the effectiveness of collaborative denoising learning since more wrongly labeled tokens (e.g., false negative tokens) can be explored via the peer dynamic update of noisy labels. (3) Meanwhile, removing two teacher models (i.e., w/o θ_{t_1}, θ_{t_2}) leads to the decline on both precision and recall. Because this simplification impairs the devised teacher-student network. It uses the predictions of each student to support the token selection strategies and the mutual update of noisy labels, which loses the stable optimization ability of EMA and leads to unreliable token selection. (4) Learning from noisy annotations benefits from soft labels since they contain the uncertainty of predicted results and are more tolerant to the noise compared to the hard ones.

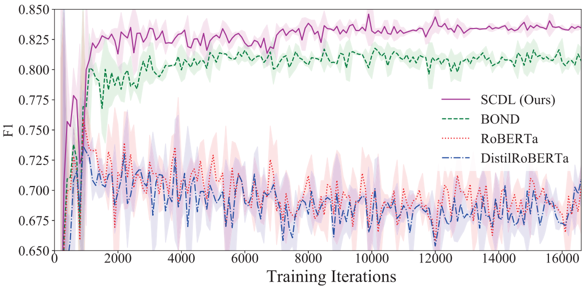

Figure 3: Learning curves of SCDL and other baselines about F1 score vs. training iterations on CoNLL03.

Learning Curve of SCDL. To evaluate the advantage of the proposed framework in handling noisy labels during training, we show the F1 score vs. training iterations on CoNLL03 test set in Figure 3. Compared to RoBERTa and DistilRoBERTa, the performance of SCDL and BOND remains stable as the training goes. Because of the memorization effect of networks, the F1 score of RoBERTa and DistilRoBERTa first reach a high level and then gradually decrease. Moreover, SCDL consistently achieves better performance than other baselines at almost any training stage, which again confirms the effectiveness of our denoising framework.

Robustness to Different Noise Ratio. To study the robustness of the proposed method in different noise ratio, we randomly replace $k\%$ entity mention labels in the corpus with other entity types or non-entity to construct different proportions of label noise and report the test F1 score on CoNLL03 in Figure 4. The pre-trained language models (e.g., RoBERTa) are robust to low level noise (less than

		Case 1	Case 2
Sentence		The girl , Abyss DeJesus , suffers ··· the St. Christopher Children 's Hospital said .	Thai poll shows military wants PM Banharn out .
Golden Labels		O O O *B-PER* *I-PER* O O ··· O *B-ORG* *I-ORG* *I-ORG* *I-ORG* *I-ORG* O O	*B-MISC* O O O O O *B-PER* O O
Initial Noisy Labels		O O O ***O*** ***O*** O O ··· O ***O*** ***B-PER*** ***O*** ***O*** ***O*** O O	***O*** O O O O O ***O*** O O
Teacher-Student Network 1	Pseudo Labels	O O O *B-PER* *I-PER* O O ··· O *B-ORG* *I-ORG* *I-ORG* *I-ORG* *I-ORG* O O	***B-LOC*** O O O O O *B-PER* O O
	Reliable Labels	O O O *B-PER* *I-PER* O O ··· O *B-ORG* *I-ORG* *I-ORG* *I-ORG* *I-ORG* O O	# O O O O O *B-PER* O O
Teacher-Student Network 2	Pseudo Labels	O O O *B-PER* *I-PER* O O ··· O ***B-PER*** ***I-PER*** ***O*** ***O*** ***O*** O O	*B-MISC* O O O O O *B-PER* O O
	Reliable Labels	O O O *B-PER* *I-PER* O O ··· O # # # # # O O	*B-MISC* O O O O O *B-PER* O O

Table 4: Case studies. Wrong labels are marked in red and # means the masked token (i.e., not selected).

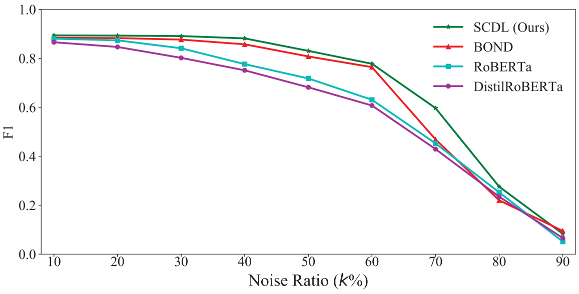

Figure 4: F1 on CoNLL03 with different noise ratio.

20%) due to their strong expressive power. When the noise ratio is between 30% and 80%, SCDL is more robust and exhibits satisfactory denoising ability, since the training data still has reasonable entity type knowledge and SCDL can learn from it to refine noisy labels. However, both SCDL and BOND degenerate to the basic model in the hardest case (more than 80%) which may not exist in reality and needs further studies in the future.

Effectiveness of Noisy Label Refinery. As the noisy labels are updated dynamically during training to explore the full dataset, we compare the F1 score before and after denoising on training set, as shown in Table 3. In detail, SCDL refines noisy labels on CoNLL03 and Twitter training set, from 70.97 to 81.22, 37.73 to 50.82 respectively, which surpasses BOND. The reason may be that BOND mainly depends on self-training which suffers from confirmation bias, while SCDL can bypass this issue by the devised teacher-student network and co-training paradigm and then improves both precision and recall significantly. Overall, the comparison before and after denoising demonstrates that SCDL indeed refines the training noisy labels to a certain extent, leading to the better use of the mislabeled data and outstanding performance on test.

Case Study. Different from most prior denoising studies on DS-NER, our proposed framework SCDL can not only handle two kinds of label noise (i.e., inaccurate and incomplete annotations) without any assumption, but also make full use of the whole training set. High F1 score in Table 1 and the effectiveness of noisy label refinery in Table 3 have proved the feasibility of SCDL quantitatively. For better understanding intuitively, we give two samples from CoNLL03 after two periodic updates to show the denoising ability of SCDL in Table 4. For case 1 with two kinds of label noise, the person name "*Abyss DeJesus*" and organization name "*St. Christopher Children 's Hospital*" are not correctly annotated by DS-NER. After denoising, "*Abyss DeJesus*" is corrected and transformed into a useful instance. Though the hospital name is still not corrected in the *teacher-student network 2*, but *network 2* selects reliable annotations successfully for training student. It shows that SCDL can not only exploit reliable instances but also explore unreliable ones. Similar situations also occur in case 2, while the *network 2* has better capability which demonstrates the validity of co-training paradigm.

	CoNLL03			Twitter		
	P	R	F1	P	R	F1
Distant-Supervision	82.38	62.33	70.97	46.71	31.64	37.73
BOND-Denoising	80.42	**76.46**	78.39	53.76	34.82	42.27
SCDL-Denoising	**87.42**	75.85	**81.22**	**54.86**	**47.33**	**50.82**

Table 3: Comparison of denoising ability of SCDL and BOND on training set.

Efficiency Analysis. In training stage, with the same batch size, the serial efficiency of our method is about 1.5 batches per second on single GPU Tesla T4, other baselines like BOND is 2.6, Co-teaching+ is 1.8, JoCoR is 1.9. The memory usage of our method is equivalent to Co-training models (e.g., Co-teaching+). Although there are two student and two teacher models in our method, only two students need back-propagation which occupies the main computational overhead (time and memory usage), while two teachers updated with EMA only need forward-propagation which occupies less computational overhead. It's worth noting that the two teacher-student networks in our framework can be trained in parallel, which will further accelerate the training. What's more, compared with other baselines, the test efficiency of our method is the same because we only use one model for predicting.

6 Conclusion and Future Work

This paper proposes SCDL to handle two kinds of label noise in DS-NER without any assumption. With devised teacher-student network and co-training paradigm, SCDL can not only exploit more reliable annotations to avoid the negative effect of noisy labels but also explore more useful information from the mislabeled data. Experimental results confirm its effectiveness and robustness in dealing with the label noise. For future work, data augmentation is worth exploring in our framework. Besides, SCDL can also be adapted to other NLP denoising tasks, e.g., classification and matching.

Acknowledgements

We would like to thank the anonymous reviewers for their insightful comments and constructive suggestions. This research is supported by the National Key Research and Development Program of China (grant No.2016YFB0801003) and the Strategic Priority Research Program of Chinese Academy of Sciences (grant No.XDC02040400).

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 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources (People's Web)*, pages 10–18, Suntec, Singapore. Association for Computational Linguistics.

Yoshua Bengio. 2014. Evolving culture vs local minima. In *Growing Adaptive Machines*.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, volume 382 of *ACM International Conference Proceeding Series*, pages 41–48. ACM.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *COLT*.

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)*, pages 261–270, Hong Kong, China. Association for Computational Linguistics.

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

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*, pages 146–153, Beijing, China. Association for Computational Linguistics.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. 2020. Bootstrap your own latent - A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2014. Distilling the knowledge in a neural network. In *NIPS Workshop*.

Jinchi Huang, Lie Qu, Rongfei Jia, and Binqiang Zhao. 2019. O2u-net: A simple noisy label detection approach for deep neural networks. In *2019 IEEE/CVF*

International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, pages 3325–3333. 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*.

Svetlana Kiritchenko and Stan Matwin. 2011. Email classification with co-training. In *Proceedings of the 2011 Conference of the Center for Advanced Studies on Collaborative Research (CASCON 2011)*, pages 301–312.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Junnan Li, Richard Socher, and Steven C. H. Hoi. 2020. Dividemix: Learning with noisy labels as semi-supervised learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. 2019. Entity-relation extraction as multi-turn question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1340–1350, Florence, Italy. Association for Computational Linguistics.

Yangming Li, Lemao Liu, and Shuming Shi. 2021. Empirical analysis of unlabeled entity problem in named entity recognition. In *ICLR*.

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.

Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2017. Neural relation extraction with multi-lingual attention. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 34–43, Vancouver, Canada. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, and et al. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv.

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.

Stephen Mayhew, Snigdha Chaturvedi, Chen-Tse Tsai, and Dan Roth. 2019. Named entity recognition with partially annotated training data. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 645–655, Hong Kong, China. Association for Computational Linguistics.

Jian Ni, Georgiana Dinu, and Radu Florian. 2017. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1470–1480, Vancouver, Canada. Association for Computational Linguistics.

Kamal Nigam and Rayid Ghani. 2000. Analyzing the effectiveness and applicability of co-training. In *CIKM*.

Farhad Nooralahzadeh, Jan Tore Lønning, and Lilja Øvrelid. 2019. Reinforcement-based denoising of distantly supervised NER with partial annotation. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 225–233, Hong Kong, China. Association for Computational Linguistics.

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

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.

Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, and Jiawei Han. 2018a. Automated phrase mining from massive text corpora. In *IEEE Transactions on Knowledge and Data Engineering*, pages 1825–1837.

Jingbo Shang, Liyuan Liu, Xiaotao Gu, Xiang Ren, Teng Ren, and Jiawei Han. 2018b. 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.

Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 1195–1204.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Shaolei Wang, Zhongyuan Wang, Wanxiang Che, and Ting Liu. 2020. Combining self-training and self-supervised learning for unsupervised disfluency detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1813–1822, Online. Association for Computational Linguistics.

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. IEEE.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, and et al. 2013. Ontonotes release 5.0 ldc2013t19. In *Linguistic Data Consortium, Philadelphia, PA 23*.

Qizhe Xie, Minh-Thang Luong, Eduard H. Hovy, and Quoc V. Le. 2020. Self-training with noisy student improves imagenet classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10684–10695. IEEE.

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*, pages 2159–2169, Santa Fe, New Mexico, USA. 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.

A Appendix

A.1 Derivation of EMA Update

In this appendix, we give detailed derivation of reorganizing exponential moving average (EMA) as the form of backpropagation. The student θ_s optimized via back-propagation in the i-th iteration is shown in Equation 9, and Equation 10 represents the update process of the teacher θ_t with EMA.

$$\theta_s^i = \theta_s^{i-1} - \gamma\frac{\partial\mathcal{L}}{\partial\theta_s^{i-1}} \tag{9}$$

$$\theta_t^i = \alpha\theta_t^{i-1} + (1-\alpha)\theta_s^i \tag{10}$$

Based on Equation 9 and Equation 10, the teacher θ_t in the i-th iteration can be represented as follows:

$$\begin{aligned}
\theta_t^i &= \alpha\theta_t^{i-1} + (1-\alpha)\theta_s^i \\
&= \alpha^i\theta_t^0 + \alpha^{i-1}(1-\alpha)(\theta_s^0 - \gamma\frac{\partial\mathcal{L}}{\partial\theta_s^0}) + ...+ \\
&+ (1-\alpha)(\theta_s^0 - \gamma\frac{\partial\mathcal{L}}{\partial\theta_s^0} - ... - \gamma\frac{\partial\mathcal{L}}{\partial\theta_s^{i-1}}) \\
&= \alpha^i\theta_t^0 + (1-\alpha)\sum_{j=0}^{i-1}\alpha^j\theta_s^0 - \gamma(1-\alpha)(\\
&\sum_{j=0}^{i-1}\alpha^j\frac{\partial\mathcal{L}}{\partial\theta_s^0} + \sum_{j=0}^{i-2}\alpha^j\frac{\partial\mathcal{L}}{\partial\theta_s^1} + ... + \sum_{j=0}^{0}\alpha^j\frac{\partial\mathcal{L}}{\partial\theta_s^{i-1}}) \\
&= \alpha^i\theta_t^0 + (1-\alpha)\frac{1-\alpha^i}{1-\alpha}\theta_s^0 - \gamma(1-\alpha)(\\
&\frac{1-\alpha^i}{1-\alpha}\frac{\partial\mathcal{L}}{\partial\theta_s^0} + \frac{1-\alpha^{i-1}}{1-\alpha}\frac{\partial\mathcal{L}}{\partial\theta_s^1} + ... + \frac{\partial\mathcal{L}}{\partial\theta_s^{i-1}}) \\
&= \alpha^i\theta_t^0 + \theta_s^0 - \alpha^i\theta_s^0 - \gamma[(1-\alpha^i)\frac{\partial\mathcal{L}}{\partial\theta_s^0} + \\
&+ (1-\alpha^{i-1})\frac{\partial\mathcal{L}}{\partial\theta_s^1} + ... + (1-\alpha)\frac{\partial\mathcal{L}}{\partial\theta_s^{i-1}}] \\
&= \theta_s^0 - \gamma\sum_{j=0}^{i-1}(1-\alpha^{i-j})\frac{\partial\mathcal{L}}{\partial\theta_s^j} \\
&= \bar{\theta} - \gamma\sum_{j=0}^{i-1}(1-\alpha^{i-j})\frac{\partial\mathcal{L}}{\partial\theta_s^j} \\
&w.r.t.\quad \theta_s^0 = \theta_t^0 = \bar{\theta}
\end{aligned} \tag{11}$$

Therefore,

$$\theta_t^{i-1} = \bar{\theta} - \gamma\sum_{j=0}^{i-2}(1-\alpha^{i-1-j})\frac{\partial\mathcal{L}}{\partial\theta_s^j} \tag{12}$$

As we tend to derive the form of back-propagation as follows:

$$\theta_t^i = \theta_t^{i-1} - \nabla \tag{13}$$

Thus,

$$\begin{aligned}
\nabla &= \theta_t^{i-1} - \theta_t^i \\
&= Equation4 - Equation3 \\
&= \gamma \sum_{j=0}^{i-1} (1 - \alpha^{i-j}) \frac{\partial \mathcal{L}}{\partial \theta_s^j} - \gamma \sum_{j=0}^{i-2} (1 - \alpha^{i-1-j}) \frac{\partial \mathcal{L}}{\partial \theta_s^j} \\
&= \gamma \sum_{j=0}^{i-1} \alpha^{i-1-j} (1 - \alpha) \frac{\partial \mathcal{L}}{\partial \theta_s^j} \\
&= \gamma (1 - \alpha) \sum_{j=0}^{i-1} \alpha^{i-1-j} \frac{\partial \mathcal{L}}{\partial \theta_s^j}
\end{aligned} \tag{14}$$

In the end, we get the back-propagation formula of EMA based on Equation 13 and 14, denoted as:

$$\theta_t^i = \theta_t^{i-1} - \gamma(1 - \alpha) \sum_{j=0}^{i-1} \alpha^{i-1-j} \frac{\partial \mathcal{L}}{\partial \theta_s^j} \tag{15}$$

where γ is the learning rate and $(1 - \alpha)$ is a small number because α is generally assigned a value close to 1 (e.g., 0.995). Therefore, the optimization of EMA is equivalent to multiplying a small coefficient on the weighted sum of student's past gradients. With this conservative and ensemble property, the application of EMA can contribute to a more reliable and robust model.

We adopt EMA in SCDL based on the following reasons: (1) The teacher model updated with EMA can quantify the fluctuation of label noise and contributes to consistency predictions. (2) As we justify above, EMA contributes to unbiased predictions with the conservative and ensemble property. (3) EMA doesn't need back-propagation (BP), which reduces the computational overhead, because BP needs to build the computation graph to compute the gradient.

A.2 Statistics of Datasets

The detailed statistics of five publicly available NER 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 datasets.

Hyper Param.	CoNLL03	ON5.0	Webpage	Wikigold	Twitter
Batch	16	32	16	16	16
Epoch	50	50	50	50	50
LR	1e-5	2e-5	1e-5	1e-5	2e-5
Sche. Warmup	200	500	100	200	200
Pre. Epoch	1	2	12	5	6
Update Cycle (iterations)	6000	7240	300	2000	3200
EMA α	0.995	0.995	0.99	0.99	0.995
Confidence Threshold δ	0.9	0.9	0.9	0.9	0.9

Table 6: Hyper-parameter settings.

A.3 Hyper-parameter and Baseline Settings

Detailed hyper-parameter settings for each dataset are shown in Table 6. Specifically, we firstly tune the partial hyper-parameters with Grid-Search for student models (i.e., θ_{s_1} and θ_{s_2}) (e.g., learning rate chosen from {1e-5, 2e-5, 5e-5, 1e-4}, training epoch from {20, 50, 100}, batch size from {16, 32}). Pretraining epoch is determined when the F1 score on development dataset doesn't increase. The number of steps for the scheduler warmup is chosen from {100, 200, 500}. Then we tune EMA α from {0.9, 0.99, 0.995, 0.998} for teacher models (i.e., θ_{t_1} and θ_{t_2}). Finally, we tune update cycle range from 100 to 8000 according to the size of dataset. The confidence threshold is set to 0.9. The rest parameters are default in huggingface Transformers[4].

For fair comparison, NegSampling and BOND adopt RoBERTa as the basic model. Co-teaching+ and JoCoR adopt RoBERTa, DistilRoBERTa as the basic models. For NegSampling, we run the officially released code using suggested hyperparameters in the original paper. For Co-teaching+ and JoCoR, noise rate τ is calculated by distantly supervised and original training set.

We conduct the experiments on NVIDIA Tesla T4 GPU. It is worth noting that only the best model $\theta \in \{\theta_{t_1}, \theta_{s_1}, \theta_{t_2}, \theta_{s_2}\}$ is adopted for predicting in our SCDL framework. Therefore, the complexity of our model is not increased during the test stage.

[4] https://huggingface.co/transformers/