Distantly-Supervised Named Entity Recognition with Adaptive Teacher Learning and Fine-Grained Student Ensemble

Xiaoye Qu[1], Jun Zeng[2], Daizong Liu[3], Zhefeng Wang[1]*, Baoxing Huai[1], Pan Zhou[4]*

[1]Huawei Cloud
[2]School of Software Engineering, Huazhong University of Science and Technology
[3]Peking University
[4]Hubei Key Laboratory of Distributed System Security, Hubei Engineering Research Center on Big Data Security, School of Cyber Science and Engineering, Huazhong University of Science and Technology
{quxiaoye, wangzhefeng, huaibaoxing}@huawei.com, dzliu@stu.pku.edu.cn, {junzeng, panzhou}@hust.edu.cn

Abstract

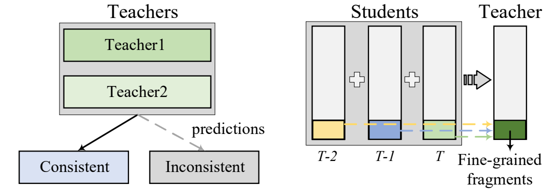
Distantly-Supervised Named Entity Recognition (DS-NER) effectively alleviates the data scarcity problem in NER by automatically generating training samples. Unfortunately, the distant supervision may induce noisy labels, thus undermining the robustness of the learned models and restricting the practical application. To relieve this problem, recent works adopt self-training teacher-student frameworks to gradually refine the training labels and improve the generalization ability of NER models. However, we argue that the performance of the current self-training frameworks for DS-NER is severely underestimated by their plain designs, including both inadequate student learning and coarse-grained teacher updating. Therefore, in this paper, we make the first attempt to alleviate these issues by proposing: (1) adaptive teacher learning comprised of joint training of two teacher-student networks and considering both consistent and inconsistent predictions between two teachers, thus promoting comprehensive student learning. (2) fine-grained student ensemble that updates each fragment of the teacher model with a temporal moving average of the corresponding fragment of the student, which enhances consistent predictions on each model fragment against noise. To verify the effectiveness of our proposed method, we conduct experiments on four DS-NER datasets. The experimental results demonstrate that our method significantly surpasses previous SOTA methods. The code is available at https://github.com/zenhjunpro/ATSEN.

Introduction

Named Entity Recognition (NER) aims to detect entity mentions in the text and classify them into predefined types, such as person, location, and organization. It is a fundamental task in information extraction and benefits many downstream NLP applications (e.g., relation extraction (Cheng et al. 2021), co-reference resolution (Clark and Manning 2016), entity linking (Gu et al. 2021) and event extraction (Zhu et al. 2022)). In recent years, deep supervised models (Li et al. 2022; Gu et al. 2022; Li et al. 2020) have achieved superior success in the NER field. However, these supervised NER methods demand a large amount of high-quality annotation, which is extremely labor-intensive and time-consuming as NER demands token-level annotation.

	Tom	visits	the	White	House	with	Washington
Gold Labels	B-PER	O	O	B-LOC	I-LOC	O	B-PER
Noisy Labels	B-PER	O	O	O	O	O	B-LOC

(a) An example of distant labels



(b) Inadequate student learning (c) Coarse-grained Ensemble

Figure 1: (a) *White House* and *Washington* are incomplete and inaccurate labels. (b) Previous method only considers the consistent prediction parts from teachers, leading to incomprehensible student learning. (c) Coarse-grained student ensemble absorbs a whole student without further considering fine-grained fragments in the model.

To solve this problem, Distantly-Supervised Named Entity Recognition (DS-NER) has attracted increasing attention. It automatically annotates training data based on external knowledge such as easily-obtained dictionaries and knowledge bases, which effectively relieves the annotation difficulty. Unfortunately, such a distant labeling procedure naturally introduces incomplete and inaccurate labels. As depicted in Figure 1 (a), “White House” is unlabeled because the distant supervision source has limited coverage of the entity mentions. Meanwhile, “Washington” is inaccurately labeled as this entity belongs to location types in the distant supervision source. Due to the existence of such noise in the distantly labeled data, straightforward application of supervised learning will yield deteriorated performance as deep neural models have a strong capacity of fitting the given noisy data. Thus, the robustness and generalization of learned DS-NER models are restricted.

[*]Zhefeng Wang and Pan Zhou are corresponding authors. The first two authors contribute equally.

To address the above challenges, several DS-NER models have been proposed. Shang et al. (2018) obtained high-quality phrases and designed TieOrBreak architecture to model those phrases that may be potential entities. Peng et al. (2019) adopt PU learning to perform classification using only limited labeled positive data and unlabeled data. However, these works mainly focus on designing network architectures that can cope with the incomplete annotations to partially alleviate the impact of the noisy annotations. Recently, the self-training teacher-student framework is applied to DS-NER tasks (Liang et al. 2020; Zhang et al. 2021) to reduce the negative effect of both incomplete and inaccurate labels. This self-looping framework first selects high-confidence annotations from noisy labels to train the student network, and then updates a new teacher by the trained student. In this way, the training labels are gradually refined and model generalization can be improved.

However, the above self-training methods have the following shortcomings: (1) inadequate student learning. As shown in Figure 1 (b), previous methods only focus on the consistent prediction from two teachers (Zhang et al. 2021) or simply consider the high-confidence part from a single teacher (Liang et al. 2020). In this way, these models tend to learn uncomplicated mentions, and the entity recall rate will decrease. (2) coarse-grained teacher updating. In Figure 1 (c), previous works absorb a whole student by exponential moving average (EMA) (Zhang et al. 2021) or directly copy the student as a new teacher (Liang et al. 2020) when updating the teacher. Such coarse-grained ensemble methods treat each model fragment equally while the noise sensitivity is diverse among different model fragments.

In this paper, we try to reconcile the above shortcomings with our newly proposed **A**daptive **T**eacher Learning and Fine-grained **S**tudent **EN**semble (ATSEN) for DS-NER. Specifically, we first apply two teacher networks to provide multi-view predictions on training samples. Then we propose an adaptive teacher learning which supervises agreement predictions by cross-entropy loss and accommodates disagreement parts with adaptive distillation. In this way, the student can be trained with more comprehensive knowledge. Subsequently, we update the new teacher with a fine-grained student ensemble, which updates a fragment of the teacher model with a temporal moving average of the corresponding fragment of the student. Therefore, the teacher model achieves more robustness for noise. With both adaptive learning and fine-grained ensemble, ATSEN is more effective than previous methods. We evaluate ATSEN on four DS-NER datasets. Experimental results demonstrate that our method significantly outperforms previous approaches.

To sum up, the main contributions of this paper are:

- To our best knowledge, this paper presents the first attempt to explore both agreement and conflicts among multiple teachers for the DS-NER by adaptive teacher learning, promoting comprehensive student learning.
- To further enhance the consistent prediction of model fragments, we devise a novel fine-grained student ensemble that stitches different fragments of previous student models into a unity. In this way, the updated teacher achieves a more robust generalization ability.
- On four benchmark DS-NER datasets (Conll03, OntoNotes 5.0, WebPage, and Twitter), our ATSEN outperforms existing approaches by significant margins.

Related Work

Traditionally, many works have been proposed for supervised named entity recognition. For instance, Huang, Xu, and Yu (2015) utilized the BiLSTM as an encoder to learn the contextual representation and then exploited Conditional Random Field (CRF) as a decoder to label the tokens. More recently, deep learning methods (Xiao et al. 2020; Qu et al. 2019) are introduced to different NLP fields, and strong pre-trained language models such as ELMo (Peters et al. 2018) and BERT (Devlin et al. 2018) are incorporated to further enhance the performance of NER. However, most of these works rely on high-quality labels, which are expensive. Meanwhile, the reliance on labeled data also limits their applications in open situations.

DS-NER To address the labeled data scarcity problem, distantly-supervised named entity recognition methods are proposed. AutoNER (Shang et al. 2018) proposed a sequence labeling framework TieOrBreak and modify the standard CRF for adapting to the scenario of label noise. Cao et al. (2019) promoted the quality of data by exploiting labels in Wikipedia. AdaPU (Peng et al. 2019) employed Positive-Unlabeled Learning to obtain unbiased estimation of the loss value. Conf-MPU (Zhou, Li, and Li 2022) further formulated the DS-NER problem via Multi-class Positive and Unlabeled (MPU) learning. BOND (Liang et al. 2020) adopted a teacher-student network to drop distant labels and use pseudo labels to gradually improve the model generalization ability. Similar to BOND, SCDL (Zhang et al. 2021) co-trained two teacher-student networks to form inner and outer loops for coping with label noise. In this paper, we propose a novel self-training framework to adaptively learn from multiple teachers and achieve a fine-grained student ensemble. In this way, our method achieves a more robust ability for noise in the DS-NER task.

Teacher-Student Framework The teacher-student framework is a popular architecture in many semi-supervised (Huo et al. 2021) and self-supervised (Abbasi Koohpayegani, Tejankar, and Pirsiavash 2020) learning tasks, as well as knowledge distillation (Hinton et al. 2015). Recently, teacher-student framework attracts increasing attention in both computer vision (He et al. 2020; Grill et al. 2020) and natural language processing (Liang et al. 2020; Zhang et al. 2021). The teacher selects reliable annotations with devised strategies for student training and then the new teacher is updated based on the trained student. The optimization goal is to ensure the prediction consistency between the student and the teacher. In particular, there are several variants of teacher-student networks proposed for DS-NER. BOND devised a self-training teacher-student strategy that copies the student as a new teacher. With this self-training loop, the training pseudo labels are gradually refined. To improve the quality of pseudo labels and remove noise, SCDL designs two teachers and reaches an agreement between them to

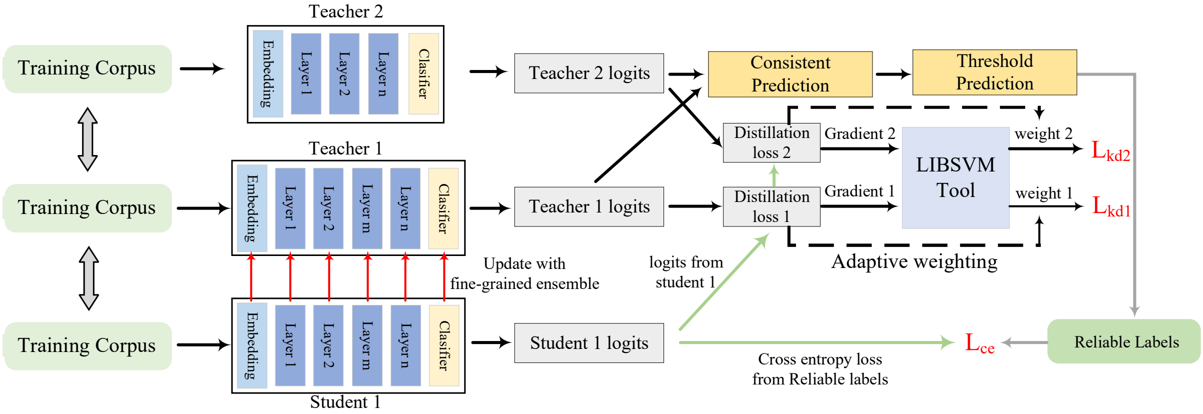


Figure 2: Overview of our proposed ATSEN. Only the updating process of student 1 and teacher 1 is shown and the renewing process of student 2 and teacher 2 is similar. Specifically, the training corpus is first fed to two teachers and one student to obtain corresponding logits. Then the reliable labels are selected to supervise the student with cross-entropy loss. Meanwhile, adaptive distillation is adopted to further consider the disagreement tokens between teachers. Subsequently, a fine-grained ensemble is applied to the trained students to obtain a new teacher model.

generate pseudo labels. Meanwhile, SCDL uses exponential moving average (EMA) to update the teacher based on the re-trained student. Following the self-training framework, we further improve the training process of both the teacher and student network to alleviate the noise problem.

Preliminaries

Here we briefly describe the task definition of DS-NER. Formally, given the training data D, where each sentence is denoted as (X^i, Y^i). X^i is a token list that represents each word, and Y^i is the corresponding tag list in the form of BIO schema. For DS-NER, we do not have access to human-annotated true labels, but only distant labels by matching unlabeled sentences with external dictionaries or knowledge bases (KBs). Thus, Y^i may not be the underlying correct one. To generate distant labels, in this work, we follow the previous work (Liang et al. 2020). The biggest challenge in DS-NER is how to reduce the label noise in the training samples and train a robust NER model as there is much ambiguity and limited coverage over entity types.

Method

In this work, considering the memory capacity and model efficiency, we train two sets of teacher-student networks instead of more pairs while our method can easily extend to more pairs. The main procedure is shown in Figure 2.

Overall Framework

The training procedure can be divided into three stages:
(1) **Pretraining with initial noisy labels.** In this stage, we train two NER models (θ_1, θ_2) using the distant labels. These two models have different architectures in this work. Then, we duplicate these two models for the initialization of two sets of teacher networks, namely $\theta_{t1} = \theta_1$ and $\theta_{t2} = \theta_2$.

The training target of θ_1 and θ_2 is:

$$L(\theta) = -\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\overset{*}{y}{}_j^i log(p(y_j^i|X^i;\theta)) \qquad (1)$$

where M is the number of sentences in the training corpus and N is the token number in each sentence. $\overset{*}{y}{}_j^i$ means the distant label of j-th token of the i-th sentence.
(2) **Training student with adaptive teacher learning.** In this phrase, we select reliable labels by predictions of teachers from the first stage and supervise the students with cross-entropy loss. Meanwhile, considering the potential conflicts or competitions that exist among teachers, we investigate the diversity of teachers in the gradient space and recast the knowledge distillation from two teachers as a multi-objective optimization problem so that we can determine a better optimization direction for the training of student. To this end, an adaptive knowledge distillation loss is also adopted in this stage.
Reliable Labels Selection. Without any prior knowledge about which tokens are mislabeled or unlabeled, it is challenging to automatically detect them. Here we adopt two strategies to select reliable labels. (i) **Consistent Prediction.** The first token selection strategy is based on the pseudo labels prediction consistency between two teachers.

$$(X^i, Y^i)_{\text{CP}} = \{(x_j, y_j)|y_j = (y_{j,t1} == y_{j,t2})\} \qquad (2)$$

where $y_{j,t1}, y_{j,t2}$ are predicted one-hot pseudo labels on training corpus for two teachers. If two teacher models predict the same labels on specific tokens, then the labels of these tokens are set to corresponding labels. Meanwhile, if two teacher models have different predictions, the labels of tokens will be set to the “O” label. (ii) **Threshold Prediction**. We propose a simple threshold-based strategy to further filter reliable labels as the tokens with high confidence are more likely to be reliable. For teacher t_1,

$$(X^i, Y^i)_{\text{TP}} = \{(x_j, y_j)|\max(p_{j,t1}) > \sigma_1\} \qquad (3)$$

where σ_1 is the confidence threshold, $p_{j,t1}$ is the label distribution of the j-th token predicted by the teacher t_1. Thus, the tokens with label confidence lower than σ_1 will also be set to "O" labels. After these two steps, we can obtain reliable labels $\overline{Y}$. With these reliable labels, we can supervise the student models with the cross-entropy loss as follows:

$$L_{ce}(\theta) = -\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\overline{y}_j^i log(p(y_j^i|X_i;\theta)) \tag{4}$$

Adaptive Distillation The above selection procedure only considers consistent parts between two teachers while the conflicts among teachers are not squared up. To handle the inner conflicts, we formulate ensemble knowledge from teachers as a multi-objective optimization (MOO) problem (Sener and Koltun 2018) and use multiple gradient descent algorithms (MGDA) to probe a Pareto optimal solution that accommodates all teachers as much as possible.

Specifically, we first formally introduce the standard knowledge distillation loss which encourages the logits of the student network to mimic the teacher network:

$$\begin{aligned} L_{kd}^t(\theta) = H(p^s, p^t) = H(\sigma(a^s;T), \sigma(a^t;T)) = \\ -\sum_{k=1}^{K} p^t \mathrm{log} p^s[k] = -\left\langle p^t, \mathrm{log} p^s \right\rangle \end{aligned} \tag{5}$$

where σ is softmax operation, a^s and a^t are the logits of student and teacher networks, T is the temperature to soften the logits. K is the number of classification types. $H(\cdot,\cdot)$ is the cross-entropy loss to measure the discrepancy of softened probabilistic output between the student and teacher. In this work, we have two teachers, thus the naive solution for distilling from two teachers is:

$$L_{kd}(\theta) = L_{kd}^{t1}(\theta) + L_{kd}^{t2}(\theta) = H(p^s, p^{t1}) + H(p^s, p^{t2}) \tag{6}$$

However, conservatively accepting the directions from all teachers, *i.e.*, accumulating the separate distillation loss from each teacher, is not a good option, since the diversity of teachers could be significant and there might be some weak or noisy teachers mingled in the ensemble. When distilling knowledge from multiple teachers, we need to incorporate the disagreement into the determination of the descent direction. Recently, a novel method is proposed to find one single Pareto optimal solution with a good trade-off among conflicting optimization targets. Following (Sener and Koltun 2018; Lin et al. 2019), we can reformulate the Pareto solution of learning from two teachers as a linear scalarization of tasks with adaptive weight assignment as follows:

$$L(\theta) = \alpha_1 L_{kd}^{t1} + \alpha_2 L_{kd}^{t2} \tag{7}$$

where we adaptively assign the weights α_m by solving the following problem in each iteration:

$$\begin{aligned} \min \frac{1}{2}||\sum_{m=1}^{M}\alpha_m \nabla_\theta L_{kd}^m(\theta^\tau)||^2, s.t. \\ \sum_{m=1}^{M}\alpha_m = 1,\ 0 \leq \alpha_m \leq C, \forall m \in [1:M] \end{aligned} \tag{8}$$

where $C > 0$ is the regularization parameter, and M is the number of teachers. $L_{kd}^m(\theta^\tau)$ is the knowledge distillation loss at Eq. 5 corresponding to the student and m-th teacher. θ^τ is the parameter of the student network at iteration τ. Considering that calculating the gradient over parameters θ^τ can be fairly time-consuming. Following (Sener and Koltun 2018), we turn to its upper bound:

$$\min \frac{1}{2}||\sum_{m=1}^{M}\alpha_m \nabla_Z L_m(\theta^\tau)||^2, s.t. \tag{9}$$

where $\sum_{m=1}^{M}\alpha_m = 1,\ 0 \leq \alpha_m \leq C, \forall m \in [1:M]$, Z is the feature over the corresponding teacher. In this way, Eq. 9 is a typical One-class SVM problem and can be solved by LIBSVM (Chang and Lin 2011). More intuitively, as shown in Figure 2, we first compute the standard distillation loss according to the student logit and each teacher logit. Through the back-propagation algorithm, we can obtain the gradients corresponding to each teacher for the student model. Subsequently, we solve the loss weights through the gradients with the LIBSVM tool.

Finally, the total training loss for the student model in the second stage is:

$$L(\theta) = L_{ce}(\theta) + \alpha L_{kd}^{t1}(\theta) + (1-\alpha)L_{kd}^{t2}(\theta) \tag{10}$$

(3) **Updating teacher with fine-grained student ensemble.** After training the students, we devise a fine-grained student ensemble to update the parameters of the teachers. Before describing the concrete fine-grained ensemble, we first introduce a preliminary version, named **segment ensemble (SE)**. During each iteration, the segment ensemble picks up some units of the student model to replace the corresponding units of the teacher model, leaving the remaining parts of the teacher unchanged. Formally, at iteration τ,

$$\theta^t(\tau) = \{|P_i < \sigma_2|\theta_i^t(\tau-1) + (1 - |P_i < \sigma_2|)\theta_i^s(\tau)\} \tag{11}$$

where P_i is random probability distribution in [0,1] for the i-th unit of the teacher which is independent of each other. If $P_i < \sigma_2$, then $|P_i < \sigma_2| = 1$, the i-th unit parameter of teacher is to be preserved. In our paper, each unit corresponds to one network layer of the student network. The motivation of our segment ensemble is from Dropout (Srivastava et al. 2014) while Dropout works when training a network.

Subsequently, the segment ensemble can further integrate with EMA to incorporate temporal property. Here we first review the traditional EMA strategy:

$$\theta^t(\tau) = \{m\theta^t(\tau-1) + (1-m)\theta^s(\tau)\} \tag{12}$$

where m denotes the smoothing coefficient. As shown in this equation, EMA treats the model as a whole. We can integrate these two ensemble methods as fine-grained ensemble:

$$\begin{aligned} \theta^\tau(\tau) = \{|P_i < \sigma_2|\theta^t(\tau-1) + (1 - |P_i < \sigma_2|)m\theta_i^t(\tau-1) \\ + (1 - |P_i < \sigma_2|)(1-m)\theta_i^s(\tau)\} \end{aligned} \tag{13}$$

Algorithm 1: ATSEN training.

```
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.
$\theta_{t_1} \leftarrow \theta_1, \theta_{t_2} \leftarrow \theta_2, step \leftarrow 0$.
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-Training.
    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 reliable tokens by Eq. 2 and Eq. 3:
    $\mathcal{T}_I^{(b)} \leftarrow$ TokenSelection$(Y_I^{(b)}, \tilde{Y}_I^{(b)})$,
    $\mathcal{T}_{II}^{(b)} \leftarrow$ TokenSelection$(Y_{II}^{(b)}, \tilde{Y}_{II}^{(b)})$.
    Update the student $\theta_{s_1}$ and $\theta_{s_2}$ by Eq. 10.
    Update the teacher $\theta_{t_1}$ and $\theta_{t_2}$ by Eq. 13.
    Update noisy labels mutually:
    $Y_I = \{Y_i \leftarrow f(X_i; \theta_{t_2})\}_{i=1}^{M}$,
    $Y_{II} = \{Y_i \leftarrow f(X_i; \theta_{t_1})\}_{i=1}^{M}$.
end while
Evaluate models $\theta_{t_1}, \theta_{s_1}, \theta_{t_2}, \theta_{s_2}$ on Dev set.
return The best model $\theta \in \{\theta_{t_1}, \theta_{s_1}, \theta_{t_2}, \theta_{s_2}\}$
```

when $m = 0$, it becomes segment ensemble. Similarly, it degenerates to EMA when $\sigma_2 = 0$. In this manner, the fine-grained ensemble not only possesses the temporal property of traditional EMA, but also enhances the robustness of each segment to noise. As a result, the teacher tends to generate more reliable pseudo labels, which can be used as new supervision signals in the next round self-training.

To sum up, the first stage is executed once for a moderate initialization with distant labels. The second and third phases will be conducted alternately in a loop for better student and teacher models. Finally, only the best model $\theta \in \{\theta_{t1}, \theta_{t2}, \theta_{s1}, \theta_{s2}\}$ will be used for prediction.

The details of our model are presented in Algorithm 1.

Experiments

Dataset		Train	Dev	Test
CoNLL03	**Sentence**	14041	3250	3453
	Token	203621	51362	46435
OntoNotes5.0	**Sentence**	115812	15680	12217
	Token	2200865	304701	230118
Webpage	**Sentence**	385	99	135
	Token	5293	1121	1131
Twitter	**Sentence**	2393	999	3844
	Token	44076	15262	58064

Table 1: The statistics of four DS-NER datasets.

Datasets

To verify the effectiveness of our proposed ATSEN, we conduct experiments on four DS-NER datasets. Here we give a short description of them as follows:

CoNLL03 (Sang and De Meulder 2003) consists of 1393 English news articles and is annotated with four entity types: person, location, organization, and miscellaneous.

OntoNotes 5.0 (Weischedel et al. 2013) contains documents from multiple domains, including broadcast conversation, P2.5 data, and Web data. It consists of 18 entity types.

Webpage (Ratinov and Roth 2009) comprises of personal, academic, and computer science conference webpages. It consists of 20 webpages that cover 783 entities.

Twitter (Godin et al. 2015) is from the WNUT 2016 NER shared task. It consists of 10 entity types.

The detailed statistics of each dataset are listed in Table 1.

Compared Methods

We compare our ATSEN with a wide range of state-of-the-art DS-NER methods and supervised methods. Fully supervised methods use the ground truth annotation for model training. DS-NER methods use the distantly-labeled training set provided in (Liang et al. 2020).

Fully-supervised Methods. We include two supervised NER methods for comparison. (1) RoBERTa (Liu et al. 2019) adopts RoBERTa model as backbone and a top linear layer for token-level classification. (2) BiLSTM-CRF (Ma and Hovy 2016) uses bi-directional LSTM with character-level CNN to produce token embeddings, which are then fed into a CRF layer to predict token labels.

Distantly-supervised Methods. (1) KB-Matching reports the distant supervision quality. (2) Distant BiLSTM-CRF, Distant DistilRoBERTa, and Distant RoBERTa fine-tune the corresponding models on distantly-labeled data as if they are ground truth with the standard supervised learning. (3) AutoNER (Shang et al. 2018) trains the model by assigning ambiguous tokens with all possible labels and then maximizing the overall likelihood using a fuzzy CRF model. LRNT (Cao et al. 2019) applies partial-CRFs on high-quality data with non-entity sampling. Co-teaching+ (Yu et al. 2019) is a classic de-nosing method in computer vision. NegSampling (Li, Shi et al. 2020) only handles incomplete annotations by negative sampling. BOND and SCDL both adopt self-training strategies that are straightforward competitors to ATSEN.

Implementation Details

The architecture of the teachers is the backbone language model and a top classification layer for token-level classification. Specifically, we adopt RoBERTa and DistilRoBERTa as backbone for teacher 1 and teacher 2. The corresponding student has the same architecture as their teacher. The max training epoch is 50 for all datasets. The training batch size is 16 for CoNLL03, Webpage, and Twitter and 32 for OntoNotes 5.0. The learning rate is set to 1e-5 for CoNLL03 and Webpage, and 2e-5 for OntoNotes 5.0 and Twitter. For the pretraining stage with noisy labels, we separately train 1, 2, 12, and 6 epochs for CoNLL03, OntoNotes 5.0, Webpage,

Method	CoNLL03			OntoNotes 5.0			Webpage			Twitter		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BiLSTM-CRF♣	91.35	91.06	91.21	85.99	86.36	86.17	50.07	54.76	52.34	60.01	46.16	52.18
RoBERTa♣*	90.61	91.72	91.22	84.59	87.88	86.20	66.29	79.73	72.39	57.32	51.85	54.45
KB-Matching	81.13	63.75	71.40	63.86	55.71	59.51	62.59	45.14	52.45	40.34	32.22	35.83
Diatant BiLSTM-CRF	75.50	49.10	59.50	**68.44**	64.50	66.41	58.05	34.59	43.34	46.91	14.18	21.77
Distant DistilRoBERTa	77.87	69.91	73.68	66.83	68.81	67.80	56.05	59.46	57.70	45.72	43.85	44.77
Distant RoBERTa	82.29	70.47	75.93	66.99	69.51	68.23	59.24	62.84	60.98	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	43.26	18.69	26.10
LRNT	79.91	61.87	69.74	67.36	68.02	67.69	46.70	48.83	47.74	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	51.67	42.66	46.73
NegSampling	80.17	77.72	78.93	64.59	**72.39**	68.26	70.16	58.78	63.97	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	53.16	43.76	48.01
SCDL	**87.96**	79.82	83.69	67.49	69.77	68.61	68.71	68.24	68.47	59.87	44.57	51.09
ATSEN	86.14	**85.05**	**85.59**	66.97	71.05	**68.95**	**71.08**	**70.03**	**70.55**	**62.32**	**45.30**	**52.46**

Table 2: Main results on four benchmark datasets measured by precision (P), recall (R) and F1 scores. Baselines are reported by (Zhang et al. 2021). ♣ marks the model trained on the fully clean dataset. * denotes models implemented by us.

and Twitter datasets. For adaptive teacher learning, the confidence threshold σ_1 is 0.9 for all datasets. In the fine-grained student ensemble, m are 0.995, 0.995, 0.99, 0.995 and σ_2 is set to 0.8, 0.995, 0.8, and 0.75 for dataset CoNLL03, OntoNotes 5.0, Webpage, and Twitter, respectively.

Main Results

Table 2 presents the performance of all methods measured by precision, recall, and F1 scores. The results are summarized as follows: On all four datasets, ATSEN achieves the best performance among all distantly-supervised methods. Specifically, the distant DistilRoBERTa and RoBERTa only slightly improve the distant labeling performance compared to the naive KB-Matching, showing that directly applying supervised learning to distantly-labeled data will lead to poor model generalization. In addition, ATSEN performs much better than previous studies which consider the noisy labels in NER, including AutoNER, LRNT, Co-teaching+, and NegSampling. When compared to strong self-training methods BOND and SCDL, our ATSEN achieves new state-of-the-art performance, demonstrating the superiority of our proposed adaptive teacher learning and fine-grained student ensemble when trained on distantly-labeled data. Concretely, on CoNLL03, ATSEN achieves 1.90 absolute F1 improvements over the strong method SCDL. On the biggest and most difficult dataset OntoNotes V5.0, we obtain a decent improvement compared to the SOTA approach SCDL by 0.34 F1 score. In addition, we get 2.08 and 1.37 F1 scores improvement on Webpage and Twitter respectively.

Ablation Study

To further validate the effectiveness of each component in our ATSEN, we compare ATSEN with the following ablations by removing specific components: (1) remove the consistent prediction (w/o CP) in Eq.2. (2) remove the threshold prediction (w/o TP) in Eq.3. (3) do not perform cross-entropy loss (w/o CE). (4) do not perform adaptive distillation (w/o AD), namely, only cross-entropy loss is adopted in Eq.10. (5) do not perform fine-grained ensemble (w/o FE), namely, directly copy the trained student as a new teacher.

Ablations	**Precision**	**Recall**	**F1**
ATSEN	86.14	85.05	85.59
w/o CP	83.48	81.67	82.56
w/o TP	84.66	85.83	83.54
w/o CE	86.05	84.37	85.20
w/o AD	87.96	82.14	84.95
w/o FE	83.57	84.66	84.11

Table 3: Ablation study on CoNLL03 dataset. We compare ATSEN with ablations by removing specific components.

As shown in Table 3, it can be observed that w/o CP and w/o TP lead to a significant performance drop, indicating these strategies are important for cross-entropy learning. Meanwhile, the result of w/o CE do not cause huge performance as adaptive distillation also considers the agreement part between teachers. The results from w/o CP, w/o TP, and w/o CE also imply that the cross-entropy loss from ambiguous labels may damage the performance.

In addition, the result of w/o AD decreases the recall largely compared to ATSEN. It shows that considering knowledge from the disagreement part of two teachers can effectively help comprehensive student learning. Finally, w/o FE significantly reduces performance, showing that our fine-grained ensemble indeed benefits the model's generalization ability.

Study of Adaptive Distillation

In this section, we study the effectiveness of adaptive distillation for the student training process. Here we implement several ablations as shown in Table 4. The baseline is the adaptive teacher learning used in Eq.10, where the weight α is computed by LIBSVM from the gradients of two teachers corresponding to the student. We devise four variants. The first variant is averaging the distillation loss to substi-

Ablations	F1 Score
Baseline (adaptive distillation)	85.59
average distillation	85.19
manually weighted distillation	85.22
dynamically weighted distillation	85.26
disagreement distillation	84.86

Table 4: F1 scores of different variants on CoNLL03 dataset.

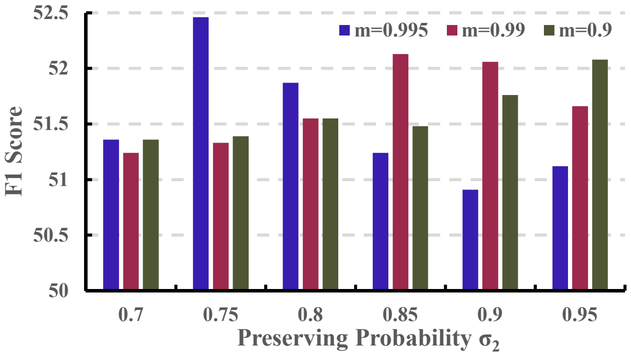


Figure 3: F1 score on Twitter dataset of different variants.

tute the adaptive distillation. It decreases by about 0.4 F1 scores. We also devise the second variant by manually setting the weights. Here we set 0.7 and 0.3 for distillation loss from teacher 1 (Roberta) and teacher 2 (DistilRoBERTa). This variant still performs worse than adaptive distillation. We have also tried other weight combinations such as 0.8 and 0.2 but achieved even worse results. Furthermore, we directly learn a dynamic weight α and achieve similar results with the manual setting. Finally, we consider a variant that only considers the disagreement part between two teachers during distillation, thus the training tokens may not be continuous and complete. The result presents a performance drop, indicating that the gradient optimization for adaptive distillation should be conducted for the whole sentence.

Study of Fine-grained Ensemble

We investigate the effectiveness of different student ensemble methods. For comprehensive evaluation, we experiment on a relatively smaller dataset Twitter instead of CoNLL03. As shown in Table 5: (1) remove all ensemble strategies (w/o all) and directly copy the student as a new teacher. (2) remove the segment ensemble (w/o SE), namely $\sigma_2 = 0$ in Eq.13. (3) remove the EMA (w/o EMA), namely $m = 0$ in Eq.13. As shown in Table 5, w/o all lead to the most significant performance drop. Meanwhile, removing either SE or EMA cause decreased results, demonstrating these two kinds of ensemble method can complement each other. It is worth noting ATSEN achieves significantly better precision than variants, indicating fine-grained ensemble can effectively enhance consistent predictions by performing on model fragments. Furthermore, we investigate the parameter influence of fine-grained ensemble in Fig. 3. As shown in this figure, we can observe $m = 0.995$ and $\sigma_2 = 0.75$ achieve the best performance. We also notice an interesting fact is that with the increase of m, the model achieves its best performance at a relatively smaller value of σ_2.

Ablations		**Precision**	**Recall**	**F1**
ATSEN		62.32	45.30	52.46
	w/o all	55.85	42.30	48.14
	w/o SE	58.90	45.35	51.25
	w/o EMA	59.67	46.65	52.36

Table 5: Ablation study on Twitter dataset. We compare our full method ATSEN with several ensemble strategy variants.

Distant: Johnson$_{\text{PER}}$ is to be hospitalized after California Angels$_{\text{PER}}$ skipper [John]$_{\text{PER}}$ McNamara was admitted to New [York]$_{\text{PER}}$'s [Columbia]$_{\text{PER}}$ Hospital. **Golden**: [Johnson]$_{\text{PER}}$ is to be hospitalized after [California Angels]$_{\text{ORG}}$ skipper [John McNamara]$_{\text{PER}}$ was admitted to [New York]$_{\text{LOC}}$'s [Columbia Hospital]$_{\text{ORG}}$.
BOND: [Johnson]$_{\text{PER}}$ is to be hospitalized after [California]$_{\text{LOC}}$ [Angels]$_{\text{PER}}$ skipper [John McNamara]$_{\text{PER}}$ was admitted to [New York]$_{\text{LOC}}$'s [Columbia]$_{\text{PER}}$ Hospital. **SCDL**: [Johnson]$_{\text{PER}}$ is to be hospitalized after [California]$_{\text{LOC}}$ [Angels]$_{\text{PER}}$ skipper [John McNamara]$_{\text{PER}}$ was admitted to [New York]$_{\text{LOC}}$'s [Columbia Hospital]$_{\text{ORG}}$. **ATSEN**: [Johnson]$_{\text{PER}}$ is to be hospitalized after [California Angels]$_{\text{ORG}}$ skipper [John McNamara]$_{\text{PER}}$ was admitted to [New York]$_{\text{LOC}}$'s [Columbia Hospital]$_{\text{ORG}}$.

Table 6: Case study. The sentence is from CoNLL03 dataset.

Case Study

We perform case study to understand the advantage of our proposed ATSEN in Table 6. We show the prediction result of BOND, SCDL, and ATSEN on a sentence with label noise. BOND can slightly generalize to unseen mentions and relieve partial incomplete annotation. For example, BOND can locate the "John McNamara" and "New York" while distant labels only can match partial person names. SCDL is able to generalize better for more accurate entity detection because it has a co-training step. For instance, SCDL can further locate the entity "Columbia Presby Hospital". However, it is still impacted by label noise. For comparision, for hard labels "California Angels", our ATSEN is able to detect them with both adaptive teacher learning and fine-grained student ensemble, instead of relying purely on distant labels.

Conclusion

In this paper, we present a novel self-training framework ATSEN for DS-NER. Specifically, ATSEN adopts adaptive teacher learning to train student networks, considering both consistent and inconsistent predictions between them. Furthermore, we devise a fine-grained student ensemble to update the teacher model. With it, each fragment of the teacher benefits from a temporal moving average of the corresponding fragment of the student. The experiment results illustrate that ATSEN significantly outperforms SOTA methods.

Acknowledgments

Pan Zhou is funded by the National Natural Science Foundation of China (NSFC) with grant numbers 61972448.

References

Abbasi Koohpayegani, S.; Tejankar, A.; and Pirsiavash, H. 2020. Compress: Self-supervised learning by compressing representations. *Advances in Neural Information Processing Systems*, 33: 12980–12992.

Cao, Y.; Hu, Z.; Chua, T.-s.; Liu, Z.; and Ji, H. 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)*, 261–270. Hong Kong, China: Association for Computational Linguistics.

Chang, C.-C.; and Lin, C.-J. 2011. LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3): 1–27.

Cheng, Q.; Liu, J.; Qu, X.; Zhao, J.; Liang, J.; Wang, Z.; Huai, B.; Yuan, N. J.; and Xiao, Y. 2021. HacRED: A Large-Scale Relation Extraction Dataset Toward Hard Cases in Practical Applications. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2819–2831.

Clark, K.; and Manning, C. D. 2016. Improving Coreference Resolution by Learning Entity-Level Distributed Representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 643–653. Berlin, Germany: Association for Computational Linguistics.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Godin, F.; Vandersmissen, B.; De Neve, W.; and Van de Walle, R. 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*, 146–153.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33: 21271–21284.

Gu, Y.; Qu, X.; Wang, Z.; Huai, B.; Yuan, N. J.; and Gui, X. 2021. Read, retrospect, select: An MRC framework to short text entity linking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 12920–12928.

Gu, Y.; Qu, X.; Wang, Z.; Zheng, Y.; Huai, B.; and Yuan, N. J. 2022. Delving Deep into Regularity: A Simple but Effective Method for Chinese Named Entity Recognition. *arXiv preprint arXiv:2204.05544*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

Hinton, G.; Vinyals, O.; Dean, J.; et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7).

Huang, Z.; Xu, W.; and Yu, K. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Huo, X.; Xie, L.; He, J.; Yang, Z.; Zhou, W.; Li, H.; and Tian, Q. 2021. ATSO: Asynchronous teacher-student optimization for semi-supervised image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1235–1244.

Li, J.; Fei, H.; Liu, J.; Wu, S.; Zhang, M.; Teng, C.; Ji, D.; and Li, F. 2022. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 10965–10973.

Li, X.; Feng, J.; Meng, Y.; Han, Q.; Wu, F.; and Li, J. 2020. A Unified MRC Framework for Named Entity Recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5849–5859.

Li, Y.; Shi, S.; et al. 2020. Empirical Analysis of Unlabeled Entity Problem in Named Entity Recognition. In *International Conference on Learning Representations*.

Liang, C.; Yu, Y.; Jiang, H.; Er, S.; Wang, R.; Zhao, T.; and Zhang, C. 2020. Bond: Bert-assisted open-domain named entity recognition with distant supervision. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1054–1064.

Lin, X.; Zhen, H.-L.; Li, Z.; Zhang, Q.-F.; and Kwong, S. 2019. Pareto multi-task learning. *Advances in neural information processing systems*, 32.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ma, X.; and Hovy, E. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*.

Peng, M.; Xing, X.; Zhang, Q.; Fu, J.; and Huang, X. 2019. Distantly supervised named entity recognition using positive-unlabeled learning. *arXiv preprint arXiv:1906.01378*.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep Contextualized Word Representations. 2227–2237. New Orleans, Louisiana: Association for Computational Linguistics.

Qu, X.; Zou, Z.; Cheng, Y.; Yang, Y.; and Zhou, P. 2019. Adversarial category alignment network for cross-domain sentiment classification. 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)*, 2496–2508.

Ratinov, L.; and Roth, D. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the thirteenth conference on computational natural language learning (CoNLL-2009)*, 147–155.

Sang, E. F.; and De Meulder, F. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Sener, O.; and Koltun, V. 2018. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31.

Shang, J.; Liu, L.; Ren, X.; Gu, X.; Ren, T.; and Han, J. 2018. Learning named entity tagger using domain-specific dictionary. *arXiv preprint arXiv:1809.03599*.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1): 1929–1958.

Weischedel, R.; Palmer, M.; Marcus, M.; Hovy, E.; Pradhan, S.; Ramshaw, L.; Xue, N.; Taylor, A.; Kaufman, J.; Franchini, M.; et al. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 23.

Xiao, L.; Qu, X.; Li, R.; Wang, J.; Zhou, P.; and Li, Y. 2020. Fine-Grained Text Sentiment Transfer via Dependency Parsing. In *ECAI 2020*, 2228–2235. IOS Press.

Yu, X.; Han, B.; Yao, J.; Niu, G.; Tsang, I.; and Sugiyama, M. 2019. How does disagreement help generalization against label corruption? In *International Conference on Machine Learning*, 7164–7173. PMLR.

Zhang, X.; Yu, B.; Liu, T.; Zhang, Z.; Sheng, J.; Mengge, X.; and Xu, H. 2021. Improving Distantly-Supervised Named Entity Recognition with Self-Collaborative Denoising Learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 10746–10757. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Zhou, K.; Li, Y.; and Li, Q. 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)*, 7198–7211. Dublin, Ireland: Association for Computational Linguistics.

Zhu, T.; Qu, X.; Chen, W.; Wang, Z.; Huai, B.; Yuan, N.; and Zhang, M. 2022. Efficient Document-level Event Extraction via Pseudo-Trigger-aware Pruned Complete Graph. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, 4552–4558. International Joint Conferences on Artificial Intelligence Organization.