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.

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.

---

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 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) 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 Adaptive Teacher Learning and Fine-grained Student Ensemble (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
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 ($\theta_1$, $\theta_2$) 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 $\theta_{1t} = \theta_1$ and $\theta_{2t} = \theta_2$. (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, 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. (3) **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 $t_1$,

$$\{(X^i, Y^i)_{TP} = \{(x_j, y_j) | \max(p_{j, t_1}) > \sigma_1\} \quad (3)$$

The training target of $\theta_1$ and $\theta_2$ is:

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

(1)

where $M$ is the number of sentences in the training corpus and $N$ is the number token in each sentence. $y^i_j$ 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.

$$\{(X^i, Y^i)_{CP} = \{(x_j, y_j) | y_j = (y_{j, t_1} == y_{j, t_2})\} \quad (2)$$

where $y_{j, t_1}, y_{j, t_2}$ 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 $t_1$, .

$$\{(X^i, Y^i)_{TP} = \{(x_j, y_j) | \max(p_{j, t_1}) > \sigma_1\} \quad (3)$$
where $\sigma_1$ is the confidence threshold, $p_{j\mid t_1}$ is the label distribution of the $j$-th token predicted by the teacher $t_1$. Thus, the tokens with label confidence lower than $\sigma_1$ 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:

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

(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:

$$L_{kd}(\theta) = H(p^s, p^t) = H(\sigma(a^s; T), \sigma(a^t; T)) = - \sum_{k=1}^{K} p^s_k \log p^t_k = - \langle p^s, \log p^t \rangle$$

(5)

where $\sigma$ 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:

$$L_{kd}(\theta) = L_{kd}^{1}(\theta) + L_{kd}^{2}(\theta) = H(p^s, p^{1\mid t}) + H(p^s, p^{2\mid t})$$

(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:

$$L(\theta) = \alpha_1 L_{kd}^{1}(\theta) + \alpha_2 L_{kd}^{2}(\theta)$$

(7)

where we adaptively assign the weights $\alpha_m$ by solving the following problem in each iteration:

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

(8)

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

(9)

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. $\theta^{\tau}$ is the parameter of the student network at iteration $\tau$. Considering that calculating the gradient over parameters $\theta^{\tau}$ can be fairly time-consuming. Following (Sener and Koltun 2018), we turn to its upper bound:

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

(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 $\tau$,

$$\theta^{\tau}(\tau) = \{P_i < \sigma_2 \theta^s_i(\tau - 1) + (1 - P_i < \sigma_2) \theta^s_i(\tau)\}$$

(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:

$$\theta^e(\tau) = \{m \theta^{\tau} - (1 - m) \theta^e(\tau)\}$$

(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:

$$\theta^{\tau}(\tau) = \{P_i < \sigma_2 \theta^{\tau}(\tau - 1) + (1 - P_i < \sigma_2) m \theta^s_i(\tau - 1) + (1 - P_i < \sigma_2)(1 - m) \theta^s_i(\tau)\}$$

(13)
Algorithm 1: ATSEN training.

**Input:** Training corpus \( D = \{ (X_i, Y_i) \}_{i=1}^{M} \) with noisy labels

**Parameter:** Two network parameters \( \theta_1, \theta_2, \theta_{s1}, \theta_{s2} \), and \( \theta_{s3} \)

**Output:** The best model

1. Pre-training two models \( \theta_1, \theta_2 \) with \( D \). \( \triangleright \)Pre-Training.
2. \( \theta_{s1} \leftarrow \theta_1, \theta_{s2} \leftarrow \theta_2, \) step \( \leftarrow 0 \).
3. Initialize noisy labels: \( Y_{I} \leftarrow Y, Y_{II} \leftarrow Y \).
4. **while** not reach max training epochs **do**
5. Get a batch \( \{ (X^{(b)}, Y^{(b)}_I), Y^{(b)}_{II} \} \) from \( D \), step \( \leftarrow \) step + 1. \( \triangleright \)Self-Training.
6. Get pseudo-labels via the teacher \( \theta_{s1}, \theta_{s2} \):
   \( \hat{Y}^{(b)}_I \leftarrow f(X^{(b)}; \theta_{s1}), \)
   \( \hat{Y}^{(b)}_{II} \leftarrow f(X^{(b)}; \theta_{s2}). \)
7. Get reliable tokens by Eq. 2 and Eq. 3:
   \( T^{(b)}_I \leftarrow \text{TokenSelection}(Y^{(b)}_I, \hat{Y}^{(b)}_I), \)
   \( T^{(b)}_{II} \leftarrow \text{TokenSelection}(Y^{(b)}_{II}, \hat{Y}^{(b)}_{II}). \)
8. Update the student \( \theta_{s1} \) and \( \theta_{s2} \) by Eq. 10.
9. Update the teacher \( \theta_1 \) and \( \theta_2 \) by Eq. 13.
10. Update noisy labels mutually:
    \( Y_I = \{ Y_i - f(X_i; \theta_{s1}) \}_{i=1}^{M}, \)
    \( Y_{II} = \{ Y_i - f(X_i; \theta_{s2}) \}_{i=1}^{M}. \)
11. **end while**
12. Evaluate models \( \theta_1, \theta_2, \theta_{s1}, \theta_{s2} \) on Dev set.
13. **return** The best model \( \theta \in \{ \theta_1, \theta_{s1}, \theta_{s2} \} \)

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 \( \theta \in \{ \theta_1, \theta_{s1}, \theta_{s2} \} \) will be used for prediction.

The details of our model are presented in Algorithm 1.

**Experiments**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CoNLL03</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentence</td>
<td>14041</td>
<td>3250</td>
<td>3453</td>
</tr>
<tr>
<td>Token</td>
<td>203621</td>
<td>51362</td>
<td>46435</td>
</tr>
<tr>
<td><strong>OntoNotes5.0</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentence</td>
<td>115812</td>
<td>20866</td>
<td>40701</td>
</tr>
<tr>
<td>Token</td>
<td>385</td>
<td>99</td>
<td>135</td>
</tr>
<tr>
<td><strong>Webpage</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentence</td>
<td>5293</td>
<td>1121</td>
<td>1131</td>
</tr>
<tr>
<td>Token</td>
<td>2393</td>
<td>999</td>
<td>3844</td>
</tr>
<tr>
<td><strong>Twitter</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentence</td>
<td>44076</td>
<td>15262</td>
<td>58064</td>
</tr>
<tr>
<td>Token</td>
<td>203621</td>
<td>51362</td>
<td>46435</td>
</tr>
</tbody>
</table>

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 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 DistillRoBERTa 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,
and Twitter datasets. For adaptive teacher learning, the confidence threshold $\sigma_1$ is 0.9 for all datasets. In the fine-grained student ensemble, $m$ are 0.995, 0.995, 0.99, 0.995 and $\sigma_2$ 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 (P), recall (R) 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.

### 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 $\alpha$ 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 substitute

<table>
<thead>
<tr>
<th>Method</th>
<th>CoNLL03 P</th>
<th>CoNLL03 R</th>
<th>CoNLL03 F</th>
<th>OntoNotes 5.0 P</th>
<th>OntoNotes 5.0 R</th>
<th>OntoNotes 5.0 F</th>
<th>Webpage P</th>
<th>Webpage R</th>
<th>Webpage F</th>
<th>Twitter P</th>
<th>Twitter R</th>
<th>Twitter F</th>
</tr>
</thead>
<tbody>
<tr>
<td>BiLSTM-CRF♣</td>
<td>91.35</td>
<td>91.06</td>
<td>91.21</td>
<td>85.99</td>
<td>86.36</td>
<td>86.17</td>
<td>30.07</td>
<td>54.76</td>
<td>52.34</td>
<td>60.01</td>
<td>46.16</td>
<td>52.18</td>
</tr>
<tr>
<td>RoBERTa♣*</td>
<td>90.61</td>
<td>91.72</td>
<td>91.22</td>
<td>84.59</td>
<td>87.88</td>
<td>86.20</td>
<td>66.29</td>
<td>79.73</td>
<td>72.39</td>
<td>57.32</td>
<td>51.85</td>
<td>54.45</td>
</tr>
<tr>
<td>KB-Matching</td>
<td>81.13</td>
<td>63.75</td>
<td>71.40</td>
<td>63.86</td>
<td>55.71</td>
<td>59.51</td>
<td>62.59</td>
<td>45.14</td>
<td>52.45</td>
<td>40.34</td>
<td>32.22</td>
<td>35.83</td>
</tr>
<tr>
<td>Distant BiLSTM-CRF</td>
<td>75.50</td>
<td>49.10</td>
<td>59.50</td>
<td>68.44</td>
<td>64.50</td>
<td>66.41</td>
<td>58.05</td>
<td>34.59</td>
<td>43.34</td>
<td>46.91</td>
<td>14.18</td>
<td>21.77</td>
</tr>
<tr>
<td>Distant DistilRoBERTa</td>
<td>77.87</td>
<td>69.91</td>
<td>73.68</td>
<td>66.83</td>
<td>68.81</td>
<td>67.80</td>
<td>56.05</td>
<td>59.46</td>
<td>57.70</td>
<td>45.72</td>
<td>43.85</td>
<td>44.77</td>
</tr>
<tr>
<td>Distant RoBERTa</td>
<td>82.29</td>
<td>70.47</td>
<td>75.93</td>
<td>69.99</td>
<td>69.51</td>
<td>68.23</td>
<td>59.24</td>
<td>62.84</td>
<td>60.98</td>
<td>50.97</td>
<td>42.66</td>
<td>46.45</td>
</tr>
<tr>
<td>AutoNER</td>
<td>75.21</td>
<td>60.40</td>
<td>67.00</td>
<td>64.63</td>
<td>69.95</td>
<td>67.18</td>
<td>48.82</td>
<td>54.23</td>
<td>51.39</td>
<td>43.26</td>
<td>18.69</td>
<td>26.10</td>
</tr>
<tr>
<td>LRNT</td>
<td>79.91</td>
<td>61.87</td>
<td>69.74</td>
<td>67.36</td>
<td>68.02</td>
<td>67.69</td>
<td>46.70</td>
<td>48.83</td>
<td>47.74</td>
<td>46.94</td>
<td>15.98</td>
<td>23.84</td>
</tr>
<tr>
<td>Co-teaching+</td>
<td>86.04</td>
<td>68.74</td>
<td>76.42</td>
<td>66.63</td>
<td>69.32</td>
<td>67.95</td>
<td>61.65</td>
<td>55.41</td>
<td>51.67</td>
<td>42.66</td>
<td>46.66</td>
<td>46.73</td>
</tr>
<tr>
<td>NegSampling</td>
<td>80.17</td>
<td>77.72</td>
<td>78.93</td>
<td>64.59</td>
<td>72.39</td>
<td>68.26</td>
<td>70.16</td>
<td>58.78</td>
<td>63.97</td>
<td>50.25</td>
<td>44.95</td>
<td>47.45</td>
</tr>
<tr>
<td>BOND</td>
<td>82.05</td>
<td>80.92</td>
<td>81.48</td>
<td>67.14</td>
<td>69.61</td>
<td>68.35</td>
<td>67.37</td>
<td>64.19</td>
<td>65.74</td>
<td>53.16</td>
<td>43.76</td>
<td>48.01</td>
</tr>
<tr>
<td>SCDL</td>
<td>87.96</td>
<td>79.82</td>
<td>83.69</td>
<td>67.49</td>
<td>69.77</td>
<td>68.61</td>
<td>68.71</td>
<td>68.24</td>
<td>64.87</td>
<td>59.87</td>
<td>44.57</td>
<td>51.09</td>
</tr>
</tbody>
</table>

| 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.

<table>
<thead>
<tr>
<th>Ablations</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATSEN</td>
<td>86.14</td>
<td>85.05</td>
<td>85.59</td>
</tr>
<tr>
<td>w/o CP</td>
<td>83.48</td>
<td>81.67</td>
<td>82.56</td>
</tr>
<tr>
<td>w/o TP</td>
<td>84.66</td>
<td>85.83</td>
<td>83.54</td>
</tr>
<tr>
<td>w/o CE</td>
<td>86.05</td>
<td>84.37</td>
<td>85.20</td>
</tr>
<tr>
<td>w/o AD</td>
<td>87.96</td>
<td>82.14</td>
<td>84.95</td>
</tr>
<tr>
<td>w/o FE</td>
<td>83.57</td>
<td>84.66</td>
<td>84.11</td>
</tr>
</tbody>
</table>

Table 3: Ablation study on CoNLL03 dataset. We compare ATSEN with ablations by removing specific components.
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 $\sigma_2$.

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 comparison, 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


