SANTA: Separate Strategies for Inaccurate and Incomplete Annotation Noise in Distantly-Supervised Named Entity Recognition

Shuzheng Si[1,2*], **Zefan Cai**[1,2*], **Shuang Zeng**[1,2],
Guoqiang Feng[1,2], **Jiaxing Lin**[1,2] **and Baobao Chang**[1†]
[1]National Key Laboratory for Multimedia Information Processing, Peking University
[2]School of Software and Microelectronics, Peking University, China
{sishuzheng,jxlin,fgq}@stu.pku.edu.cn, zefncai@gmail.com
{zengs,chbb}@pku.edu.cn

Abstract

Distantly-Supervised Named Entity Recognition effectively alleviates the burden of time-consuming and expensive annotation in the supervised setting. But the context-free matching process and the limited coverage of knowledge bases introduce inaccurate and incomplete annotation noise respectively. Previous studies either considered only incomplete annotation noise or indiscriminately handle two types of noise with the same strategy. In this paper, we argue that the different causes of two types of noise bring up the requirement of different strategies in model architecture. Therefore, we propose the **SANTA** to handle these two types of noise separately with (1) Memory-smoothed Focal Loss and Entity-aware KNN to relieve the entity ambiguity problem caused by inaccurate annotation, and (2) Boundary Mixup to alleviate decision boundary shifting problem caused by incomplete annotation and a noise-tolerant loss to improve the robustness. Benefiting from our separate tailored strategies, we confirm in the experiment that the two types of noise are well mitigated. SANTA also achieves a new state-of-the-art on five public datasets.

1 Introduction

As a fundamental task in NLP, Named Entity Recognition (NER) aims to locate and classify named entities in text, which plays an important role in many tasks such as knowledge graph construction (Peng et al., 2022; Li et al., 2022c) and relation extraction (Zeng et al., 2021; Wang et al., 2022). To alleviate the burden of annotation in the supervised setting, Distantly-Supervised Named Entity Recognition (DS-NER) is widely used in real-world scenarios. It can automatically generate labeled training data by matching entities in existing knowledge bases with snippets in plain text. However, DS-NER suffers from two inherent issues which introduce many noisy samples: (1) **inaccurate annotation**: the entity with multiple types in the knowledge bases may be labeled as an inaccurate type in the text, due to the context-free matching process, and (2) **incomplete annotation**: the knowledge bases with limited coverage of entities cannot label all entities in the text. As shown in Figure 1, the entity types of "Amazon" and "Washington" are wrongly labeled owing to context-free matching, and "410 Terry Ave N" is not recognized due to the limited coverage of knowledge bases.

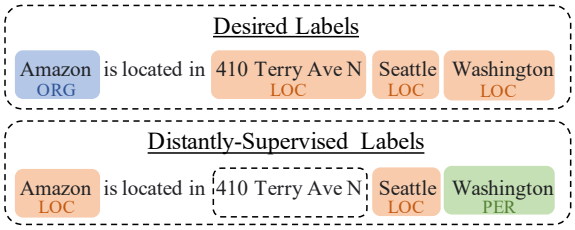

Figure 1: A sample generated by Distant Supervision. "Amazon" and "Washington" are inaccurate annotations. "410 Terry Ave N" is the incomplete annotation.

Due to the sensitivity to the noise, the original supervised methods achieve poor performances in DS-NER. Therefore, many works have been proposed to handle the issue. Some works attempted to focus on solving incomplete annotation noise in DS-NER, including positive-unlabeled (PU) learning (Peng et al., 2019; Zhou et al., 2022b), negative sampling (Li et al., 2021, 2022a), and retrieval augmented inference with contrastive learning (Si et al., 2022). However, the ignorance of inaccurate annotation noise limits the model to further improve the performance. Recently, Zhang et al. (2021a) jointly trained two teacher-student networks to handle the two types of noise with the same strategy. Meng et al. (2021) adopted a noise-robust learning scheme and self-training for the whole training set to avoid overfitting in noise. However, by handling both types of noise with the same strategy, these

*Equal contribution.
†Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2023, pages 3883–3896
July 9-14, 2023

methods failed to address the unique characteristics of different type of noise, thereby limiting their ability to effectively handle the noise in DS-NER.

As the causes of the two types of noise are different, both of two noise may lead to different problems in DS-NER task. Inaccurate annotation noise in training data can lead to serious entity ambiguity problem in the DS-NER. As exampled in Figure 1, when a model is trained on data where "Washington" is consistently labeled as "PER" (person) due to context-free matching, the model may continue to predict "Washington" as a "PER" even in contexts where it should be labeled as a "LOC" (location), such as when it refers to the city of Washington. Incomplete annotation noise can lead to the decision boundary shifting problem (Si et al., 2022) in the DS-NER task. This problem occurs when the model is trained on data where some entity spans are not labeled, causing the model to shift its decision boundary, making it more likely to predict an entity span as a non-entity type. Therefore, the noise in the spans labeled as entities by distant supervision leads to the ambiguity problem, and the noise in the spans labeled as non-entities leads decision boundary shifting problem. To further improve the performance in DS-NER, we argue that the two types of noise should be handled separately with specialized designs in model architecture. This can help the model to address the specific problems posed by each type of noise and lead to the better overall performance of the model.

In this paper, we propose the **S**eparate str**A**tegies for i**N**accurate and incomple**T**e **A**nnotation noise in DS-NER (**SANTA**). Unlike previous works in DS-NER, we introduce different strategies to handle the two types of noise respectively. For inaccurate annotation, we propose Memory-smoothed Focal Loss (MFL) and Entity-aware KNN. These strategies aim to address the entity ambiguity problem posed by inaccurate annotation noise. For incomplete annotation, we propose Boundary Mixup to handle decision boundary shifting problem caused by incomplete annotation, which generates augmented instances by combining the instances around the boundary and the entity instances. Due to further training on the augmented instances, the biased decision boundary can be pushed towards right (fully supervised) side as the augmented instances exist between the biased boundary and the right boundary. Meanwhile, we empirically analyze the characteristics of incomplete annotation noise, then adopt a noise-tolerant loss to further improve the model's robustness to this type of noise.

Experiments show SANTA achieves state-of-the-art on five public DS-NER datasets. Further analysis shows the effectiveness of each designed module and separate handling.

2 Related Work

To address data scarcity problem, several studies attempted to annotate datasets via distant supervision. Using external knowledge bases can easily get training data through string matching, but introduces two issues: inaccurate annotation noise and incomplete annotation noise. To address these issues, various methods have been proposed.

Only Focusing on Incomplete Annotation. Several studies (Shang et al., 2018; Yang et al., 2018; Jie et al., 2019) modified the standard CRF to get better performance under the noise, e.g., Partial CRF. LRNT (Cao et al., 2019) leaved training data unexplored fully to reduce the negative effect of noisy labels. AdaPU (Peng et al., 2019) employed PU learning to obtain unbiased estimation of the loss value. Furthermore, Conf-MPU (Zhou et al., 2022b) used multi-class PU learning to further improve the performance. Li et al. (2021) performed uniform negative sampling to mitigate the misguidance from unlabeled entities. Li et al. (2022a) then proposed a weighted sampling distribution to introduce direction to incomplete annotation when negative sampling. Si et al. (2022) adopt supervised contrastive-learning loss and retrieval-augmented inference to mitigate the decision boundary shifting problem. However, these studies only addressed incomplete annotation noise, ignoring inaccurate annotation noise, which also exists in DS-NER.

Handling Two Types of Noise With the Same Strategy. To further exploring the information in DS-NER text, many studies attempted to consider both inaccurate and incomplete annotation noise. BOND (Liang et al., 2020) designed a teacher-student network to drop unreliable labels and use pseudo labels to get more robust model. SCDL (Zhang et al., 2021b) further improved the performance by jointly training two teacher-student network and refining the distant labels. RoSTER (Meng et al., 2021) adopt a noise-robust learning scheme and self-training to improve the robustness. CReDEL (Ying et al., 2022) trained an automatic distant label refinement model via contrastive learn-

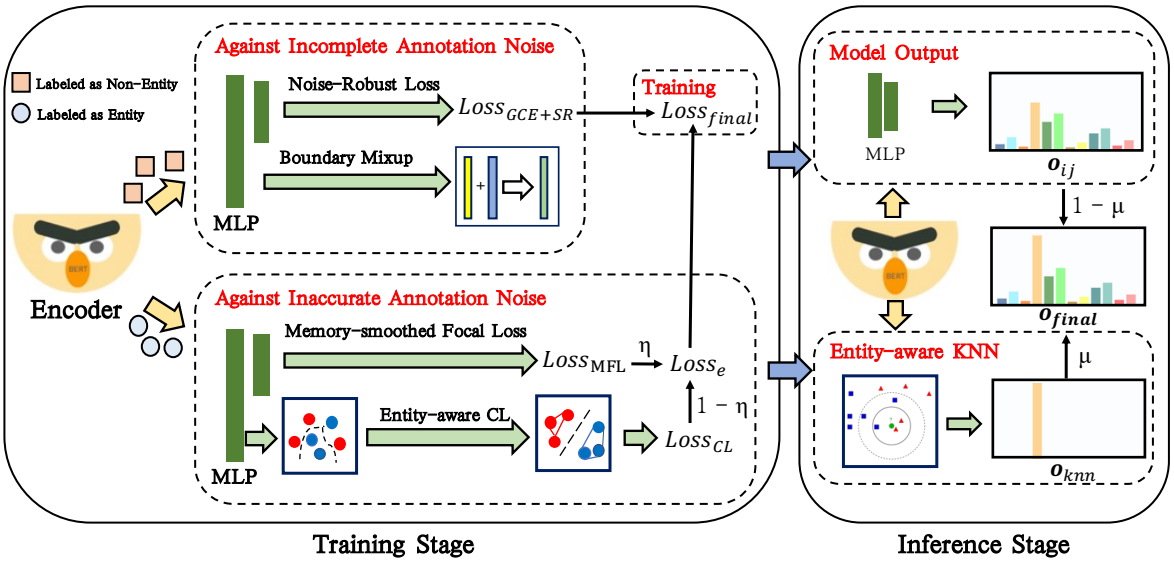

Figure 2: General architecture of SANTA.

ing as a plug-in module for other DS-NER models. Although these works jointly considered the two types of noise, they did not take into account the difference of the two types of noise. Due to the different causes and frequencies of the two different types of noise, we argue that they should be handled separately with specialized strategies, which have not been considered in previous work.

3 Method

As shown in Figure 2, SANTA separately handles the spans labeled as entities and labeled as non-entities. SANTA uses the MFL and Entity-aware KNN to address the entity ambiguity problem caused by inaccurate annotation. SANTA adopts noise-tolerant GCE + SR loss and Boundary Mixup to handle the incomplete annotation.

3.1 Span-based NER Model

We follow the same span-based NER model as Li et al. (2022b, 2020) and Si et al. (2022). For sentence $[x_1, x_2, ..., x_n]$, we use a pre-trained language model as an encoder to get the representations for every token x in the sentence:

$$[\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_n] = \text{Encoder}([\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]) \quad (1)$$

where h_i is the representation for token x_i.

For each span $s_{i,j}$ ranging from i-th token to j-th token, the span representation $\mathbf{s}_{i,j}$ is calculated as:

$$\mathbf{s}_{i,j} = \mathbf{h}_i \oplus \mathbf{h}_j \oplus (\mathbf{h}_i - \mathbf{h}_j) \oplus (\mathbf{h}_i \odot \mathbf{h}_j) \quad (2)$$

where $\oplus$ is the concatenation operation and $\odot$ is the element-wise product operation.

Then, we use linear layer and activation function to get a more dense representation:

$$\mathbf{r}_{i,j} = \tanh(W\mathbf{s}_{i,j}) \quad (3)$$

Finally, we can obtain the entity label distribution $o_{i,j}$ for every span $s_{i,j}$ as:

$$\mathbf{o}_{i,j} = \text{softmax}(V\mathbf{r}_{i,j}) \quad (4)$$

where W and V are trainable parameter.

3.2 Against Inaccurate Annotation Noise

We observe that the inaccurate annotation in spans labeled as entities leads to two problems: (1) it can cause fluctuations in the training process and make it difficult to achieve consistent predictions; (2) it can lead to the entity ambiguity problem.

3.2.1 Memory-smoothed Focal Loss

In the training process, the inaccurately labeled entities and the similar true entities supervise the model back and forth and cause fluctuations in the model's learning. For example, the "Washington" in Figure 1 is inaccurately labeled as "PER" (person), the model trained with it tends to predict "Washington" as "PER" instead of "LOC" (location). However, if the model is also exposed to similar and correctly labeled entities, such as "Seattle" labeled as "LOC", the model may also learn to generalize "Washington" as a "LOC". This back and forth supervision can make the model's performance being less consistent and less accurate.

Intuitively, if we can smooth out such fluctuations, the performance of model can be further improved.

This observation motivates us to propose Memory Label Smoothing (MLS) to reduce the fluctuation. MLS memorizes model predictions to continuously update the soft labels during the training stage. The memorized soft labels are used to smooth the training of every sample, which tries to enable the model jointly consider the label information from previous predictions, so that learning for a particular sample does not fluctuate tremendously. Given each span $x_{i,j}$ labeled as entity, if its prediction is correct, the soft label corresponding to the entity type $y_{i,j}$ will be updated using the model output $o_{i,j}$. Specifically, in training total T epochs, we construct $\hat{Y} = [\hat{Y}^0, ..., \hat{Y}^t, ..., \hat{Y}^T]$ as the memory soft labels at different training epochs. $\hat{Y}^t$ is a matrix with $|L|$ rows and $|L|$ columns, and each column in $\hat{Y}^t$ corresponds to the vector with latitude $|L|$ as soft label for one category. L denotes the set of entity types. For epoch $t > 0$, $\hat{Y}^t$ is defined as:

$$\hat{Y}^t_{y_{i,j},l} = \frac{1}{N} \sum_{l \in L} \mathbb{I}\{l = y_{i,j}\} o_{i,j} \tag{5}$$

where N denotes the number of correctly predicted entities with label $y_{i,j}$. $\mathbb{I}$ is indicator function. $\hat{Y}^0$ is initialize as identity matrix.

Then the updated soft labels will be utilized to supervise the model in the next epoch and smooth the learning curve of the model. In training epoch t, given a span $(x_{i,j}, y_{i,j})$, we use the previous G soft label $[\hat{Y}^{t-G}, ..., \hat{Y}^{t-1}]$ to supervise the model:

$$Y^t_{final} = \sum_{l \in L} (\lambda \mathbb{I}\{l = y_{i,j}\} + (1 - \lambda) \frac{1}{G} \sum_{g=1}^{G} \hat{Y}^{t-g}_{y_{i,j},l}) \tag{6}$$

where G and λ are hyperparameters. For $t < G$, we use all the previous soft labels to get Y^t_{final}.

Meanwhile, we adopt Focal Loss (Lin et al., 2020) to handle the entity ambiguity problem, which can be calculated as:

$$\mathcal{L}_{FL} = -\sum_{l \in L} \alpha (1 - o_{i,j})^{\gamma} \log(o_{i,j}) \tag{7}$$

where α and γ are hyperparameters. It can be seen that Focal Loss has a greater weight for ambiguous samples with low confidence.

Combined with the capability from MLS to reduce the fluctuation and the capability from Focal Loss to handle the ambiguous samples, the Memory-smoothed Focal Loss (MFL) can be defined as:

$$\mathcal{L}_{MFL} = \sum_{x_{i,j} \in D_e} \sum_{l \in L} \alpha (Y^t_{final})_l (1 - o_{i,j})^{\gamma} \log(o_{i,j}) \tag{8}$$

In addition, MFL is only performed on spans labeled as entities D_e to focus on the problems caused by inaccurate annotation.

3.2.2 Entity-aware KNN

The proposed MFL method may not completely solve the entity ambiguity caused by inaccurate annotation noise, and therefore relying solely on the output of the trained model may still result in limited performance. To address this issue, we propose Entity-aware KNN during the inference stage to facilitate the decoding process by retrieving similar labeled samples in the training set, which further improves the overall performance.

Entity-aware KNN consists of two parts, including Entity-aware Contrastive Learning (Entity-aware CL) and KNN-augmented Inference. Specifically, we use Entity-aware CL to close the distance between the span labeled as entity with the same type, pull the distance of different types. Therefore, we could get the better representation to easily use KNN-augmented Inference to get a retrieved distribution o_{knn}. Finally, we interpolate the output $o_{i,j}$ from model with o_{knn} to further handle the entity ambiguity problem.

To improve the performance of KNN retrieving, Entitiy-aware CL pulls spans belonging to the same entity type together in representation space, while simultaneously pushing apart clusters of spans from different entity types. Therefore, the type of entities could be better distinguished. We use the cosine similarity as metric between the representations $r_{i,j}$ and $r_{\hat{i},\hat{j}}$ of span $x_{i,j}$ and $x_{\hat{i},\hat{j}}$:

$$d_{s_{i,j}, s_{\hat{i},\hat{j}}} = \frac{\mathbf{r}_{i,j} \cdot \mathbf{r}_{\hat{i},\hat{j}}}{|\mathbf{r}_{i,j}||\mathbf{r}_{\hat{i},\hat{j}}|} \tag{9}$$

Then the Entity-aware CL $\mathcal{L}_{CL}$ is defined as:

$$\mathcal{L}_{CL} = -\sum_{l \in L_e} \sum_{r_{i,j} \in E_l} \frac{1}{N_l - 1} \sum_{r_{\hat{i},\hat{j}} \in E_{\bar{l}}} F(\mathbf{r}_{i,j}, \mathbf{r}_{\hat{i},\hat{j}}) \tag{10}$$

where L_e is the entity label set; N_l is the total number of spans with the same entity label l in the batch; E_l is the collection of all training spans with

l-th entity label. $F(\mathbf{r}_{i,j}, \mathbf{r}_{\hat{i},\hat{j}})$ is calculated as:

$$F(\mathbf{r}_{i,j}, \mathbf{r}_{\hat{i},\hat{j}}) = \log \frac{\exp(d_{r_{i,j}, r_{\hat{i},\hat{j}}}/\tau)}{\sum_{r_{m,n} \in E_{\bar{l}}} \exp(d_{r_{i,j}, r_{m,n}}/\tau)} \tag{11}$$

where τ is the temperature. $E_{\bar{l}}$ is the collection of labeled entity spans not with entity label l.

After we get the easily distinguishable entity span representations, we propose KNN-augmented Inference to further relief the ambiguity problem by augmenting the trained model output. KNN-augmented Inference can be split into three parts:

(i) Firstly, we cache each entity representation r_{key} in training set and its label l_{value} to construct a pair (key,value) $\in$ DataStore.

(ii) We calculate the cosine similarity as Eq. 9 between the representation $r_{i,j}$ from span $x_{i,j}$ and each cached representation from DataStore.

(iii) Then, we select the top K most similar retrieved entities D_K and then convert them into a one-hot distribution based on the KNN majority voting mechanism.

$$\begin{aligned} y_{knn} &= \arg\max_{l} \sum_{D_K} \mathbb{I}(l_{value} = l), \forall l \in L \\ o_{knn} &= \text{onehot}(y_{knn}) \end{aligned} \tag{12}$$

(iv) Finally, we interpolate the $o_{i,j}$ from model with o_{knn} to get the final distribution o_{final}:

$$o_{final} = (1 - \mu) * o_{i,j} + \mu * o_{knn} \tag{13}$$

where μ is a hyperparameter to make a balance between two distributions.

In this way, we could use cached similar entities in the training set as memory to adjust the output of the trained model, therefore further mitigating the entity ambiguity problem.

3.3 Against Incomplete Annotation Noise

We observe that the incomplete annotation in spans labeled as non-entities leads to two problems: (1) the decision boundary shifting problem, and (2) the high asymmetric noise rate.

3.3.1 Boundary Mixup

When the model is trained on data where some entity spans are not labeled, the learned decision boundary can be biased and the model tends to predict an entity span as a non-entity type. As shown in Figure 3, the learned decision boundary tends to shift from the fully supervised boundary (our desired boundary) towards the entity side. Therefore, we propose Boundary Mixup to push the decision boundary to the unbiased (desired) side.

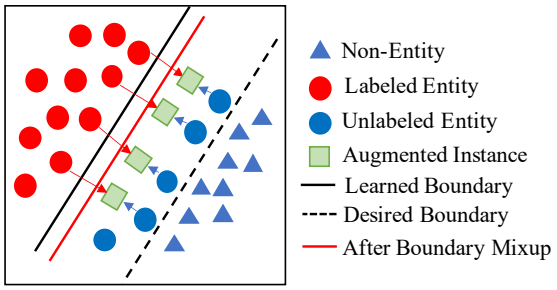

Figure 3: A toy case for decision boundary shifting problem and Boundary Mixup.

The model always makes a wrong prediction to the instances around the learned decision boundary due to the decision boundary shifting. Motivated by this, if we can find the instances around the learned decision boundary, we can further find the location of the learned decision boundary. Then we can utilize Mixip (Zhang et al., 2018) to generate the augmented instances to modify the location of the learned decision boundary. Specifically, if the span $x_{i,j}$ is predicted as non-entity l_n with a confidence lower than ε, it may actually be an instance around the learned decision boundary. Therefore, we randomly sample a entity e with the most possible entity label l_e according to $o_{i,j}$. Then use Mixup between (r_e, l_e) and $(r_{i,j}, l_{i,j})$ to generate an augmented instance $(\hat{r}, \hat{y})$:

$$(\hat{r}, \hat{y}) = (\theta' r_{i,j} + (1 - \theta') r_e, \theta' l_n + (1 - \theta') l_e) \tag{14}$$

where θ' is calculated as:

$$\theta' = \max(\theta, 1 - \theta) \tag{15}$$

$$\theta \sim \text{Beta}(\alpha', \alpha'), \alpha' \in (0, \infty) \tag{16}$$

where α' is a hyperparameter.

As shown in Figure 3, due to further training on the augmented instances, the biased decision boundary can be pushed towards the fully supervised side, mitigating the decision boundary shifting problem.

3.3.2 Noise-Tolerant Loss

To study the characteristics of incomplete annotation noise, we conduct an empirical analysis on BC5CDR and CoNLL2003. We use the knowledge bases provided by Zhou et al. (2022a) to relabel

the two training sets using string matching, then compare the matching results with original well-labeled training sets to get the rate of inaccurate and incomplete annotation for different entity types. As shown in Table 1, incomplete annotation noise accounts for the majority of the noise in DS-NER as the Incomplete Rate is much higher than Inaccurate Rate. Meanwhile, the incomplete annotation noise is asymmetric, which means the incomplete annotation noise is unbalanced in different entity types, because the knowledge bases always have different limited coverage of different entity types.

Recently, Meng et al. (2021) adopts generalized cross entropy (GCE) (Zhang and Sabuncu, 2018) instead of cross entropy (CE) to improve the robustness in DS-NER. GCE calculated on spans labeled as non-entities D_n can be described as following:

$$\mathcal{L}_{\mathrm{GCE}} = \sum_{x_{i,j} \in D_n} \frac{1 - o_{i,j}^{q}}{q} \tag{17}$$

where q denotes a hyperparameter; when $q \to 1$ GCE approximates Mean Absolute Error Loss (MAE), which is widely used in regression task; when $q \to 0$, GCE approximates CE (using L'Hôpital's rule).

Ghosh et al. (2017) theoretically proved that a loss function that satisfies the symmetric condition would be inherently tolerant to symmetric label noise. However, the derived loss functions according to this design principle such as MAE, suffer from the underfitting (Charoenphakdee et al., 2019). GCE attempts to perform the trade-off between CE and symmetric loss MAE, but is performs poorly on asymmetric noise as it attempts to satisfy symmetric condition (Ghosh et al., 2017). As the incomplete annotation noise is always asymmetric, we adopt Sparse Regularization (SR) to improve the GCE performance, which is proved to be more robust to asymmetric noise (Zhou et al., 2021). GCE + SR only calculated on D_n can be defined as:

$$\mathcal{L}_{\mathrm{GCE+SR}} = \sum_{x_{i,j} \in D_n} \frac{1 - o_{i,j}^{q}}{q} + \|(\mathbf{o}_{i,j})\|_p^p \tag{18}$$

where $p \leq 1$ denotes a hyperparameter.

In this way, we can achieve better robustness to asymmetric noise caused by incomplete annotation.

Dataset	Type	Inaccurate Rate	Incomplete Rate
BC5CDR	Chemical	2.01	36.86
	Disease	1.64	53.27
CoNLL2003	PER	17.64	17.89
	LOC	0.02	34.80
	ORG	9.53	39.41
	MISC	0.00	79.93

Table 1: The token-level quality of distant labels on training sets in Zhou et al. (2022b) settings.

Dataset	Types	Train	Test
CoNLL2003	4	14041	3453
OntoNotes5.0	18	115812	12217
Webpage	4	385	135
BC5CDR	2	4560	4797
EC	5	3657	798

Table 2: Statistics of five DS-NER datasets.

3.4 Training

We weighted the losses as follows:

$$\mathcal{L}_{\mathrm{Final}} = \eta \mathcal{L}_{\mathrm{MFL}} + (1 - \eta)\mathcal{L}_{\mathrm{CL}} + \mathcal{L}_{\mathrm{GCE+SR}} \tag{19}$$

η is hyperparameter to control the weight between $\mathcal{L}_{\mathrm{MFL}}$ and $\mathcal{L}_{\mathrm{CL}}$ only calculated on spans labeled as entities. As introduced before, $\mathcal{L}_{\mathrm{GCE+SR}}$ is only calculated on spans labeled as non-entities.

4 Experiment

Compared with extensive baselines, SANTA achieves significant improved performance in five datasets. We also conduct experiments and provide analyses to justify the effectiveness of SANTA.

4.1 Dataset

We conduct experiments on five benchmark datasets, including CoNLL2003 (Tjong Kim Sang and De Meulder, 2003), Webpage (Ratinov and Roth, 2009), OntoNotes5.0 (Weischedel et al., 2013), BC5CDR (Li et al., 2015) and EC (Yang et al., 2018). For CoNLL2003, OntoNotes5.0, Webpage, Liang et al. (2020) re-annotates the training set by distant supervision, and uses the original development and test set. We keep the same knowledge bases as Shang et al. (2018) in BC5CDR. For EC, Yang et al. (2018) uses the distant supervision to get training data, and labels development and test set by crowd-sourcing. Statistics of five datasets are shown in Table 2.

Method	CoNLL2003			Webpage			OntoNotes5.0			BC5CDR			EC		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
KB-Matching	81.13	63.75	71.40	62.59	45.14	52.45	63.86	55.71	59.51	85.90	48.20	61.70	-	-	44.02
BiLSTM-CRF	75.50	49.10	59.50	58.05	34.59	43.34	68.44	64.50	66.41	83.60	52.40	64.40	-	-	54.59
RoBERTa	82.29	70.47	75.93	59.24	62.84	60.98	66.99	69.51	68.23	79.30	66.50	72.30	-	-	-
LRNT†	79.91	61.87	69.74	46.70	48.83	47.74	67.36	68.02	67.69	-	-	-	-	-	-
Co-teaching+*	86.04	68.74	76.42	61.65	55.41	58.36	66.63	69.32	67.95	-	-	-	-	-	-
JoCoR*	83.65	69.69	76.04	62.14	58.78	60.42	66.74	68.74	67.73	-	-	-	-	-	-
NegSampling†	80.17	77.72	78.93	70.16	58.78	63.97	64.59	**72.39**	68.26	-	-	-	-	-	66.17
NegSampling+†	-	-	-	-	-	-	-	-	-	-	-	-	-	-	67.03
SCDL*	**87.96**	79.82	83.69	68.71	**68.24**	68.47	67.49	69.77	68.61	-	-	-	-	-	-
AutoNER†	75.21	60.40	67.00	48.82	54.23	51.39	64.63	69.95	67.18	79.80	58.60	67.50	-	-	-
BOND*	82.05	80.92	81.48	67.37	64.19	65.74	67.14	69.61	68.35	78.80	66.60	72.10	-	-	-
RoSTER*	85.90	84.90	85.40	-	-	-	-	-	-	73.30	72.60	72.90	-	-	-
Conf-MPU†	78.58	79.75	79.16	-	-	-	-	-	-	69.79	**86.42**	77.22	-	-	-
CReDEL*	-	-	-	-	-	-	-	-	-	65.20	80.60	72.10	-	-	-
SCL-RAI†	-	-	-	-	-	-	-	-	-	-	-	-	-	-	71.24
Ours	86.25	**86.95**	**86.59**	**78.40**	66.22	**71.79**	**69.24**	70.21	**69.72**	**81.74**	76.88	**79.23**	**73.13**	**73.53**	**73.33**

Table 3: Results on BC5CDR, CoNLL2003, OntoNotes5.0, Webpage. We report the baseline results from Zhang (2022), Zhang et al. (2021a), Meng et al. (2021), and Si et al. (2022). †: methods that only consider the incomplete annotation. ∗: methods that consider two types of noise equally.

Method	Webpage		
	P	**R**	**F1**
Focal Loss	66.67	55.77	60.74
CE	67.32	56.41	61.38
GCE + SR	67.45	57.17	61.88
CE & GCE + SR	68.46	58.32	62.98
SANTA	78.40	66.22	**71.79**
w/o MFL	-	-	-
w. GCE + SR	77.78	61.49	68.68
w. CE	73.33	66.89	69.96
w/o Memory Label Smoothing in MLF	74.02	63.51	68.36
w. Label Smoothing in MLF	74.05	65.54	69.53
w/o Entity-aware KNN	77.66	49.32	60.33
w. KNN-augmented Inference	67.94	60.14	63.80
w. Entity-aware CL	87.84	43.92	58.56
w/o GCE + SR	-	-	-
w. GCE	77.52	66.03	71.32
w. CE	76.56	66.22	71.01
w. MFL	72.31	66.43	69.25
w/o Boundary Mixup	70.71	47.30	56.68
w. Mixup	76.98	65.54	70.80

Table 4: Ablation study on Webpage.

4.2 Evaluation Metrics and Baselines

To compare with baselines, we use Precision (P), Recall (R), and F1 score as the evaluation metrics. We compare SANTA with different groups of baseline methods, including supervised methods and distantly supervised methods. We also present the results of **KB-Matching**, which directly uses knowledge bases to annotate the test sets.

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

Distantly-Supervised Methods We compare several DS-NER baselines, including: (1) methods that only consider the incomplete annotation: **Conf-MPU** (Zhou et al., 2022b) uses multi-class PU-learning loss to better estimate the loss. **AutoNER** (Shang et al., 2018) modifies the standard CRF to get better performance under the noise. **LRNT** (Cao et al., 2019) tries to reduce the negative effect of noisy labels, leaving training data unexplored fully. **NegSampling** (Li et al., 2021) and **Weighted NegSampling** (Li et al., 2022a) uses down-sampling in non-entities to relief the misleading from incomplete annotation. **SCL-RAI** (Si et al., 2022) uses span-based supervised contrastive-learning loss and designed inference method to improve the robustness against incomplete annotation. (2) methods that consider two types of noise equally: **Co-teaching+** (Yu et al., 2019) and **JoCoR** (Wei et al., 2020) are two classical methods to handle noisy labels in computer vision area. **BOND** (Liang et al., 2020) and **SCDL** (Zhang et al., 2021b) proposes teacher-student network to reduce the noise from distant labels. **RoSTER** (Meng et al., 2021) adopts GCE loss, self-training and noisy label removal step to improve the robustness. **CReDEL** (Ying et al., 2022) proposes an automatic distant label refinement model via contrastive-learning as a plug-in module.

4.3 Experimental Settings

For a fair comparison, (1) we use BERT-base (Devlin et al., 2019) as the encoder the same as Si et al. (2022) and Zhang (2022) for CoNLL2003,

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

Table 5: Case study with SANTA and baselines.

OntoNotes5.0, Webpage and EC; (2) for BC5CDR in the biomedical domain, we use BioBERT-base (Lee et al., 2020) the same as Zhang (2022). We use Adam (Kingma and Ba, 2015) as our optimizer. We list detailed hyperparameters in Table 7. Experiments are run on NVIDIA-P40 and NVIDIA-A100.

4.4 Main Results

Table 3 presents the main results of SANTA. From these results, the following four insights can be drawn. (1) on all five datasets, SANTA achieves the best F1 performance among all DS-NER baselines, and strikes a good balance between precision and recall, demonstrating superiority when trained on distantly-supervised text; (2) compared to original supervised methods, including BiLSTM-CRF and RoBERTa, SANTA improves the F1 score with an average increase of 18.48% and 7.47% respectively, which demonstrates the necessity of DS-NER models and the effectiveness; (3) compared with methods only focusing on incomplete annotation such as NegSampling, SANTA achieves more balanced precision and recall, showing the necessity to handle two types of noise. (4) compared with methods handling two types of noise with the same strategy such as RoSTER, SANTA achieves better performance both in precision and recall, demonstrating the effectiveness of separate handling.

4.5 Analysis

Ablation Study. To further validate the effectiveness of each component, we compare SANTA with the fine-grained ablations by removing one component at a time in Table 4. (1) MFL improves the precision and recall compared with using noise-tolerant GCE + SR, indicating that MFL can help model to avoid underfitting. When using CE instead of MFL, percision is significantly reduced showing MFL can better handle entity ambiguity problem. Meanwhile, designed MLS can reduce

Distant Match: [Arafat]$_{\text{PER}}$ to meet [Peres]$_{\text{PER}}$ in [Gaza]$_{\text{PER}}$ after flight ban.
Ground Truth: [Arafat]$_{\text{PER}}$ to meet [Peres]$_{\text{PER}}$ in [Gaza]$_{\text{LOC}}$ after flight ban.
Span-based NER Model: [Arafat]$_{\text{PER}}$ to meet [Peres]$_{\text{PER}}$ in [Gaza]$_{\text{PER:0.42}}$ after flight ban.
w. Entity-aware KNN: [Arafat]$_{\text{PER}}$ to meet [Peres]$_{\text{PER}}$ in [Gaza]$_{\text{LOC:0.46}}$ after flight ban.
Distant Match: [EC]$_{\text{ORG}}$ rejects German call to boycott British lamb.
Ground Truth: [EC]$_{\text{ORG}}$ rejects [German]$_{\text{MISC}}$ call to boycott [British]$_{\text{MISC}}$ lamb.
Span-based NER Model: [EC]$_{\text{ORG}}$ rejects [German]$_{\text{MISC}}$ call to boycott [British]$_{\text{MISC:0.31}}$ lamb.
w. Boundary Mixup: [EC]$_{\text{ORG}}$ rejects [German]$_{\text{MISC}}$ call to boycott [British]$_{\text{MISC:0.56}}$ lamb.

Table 6: Exploring Entity-aware KNN & Boundary Mixup. We give the selected scores from SANTA.

the fluctuation of Focal Loss to achieve better performance, and better than static Label Smoothing; (2) Entity-aware KNN consists of KNN-augmented Inference and Entity-aware CL, the significantly reduction of F1 score and only using one of them achieves poor performance, both showing the design of Entity-aware KNN is necessary; (3) compared with GCE, CE and MFL, GCE + SR shows the strong capability to handle the incomplete annotation noise, due to the improvement of robustness and asymmetric noise condition; (4) Boundary Mixup can significantly improve the recall, which indicates it can help model to make more correct prediction of the samples around the decision boundary and alleviate the decision boundary shifting problem. Boundary Mixup reduces randomness of sampling in Mixup and focus on examples around the decision boundary to better handle the incomplete annotation in DS-NER.

Case Study. We also perform case study to better understand the advantage of SANTA in Table 5. We show the prediction of Conf-MPU only focusing on

incomplete annotation noise, and SCDL handling two types of noise with same strategy. Conf-MPU is able to learn from labeled DS-NER text and slightly learn to generalize, such like “Johnson" and “John McNamara" can be recognized correctly. But the limited generalization capability leads to memorize “Columbia" and “Angels" with wrong labels. Meanwhile, Conf-MPU can not handle entity ambiguity well because it ignores the inaccurate annotation noise, such as wrongly predicting the type of span “California Angels" and “Columbia Presby Hospital". SCDL is able to generalize better and slightly handle entity ambiguity, due to the comprehensive consideration of two types of noise. But it is still impacted by entity ambiguity problem in difficult span “California Angels" . SANTA can further detect the noisy labels via separate strategies and correctly recognize all the entities in this sentence from CoNLL2003 training set.

Strategies Exploration. We further explore our strategies as following: (1) as shown in Table 4, we simply use Focal Loss, CE, GCE + SR in Span-based NER model as baselines. Meanwhile, we use CE & GCE + SR separately for the spans labeled as entities suffered from inaccurate annotation & the spans labeled as non-entities suffered from incomplete annotation. CE & GCE + SR and SANTA achieve better performance, showing that our motivation of separately handing is helpful and well-designed separate strategies is powerful. (2) as shown cases in Table 6, we can observe that only use Entity-aware KNN and only use Boundary Mixup can both help model to make more accurate predictions. Entity-aware KNN augments the score from model and finally get the right label prediction, alleviating the entity ambiguity problem. Boundary Mixup correctly help model to recall span “British", which indicates decision boundary shifting problem is relieved.

5 Conclusion

Inaccurate and incomplete annotation noise are two types of noise in DS-NER. We propose SANTA to use separate strategies for two types of noise. For inaccurate annotation, we propose Memory-smoothed Focal Loss and Entity-aware KNN to relief the ambiguity problem. For incomplete annotation, we utilize noise-robust loss GCE + SR and propose Boundary Mixup to improve the robustness and mitigate the decision boundary shifting problem. Experiments show that SANTA achieves state-of-the-art methods on five DS-NER datasets and the separate strategies are effective.

Name	BC5CDR	CoNLL2003	OntoNotes5.0	Webpage	EC
Learning Rate	1e-5	1e-5	1e-5	1e-5	1e-5
Batch Size	16	16	16	12	16
dim of W in 3	256	256	256	128	256
K in Entity-aware KNN	64	64	64	16	64
G in eq. 6	1	3	1	1	1
ϵ in Boundary Mixup	0.5	0.5	0.5	0.5	0.5
λ in eq. 6	0.8	0.8	0.8	0.8	0.8
α in eq. 8	0.5	0.5	0.5	0.5	0.5
γ in eq. 8	2	2	2	2	2
τ in eq. 11	0.05	0.05	0.05	0.05	0.05
μ in eq. 13	0.7	0.3	0.3	0.7	0.7
p in eq. 18	0.5	0.5	0.5	0.5	0.5
q in eq. 18	0.3	0.3	0.3	0.3	0.3
α' in eq. 16	0.2	0.2	0.2	0.2	0.2
η in eq. 19	0.9	0.9	0.9	0.9	0.9

Table 7: Hyperparameters

Limitations

Our proposed work is dedicated to considering the noise in DS-NER, and our noise-specific analyses are all based on this task. Therefore, if it were not for DS-NER task, our model would not necessarily be robust compared to other task-specific methods. Also, our approach is based entirely on previous experimental settings in DS-NER, so we do not consider how to reduce noise from the distant supervision process, e.g., designing models to help the annotation process rather than learning to reduce noise from the distantly-supervised text. Designing models to help the distant supervision process could be a direction for future study.

Acknowledgements

This paper is supported by the National Key R&D Program of China under Grand No.2018AAA0102003, the National Science Foundation of China under Grant No.61936012. Our code can be found in https://github.com/PKUnlp-icler/SANTA.

References

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

Nontawat Charoenphakdee, Jongyeong Lee, and Masashi Sugiyama. 2019. On symmetric losses for learning from corrupted labels. 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 961–970. PMLR.

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.

Aritra Ghosh, Himanshu Kumar, and P. S. Sastry. 2017. Robust loss functions under label noise for deep neural networks. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 1919–1925. AAAI Press.

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*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform.*, 36(4):1234–1240.

Jiao Li, Yueping Sun, R Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2015. Annotating chemicals, diseases, and their interactions in biomedical literature. In *Proceedings of the fifth BioCreative challenge evaluation workshop*, pages 173–182. The Fifth BioCreative Organizing Committee.

Yangming Li, Lemao Liu, and Shuming Shi. 2020. Empirical analysis of unlabeled entity problem in named entity recognition. *arXiv preprint arXiv:2012.05426*.

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

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

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

Yunshui Li, Junhao Liu, Min Yang, and Chengming Li. 2022c. Self-distillation with meta learning for knowledge graph completion. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2048–2054, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2020. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(2):318–327.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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

Yu Meng, Yunyi Zhang, Jiaxin Huang, Xuan Wang, Yu Zhang, Heng Ji, and Jiawei Han. 2021. Distantly-supervised named entity recognition with noise-robust learning and language model augmented self-training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10367–10378, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Miao Peng, Ben Liu, Qianqian Xie, Wenjie Xu, Hua Wang, and Min Peng. 2022. SMiLE: Schema-augmented multi-level contrastive learning for knowledge graph link prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4165–4177, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

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

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

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

Peiyi Wang, Yifan Song, Tianyu Liu, Binghuai Lin, Yunbo Cao, Sujian Li, and Zhifang Sui. 2022. Learning robust representations for continual relation extraction via adversarial class augmentation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6264–6278, Abu Dhabi, United Arab Emirates. 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. Computer Vision Foundation / IEEE.

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

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

Huaiyuan Ying, Shengxuan Luo, Tiantian Dang, and Sheng Yu. 2022. Label refinement via contrastive learning for distantly-supervised named entity recognition. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2656–2666, Seattle, United States. 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.

Shuang Zeng, Yuting Wu, and Baobao Chang. 2021. SIRE: Separate intra- and inter-sentential reasoning for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 524–534, Online. Association for Computational Linguistics.

Bryan Zhang. 2022. Improve MT for search with selected translation memory using search signals. In *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 2: Users and Providers Track and Government Track)*, pages 123–131, Orlando, USA. Association for Machine Translation in the Americas.

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

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

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

Zhilu Zhang and Mert R. Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with 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 8792–8802.

Hao Zhou, Gongshen Liu, and Kewei Tu. 2022a. Improving constituent representation with hypertree neural networks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1682–1692, Seattle, United States. Association for Computational Linguistics.

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

Xiong Zhou, Xianming Liu, Chenyang Wang, Deming Zhai, Junjun Jiang, and Xiangyang Ji. 2021. Learning with noisy labels via sparse regularization. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 72–81. IEEE.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
We discuss the limitations in Sec. Limitations

☒ A2. Did you discuss any potential risks of your work?
Our work focuses on Distantly-Supervised Named Entity Recognition and public datasets, which does not pose a risk.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
We discuss abstract and introduction in Sec. Abstract and Sec. Introduction

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

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
No response.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

Left blank.

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

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We discuss them in Sec. Experiment and Appendix

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

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Not applicable. we do not use them.

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

Left blank.

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

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

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

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

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