CoLaDa: A Collaborative Label Denoising Framework for Cross-lingual Named Entity Recognition

Tingting Ma[1*], **Qianhui Wu**[2], **Huiqiang Jiang**[2],
Börje F. Karlsson[2], **Tiejun Zhao**[1†], **Chin-Yew Lin**[2]
[1]Harbin Institute of Technology, Harbin, China [2]Microsoft
hittingtingma@gmail.com
{qianhuiwu, hjiang, borjekar, cyl}@microsoft.com
tjzhao@hit.edu.cn

Abstract

Cross-lingual named entity recognition (NER) aims to train an NER system that generalizes well to a target language by leveraging labeled data in a given source language. Previous work alleviates the data scarcity problem by translating source-language labeled data or performing knowledge distillation on target-language unlabeled data. However, these methods may suffer from label noise due to the automatic labeling process. In this paper, we propose **CoLaDa**, a **Co**llaborative **La**bel **D**enoising Fr**a**mework, to address this problem. Specifically, we first explore a *model-collaboration*-based denoising scheme that enables models trained on different data sources to collaboratively denoise pseudo labels used by each other. We then present an *instance-collaboration*-based strategy that considers the label consistency of each token's neighborhood in the representation space for denoising. Experiments on different benchmark datasets show that the proposed CoLaDa achieves superior results compared to previous methods, especially when generalizing to distant languages.[1]

1 Introduction

The named entity recognition (NER) task aims to locate and classify entity spans in a given text into predefined entity types. It is widely used for many downstream applications, such as relation extraction and question answering. Deep neural networks have made significant progress on this task leveraging large-scale human-annotated data for training. However, fine-grained token-level annotation makes it costly to collect enough high-quality labeled data, especially for low-resource languages. Such scenarios motivate the research on *zero-shot* cross-lingual NER, which attempts to leverage labeled data in a rich-resource source language to solve the NER task in a target language without annotated data.

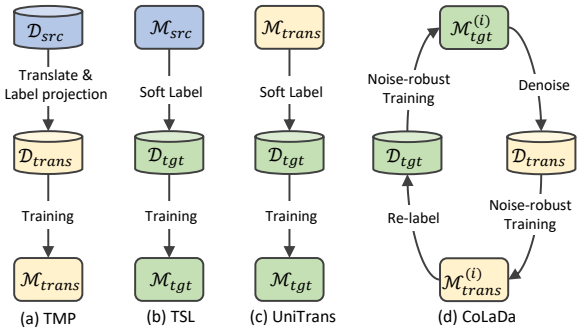

Figure 1: Comparison between previous methods (**a/b/c**) and our CoLaDa at the i-th iteration (**d**) CoLaDa starts at $\mathcal{M}_{tgt}^{0}$ and performs denoising iteratively. $\mathcal{D}_{src}$: Source-language labeled data. $\mathcal{D}_{trans}$: Translation data. $\mathcal{D}_{tgt}$: Target-language unlabeled data with pseudo-labels generated by NER models. $\mathcal{M}_{src/trans/tgt}$: NER model learned on $\mathcal{D}_{src/trans/tgt}$.

Recent attempts at cross-lingual NER can be roughly categorized from two aspects: learning language-independent features via feature alignment (Huang et al., 2019; Keung et al., 2019) and learning language-specific features from automatically labeled target-language data (Wu et al., 2020c,b). Despite bringing great success to cross-lingual NER, the former line of research misses exploiting language-specific features and thus shows substandard performance, especially when transferring to distant languages, *e.g.*, from English to Arabic (Fu et al., 2023). Hence, a series of studies focuses on the latter category, which typically creates pseudo-labeled target-language data and uses it to perform conventional supervised learning or teacher-student learning. For example, as shown in Fig 1(a), earlier studies (Ehrmann et al., 2011; Mayhew et al., 2017; Xie et al., 2018; Jain et al., 2019), such as TMP (Jain et al., 2019), first translate labeled data in the source language and then perform label projection. Recently, several approaches have

[*]Work during internship at Microsoft.
[†]Corresponding author.
[1]Our code is available at https://github.com/microsoft/vert-papers/tree/master/papers/CoLaDa.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 5995–6009
July 9-14, 2023

utilized a weak model, which could be an NER model either trained on the source language's labeled data as in TSL (Wu et al., 2020c), or further finetuned on the generated translation data as in UniTrans (Wu et al., 2020b), to annotate the unlabeled target-language data for improvement, as shown in Fig 1(b) and Fig 1(c).

Unfortunately, these methods inevitably suffer from the label noise induced by inaccurate translation and label projection, or the weak model's limited capability. Although some methods are proposed to mitigate the label noise problem by additionally training an instance selector (Liang et al., 2021; Chen et al., 2021) or designing heuristic rules for data selection (Ni et al., 2017), they independently manipulate either the translation data ($\mathcal{D}_{trans}$) (Ni et al., 2017) or the target-language data ($\mathcal{D}_{tgt}$) pseudo-labeled by NER models trained in the source language (Liang et al., 2021; Chen et al., 2021). Hence, all these methods ignore the complementary characteristics between both for denoising. Particularly, from the *text view*, $\mathcal{D}_{tgt}$ is collected from a natural text distribution of the target-language data, while $\mathcal{D}_{trans}$ can be regarded as a way of data augmentation to provide more lexicon variants. From the *labeling function view*, labels of $\mathcal{D}_{trans}$ are obtained via the label projection algorithm, which have little association with those of $\mathcal{D}_{tgt}$ generated by NER models.

With such consideration, we propose **a *model-collaboration*-based denoising scheme**, which incorporates models trained on both data sources to mutually denoise the pseudo-labels of both data sources in an iterative way. As shown in Fig 1(d), we first leverage $\mathcal{M}_{tgt}$ trained on the pseudo-labeled target-language data $\mathcal{D}_{tgt}$ to denoise the translation data annotated by label projection. In this way, the learned model $\mathcal{M}_{trans}$ will be less affected by noise in the translation data. We then employ the improved $\mathcal{M}_{trans}$ to re-label the target-language unlabeled data $\mathcal{D}_{tgt}$. It is expected that there is less noise in the relabeled data, and thus we can produce a more powerful $\mathcal{M}_{tgt}$. We perform this procedure for several iterations, so that all the involved data sources and models can be improved in an upward spiral.

Moreover, borrowing the idea from anomaly detection (Gu et al., 2019) that a given data point's neighborhood information can be used to measure its anomalism, here we find that the similar tokens in the feature space can also collaborate for denoising. Previous studies (Zhai and Wu, 2019; Xu et al., 2020) have shown that instances with the same label are more likely to locate close to each other in the representation space. Our intuition is that, if a token's label conflicts a lot with labels of other tokens in its neighborhood, then this label is probably noisy. Therefore, we further propose **an *instance-collaboration*-based denoising strategy** to explore the neighborhood structure of each token for denoising, as shown in Figure 2. Specifically, we utilize the label consistency of each token's neighborhood in the representation space to re-weight the soft-labeled examples in knowledge distillation.

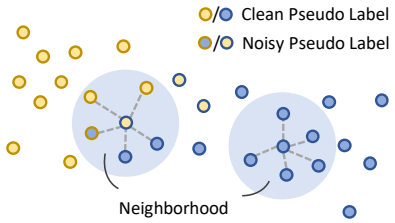

Figure 2: Illustration of the instance collaboration for denoising. Different colors depict different entity types.

We integrate the instance-collaboration-based denoising strategy into the model-collaboration-based denoising scheme and propose a **Co**llaborative **La**bel **D**enoising fra**m**ework, *i.e.*, **CoLaDa**, for cross-lingual NER. We conduct extensive experiments on two popular benchmarks covering six languages for evaluation. Experimental results show that our method outperforms existing state-of-the-art methods. Qualitative and quantitative analyses further demonstrate the effectiveness of our framework in reducing the data noise.

2 Problem Formulation

Here we take the typical sequence labeling formulation for the named entity recognition task. Given a sequence with L tokens $\boldsymbol{x} = (x_1, \ldots, x_L)$ as the input text, an NER system is excepted to assign each token x_i with a label y_i.

In this paper, we assume to have the labeled training data $\mathcal{D}_{src} = \{(\boldsymbol{x}^s, \boldsymbol{y}^s)\}$ in the source language, the unlabeled data $\mathcal{D}_{tgt} = \{\boldsymbol{x}^u\}$ from the target language, and translation data $\mathcal{D}_{trans} = \{(\boldsymbol{x}^t, \boldsymbol{y}^t)\}$ obtained by data projection from $\mathcal{D}_{src}$. Our goal is to train an NER model $\mathcal{M}$ that can generalize well to the target language utilizing these resources.

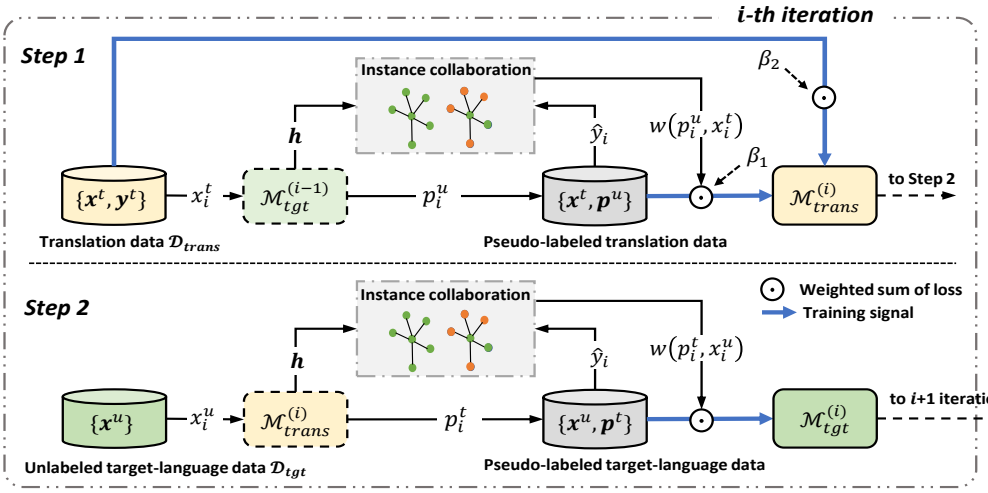

Figure 3: Framework of *CoLaDa*, which is an **iterative model-collaboration** process with two steps: 1) Step 1: noise-robust training on translation data with the collaborator $\mathcal{M}_{tgt}^{(i-1)}$, 2) Step 2: noise-robust training on unlabeled target-language data with the collaborator $\mathcal{M}_{trans}^{(i)}$. The **instance-collaboration** is used to re-weight the noisy labels from a teacher model in both steps. $\mathcal{M}_{trans}^{(i)}/\mathcal{M}_{tgt}^{(i)}$: model trained on $\mathcal{D}_{trans}/\mathcal{D}_{tgt}$ at i-th iteration.

3 CoLaDa Framework

Figure 3 depicts an overview of the CoLaDa framework. It is an iterative model-collaboration-based denoising framework which consists of two steps: noise-robust learning on translation data and noise-robust learning on unlabeled target-language data. An instance-collaboration-based denoising strategy (Sec 3.1) is then integrated into the model-collaboration-based denoising procedure (Sec 3.2).

3.1 Instance Collaboration for Denoising

Previous work (Zhai and Wu, 2019; Xu et al., 2020) indicates that tokens with the same labels are more likely to locate close to each other in the representation space of a deep neural network. If the label of a given token is inconsistent with lots of its neighbors, this token would be isolated from other tokens with the same label in the feature space, and hence its label is more likely to be noisy. Therefore, we propose instance-collaboration-based denoising, which evaluates the reliability of a given token's label by measuring the label consistency of its neighborhood, and then uses the reliability score to weight the noisy labels from a teacher model $\mathcal{M}$ for knowledge distillation on data $\mathcal{D} = \{\boldsymbol{x}\}$. Noisy labels are expected to have lower weights than clean ones.

Create a memory bank. We leverage the feature extractor $\mathcal{F}$ of the NER model $\mathcal{M}$ to obtain the hidden representations $\boldsymbol{h} = \{h_i\}_{i=1}^{L}$ of each sentence $\boldsymbol{x} = \{x_i\}_{i=1}^{L} \in \mathcal{D}$:

$$\boldsymbol{h} = \mathcal{F}(\boldsymbol{x}). \tag{1}$$

We then construct a memory bank $\mathcal{B}_{\mathcal{D}} = \{h\}$ to store the hidden representations of all tokens in $\mathcal{D}$.

Compute label consistency. Given a token x_i, we retrieve its K-nearest neighbors $\mathcal{N}_K(x_i)$ in $\mathcal{B}_{\mathcal{D}}$ using cosine similarity. Let p_i denote the soft label (*i.e.*, the probability distribution over the entity label set) assigned by the teacher model $\mathcal{M}$ for x_i. We measure the label consistency of x_i, *i.e.*, $\lambda(p_i; x_i)$, by calculating the fraction of x_i's neighbors that are assigned with the same labels as x_i in $\mathcal{N}_K(x_i)$:

$$\lambda(p_i; x_i) = \frac{1}{K} \sum_{x_j \in \mathcal{N}_k(x_i)} I(\hat{y}_j = \hat{y}_i), \tag{2}$$

where $\hat{y}_i = \arg\max(p_i)$ is the pseudo entity label corresponding to the maximum probability in p_i. Similarly, $\hat{y}_j$ is the pseudo entity label corresponding to x_j. I is the indicator function.

Produce a reliability score. We use the label consistency $\lambda(p_i; x_i)$ to compute the reliability score of the soft label p_i, which is further used as the weight of p_i during model learning (see 3.2). Considering that different entity types may contain different levels of label noise and show different statistics on label consistency, here we present a class-adaptive reliability score for weighting:

$$w(p_i; x_i) = \text{Sigmoid}\left(\alpha\left(\lambda(p_i; x_i) - \mu(\hat{y}_i)\right)\right), \tag{3}$$

where $\mu(\hat{y}_i)$ denote the mean of all $\lambda(p_j; x_j)$ where $\arg\max(p_j) = \hat{y}_i$ and $x_j \in \mathcal{D}$. $\alpha > 0$ is a hyper-parameter that controls the sharpness of the weighting strategy. If $\alpha \to 0$, all tokens have equal weights. If $\alpha \to \infty$, tokens whose label consistency is larger than the average label consistency *w.r.t.* its pseudo label will be weighted with 1 and those with smaller consistency will be dropped.

3.2 Model Collaboration for Denoising

Here we elaborate on the details of the two noise-robust training processes. Algorithm 1 depicts the overall training procedure of CoLaDa.

Noise-robust training on translation data. Assuming the availability of a collaborator $\mathcal{M}_{tgt}$[2] trained on pseudo-labeled target-language data $\mathcal{D}_{tgt}$, here we focus on leveraging $\mathcal{M}_{tgt}$ to reduce the label noise in the translation data $\mathcal{D}_{trans} = \{(\boldsymbol{x}^t, \boldsymbol{y}^t)\}$, with which we further deliver a more powerful model $\mathcal{M}_{trans}$.

Specifically, given a sentence $(\boldsymbol{x}^t, \boldsymbol{y}^t) \in \mathcal{D}_{trans}$, we first obtain the soft label p_i^u of each $x_i^t \in \boldsymbol{x}^t$ from the collaborator $\mathcal{M}_{tgt}$. Then, we take both the one hot label y_i^t and the soft label p_i^u as the supervision to train the model $\mathcal{M}_{trans}$.[3] Denote the output probability distribution of $\mathcal{M}_{trans}$ for x_i^t as $\hat{p}_i^t$. The loss function *w.r.t.* $\boldsymbol{x}^t$ is defined as:

$$\mathcal{L}^{\boldsymbol{x}^t} = \frac{1}{L}\sum_{i=1}^{L}\left(\beta_1 \text{CE}(\hat{p}_i^t, p_i^u) + \beta_2 \text{CE}(\hat{p}_i^t, y_i^t)\right), \tag{4}$$

where $\text{CE}(\cdot,\cdot)$ denotes the cross-entropy loss, L is the sentence length, β_1 and β_2 are weighting scalars. Here we further incorporate the instance-collaboration-based denoising strategy (3.1) to provide a token-level reliability evaluation to the supervision from the collaborator $\mathcal{M}_{tgt}$ via:

$$\beta_1(x_i^t) \leftarrow \beta_1 * w(p_i^u, x_i^t), \tag{5}$$

where $w(p_i^u, x_i^t)$ is calculated by Eq. (3).

Noise-robust training on target-language unlabeled data. Here we leverage $\mathcal{M}_{trans}$ obtained via the above noise-robust training on translation data to provide high-quality supervision for $\mathcal{D}_{tgt} = \{\boldsymbol{x}^u\}$. By performing knowledge distillation on $\mathcal{D}_{tgt}$, the student model $\mathcal{M}_{tgt}$ is supposed to benefit from the unlabeled data drawn from the real text distribution in the target language with the knowledge from the teacher model $\mathcal{M}_{trans}$.

Specifically, given a sentence $\boldsymbol{x}^u \in \mathcal{D}_{tgt}$, we first utilize $\mathcal{M}_{trans}$ to predict soft label p_i^t for each token $x_i^u \in \boldsymbol{x}^u$. Then, we integrate the instance-collaboration-based denoising technique into the learning process. The loss function *w.r.t.* $\boldsymbol{x}^u$ to train the student model $\mathcal{M}_{tgt}$ can be formulated as:

$$\mathcal{L}^{\boldsymbol{x}^u} = \frac{1}{L}\sum_{i=1}^{L} w(p_i^t, x_i^u) \cdot \text{CE}(\hat{p}_i^u, p_i^t), \tag{6}$$

where $\hat{p}_i^u$ denotes the output probability distribution of $\mathcal{M}_{tgt}$ for the i-th token x_i^u and $w(p_i^t, x_i^u)$ is calculated by Eq. (3).

Algorithm 1 Pseudo code of CoLaDa.

Input: an NER model $\mathcal{M}_{src}$ trained on $\mathcal{D}_{src}$, translation data $\mathcal{D}_{trans}$, the unlabeled data $\mathcal{D}_{tgt}$, the maximum iteration T.

$\mathcal{M}_{tgt}^{(0)} \leftarrow \mathcal{M}_{src}$ ▷ Initialization
for $i = 1, 2, \ldots,$ T **do**
  # Step 1: Noise-robust training on D_{trans}
  Inference $\mathcal{M}_{tgt}^{(i-1)}$ on $\mathcal{D}_{trans} = \{(\boldsymbol{x}^t, \boldsymbol{y}^t)\}$ to get the predictions $\widehat{\mathcal{D}}_{trans} = \{(\boldsymbol{x}^t, \boldsymbol{p}^u)\}$
  Get $\boldsymbol{w}$ for $(\boldsymbol{x}^t, \boldsymbol{p}^u) \in \widehat{\mathcal{D}}_{trans}$ with $\mathcal{M}_{tgt}^{(i-1)}$, Eq.(3)
  Train $\mathcal{M}_{trans}^{(i)}$ with loss on $(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{p}^u, \boldsymbol{w})$, Eq.(4)
  # Step 2: Noise-robust training on D_{tgt}
  Inference $\mathcal{M}_{trans}^{(i)}$ on $\mathcal{D}_{tgt} = \{\boldsymbol{x}^u\}$ to get the predictions $\widehat{\mathcal{D}}_{tgt} = \{(\boldsymbol{x}^u, \boldsymbol{p}^t)\}$
  Get $\boldsymbol{w}'$ for $(\boldsymbol{x}^u, \boldsymbol{p}^t) \in \widehat{D}_{tgt}$ with $\mathcal{M}_{trans}^{(i)}$, Eq.(3)
  Train $\mathcal{M}_{tgt}^{(i)}$ with loss on $(\boldsymbol{x}^u, \boldsymbol{p}^t, \boldsymbol{w}')$, Eq.(6)
end for

Output: an NER model $\mathcal{M}_{tgt}^{(T)}$.

4 Experiments

4.1 Experiment Settings

Datasets We conduct experiments on two standard cross-lingual NER benchmarks: CoNLL (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003) and WikiAnn (Pan et al., 2017). CoNLL contains four languages: English (en) and German (de) from the CoNLL-2003[4] NER shared task (Tjong Kim Sang and De Meulder, 2003), and Spanish (es) and Dutch (nl) from the CoNLL-2002[5] NER shared task (Tjong Kim Sang, 2002). This dataset is annotated with four entity types: PER, LOC, ORG, and MISC. WikiAnn

[2]For the first iteration, we use an NER model trained on the source language labeled data $\mathcal{D}_{src}$. For the later iterations ($i > 1$), we use the model from the noise-robust-training on target-language unlabeled data in the previous iteration ($i-1$).

[3]The student model $\mathcal{M}_{trans}$ is initialized from $\mathcal{M}_{tgt}$ to equip the knowledge of real target-language text distribution for better generalization during test.

[4]https://www.clips.uantwerpen.be/conll2003/ner/

[5]https://www.clips.uantwerpen.be/conll2002/ner/

contains an English dataset and datasets in three non-western languages: Arabic (ar), Hindi (hi), and Chinese (zh). Each dataset is annotated with 3 entity types: PER, LOC, and ORG. All datasets are annotated with the BIO tagging scheme. We use the train, development, and test splits as previous work (Wu and Dredze, 2019; Wu et al., 2020b).

We take English as the source language and other languages as the target language, respectively. We remove the labels of the training data for the target language and take it as the unlabeled target language data. For the CoNLL benchmark, we use the word-to-word translation data provided in UniTrans (Wu et al., 2020b) for a fair comparison. For the WikiAnn benchmark, we translate the source data to the target language with the public M2M100 (Fan et al., 2020) translation system and conduct label projection with the marker-based alignment algorithm as Yang et al. (2022).

Evaluation The entity-level micro-F1 on test set of the target language is used as the evaluation metric. We report the mean value of 5 runs with different seeds for all the experiments.

Implementation Details For the base NER model, we stack a linear classifier with softmax over a base encoder such as mBERT. We implement our framework with Pytorch 1.7.1[6], the *HuggingFace* transformer library (Wolf et al., 2020), and use FAISS (Johnson et al., 2019) for embedding retrieval. Following Wu and Dredze (2019) and Zhou et al. (2022), we use the multilingual BERT base model (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) large model as our base encoders. Most of our hyper-parameters are set following Wu et al. (2020b). We use AdamW (Loshchilov and Hutter, 2019) as optimizer and train the model on source NER data with the learning rate of 5e-5 for 3 epochs. The dropout rate is 0.1. For teacher-student learning, we train the model with a learning rate of 2e-5 for 10 epochs. We freeze the bottom three layers as Wu and Dredze (2019). Following Keung et al. (2019), we choose other hyper-parameters according to the target language dev set. We set K in Eq. (2) to 500 and α in Eq. (3) to 6. For the first iteration, we start with an NER model trained on the source-language data to denoise the translation data with β_1 and β_2 in Eq. (5) setting to 0.5. For the following iterations, β_1 is set to 0.9 and β_2 is set to 0.1. The maximum number of iterations is 8.

[6] https://pytorch.org/

Method	de	es	nl	avg
mBERT based methods:				
mBERT (Wu and Dredze, 2019)	69.56	74.96	77.57	73.57
AdvCE (Keung et al., 2019)	71.90	74.3	77.60	74.60
TSL (Wu et al., 2020c)	73.16	76.75	80.44	76.78
UniTrans (Wu et al., 2020b)	74.82	79.31	82.90	79.01
TOF (Zhang et al., 2021)	76.57	80.35	82.79	79.90
AdvPicker (Chen et al., 2021)	75.01	79.00	82.90	78.97
RIKD (Liang et al., 2021)	75.48	77.84	82.46	78.59
MTMT (Li et al., 2022)	76.80	**81.82**	83.41	80.67
CoLaDa (ours)	**77.30**	80.43	**85.09**	**80.94**
XLM-R based methods:				
MulDA (Liu et al., 2021)	74.55	78.14	80.22	77.64
xTune (Zheng et al., 2021)	74.78	80.03	81.76	78.85
ConNER (Zhou et al., 2022)	77.14	80.50	83.23	80.29
CoLaDa (ours)	**81.12**	**82.70**	**85.15**	**82.99**

Table 1: F1 scores on CoNLL.

Method	ar	hi	zh	avg
mBERT based methods:				
BERT-align (Wu and Dredze, 2020)	42.30	67.60	52.90	54.26
TSL (Wu et al., 2020c)	43.12	69.54	48.12	53.59
RIKD (Liang et al., 2021)	45.96	70.28	50.40	55.55
MTMT (Li et al., 2022)	52.77	70.76	52.26	58.59
UniTrans† (Wu et al., 2020b)	42.90	68.76	56.08	55.91
CoLaDa (ours)	**54.26**	**72.42**	**60.77**	**62.48**
XLM-R based methods:				
XLM-R (Conneau et al., 2020)	50.84	72.17	39.23	54.08
ConNER (Zhou et al., 2022)	59.62	74.49	39.17	57.76
CoLaDa (ours)	**66.94**	**76.69**	**60.08**	**67.90**

Table 2: F1 scores on WikiAnn. † denotes results obtained by running their public code on our data.

4.2 Main Results

Baselines We compare our method to previous start-of-the-art baselines as follows: i) feature alignment based methods: mBERT (Wu and Dredze, 2019), XLM-R (Conneau et al., 2020), BERT-align (Wu and Dredze, 2020), AdvCE (Keung et al., 2019), and AdvPicker (Chen et al., 2021); ii) translation based methods: MulDA (Liu et al., 2021), UniTrans (Wu et al., 2020b), and TOF (Zhang et al., 2021)); iii) knowledge distillation based methods: TSL (Wu et al., 2020c), RIKD (Liang et al., 2021), and MTMT (Li et al., 2022); iv) consistency based methods: xTune (Zheng et al., 2021) and ConNER (Zhou et al., 2022).

Performance Comparison Tables 1 and 2 show the performance comparison of the proposed CoLaDa and prior start-of-the-art baselines on CoNLL and Wikiann, respectively. It can be seen that

Method	de	es	nl	ar	hi	zh
CoLaDa	**77.30**	**80.43**	**85.09**	**54.26**	**72.42**	**60.77**
1) CoLaDa w/o instance collaboration	76.08	79.94	83.86	50.98	71.31	59.64
2) CoLaDa w/o translation data denoise	76.17	79.22	83.10	41.41	71.10	55.04
3) CoLaDa w/o iteratively denoise	75.77	79.64	83.50	47.82	71.31	57.64
4) CoLaDa w/o model collaboration	75.64	78.99	82.98	46.51	71.09	55.25
5) CoLaDa w/o instance & model collaboration	74.54	79.94	82.97	42.33	70.39	55.55

Table 3: Ablation study on CoNLL and WikiAnn.

CoLaDa outperforms prior methods with both encoders, achieving a significant improvement of 2.70 F1 scores on average for CoNLL and 10.14 F1 scores on average for WikiAnn with XLM-R as the encoder. This well demonstrates the effectiveness of our approach. Interestingly, CoLaDa shows more significant superiority when transferring to distant target languages in WikiAnn. The knowledge distillation based baselines (*i.e.*, TSL, RIKD, MTMT) struggle on distant languages such as Chinese (zh) due to the noisy predictions from the weak teacher model $\mathcal{M}_{src}$ trained in the source language. UniTrans, which is developed with the same data sources as ours, shows poor performance, especially in distant languages such as Arabic (ar). We conjecture that the problem of label noise is even more critical in these distant languages. Our CoLaDa can better handle noise in both translation data and unlabeled target-language data, thus leading to significant performance gains.

5 Analysis

5.1 Ablation Study

To further validate the effectiveness of each mechanism in the proposed framework, we introduce the following variants of CoLaDa in an ablation study: 1) *CoLaDa w/o instance collaboration*, where we directly set the reliability score in Eq. (3) to 1 for all tokens. 2) *CoLaDa w/o translation data denoise*, where we set β_1 in Eq. (4) to 0. 3) *CoLaDa w/o iteratively denoise*, where we remove the iterative enhancement and only conduct the denoising process for one iteration. 4) *CoLaDa w/o model collaboration*, where we set β_1 in Eq. (4) to 0, remove the iteration mechanism, and directly take the model finetuned on $\mathcal{D}_{trans}$ as the teacher model to train a student model with instance-collaboration-based denoising on $\mathcal{D}_{tgt}$. 5) *CoLaDa w/o instance & model collaboration*, which further drops the instance-collaboration-based denoising from 4).

Table 3 shows the ablation results. We can draw some in-depth conclusions as follows.

1) *CoLaDa* outperforms *CoLaDa w/o instance collaboration*, which highlights the effectiveness of leveraging neighborhood information to reduce label noise in knowledge distillation.

2) *CoLaDa* outperforms *CoLaDa w/o translation data denoise*, which emphasizes the importance of using the collaborator $\mathcal{M}_{tgt}$ to refine labels of translation data, especially in distant languages where the translation data is noisier (*e.g.*, 12.8 F1 drop on Arabic and 5.7 F1 drop on Chinese).

3) *CoLaDa* outperforms *CoLaDa w/o iteratively denoise*, which indicates the necessity of iterative learning: models obtained from the previous iteration should be re-used as the collaborator to further improve label quality in the next iteration.

4) *CoLaDa w/o instance & model collaboration*, which eliminates all denoising strategies from *CoLaDa*, leads to a significant performance drop, demonstrating the essentiality of label denoising for cross-lingual NER.

5.2 Analysis of Model Collaboration

Here we attempt to understand how the two models, *i.e.*, $\mathcal{M}_{trans}$ and $\mathcal{M}_{tgt}$, collaboratively improve each other.

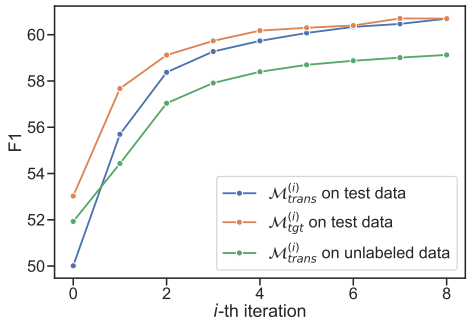

Figure 4: F1 scores of $\mathcal{M}^{(i)}_{trans}$ and $\mathcal{M}^{(i)}_{tgt}$ on the test data and the target-language unlabeled data. $\mathcal{M}^{0}_{tgt}$: model trained on source-language data. $\mathcal{M}^{0}_{trans}$: model trained on original translation data.

As shown in Figure 4, F1 scores of $\mathcal{M}_{trans}$ and $\mathcal{M}_{tgt}$ consistently improve as iterations go on, and finally converge at the last iteration. This indicates that both models benefit from the proposed model collaboration scheme. Two reasons are speculated: i) An improved $\mathcal{M}_{tgt}$ can provide more accurate labels on the translation data, which further help to improve $\mathcal{M}_{trans}$ via noise-robust learning on such translation data. For example, at the initial step ($i = 0$), the F1 score of the model $\mathcal{M}^0_{trans}$ trained on the original translation labels is 50.0. With the additional supervision from the collaborator $\mathcal{M}^0_{tgt}$, $\mathcal{M}^1_{trans}$ achieves a performance gain of 5.7 F1. ii) An improved $\mathcal{M}_{trans}$ predicts pseudo labels with higher quality on the target-language unlabeled data, which further benefits the learning of $\mathcal{M}_{tgt}$. As in Figure 4, the quality of pseudo-labeled $\mathcal{D}_{tgt}$ (the green line) grows as $\mathcal{M}_{trans}$ improves. In this way, both $\mathcal{M}_{trans}$ and $\mathcal{M}_{tgt}$ are providing more and more reliable labels for each other to learn as the iterations progress.

5.3 Analysis of Instance Collaboration

This subsection dives into the working mechanism of the instance-collaboration-based denoising.

Reliability scores v.s. label quality. To study the relationship between reliability score and label quality, we partition tokens in the target-language unlabeled data, $x_i \in \mathcal{D}_{tgt}$ into several bins according to their reliability scores $w(p^t_i, x_i)$ calculated via $\mathcal{M}^{(1)}_{trans}$. Then, we compute the token-level F1 over each bin by comparing pseudo labels $\hat{y}_i = \arg\max(p^t_i)$ to the ground-truth ones. As shown in Figure 5, the label quality is proportional to the reliability score, which well demonstrates the effectiveness of our instance-collaboration-based denoising strategy.

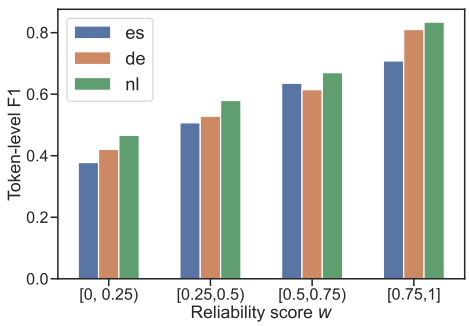

Figure 5: Illustration of the relationship between reliability score and label quality.

Analysis of Label Consistency. We also study the characteristics of label consistency *w.r.t.* different entity types and representation spaces of the memory bank. Figure 6 shows the results. We can draw some in-depth observations as follows.

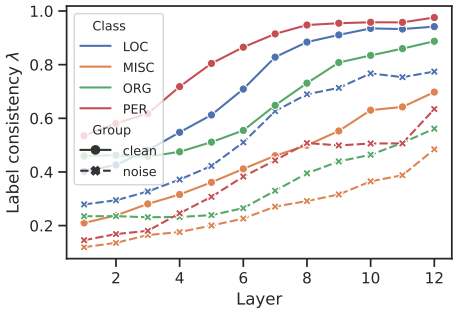

Figure 6: Mean label consistency calculated from different conditions (*i.e.*, different entity types, representation spaces of different layers, clean/noisy tokens) on the German dataset.

i) Clean tokens show a larger average consistency than noisy tokens *w.r.t.* all entity types, demonstrating the effectiveness of our label consistency based denoising strategy again.

ii) Different entity types lead to different distributions of label consistency, which validates the necessity of our design for *class-adaptive* reliability score for weighting as Eq.(3).

iii) Label consistencies calculated with token representations from the upper layers are generally larger than those corresponding to the bottom layers. Also, the label consistency gap between clean tokens and noisy tokens gets larger from the bottom to the top (*e.g.*, the gap between two orange lines). This may be attributed to the fact that representations from upper layers are more task-specific (Muller et al., 2021), hence they can better discriminate between noisy and clean tokens.

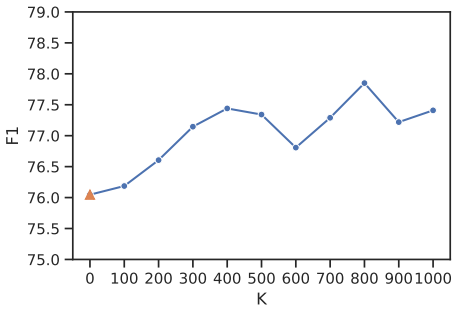

Figure 7: F1 scores of CoLaDa with different K for neighborhood information on German dataset.

Original English:	** '' Duchy of Aquitaine[LOC] '' ' – William V[PER] (995–1030)	***Inaccurate translation*** Translate "Duchy of Aquitaine" into "爱因公爵" (Duke of Ain) incorrectly.
Original Translation:	** " 爱因公爵[LOC] " – 威廉五世[PER] (995)	
Translate-train:	** " 爱因公爵[LOC] " – 威廉五世[PER] (995)	
Ours:	** " 爱因公爵[PER] " – 威廉五世[PER] (995)	
Original English:	There have been many tenants, including The Sunday Times[ORG] '' and The Daily News[ORG] ''.	***Inaccurate alignment boundary*** The symbols 《 and 》 are corner brackets used to enclose the name of a newspaper, and other works.
Original Translation:	有许多租户，包括《星期日时报[ORG]》和《每日新闻[ORG]》。	
Translate-train:	有许多租户，包括《星期日时报[ORG]》和《每日新闻[ORG]》。	
Ours:	有许多租户，包括《星期日时报[ORG]》和《每日新闻[ORG]》。	
Original English:	It is found in Democratic Republic of Congo[ORG] , Kenya[LOC] , Tanzania[LOC]	***Label noise in English data*** Wrong entity type in original English data for "Democratic Republic of China".
Original Translation:	它位于刚果民主共和国[ORG]， 肯尼亚[LOC]， 坦桑尼亚[LOC] ...	
Translate-train:	它位于刚果民主共和国[ORG]， 肯尼亚[LOC]， 坦桑尼亚[LOC] ...	
Ours:	它位于刚果民主共和国[LOC]， 肯尼亚[LOC]， 坦桑尼亚[LOC] ...	

Figure 8: Case study on translation data in Chinese. The blue (red) texts denote the correct (incorrect) entity labels. The *original translation* lines display the translation texts and labels obtained by data projection. *Translate-train* and *Ours* illustrate the predictions from the translate-train method ($\mathcal{M}^0_{trans}$) and our CoLaDa, respectively.

Choice of K for neighborhood information. Figure 7 shows the performance of CoLaDa using different K in Eq. (2). Generally speaking, CoLaDa is robust to the choice of K. Any value for $K > 0$ leads to a better performance compared with removing the instance collaboration, *i.e.*, K = 0. A smaller K may lead to a slight performance drop due to limited neighborhood information.

5.4 Case Study

To better illustrate the kinds of label noise presented in the data and the capability of CoLaDa to address such noise, we conduct a case study on the Chinese translation data from the WikiAnn English data. As shown in Figure 8, there are three typical cases of noisy labels in the translation data: noisy labels induced by inaccurate translations, alignment errors, and annotation errors in the original source-language data.[7] Figure 8 shows that the translate-train model, finetuned on the original translation data, overfits the noisy labels. However, CoLaDa is less affected by such noise and makes correct predictions.

6 Related Work

6.1 Cross-lingual NER

Prior work on cross-lingual NER mainly falls into two major categories: feature-based and data-based transfer.

Feature-based These methods learn language-independent features so that the model trained on the source language can directly adapt to the target language. Earlier work exploits word clusters (Täckström et al., 2012), gazetteers (Zirikly and Hagiwara, 2015), Wikifier features (Tsai et al., 2016), and cross-lingual word embedding (Ni et al., 2017), *etc*. More recently, with the fast growth of multilingual pre-trained language models (Devlin et al., 2019; Conneau et al., 2020) and their promising results on cross-lingual transfer (Wu and Dredze, 2019), lots of studies build upon such pre-trained models and further promote the learning of language-independent features via meta-learning (Wu et al., 2020d), contrastive alignment (Wu and Dredze, 2020), adversarial learning (Keung et al., 2019; Chen et al., 2021), and by integrating other resources (Fetahu et al., 2022). Despite the great success, they mostly ignore language-specific features, which are especially important when transferring to distant languages (Fu et al., 2023).

Data-based These approaches learn language-specific features via automatically labeled target-language data and can be further divided into *translation-based* and *knowledge distillation-based* methods.

Translation-based methods first translate the source-language data to the target language, then perform label projection from the source side to the target side. Some prior studies have proposed to use cheap translation such as word-to-word (Xie et al., 2018) or phrase-to-phrase (Mayhew et al., 2017) translation. Jain et al. (2019) propose an entity projection algorithm to utilize the Google translation system. Recently, Liu et al. (2021) and Yang et al. (2022) propose to translate sentences with pre-defined markers for label projection. And Ni et al. (2017) design heuristic rules to select high-

[7] Due to the short entity context information in many sentences in WikiAnn, the translation quality of entity mentions with M2M100 is less than satisfactory on the dataset.

quality translation data. However, both data noise and artifacts (Artetxe et al., 2020) in the translation data still limit the performance of such methods (García-Ferrero et al., 2022).

Knowledge distillation-based methods train a student model on unlabeled target-language data with the soft labels from a teacher model (Wu et al., 2020c). Li et al. (2022) improve the single task based teacher-student learning with entity similarity as an auxiliary task. To mitigate the label noise from the teacher model, Chen et al. (2021) propose AdvPicker, which trains a language discriminator to select the less language-dependent unlabeled data for knowledge distillation; Liang et al. (2021) design a reinforcement learning algorithm to train an instance selector according to features such as model confidence to select reliable pseudo labels iteratively.

While most previous work leverages either translation data or unlabeled data, UniTrans (Wu et al., 2020b) utilizes the model trained on translation data to perform teacher-student learning on unlabeled data. But it still suffers from the data noise problem. More recently, consistency training (Zheng et al., 2021; Zhou et al., 2022) has also been explored to leverage both unlabeled data and translation data without explicit label annotation.

To the best of our knowledge, we are the first to propose a unified denoising framework to handle data noise in both translation and unlabeled data *collaboratively* from the model and instance levels for cross-lingual NER.

6.2 Learning with Label Noise

Previous studies mainly address the label noise via re-weighting examples (Shu et al., 2019), designing noise-robust loss functions (Ma et al., 2020), and selecting clean instances (Bahri et al., 2020; Wu et al., 2020a), *etc*. However, these methods only consider the corrupted labels that naturally occur in one data source. In this work, we consider the *complementary* characteristics of translation and unlabeled data, and design a model-collaboration-based denoising scheme. While Xu et al. (2023) target at the few-shot learning scenario and leverage the neighborhood information among the *labeled examples* to hard-select the reliable pseudo labels in self-training, we focus on the *zero-shot* cross-lingual setting and softly re-weight the noisy pseudo-labels in knowledge distillation without any clean labeled data in target language.

7 Conclusion

To address the problem of label noise in cross-lingual NER, this paper presents CoLaDa, a collaborative label denoising framework. We propose a model-collaboration-based denoising scheme to make two models trained on different data sources to denoise the labels of each other and hence promote each other's learning. We further propose an instance-collaboration-based strategy that collaboratively considers the label consistency among similar tokens in the feature space to re-weight the noisy labels assigned by a teacher model in knowledge distillation. By integrating the instance-collaboration strategy into the model-collaboration denoising scheme, our final framework CoLada achieves superior performance over prior start-of-the-art methods by benefiting from better handling the data noise.

Limitations

Our framework relies on the availability of translation system and unlabeled data in the target language, which can not be applied to languages without any unlabeled text or translation text. The knowledge distillation step requires a certain amount of unlabeled text, while it may struggle in cases where only few hundreds of unlabeled sentences are available. It would be interesting to combine our label denoising framework with data augmentation techniques in such scenarios. Besides, the boarder application to other low-resource languages, such as MasakhaNER 2.0 (Adelani et al., 2022), and other cross-lingual sequence labeling tasks are left for exploration in future work.

References

David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo Lerato Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Oluwaseun Adeyemi, Gilles Quentin Hacheme,

Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu, and Dietrich Klakow. 2022. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

Dara Bahri, Heinrich Jiang, and Maya Gupta. 2020. Deep k-NN for noisy labels. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 540–550. PMLR.

Weile Chen, Huiqiang Jiang, Qianhui Wu, Börje F. Karlsson, and Yi Guan. 2021. AdvPicker: Effectively Leveraging Unlabeled Data via Adversarial Discriminator for Cross-Lingual NER. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 743–753, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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

Maud Ehrmann, Marco Turchi, and Ralf Steinberger. 2011. Building a multilingual named entity-annotated corpus using annotation projection. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 118–124, Hissar, Bulgaria. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Çelebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22:107:1–107:48.

Besnik Fetahu, Anjie Fang, Oleg Rokhlenko, and Shervin Malmasi. 2022. Dynamic gazetteer integration in multilingual models for cross-lingual and cross-domain named entity recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2777–2790, Seattle, United States. Association for Computational Linguistics.

Yingwen Fu, Nankai Lin, Boyu Chen, Ziyu Yang, and Shengyi Jiang. 2023. Cross-lingual named entity recognition for heterogenous languages. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:371–382.

Iker García-Ferrero, Rodrigo Agerri, and German Rigau. 2022. Model and data transfer for cross-lingual sequence labelling in zero-resource settings. In *Findings of the Association for Computational Linguistics: EMNLP 2022*.

Xiaoyi Gu, Leman Akoglu, and Alessandro Rinaldo. 2019. Statistical analysis of nearest neighbor methods for anomaly detection. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 10921–10931.

Lifu Huang, Heng Ji, and Jonathan May. 2019. Cross-lingual multi-level adversarial transfer to enhance low-resource name tagging. 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 3823–3833, Minneapolis, Minnesota. Association for Computational Linguistics.

Alankar Jain, Bhargavi Paranjape, and Zachary C. Lipton. 2019. Entity projection via machine translation for cross-lingual NER. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1083–1092, Hong Kong, China. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Phillip Keung, Yichao Lu, and Vikas Bhardwaj. 2019. Adversarial learning with contextual embeddings for zero-resource cross-lingual classification and NER. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1355–1360, Hong Kong, China. Association for Computational Linguistics.

Zhuoran Li, Chunming Hu, Xiaohui Guo, Junfan Chen, Wenyi Qin, and Richong Zhang. 2022. An unsupervised multiple-task and multiple-teacher model for cross-lingual named entity recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 170–179, Dublin, Ireland. Association for Computational Linguistics.

Shining Liang, Ming Gong, Jian Pei, Linjun Shou, Wanli Zuo, Xianglin Zuo, and Daxin Jiang. 2021. Reinforced iterative knowledge distillation for cross-lingual named entity recognition. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 3231–3239, New York, NY, USA. Association for Computing Machinery.

Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq Joty, Luo Si, and Chunyan Miao. 2021. MulDA: A multilingual data augmentation framework for low-resource cross-lingual NER. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5834–5846, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. 2020. Normalized loss functions for deep learning with noisy labels. In *ICML*.

Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. 2017. Cheap translation for cross-lingual named entity recognition. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2536–2545, Copenhagen, Denmark. Association for Computational Linguistics.

Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 2021. First align, then predict: Understanding the cross-lingual ability of multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2214–2231, Online. Association for Computational Linguistics.

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

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. 2019. Meta-weight-net: Learning an explicit mapping for sample weighting. In *NeurIPS*.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 477–487, Montréal, Canada. Association for Computational Linguistics.

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

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 CoNLL-2003*, pages 142–147. Edmonton, Canada.

Chen-Tse Tsai, Stephen Mayhew, and Dan Roth. 2016. Cross-lingual named entity recognition via wikification. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 219–228, Berlin, Germany. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Pengxiang Wu, Songzhu Zheng, Mayank Goswami, Dimitris Metaxas, and Chao Chen. 2020a. A topological filter for learning with label noise. In *Advances in Neural Information Processing Systems*.

Qianhui Wu, Zijia Lin, Börje F. Karlsson, Biqing Huang, and Jianguang Lou. 2020b. Unitrans : Unifying model transfer and data transfer for cross-lingual named entity recognition with unlabeled data. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3926–3932. ijcai.org.

Qianhui Wu, Zijia Lin, Börje F. Karlsson, Jian-Guang Lou, and Biqing Huang. 2020c. Single-/multi-source cross-lingual NER via teacher-student learning on unlabeled data in target language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6505–6514, Online. Association for Computational Linguistics.

Qianhui Wu, Zijia Lin, Guoxin Wang, Hui Chen, Börje F. Karlsson, Biqing Huang, and Chin-Yew Lin. 2020d. Enhanced meta-learning for cross-lingual named entity recognition with minimal resources. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9274–9281. AAAI Press.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Do explicit alignments robustly improve multilingual encoders? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4471–4482, Online. Association for Computational Linguistics.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 369–379, Brussels, Belgium. Association for Computational Linguistics.

Hong Xu, Keqing He, Yuanmeng Yan, Sihong Liu, Zijun Liu, and Weiran Xu. 2020. A deep generative distance-based classifier for out-of-domain detection with mahalanobis space. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1452–1460, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ran Xu, Yue Yu, Hejie Cui, Xuan Kan, Yanqiao Zhu, Joyce C. Ho, Chao Zhang, and Carl Yang. 2023. Neighborhood-regularized self-training for learning with few labels. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*.

Jian Yang, Shaohan Huang, Shuming Ma, Yuwei Yin, Li Dong, Dongdong Zhang, Hongcheng Guo, Zhoujun Li, and Furu Wei. 2022. Crop: Zero-shot cross-lingual named entity recognition with multilingual labeled sequence translation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*.

Andrew Zhai and Hao-Yu Wu. 2019. Classification is a strong baseline for deep metric learning. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, page 91. BMVA Press.

Ying Zhang, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2021. Target-oriented fine-tuning for zero-resource named entity recognition. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1603–1615, Online. Association for Computational Linguistics.

Bo Zheng, Li Dong, Shaohan Huang, Wenhui Wang, Zewen Chi, Saksham Singhal, Wanxiang Che, Ting Liu, Xia Song, and Furu Wei. 2021. Consistency regularization for cross-lingual fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3403–3417, Online. Association for Computational Linguistics.

Ran Zhou, Xin Li, Lidong Bing, Erik Cambria, Luo Si, and Chunyan Miao. 2022. ConNER: Consistency training for cross-lingual named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8438–8449, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ayah Zirikly and Masato Hagiwara. 2015. Cross-lingual transfer of named entity recognizers without parallel corpora. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 390–396, Beijing, China. Association for Computational Linguistics.

A Appendix

A.1 Dataset Statistics

Table A.1 reports the dataset statistics for CoNLL and WikiAnn.

A.2 Other Implementation Details

All experiments are conducted on a Tesla V100 (32GB). The total of trainable parameters ($\mathcal{M}_{trans}$ and $\mathcal{M}_{tgt}$) for our model with mBERT-base-cased as the encoder is 172M and the training time is about 35 mins for one iteration. With XLM-R-large as our base encoder, the total of trainable parameters are 822M and the training takes about 90 mins for one iteration.

A.3 Baselines

We consider the following start-of-the-art baselines:

Language	Statistic	Train	Dev	Test
English (en)	N_S	14,042	3,252	3,454
(CoNLL-2003)	N_E	23,499	5,942	5,648
German (de)	N_S	12,167	2,875	3,009
(CoNLL-2003)	N_E	11,851	4,833	3,673
Spanish (es)	N_S	8,405	1,926	1,524
(CoNLL-2002)	N_E	18,798	4,351	3,558
Dutch (nl)	N_S	15,836	2,895	5,202
(CoNLL-2002)	N_E	13,344	2,616	3,941
English (en)	N_S	20,000	10,000	10,000
(WikiAnn)	N_E	27,931	14,146	13,958
Arabic (ar)	N_S	20,000	10,000	10,000
(WikiAnn)	N_E	22,501	11,267	11,259
Hindi (hi)	N_S	5,000	1,000	1,000
(WikiAnn)	N_E	6,124	1,226	1,228
Chinese (zh)	N_S	20,000	10,000	10,000
(WikiAnn)	N_E	24,135	12,017	12,049

Table A.1: Dataset statistics. N_S: the number of sentences, N_E: the number of entities.

mBERT (Wu and Dredze, 2019) and **XLM-R** (Conneau et al., 2020) directly train an NER model on the labeled data in the source language, with mBERT and XLM-R as the basic encoder, respectively.

BERT-align (Wu and Dredze, 2020) tries to explicitly add word-level contrastive alignment loss to enhance the mBERT representation.

AdvCE (Keung et al., 2019) exploits adversarial learning on source- and target-language text to avoid learning language-specific information.

AdvPicker (Chen et al., 2021) leverages adversarial learning to learn language-shared features and then selects the less language-specific sentences in target-language unlabeled text for knowledge distillation.

MulDA (Liu et al., 2021) proposes the labeled sequence translation method for data projection from source-language NER data, a generative model is further applied to augment more diverse examples in the target language.

UniTrans (Wu et al., 2020b) unifies model- and translation-data-based-transfer via knowledge distillation.

TOF (Zhang et al., 2021) leverages the labeled data for machine reading comprehension task on target language to help the NER task in cross-lingual transfer.

TSL (Wu et al., 2020c) proposes knowledge distillation to use unlabeled target-language data for cross-lingual NER.

RIKD (Liang et al., 2021) proposes a reinforcement learning algorithm to iteratively select reliable pseudo-labels for knowledge distillation.

MTMT (Li et al., 2022) proposes multi-task multi-teacher knowledge distillation, which further leverages the entity similarity task.

xTune (Zheng et al., 2021) leverages unlabeled translation text and other word-level data augmentation techniques for consistency training.

ConNER (Zhou et al., 2022) conducts span-level consistency training on unlabeled target-language data using translation and further applies dropout-based consistency training on the source-language data.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
"Limitations" section.

☐ A2. Did you discuss any potential risks of your work?
Not applicable. We study cross-lingual NER task on public datasets, our work doesn't have potential risks.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract & Sec 1.

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

B ☑ Did you use or create scientific artifacts?

Sec 4

☑ B1. Did you cite the creators of artifacts you used?
Sec 4

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Sec 4

☐ 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)?
Not applicable. These datasets are all public for research purpose.

☐ 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?
Not applicable. We conduct experiments on public datasets.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Sec 4.

☑ 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.
Appendix.

C ☑ Did you run computational experiments?

Sec 4.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
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?
Sec 4.

☑ 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?
Sec4.

☑ 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.)?
Sec 4.

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.