Discrepancy and Uncertainty Aware Denoising Knowledge Distillation for Zero-Shot Cross-Lingual Named Entity Recognition

Ling Ge[1], Chunming Hu[1,2,3,*], Guanghui Ma[1], Jihong Liu[4,*], Hong Zhang[5]
[1] School of Computer Science and Engineering, Beihang University, Beijing, China
[2] College of Software, Beihang University, Beijing, China
[3] Zhongguancun Laboratory, Beijing, China
[4]School of Mechanical Engineering and Automation, Beihang University, Beijing, China
[5]National Computer Network Emergency Response Technical Team / Coordination Center of China, Beijing, China
{geling, hucm, maguanghui, ryukeiko}@buaa.edu.cn, zhangh@isc.org.cn

Abstract

The knowledge distillation-based approaches have recently yielded state-of-the-art (SOTA) results for cross-lingual NER tasks in zero-shot scenarios. These approaches typically employ a teacher network trained with the labelled source (rich-resource) language to infer pseudo-soft labels for the unlabelled target (zero-shot) language, and force a student network to approximate these pseudo labels to achieve knowledge transfer. However, previous works have rarely discussed the issue of pseudo-label noise caused by the source-target language gap, which can mislead the training of the student network and result in negative knowledge transfer. This paper proposes an discrepancy and uncertainty aware Denoising Knowledge Distillation model (DenKD) to tackle this issue. Specifically, DenKD uses a discrepancy-aware denoising representation learning method to optimize the class representations of the target language produced by the teacher network, thus enhancing the quality of pseudo labels and reducing noisy predictions. Further, DenKD employs an uncertainty-aware denoising method to quantify the pseudo-label noise and adjust the focus of the student network on different samples during knowledge distillation, thereby mitigating the noise's adverse effects. We conduct extensive experiments on 28 languages including 4 languages not covered by the pre-trained models, and the results demonstrate the effectiveness of our DenKD.

Introduction

Named Entity Recognition (NER) is a fundamental information extraction task that aims to identify and classify text spans into predefined entity classes. Recently, tremendous advances in deep learning have propelled NER towards SOTA performance (Ge et al. 2023; Huang et al. 2023). However, the success of these deep learning-based methods depends on large-scale, manually annotated data. Due to the lack of labeled training data, most languages have yet to benefit from these technical advances. In consequence, researchers begin to focus on cross-lingual transfer learning in zero-shot (no annotated data) scenarios, proposing various approaches (Huang, May, and Peng 2019; Bari, Joty, and Jwalapuram 2020; Plank 2021; Guo et al. 2022). These approaches effectively transfer language-independent knowledge from high-resource (source) languages to zero-shot (target) languages and are expected to achieve good performance for target languages.

Among these, the knowledge distillation-based approaches have yielded the most remarkable results and gained widespread attention (Wu et al. 2020a; Liang et al. 2021; Chen et al. 2021; Zeng et al. 2022; Ma et al. 2022; Ge et al. 2023). These approaches typically train a teacher network with the source language to infer pseudo-soft labels for the target language, and then force the student network to mimic the teacher's inference to transfer knowledge. Since soft labels contain valuable information (e.g., inter-label relations) (Hinton et al. 2015), the student network can perform better than the teacher network in the target language.

However, due to the language gap between source and target languages, the teacher network that only accesses the ground-truth labels of the source language inevitably infers low-quality (noisy) pseudo labels for the target language. If these noisy pseudo labels are directly utilized for knowledge distillation, they will mislead the training process of the student network, resulting in negative knowledge transfer (Feng et al. 2021). Only a few works have focused on this issue. RIKD (Liang et al. 2021) combines multiple data features to filter low-valuable target data, and AdvPicker (Chen et al. 2021) picks target samples similar to the source data for knowledge distillation. These methods attempt to reduce noise via data selection. Unfortunately, the pseudo labels of the selected samples still contain different-level noise (Xu et al. 2022; Qin et al. 2022), which can confound the optimization of the student network. Furthermore, these methods neglect to improve the quality of pseudo labels that determines the accuracy degree of knowledge transfer.

In this paper, we propose an **Den**oising **K**nowledge **D**istillation model (named **DenKD**) to mitigate the noise issue via enhancing the pseudo labels' quality and adjusting the student network's focus on different samples. Firstly, DenKD introduces a discrepancy-aware denoising representation learning method to enable the teacher network to learn discriminative representations for the target language. In general, better representations lead to better classifiers (Zhou et al. 2020; Wang et al. 2021b), yielding higher-quality pseudo labels. In addition, to further mitigate the

*Corresponding authors.
Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

damage from noise contained in pseudo labels, DenKD employs an uncertainty-aware denoising method to quantify the noise of pseudo labels and precisely adjust their impacts on the student network.

Specifically, studies based on learning from noisy data reveal that, if different neural networks have large prediction discrepancies for the same sample, this sample tends to be a noisy sample, i.e., one with a noisy pseudo label (Venkat et al. 2020; Du et al. 2021). Therefore, we adopt a two-classifier teacher network and employ the prediction discrepancy between two classifiers as guiding signals to identify noisy samples and optimize their representations inspired by (Saito et al. 2018). On the one hand, we encourage these two classifiers to maximize the prediction discrepancy to identify as many noisy samples as possible. On the other hand, we force the encoder of the teacher network to minimize the prediction discrepancy to avoid being identified as noise samples. With this max-min adversarial training mechanism, the teacher network will be forced to create discriminative representations for the target language, thereby reducing the target language pseudo-labelling noise.

Additionally, we rely on uncertainty to quantify the pseudo-label noise for each target sample, with higher uncertainty values indicating higher pseudo-label noise and poorer reliability. In detail, we apply Monte-Carlo Dropout (Gal and Ghahramani 2016; Mukherjee et al. 2020) to our teacher network and use the variance of the teacher network's multiple predictions as the uncertainty of pseudo labels. During the knowledge distillation, the pseudo-label uncertainty is employed as the weight of distillation loss to dynamically adjust the focus of the student network on different samples, thereby mitigating the negative impact of noisy samples on the student network.

Contributions: (1) We propose the DenKD model for zero-shot cross-lingual NER tasks, which improves the model's performance on the target language by mitigating the negative knowledge transfer issue caused by the pseudo-label noise. (2) We present a divergence-aware denoising representation learning method to improve the quality of pseudo labels provided by the teacher network for the target language. (3) We propose an uncertainty-aware denoising method to reduce the negative impact of pseudo-label noise on the student network during knowledge distillation. (4) Experimental results on 28 languages, including 4 languages not covered by the pre-trained models, validate the effectiveness of our DenKD model.

Related Works

Cross-Lingual NER

Various strategies have been investigated to tackle the zero-resource challenge for cross-lingual NER, such as translation-based (Xie et al. 2018; Liang et al. 2021), direct transfer-based (Wu and Dredze 2019; Wu et al. 2020c), and knowledge distillation-based approaches (Wu et al. 2020a; Liang et al. 2021; Chen et al. 2021; Zeng et al. 2022; Fu et al. 2022). The direct transfer-based methods yield inferior results, since the target data has not been effectively utilized. The translation-based approaches rely on high-quality translation resources, which is arduous in practice. The knowledge distillation-based methods, which encourage the student network to learn knowledge from the teacher network, achieve the most appealing results, with recent advances such as MSD (Ma et al. 2022) and ProKD (Ge et al. 2023).

Only a limited number of works have considered the pseudo-label noise issue in knowledge distillation. RIKD (Liang et al. 2021) employs various data features to select valuable target data, but the majority of the filtering factors, like sentence length and entity number, have no correlation with the label noise. AdvPicker (Chen et al. 2021) uses the degree of feature alignment with the source language for target data selection. Unfortunately, the selected samples are all similar to the source data, which can exacerbate the bias of the teacher network towards the source language and thus hinder the model's performance in the target language. ContProto (Zhou et al. 2023) corrects pseudo-labels generated by the teacher via prototype learning, which is essentially clustering-based representation learning, and will inevitably introduce noise. Moreover, it only adjusts the pseudo-labels already produced, neglecting to improve the ability of the teacher network to produce high-quality pseudo-labels, which we believe is essential.

Knowledge Distillation

Knowledge distillation(Hinton et al. 2015) enables knowledge transfer from the teacher network to the student network via enforcing the student network to mimic the logit output of the teacher network. It has now been widely used in areas such as model compression(Liu et al. 2022), machine translation (Wang et al. 2021a), and relation extraction (Tan et al. 2022). Wu et al. (2020a) pioneers the use of knowledge distillation architectures for cross-lingual NER tasks and achieves surprising performance. Since then, researchers in this field have focused on knowledge distillation-based approaches and developed various improvements(Zeng et al. 2022; Fu et al. 2022). Following the previous works above, this paper employs the knowledge distillation architecture as the backbone for cross-lingual NER. The main issue addressed in this paper is the pseudo-label noise in knowledge distillation resulting from the language gap, which is rarely mentioned in previous works.

Methodology

This paper follows the previous works (Wu et al. 2020a; Ge et al. 2023) and models the NER task as a sequence labeling problem, i.e., given a sentence $\boldsymbol{x} = \{x_i\}_{i=0}^{L}$, assigning a single label y_i to each token x_i and obtaining $\boldsymbol{y} = \{y_i\}_{i=0}^{L}$ with L denoting the sentence length. For cross-lingual NER in zero-shot scenarios, given labelled source language data $\mathcal{D}_s = \{(\boldsymbol{x}_m^s, \boldsymbol{y}_m^s)\}_{m=1}^{N_s}$ and unlabeled target language data $\mathcal{D}_t = \{(\boldsymbol{x}_m^t)\}_{m=1}^{N_t}$, the goal of this task is to build a good NER model in the target language, with the training on labeled source language data $\mathcal{D}_s$ and unlabeled target data$\mathcal{D}_t$.

Overall Architecture

As shown in Figure 1, our model DenKD follows a knowledge distillation architecture and consists of two main com-

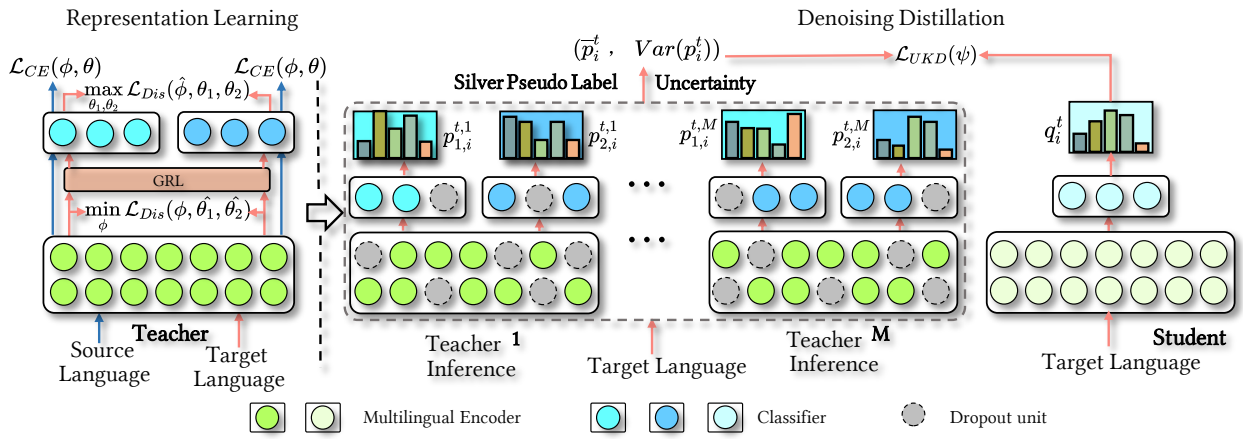

Figure 1: Model Architecture. The DenKD mitigates the noise issue from two aspects: (1) enhancing the pseudo labels' quality, and (2) adjusting the student network's focus on different samples.

ponents, namely (1) the denoising representation learning module for the teacher network, which utilizes a discrepancy-aware representation learning method to reduce the noise of pseudo-labels inferred by the teacher network for the target language, and (2) the denoising distillation module for the student network, which uses an uncertainty-aware denoising method to further mitigate the adverse effects of noise on the student network. Afterward, we elaborate on our DenKD using the two primary methods in these components as clues.

Discrepancy-aware Denoising Representation Learning

In this part, we adopt a teacher network with two classifiers and exploit the prediction discrepancy between them for target samples to guide the teacher network's representation learning through adversarial training between the encoder and two classifiers, inspired by (Saito et al. 2018). Thus, we can repeatedly identify noisy samples and optimize their representations. Eventually, our teacher network can generate discriminative representations for target language.

Concretely, our teacher network contains a multilingual encoder $\boldsymbol{E}$ and two classifiers $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ with the same structure but distinct initializations.

Encoder Formally, given an input sequence $\boldsymbol{x} = \{x_i\}_{i=0}^{L}$, the encoder maps this sentence into a multilingual shared semantic space, and outputs its semantic representation $\boldsymbol{h}$:

$$\boldsymbol{h} = \boldsymbol{E}(\boldsymbol{x};\phi) \tag{1}$$

where $\boldsymbol{h} = \{h_i\}_{i=0}^{L}$ with h_i denoting the representation of token x_i, and ϕ is the encoder parameters.

Classifiers We pass this representation $\boldsymbol{h}$ into $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$, which are both a two-layer MLP followed by a softmax function. Then, two sequences of probability distributions are obtained:

$$\boldsymbol{p_1} = \boldsymbol{C}_1(\boldsymbol{h};\theta_1),\ \ \boldsymbol{p_2} = \boldsymbol{C}_2(\boldsymbol{h};\theta_2) \tag{2}$$

where $\boldsymbol{p_1} = \{p_{1,i}\}_{i=0}^{L}$ and $\boldsymbol{p_2} = \{p_{2,i}\}_{i=0}^{L}$. The $p_{1,i}$ and $p_{2,i}$ represent the probability produced by classifier $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ for token x_i. The θ_1 and θ_2 indicate the learning parameters of two classifiers.

Source Language Loss To optimize the encoder and two classifiers of the teacher network, firstly, we resort to the gold label of the source language and employ the cross-entropy loss to perform training, expressed as:

$$\mathcal{L}_{CE}(\phi,\theta) = -\frac{1}{N_s L}\sum_{D_s}\sum_{i=0}^{L} y_i^s log(p_i^s) \tag{3}$$

where N_s indicates the number of the sentences in dataset D_s, and y_i^s represents the gold label for token x_i^s. Note that, for classifiers $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$, we optimize them separately using the Equation 3. Subsequently, we can obtain the trained teacher network with two classifiers.

Prediction Discrepancy Loss Since the two classifiers are initialized differently and optimized separately, they naturally generate a prediction discrepancy for the same input. Intuitively, given a target language token x_i^t, relying on Equation 1 and 2, we can obtain its two different soft labels $p_{1,i}^t$ and $p_{2,i}^t$. Thus, the prediction discrepancy loss of these two classifiers for the target samples can be defined as:

$$\mathcal{L}_{Dis}(\phi,\theta_1,\theta_2) = \sum_{D_t}\sum_{i=1}^{i=L} d(p_{1,i}^t, p_{2,i}^t) \tag{4}$$

where $d(,)$ is the metric function of the prediction discrepancy for one token. This paper models the prediction discrepancy as the $L1$-distance between the probability outputs of two classifiers. Note that, we can also use other metrics to measure this discrepancy, but it is not this paper's focus.

Afterward, we can incorporate the unlabeled target samples into the training of the teacher in an adversarial way.

Maximize the Prediction Discrepancy In general, the target samples near the decision boundaries of $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$,

are more likely to be noisy samples, which are easily misclassified by different classifiers (Han et al. 2018; Saito et al. 2018; Zheng et al. 2021). Also, the predictions of the above two classifiers for these samples differ significantly. Based upon this, we resort to updating the parameters of these two classifiers to maximize their prediction discrepancy to detect as many noise samples as possible:

$$(\hat{\theta_1}, \hat{\theta_2}) = \arg\max_{\theta_1,\theta_2} \mathcal{L}_{Dis}(\hat{\phi}, \theta_1, \theta_2) \tag{5}$$

where $\hat{\theta_1}$, $\hat{\theta_2}$ and $\hat{\phi}$ indicate the optimal parameters.

Minimize the Prediction Discrepancy Meanwhile, the target samples far from the decision boundaries tend to be samples with high-quality pseudo labels, for which the two classifiers differ less in their predictions, or even perform consistently. Considering this, to avoid being detected as noise samples by these two classifiers, we depend on training the encoder to minimize the discrepancy to generate target features away from two decision boundaries. The objective can be expressed as:

$$\hat{\phi} = \arg\min_{\phi} \mathcal{L}_{Dis}(\phi, \hat{\theta_1}, \hat{\theta_2}) \tag{6}$$

Ideally, according to Equation 5 and 6, we can iteratively train the encoder E and two classifiers $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$. In practical implementation, the above adversarial learning can be realized by inserting a gradient reversal layer (GRL) (Ganin et al. 2016) between the encoder and two classifiers. The GRL is a constant transform in forward propagation, while in backward propagation, it reverses the gradient signal by multiplying a fact $-\lambda$ to the subsequent layer gradient. As such, updates for maximizing the discrepancy via classifiers and minimizing it via the encoder can be performed simultaneously when the gradient passes through. Note that, to ensure stable training (Long et al. 2018), we follow the previous work (Ganin et al. 2016) and let $\lambda = \frac{2}{1+\exp(-10\cdot(e/Epo))} - 1$. The Epo is the number of total training epochs, and e is the current epoch. As e increases, the λ changes from 0 to 1.

To this end, the total loss function for training the teacher network with labeled source data and unlabeled target data can be expressed as:

$$\begin{aligned}\mathcal{L}_{tea}(\phi, \theta_1, \theta_2) = \mathcal{L}_{CE}(\phi, \theta_1, \theta_2) \\ - \mathcal{L}_{Dis}(\theta_1, \theta_2) + \lambda \mathcal{L}_{Dis}(\phi)\end{aligned} \tag{7}$$

Uncertainty-aware Denoising Knowledge Distillation

In this section, we employ an uncertainty-aware pseudo-label denoising strategy in the student network to further mitigate the negative impact of pseudo-label noise during knowledge distillation.

Uncertainty Estimation Specifically, we rely on uncertainty to model the pseudo-label noise, where more significant uncertainty indicates higher noise. We follow the same motivation for training teacher networks: different neural networks have more difficulty making consistent predictions for noisy samples than for clean samples (samples with high-quality pseudo labels). Consequently, we propose using the variance of the teacher network's multiple predictions to estimate the uncertainty of pseudo labels.

To achieve this, we apply Monte-Carlo Dropout (Gal and Ghahramani 2016), the recent advances in uncertainty estimation (Mukherjee et al. 2020), to our two-classifier teacher network, to obtain multiple prediction values.

In more detail, we use dropout after different hidden layers in our trained teacher network and forward propagate M times to perform inference. This allows us to sample M masked model weights $\{\phi_m, \theta_m\}_{m=1}^{M}$ obeying the Dropout distribution (Srivastava et al. 2014), and each m corresponds to a sub-model $Teacher^m$. This sub-model, integrated with the two classifiers $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$, yields 2 predictions $p_{1,i}^{t,m}$ and $p_{2,i}^{t,m}$ for one input target language token x_i^t. The above process can be formalized as follows:

$$p_{1,i}^{t,m}, p_{2,i}^{t,m} = Teacher^m(x_i^t) \tag{8}$$

Overall, M times of inferences produce $2M$ prediction values. Afterward, we use the mean of these predictions as the pseudo-silver label $\overline{p}_i^t$ for the token x_i^t to guide the knowledge distillation learning. The pseudo-silver label $\overline{p}_i^t$ can be denoted as:

$$\overline{p}_i^t = \frac{1}{2M} \sum_{m=1}^{M} (p_{1,i}^{t,m} + p_{2,i}^{t,m}) \tag{9}$$

Meanwhile, we use the variance of these predictions $Var(p_i^t)$ to estimate the uncertainty of the pseudo label of token x_i^t , denoted as:

$$Var(p_i^t) = \sum_{m=1}^{M} [(p_{1,i}^{t,m} - \overline{p}_i^t)^2 + (p_{2,i}^{t,m} - \overline{p}_i^t)^2] \tag{10}$$

Here, when performing uncertainty estimation, we utilize prediction variance generated by Monte-Carlo Dropout (for short Monte-Carlo variance) instead of two-classifier discrepancy. The reason is that our two classifiers, trained adversarially, will make consistent predictions for the samples. In contrast, since Monte-Carlo Dropout is equivalent to integrating multiple models for prediction, the Monte-Carlo variance is more sensitive to label noise and fits more accurately (see Section Uncertainty-Estimation Strategy Study for details). As the Monte-Carlo variance involves forward inference T times, requireing some time cost, we did not employ it for teacher network training.

Denoising Knowledge Distillation To transfer NER knowledge to the student network, we depend on the classical knowledge distillation learning (Hinton et al. 2015) to force the student network to mimic the pseudo silver label $\overline{p}_i^t$. In particular, the student network utilizes the same design as the teacher network, with the difference that only one classifier is added after the encoder.

To further reduce the negative impact of noisy target samples, we rely on the pseudo-label uncertainty of each

Methods	az	bg	cy	da	eu	hr	hy	hu	ko	ms	nl	pl	ro	sh
FTDT (2019)	66.03	77.80	58.85	83.10	62.98	78.90	52.12	76.03	57.38	70.74	82.55	80.65	76.56	57.25
Single-TS (2020a)	67.23	78.98	59.48	84.47	64.16	80.02	53.56	76.72	58.94	72.22	83.80	81.86	78.53	59.81
RIKD (2020a)	68.11	79.71	60.85	84.83	65.71	80.13	55.17	77.32	58.03	72.92	84.65	82.14	78.54	59.51
AdvPicker (2021)	68.67	78.02	57.39	85.93	63.51	76.78	52.49	76.13	59.25	70.90	84.27	82.28	76.77	59.28
DualNER (2022)	68.64	79.20	61.78	85.11	65.13	80.27	55.92	77.46	57.48	71.10	84.36	82.11	79.91	60.71
MSD (2022)	71.84	79.03	62.24	84.71	65.41	82.12	56.22	76.57	61.44	71.61	84.23	82.46	76.81	58.35
ProKD (2023)	71.79	80.18	61.71	85.85	71.89	82.85	62.58	80.40	61.31	**76.53**	84.73	82.60	80.82	74.50
DenKD (Ours)	**75.75**	**80.84**	**73.19**	**86.33**	**74.54**	**83.53**	**65.20**	**80.71**	**62.51**	74.04	**85.69**	**83.39**	**83.20**	**75.17**
Methods	**fr**	**es**	**ru**	**zh**	**de**	**ja**	**pt**	**hi**	**ka**	**af**	**eo**	**no**	**zh-yue**	**Avg**
FTDT (2019)	80.20	74.55	64.09	43.85	78.64	29.82	80.95	64.79	64.68	77.29	58.02	76.50	42.68	67.30
Single-TS (2020a)	80.38	77.18	66.02	45.60	79.96	31.19	82.26	65.26	66.20	78.14	59.64	79.72	45.53	68.22
RIKD (2020a)	81.20	77.79	65.63	47.38	80.20	31.49	82.66	65.69	66.83	78.00	59.30	76.67	45.34	69.10
AdvPicker (2021)	79.91	77.81	68.28	53.02	79.72	37.62	80.92	70.00	68.37	70.00	60.58	73.65	50.47	68.59
DualNER (2022)	80.92	78.42	65.06	47.84	80.17	31.07	82.87	66.24	67.28	77.29	59.46	76.92	45.46	69.20
MSD (2022)	81.16	75.75	67.71	47.97	80.62	33.34	82.82	66.34	69.23	79.68	61.33	79.74	48.31	69.89
ProKD (2023)	81.45	79.19	65.59	51.80	79.74	33.72	83.83	**70.72**	69.07	68.26	59.94	77.29	**50.71**	71.45
DenKD (Ours)	**82.34**	**84.68**	**69.35**	**55.62**	**82.50**	**37.90**	**85.30**	69.76	**69.30**	**80.40**	**61.39**	**82.95**	50.19	**73.92**

Table 1: The comparison in token-level F1 value (%) of different methods. The best results are displayed in bold.

sample to dynamically adjust the weight of distillation loss. In detail, for each target sample involved in distillation, we weight its original distillation loss using a factor $\exp(-Var(p_i^t))$ that is negatively correlated with uncertainty. If a sample's pseudo label has high uncertainty, this factor will return a small value to reduce its loss. Thus, the distillation loss can be formulated as:

$$\mathcal{L}_{DKD}(\psi) = \frac{1}{N_t L} \sum_{x \in D_t} \sum_{i=1}^{L} (\overline{p}_i^t - q_i^t)^2 \cdot \exp(-Var(p_i^t)) \tag{11}$$

where q_i^t denotes the probability distribution produced by the student network. ψ is the student network's learning parameters. Following previous works (Yang et al. 2020; Wu et al. 2020a), we use the MSE loss to measure the prediction difference between the teacher and student networks.

After the weighting operation, the student can focus more on samples in which the teacher is more certain (lower variance) and ignore more uncertain samples (higher variance), thereby mitigating the negative impact of noisy samples.

Experiments and Analysis

Experiment Setting

Datasets We employ a wildly-used benchmark NER dataset **Wikiann** (Rahimi, Li, and Cohn 2019) for experiments, which is annotated with LOC (location), PER (person), and ORG (organisation) tags. For each language-specific dataset, we have the standard training, development, and evaluation sets. We chose 28 language datasets for our experiments, containing 4 low-resource languages, namely Esperanto (eo), Norwegian (no), Serbo-Croatian (sh), and Cantonese (zh-yue), which the pre-trained model has not been specifically trained on.

For preprocessing, we utilize the word piece (Wu et al. 2016) to tokenize words and label sub-words via the BIO scheme. We take English as the source language and others as the target language, respectively. To simulate the zero-shot scenario, we train the model on the labeled source and unlabeled target language training sets, validate the model on the source language development set, and evaluate the model on the target language test set.

Implementation Details Following previous work (Wu et al. 2020a; Ge et al. 2023), we adopt the pre-trained multilingual BERT (Pires et al. 2019) as the language encoder and utilize the token-level F1 value as the evaluation metric. For all experiments, we perform five runs to present the average performance. We leverage Adam (Kingma and Ba 2015) as the optimizer, and set the maximum sequence length to 128, the dropout to 0.5, and the batch size to 128. We utilize the grid-search way to obtain the optimal super-parameters, including the learning rate selected from $\{1e-5, 5e-5\}$ and the number of Monte Carlo inferences M selected from $\{10, 20, 30, 40, 50\}$. Following previous studies (Chen et al. 2021; Ge et al. 2023), we freeze the parameters of the embedding layer and the bottom three layers of the multilingual BER during training, and we only consider the first sub-word tokenized by word-piece in our loss function. For the implementation of the baseline and SOTA models, we utilize the original open source code from their papers, including the FTDT [1], Sing-TS [2], AdvPicker [3] and MSD [4]. For RIKD, DualNER, and ProKD, we reproduce the model according to the description of their papers.

[1] https://github.com/shijie-wu/crosslingual-nlp

[2] https://github.com/microsoft/vert-papers/tree/master/papers/SingleMulti-TS

[3] https://github.com/microsoft/vert-papers/tree/master/papers/AdvPicker

[4] https://github.com/Mckysse/MSD

Methods	fr	es	ru	zh	de	ja	pt	hi	ka	af	eo	no	zh-yue	Avg
DenKD	**82.34**	**84.68**	**69.35**	**55.62**	**82.50**	**37.90**	**85.30**	**69.76**	**69.30**	**80.40**	**61.39**	**82.95**	**50.19**	**70.12**
DenKD$_{w/o\ \mathrm{RL}}$	81.59	79.86	66.28	46.95	81.12	33.85	83.68	67.07	68.32	78.64	60.22	79.78	46.36	67.21 (↓2.91)
DenKD$_{w/o\ \mathrm{UD}}$	82.09	82.53	68.61	54.47	81.38	36.73	84.52	68.23	68.77	79.76	60.35	82.01	49.48	69.13 (↓0.99)
DenKD$_{w/\ \mathrm{TCD}}$	81.88	84.59	69.10	55.46	79.86	37.36	85.02	69.63	69.13	79.93	60.42	82.59	49.77	69.60 (↓0.52)
DenKD$_{w/\ \mathrm{OI}}$	82.17	84.14	69.03	53.88	81.84	37.86	85.28	69.33	69.08	80.30	61.21	82.67	49.88	69.74 (↓0.38)

Table 2: Ablation studies on model components.

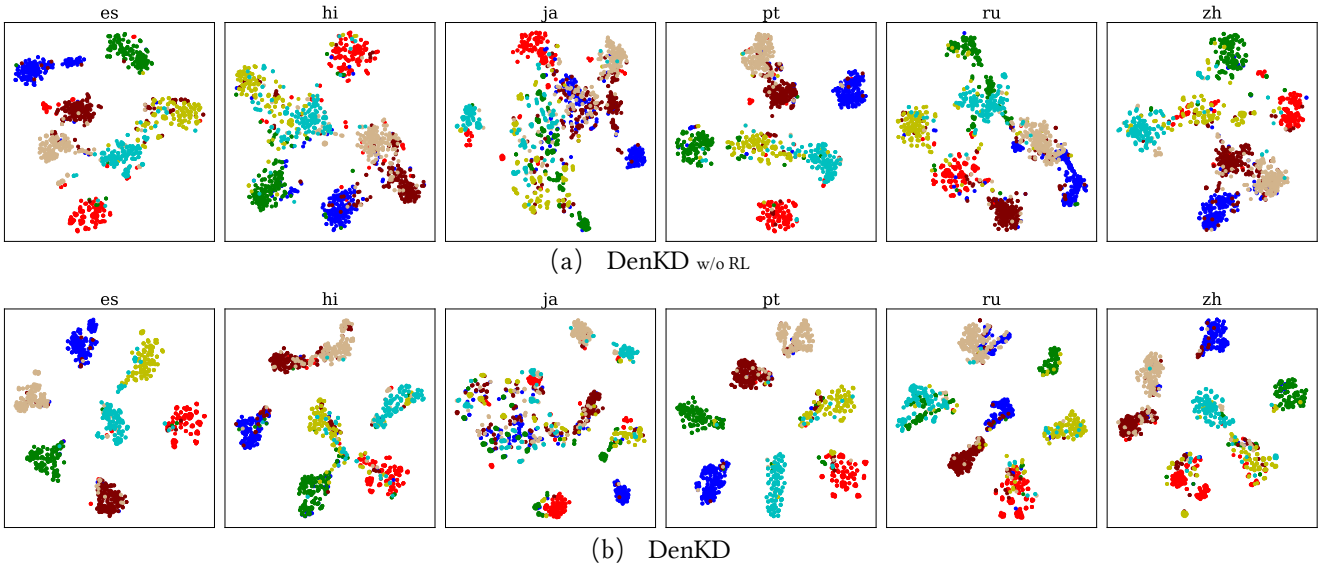

Figure 2: DenKD can enable the teacher network to learn better discriminative class representations, thereby reducing the target language pseudo-labelling noise. Different colours represent different classes.

Evaluation Results and Analysis

Baseline and SOTA We compare our DenKD with several related methods, (1) **FTDT** (Wu and Dredze 2019): a baseline method that only uses the pre-trained model to direct transfer; (2) **Single-TS** (Wu et al. 2020a): a vanilla knowledge distillation method; (3) **RIKD** (Liang et al. 2021): a denoising knowledge distillation method that iteratively selects valuable data for training; (4) **AdvPicker** (Chen et al. 2021): a SOTA denoising knowledge distillation method that selects target samples similar to the source language; (5) **DualNER** (Zeng et al. 2022): a multi-task knowledge distillation method that combines sequence tagging and span prediction tasks into a unified framework; (6) **MSD** (Ma et al. 2022): a multi-channel knowledge distillation method that constructs multiple channels between the teacher and student. (7) **ProKD** (Ge et al. 2023): a SOTA knowledge distillation method that employs the prototypical class-wise alignment and prototypical self-training to acquire language-independent and language-specific knowledge. Note that, we do not compare our DenKD with methods that use additional resources, e.g., UniTrans (Wu et al. 2020b) utilizes machine translation to augment data, ContProto (Zhou et al. 2023) employs large amounts of unlabelled data to improve the generalisability of the model.

Performance Comparison As shown in Table 1, DenKD yields the best results for most target languages. Compared to the most competitive method, ProKD, our method improves on the average F1 by 2.47%. For instance, for the Spanish(es) language, our method outperforms ProKD by 5.49%. In particular, compared to the denoising SOTA method, Advpicker, our average F1 is 6.87% higher. Analytically, ProKD overemphasizes class prototype alignment and disregards class representations learning, which prevents it from producing high-quality pseudo labels for target samples. Advpicker attempts to reduce noise by data filtering, however, it disregards that picked samples still contain noise. Our method, in contrast, considering both the quality of the pseudo labels and the pseudo-label noise of each sample engaged in distillation, achieves superior results.

Ablation Study To explore contributions of different factors, we conduct ablation experiments with four variant models. (1) DenKD$_{w/o\ \mathrm{RL}}$ removes the discrepancy-aware representation learning in teacher network training. (2) DenKD$_{w/o\ \mathrm{UD}}$ wipes out the uncertainty weighting for sample loss during knowledge distillation. (3) DenKD$_{w/\ \mathrm{TCD}}$ utilizes two-classifier discrepancy instead of Monte-Carlo variance to estimate pseudo-label uncertainty. (4) DenKD$_{w/\ \mathrm{OI}}$ employs one inference of the teacher net-

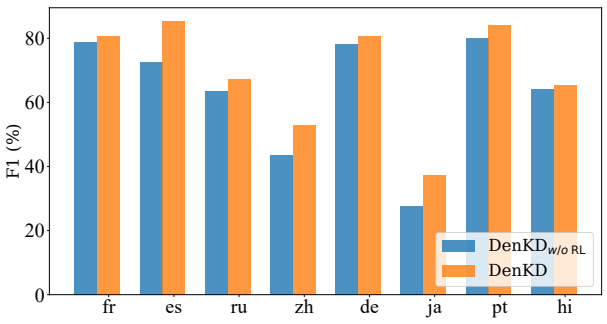

Figure 3: Pseudo-Label Quality Study.

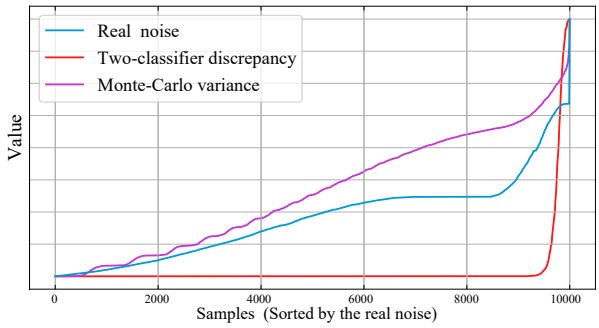

Figure 4: Noise and Uncertainty Curves.

work as the silver pseudo label instead of multiple prediction's mean.

As observed in Table 2, the F1 score of DenKD$_{w/o\ \mathrm{RL}}$ drops by 2.91% compared to DenKD. This indicates that model's discriminative representation learning for target data can effectively improve the model's generalization to this language, as it can improve the quality of the pseudo labels, benefiting the optimization of the student network. The F1 value of DenKD$_{w/o\ \mathrm{UD}}$ decreases by 0.99% compared to DenKD, which well validates the effectiveness of noise reduction using pseudo-label uncertainty. DenKD $_{w/\ \mathrm{TCD}}$ is 0.52% lower than DenKD, indicating that our Monte-Carlo variance can model noise better. The slightly lower F1 value for DenKD $_{w/\ \mathrm{OI}}$ than DenKD suggests that combining multiple predictions can result in higher quality pseudo labels than single prediction.

Visualize Token-Level Representations To demonstrate that our teacher can learn better representations for the target language, we randomly select 200 token samples for each class from target data for visualisation experiments. We feed them to the teacher of DenKD$_{w/o\ \mathrm{RL}}$ and DenKD to obtain token-level representations, respectively. Then we visualize these representations using T-SNE (Van der Maaten and Hinton 2008) and display the results in Figure 2, where different colors represent different classes. As observed, the representations obtained by DenKD$_{w/o\ \mathrm{RL}}$ have significant overlap between different classes, resulting in indistinguishable classes. Even in some languages, most classes are overlapping and miscellaneous, e.g., Japanese (ja). Owing to discrepancy-aware representation learning, our teacher can reduce sample overlap between classes and produce more distinct class representation boundaries.

Pseudo-Label Quality Study To verify that our teacher network can improve the quality of pseudo label for the target language, we report the predicted F1 of the teacher networks of DenKD and DenKD$_{w/o\ \mathrm{RL}}$ for eight target languages (Figure 3). For a teacher, a higher F1 value indicates a closer prediction to the gold label, and then the pseudo-label produced by this teacher will be more accurate. As observed, the teacher of DenKD outperforms DenKD$_{w/o\ \mathrm{RL}}$ in F1 value across the board, for example, by a large margin on Spanish (es) and Chinese (zh) language. This is because our teacher can learn discriminative representations for target samples by adversarially identifying and reducing noisy samples, thus improving the quality of pseudo labels.

Uncertainty-Estimation Strategy Study This part compares two strategies for uncertainty estimation: the two-classifier discrepancy and Monte-Carlo variance. Firstly, we employ our trained teacher network to generate pseudo labels for the target samples. Then we compute the LI distance between the obtained pseudo label and gold label for each sample, and use it as the pseudo-label noise (real noise). Taking French (fr) as an example, we select the top 10000 noisy samples sorted by the real noise to depict the noise curve (blue). Also, we utilize the two-classifier discrepancy and Monte-Carlo variance strategies to estimate the pseudo-label uncertainty of these samples, with the uncertainty curves shown in Figure 4. As observed, the two-classifier discrepancy (red) essentially fails to model noise, as the discrepancies between the two classifiers approach 0 after max-min adversarial training. In contrast, the uncertainty curve of Monte-Carlo variance (purple) follows the same trend as the noise curve, indicating that Monte-Carlo variance can accurately measure pseudo-label noise. This conclusion is also supported by the ablation experiment with DenKD$_{w/\ \mathrm{TCD}}$ 0.49% lower than DenKD in the average F1 value.

Conclusions and Discussions

This paper proposes a novel cross-lingual NER model DenKD to alleviate the pseudo-label noise caused by the language gap in knowledge distillation. DenKD proposes a discrepancy-aware representation learning method to improve the quality of pseudo labels. Furthermore, DenKD proposes an uncertainty-aware denoising method to mitigate noise's adverse effects. Extensive experiments on 28 languages demonstrate the effectiveness of our approach.

There are some potential limitations in this study. First, we employ a Monte Carlo approach to perform M-inference when performing pseudo-label uncertainty evaluation, which inevitably results in additional time costs. Second, the core of our approach is a knowledge distillation architecture consisting of the teacher and student networks. Compared to traditional direct transfer-based and translation-based methods, our approach will additionally increase the memory consumption of the GPU. However, these limitations do not affect the effectiveness of our model.

Acknowledgments

We sincerely thank the reviewers for their insightful comments and valuable suggestions. This work is funded by the National Key Research and Development Program of the Ministry of Science and Technology of China (No. 2021YFB1716201). Thanks for the computing infrastructure provided by Beijing Advanced Innovation Center for Big Data and Brain Computing.

References

Bari, M. S.; Joty, S. R.; and Jwalapuram, P. 2020. Zero-Resource Cross-Lingual Named Entity Recognition. 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*, 7415–7423. AAAI Press.

Chen, W.; Jiang, H.; Wu, Q.; Karlsson, B.; and Guan, Y. 2021. AdvPicker: Effectively Leveraging Unlabeled Data via Adversarial Discriminator for Cross-Lingual NER. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, 743–753. Association for Computational Linguistics.

Du, Z.; Li, J.; Su, H.; Zhu, L.; and Lu, K. 2021. Cross-Domain Gradient Discrepancy Minimization for Unsupervised Domain Adaptation. In *Proc. of CVPR*.

Feng, L.; Qiu, M.; Li, Y.; Zheng, H.; and Shen, Y. 2021. Wasserstein Selective Transfer Learning for Cross-domain Text Mining. In *Proc. of EMNLP*.

Fu, Y.; Lin, N.; Yang, Z.; and Jiang, S. 2022. A Dual-Contrastive Framework for Low-Resource Cross-Lingual Named Entity Recognition. *CoRR*, abs/2204.00796.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proc. of ICML*.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. S. 2016. Domain-Adversarial Training of Neural Networks. *J. Mach. Learn. Res.*, 17: 59:1–59:35.

Ge, L.; Hu, C.; Ma, G.; Zhang, H.; and Liu, J. 2023. ProKD: An Unsupervised Prototypical Knowledge Distillation Network for Zero-Resource Cross-Lingual Named Entity Recognition. In *Proc. of AAAI*.

Guo, Y.; Li, L.; Jiang, X.; and Liu, Q. 2022. FreeTransfer-X: Safe and Label-Free Cross-Lingual Transfer from Off-the-Shelf Models. In Carpuat, M.; de Marneffe, M.; and Ruíz, I. V. M., eds., *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, 217–228. Association for Computational Linguistics.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I. W.; and Sugiyama, M. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Proc. of NeurIPS*.

Hinton, G. E.; et al. 2015. Distilling the Knowledge in a Neural Network. *CoRR*.

Huang, X.; May, J.; and Peng, N. 2019. What Matters for Neural Cross-Lingual Named Entity Recognition: An Empirical Analysis. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *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*, 6394–6400. Association for Computational Linguistics.

Huang, Y.; Liu, W.; Zhang, X.; Lang, J.; Gong, T.; and Li, C. 2023. PRAM: An End-to-end Prototype-based Representation Alignment Model for Zero-resource Cross-lingual Named Entity Recognition. In Rogers, A.; Boyd-Graber, J. L.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, 3220–3233. Association for Computational Linguistics.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Liang, S.; Gong, M.; Pei, J.; Shou, L.; Zuo, W.; Zuo, X.; and Jiang, D. 2021. Reinforced Iterative Knowledge Distillation for Cross-Lingual Named Entity Recognition. In Zhu, F.; Ooi, B. C.; and Miao, C., eds., *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, 3231–3239. ACM.

Liu, C.; Tao, C.; Feng, J.; and Zhao, D. 2022. Multi-Granularity Structural Knowledge Distillation for Language Model Compression. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, 1001–1011. Association for Computational Linguistics.

Long, M.; Cao, Z.; Wang, J.; and Jordan, M. I. 2018. Conditional Adversarial Domain Adaptation. In *Proc. of NeurIPS*.

Ma, J.-Y.; Chen, B.; Gu, J.-C.; Ling, Z.-H.; Guo, W.; Liu, Q.; Chen, Z.; and Liu, C. 2022. WIDER & CLOSER: Mixture of Short-channel Distillers for Zero-shot Cross-lingual Named Entity Recognition. *arXiv preprint arXiv:2212.03506*.

Mukherjee, S.; et al. 2020. Uncertainty-aware Self-training for Few-shot Text Classification. In *Proc. of NeurIPS*.

Pires, T.; et al. 2019. How Multilingual is Multilingual BERT? In Korhonen, A.; Traum, D. R.; and Màrquez, L., eds., *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, 4996–5001. Association for Computational Linguistics.

Plank, B. 2021. Cross-Lingual Cross-Domain Nested Named Entity Evaluation on English Web Texts. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Findings of the*

Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021, volume ACL/IJCNLP 2021 of *Findings of ACL*, 1808–1815. Association for Computational Linguistics.

Qin, C.; et al. 2022. Robust Semi-supervised Domain Adaptation against Noisy Labels. In *Proc. of CIKM*.

Rahimi, A.; Li, Y.; and Cohn, T. 2019. Massively Multilingual Transfer for NER. In Korhonen, A.; Traum, D. R.; and Màrquez, L., eds., *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, 151–164. Association for Computational Linguistics.

Saito, K.; Watanabe, K.; Ushiku, Y.; and Harada, T. 2018. Maximum Classifier Discrepancy for Unsupervised Domain Adaptation. In *Proc. of CVPR*.

Srivastava, N.; Hinton, G. E.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1): 1929–1958.

Tan, Q.; He, R.; Bing, L.; and Ng, H. T. 2022. Document-Level Relation Extraction with Adaptive Focal Loss and Knowledge Distillation. In *Proc. of ACL Findings*.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Venkat, N.; Kundu, J. N.; Singh, D. K.; Revanur, A.; and R., V. B. 2020. Your Classifier can Secretly Suffice Multi-Source Domain Adaptation. In *Proc. of NeurIPS*.

Wang, F.; Yan, J.; Meng, F.; and Zhou, J. 2021a. Selective Knowledge Distillation for Neural Machine Translation. In *Proc. of ACL*.

Wang, P.; Han, K.; Wei, X.; Zhang, L.; and Wang, L. 2021b. Contrastive Learning Based Hybrid Networks for Long-Tailed Image Classification. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 943–952. Computer Vision Foundation / IEEE.

Wu, Q.; Lin, Z.; Karlsson, B.; Lou, J.; and Huang, B. 2020a. Single-/Multi-Source Cross-Lingual NER via Teacher-Student Learning on Unlabeled Data in Target Language. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J. R., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, 6505–6514. Association for Computational Linguistics.

Wu, Q.; Lin, Z.; Karlsson, B. F.; Huang, B.; and Lou, J. 2020b. UniTrans : Unifying Model Transfer and Data Transfer for Cross-Lingual Named Entity Recognition with Unlabeled Data. In Bessiere, C., ed., *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, 3926–3932. ijcai.org.

Wu, Q.; Lin, Z.; Wang, G.; Chen, H.; Karlsson, B. F.; Huang, B.; and Lin, C. 2020c. Enhanced Meta-Learning for Cross-Lingual Named Entity Recognition with Minimal Resources. In *Proc. of AAAI*.

Wu, S.; and Dredze, M. 2019. Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *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*, 833–844. Association for Computational Linguistics.

Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; Klingner, J.; Shah, A.; Johnson, M.; Liu, X.; Kaiser, L.; Gouws, S.; Kato, Y.; Kudo, T.; Kazawa, H.; Stevens, K.; Kurian, G.; Patil, N.; Wang, W.; Young, C.; Smith, J.; Riesa, J.; Rudnick, A.; Vinyals, O.; Corrado, G.; Hughes, M.; and Dean, J. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *CoRR*, abs/1609.08144.

Xie, J.; Yang, Z.; Neubig, G.; Smith, N. A.; and Carbonell, J. G. 2018. Neural Cross-lingual Named Entity Recognition with Minimal Resources. In Riloff, E.; Chiang, D.; Hockenmaier, J.; and Tsujii, J., eds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, 369–379. Association for Computational Linguistics.

Xu, Z.; Wei, P.; Zhang, W.; Liu, S.; Wang, L.; and Zheng, B. 2022. UKD: Debiasing Conversion Rate Estimation via Uncertainty-regularized Knowledge Distillation. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*.

Yang, Z.; Shou, L.; Gong, M.; Lin, W.; and Jiang, D. 2020. Model Compression with Two-stage Multi-teacher Knowledge Distillation for Web Question Answering System. In Caverlee, J.; Hu, X. B.; Lalmas, M.; and Wang, W., eds., *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020*, 690–698. ACM.

Zeng, J.; Jiang, Y.; Yin, Y.; Wang, X.; Lin, B.; and Cao, Y. 2022. DualNER: A Dual-Teaching framework for Zero-shot Cross-lingual Named Entity Recognition. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, 1837–1843. Association for Computational Linguistics.

Zheng, Z.; et al. 2021. Rectifying Pseudo Label Learning via Uncertainty Estimation for Domain Adaptive Semantic Segmentation. *Int. J. Comput. Vis.*

Zhou, B.; Cui, Q.; Wei, X.; and Chen, Z. 2020. BBN: Bilateral-Branch Network With Cumulative Learning for Long-Tailed Visual Recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 9716–9725. Computer Vision Foundation / IEEE.

Zhou, R.; Li, X.; Bing, L.; Cambria, E.; and Miao, C. 2023. Improving Self-training for Cross-lingual Named Entity Recognition with Contrastive and Prototype Learning. In Rogers, A.; Boyd-Graber, J. L.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, 4018–4031. Association for Computational Linguistics.