PDALN: Progressive Domain Adaptation over a Pre-trained Model for Low-Resource Cross-Domain Named Entity Recognition

Tao Zhang[1], **Congying Xia**[1], **Philip S. Yu**[1], **Zhiwei Liu**[1], **Shu Zhao**[2]
[1]University of Illinois at Chicago, Chicago, IL, USA;
[2]Anhui University, Hefei, Anhui, China
{tzhang90,cxia8,psyu,zliu213}@uic.edu; zhaoshuzs2002@hotmail.com

Abstract

Cross-domain Named Entity Recognition (NER) transfers the NER knowledge from high-resource domains to the low-resource target domain. Due to limited labeled resources and domain shift, cross-domain NER is a challenging task. To address these challenges, we propose a progressive domain adaptation Knowledge Distillation (KD) approach – PDALN. It achieves superior domain adaptability by employing three components: (1) Adaptive data augmentation techniques, which alleviate cross-domain gap and label sparsity simultaneously; (2) Multi-level Domain invariant features, derived from a multi-grained MMD (Maximum Mean Discrepancy) approach, to enable knowledge transfer across domains; (3) Advanced KD schema, which progressively enables powerful pre-trained language models to perform domain adaptation. Extensive experiments on four benchmarks show that PDALN can effectively adapt high-resource domains to low-resource target domains, even if they are diverse in terms and writing styles. Comparison with other baselines indicates the state-of-the-art performance of PDALN.

1 Introduction

Named Entity Recognition (NER) is typically framed as a sequence labeling task that targets to locate and classify named entities in text into predefined semantic types, such as *Person*, *Organization*, *Location*, etc. NER is a fundamental task in information extraction (Karatay and Karagoz, 2015) and text understanding (Krasnashchok and Jouili, 2018). The effectiveness of most existing NER models depends on sufficient labeled data, which is time-consuming and labor-intensive. Current research proposes cross-domain NER, which enables NER on the low-resource target domain by transferring knowledge from other high-resource source domains.

However, it is challenging to build a cross-domain NER component with high precision and recall, due to the domain shift problem (Ben-David et al., 2010). When casting the cross-domain NER as a transfer learning problem, most solutions (He and Sun, 2017; Yang et al., 2017; Aguilar et al., 2017; Lee et al., 2018; Liu et al., 2020b) require high-quality cross-domain features for knowledge transfer. Limited labeled data prohibit transfer learning from extracting informative features. Besides, it is hard to find a single training dataset covering all the required NER types. Even if words overlap across domains, their combination or usage is different from each other.

Domain adaptation (Sun et al., 2015) is widely studied to solve the domain shift issue. Existing approaches mainly introduce either word-level or discourse-level domain adaptations to enable cross-domain NER. To mitigate the word-level discrepancy, previous endeavors propose distributed word embedding (Kulkarni et al., 2016), label-aware maximum mean discrepancy estimation (Wang et al., 2018), and projecting learning (Lin and Lu, 2018). As to the discourse-level discrepancy, existing approaches introduce multi-level adaptation layers (Lin and Lu, 2018), tensor decomposition (Jia et al., 2019), and multi-task learning with external information (Liu et al., 2020b; Aguilar et al., 2017). However, those methods require sufficient labeled data, which hinders their performances under low-resource scenarios. To tackle both label sparsity and domain shift problem, existing approaches (Liang et al., 2020; Simpson et al., 2020; Cao et al., 2020) exploit external resources to generate pseudo labels for the low-resource domain. Nevertheless, the less confident labels may deteriorate the robustness of models because of noise.

In this paper, we propose a progressive domain adaptation cross-domain NER model PDALN. It introduces a novel domain adaptation component, which is enhanced by a progressive KD framework.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 5441–5451

PDALN addresses both word- and discourse-level domain adaptation on two low-resource scenarios: unsupervised and semi-supervised cross-domain NER. We first augment mix-domain training data by cross-domain anchor pairs, which alleviates the sparsity of annotated target domain. Next, we enable knowledge transfer across domains through domain invariant features learned from a multi-grained MMD adaptation metric. Additionally, we fuse contrastive learning (Hadsell et al., 2006) with a pre-trained model to extract robust features. Finally, instead of directly fine-tuning the model on the augmented adaptive data under the MMD-based metric, we integrate the cross-domain NER model into a sequential KD framework to learn a *low-capacity* student model. The *low-capacity* student can avoid over-fitting on limited annotated data because it progressively only cares about general cross-domain features retrieved by its sequential teachers to increase model confidence over domain invariant features. Our main contributions are summarized as follows:

- We propose a low-resource cross-domain NER model, PDALN, to transfer multi-level domain invariant knowledge from high-resource source domain to minimal-resource target domain without external retrieval auxiliary. Besides, PDALN can perform on both zero-resource and minimal-resource scenarios.
- We design an adaptive data augmentation for the low-resource domains. Moreover, we propose a multi-grained domain adaptation metric on the adaptive data to explore both word-level and discourse-level domain invariant features. We exploit a contrastive-learning fused pre-trained language model in a progressive self-training manner to enhance feature extraction.
- We conduct extensive experiments on four benchmarks to show our new state-of-the-art performance on two low-resource settings, including unsupervised and semi-supervised cross-domain NER.

2 Problem Definition

NER is typically formulated as a sequence labeling task. Based on the BIO schema [1], NER is to assign a sequence of labels $\mathcal{Y} = [y_1, ..., y_N]$ to a given sentence $\mathcal{X} = [x_1, ..., x_N]$ with N tokens. An entity is a span of tokens $\mathbf{e} = [x_i, ...x_j] (1 \le i \le j \le N)$ associated with an entity type.

[1] https://en.wikipedia.org/wiki/Inside-outside-beginning_(tagging)

In unsupervised NER domain adaptation, we are given source domain $\{(\mathcal{X}^s_m, \mathcal{Y}^s_m)\}^{N_s}_{m=1}$ with N_s labeled examples, and target domain data $\{\mathcal{X}^t_m\}^{N_t}_{m=1}$ with N_t unlabeled testing examples. The source domain and target domain are characterized by probability distributions P_s and P_t, respectively. We aim to construct a model which can learn transferable features to bridge the cross-domain discrepancy, and build a classifier $\mathcal{F} = f(\mathcal{X}; \theta)$ which can minimize target prediction error using source supervision. For low-resource cross-domain NER, it is a semi-supervised adaptation where the target has a few labeled examples. We denote the source domain data $\mathbf{D}^s = \{(\mathcal{X}^s_m, \mathcal{Y}^s_m)\}^{N_s}_{m=1}$, unannotated target data $\mathbf{D}^{t_u} = \{\mathcal{X}^{t_u}_i\}^{N_u}_{i=1}$, and annotated target data $\mathbf{D}^{t_a} = \{(\mathcal{X}^{t_a}_j, \mathcal{Y}^{t_a}_j)\}^{N_a}_{j=1}$. $\mathbf{D}^t = \mathbf{D}^{t_u} \cup \mathbf{D}^{t_a}$ is the total target data.

3 Preliminary

3.1 Base Model

To obtain expressive sentence features, we adopt a pre-trained language model (e.g. BERT(Devlin et al., 2018)) to encode the sentence $\mathcal{X} = [x_{\text{CLS}}, x_1, ..., x_N, x_{\text{SEP}}]$ (after padding tokens in BERT) into sentence representation $\mathbf{h} = [h_{\text{CLS}}, h_1, ..., h_N, h_{\text{SEP}}]$. The task objective is denote as CRF loss, where $\mathcal{L}_{\text{crf}} = \log p(\mathcal{Y}|\mathcal{X})$.

$$p(\mathcal{Y}|\mathcal{X}) = \frac{1}{\mathbf{Z}} \prod_{i=1}^{N} \phi_n(y_i|h_i, \mathbf{V}) \prod_{i=1}^{N-1} \phi_e(y_{i,i+1}|\mathbf{A}), \tag{1}$$

$$\mathcal{L}_{\text{crf}} = \sum_{i=1}^{N} \phi_n(y_i|h_i, \mathbf{V}) + \sum_{i=1}^{N-1} \mathbf{A}_{y_i, y_{i+1}} + \log \mathbf{Z}, \tag{2}$$

where $\log \phi_n(y_i = j|h_i, \mathbf{V}) = \exp(\mathbf{V}_j^T h_i)$, h_i is the encoded contextualized word vector, $\mathbf{V}$ is the weight matrix. $\mathbf{A}$ is the parameter for the transition matrix ϕ_e. $\mathbf{Z}(\cdot)$ is the normalization constant.

3.2 Maximum Mean Discrepancy (MMD) Measurement

The MMD is defined in particular function spaces $\mathcal{H}_k$ that measures the difference in cross domain distributions (P_s, P_t). $\mathcal{H}_k$ is the Reproducing Kernel Hilbert Space (RKHS) endowed with a characteristic kernel k. The squared formulation of MMD, $d^2_k(P_s, P_t)$, is defined as

$$d^2_k(P_s, P_t) = \|\mathbf{E}_{P_s}[\varphi(\mathbf{D}^s)] - \mathbf{E}_{P_t}[\varphi(\mathbf{D}^t)]\|^2_{\mathcal{H}_k}, \tag{3}$$

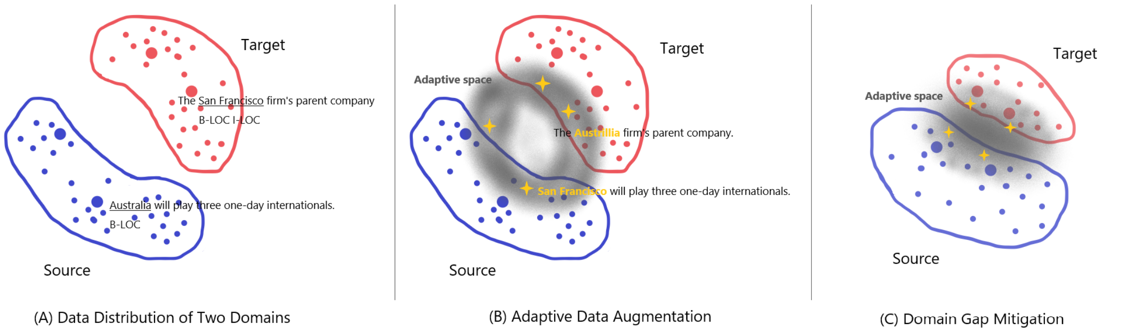

Figure 1: Toy illustration of the method. (A) There are two distributions of sentence embeddings. Data points in red represent the source dataset, and those in blue are the target. The oversized dots are the samples selected from each domain to construct the adaptive data. (B) The adaptive data is the yellow stars that form the adaptive space in gray. Each yellow star corresponds to an oversized dot near it. The adaptive data usually share the same sentence feature but perform cross-domain word replacement, like switched words in yellow. (C) Finally, we fine-tune the pre-trained model by adaptive data and MMD-based domain invariant features. In effect, the adaptive space works to guide the model to explore the target domain space as much as possible. The MMD-based domain adaptation approach gathers data points with similar sentence features. The domain-shared knowledge is the domain invariant features learned from these gathering points nearby the bridge.

where $\varphi : \mathcal{X} \to \mathcal{H}_k$. The most important property is that $P_s = P_t$ iff $d_k^2(P_s, P_t) = 0$. The characteristic kernel associated with the feature map φ and Gaussian Kernel $k(\mathbf{D}^s, \mathbf{D}^t)$.

To calculate MMD loss in cross-domain NER, we first compute the squared formulation of MMD between the BERT representations of source/target samples:

$$d_k^2(\mathbf{H}^s, \mathbf{H}^t) = \frac{1}{(N^s)^2} \sum_{i,j=1}^{N^s} k(h_i^s, h_j^s) + \frac{1}{(N^t)^2} \sum_{i,j=1}^{N^t} k(h_i^t, h_j^t) - \frac{2}{N^s N^t} \sum_{i,j=1}^{N^s N^t} k(h_i^s, h_j^t), \tag{4}$$

where $\mathbf{H}^s$ and $\mathbf{H}^t$ are sets of the BERT embeddings h^s and h^t with corresponding number N^s and N^t.

4 The Proposed Model

In this section, we present the structure of the proposed model. We first introduce domain adaptation components. On the one hand, we design an adaptive data augmentation to tackle the label sparsity issue. On the other hand, we introduce a multi-grained MMD metric on the augmented adaptive data to extract domain invariant features. There is an intuitive illustration in Figure 1 to show how our domain adaption approach mitigates the domain shifting. Besides, we exploit the power of the pre-trained model to capture expressive data features. We integrate a sequential self-training strategy to progressively and effectively perform our domain adaption components, as shown in Figure 2. We describe the details of cross-domain adaptation in Section 4.1 and progressive self-training for low-resource domain adaptation in Section 4.2.

4.1 Cross-domain Adaptation

When labels are insufficient in the target domain, most cross-domain NER models are vulnerable to over-fitting, thus yielding unsatisfactory performance. Therefore, we augment mix-domain data by *Cross-Domain Anchor* pairs. Those augmented data is defined as **adaptive data**, which can alleviate the data insufficiency problem. Our adaptive data is designed to simultaneously mitigate the domain gaps on both word-level and discourse-level. Those adaptive data form an adaptive space, as shown in Figure 1, which bridge two domains for cross-domain knowledge transferring.

4.1.1 Adaptive Data Augmentation

We first give the definition of *Cross-Domain Anchor*. An entity in source domain is denoted by $\mathbf{e}^s$ whose labels are $[y_{i^s}^s, ...y_{j^s}^s]$. A target entity is $\mathbf{e}^t$ whose labels are $[y_{i^t}^t, ...y_{j^t}^t]$. *Cross-Domain Anchor* pairs are $\mathcal{M}_{Anchor} = \{(\mathbf{e}^s, \mathbf{e}^t), y_{i^s}^s = y_{i^t}^t\}$. The cross-domain anchor is a relationship between two entities from different domains. $y_{i^s}^s = y_{i^t}^t$ denotes two entities belong to same entity type when their first label is the same. Intuitively, the anchor pairs address the cross-domain word discrepancy by sharing words per NER type cross domains.

Then, we use the cross-domain anchor pairs $\mathcal{M}_{Anchor}$ to create adaptive data $\mathbf{D}^{aug}$. Suppose we have $\mathbf{e}^p$, where $p \in \{s, t\}$ and $\mathbf{e}^p \in \mathcal{X}^p = [x_1^p, ..., x_{i^p}^p, ...x_{j^p}^p, ..., x_{|\mathcal{X}^p|}^p]$. Given an anchor pair

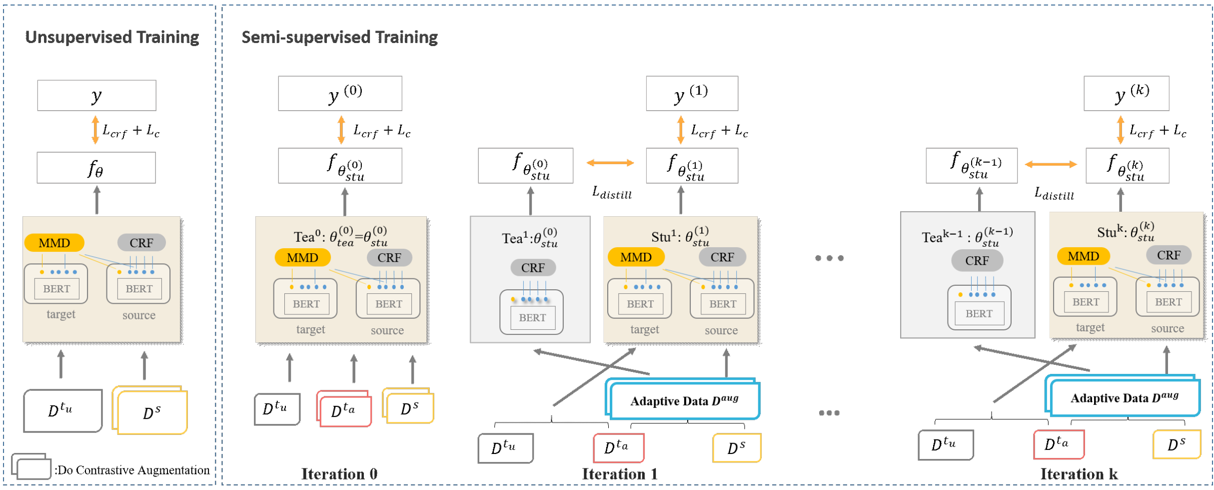

Figure 2: Unsupervised and Semi-supervised Training Schema. Semi-supervised Training mainly contains progressive Knowledge Distillation strategy with adaptive data augmentation in 4.2. The Contrastive augmentation denotes the data augmentation mentioned in 4.2.1.

$(\mathbf{e}^p, \mathbf{e}^q) \in \mathcal{M}_{Anchor}$, where $q \in \{s, t\}$ and $q \neq p$, we replace e^p in $\mathcal{X}^p$ with $\mathbf{e}^q$ as the augmented adaptive data $\mathcal{X}^{p\prime} = [x_1^p, ..., x_{iq}^q, ...x_{jq}^q, ..., x_{|\mathcal{X}^p|}^p]$. Finally, we obtain the adaptive data $\mathbf{D}^{aug} = \{\mathcal{X}^{p\prime}\}$.

Intuitively, the augmented adaptive sentences are regarded as mix-domain augmented data that share sentence pattern cross domains. Such semantically or syntactically similar sentences are the adaptive data, which can explore the unknown area in the target domain. The grey space, shown in Figure 1 (b), denotes the adaptive space, which is comprised of adaptive sentences like " The Australia firm's parent company." and "San Francisco will play three one-day internationals.". These two sentences are augmented by the *Cross-Domain Anchor* pair ("Australia", "San Francisco") which are both assigned to the label "LOC". When model fine-tuning is processed on the adaptive data, the model can benefit from the cross-domain features acquired from the adaptive space to improve model generalizability on the low-resource target domain.

4.1.2 Multi-grained MMD for Domain-invariant Features

As aforementioned, the adaptive space function is regarded as a cross-domain bridge. In this part, we seek to strengthen its domain adaptability and further aggregate the cross-domain features. We adapt domain-adaptation MMD (Long et al., 2015) to gather data points with similar word and sentence features, as shown in Figure 1 (c). Since MMD is to compute the norm of the difference between two domain means, MMD-based NER objective can thus learn both discriminative and domain invariant representations. We propose the multi-grained MMD method to simultaneously alleviate both the word-level and discourse-level discrepancy.

To distinguish the adaptation on word-level and discourse-level, we propose word MMD loss and sentence MMD loss, denoted by L^w_{MMD} and L^d_{MMD} respectively.

$$\mathcal{L}^d_{\text{MMD}}(\mathbf{D}^s, \mathbf{D}^t) = d_k^2(\mathbf{H}^s_{\text{CLS}}, \mathbf{H}^t_{\text{CLS}}), \quad (5)$$

where $\mathbf{H}_{\text{CLS}}$ is the set of CLS token embeddings. CLS is the sentence pool output for the token CLS in pre-trained language model. The word level MMD loss is denoted by the same label $y \in \text{label} = \{\text{B-X}, \text{I-X}, \text{O}\}$:

$$\mathcal{L}^w_{\text{MMD}}(\mathbf{D}^s, \mathbf{D}^t) = \sum_{y \in \text{label}} \mu_y d_k^2(\mathbf{H}_y(\mathbf{D}^s), \mathbf{H}_y(\mathbf{D}^t)), \quad (6)$$

where μ_y is the corresponding coefficient. $\mathbf{H}_y$ are the set of token embeddings with the label y.

Finally, the representations of a sentence and its tokens are the domain invariant features, which capture the cross-domain knowledge under the guide of $\mathcal{L}^d_{\text{MMD}}$ and $\mathcal{L}^w_{\text{MMD}}$. As shown in Figure 1 (c), the domain invariant features work to gather samples around the adaptive space to assist adaptation on both source and target domains.

4.2 Self-training for Low-Resource Domain Adaptation(DA)

4.2.1 Robust Feature Adaptation

Considering limited vocabulary and noise data samples on both source and target domains, we adopt contrastive learning (Hadsell et al., 2006; Ye et al., 2020; Chen et al., 2020; Liu et al., 2020a; Wu et al., 2020) to extract robust features through text augmentation like synonym replacement(Wu et al., 2020) and span deletion (Wei and Zou, 2019). We construct a distorted dataset $\mathbf{D}^c = \{(\mathcal{X}', \mathcal{Y}')\}$ over a given dataset $\mathbf{D} = \{(\mathcal{X}, \mathcal{Y})\}$.

$$\mathcal{L}_c = -\log \frac{\exp(\mathbf{z} \cdot \bar{\mathbf{z}})/\tau}{\sum_{\mathbf{z}_i \in \{\bar{\mathbf{z}}\} \cup \mathbf{Z}^{neg}} \exp(\mathbf{z} \cdot \mathbf{z}_i/\tau)}, \quad (7)$$

where $\mathbf{z} = \mathbf{W}^\top h_{CLS}$ is a mapping vector of a sentence $\mathcal{X}$. $\mathbf{W}$ is a trainable parameter. $\bar{\mathbf{z}} = \mathbf{W}^\top \bar{h}_{CLS}$ is the mapping vector of $\mathcal{X}'$ that is augmented by operating synonym replacement or span deletion on $\mathcal{X}$. $\mathbf{Z}^{neg}$ is constructed by other sentences in $\mathbf{D} \cup \mathbf{D}^c$ except $\mathcal{X}$ and $\mathcal{X}'$. τ is a temperature hyper-parameter.

4.2.2 Low-Resource Objectives

To address the low-resource scenarios, we consider both zero-resource and minimal-resource cross-domain NER training settings. We first perform the base model on both the source domain and target domain to seek the cross-domain bridge through multi-grained MMD adaptation. The **unsupervised cross-domain NER** loss is denoted as:

$$\mathcal{L}_{\text{unDA}} = \alpha' L^d_{\text{MMD}}(\mathbf{D}^s, \mathbf{D}^{t_u}) + \mathcal{L}_{\text{crf}} + \mathcal{L}_c. \quad (8)$$

which is free of any annotated target examples but still enables domain adaptation by $L^d_{\text{MMD}}(\mathbf{D}^s, \mathbf{D}^{t_u})$. The **semi-supervised cross-domain NER** objective is denoted as:

$$\begin{aligned} \mathcal{L}_{\text{semiDA}} = \alpha \cdot L^d_{\text{MMD}}(\mathbf{D}^s, \mathbf{D}^t) + \\ \beta \cdot \mathcal{L}^w_{\text{MMD}}(\mathbf{D}^s, \mathbf{D}^{t_a}) + \mathcal{L}_{\text{crf}} + \mathcal{L}_c, \end{aligned} \quad (9)$$

where α and β are the hyperparameters to balance the multi-grained MMD loss terms.

4.2.3 Progressive Joint KD and DA

We propose a progressive domain adaptation by integrating a sequential teacher-student framework to prevent the model from over-fitting on limited labeled data and augmented adaptive data. The intuition is that the student easily overlooks “problematic" examples but learns things that generalize well. Therefore, the KD framework enjoys the merits that it progressively improves the domain adaptation confidence over data.

The cross-domain NER loss over adaptive data is denoted as:

$$\begin{aligned} \mathcal{L}_{\text{semiDA}} = \alpha \cdot \mathcal{L}^d_{\text{MMD}}(\mathbf{D}^{aug}, \mathbf{D}^t) + \\ \beta \cdot \mathcal{L}^w_{\text{MMD}}(\mathbf{D}^{aug}, \mathbf{D}^{t_a}) + \mathcal{L}_{\text{crf}} + \mathcal{L}_c. \end{aligned} \quad (10)$$

In the progressive KD framework, we use $f_{\theta_{tea}}$ and $f_{\theta_{stu}}$ to denote teacher and student models, respectively. Suppose $f_{\hat{\theta}}$ is the base model learned by the objective in Equation 9, we initial the teacher model and the student model as: $\theta^{(0)}_{tea} = \theta^{(0)}_{stu} = \hat{\theta}$.

At t-th iteration, the student model loss is denoted as:

$$\begin{aligned} \mathcal{L}_{\text{distill}} = (1-\gamma) \cdot \mathcal{L}_{\text{semiDA}} + \\ \gamma \cdot \frac{1}{N} \sum_{n=1}^{N} -f_{\theta^{(t)}_{tea}, n}(\mathcal{X}) \log f_{\theta_{stu}, n}(\mathcal{X}), \end{aligned} \quad (11)$$

Where $\mathcal{X} \in \mathbf{D}^{aug}$, containing N entities. $f_{\cdot, n}(\mathcal{X})$ means the output of entity n.

The updated model is $\hat{\theta}^{(t)}_{stu} = \arg\min_{\theta_{stu}} \mathcal{L}_{\text{distill}}$. Finally, we update the teacher-student model for the $(t+1)$-th iteration by: $\theta^{(t+1)}_{tea} = \theta^{(t+1)}_{stu} = \hat{\theta}^{(t)}_{stu}$.

5 Experiments

In this section, we evaluate PDALN and other baselines on four public benchmarks. We conduct two groups of comparison experiments for unsupervised and semi-supervised cross-domain NER separately. We also conduct further ablation studies and hyperparameter studies to validate the efficacy of the domain adaptation approaches.

5.1 Datasets

The datasets in the source and target domains contain the same four types of entities, namely, PER (person), LOC (location), ORG (organization), and MISC (miscellaneous). Our source domain is CoNLL-2003 English NER data (Sang and De Meulder, 2003) containing 15.0K/3.5K/3.7K samples for the training/validation/test sets. We consider four target doamins: (1) **SciTech** (Jia et al., 2019) News with 2K sentences; (2) **WNUT 2016** (Strauss et al., 2016) containing 2400 tweets (comprising 34k tokens) with 10 entity types; (3) **Webpage** (Ratinov and Roth, 2009) comprising 20 webpages and 783 entities with documents varying from personal, academic, to computer science conference; (4) **Wikigold** (Balasuriya et al., 2009), a set of Wikipedia articles with 40k tokens. To make the datasets consistent, we convert 10 types in WNUT 2016 NER into four CoNLL03 entity types.

5.2 Baselines

We compare PDALN with the following state-of-the-art cross-domain NER models:

BiLSTM+CRF (Lample et al., 2016) harnesses character-level Bi-LSTMs to capture the morphological and orthographic features and word-level

Bi-LSTMs to integrate the sentence grammar feature. At last, the model stacks a CRF layer to predict the labels considering their dependencies.

BERT+CRF replaces traditional BiLSTM component with the powerful pre-trained language model BERT to obtain more informative and contextual enhanced word representations.

La-DTL (Simpson et al., 2020) proposes the label-aware MMD metric learning to mitigate the word distribution discrepancy.

DATNet (Zhou et al., 2019) proposes a generalized resource-adversarial discriminator to capture the share feature space across different domains. Then the domain shared space guides the target domain prediction on NER task.

JIA2019 (Jia et al., 2019) combines language model and NER task to construct multi-task learning structure, and then exploits tensor decomposition to learn the task embedding for cross-domain NER prediction over such task embeddings.

Multi-Cell (Jia and Zhang, 2020) proposes a multi-cell compositional LSTM structure for cross-domain NER under the multi-task learning strategy.

In addition, we compare the evaluation of two variants of PDALN. We replace the sequential KD framework in the self-training stage with **MT** and **VAT**, Mean Teacher strategy (Tarvainen and Valpola, 2017) and Virtual Adversarial Training (Miyato et al., 2018), respectively.

5.3 Training and Implementation Details

We adopt the Adam optimization algorithm with a decreasing learning rate of 0.00005. We utilize the pre-trained BERT (BERT-base, cased) where the number of transformer blocks is 12, the hidden layer size is 768, and the number of self-attention heads is 12. Each batch contains 32 examples, with a maximum encoding length of 128. The coefficient μ_y in Equation 6 is 0.25. The temperature hyper-parameter $\tau = 0.05$. We choose 100 labeled target examples and 500 labeled source examples to augment adaptive data in the size of 1400 (100*4+500*2). Each target example operates 4 times anchor word replacement into 4 augmented sentences, while 2 replacements for each source example. Particularly, we take 10/100/240 as target/source/adaptation examples in the Webpage dataset, due to its insufficient target examples.

5.4 Results and Discussion

Domain Adaptation on Unsupervised NER The unsupervised NER follows the zero-shot paradigm, preventing model training from any testing labeled data. Compared with other unsupervised NER baselines, PDALN achieves the best F-1 on all benchmarks, even suffering failure on the precision scores. As the unsupervised NER results are shown in Table 1, PDALN and BERT+CRF both attain competitive performance on the recall scores, which benefits from the powerful contrastive-learning fused pre-trained language model. But for WNUT2016 and Wikigold, PDALN surpassing BERT+CRF shows the benefits from sentence-level domain adaptation through $\mathcal{L}^d_{\text{MMD}}$ and robust feature extraction through $\mathcal{L}_c$.

Evaluation on Semi-supervised NER As shown in Table 1, most of the baselines cannot achieve decent performance gain by taking in limited annotated resources. But PDALN outperforms the best public baseline range from 1.5% to 4.0% on all benchmarks. Most of the existing approaches adopt BiLSTM as their fundamental component to aggregate input information. Unfortunately, BiLSTM cannot capture expressive sentence features due to its intrinsic shortcomings, vanishing or exploding gradient problems. Therefore, these approaches are prone to increasing false-positive predictions and suffer unsatisfied recall scores. Even though pre-trained language models can attain stunning recall scores, their precision scores dramatically fall behind the baselines. The main reason is that such a powerful pre-trained model is prone to over-fitting on small annotated data. Compared with BERT+CRF, our promising precision gain and increasing recall scores show that our model can make a successful tradeoff between the precisions and recalls. Besides, we compare with two variants (w/ **MT** and w/ **VAT**) of our model with different KD strategies, like Mean Teacher and Virtual Adversarial Training. Their performance is close to ours on the high-quality labeled data, SciTech. But their performance on the other domains shows they are vulnerable to the noise and easily overfit on limited annotated samples. PDALN overcomes that well by the progressive domain adaptation with moderate knowledge distillation from the teachers.

Ablation Study We conduct ablation studies that quantify the contribution of each adaptation component in PDALN. As Table 2 shows, the removal of augmented data causes dramatic performance decreases on all four benchmarks. That indicates adaptive data augmentation plays the most vital role in the low-resource cross-domain NER task.

Baselines	SciTech F1 (Pre/Rec)	WNUT 2016 F1 (Pre/Rec)	Webpage F1 (Pre/Rec)	Wikigold F1 (Pre/Rec)
Un-supervised NER				
BiLSTM+CRF	67.01 (73.53 / 61.56)	24.76 (47.01 / 16.81)	43.34 (58.05 / 34.59)	42.92 (47.55 / 39.11)
BERT+CRF	74.26 (68.57 / 80.97)	44.37 (34.39 / 62.50)	55.94 (58.29 / 53.78)	47.99 (44.13 / 52.61)
JIA2019	73.58 (74.28 / 72.91)	38.16 (47.26 / 32.00)	46.96 (51.61 / 43.08)	45.18 (**48.68** / 42.15)
Multi-Cell	75.01 (**77.10** / 73.03)	41.07 (**47.96** / 35.91)	48.62 (58.27 / 41.72)	46.04 (47.94 / 44.29)
PDALN	**75.80** (70.21 / **82.36**)	**46.12** (36.00 / **64.19**)	**56.93** (**58.36** / **55.57**)	**49.73** (45.39 / **54.99**)
	75.56 ± 0.41	**45.93 ± 0.35**	**57.25 ± 0.31**	**49.55 ± 0.44**
Semi-supervised NER				
BiLSTM+CRF	67.83 (72.95 / 63.39)	27.61 (48.56 / 19.29)	44.46 (58.88 / 35.72)	44.65 (48.40 / 41.44)
BERT+CRF	75.29 (70.23 / 81.14)	45.31 (35.15 / 63.77)	56.78 (58.71 / 54.99)	48.45 (44.02 / 53.88)
La-DTL	73.30 (74.10 / 72.52)	35.97 (37.22 / 34.78)	51.39 (48.81 / 54.23)	47.74 (46.70 / 48.83)
DATNet	69.22 (65.14 / 73.84)	32.67 (35.56 / 30.21)	47.71 (47.53 / 47.90)	37.92 (36.90 / 39.00)
JIA2019	74.65 (75.65 / 74.01)	39.14 (**48.89** / 32.64)	47.39 (52.19 / 43.40)	45.77 (**49.24** / 42.76)
Multi-Cell	75.89 (**76.89** / 74.92)	42.19 (47.83 / 37.74)	49.45 (59.94 / 42.09)	46.45 (45.29 / 47.67)
PDALN w/ **MT**	77.80 (72.93 / 83.38)	46.45 (36.11 / 65.10)	57.43 (58.69 / 56.24)	51.74 (47.39 / 56.97)
PDALN w/ **VAT**	77.33 (73.10 / 82.08)	46.68 (36.46 / 64.87)	57.14 (58.26 / 56.07)	51.08 (46.88 / 56.13)
PDALN	**78.23** (73.58 / **83.51**)	**48.22** (37.78 / **66.66**)	**58.56** (**59.99/ 57.20**)	**53.06** (48.77 / **58.19**)
	77.31 ± 0.59	**47.63 ± 0.61**	**58.25 ± 0.34**	**52.48 ± 0.49**

Table 1: Model evaluation on four benchmarks: F1 Score (Precision/Recall) (in %). PDALN's performance contains two parts: the best score of five runs in the top, average F-1 score with deviation in the bottom.

Baselines	SciTech F1 (Pre/Rec)	WNUT 2016 F1 (Pre/Rec)	Webpage F1 (Pre/Rec)	Wikigold F1 (Pre/Rec)
PDALN	78.23 (73.58 / 83.51)	48.22 (37.78 / 66.66)	58.56 (59.99 / 57.20)	53.06 (48.77 / 58.19)
w/o $\mathcal{L}_c$	-0.56 (-0.56 / -0.57)	-0.72 (-0.64 / -0.81)	-0.27 (-0.35 / -0.21)	-1.20 (-1.22 / -1.18)
w/o $\mathcal{L}^d_{\text{MMD}}$	-1.25 (-1.42 / -1.02)	-1.21 (-0.93 / -1.77)	-0.80 (-0.64 / -0.95)	-1.46 (-1.33 / -1.64)
w/o $\mathcal{L}^w_{\text{MMD}}$	-1.59 (-1.91 / -1.14)	-1.39 (-1.23 / -1.53)	-0.98 (-0.81 / -1.14)	-1.51 (-1.49 / -1.54)
w/o $\mathcal{L}_{\text{distill}}$	-1.94 (**-2.57** / -1.10)	-1.68 (**-1.60** / -1.46)	-1.38 (-1.30 / -1.46)	-1.56 (**-1.68** / -1.37)
w/o $\mathbf{D}^{aug}$	**-1.96** (-2.36 / **-1.44**)	**-1.79** (-1.56 / **-2.02**)	**-2.17** (**-2.16** / **-2.19**)	**-1.64** (-1.51 / **-1.81**)

Table 2: Ablation study. All the results are percentages. The minus number means performance drop after removing or replacing the methods. (w/o $\mathcal{L}^c$) means the removal of robust feature extraction by Equation 7. (w/o $\mathcal{L}^d_{\text{MMD}}$) and (w/o $\mathcal{L}^w_{\text{MMD}}$) mean the removal of the sentence-level MMD loss and word-level MMD loss in Equation 9, respectively. (w/o $\mathcal{L}_{\text{distill}}$) means the removal of progressive knowledge distillation loss in Equation 11.

PDALN	PER F1 (Pre/Rec)	LOC F1 (Pre/Rec)	ORG F1 (Pre/Rec)	MISC F1 (Pre/Rec)
SciTech	91.42 (92.25 / 90.61)	71.36 (64.21 / 80.31)	68.76 (60.56 / 79.54)	48.81 (45.12 / 53.17)
WNUT	86.27 (84.51 / 88.12)	48.57 (44.33 / 53.71)	46.81 (41.11 / 54.36)	27.90 (21.48 / 39.79)
Webpage	80.34 (78.45 / 82.34)	45.75 (41.50 / 50.97)	45.60 (43.12 / 48.39)	42.48 (39.61 / 45.81)
Wikigold	84.95 (85.69 / 84.24)	43.36 (39.45 / 48.14)	42.12 (35.94 / 50.89)	37.53 (32.11 / 45.16)

Table 3: PDALN's performance on each entity type.

Our progressive KD framework shows its importance on precision gain as w/o $\mathcal{L}_{\text{distill}}$ causes the worst precision drop. Our multi-grained MMD (either the sentence-level or word-level MMD) methods play noteworthy contributions for cross-domain NER adaptation as well, as their removals also cause serious performance loss. The removal of $\mathcal{L}_c$ attests the robust feature extraction works well

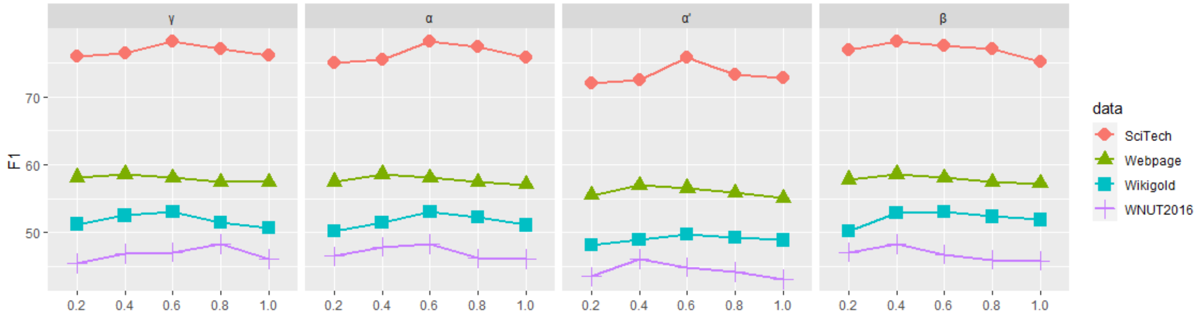

Figure 3: Hyper-parameter (γ, α, α',β) study on four benchmarks.

when the annotated data (e.g. Wikigold) are not very precise.

Evaluation on Entity Type We provide PDALN's performance on each entity type in Table 3. The performance on PER is more stable than the other three, LOC, ORG, and MISC. The other three types mainly exhibits the difference of entity distributions and topics between the four benchmarks. Besides, the long-chunk entities (e.g. "New[B-ORG] Jersey[I-ORG] Department[I-ORG] of[I-ORG] Public[I-ORG] Safety[I-ORG].") easily cause mixed or incomplete labeling, which degrades the evaluation scores, especially for dataset Webpage and Wikigold even on all four types. A large group of MISC entities in WNUT also impede the model performance.

Parameter Study We investigate the effects of each adaptation component by its coefficients. When the coefficient is under evaluation, the others are assigned with default values. We tune the best score of coefficients on each domain, as shown in Figure 3. Besides, we conduct experiments to investigate model behavior over the different sizes of augmented data. We fix the number of target labeled examples but provide a range of source samples. From Table 4, performance on all datasets increases fast at the first three augmented groups. But the increasing speed cools down at the last group. Hence, adaptive data can benefit domain adaptation but be careful to avoid overusing.

Error Analysis Thanks to the power of being pre-trained on large corpora, BERT is easy to assign specific labels to a roughly-labeled sentence. For example, the test example is "Chinese[B-MISC] President[O] Xi[B-PER] Jinping[I-PER] at[O] the[O] G-20[B-MISC] summit[O] in[O] Argentina[O].". Even if the ground truth for the entity "Argentina" is [O], BERT correctly assigns it with [B-LOC], but not by BiLSTM-CRF. Therefore, the pre-trained model group achieves much higher recalls but lower precisions. Apart from data annotation errors in datasets, occasionally over labelling occurs in PDALN, like the test sentence, "Jared[B-PER] put[O] together[O] this[O] thing[O] called[O] Environmentor[O]". PDALN prefers to label the last entity with "Environmentor[B-ORG]".

Method	SciTech	WNUT2016	Wikigold
Ours-MMD^d	75.80	46.12	49.73
Ours(DA[1:1])	76.29	46.55	50.16
Ours(DA[1:3])	77.38	47.74	51.81
Ours(DA[1:5])	78.23	48.22	53.06
Ours(DA[1:7])	78.38	48.59	53.71

Table 4: Evaluation on augmented data size (F1 score in %). DA[1:X] means data augmentation uses the ratio 1:X over the target and source samples. Then anchor word replacement operates on them to make the ratio be (1*4):(X*2). 4 / 2 means generated examples of each sample after the replacement.

6 Related Work

Recently, label sparsity has achieved great success in many research frontiers (Liu et al., 2019, 2021; Zhang et al., 2020c; Xia et al., 2018, 2020, 2021; Zhang et al., 2020a,b). One of the widely adopted strategies is a cross-domain transfer which mainly deals with the domain shift problem. The causes for domain shift in NER are mainly twofold including the discrepancies of word distributions and sentence patterns between source and target domain.

On the one hand, word distributions are not compatible between different domain datasets. Therefore, existing works equip the model with diverse domain adaptation components to alleviate domain shift. Kulkarni et al. (2016) propose distributed word embedding methods to leverage domain-specific knowledge to boost their cross-domain NER performance. Wang et al. (2018) introduce a label-aware mechanism into maximum mean discrepancy (MMD) to explicitly reduce domain shift between the same labels across domains in medical data. Lin and Lu (2018) employ projecting learning to obtain a transfer matrix that maps target domain words into the word space of the source domain.

On the other hand, diverse sentence patterns are usually caused by various factors, like written styles, publication categories, data quality, etc. The solutions for mitigating the discourse-level discrepancy mainly include multi-level adaptation layers (Lin and Lu, 2018), tensor decomposition (Jia et al., 2019) and multi-task learning with external information (Liu et al., 2020b; Aguilar et al., 2017). As we mentioned before, Lin and Lu (2018) construct the word adaptation component in their model. Besides, they construct another sentence-adaptation layer, which takes in the adapted word embedding to extract another adaptation sentence feature. Jia et al. (2019) use multi-task learning and tensor decomposition to extract latent factors. Through latent factors, knowledge can be transferred across source and target domains. Liu et al. (2020b) employ NER label experts to guide model learning between domains. The label-aware guidance layer is key to enable domain adaptation. Jia and Zhang (2020) a multi-cell compositional LSTM structure for cross-domain NER under the multi-task learning strategy. Besides, those (Liang et al., 2020; Simpson et al., 2020; Cao et al., 2020) exploit external resources to generate pseudo labels for the low-resource domain with the assistance of a pre-trained language model.

However, those methods either lack the capability to capture expressive text features for the adaptation or require sufficient labeled target data, which impedes their performances under both zero-resource and minimal-resource scenarios. For the pre-trained model assisted approaches mainly rely on external knwledge bases which introduces too much noise.

7 Conclusion

In this paper, we propose a progressive adaptation knowledge distillation framework, including anchor-guided adaptive data to address data sparsity, multi-grained MMD to bridge the domain adaptation, and progressive KD to stably distill cross-domain knowledge. The results exhibit the model's superiority over the most state-of-the-arts.

8 Acknowledgments

We thank the reviewers for their valuable comments. This work is supported in part by NSF under grants III-1763325, III-1909323, III-2106758, and SaTC-1930941.

References

Gustavo Aguilar, Suraj Maharjan, Adrian Pastor López-Monroy, and Thamar Solorio. 2017. A multi-task approach for named entity recognition in social media data. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 148–153.

Dominic Balasuriya, Nicky Ringland, Joel Nothman, Tara Murphy, and James R Curran. 2009. Named entity recognition in wikipedia. In *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources (People's Web)*, pages 10–18.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine learning*, 79(1):151–175.

Yixin Cao, Zikun Hu, Tat Seng Chua, Zhiyuan Liu, and Heng Ji. 2020. Low-resource name tagging learned with weakly labeled data. In *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 261–270. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE.

Hangfeng He and Xu Sun. 2017. A unified model for cross-domain and semi-supervised named entity recognition in chinese social media. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Chen Jia, Xiaobo Liang, and Yue Zhang. 2019. Cross-domain ner using cross-domain language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2464–2474.

Chen Jia and Yue Zhang. 2020. Multi-cell compositional LSTM for NER domain adaptation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5906–5917. Association for Computational Linguistics.

Deniz Karatay and Pinar Karagoz. 2015. User interest modeling in twitter with named entity recognition. In *5th Workshop on Making Sense of Microposts*.

Katsiaryna Krasnashchok and Salim Jouili. 2018. Improving topic quality by promoting named entities in topic modeling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 247–253.

Vivek Kulkarni, Yashar Mehdad, and Troy Chevalier. 2016. Domain adaptation for named entity recognition in online media with word embeddings. *arXiv preprint arXiv:1612.00148*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*, pages 260–270.

Ji Young Lee, Franck Dernoncourt, and Peter Szolovits. 2018. Transfer learning for named-entity recognition with neural networks. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Chen Liang, Yue Yu, Haoming Jiang, Siawpeng Er, Ruijia Wang, Tuo Zhao, and Chao Zhang. 2020. Bond: Bert-assisted open-domain named entity recognition with distant supervision. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1054–1064.

Bill Yuchen Lin and Wei Lu. 2018. Neural adaptation layers for cross-domain named entity recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2012–2022.

Xiao Liu, Fanjin Zhang, Zhenyu Hou, Zhaoyu Wang, Li Mian, Jing Zhang, and Jie Tang. 2020a. Self-supervised learning: Generative or contrastive. *arXiv preprint arXiv:2006.08218*, 1(2).

Zhiwei Liu, Ziwei Fan, Yu Wang, and Philip S. Yu. 2021. Augmenting sequential recommendation with pseudo-prior items via reversely pre-training transformer. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Zhiwei Liu, Lei Zheng, Jiawei Zhang, Jiayu Han, and S Yu Philip. 2019. Jscn: Joint spectral convolutional network for cross domain recommendation. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 850–859. IEEE.

Zihan Liu, Genta Indra Winata, and Pascale Fung. 2020b. Zero-resource cross-domain named entity recognition. *arXiv preprint arXiv:2002.05923*.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. 2015. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 147–155.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Edwin Simpson, Jonas Pfeiffer, and Iryna Gurevych. 2020. Low resource sequence tagging with weak labels. In *AAAI*, pages 8862–8869.

Benjamin Strauss, Bethany Toma, Alan Ritter, Marie-Catherine De Marneffe, and Wei Xu. 2016. Results of the wnut16 named entity recognition shared task. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 138–144.

Shiliang Sun, Honglei Shi, and Yuanbin Wu. 2015. A survey of multi-source domain adaptation. *Information Fusion*, 24:84–92.

Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780*.

Zhenghui Wang, Yanru Qu, Liheng Chen, Jian Shen, Weinan Zhang, Shaodian Zhang, Yimei Gao, Gen Gu, Ken Chen, and Yong Yu. 2018. Label-aware double transfer learning for cross-specialty medical named entity recognition. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1–15.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. Clear: Contrastive learning for sentence representation. *arXiv preprint arXiv:2012.15466*.

Congying Xia, Caiming Xiong, S Yu Philip, and Richard Socher. 2020. Composed variational natural language generation for few-shot intents. In

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, pages 3379–3388.

Congying Xia, Wenpeng Yin, Yihao Feng, and S Yu Philip. 2021. Incremental few-shot text classification with multi-round new classes: Formulation, dataset and system. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1351–1360.

Congying Xia, Chenwei Zhang, Xiaohui Yan, Yi Chang, and Philip S Yu. 2018. Zero-shot user intent detection via capsule neural networks. *arXiv preprint arXiv:1809.00385*.

Zhilin Yang, Ruslan Salakhutdinov, and William W Cohen. 2017. Transfer learning for sequence tagging with hierarchical recurrent networks. *arXiv preprint arXiv:1703.06345*.

Hai Ye, Qingyu Tan, Ruidan He, Juntao Li, Hwee Tou Ng, and Lidong Bing. 2020. Feature adaptation of pre-trained language models across languages and domains with robust self-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7386–7399.

Jianguo Zhang, Kazuma Hashimoto, Wenhao Liu, Chien-Sheng Wu, Yao Wan, S Yu Philip, Richard Socher, and Caiming Xiong. 2020a. Discriminative nearest neighbor few-shot intent detection by transferring natural language inference. In *EMNLP*, pages 5064–5082.

Jianguo Zhang, Kazuma Hashimoto, Chien-Sheng Wu, Yao Wang, S Yu Philip, Richard Socher, and Caiming Xiong. 2020b. Find or classify? dual strategy for slot-value predictions on multi-domain dialog state tracking. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 154–167.

Tao Zhang, Congying Xia, Chun-Ta Lu, and S Yu Philip. 2020c. Mzet: Memory augmented zero-shot fine-grained named entity typing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 77–87.

Joey Tianyi Zhou, Hao Zhang, Di Jin, Hongyuan Zhu, Meng Fang, Rick Siow Mong Goh, and Kenneth Kwok. 2019. Dual adversarial neural transfer for low-resource named entity recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3461–3471.