Domain-Invariant Feature Progressive Distillation with Adversarial Adaptive Augmentation for Low-Resource Cross-Domain NER

TAO ZHANG and CONGYING XIA, University of Illinois at Chicago, USA
ZHIWEI LIU, Salesforce AI Research, USA
SHU ZHAO, Anhui University, China
HAO PENG, Beihang University, China
PHILIP YU, University of Illinois at Chicago, USA

Considering the expensive annotation in **Named Entity Recognition (NER)**, Cross-domain NER enables NER in low-resource target domains with few or without labeled data, by transferring the knowledge of high-resource domains. However, the discrepancy between different domains causes the domain shift problem and hampers the performance of cross-domain NER in low-resource scenarios. In this article, we first propose an adversarial adaptive augmentation, where we integrate the adversarial strategy into a multi-task learner to augment and qualify domain adaptive data. We extract domain-invariant features of the adaptive data to bridge the cross-domain gap and alleviate the label-sparsity problem simultaneously. Therefore, another important component in this article is the progressive domain-invariant feature distillation framework. A multi-grained **MMD (Maximum Mean Discrepancy)** approach in the framework to extract the multi-level domain invariant features and enable knowledge transfer across domains through the adversarial adaptive data. Advanced **Knowledge Distillation (KD)** schema processes progressively domain adaptation through the powerful pre-trained language models and multi-level domain invariant features. Extensive comparative experiments over four English and two Chinese benchmarks show the importance of adversarial augmentation and effective adaptation from high-resource domains to low-resource target domains. Comparison with two vanilla and four latest baselines indicates the state-of-the-art performance and superiority confronted with both zero-resource and minimal-resource scenarios.

CCS Concepts: • **Information systems → Information retrieval**; **Retrieval tasks and goals**; **Information extraction;**

A preliminary version [56] of this article appeared in the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Pages 5441-5451 (EMNLP'21).
This work is supported by National Key R&D Program of China through grant 2021YFB1714800, NSF under grants III-1763325, III-1909323, III-2106758, SaTC-1930941, S&T Program of Hebei through grants 20310101D and 21340301D, NSFC through grant 62002007, Beijing Natural Science Foundation through grant 4222030, and the Fundamental Research Funds for the Central Universities.
Authors' addresses: T. Zhang, C. Xia, Z. Liu, and P. S. Yu Department of Computer Science, University of Illinois at Chicago, 851 South Morgan Street, Chicago, IL 60607-7053, USA; email: {tzhang, cxia8, liu213, psyu}@uic.edu; S. Zhao, School of Computer Science and Technology, Anhui University, No. 111 Jiulong Road, Hefei, Anhui, 230601, China; email: zhaoshuzs2002@hotmail.com; H. Peng (corresponding author), Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, No. 37 Xue Yuan Road, Haidian District, Beijing, 100191, China; email: penghao@buaa.edu.cn.

2375-4699/2023/04-ART76 $15.00
https://doi.org/10.1145/3570502

Additional Key Words and Phrases: NER, adversarial augmentation, cross-domain, domain adaptation, low-resource, knowledge distillation

ACM Reference format:
Tao Zhang, Congying Xia, Zhiwei Liu, Shu Zhao, Hao Peng, and Philip Yu. 2023. Domain-Invariant Feature Progressive Distillation with Adversarial Adaptive Augmentation for Low-Resource Cross-Domain NER. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 22, 3, Article 76 (April 2023), 21 pages.
https://doi.org/10.1145/3570502

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*, and so on. As one of the fundamental tasks in **natural language processing (NLP)**, NER has been explored for decades to serve other advanced downstream tasks, like information extraction [13, 50], text understanding [14, 33], and so on. Most existing NER models are trained in a supervised manner, which depends on sufficient labeled data. As we know, well-annotated data has expensive accessibility and is not always feasible in the real world. Although distant supervision [2, 28] can bring alleviation and provide weak annotation, data denoising is another time-consuming and labor-intensive issue caused by distant supervision techniques. Current research [25] prefers cross-domain NER for learnt NER knowledge transferred from the high-resource source domains to low-resource target domains. Like the transfer learning paradigm [30], cross-domain NER trains a NER learner on a well-labeled source domain but enables it to perform on a target domain without enough labeled examples.

However, cross-domain NER is faced with the challenge of decent solutions on the domain shift problem [3, 7, 36]. Most solutions [1, 10, 17, 24, 32, 51] rely on high-quality cross-domain features during cross-domain knowledge transfer. For low-resource scenarios, insufficient labeled data usually show undesired performance in cross-domain features extraction. Another challenge is the high quality of source domain data. The source domain should be enough assorted and resourceful to make a single training dataset cover all the required NER types. The semantic ambiguous words may appear in multiple domains and assigned different NER types. Because their combination or usage is different across different domains, domain adaptation [27, 43] is an intensively-explored solution among recent researches. Existing approaches follow two categories, either word-level or discourse-level, in domain adaptations to enable cross-domain NER. When mitigating the word-level discrepancy, previous endeavors introduce distributed word embedding [15], label-aware maximum mean discrepancy estimation [45], and projecting learning [20]. When alleviating the discourse-level discrepancy, researchers propose multi-level adaptation layers [20], tensor decomposition [11], and multi-task learning with external information [1, 24]. However, promising results in these methods all rely on valuable cross-domain features derived from sufficient labeled data, which hinders their performance for low-resource scenarios. To tackle both insufficient labeled data and domain shift problems, recent approaches [4, 19, 41, 50] leverage external resources to generate pseudo labels for the compensation. Nevertheless, the less confident labels not only deteriorate the robustness of models due to noise but also consume additional computational resources.

To address above mentioned limitations of existing methods, we propose a domain-invariant feature progressive distillation framework, PDALN†. We propose both word- and discourse-level domain adaptation on two low-resource scenarios: unsupervised and semi-supervised cross-domain NER. PDALN† works on both insufficient labeled data and domain shift simultaneously through adversarial adaptive data augmentation and adaptive feature progressive distillation. Components in

PDALN† work one after another to uncover domain-invariant features. To alleviate the sparsity of annotated target domain, we augment mix-domain training data with cross-domain anchor pairs. Then we design to qualify and select the augmented data. We adopt adversarial training with a domain discriminator to explore domain-invariant space across the source and target domains. Such a discriminator ranks the augmented data and selects the most adaptive set located in or closed to the domain-invariant space. Next, we extract multi-level domain-invariant features through a multi-grained **Maximum Mean Discrepancy (MMD)** adaptation metric to enable knowledge transfer across domains. Besides, we boost the robustness of a pre-trained model through examples from contrastive learning [9, 21, 34, 39]. Finally, a sequential teacher-student **Knowledge Distillation (KD)** framework works to progressively perform domain adaption increasing model robustness and its confidence over domain invariant features.

The source code and datasets are publicly available at https://github.com/towermxt/cross_doamin_ner. Our main contributions are summarized as follows:

- We propose an adversarial augmentation-assisted cross-domain NER model, PDALN†, which is mainly used for both zero-resource and minimal-resource scenarios. PDALN† can transfer multi-level domain invariant knowledge from high-resource source domain to low-resource target domain without external retrieval auxiliary.
- We introduce an adversarial adaptive data augmentation to qualify and refine the augmentation. Moreover, we learn the word-level and discourse-level domain invariant features by a multi-grained domain adaptation metric on the refined adaptive data. We propose a self-trainer to progressively boost invariant feature extraction.
- We conduct extensive experiments on four English benchmarks and two Chinese benchmarks to show our new state-of-the-art performance in two low-resource settings, including unsupervised and semi-supervised cross-domain NER.

We expand on our preliminary work, PDALN [56], by extending a multi-task leaner with the adversarial strategy to denoise the adaptive data and explore a more reliable cross-domain adaptive space. Specifically, the improvements encompass: (1) supporting augmented data qualification and refinement. Unbefitting augmentation can cause catastrophic error accumulation during learning; (2) providing weakly-labeled target data for the zero-resource scenario. The discriminator created in the adversarial strategy can select unlabeled target samples located in the adaptive space, and assign them pseudo labels by the NER classifier in the multi-task learner; (3) exhibiting detailed performance on each NER type; and (4) evaluating the cross-domain adaptability on two Chinese benchmarks. Individual achievements indicate the model's capability to generalize on different domains.

2 BACKGROUND AND OVERVIEW

In this section, we introduce the problem definition and related concepts. Then we discuss the problem scope and challenges in the cross-domain NER task.

2.1 Problem Definition

Based on the BIO schema[1], NER is to assign a sequence of labels $\mathcal{Y} = [y_1, \ldots, y_N]$ to a given sentence $\mathcal{X} = [x_1, \ldots, x_N]$ with N tokens. An entity is a span of tokens $\mathbf{e} = [x_i, \ldots x_j](1 \leq i \leq j \leq N)$ associated with an entity type. Specifically, the first token of an entity mentioned in the sentence with type X is labeled as $B-X$, the other tokens inside that entity mention are labeled as $I-X$, and the non-entity tokens are labeled as O. Therefore, $y_i \in$ label = {B-X, I-X, O}, where X is a

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

NER type such as "PER", "LOC", "ORG" and so on. While cross-domain NER addresses domain shift issues that occurred in NER for the sake of better domain adaptability. Domain adaptation is to adapt a learner to the new/target domain using the labeled data available from the original/source domain. The low-resource cross-domain NER task mainly studies the domain adaptation under un-supervision or semi-supervision on the target domain in this article, while most of the existing works [11, 12, 20, 45] lack evaluation on un-supervised cross-domain adaptation.

Unsupervised NER domain adaptation attempts to tackle the cross-domain discrepancy problem without any supervision on the target domain, assuming that no labels for examples are available from the target domain in training. In contrast, semi-supervised domain adaptation relaxes the strict constraint, using a small number of additional labels on the target data. We are given source domain $\{(\mathcal{X}_m^s, \mathcal{Y}_m^s)\}_{m=1}^{N_s}$ with N_s labeled examples, and target domain data $\{\mathcal{X}_m^t\}_{m=1}^{N_t}$ with N_t unlabeled testing examples. We assume the source domain is characterized by probability distributions P_s, while P_t represents target domain distribution. 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 optimize target prediction using source supervision. We denote the source domain data $\mathbf{D}^s = \{(\mathcal{X}_m^s, \mathcal{Y}_m^s)\}_{m=1}^{N_s}$, unannotated target data $\mathbf{D}^{t_u} = \{\mathcal{X}_i^{t_u}\}_{i=1}^{N_u}$, and annotated target data $\mathbf{D}^{t_a} = \{(\mathcal{X}_j^{t_a}, \mathcal{Y}_j^{t_a})\}_{j=1}^{N_a}$. $\mathbf{D}^t = \mathbf{D}^{t_u} \cup \mathbf{D}^{t_a}$ is the total target data.

2.2 Problem Scope and Challenges

Sequence labeling is a general but fundamental approach encompassing various natural language processing (NLP) tasks including word segmentation [26], **part-of-speech tagging (POS)** [38], and named entity recognition [16]. NER studies take a significant portion in the development of the sequence labeling family. Typically, existing NER methods follow the supervised learning paradigm and require high-quality annotations. While gold standard annotation is labor-intensive and time-consuming, imperfect annotations are relatively easier to obtain from crowd-sourcing or distant supervision manner but bring data noise issues. Researchers [1, 10, 17, 24] seek advanced techniques for out-of-domain knowledge transfer. They require fewer annotation efforts but perform similarly to gold annotations.

Even though the benefit of well-learned cross-domain knowledge is prevalent in low-resource annotation NER tasks, there are three main challenges faced by most of the state-of-the-art works.

(1) How to evaluate the domain discrepancy for better mitigation. Existing approaches [1, 4, 15, 19, 20, 24, 41, 45] form domain shift into either word-level or discourse-level discrepancy. In other words, an entity mention can appear and be assigned with different types across domains. Besides, different types of entities show imbalance frequency. For example, a large number of location names are shared in the political news domain and the sports domain, but the case is very different for organization names across these domains. Moreover, written styles are distinct among different data resources, like science news and tweets. Approaches in the literature are diverse in the adaptation techniques to mitigate the gap of either token distribution or sentence structure. The open-mind works [4, 19, 41] focus on token-level consistency and complement with the help of auxiliary knowledge base linking, which introduces too much noise. There are many studies [11, 12, 20] discuss how to cope with both word- and discourse-level discrepancies. However, those methods either lack the capability to capture expressive text features for the adaptation or should consume sufficient labeled target data.
(2) How to learn sufficient entity features in the source domain. Usually, we expect the model to take in as much knowledge of the NER classes as possible. But, it is hard to find a single

training dataset that exactly covers all the required NER types. For each type, sufficient mention instances are another consideration of source domain selection. Even though words overlap across domains, their combination or usage is different. In low-resource scenarios, supervised models incline to lose their behavior due to a lack of enough annotated data to support the cross-domain bridge. The self-training strategy [19] and Knowledge Base assistance [4, 41, 50] work to alleviate the issue during the training. More intuitively, those methods harness the idea of data augmentation through either the pseudo label prediction in self-training or the additional token linking from knowledge bases. The augmented data truly benefits domain adaptation, but still suffers the lack of qualification. As we know, noise data usually brings inevitable distraction to the model training.

(3) How to extend and guarantee the model's ability to minimal or no annotation scenarios. As mentioned above, data augmentation is explored to work as a compliment when transferring knowledge from the high-resource source domain to the low-resource target domain. But this leaves another consideration of data qualification. Our preliminary work [56] constructs adaptive data under the guidance of cross-domain anchors. The results show its superiority but raise further thinking about automatically adaptive data refinement. Even though limited augmented data bring decent performance increases, we still expect more improvement with more adaptive data that exactly form the cross-domain bridge.

3 PRELIMINARY

In this section, we introduce the base model in NER and the maximum mean discrepancy measurement.

3.1 Base Model

We adopt an expressive pre-trained language model (e.g., BERT [6]) to encode the sentence $\mathcal{X} = [x_{\text{CLS}}, x_1, \ldots, x_N, x_{\text{SEP}}]$ into a sequence hidden states $\mathbf{h} = [h_{\text{CLS}}, h_1, \ldots, h_N, h_{\text{SEP}}]$, as the sentence representation.

Encoder. We encode the input example $\mathcal{X}$ through the encoder **Encoder** to extract its features $\mathbf{h}$:

$$\mathbf{h} = \textbf{Encoder}(\mathcal{X}). \tag{1}$$

NER Classifier. We describe the NER task objective as CRF loss, where $\mathcal{L}_{\text{crf}} = \log p_{crf}(\mathcal{Y}|\mathcal{X})$.

$$p_{crf}(\mathcal{Y}|\mathcal{X}) = \frac{1}{\mathbf{Z}} \prod_{i=1}^{N} \phi_n(y_i|h_i, \mathbf{V}) \prod_{i=1}^{N-1} \phi_e(y_{i,i+1}|\mathbf{A}), \tag{2}$$

$$\mathcal{L}_{\text{crf}} = \sum_{i=1}^{N} \phi_n(y_i|h_i, \mathbf{V}) + \sum_{i=1}^{N-1} \mathbf{A}_{y_i, y_{i+1}} + \log \mathbf{Z}, \tag{3}$$

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

3.2 Maximum Mean Discrepancy (MMD) Measurement

The MMD measures the difference of two distributions (P_s, P_t) in a pre-defined function space $\mathcal{H}_k$. Usually, the **Reproducing Kernel Hilbert Space (RKHS)** works as the function space $\mathcal{H}_k$ with a kernel k. The MMD measurement computes the squared formulation, $d_k^2(P_s, P_t)$ which is denoted as:

$$d_k^2(P_s, P_t) = \|\mathbf{E}_{P_s}[\varphi(\mathbf{D}^s)] - \mathbf{E}_{P_t}[\varphi(\mathbf{D}^t)]\|_{\mathcal{H}_k}^2, \tag{4}$$

where the mapping function, $\varphi : \mathcal{X} \rightarrow \mathcal{H}_k$. And $P_s = P_t$ iff $d_k^2(P_s, P_t) = 0$. Gaussian Kernel $k(\mathbf{D}^s, \mathbf{D}^t)$ is used in φ. When adapting MMD metric on cross-domain NER, we compute the squared version of MMD between source/target samples feature vectors:

$$d_k^2(\mathbf{H}^s, \mathbf{H}^t) = \frac{1}{(N^s)^2} \sum_{i,j=1}^{N^s} k\left(h_i^s, h_j^s\right) + \frac{1}{(N^t)^2} \sum_{i,j=1}^{N^t} k\left(h_i^t, h_j^t\right) - \frac{2}{N^s N^t} \sum_{i,j=1}^{N^s N^t} k\left(h_i^s, h_j^t\right), \quad (5)$$

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

4 THE PROPOSED MODEL

In this section, we present the details of model design in PDALN†. Above all, domain adaptation components contains adaptive data augmentation and selection, and the multi-level MMD metrics for domain-invariant feature extraction. Firstly, adaptive data augmentation works to tackle the labeled data insufficiency issue. Secondly, we construct a multi-task schema with both the NER classifier and adversarial domain discriminator. This schema can explore the potential adaptive space for augmented data qualification and selection. Thirdly, a multi-grained MMD metric works on the augmented adaptive data to extract domain invariant features. There is an intuitive illustration in Figure 1 to describe the motivation and function of each domain adaptation components, and interpret how it solves the domain shift problems. Besides, we exploit a pre-trained model to capture feature embedding, due to its impressive success in word contextual embedding. We investigate a self-training strategy to progressively and effectively perform our domain adaption components, as shown in Figure 3. 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

Cross-domain NER models suffer over-fitting on limited labeled data. Thus, we exploit *Cross-Domain Anchor* pairs to synthesize mix-domain data, so-called **adaptive data**. The adaptive data is the cure for alleviating both word-level and discourse-level cross-domain gaps. Because we can explore the adaptive space (in Figure 1) via those adaptive data, which is exactly the cross-domain bridge transferring knowledge. To avoid including noise adaptive data, an adaptation capability qualifier works to rank and select the most adaptive augmented data fed into the downstream component.

4.1.1 Adaptive Data Augmentation. Pioneers [1, 20, 24, 45] mostly address domain shift problem by reducing the word-level and discourse-level cross-domain discrepancy. We synthesize the adaptive data to provide shared features and bridge cross domain gaps on both word-level and discourse-level.

We first introduce *Cross-Domain Anchor* used in adaptive data synthesis. We denote a source domain entity by $\mathbf{e}^s$ whose labels are $[y_{i^s}^s, \ldots y_{j^s}^s]$, and a target entity by $\mathbf{e}^t$ whose labels are $[y_{i^t}^t, \ldots y_{j^t}^t]$. Therefore, *Cross-Domain Anchor* pairs are defined as $\mathcal{M}_{Anchor} = \{(\mathbf{e}^s, \mathbf{e}^t), y_{i^s}^s = y_{i^t}^t\}$. $y_{i^s}^s = y_{i^t}^t$ denotes two entities belonging to the same entity type when their first label is the same. Intuitively, the anchor pairs works to alleviate the cross-domain word-level discrepancy. 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 $\mathbf{e}^p \in \mathcal{X}^p = [x_1^p, \ldots, x_{i^p}^p, \ldots x_{j^p}^p, \ldots, x_{|\mathcal{X}^p|}^p]$. Given an anchor pair $(\mathbf{e}^p, \mathbf{e}^q) \in \mathcal{M}_{Anchor}$, where $q \in \{s, t\}$ and $q \neq p$, we replace $\mathbf{e}^p$ in $\mathcal{X}^p$ with $\mathbf{e}^q$ as the augmented adaptive data $\mathcal{X}^{p\prime} = [x_1^p, \ldots, x_{i^q}^q, \ldots x_{j^q}^q, \ldots, x_{|\mathcal{X}^p|}^p]$. Finally, we obtain the adaptive data $\mathbf{D}^{aug} = \{\mathcal{X}^{p\prime}\}$. Intuitively, the augmented mix-domain sentences are considered as adaptive data because of their

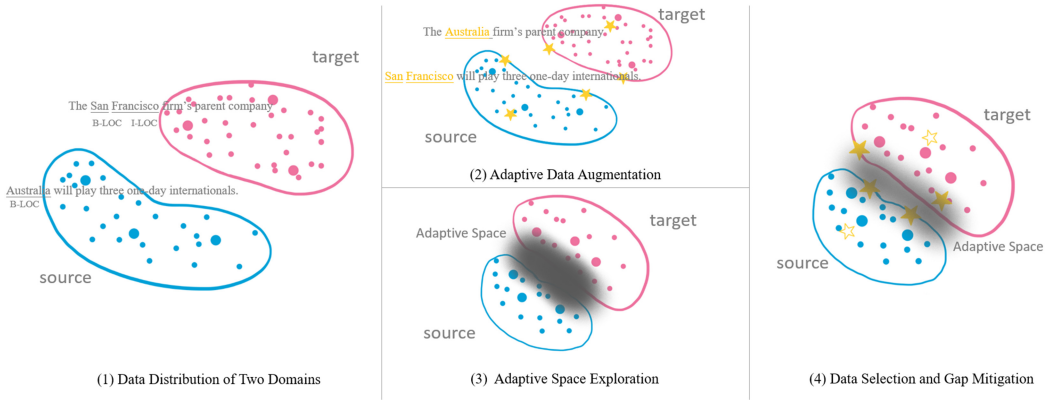

Fig. 1. Toy illustration of the method. (1) Data Distribution of Two Domains. (2) Adaptive Data Augmentation. (3) Adaptive SPace Exploration. (4) Data Selection and Gap Mitigation.

shared sentence pattern across domains. For example, in Figure 1(2), adaptive sentences are "The Australia firm's parent company." and "San Francisco will play three one-day internationals.". We exchange the words referring The *Cross-Domain Anchor* pair ("Australia", "San Francisco") which are both assigned to the label "LOC".

4.1.2 Adversarial Selection. Mix-domain data possibly introduce undesired noise into the adaptive data, which draws down the performance instead of adaptation benefit. Therefore, we seek a way to qualify and select the adaptive data and expect the selected adaptive data to be more valuable when extracting domain-invariant features. We adopt the adversarial strategy to learn the adaptive space and filter out unconfident data around the boundary.

Multiple domains adversarial networks [8, 31] achieve great success in extracting transferable knowledge to enable domain adaptation. Usually, the adversarial learning procedure is a two-player game. The first player is a binary classifier, the domain discriminator, trained to distinguish the source domain from the target domain. The second player is the domain-invariant feature extractor to confuse the domain discriminator.

Domain Discriminator. Two linear neural networks followed by a ReLU function comprise the domain discriminator. The domain discriminator takes in the token embeddings and passes down domain recognition, based on which the sigmoid function predicts the probability of whether the input belongs to the source domain,

$$p(\mathcal{Y}^d|\mathcal{X}) = \sigma(\mathbf{W}^1 \cdot \mathrm{ReLU}(\mathbf{W}^2 \cdot \mathbf{Encoder}(\mathcal{X}))), \tag{6}$$

where $\mathbf{W}^2 \in \mathbb{R}^{d_d \times d_e}$ and $\mathbf{W}^1 \in \mathbb{R}^{d_l \times d_d}$. d_e is the hidden dimension of the encoder. d_d is the hidden dimension of the discriminator. d_l is the label size of the domain classification task. σ is the sigmoid function to obtain the domain probability of each word. $\mathcal{Y}^d \in \{0,1\}^{d_l}$ is the domain prediction. The loss function in the domain discriminator is denoted by:

$$\mathcal{L}_{dis} = \mathrm{CrossEncropy}(p(\mathcal{Y}^d|\mathcal{X}), \mathcal{Y}^{dis}), \tag{7}$$

where $\mathcal{Y}^{dis} \in \{0,1\}^{d_l}$ is the ground truth of the domain classification task. In contrast, the domain-invariant feature extractor loss is denoted by:

$$\mathcal{L}_{fea} = -\mathrm{CrossEncropy}(p(\mathcal{Y}^d|\mathcal{X}), \mathcal{Y}^{dis}). \tag{8}$$

ALGORITHM 1: Adversarial Training for Domain Adaptive Space

Require: Source domain sentences $\mathcal{X}^s$, their NER labels $\mathcal{Y}^s$, and their binary domain label $\mathcal{Y}^{d_s}$. Target domain sentences $\mathcal{X}^t$, and their binary domain label $\mathcal{Y}^{d_t}$. Adaptive data, $\mathbf{D}^{aug}$. Number of batches, *Batch*. Slack factor, ξ. Selection threshold: n for adaptive data $\mathbf{D}^{aug}$; ρ for weakly-labeled target data $\mathbf{D}^{pseudo}$.

Ensure: Selected Adaptive data $\mathbf{D}^{aug}_{\xi}$, and Weakly-labeled target data $\mathbf{D}^{pseudo}$.

procedure ADV-SELECT($\mathcal{X}^s, \mathcal{X}^t, \mathcal{Y}^s, \mathcal{Y}^{d_s}, \mathcal{Y}^{d_t}, Batch$)
  for $i = 1, \ldots, Batch$ **do**
    for model m in {NER, Fea, Dis} **do** // Three objectives in multi-task framework.
      if model m is NER **then**
        Sample batch $\{x_i^s, y_i^s\}$. // Where $x_i^s \in \mathcal{X}^s$, and $y_i^s \in \mathcal{Y}^s$.
        Compute $\mathcal{L}_{crf} \leftarrow$ Equation (3).
        Update parameters θ_{NER} in Encoder and CRF layer. // θ_{NER} contains both parameter in Encoder θ_e and CRF layer θ_{crf}.
      else
        Sample batch $\{x_i, y_i^d\}$. // Where $x_i \in \mathcal{X}^s \bigcup \mathcal{X}^t$, and $y_i^d \in \mathcal{Y}^{d_s} \bigcup \mathcal{Y}^{d_t}$.
        if model m is Dis **then**
          Compute $\mathcal{L}_{dis} \leftarrow$ Equation (7).
        else
          Compute $\mathcal{L}_{fea} \leftarrow$ Equation (8).
        Update parameters θ_{bin} in Encoder and Discriminator layers. // θ_{bin} contains both parameter in Encoder θ_e and discriminator layer θ_{dis}.
  for $i = 1, \ldots, len(\mathbf{D}^{aug})$ **do**
    Calculate $I_{domain}(x_i^{aug}, \xi) \leftarrow$ Equation (9). // Where $x_i^{aug} \in \mathbf{D}^{aug}$.
  Rank $\{I_{domain}(x_i^{aug}, \xi)\}$ and Select top n to be $\mathbf{D}^{aug}_{\xi}$.
  for $i = 1, \ldots, len(\mathcal{X}^t)$ **do**
    Calculate $I_{domain}(x_i^t, \xi) \leftarrow$ Equation (9). // Where $x_i^t \in \mathcal{X}^t$.
  Rank $\{I_{domain}(x_i^t, \xi)\}$ and Select top ρ to be $\mathcal{X}^{t\prime}$.
  Predict pseudo label $\mathcal{Y}^{t\prime}$ of $\mathcal{X}^{t\prime}$ by NER model with parameter θ_{NER}, and $\mathbf{D}^{pseudo} = \{\mathcal{X}^{t\prime}, \mathcal{Y}^{t\prime}\}$.

We formulate the adversarial selection model as a multi-task framework, including the NER task and two binary domain classification tasks. We describe the training details in Algorithm 1 and Figure 2. To better explore the target domain, we make the model take in both source and target domain data. Since there are many unannotated data $\mathbf{D}^{t_u}$ in target, they are only allowed to update the discriminator and feature extraction task but not the NER task. The domain discriminator tries to make the encoder unable to distinguish the domain of a token through confrontation. In this way, the encoder should pay more attention to features that are less related to the source domain when learning the NER task. After adversarial training, the domain discriminator can still correctly classify certain sentences with a high probability. We define these as domain-adhered samples. Other samples are ambiguous regarding domain (for example, sentences with a probability close to 0.5), and they are defined as samples that are more domain-invariant. To select the most adaptive data in $\mathbf{D}^{aug} = \{\mathcal{X}^{p\prime}\}$, we calculate domain-invariant score, I_{domain}, for each of the adaptive samples. It is denoted by:

$$I_{domain} = \max(1 - |p(y_s^d) - 0.5|, 1 - |p(y_t^d) - 0.5| + \xi), \tag{9}$$

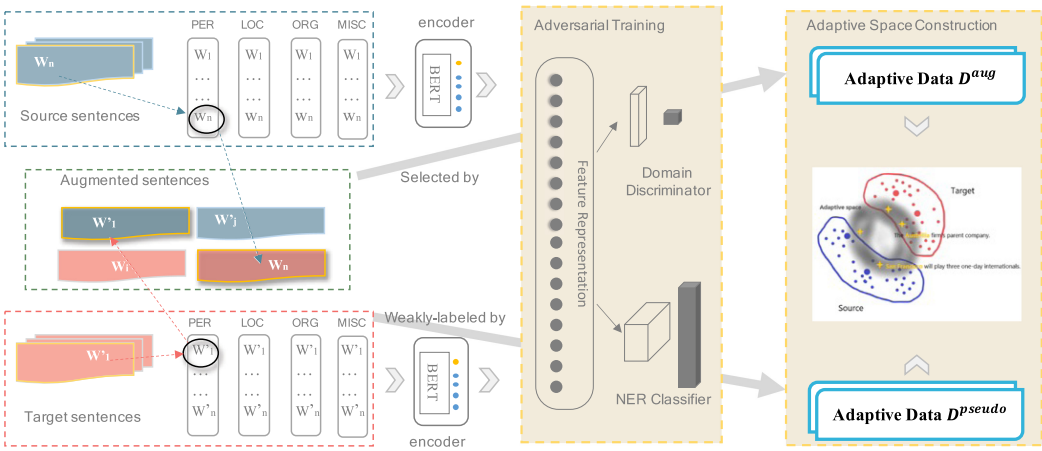

Fig. 2. Multi-task learning strategy with both NER task and adversarial domain-invariant feature space learning task. The learner through adversarial training helps to select and refine the adaptive data augmentation proposed in Section 4.1.1.

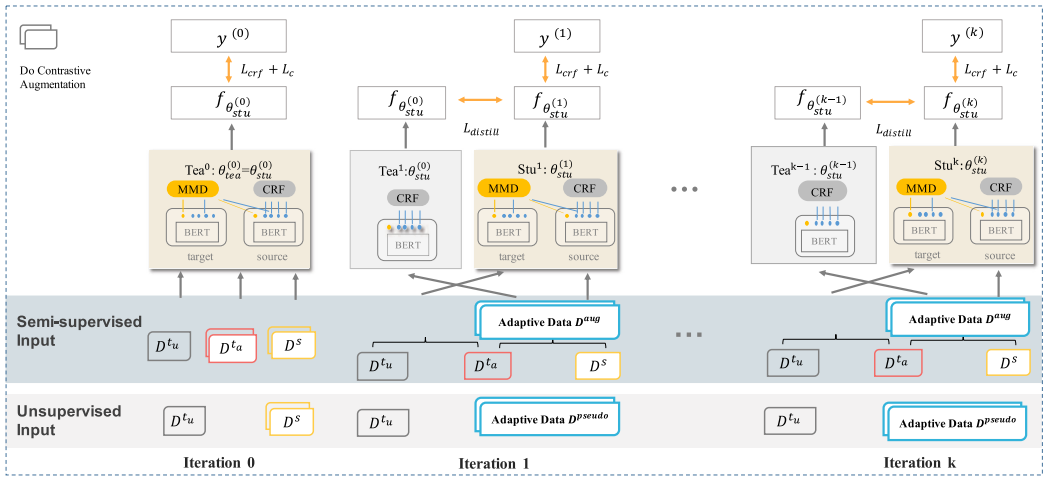

Fig. 3. Low-resource cross-domain NER Training Schema. Semi-supervised Training mainly extract the domain-invariant feature in the adaptive data by a progressive knowledge distillation strategy.

where $p(y_s^d)$ and $p(y_t^d)$ are predicted probabilities on source and target domain through the domain discriminator. I_{domain} indicates how much the example is independent of either source or the target domain. If it is hard to tell which domain the example should be, the example is regarded as the domain-invariant feature carrier. Therefore, we select samples with the highest I_{domain} in the top n to refine set $\mathbf{D}^{aug} = \{\mathcal{X}^{p'}\}$. ξ is a slack factor for releasing the constraints on the target domain.

The adaptive data $\mathbf{D}^{aug}$ serves only the semi-supervised cross-domain NER task because it exploits the label information to construct cross-domain anchor pairs. For the unsupervised cross-domain NER task, we can use the multi-task schema to create weakly labeled target data. As we discussed above, the adversarial discriminator can explore the potential adaptive space, sharing the

domain-invariant features among the samples located in it. Therefore, we can search the whole unlabeled target examples and pick out those located in or close to the adaptive space. I_{domain} measures how much the sample is close to the adaptive space. We rank the samples by I_{domain} and select the top ρ samples to be the domain-adaptive pseudo-labeled dataset $\{\mathcal{X}^{t\prime}\}$. ρ is the hyper-parameter to decide the ratio of selected data. After deciding the domain-invariant sample set in the target domain, the NER classifier can predict pseudo-NER labels by $p_{crf}(\mathcal{Y}^{t\prime}|\mathcal{X}^{t\prime})$. Finally, the weakly-labeled domain-invariant sample set is $\mathbf{D}^{pseudo} = \{\mathcal{X}^{t\prime}, \mathcal{Y}^{t\prime}\}$.

4.1.3 Multi-grained MMD for Domain-invariant Features. In this part, we present how to extract multi-level domain-invariant features through multi-grained MMD metrics. Therefore, data points sharing same word and sentence features can move together, like shown in Figure 1(3). We devise the multi-grained MMD objectives, L^{w}_{MMD} and L^{d}_{MMD}, to measure and alleviate the word-level and discourse-level gaps.

$$\mathcal{L}^{d}_{\text{MMD}}(\mathbf{D}^s, \mathbf{D}^t) = d^2_k\left(\mathbf{H}^s_{\text{CLS}}, \mathbf{H}^t_{\text{CLS}}\right), \tag{10}$$

$$\mathcal{L}^{w}_{\text{MMD}}(\mathbf{D}^s, \mathbf{D}^t) = \sum_{y\in\text{label}} \mu_y d^2_k(\mathbf{H}_y(\mathbf{D}^s), \mathbf{H}_y(\mathbf{D}^t)), \tag{11}$$

where $\mathbf{H}_{\text{CLS}}$ is the set of CLS token embeddings in a pre-trained model. $\mathbf{H}_y$ are the set of token embeddings assigned with label y. μ_y is the corresponding coefficient.

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

4.2.1 Robust Feature Adaptation. Considering limited vocabulary and noise data samples, we leverage the contrastive learning [5, 9, 34, 52] to boost feature extraction. We follow the positive and negative sample construction in [46] to construct a distorted dataset $\mathbf{D}^c = \{(\mathcal{X}', \mathcal{Y}')\}$ over a given dataset $\mathbf{D} = \{(\mathcal{X}, \mathcal{Y})\}$.

$$\mathcal{L}_c = -\log \frac{\exp(\mathbf{z}\cdot\bar{\mathbf{z}})/\tau}{\sum_{\mathbf{z}_i\in\{\bar{\mathbf{z}}\}\cup\mathbf{Z}^{neg}}\exp(\mathbf{z}\cdot\mathbf{z}_i/\tau)}, \tag{12}$$

where $\mathbf{z} = \mathbf{W}^\top h_{CLS}$, $\bar{\mathbf{z}} = \mathbf{W}^\top \bar{h}_{CLS}$ are mapping vectors of a sentence $\mathcal{X}$ and $\mathcal{X}'$, respectively. $\mathbf{Z}^{neg}$ is the negative samples from $\mathbf{D}\cup\mathbf{D}^c$ except $\mathcal{X}$ and $\mathcal{X}'$. τ is a temperature hyper-parameter.

4.2.2 Low-Resource Objectives. Low-resource cross-domain NER includes both zero-resource and minimal-resource scenarios. For the zero-resource scenario, we use all unlabeled target data to compute the adaptation on the discourse level. Since the word-level MMD loss needs pre-provided token label information, we evaluate it on the domain-adaptive and pseudo-labeled data $\mathbf{D}_{\textbf{pseudo}}$. Therefore, the **unsupervised cross-domain NER** loss is denoted as:

$$\mathcal{L}_{\text{unDA}} = \mathcal{L}_{\text{crf}} + \alpha' L^{d}_{\text{MMD}}(\mathbf{D}^s, \mathbf{D}^{t_u}) + (1-\alpha')\cdot\mathcal{L}^{w}_{\text{MMD}}(\mathbf{D}^s, \mathbf{D}^{\textbf{pseudo}}) + \mathcal{L}_c. \tag{13}$$

For the minimal-resource scenario, we can use a small size of labeled target data $\mathbf{D}^{t_a}$ to estimate the word-level adaptation. Then, the **semi-supervised cross-domain NER** objective is denoted as:

$$\mathcal{L}_{\text{semiDA}} = \mathcal{L}_{\text{crf}} + \alpha\cdot L^{d}_{\text{MMD}}(\mathbf{D}^s, \mathbf{D}^t) + \beta\cdot\mathcal{L}^{w}_{\text{MMD}}(\mathbf{D}^s, \mathbf{D}^{t_a}) + \mathcal{L}_c, \tag{14}$$

where α and β are the hyper-parameters.

4.2.3 Progressive Joint KD and DA. First of all, we train the base model on both source and target domains to explore a rough adaptive space. Then to further refine it, we exploit a progressive teacher-student framework to perform the pre-trained model on adaptive data. The sequential teacher-student framework works to prohibit from over-fitting on limited augmented adaptive data. The sequential students can progressively overlook "problematic" examples but learns things

that generalize well from its teacher. Therefore, the KD framework can improve the domain adaptation confidence over the adaptive data. The cross-domain NER loss used for adaptive data is denoted as:

$$\mathcal{L}_{\text{unDA}} = \mathcal{L}_{\text{crf}} + \alpha' L^{d}_{\text{MMD}}(\mathbf{D}^{pseudo}, \mathbf{D}^{t_u}) + \mathcal{L}_c, \tag{15}$$

$$\mathcal{L}_{\text{semiDA}} = \mathcal{L}_{\text{crf}} + \alpha \cdot \mathcal{L}^{d}_{\text{MMD}}(\mathbf{D}^{aug}, \mathbf{D}^{t}) + \beta \cdot \mathcal{L}^{w}_{\text{MMD}}(\mathbf{D}^{aug}, \mathbf{D}^{t_a}) + \mathcal{L}_c. \tag{16}$$

We use $f_{\theta_{tea}}$ and $f_{\theta_{stu}}$ to denote teacher and student models in the progressive KD framework, respectively. $f_{\hat{\theta}}$ is the base model learned by Equation (14). Then, we initial the teacher and the student model as: $\theta^{(0)}_{tea} = \theta^{(0)}_{stu} = \hat{\theta}$. The t-th student computes the loss:

$$\mathcal{L}_{\text{distill}} = (1-\gamma) \cdot \mathcal{L}_{\text{DA}} + \gamma \cdot \frac{1}{N} \sum_{n=1}^{N} -f_{\theta^{(t)}_{tea}, n}(\mathcal{X}) \log f_{\theta_{stu}, n}(\mathcal{X}), \tag{17}$$

where $\mathcal{L}_{\text{DA}} \in \{\mathcal{L}_{\text{unDA}}, \mathcal{L}_{\text{semiDA}}\}$, $\mathcal{X} \in \mathbf{D}^{aug}$, 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 $(t+1)$-th iteration by: $\theta^{(t+1)}_{tea} = \theta^{(t+1)}_{stu} = \hat{\theta}^{(t)}_{stu}$.

5 EXPERIMENTS

In this section, we make a comparison between PDALN† and other baselines on four English and two Chinese public benchmarks. Besides, to better evaluate the domain adaptability on each NER label, we present a comparative analysis of each label type with our previous work, PDALN [56].

5.1 Datasets

All datasets contain the four common entity types, including **PER (person)**, **LOC (location)**, **ORG (organization)**, and **MISC (miscellaneous)**. For the English group, the source comes from CoNLL-2003 English NER data [40] comprising 21.0K/3.5K/3.7K samples for the training/validation/test sets. The target domains are:

(1) **SciTech** [11] contains a set of science and technology news with 3K sentences.
(2) **WNUT 2016** [42] contains 2,400 tweets (comprising 61K sentences and 34k tokens) with 10 entity types. We convert 10 types in WNUT 2016 into four CoNLL03 entity types for evaluation consistency.
(3) **Webpage** [37] comprises 783 entities from 20 webpages, including many long sentences.
(4) **Wikigold** [2] contains distant supervised examples derived from Wikipedia articles with 40k tokens.

The two Chinese datasets are the SIGHAN2006 NER dataset (Sighan) [18] and the Weibo NER dataset (Weibo) [35]. Sighan contains three types, including person, location, and organization, while Weibo contains additional types (geopolitical entity) apart from these three. We convert four-type Weibo into a three-type dataset by merging type geopolitical entity into type location. We take Sighan (only training set) as the source domain and the entire small Weibo to be the target.

Table 1 shows the data distributions on each NER type. SciTech has a comparative annotation advantage on these four types since specific types (i.e., "PER", "LOC", "ORG") take the major parts like the case in the source domain. However, WNUT 2016 suffers the flood of type "MISC" after reducing 10 types to four CoNLL03 entity types. Webpage lacks balanced annotations. Sighan, as the source, contains sufficient examples of each type.

5.2 Baselines

BiLSTM+CRF [16], early vanilla base model in NER.

Table 1. The Distributions of Sentence and Each Entity in the Seven Datasets Mentioned in 5.1

Datasets	#Sentence	#PER	#LOC	#ORG	#MISC
CoNLL	217662	6600	7140	6321	3438
SciTech	34733	794	266	538	228
WNUT 2016	61908	669	867	760	2092
Webpage	5678	233	32	69	59
Wikigold	6812	140	165	176	126
Sighan	41728	9028	18522	10261	–
Weibo	1890	1919	425	333	–

BERT+CRF, advanced vanilla base model, which only replaces with the pre-trained language model BERT.
La-DTL [41] introduces the label-aware MMD metric for alleviating word-level discrepancy.
DATNet [57] devises a generalized resource-adversarial discriminator to capture the cross-domain features.
JIA2019 [11] constructs multi-task architecture for cross-domain NER knowledge transfer. Its key idea is the tensor decomposition to learn the task embedding.
Multi-Cell [12] proposes another multi-task learning strategy. The authors devise a multi-cell compositional LSTM structure for cross-domain NER.

We evaluate the performance on two variants of PDALN†. Following PDALN, we replace the sequential KD framework in the self-training stage with **MT** and **VAT**, Mean Teacher strategy [44] and Virtual Adversarial Training [29], respectively.

5.3 Training and Implementation Details

Our optimizer is Adam with a decay learning rate of 0.00005. We adopt cased BERT-base and BERT-base-Chinese, with 12 transformer blocks and the self-attention heads, 768 of the hidden layer size, and 32 of the batch size. The slack factor in Equation (9) is $\xi = 0.1$. The temperature hyper-parameter is $\tau = 0.05$. The coefficient μ_y in Equation (11) is 0.25. We pick out 200/1000 labeled target/source examples to synthesize adaptive data, obtaining 2,800 (200*4+1000*2) examples. Through the discriminator learned in Section 4.1.2, we choose the top 1,400 examples with the highest I_{domain} scores.

5.4 Results and Discussion

5.4.1 Domain Adaptation on Unsupervised NER. Our main results are shown in Tables 2 and 3 for English and Chinese benchmarks, respectively. These two tables both exhibit two groups of results, one for un-supervised cross-domain NER and the other for semi-supervised cross-domain NER. Unsupervised cross-domain NER represents the zero-resource scenarios, where model training is blind to labeled target examples. Not all baselines can extend performance on the zero-shot paradigm. Compared with the baselines with zero-shot capability, PDALN† achieves the best F-1 scores on all benchmarks, beating PDALN. Its performance gain ranges 0.6%-1.3% over PDALN. Particularly, PDALN† attains significantly new state-of-the-art performance. Adversarial adaptive data augmentation plays a core role in domain adaptation when there is no accessibility to labeled data. Moreover, results show that approaches (BERT+CRF, PDALN, and PDALN†) integrated with a

Table 2. Comparison with four English Benchmarks under the Evaluation Metrics, F1 Score (Precision/Recall) (in %)

Baselines	SciTech	WNUT 2016	Webpage	Wikigold
	F1 (Pre/Rec)	F1 (Pre/Rec)	F1 (Pre/Rec)	F1 (Pre/Rec)
Un-supervised NER				
BiLSTM+CRF	67.01 (73.53/61.56)	26.54 (47.79/18.37)	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
PDALN†	**76.71** (71.46/**82.78**)	**47.27** (37.44/64.11)	**58.29** (**59.09**/**57.51**)	**50.34** (46.03/**55.55**)
	76.65 ± 0.07	47.10 ± 0.11	58.21 ± 0.10	50.28 ± 0.11
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/**MT**	77.76 (73.05/83.12)	48.34 (38.21/65.79)	58.38 (59.88/56.97)	51.89 (47.54/57.13)
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† w/**VAT**	77.85 (73.25/83.08)	47.86 (38.01/64.61)	58.69 (59.64/57.77)	51.30 (47.09/56.35)
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
PDALN†	78.17 (73.27/**83.77**)	**49.21** (38.84/**67.17**)	**59.17** (**60.39**/**58.01**)	**53.44** (49.23/**58.44**)
	78.05 ± 0.09	49.00 ± 0.10	59.12 ± 0.05	53.27 ± 0.13

Only the performances of PDALN and PDALN† comprise two parts: the best score among five runs in the top, average F-1 score with deviation score in the bottom.

pre-trained language model obtain decent recall but suffer a failure on the precision scores. Though showing competitive performance on the recall scores, PDALN† and PDALN surpass BERT+CRF overall due to benefits from multi-level domain adaptation with the contrastive-learning fused pre-trained language model.

5.4.2 Domain Adaptation on Semi-supervised NER. As shown in Tables 2 and 3, most of the baselines in the group of unsupervised cross-domain NER cannot achieve decent performance gain by only taking in limited annotated resources. But PDALN† and PDALN show their superiority over not only the unsupervised approaches but also those in supervision. Their outperformance over the most state-of-the-art shows overall 2%-4% improvements among all benchmarks.

Table 3. Chinese Benchmark Evaluation under the Evaluation Metrics: F1 Score (Precision/Recall) (in %)

Baselines	Un-supervised NER F1 (Pre/Rec)	Semi-supervised NER F1 (Pre/Rec)
BiLSTM+CRF	38.01 (51.34/30.26)	41.45 (52.66/34.18)
BERT+CRF	52.43 (55.37/49.78)	54.83 (57.70/52.24)
La-DTL	–	43.87 (53.64 /37.11)
DATNet	–	45.37 (53.88/39.18)
JIA2019	37.35 (50.94/29.48)	41.88 (52.19/34.97)
Multi-Cell	40.14 (50.19/33.44)	42.03 (52.93/34.85)
PDALN w/**MT**	–	55.62 (59.98/51.85)
PDALN† w/**MT**	–	55.92 (60.49/51.99)
PDALN w/**VAT**	–	56.61 (60.64/53.08)
PDALN† w/**VAT**	–	57.45 (61.37/**54.00**)
PDALN	53.83 (56.85/51.11) 52.99 ± 0.76	55.80 (60.21/52.00) 54.57 ± 0.80
PDALN†	**53.99** (**57.13**/**51.17**) 53.41 ± 0.13	**57.59** (**61.82**/53.91) 56.76 ± 0.55

Only the performances of PDALN and PDALN† comprise two parts: the best score among five runs in the top, and the average F-1 score with deviation score in the bottom.

Since many baselines adopt an RNN-based network to encode input sentences, it is a challenge to overcome model intrinsic shortcomings, i.e., vanishing and exploding gradient problems. These approaches lack expressive contextual information to assist recall score rise, which contrarily is the merit of most pre-trained language models. Consequently, they are prone to increase false-positive predictions.

Even though approaches integrated with a language model achieve stunning recall scores, their precision scores dramatically fall behind the baselines without one. The pre-trained model group shows too much power on limited annotated data, easily causing overfitting. PDALN† and PDALN take a breakthrough in moderating the pre-trained language model's capability to make a trade-off between the precisions and recalls. They attain promising precision gain and increasing recall scores compared with BERT+CRF. Their breakthroughs mainly benefit from the progressive domain adaptation with moderate knowledge distillation from the teachers.

Besides, we compare two variants (w/**MT** and w/**VAT**) of PDALN† and PDALN with different KD strategies, one for Sequential Mean Teacher and the other for teacher-student Virtual Adversarial Training. Their performance approaches to PDALN† and PDALN on the high-quality labeled data, SciTech. But they are vulnerable to noise data in WNUT 2016 and easily overfit on limited and incomplete annotated samples in Webpage or Wikigold.

At last, we describe the evaluation of improvement between PDALN† and PDALN. Overall, PDALN† brings the model's stability with significantly-reduced performance deviation. The model's stability shows the practicability and reliability of adversarial adaptive data augmentation.

Table 4. Evaluation on the four Entity Types

PDALN / 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)
	91.34 (92.11/90.59)	71.40 (64.35/80.20)	68.45 (60.55/78.72)	**49.13** (45.61/53.25)
WNUT 2016	86.27 (84.51/88.12)	48.57 (44.33/53.71)	46.81 (41.11/54.36)	27.90 (21.48/39.79)
	86.39 (85.77/87.02)	**49.07** (44.91/54.08)	**47.01** (41.12/54.88)	**28.34** (22.00/39.81)
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)
	80.96 (79.70/82.27)	45.18 (40.98/50.34)	**46.28** (43.88/48.93)	42.48 (39.30/46.22)
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)
	84.97 (84.95/85.00)	**43.65** (40.01/48.02)	41.58 (35.37/50.44)	**37.97** (32.34/45.98)

PDALN† achieve outperformance on three of the four English benchmarks. The adversarial data selection serves for noteworthy improvement on the datasets suffering low annotation quality.

5.4.3 Evaluation on Entity Type. We provide PDALN†'s and PDALN's performance on each entity type in Table 4. The performance on type "PER" is more stable and well-performed than the other three, "LOC", "ORG", and "MISC". In other words, domain discrepancy is mainly caused by the other three types, which convey the difference in entity distributions and topics between the four benchmarks. Entity performances also vary among those four benchmarks. SciTech takes the best performer thanks to its comparative annotation quality. Most of the entities in SciTech are in the group of type "PER", which have been well recognized almost approaching 91% F1 scores. Thus, domain adaptation components in PDALN only provide limited performance gains. That means the other three benchmarks leave more space for domain discrepancy mitigation by both PDALN† and PDALN. They obtain significant performance gains compared with SciTech. Compared with PDALN, PDALN† shows its advantage on type "MISC", due to the adversarial adaptive data selection.

5.4.4 Ablation Study. As Table 5 shows, the adaptive data is the most important component contributing to the performance improvements. The progressive KD framework plays a secondarily important role. Similar findings with PDALN, the multi-grained MMD and $\mathcal{L}_c$ are still crucial for robust feature extraction, especially for the noisy dataset, like distant supervised Wikigold. PDALN†'s success depends more on the adversarial adaptive data selection than PDALN does on adaptive data without selection.

The ablation study on each individual entity type in Table 6 shows that data augmentation and progressive self-training perform more impressive results for domain adaptation than the multi-grained MMD method. Mostly, domain gap mitigation works more significantly on "LOC", "ORG", and "MISC" than "PER". As we discussed in the paper, the BERT-based models are prone to overfitting small annotated data, especially for imbalanced data. Even if the powerful pre-trained language model brings high recalls but ruins the precision by increasing false positive examples. But D^{aug} and $L_{distill}$ help the precision increase for most entity types as well as achievement in the recall in PDALN†.

5.4.5 Case Study. In Figure 4, we compare the predictions on the sample sentences among ground-truth, best baseline model, and ours. There are some complex Chinese samples to show the

Table 5. Ablation Study

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**)
PDALN†	78.17 (73.27/83.77)	49.21 (38.84/67.17)	59.17 (60.39/58.01)	53.44 (49.23/58.44)
w/o $\mathcal{L}_c$	−0.51 (−0.25/−0.83)	−0.92 (−0.90/−0.72)	−0.62 (−0.75/−0.52)	−1.14 (−1.28/−0.93)
w/o $\mathcal{L}^d_{\text{MMD}}$	−1.19 (−1.11/−1.28)	−1.57 (−1.39/−1.68)	−0.74 (−0.84/−0.66)	−1.50 (−1.49/−1.49)
w/o $\mathcal{L}^w_{\text{MMD}}$	−1.52 (−1.60/−1.40)	−1.85 (−1.79/−1.54)	−1.23 (−0.91/−1.55)	−1.45 (−1.35/−1.59)
w/o $\mathcal{L}_{\text{distill}}$	−1.88 (**−2.26**/−1.36)	−1.82 (−1.86/−1.18)	−1.58 (−1.40/−1.77)	−1.41 (−1.54/−1.22)
w/o $\mathbf{D}^{aug}$	**−1.90** (−2.05/**−1.70**)	**−2.11** (**−1.92**/**−2.13**)	**−1.93** (**−1.86**/**−2.00**)	**−1.66** (**−1.57**/**−1.76**)

The minus number indicates performance drops (in percentage) after removing or replacing the methods. (w/o $\mathcal{L}^c$): the removal of robust feature extraction by Equation (12). (w/o $\mathcal{L}^d_{\text{MMD}}$) and (w/o $\mathcal{L}^w_{\text{MMD}}$): the removal of the sentence-level and word-level MMD loss in Equation (14). (w/o $\mathcal{L}_{\text{distill}}$): the removal of progressive knowledge distillation loss in Equation (17).

challenge of NER. All models fail on the first sentence. The mask words can be a school (tagged with "ORG") or a specific place (tagged with "LOC") to learn how to drive. Most of the learners think they refer to an organization while the distance supervised tagger assigns them location labels. From the second sample, most of the models have the ability to undercover the correct knowledge that was missed by the ground truth. Our model shows the advantage of understanding fine-grained parts in a long-term entity mention, like the example in Sentence 3. The reason is that the anchor-driven data augmentation makes the model aware of the small segments. For the English group, ours can outperform the baselines in Sentence 4, correct the ground truth in Sentence 5, or behavior the same with the ground truth in Sentence 6.

6 RELATED WORK

Recently, label sparsity is under intensive exploration and has obtained remarkable success in many research frontiers [22, 23, 47–49, 53–55]. In its literature, cross-domain transfer plays a great role when dealing with the domain shift problem. Most of existing approaches in cross-domain NER are summarized into two categories by how they solve the domain shift issue.

The first group of approaches is to address the word discrepancy across different domains. Word-level discrepancy means that word distributions are not compatible across different domain datasets. This work [15] devises distributed word embedding methods to adopt and aggregate domain-specific knowledge. Therefore, the domain-specific knowledge works to boost the cross-domain NER performance. Some researchers [45] introduce label-aware MMD to solve domain shift by shared knowledge located in the words assigned with the same labels across domains. Others [20] introduce a simple projecting function on the word level to map target domain words into source domain word space.

Table 6. Ablation Study on Each NER Type

Baselines	PER	LOC	ORG	MISC
	F1 (Pre/Rec)	F1 (Pre/Rec)	F1 (Pre/Rec)	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)
w/o $\mathbf{D}^{aug}$	−0.36 (−0.34/−0.39)	−1.38 (−1.81/−0.65)	−2.38 (−2.81/−1.49)	−1.69 (−2.00/−1.23)
w/o $\mathcal{L}_{\text{MMD}}$	−0.28 (−0.50/−0.07)	−1.41 (−1.51/−1.21)	−1.08 (−1.16/−0.88)	−1.40 (−1.05/−1.88)
w/o $\mathcal{L}_{\text{distill}}$	−0.58 (−0.48/−0.69)	−1.66 (−2.02/−1.03)	−1.83 (−1.93/−1.57)	−1.24 (−1.28/−1.17)
WNUT 2016	86.27 (84.51/88.12)	48.57 (44.33/53.71)	46.81 (41.11/54.36)	27.90 (21.48/39.79)
w/o $\mathbf{D}^{aug}$	−0.72 (−0.67/−0.78)	−1.27 (−1.06/−1.56)	−1.48 (−1.51/−1.35)	−2.07 (−1.62/−2.88)
w/o $\mathcal{L}_{\text{MMD}}$	−0.53 (−0.52/−0.53)	−1.11 (−1.36/−0.71)	−1.47 (−1.33/−1.64)	−1.04 (−0.74/−1.70)
w/o $\mathcal{L}_{\text{distill}}$	−1.15 (−1.23/−1.08)	−1.59 (−1.88/−1.12)	−1.26 (−1.05/−1.57)	−2.03 (−1.93/−1.56)
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)
w/o $\mathbf{D}^{aug}$	−1.72 (−2.01/−1.40)	−2.57 (−2.27/−2.97)	−1.40 (−1.18/−1.68)	−1.73 (−1.45/−2.10)
w/o $\mathcal{L}_{\text{MMD}}$	−0.43 (−0.51/−0.34)	−0.79 (−0.62/−1.03)	−0.43 (−0.35/−0.53)	−0.65 (−0.49/−0.87)
w/o $\mathcal{L}_{\text{distill}}$	−1.02 (−1.23/−0.78)	−0.97 (−1.04/−0.84)	−1.27 (−1.43/−1.06)	−1.36 (−1.31/−1.40)
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)
w/o $\mathbf{D}^{aug}$	−1.09 (−1.22/−0.97)	−1.95 (−2.44/−1.15)	−2.06 (−2.08/−1.85)	−1.89 (−1.76/−1.99)
w/o $\mathcal{L}_{\text{MMD}}$	−0.84 (−0.78/−0.90)	−1.45 (−1.63/−1.13)	−1.56 (−1.35/−1.85)	−1.63 (−1.64/−1.49)
w/o $\mathcal{L}_{\text{distill}}$	−1.15 (−1.43/−1.28)	−1.47 (−1.61/−1.22)	−1.35 (−1.22/−1.52)	−1.83 (−1.84/−1.64)

The other group is for alleviating the sentence-level discrepancy. Because sentences from different domains vary in the patterns, written styles, publication categories, data quality, and so on. Approaches in this group include multi-level adaptation layers [20], tensor decomposition [11], and multi-task learning with external information [1, 24]. This work [20] devises sentence-adaptation component, taking in the pre-adapted word embedding to obtain adaptive sentence features. Besides, researchers [11] design multi-task learning strategy, performing tensor decomposition to transfer cross-domain NER knowledge. Moreover, some explorers [24] make NER labels as the experts to educate model learning between domains. Others [12] propose a multi-cell compositional LSTM structure for cross-domain NER. Besides, those [4, 19, 41] seek auxiliary knowledge from external resources. Through the external resource, model can generate pseudo labels for the low-resource domain to solve the labeled data insufficiency.

However, those methods impede their performances under both zero-resource and minimal-resource scenarios, because of a lack of robust and adaptive features or sufficient labeled target data for fine-tuning. The external resource assisted models inevitably introduce too much noise.

7 CONCLUSION

In this article, we propose an adversarial assisted progressive adaptation knowledge distillation framework, including anchor-guided and adversarial qualified 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

Sentence 1	星期天的早晨七点学车，驾 校 太给力
Ground Truth	B-LOC I-LOC
Baseline	B-ORG I-ORG
Ours	B-ORG I-ORG
Sentence 2	不是像，而是就是 陵 猫 猫 哈哈
Ground Truth	O O O
Baseline	B-PER I-PER I-PER
Ours	B-PER I-PER I-PER
Sentence 3	女 主 播 陵 猫 猫 这么像你呢，哈哈
Ground Truth	B-PER I-PER I-PER B-PER I-PER I-PER
Baseline	B-PER I-PER I-PER I-PER I-PER I-PER
Ours	B-PER I-PER I-PER B-PER I-PER I-PER
Sentence 4	订购一台 小 新 Air 系 列 的笔记本，A面喷绘 怪 杰 面 具 。
Ground Truth	B-MISC I-MISC I-MISC I-MISC I-MISC B-MISC I-MISC I-MISC I-MISC
Baseline	B-MISC I-MISC I-MISC I-MISC I-MISC B-PER I-PER O O
Ours	B-MISC I-MISC I-MISC I-MISC I-MISC O O O O
Sentence 5	Riots on 17th and Page for what
Ground Truth	B-LOC B-LOC
Baseline	B-LOC O
Ours	B-LOC B-LOC
Sentence 6	The second car of the West Jersey train was also carried into the ditch
Ground Truth	B-LOC I-LOC O
Baseline	B-MISC I-MISC I-MISC
Ours	B-MISC I-MISC I-MISC
Sentence 7	the Natural Resources Defense Council, a New York City-based non-profit organization.
Ground Truth	B-ORG I-ORG I-ORG I-ORG B-LOC I-LOC I-LOC
Baseline	B-ORG I-ORG I-ORG I-ORG B-LOC I-LOC I-LOC
Ours	B-ORG I-ORG I-ORG I-ORG B-LOC I-LOC I-LOC
Sentence 8	Order a Xiaoxin Air laptop with the Geek Mask imprinted on the A-side.
Ground Truth	B-MISC I-MISC B-MISC I-MISC
Baseline	B-MISC I-MISC O O
Ours	B-MISC I-MISC O O

Fig. 4. Case Study on Both English and Chinese sentences. For each sample sentence, a comparison performs on the ground truth and the predictions of the best baseline model and PDALN†. Words under the green mask in each sentence are assigned NER tags under the human judgments to evaluate the quality of the ground truth. The prediction in red means incorrect assignments while the green is correct ones.

art. The most merit in this model is valuable domain-invariant feature extraction by the effective training framework from the well-denoised augmented mix-domain data. We expect and explore further ideas to construct more advanced and effective adaptive data or to study a more expressive and powerful adaptation space rather than an adversarial training strategy.

ACKNOWLEDGMENTS

We thank the reviewers for their valuable comments.

REFERENCES

[1] 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*. 148–153.

[2] 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)*. 10–18.

[3] 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 (2010), 151–175.

[4] 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*. Association for Computational Linguistics, 261–270.

[5] 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*. PMLR, 1597–1607.

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

[7] Hady Elsahar and Matthias Gallé. 2019. To annotate or not? Predicting performance drop under domain shift. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2163–2173.

[8] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*. PMLR, 1180–1189.

[9] 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)*, Vol. 2. IEEE, 1735–1742.

[10] Hangfeng He and Xu Sun. 2017. A unified model for cross-domain and semi-supervised named entity recognition in Chinese social media. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31. 3216–3222.

[11] 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*. 2464–2474.

[12] 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*. Association for Computational Linguistics, 5906–5917. https://www.aclweb.org/anthology/2020.acl-main.524.

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

[14] 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)*. 247–253.

[15] 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* (2016).

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

[17] 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)*. 4470–4473.

[18] Gina-Anne Levow. 2006. The third international Chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*. 108–117.

[19] 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*. 1054–1064.

[20] 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*. 2012–2022.

[21] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. 2021. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering* (2021).

[22] 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*. 1608–1612.

[23] 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)*. IEEE, 850–859.

[24] Zihan Liu, Genta Indra Winata, and Pascale Fung. 2020. Zero-resource cross-domain named entity recognition. In *Proceedings of the 5th Workshop on Representation Learning for NLP*. Association for Computational Linguistics, 1–6.

[25] Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2021. CrossNER: Evaluating cross-domain named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 13452–13460.

[26] Jin Kiat Low, Hwee Tou Ng, and Wenyuan Guo. 2005. A maximum entropy approach to Chinese word segmentation. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*. 161–164.

[27] Xiaofei Ma, Peng Xu, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2019. Domain adaptation with BERT-based domain classification and data selection. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*. 76–83.

[28] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. 1003–1011.

[29] 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 (2018), 1979–1993.

[30] Sinno Jialin Pan, Zhiqiang Toh, and Jian Su. 2013. Transfer joint embedding for cross-domain named entity recognition. *ACM Transactions on Information Systems (TOIS)* 31, 2 (2013), 1–27.

[31] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. 2018. Multi-adversarial domain adaptation. In *Thirty-second AAAI Conference on Artificial Intelligence*. 3934–3941.

[32] Hao Peng, Haoran Li, Yangqiu Song, Vincent Zheng, and Jianxin Li. 2021. Differentially private federated knowledge graphs embedding. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (Virtual Event, Queensland, Australia) *(CIKM '21)*. Association for Computing Machinery, 1416–1425.

[33] Hao Peng, Jianxin Li, Yangqiu Song, Renyu Yang, Rajiv Ranjan, Philip S. Yu, and Lifang He. 2021. Streaming social event detection and evolution discovery in heterogeneous information networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 15, 5 (2021), 1–33.

[34] Hao Peng, Ruitong Zhang, Shaoning Li, Yuwei Cao, Shirui Pan, and Philip Yu. 2022. Reinforced, incremental and cross-lingual event detection from social messages. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022), 1–1.

[35] Nanyun Peng and Mark Dredze. 2015. Named entity recognition for Chinese social media with jointly trained embeddings. In *Processings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 548–554.

[36] Piyush Rai, Avishek Saha, Hal Daumé III, and Suresh Venkatasubramanian. 2010. Domain adaptation meets active learning. In *Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*. 27–32.

[37] 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)*. 147–155.

[38] Adwait Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In *Conference on Empirical Methods in Natural Language Processing*.

[39] Nils Rethmeier and Isabelle Augenstein. 2021. A primer on contrastive pretraining in language processing: Methods, lessons learned & perspectives. *ACM Computing Surveys (CSUR)* (2021).

[40] Erik F. Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050* (2003).

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

[42] 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)*. 138–144.

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

[44] 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* (2017).

[45] 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)*. 1–15.

[46] Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 6382–6388.

[47] 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*. 3379–3388.

[48] 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*. 1351–1360.

[49] Congying Xia, Chenwei Zhang, Xiaohui Yan, Yi Chang, and Philip Yu. 2018. Zero-shot user intent detection via capsule neural networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 3090–3099.

[50] Yiying Yang, Xi Yin, Haiqin Yang, Xingjian Fei, Hao Peng, Kaijie Zhou, Kunfeng Lai, and Jianping Shen. 2021. KGSynNet: A novel entity synonyms discovery framework with knowledge graph. In *Proceedings of the Database Systems for Advanced Applications: 26th International Conference, DASFAA*. 174–190.

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

[52] 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)*. 7386–7399.

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

[54] Jianguo Zhang, Kazuma Hashimoto, Chien-Sheng Wu, Yao Wang, S. Yu Philip, Richard Socher, and Caiming Xiong. 2020. 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*. 154–167.

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

[56] Tao Zhang, Congying Xia, Philip S. Yu, Zhiwei Liu, and Shu Zhao. 2021. PDALN: Progressive domain adaptation over a pre-trained model for low-resource cross-domain named entity recognition. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 5441–5451. https://doi.org/10.18653/v1/2021.emnlp-main.442.

[57] 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*. 3461–3471.

Received 24 March 2022; revised 24 September 2022; accepted 24 October 2022