PRAM: An End-to-end Prototype-based Representation Alignment Model for Zero-resource Cross-lingual Named Entity Recognition

Yucheng Huang[2,3†*] **Wenqiang Liu**[1‡†] **Xianli Zhang**[1] **Jun Lang**[1]
Tieliang Gong[2,3] **Chen Li**[2,3‡]
[1]Interactive Entertainment Group, Tencent Inc., Shenzhen, China
[2]School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, China
[3]Shaanxi Provincial Key Laboratory of Big Data Knowledge Engineering, Xi'an, China
huangyucheng@stu.xjtu.edu.cn, {gongtl, cli}@xjtu.edu.cn
{masonqliu, ryanxzhang, bruceang}@tencent.com

Abstract

Zero-resource cross-lingual named entity recognition (ZRCL-NER) aims to leverage rich labeled source language data to address the NER problem in the zero-resource target language. Existing methods are built either based on data transfer or representation transfer. However, the former usually leads to additional computation costs, and the latter lacks explicit optimization specific to the NER task. To overcome the above limitations, we propose a novel prototype-based representation alignment model (PRAM) for the challenging ZRCL-NER task. PRAM models the cross-lingual (CL) NER task and transfers knowledge from source languages to target languages in a unified neural network, and performs end-to-end training, avoiding additional computation costs. Moreover, PRAM borrows the CL inference ability of multilingual language models and enhances it with a novel training objective—attribution-prediction consistency (APC)—for explicitly enforcing the entity-level alignment between entity representations and predictions, as well as that across languages using prototypes as bridges. The experimental results show that PRAM significantly outperforms existing state-of-the-art methods, especially in some challenging scenarios.

1 Introduction

Named Entity Recognition (NER) aims to identify the boundaries and categories of entities in a chunk of text (Tjong Kim Sang, 2002). Automatic NER is useful for various downstream applications, such as search engines (Cowan et al., 2015), dialogue systems (Bowden et al., 2018), and knowledge graphs (Al-Moslmi et al., 2020). Most of the recent advances in NER are achieved by deep neural networks that are trained on a large amount of labeled data (Lample et al., 2016; Chiu and Nichols, 2016; Li et al., 2020; Yu et al., 2020). However, it is not universal that every kind of language has sufficient labeled data for training a deep NER model. This motivates research on a new challenging task named zero-resource cross-lingual NER (ZRCL-NER), which aims to leverage a rich labeled source language to address the NER problem in an unlabeled target language.

Most advanced methods for the challenging task have concentrated on transferring knowledge from the rich-resource language (RRL) to the zero-resource language (ZRL). According to the type of transferred knowledge, these methods can be divided into two categories, *i.e.*, data transfer based methods and representation transfer based methods. The data transfer based methods transfer the knowledge from RRL to the ZRL by generating pseudo-labeled data (Mayhew et al., 2017; Jain et al., 2019; Zhou et al., 2022) or soft labels for unlabeled data (Chen et al., 2021; Liang et al., 2021; Li et al., 2022) in the ZRL via translation or teacher models, respectively. However, they follow a two-separated training setup that needs extra translation or teacher models, introducing additional computation costs. In contrast, the representation transfer based methods directly model the cross-lingual NER in a unified model, borrowing the cross-lingual inference ability of the multilingual pre-trained language models (mPLMs) (Devlin et al., 2019; Conneau and Lample, 2019; Conneau et al., 2020). To further enhance the cross-lingual inference ability, they enforce the mPLMs to learn language-independent features via representation alignment *w.r.t.* token (Kulshreshtha et al., 2020; Muller et al., 2021) or cross-lingual antagonistic training *w.r.t.* language (Keung et al., 2019). These optimizations are implicit for the NER task and thus show limited improvements on the ZRCL-NER.

To overcome the above limitations, we propose a novel Prototype-based Representation Alignment

*Work done during an internship at Tencent.
†Equal Contribution
‡Corresponding author

Findings of the Association for Computational Linguistics: ACL 2023, pages 3220–3233
July 9-14, 2023 ©2023 Association for Computational Linguistics

Model (PRAM) for ZRCL-NER. Our PRAM builds upon the mPLM that falls into the representation transfer based framework, and performs end-to-end training that avoids tedious training stages. This prevents excess computational costs that is not necessary for improving a model's performance on the target language. We explicitly enforce entity-level alignment between representations and predictions in both source and target languages by a new training objective called Attribution-Prediction Consistency (APC). By treating prototypes as category anchors, we can enforce the alignment of similar entity representations across languages by guiding similarity distributions of entity representations relative to prototypes through predictions. Specifically, we treat the centroid of entity representations belonging to the same class in the source language as a prototype. The APC minimizes JS divergence between predicted probability distribution (model output for each entity mention) and similarity distribution (similarities between each mention representation and all prototypes). In this way, with the predicted probability distribution becoming more discriminative in the training, the APC can impose the alignment between the entity representation and its corresponding prototype. The alignment property in the representation space, in turn, can regularize the deviation of model predictions. Probability distribution changes in source languages lead to similar changes in target languages due to mPLM's cross-lingual inference ability. Thus, the similar entity representations of the source and target languages are clustered with prototypes and this alignment can enhance cross-lingual inference.

To explore the performance of PRAM, we conduct experiments on three NER datasets under both single- and multi-source settings. The experimental results show that PRAM outperforms all state-of-the-arts (SOTAs), especially in some more challenging scenarios. Specifically, in the setting of single source transfer on the WikiAnn dataset, PRAM achieves 3.92% absolute improvements on F1-score compared with SOTA. On the other hand, PRAM achieves an average improvement of 3.88% on the F1 score compared with SOTA in the multi-source setting. Moreover, we demonstrate the APC is able to impose entity-level alignment across languages by visualizing entity representations of the same class from different languages.

The contribution of this work is three-fold:

- We propose PRAM, a novel prototype-based representation alignment model for the challenging ZRCL-NER task. The PRAM model transfers knowledge from RRL to ZRL and models the cross-lingual NER task in a unified model without tedious training steps and additional computation costs.
- To enforce the entity-level alignment between entity representations and predictions, as well as that across languages, we propose a new training objective named APC. The APC borrows the cross-lingual inference ability of mPLM and enhances it dynamically in the training process, which benefits the ZRCL-NER task.
- We conduct exhaustive experiments on three datasets under single- and multi-source settings. The experimental results show that PRAM outperforms a wide range of SOTAs by a significant margin. Moreover, we show the entity-level cross-lingual alignment property induced by APC by representation dimensional reduction visualization.

2 Related Work

2.1 Cross-lingual Named Entity Recognition

Cross-lingual NER can be roughly categorized into data transfer based and representation transfer based. The former aims to train a target-language NER model with pseudo labeled data constructed from the labeled source-language data. Specifically, translation-based data transfer constructs pseudo-labeled data by translating texts and projecting labels from the source to the target language (Jain et al., 2019; Qin et al., 2021; Zhou et al., 2022). Knowledge-distillation-based data transfer generates soft labels for unlabeled target data with the trained teacher model on the source language (Wu et al., 2020a; Gou et al., 2021; Li et al., 2022). Such methods typically follow a two-separate training setup and need additional computing costs for the training of the translation models or the teacher models. The latter aims to model the cross-lingual NER in a unified model with the help of mPLMs (Pires et al., 2019). To further exploit the language-independent features, some cross-lingual representation transfer techniques are introduced, including word-word alignment (Wang et al., 2020), adversarial training (Keung et al., 2019), and meta learning (Wu et al., 2020b), etc. These optimizations are implicit for the NER task and fail to learn

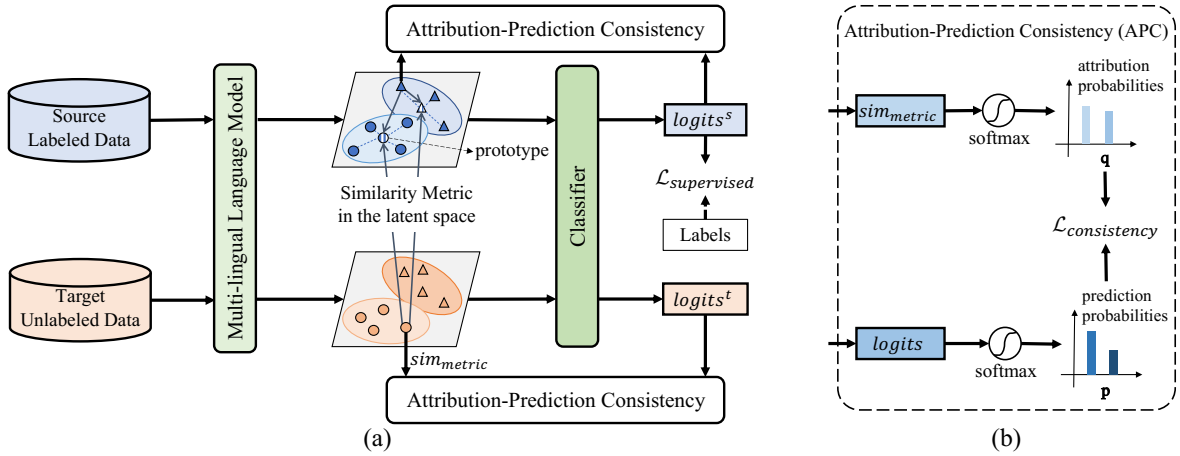

Figure 1: (a) The illustration of PRAM based on a prototype-based representation alignment in an end-to-end manner; (b) The process of Attribution-Prediction Consistency (APC).

the task-specific information, limiting their applicability on the ZRCL-NER.

2.2 Multilingual Representation Learning

Multilingual representation learning aims to create multilingual representations of different languages in a unified semantic space that can be used for various tasks across languages. With pre-trained from monolingual corpora in 104 languages, Multilingual BERT (M-BERT) (Devlin et al., 2019) can encode multilingual representations and works well on zero-resource cross-lingual transfer tasks, which motivates more work on multilingual pre-trained language models (Conneau and Lample, 2019; Conneau et al., 2020; Ouyang et al., 2021). Cao et al. (2020) demonstrate that the cross-lingual transfer ability of M-BERT is mainly due to the partial alignment of its cross-lingual representations, and such transfer ability can be further enhanced by introducing effective alignment procedures. Motivated by this, several representation alignment approaches have been proposed to improve the cross-lingual transfer in downstream tasks. Wang et al. (2019) introduce a bilingual projection of the contextual representations based on word alignment trained on parallel data for cross-lingual dependency parsing. Wang et al. (2021) align the token representations from different languages via adversarial domain adaptation to efficiently apply M-BERT in cross-lingual information retrieval. Rotational alignment (Wang et al., 2020; Kulshreshtha et al., 2020) and adversarial training (Keung et al., 2019; Bari et al., 2020) are applied in the ZRCL-NER for representation alignment. Still, these techniques need to construct parallel corpus or train additional discriminators, which require expensive labor or computing costs.

3 Methodology

This section introduces our prototype-based representation alignment model (PRAM), as illustrated in Figure 1. First, the ZRCL-NER task is described formally. Then, we introduce the basic cross-lingual NER model, which we need to remold. Subsequently, our APC module is elaborated. Finally, we describe how to train PRAM in single- and multi-source settings.

3.1 Problem Definition

The zero-resource cross-lingual NER can be formulated as a sequence labeling problem. Given a sentence $\mathbf{x} = \{x_i\}_{i=1}^{L}$ with L tokens, a NER model aims to produce a label sequence $\mathbf{y} = \{y_i\}_{i=1}^{L}$, where y_i is the inferred label of the corresponding token x_i. In the source language, the annotated training data is denoted by $\mathcal{D}^s_{train} = \{(\mathbf{x}, \mathbf{y})\}$. In the target language, only the unlabeled data denoted by $\mathcal{D}^t_u = \{\mathbf{x}\}$ is available in the training process and a small set of labeled data denoted by $\mathcal{D}^t_{test} = \{(\mathbf{x}, \mathbf{y})\}$ is left for evaluation. Formally, ZRCL-NER aims to learn a model based on $\mathcal{D}^s_{train}$ and $\mathcal{D}^t_u$ that can perform well on $\mathcal{D}^t_{test}$.

3.2 The Basic Cross-lingual NER Model

The basic cross-lingual NER model is built by adding a linear classification layer upon a mPLM, which can be formulated as:

$$\mathbf{h} = \texttt{mPLM}(\mathbf{x}), \tag{1}$$

$$\mathbf{p}_i = \texttt{softmax}(\mathbf{W}\mathbf{h}_i + b), \tag{2}$$

where $\mathbf{x} = \{x_i\}_{i=1}^{L}$ is a chunk of text with a length of L, $\mathbf{h} = \{\mathbf{h}_i\}_{i=1}^{L}$ is the list of token representations, and $\mathbf{p}_i$ is the predicted probability distribution for token x_i.

For all samples in $\mathcal{D}_{train}^{s}$, the loss function of the supervised learning is:

$$\mathcal{L}_{sup}(\theta) = -\frac{1}{NL}\sum_{i=1}^{N}\sum_{j=1}^{L} y_{i,j}\log p_{i,j}, \quad (3)$$

where N is the training data size. It is worth noting that only the source language is annotated with ground-truth labels in the training process.

3.3 The Attribution-Prediction Consistency Constraint

Without additional constraints, the basic cross-lingual NER model may be dominated by the source language, which may hinder its generalization in the target language. Such an issue is caused by the different entity representation distribution between the source and the target language (Libovický et al., 2019). And during the training, the distribution difference may become even more significant. To alleviate this issue, we propose a new optimizing objective named attribution-prediction-consistency (APC) that can enforce entity-level alignment across languages. The APC is equipped for both source and target languages, which aims to optimize the consistency between the predicted probability distribution and the similarity distribution between each representation and all source prototypes. Cooperated with the initial cross-lingual inference ability of mPLM, the APC can progressively enforce entity-level alignment across languages as the training goes on.

Firstly, we feed both the labeled data of the source language and the unlabeled data of the target language into the mPLM encoder within each batch. For each entity class, its prototype is obtained by averaging all its inclusion entity representations on the source language, where the representations are extracted according to Eq.(1). This process can be formulated as follow:

$$\mathcal{C}_k = \frac{1}{n_k}\sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}_{train}^{S}}\sum_{i=0}^{L}\mathbb{1}[y_i = k]\mathbf{h}_i, \quad (4)$$

where $\mathbb{1}[\cdot]$ is an indicator function, k represents an entity class label, and n_k denotes the number of tokens belonging to class k in the source language. In practice, the prototypes are generated from mini-batch instead of all samples to reduce computational costs. To further ensure the stability of updates, the moving average method (Xie et al., 2018) is adopted to update the prototypes:

$$\mathcal{C}_k = \lambda * \mathcal{C}_k + (1-\lambda) * \mathcal{C}_k', \quad (5)$$

where $\mathcal{C}_k'$ denotes the prototype of class k calculated from the previous moment, and $\lambda \in (0, 1)$ is the moving average coefficient.

Subsequently, we can obtain the similarity distribution between each token and all prototypes. For each token x_i, we calculate the cosine similarity between its representation $\mathbf{h}_i$ and prototype $\mathcal{C}_k$:

$$s(\mathbf{h}_i, \mathcal{C}_k) = cos(\mathbf{h}_i, \mathcal{C}_k) = \frac{\mathbf{h}_i \cdot \mathcal{C}_k}{||\mathbf{h}_i|| \cdot ||\mathcal{C}_k||}. \quad (6)$$

The similarity distribution $\mathbf{q}_i = \{q_{(i,k)}\}_{k=1}^{K}$ is produced by applying a softmax function with a temperature coefficient τ (Chen et al., 2020) on the cosine similarity score distribution:

$$q_{(i,k)} = \frac{exp(s(\mathbf{h}_i, \mathcal{C}_k))/\tau}{\sum_i^K exp(s(\mathbf{h}_i, \mathcal{C}_i))/\tau}. \quad (7)$$

Finally, the APC improves the consistency between the similarity distribution $\mathbf{q}_i$ and the predicted probability distribution $\mathbf{p}_i$ by minimizing their Jensen-Shannon divergence. Mathematically, this process can be formulated as:

$$\begin{aligned}\mathcal{L}_c(\theta) &= \frac{1}{NL}\sum_{i=0}^{N}\sum_{j=0}^{L}\texttt{JsDiv}(\mathbf{q}_{i,j}||\mathbf{p}_{i,j}),\\ \texttt{JsDiv}(\mathbf{q}||\mathbf{p}) &= \frac{1}{2}\sum \mathbf{p}\log(\frac{2\mathbf{p}}{\mathbf{p}+\mathbf{q}})\\ &\quad + \frac{1}{2}\sum \mathbf{q}\log(\frac{2\mathbf{q}}{\mathbf{p}+\mathbf{q}}).\end{aligned} \quad (8)$$

3.4 The Prototype-based Representation Alignment Model for Single-source Setting

Our PRAM is built by combining the basic cross-lingual NER model and the APC module. The total loss of the PRAM is

$$\mathcal{L}(\theta) = \mathcal{L}_{sup}(\theta) + \alpha\mathcal{L}_c^s(\theta) + \beta\mathcal{L}_c^t(\theta), \quad (9)$$

where α and β are the balancing weights for the source and target language, respectively.

Discussion: The first term $\mathcal{L}_{sup}$ in Eq.(9) leverages the supervision signals in the source language to learn the task-specific semantics. As the training goes on, the predicted probability distribution

Algorithm 1 Overall training process of PRAM

Require: $\mathcal{D}^s_{train}$: training data in the source language; $\mathcal{D}^t_u$: unlabeled data in the target language; θ_{enc}: parameters of the multilingual word encoder; θ_{cls}: parameters of the classifier; T_{steps}: the maximum steps for training.

Initialize θ_{enc} and θ_{cls}
iter=0
while iter < T_{steps} **do**
  Sample a mini-batch $\mathcal{B}$ with $(\mathbf{x}, \mathbf{y}) \in \mathcal{D}^s_{train}$ and $(\mathbf{x}) \in \mathcal{D}^t_u$
  for all samples in $\mathcal{B}$ **do**
    $\mathbf{h}_i \leftarrow f(\theta_{enc}, \mathbf{x}_i)$
    $\mathbf{p}_i \leftarrow f(\theta_{cls}, \mathbf{h}_i)$
  end for
  # supervised learning
  Calculate $\mathcal{L}_{sup}(\theta)$ on $(\mathbf{x}, \mathbf{y}) \in \mathcal{D}^s_{train}$ as in Eq.(3)
  # attribution-prediction consistency (APC)
  Obtain the class prototypes C as in Eq.(4) and (5)
  for all samples in $\mathcal{B}$ **do**
    Produce $\mathbf{q}_i$ as in Eq.(6) and (7)
  end for
  Calculate $\mathcal{L}^s_c(\theta)$ and $\mathcal{L}^t_c(\theta)$ as in Eq.(8)
  # the total loss for training
  $\mathcal{L}(\theta) \leftarrow \mathcal{L}_{sup}(\theta) + \mathcal{L}^s_c(\theta) + \mathcal{L}^t_c(\theta)$
  Update θ_{enc} and θ_{cls} via gradient back-propagation
  iter += 1
end while

becomes more discriminative in both the source and target languages. This is further leveraged by the APC, *i.e.*, $\mathcal{L}^s_c$ and $\mathcal{L}^t_c$, to impose the similarity distribution to be consistent with the probability distribution. Because the similarity distribution (in both source and target languages) measures the similarity between each entity representation and all prototypes (produced based on source language), the APC indirectly enforces the entity-level alignment between the source and target languages. The alignment property in the representation space, in turn, can revise the deviation of model predictions.

Algorithm 1 shows the pseudocode for the overall training process of PRAM.

3.5 Extend PRAM to Multi-source Setting

PRAM can be easily extended to meet the multi-source scenario. Assuming that there are n source languages, we can construct n sets of prototypes, *i.e.*, one set for one source language according to Eq.(4) and Eq.(5). For each sample in the source languages, we obtain its similarity distribution according to the prototypes of the corresponding language. While, for each sample in the target languages, we obtain n similarity distributions according to different sets of prototypes of different source languages. Figure 2 provides a visual illustration of how to derive the similarity distributions in the multi-source setting. The total loss in Eq.(9)

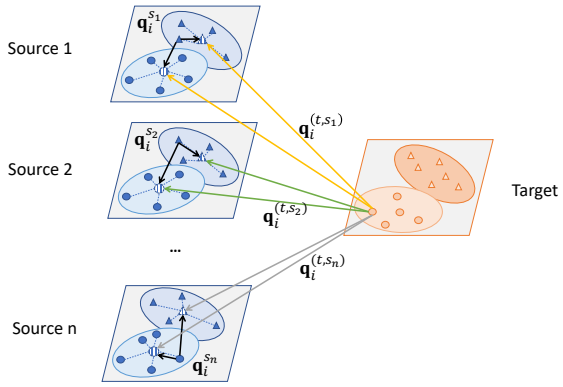

Figure 2: The similarity distributions produced in the multi-source cross-lingual transfer.

for training PRAM can be rewritten as

$$\mathcal{L}(\theta) = \mathcal{L}_{sup}(\theta) + \alpha \sum_{i=1}^{n} \mathcal{L}^{s_i}_c(\theta) + \beta \sum_{i=1}^{n} \mathcal{L}^{(t,s_i)}_c(\theta), \tag{10}$$

where $\mathcal{L}^{s_i}_c(\theta)$ denotes the consistency loss of the i-th source language, and $\mathcal{L}^{(t,s_i)}_c(\theta)$ denotes the consistency loss of samples in the target language *w.r.t.* the prototypes in i-th source language.

4 Experiment

We evaluate PRAM in both single-source and multi-source transfer settings, and compare it with state-of-the-art models. Moreover, an ablation study and several analytical experiments are conducted to demonstrate the effectiveness of our model.

4.1 Datasets and Experiment Settings

There are three benchmark datasets included in our experiments, which are CoNLL-2002 (Tjong Kim Sang, 2002), CoNLL-2003 (Tjong Kim Sang and De Meulder, 2003) and WikiAnn (Pan et al., 2017; Rahimi et al., 2019). All datasets are labeled using the BIO scheme with four entity types, which are persons (PER), locations (LOC), organizations (ORG), and miscellaneous (MISC). Each dataset is split into training/development/test sets the same as initially published. The dataset statistics are listed in Table 1.

In the single-source transfer, we treat English as the source language and the others as the target languages for both CoNLL-2002/2003 and WikiAnn. In the multi-source transfer, we follow the previous work (Wu et al., 2020a) that selects source languages in a leave-one-out manner, which means all

(a) CoNLL statistics

Language	Type	Train	Dev	Test
English-en (CoNLL-2003)	Sentence	14987	3466	3684
	Entity	23499	5942	5648
German-de (CoNLL-2003)	Sentence	8323	1915	1517
	Entity	18798	4315	3558
Spanish-es (CoNLL-2002)	Sentence	15806	2895	5195
	Entity	13344	2616	3941
Dutch-nl (CoNLL-2002)	Sentence	12705	3068	3160
	Entity	11851	4833	3673

(b) WikiAnn statistics

Language	Type	Train	Dev	Test
English-en	Sentence	20000	10000	10000
	Entity	27931	14146	13958
Arabic-ar	Sentence	20000	10000	10000
	Entity	22500	11266	11925
Hindi-hi	Sentence	5000	1000	1000
	Entity	6124	1226	1228
Chinese-zh	Sentence	20000	10000	10000
	Entity	25031	12493	12532

Table 1: Statistics of the datasets

languages are treated as source ones except for the target one. Following previous works, the training set in the target language of CoNLL-2002/2003 is used as unlabeled target data $\mathcal{D}_u^t$ after removing labels. For WikiAnn, the training set and extra set [1] with removed labels are selected as the unlabeled target data.

4.2 Implementation Details

The cross-lingual encoder is initialized with the parameters of the cased M-BERT$_{\text{base}}$ released by HuggingFace Transformers [2]. Following the previous work (Wu et al., 2020b), the parameters of the embedding layer and the bottom three layers of M-BERT are frozen. We only consider the first subword in the loss function if a word is tokenized into several subwords by word piece. To avoid the adverse effect on the cross-lingual transfer performance led by excessive non-entity, we adopt a non-entity down-sampling strategy (Li et al., 2021) as described in Appendix A. And the balancing rate γ is set with {1.0, 1.5, 2.0} in different cases.

For all experiments, we use the AdamW optimizer with a learning rate of 1e-5 for training. The batch size and the maximum sequence length are both set to 128 empirically. The early stopping strategy is adopted and the maximum training step is set to 20000. Additionally, we use the grid search technique for other hyper-parameters to obtain the optimal ones, including the temperature coefficient τ selected from 0.15 to 0.25, the moving average coefficient λ selected from 0.8 to 0.99, the loss weight α and β selected from 0.5 to 1.5 for CoNLL and from 1.0 to 3.0 for WikiANN.

We implement our approach with PyTorch 1.11.0 and all calculations are done on NVIDIA Tesla V100 GPU. The entity-level micro-F1 score is used as the evaluation metric. For all experiments, we report the average F1 scores over 5 runs with different random seeds.

4.3 Baselines

We compare our proposed method with the following SOTAs, including data transfer based methods and representation transfer based methods:

TMP (Jain et al., 2019) proposes a system that improves the entity-projection annotation by leveraging machine translation.

BERT-f (Wu and Dredze, 2019) fine-tunes the multilingual BERT in the source language and directly performs prediction in the target languages.

AdvCE (Keung et al., 2019) introduces language-adversarial training on the contextual representations for cross-lingual NER.

BERT-RA (Kulshreshtha et al., 2020) utilizes parallel corpora to supervise the rotation alignment of representations across different languages.

TSL (Wu et al., 2020a) proposes teacher-student learning to transfer task-specific knowledge from the source to the target language.

Unitrans (Wu et al., 2021) devises a pipeline to unify both model and data transfer for ZRCL-NER.

AdvPicker (Chen et al., 2021) designs an adversarial learning framework to select less language-dependent data in the target language to improve the ZRCL-NER performance.

RIKD (Liang et al., 2021) proposes a cross-lingual NER approach combining knowledge distillation and reinforcement learning.

MTMT (Li et al., 2022) introduces a similarity metric model based on knowledge distillation and multi-task learning for cross-lingual NER.

[1] https://github.com/afshinrahimi/mmner
[2] https://github.com/huggingface/transformers

Method	de	es	nl	Avg.
TMP	57.33	57.16	49.46	49.39
BERT-f	69.56	74.96	77.57	74.03
AdvCE	71.90	74.30	77.60	74.60
BERT-RA	70.48	75.84	79.52	75.28
TSL	73.16	76.75	80.44	76.78
Unitrans	73.61	81.20	77.30	77.37
AdvPicker	75.01	79.00	82.90	78.97
RIKD	75.48	77.84	82.46	78.59
MTMT	76.80	81.82	**83.41**	80.68
PRAM	**77.64**	**82.06**	83.15	**80.95**

Table 2: F1 Scores (%) on the CoNLL-2002/2003 dataset in the single-source setting. We bold the best performance and underline the second-best performance.

Method	ar	hi	zh	Avg.
BERT-f	42.30	67.60	52.90	54.27
TSL	43.12	69.54	48.12	53.59
RIKD	45.96	70.28	50.40	55.55
MTMT	52.77	70.76	52.26	58.60
PRAM	**57.44**	**74.67**	**55.46**	**62.52**

Table 3: F1 Scores (%) on the WikiAnn dataset in the single-source setting. We bold the best performance and underline the second-best performance.

4.4 Performance Comparison

Single-source Transfer: As shown in Table 2 and 3, PRAM convincingly outperforms previous SOTAs in most cases. Specifically, on the CoNLL-2002/2003 dataset (Table 2), PRAM achieves the best performance on German and Spanish and the second best on Dutch. Compared with the previous SOTA, PRAM improves the F1 score by 0.27% on average. On the WikiAnn dataset (Table 3), PRAM outperforms the previous SOTA by a large margin, with average F1-score improvements of 3.92% compared to MTMT (ranging from 3.20% for Chinese to 4.67% for Arabic). We can observe that PRAM achieves a more significant boost on the WikiAnn dataset, where the source (English) and the target languages (Arabic, Hindi, and Chinese) come from distinct language families. The huge difference between the source and target languages largely limits the transfer ability of previous methods, but PRAM still performs well in such challenging settings. Moreover, compared with the latest models, MTMT, RIKD, and AdvPicker, our method performs end-to-end, saving computational costs for training teacher models.

Method	de	es	nl	Avg.
BERT-f	73.86	76.93	80.12	76.97
TSL-avg	74.97	77.75	80.39	77.70
PRAM	**78.41**	**82.79**	**83.53**	**81.58**

Table 4: F1 Scores (%) on the CoNLL dataset in the multi-source setting. We bold the best performance.

(a) CoNLL datasets

Method	de	es	nl	Avg.
PRAM	77.64	82.06	83.15	80.95
$\text{PRAM}_{\text{w/o S-C}}$	76.62	80.63	81.57	79.61
$\text{PRAM}_{\text{w/o T-C}}$	72.38	75.91	78.56	75.62
$\text{PRAM}_{\text{w/o C}}$	70.93	74.72	77.74	74.46

(b) WikiAnn dataset

Method	ar	hi	zh	Avg.
PRAM	57.44	74.67	55.46	62.52
$\text{PRAM}_{\text{w/o S-C}}$	54.36	73.01	53.29	60.22
$\text{PRAM}_{\text{w/o T-C}}$	46.67	68.14	51.46	55.43
$\text{PRAM}_{\text{w/o C}}$	44.13	67.68	52.37	54.72

Table 5: Ablation Study

Multi-source Transfer: To in-line with the previous work (Wu et al., 2020a), on the CoNLL-2002/2003 dataset, we take German, Spanish, and Dutch as target languages. As shown in Table 4, PRAM obtains significant and consistent improvements on three target languages. Specifically, PRAM improves the F1 score by 4.61% on average compared to BERT-f and 3.88% compared to TSL-avg (TSL with averaging teacher models) (Wu et al., 2020a). Moreover, compared to the single-source setting, the cross-lingual performance of PRAM in the multi-source setting is consistently improved due to aligning target representations with the prototypes of multiple source languages.

4.5 Ablation Study

To investigate the contributions of different modules in PRAM, we conduct ablation experiments with three variants: 1) $\text{PRAM}_{\text{w/o S-C}}$ removes the APC on the source language; 2) $\text{PRAM}_{\text{w/o T-C}}$ removes the APC on the target language; 3) $\text{PRAM}_{\text{w/o C}}$ does not use any consistency strategy. As shown in Table 5, the performance of the three variants drops significantly. Compared

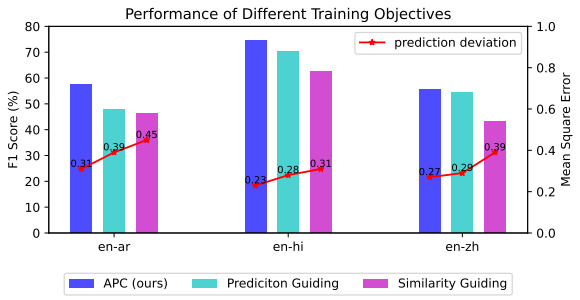

Figure 3: The performance (F1 score, bar charts) and the prediction deviations (MSE score, line charts) of different training objectives on the WikiAnn dataset.

to $\text{PRAM}_{\text{w/o S-C}}$, $\text{PRAM}_{\text{w/o T-C}}$ yields a more significant drop in the F1 scores. This indicates that the APC on the target language is more critical for improving the cross-lingual transfer ability because it helps to align the target representations with the prototypes produced in the source language. Without the APC constraints, the average F1 score of $\text{PRAM}_{\text{w/o C}}$ decreases by 6.49% on CoNLL and 7.80% on WikiAnn compared to PRAM due to lacking explicit representation optimization.

4.6 Effectiveness of Attribution-Prediction Consistency

To validate the effectiveness of the APC, we conduct experiments with two variants of this training objective: Prediction Guiding (PG) employs the predicted probability distributions to supervise the similarity distributions, stopping gradient backpropagation from the prediction path. Conversely, Similarity Guiding (SG) prevents gradient backpropagation from the similarity-metric path. As shown in Figure 3, the performance of the two variants shows a significant drop compared to the APC. The performance degradation of the SG is more significant than that of the PG, suggesting that the guidance from the predictions to the representation distributions plays a more important role. We also report the prediction deviation of the model with different strategies, which is defined as the mean square error between the predicted probability distributions and the one-hot labels: $e = \frac{1}{n}\sum(\mathbf{y}_i - \mathbf{p}_i)^2$. The APC achieves a lower prediction deviation than the PG. This demonstrates that the APC effectively improves model performance through the combined effect of aligning cross-lingual representations and using alignment properties to revise prediction deviation.

Languages	CoNLL	Languages	WikiAnn
en	91.45	en	85.21
en-de	91.55	en-ar	85.71
en-es	91.72	en-hi	85.83
en-nl	92.21	en-zh	85.56

Table 6: F1 Scores (%) on the test set of the source language in the single-source settings.

4.7 Representation Visualization

To demonstrate that PRAM can align similar entity representations across languages, we randomly select 150 samples per class from the source and the target languages and employ t-SNE (Van der Maaten and Hinton, 2008) to project their representations encoded by mPLM into a two-dimensional space. As shown in Figure 4, PRAM results in greater alignment of similar representations across languages compared to the baseline, *i.e.*, the basic cross-lingual NER model introduced in Section 3.2. When there is a huge difference between the source and the target language (from English to Arabic), many target representations from the baseline are mixed together and distributed differently from the similar source representations, which hinders the cross-lingual transfer. In contrast, PRAM significantly enhances the representation alignment across languages, especially in those classes where the baseline struggles (B-ORG, I-ORG, I-LOC, etc.). When the target language is similar to the source language (from English to German), PRAM can further optimize and align similar representations compared to the baseline, making the entities belonging to the same class more clustered.

4.8 Effect on the Source Language

To investigate the impact of cross-lingual transfer on the source language, we have undertaken a deeper evaluation of the performance delivered by PRAM in the single-source transfer setting. The results of our evaluation can be found in Table 6, which reveals that PRAM is indeed capable of boosting the performance on the source language (English or 'en') for both CoNLL and WikiAnn datasets. This demonstrates that PRAM can effectively handle both the source and target languages with a single end-to-end training. Obviously, PRAM is clearly more efficient than previous SOTAs, which required separate models to be trained in two stages for each language.

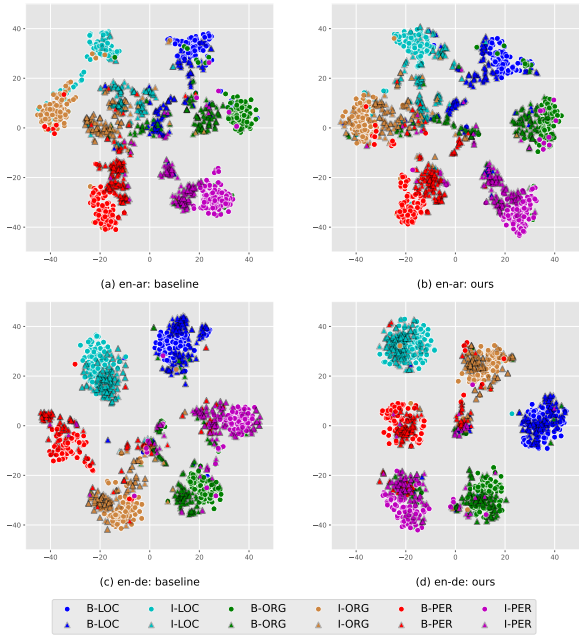

Figure 4: Two-dimensional t-SNE visualizations of the samples: ● denotes the tokens from the source language, and ▲ denotes the tokens from the target language.

5 Conclusion

This paper proposes a novel and effective prototype-based representation alignment model (PRAM) for ZRCL-NER. A novel training objective named APC is proposed to cooperate with the cross-lingual inference ability of mPLM, which can enhance the alignment of entity representations and predictions, as well as the representations of homogeneous entities across languages. The experimental results show that PRAM achieves excellent performance in both single-source and multi-source transfer settings. Last but not least, the training of PRAM is performed end-to-end and only additionally utilizes unlabeled target data.

6 Limitations

PRAM effectively handles the ZRCL-NER task but has certain limitations. Firstly, since PRAM relies on the cross-lingual inference ability of mPLM, its transfer ability may be restricted if the target language is not among the pre-trained languages of mPLM. Secondly, the high memory requirements of PRAM may occurs when the task is the multi-source transfer, where we need to set a large batch size to ensure the stable update of prototypes on different source languages. This drives us to enhance the space efficiency of our method in the future.

Acknowledgements

This work has been supported by the Innovative Research Group of the National Natural Science Foundation of China (61721002); The Key Research and Development Program of Ningxia Hui Nationality Autonomous Region (2022BEG02025); The consulting research project of the Chinese Academy of Engineering (The Online and Offline Mixed Educational Service System for The Belt and Road Training in MOOC China); Project of China Knowledge Centre for Engineering Science and Technology; The innovation team from the Ministry of Education (IRT_17R86).

References

Tareq Al-Moslmi, Marc Gallofré Ocaña, Andreas L Opdahl, and Csaba Veres. 2020. Named entity extraction for knowledge graphs: A literature overview. *IEEE Access*, 8:32862–32881.

M Saiful Bari, Shafiq Joty, and Prathyusha Jwalapuram. 2020. Zero-resource cross-lingual named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7415–7423.

Kevin Bowden, Jiaqi Wu, Shereen Oraby, Amita Misra, and Marilyn Walker. 2018. Slugnerds: A named entity recognition tool for open domain dialogue systems. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *ICLR*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

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

Jason P.C. Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco

Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Brooke Cowan, Sven Zethelius, Brittany Luk, Teodora Baras, Prachi Ukarde, and Daodao Zhang. 2015. Named entity recognition in travel-related search queries. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 3935–3941. AAAI Press.

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

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *Int. J. Comput. Vision*, 129(6):1789–1819.

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

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

Saurabh Kulshreshtha, Jose Luis Redondo Garcia, and Ching-Yun Chang. 2020. Cross-lingual alignment methods for multilingual BERT: A comparative study. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 933–942, Online. Association for Computational Linguistics.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020. A unified MRC framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5849–5859, Online. Association for Computational Linguistics.

Yangming Li, lemao liu, and Shuming Shi. 2021. Empirical analysis of unlabeled entity problem in named entity recognition. In *International Conference on Learning Representations*.

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

Shining Liang, Ming Gong, Jian Pei, Linjun Shou, Wanli Zuo, Xianglin Zuo, and Daxin Jiang. 2021. Reinforced iterative knowledge distillation for cross-lingual named entity recognition. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3231–3239.

Jindřich Libovickỳ, Rudolf Rosa, and Alexander Fraser. 2019. How language-neutral is multilingual bert? *arXiv preprint arXiv:1911.03310*.

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

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

Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE-M: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 27–38, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Papers), pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2021. Cosda-ml: multi-lingual code-switching data augmentation for zero-shot cross-lingual nlp. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3853–3860.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

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

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

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Runchuan Wang, Zhao Zhang, Fuzhen Zhuang, Dehong Gao, Yi Wei, and Qing He. 2021. Adversarial domain adaptation for cross-lingual information retrieval with multilingual bert. CIKM '21, page 3498–3502, New York, NY, USA. Association for Computing Machinery.

Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019. Cross-lingual BERT transformation for zero-shot dependency parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5721–5727, Hong Kong, China. Association for Computational Linguistics.

Zirui Wang, Jiateng Xie, Ruochen Xu, Yiming Yang, Graham Neubig, and Jaime G. Carbonell. 2020. Cross-lingual alignment vs joint training: A comparative study and a simple unified framework. In *International Conference on Learning Representations*.

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

Qianhui Wu, Zijia Lin, Börje F. Karlsson, Biqing Huang, and Jian-Guang Lou. 2021. Unitrans: Unifying model transfer and data transfer for cross-lingual named entity recognition with unlabeled data. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20, pages 3926–3932.

Qianhui Wu, Zijia Lin, Guoxin Wang, Hui Chen, Börje F Karlsson, Biqing Huang, and Chin-Yew Lin. 2020b. Enhanced meta-learning for cross-lingual named entity recognition with minimal resources. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9274–9281.

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

Shaoan Xie, Zibin Zheng, Liang Chen, and Chuan Chen. 2018. Learning semantic representations for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 5423–5432. PMLR.

Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. Named entity recognition as dependency parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6470–6476, Online. Association for Computational Linguistics.

Ran Zhou, Xin Li, Ruidan He, Lidong Bing, Erik Cambria, Luo Si, and Chunyan Miao. 2022. MELM: Data augmentation with masked entity language modeling for low-resource NER. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2251–2262, Dublin, Ireland. Association for Computational Linguistics.

A Non-entity Down-sampling Strategy

Since non-entity (O-class) samples constitute the bulk of the overall samples, it is crucial to alleviate the class imbalance caused by excessive non-entity samples. We adopt a non-entity down-sampling strategy to address this. All tokens in the input sentence participate in forward propagation, but only the entity tokens and part of the non-entity tokens are considered in the loss functions. Since the target data is unlabeled, we assign the class with the highest predicted probability output by the classifier as the label of the current token: $\hat{y}_i^t = \texttt{argmax}(p_i^t)$. With the help of the labels of the source language samples and the pseudo labels $\hat{y}^t$ of the target language samples, we randomly downsample the non-entity tokens to balance the number of non-entity tokens and entity tokens:

$$n_o^{ds} = \begin{cases} \gamma * n_e, & \text{if } \frac{n_o}{n_e} > \gamma; \\ n_o, & \text{else,} \end{cases} \tag{11}$$

where n_o and n_e are the numbers of non-entity tokens and entity tokens, respectively, n_o^{ds} denotes the number of the downsampled non-entity tokens, and γ is a balancing rate. When calculating the supervised loss $\mathcal{L}_{sup}$ and the consistency loss $\mathcal{L}_{consis}$, only the entity tokens and the sampled non-entity tokens participate.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 6

☒ A2. Did you discuss any potential risks of your work?
We don't see any potential risks in your work.

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

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

B ☑ Did you use or create scientific artifacts?

Section 4.1

☑ B1. Did you cite the creators of artifacts you used?
Section 4.1

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 4.1

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 4.1

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We don't construct a new dataset, and the used CoNLL and the WikiAnn datasets were just the same as they were initially published.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 4.1

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 4.1

C ☑ Did you run computational experiments?

Section 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4.2

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.1 and Section 4.2

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4.2

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4.2

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

Left blank.

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

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

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

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

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