Improving Self-training for Cross-lingual Named Entity Recognition with Contrastive and Prototype Learning

Ran Zhou[*1,2] **Xin Li**[1†] **Lidong Bing**[1] **Erik Cambria**[2] **Chunyan Miao**[2]
[1]DAMO Academy, Alibaba Group [2]Nanyang Technological University, Singapore
{ran.zhou, xinting.lx, l.bing}@alibaba-inc.com
{cambria, ascymiao}@ntu.edu.sg

Abstract

In cross-lingual named entity recognition (NER), self-training is commonly used to bridge the linguistic gap by training on pseudo-labeled target-language data. However, due to sub-optimal performance on target languages, the pseudo labels are often noisy and limit the overall performance. In this work, we aim to improve self-training for cross-lingual NER by combining representation learning and pseudo label refinement in one coherent framework. Our proposed method, namely ContProto mainly comprises two components: (1) contrastive self-training and (2) prototype-based pseudo-labeling. Our contrastive self-training facilitates span classification by separating clusters of different classes, and enhances cross-lingual transferability by producing closely-aligned representations between the source and target language. Meanwhile, prototype-based pseudo-labeling effectively improves the accuracy of pseudo labels during training. We evaluate ContProto on multiple transfer pairs, and experimental results show our method brings in substantial improvements over current state-of-the-art methods. [1]

1 Introduction

Cross-lingual named entity recognition (NER) (Tsai et al., 2016; Xie et al., 2018) has seen substantial performance improvement since the emergence of large-scale multilingual pretrained language models (Devlin et al., 2019; Conneau et al., 2020). However, there is still a noticeable gap between zero-shot cross-lingual transfer and monolingual NER models trained with target-language labeled data. To further bridge the linguistic gap between the source and target language, self-training is widely adopted to exploit the abundant language-specific information in unlabeled target-language data (Wu et al., 2020b; Ye et al., 2020; Chen et al., 2021). In general, self-training (sometimes referred to as teacher-student learning (Wu et al., 2020a)) uses a weak tagger (i.e. teacher model) trained on source-language data to assign pseudo labels onto unlabeled target-language data, which is then combined with labeled source-language data to train the final model (i.e. student model). Nevertheless, due to sub-optimal performances on target languages, the pseudo-labeled data contains a large number of errors and might limit the performances of NER models trained on them.

To optimize self-training for cross-lingual NER, several methods have been proposed to improve the quality of pseudo labels. One line of work focuses on selecting curated pseudo-labeled data for self-training via reinforcement learning (Liang et al., 2021a) or an adversarial discriminator (Chen et al., 2021). However, they do not fully utilize all the unlabeled data available. Wu et al. (2020a,b) exploit the full unlabeled dataset and alleviate the noise in pseudo labels by aggregating predictions from multiple teacher models. Likewise, Liang et al. (2021a) develop multi-round self-training which iteratively re-trains the teacher model to generate more accurate pseudo-labels. Despite their effectiveness, both multi-teacher and multi-round self-training impose a large computational overhead. Furthermore, the aforementioned methods are mostly data-driven and ignore the explicit modeling of cross-lingual alignment in the representation space.

In this work, we take a different approach and propose ContProto as a novel self-training framework for cross-lingual NER. Unlike existing data selection methods, ContProto sufficiently leverages knowledge from all available unlabeled target-language data. Compared with multi-teacher or multi-round self-training, our method improves pseudo label quality without training separate mod-

[*] Ran Zhou is under the Joint Ph.D. Program between Alibaba and Nanyang Technological University.

[†] Corresponding author

[1] Our code is available at https://github.com/DAMO-NLP-SG/ContProto.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4018–4031

els. Moreover, we explicitly align the representations of source and target languages to enhance the model's cross-lingual transferability. Specifically, ContProto comprises two key components, namely contrastive self-training and prototype-based pseudo-labeling. Firstly, we introduce a contrastive objective for cross-lingual NER self-training. Whereas typical supervised contrastive learning (Khosla et al., 2020) treats labeled entities of the same class as positive pairs, we further construct pseudo-positive pairs comprising of a labeled source-language entity and a target-language span predicted as the same entity type by the current model. Hence, such contrastive objective not only separates different entity classes for easier classification, but also better aligns representations of the source and target language, achieving enhanced cross-lingual transferability. Secondly, we propose a prototype-based pseudo-labeling to refine pseudo labels on-the-fly at each training step. We start with constructing class-specific prototypes based on the representations produced by contrastive self-training, which can be regarded as cluster centroids of each entity type. Then, by ranking the distances between the representation of an unlabeled span and each prototype, we gradually shift its soft pseudo label towards the closest class. As a result, errors in pseudo labels are dynamically corrected during training.

It is noteworthy that our contrastive self-training and prototype-based pseudo-labeling are mutually beneficial. On one hand, entity clusters generated by contrastive learning make it easier to determine the closest prototype and update pseudo labels correctly. In turn, the model trained on the refined pseudo labels becomes more accurate when classifying unlabeled spans, and yields more reliable positive pairs for contrastive learning.

Our contributions are summarized as follows: (1) The proposed ContProto shows competitive cross-lingual NER performance, establishing new state-of-the-art results on most of the evaluated cross-lingual transfer pairs (five out of six). (2) Our contrastive self-training produces well-separated clusters of representations for each class to facilitate classification, and also aligns the source and target language to achieve improved cross-lingual transferability. (3) Our prototype-based pseudo-labeling effectively denoises pseudo-labeled data and greatly boosts the self-training performance.

2 Preliminaries

2.1 Problem Definition

Cross-lingual named entity recognition aims to train a NER model with labeled data in a source language, and evaluate it on test data in target languages. Following previous works (Jiang et al., 2020; Ouchi et al., 2020; Yu et al., 2020; Li et al., 2020a; Fu et al., 2021), we formulate named entity recognition as a span prediction task. Given a sentence $X = \{x_1, x_2, ..., x_n\}$, we aim to extract every named entity $e_{jk} = \{x_j, x_{j+1}, ..., x_k\}$ and correctly classify it as entity type y. Under zero-shot settings, labeled data D_l^{src} is only available in the source language (src), and we leverage unlabeled data D_{ul}^{tgt} of the target language (tgt) during training.

2.2 Span-based NER

Following Fu et al. (2021), we use the span-based NER model below as our base model. Firstly, the input sentence $X = \{x_1, ..., x_n\}$ is fed through a pretrained language model to obtain its last layer representations $h = \{h_1, ..., h_n\}$. Then, we enumerate all possible spans $s_{jk} = \{x_j, ..., x_k\}$ where $1 \leq j \leq k \leq n$, to obtain the total set of spans $S(X)$. The representation for each span $s_{jk} \in S(X)$ can be the concatenation of the last hidden states of its start and end tokens $[h_j; h_k]$. We additionally introduce a span length embedding l_{k-j}, which is obtained by looking up the (k-j)$^{\text{th}}$ row of a learnable span length embedding matrix. Thus, we obtain the final representation of s_{jk} as $z_{jk} = [h_j; h_k; l_{k-j}]$. Finally, the span representation is passed through a linear classifier to obtain its probability distribution $P_\theta(s_{jk}) \in \mathbb{R}^{|\mathbb{C}|}$, where $\mathbb{C}$ is the label set comprising of predefined entity types and an "O" class for non-entity spans.

2.3 Self-training for NER

Typically, self-training (or teacher-student learning) for cross-lingual NER first trains a teacher model $\mathcal{M}(\theta_t)$ on the available source-language labeled dataset D_l^{src} using a cross-entropy loss:

$$L_{src} = -\frac{1}{N} \sum_{X \in D_l^{src}} \frac{1}{|S(X)|} \sum_{s_{jk} \in S(X)} \sum_{c \in \mathbb{C}} y_{jk}^c \log P_{\theta_t}^c(s_{jk}) \tag{1}$$

where N is the batch size, $y_{jk}^c = 1$ for the true label of span s_{jk} and 0 otherwise.

Given an unlabeled target-language sentence $X \in D_{ul}^{tgt}$, the teacher model then assigns soft pseudo label $\hat{y}_{jk} = P_{\theta_t}(s_{jk}) \in \mathbb{R}^{|\mathbb{C}|}$ to each span $s_{jk} \in X$. The student model $\mathcal{M}(\theta_s)$ will be trained on the pseudo-labeled target-language data as well, using a soft cross-entropy loss:

$$L_{tgt} = -\frac{1}{N} \sum_{X \in D_{ul}^{tgt}} \frac{1}{|S(X)|} \sum_{s_{jk} \in S(X)} \sum_{c \in \mathbb{C}} \hat{y}_{jk}^{c} \log P_{\theta_s}^{c}(s_{jk}) \tag{2}$$

The total objective for the student model in vanilla self-training is:

$$L = L_{src} + L_{tgt} \tag{3}$$

3 Methodology

In this section, we present our self-training framework ContProto for cross-lingual NER. As shown in the right part of Figure 1, ContProto mainly comprises two parts, namely: (1) contrastive self-training (Section 3.1) which improves span representations using contrastive learning; (2) prototype-based pseudo-labeling (Section 3.2) which gradually improves pseudo label quality with prototype learning.

3.1 Contrastive Self-training

In the following section, we first describe supervised contrastive learning for span-based NER, which focuses on source-language representations. Then, we introduce our pseudo-positive pairs, by which we aim to improve target-language representations as well.

Supervised contrastive learning We extend SupCon (Khosla et al., 2020) to span-based NER, which leverages label information to construct positive pairs from samples of the same class and contrasts them against samples from other classes. Firstly, to generate multiple views of the same labeled source-language sentence, each batch is passed twice through the span-based NER model described in Section 2.2. An input sentence X undergoes different random dropouts during each pass, such that each span $s_{jk} \in S(X)$ yields two representations z_{jk}, z'_{jk}. The span representations are further passed through a two-layer MLP, to obtain their projected representations ζ_{jk}, ζ'_{jk}. We denote the entire set of multi-viewed spans as $\{s_i, y_i, \zeta_i\}_{i=1}^{2m}$, where y_i is the true label of s_i and $m = \sum_X |S(X)|$ is the total number of spans in the original batch of sentences.

Then, the supervised contrastive loss is defined as follows:

$$L_{cont} = -\frac{1}{2m} \sum_{i=1}^{2m} \frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\zeta_i \cdot \zeta_p / \tau)}{\sum_{a \in A(i)} \exp(\zeta_i \cdot \zeta_a / \tau)} \tag{4}$$

where $A(i) \equiv \{1, 2, ..., 2m\} \setminus \{i\}$, and $P(i) \equiv \{p \in A(i) : y_i = y_p\}$ are indices of the positive sample set consisting of spans sharing the same label as s_i. Essentially, supervised contrastive learning helps to pull source-language entities of the same class together while pushing clusters of different classes apart, which induces a clustering effect and thereby benefits classification.

Pseudo-positive pairs As the aforementioned positive pair only involve source-language spans, it does not explicitly optimize target-language representations or promote cross-lingual alignment. Therefore, we propose to construct pseudo-positive pairs which take target-language spans into account as well.

Concretely, we expand the multi-viewed span set $\{s_i, y_i, \zeta_i\}_{i=1}^{2m}$ by adding in unlabeled target-language spans, where m denotes the total number of spans from the source- and target-language sentences. For a source-language span, y_i is still its gold label y_i^{gold}. However, as gold annotations are not available for target-language spans, we instead treat the model's prediction at the current training step as an approximation for its label y_i:

$$y_i = \begin{cases} y_i^{gold} & \text{if } s_i \in D_l^{src} \\ \operatorname{argmax} P_\theta(s_i) & \text{if } s_i \in D_{ul}^{tgt} \end{cases} \tag{5}$$

Likewise, we construct positive pairs from entities with the same (gold or approximated) label. As an example, positive pairs for the PER (person) class might be composed of: (1) two source-language PER names; (2) one source-language PER name and one target-language span predicted as PER; (3) two target-language spans predicted as PER. Therefore, apart from separating clusters of different classes, our contrastive self-training also explicitly enforces the alignment between languages, which facilitates cross-lingual transfer.

Consistency regularization We also include a consistency regularization term (Liang et al., 2021b) to further enhance the model's robustness. Recall that each sentence is passed twice through the NER model, and each span s_i yields two probability distributions $P_\theta(s_i), P'_\theta(s_i)$ that are not exactly identical due to random dropout. Therefore,

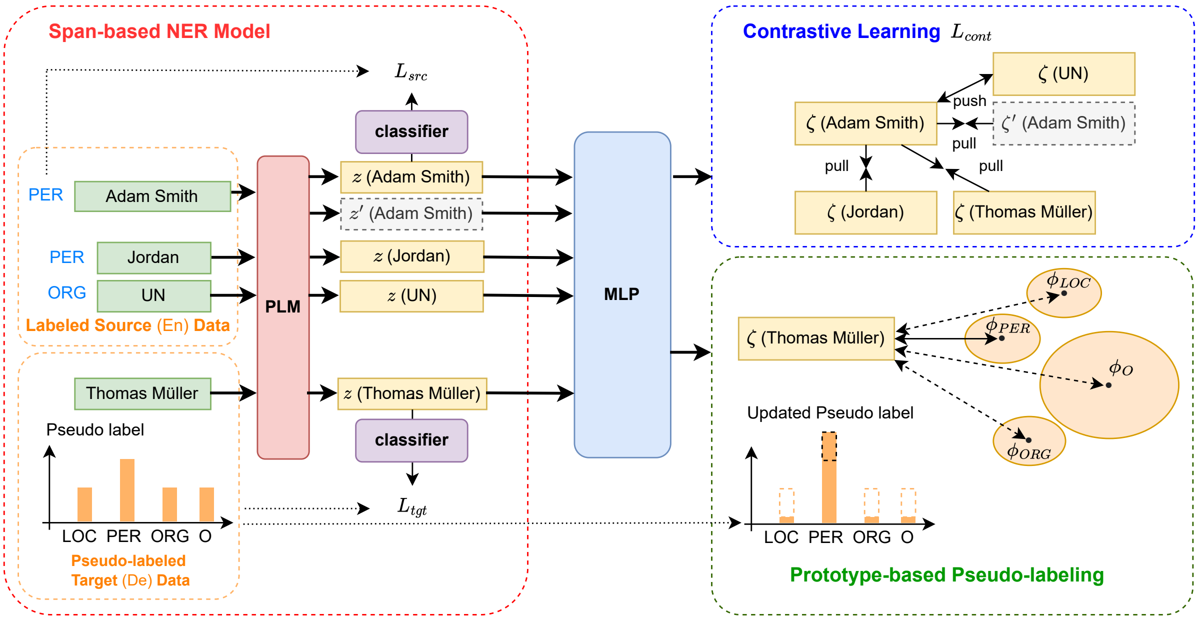

Figure 1: Illustration of ContProto. Both classifier blocks share the same parameters.

we enforce the model to output consistent predictions by minimizing the following KL divergence:

$$L_{reg} = -\frac{1}{m}\sum_{i=1}^{m} \mathrm{KL}(P_\theta(s_i) \,||\, P'_\theta(s_i)) \quad (6)$$

Finally, the total objective for ContProto is:

$$L = L_{src} + L_{tgt} + L_{cont} + L_{reg} \quad (7)$$

3.2 Prototype-based Pseudo-labeling

Benefiting from our contrastive self-training in Section 3.1, entity representations (both source- and target-language) of the same class are tightly clustered together. Intuitively, the closest cluster to an unlabeled span is likely to represent the span's true class. Therefore, we can conveniently utilize these induced clusters as guidance to infer the unlabeled span's NER label. To this end, we introduce prototype-based pseudo-labeling, which leverages prototype learning (Snell et al., 2017) to refine pseudo labels at each training step.

Class-specific prototypes To start off, we first define a series of prototypes ϕ_c, each corresponding to a class $c \in \mathbb{C}$. A prototype ϕ_c is a representation vector that can be deemed as the cluster centroid of class c. Naively, ϕ_c can be calculated by averaging representations of class c in the entire dataset at the end of an epoch. However, this means the prototypes will remain static during the next full epoch. This is not ideal as distributions of span representations and clusters are vigorously changing, especially in the earlier epochs. Hence, we adopt a moving-average style of calculating prototypes. Specifically, we iterate through a batch of mixed source- and target-language spans $\{s_i, y_i, \zeta_i\}_{i=1}^{m}$, and update prototype ϕ_c as the moving-average embedding for spans with (either gold or approximated) label c:

$$\begin{aligned} \phi_c = \text{Normalize } (\alpha\phi_c + (1-\alpha)\zeta_i), \\ \forall i \in \{i \mid y_i = c\} \end{aligned} \quad (8)$$

Same as Equation 5, y_i is either the gold label for source-language spans, or the approximated label obtained from the model's predictions for target-language spans. α is a hyperparameter controlling the update rate.

Pseudo label refinement Having obtained the prototypes, we then use them as references to refine the pseudo labels of target-language spans. Typically, prototype learning classifies an unlabeled sample by finding the closest prototype, and assigning the corresponding label. However, this may cause two problems: (1) Assigning a hard one-hot label forfeits the advantages of using soft labels in self-training. (2) As the closest prototype might differ between consecutive epochs, there is too much perturbation in pseudo labels that makes training unstable. Thus, we again take a moving-average approach to incrementally update pseudo labels at each training step. Given a target-language span $\{s, \zeta\}$ at epoch t, its soft pseudo label from previous epoch $\hat{y}_{t-1}$ is updated as follows:

$$\hat{y}_t^c = \begin{cases} \beta\hat{y}_{t-1}^c + (1-\beta) & \text{if } c = \arg\max_{\gamma\in\mathbb{C}}(\phi_\gamma \cdot \zeta) \\ \beta\hat{y}_{t-1}^c & \text{otherwise} \end{cases} \tag{9}$$

where $\hat{y}_t^c$ represents the pseudo probability on class c and β is a hyperparameter controlling the update rate. We use the dot product to calculate similarity $\phi_\gamma \cdot \zeta$, and define the distance between span representation and prototype as $(1 - \phi_\gamma \cdot \zeta)$. In other words, we find the prototype closest to the span's representation and take the corresponding class as an indication of the span's true label. Then, we slightly shift the current pseudo label towards it, by placing extra probability mass on this class while deducting from other classes. Cumulatively, we are able to rectify pseudo labels whose most-probable class is incorrect, while reinforcing the confidence of correct pseudo labels.

Margin-based criterion NER is a highly class-imbalanced task, where the majority of spans are non-entities ("O"). As a result, non-entity span representations are widespread and as later shown in Section 5.2, the "O" cluster will be significantly larger than other entity types. Therefore, a non-entity span at the edge of the "O" cluster might actually be closer to an entity cluster. Consequently, the above prototype-based pseudo-labeling will wrongly shift its pseudo label towards the entity class and eventually result in a false positive instance.

To address this issue, we further add a margin-based criterion to enhance prototype learning. Intuitively, a true entity span should lie in the immediate vicinity of a certain prototype. Thus, we do not update pseudo labels towards entity classes if the span is not close enough to any of the entity prototypes ϕ_γ, i.e., the similarity between the prototype and any span representation $(\phi_\gamma \cdot \zeta_i)$ does not exceed a margin r. Meanwhile, as non-entity spans are widely distributed, we do not apply extra criteria and update a span as "O" as long as its closest prototype is ϕ_O. Formally:

$$\beta = \begin{cases} \beta & \text{if } \arg\max_{\gamma\in\mathbb{C}}(\phi_\gamma \cdot \zeta_i) = \text{O} \\ \beta & \text{if } \max_{\gamma\in\mathbb{C}\setminus\{\text{O}\}}(\phi_\gamma \cdot \zeta_i) > r \\ 1 & \text{otherwise} \end{cases} \tag{10}$$

We notice that different entity classes of different target languages might have varying cluster tightness, and thus it is not judicious to manually set a fixed margin r universally. Instead, we automatically set class-specific margin r_c from last epoch's statistics, by calculating the averaged similarity between target-language spans predicted as class c and prototype ϕ_c:

$$r_c = \text{MEAN}(\phi_c \cdot \zeta_i), \text{where } \arg\max P_\theta(s_i) = c \tag{11}$$

Note that, at the start of training, our model does not produce well-separated clusters and the prototypes are randomly initialized. Therefore, we warm up the model by not updating pseudo labels in the first epoch.

We highlight that our contrastive learning and prototype-based pseudo-labeling are mutually beneficial. By virtue of the clustering effect from contrastive learning, the resulting representations and prototypes act as guidance for refining pseudo labels. In turn, the model trained with refined pseudo-labels predicts unlabeled spans more accurately, and ensures the validity of pseudo-positive spans for contrastive learning. To summarize, the two components work collaboratively to achieve the overall superior performance of ContProto.

4 Experiments

In this section, we verify the effectiveness of ContProto by conducting experiments on two public NER datasets with six cross-lingual transfer pairs and performing comparisons with various baseline models.

4.1 Dataset

Following previous works (Liang et al., 2021a; Li et al., 2022), we evaluate our ContProto on six cross-lingual transfer pairs from two widely used NER datasets: (1) CoNLL dataset (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003), which includes four languages, namely English (En), German (De), Spanish (Es) and Dutch (Nl); (2) WikiAnn dataset (Pan et al., 2017) of English (En), Arabic (Ar), Hindi (Hi), and Chinese (Zh). Following common settings, we use the original English training set as our source-language training data D_l^{src}, while treating others as target languages and evaluate on their test sets. Annotations on target-language training sets are removed, and they are used as our unlabeled target-language data D_{ul}^{tgt} for self-training. English development set is used for early stopping and model selection.

4.2 Baselines

We mainly benchmark against the following self-training baselines for cross-lingual NER:

TSL (Wu et al., 2020a) weights supervision from multiple teacher models based on a similarity measure as pseudo labels for self-training.

Unitrans (Wu et al., 2020b) trains a series of teacher models sequentially using source-language data or translated data, and uses a voting scheme to aggregate pseudo labels from them.

RIKD (Liang et al., 2021a) proposes a reinforced instance selector for picking unlabeled data and iteratively conducts self-training for multiple rounds.

AdvPicker (Chen et al., 2021) leverages adversarial language discriminator for picking pseudo-labeled data.

MTMT (Li et al., 2022) introduces an extra entity similarity module and trains the student model with both NER and similarity pseudo labels.

We also compare ContProto with several baseline methods that do not leverage unlabeled target-language data, including Wiki (Tsai et al., 2016), WS (Ni et al., 2017), TMP (Jain et al., 2019), BERT-f (Wu and Dredze, 2019), AdvCE (Keung et al., 2019), XLM-R$_{\text{Large}}$ (Hu et al., 2020), mT5$_{\text{XXL}}$ (Xue et al., 2021).

4.3 Implementation Details

We use XLM-R$_{\text{Large}}$ (Conneau et al., 2020) as the backbone pretrained language model of our span-based NER model. The dimension of the projected representations ζ_i for contrastive learning is set to 128. The model is trained for 10 epochs. AdamW (Loshchilov and Hutter, 2019) is used for optimization and the learning rate is set to 1e-5. We empirically set exponential moving average coefficients as $\alpha = 0.99$ and $\beta = 0.95$. The batch size for both labeled source-language data and unlabeled target-language data is set to 16.

4.4 Main Results

CoNLL results We present the experimental results on CoNLL dataset in Table 1. Overall, our ContProto achieves the best results in terms of averaged F1 over the target languages, with a +1.03 improvement compared to the previous state-of-the-art MTMT. Compared with methods that do not use unlabeled data, ContProto presents substantial improvements, suggesting that incorporating target-language unlabeled data is indeed beneficial to cross-lingual NER. Furthermore, our method outperforms both multi-teacher (i.e., TSL, Unitrans) and multi-round (i.e., Unitrans, RIKD) self-training. This shows our prototype learning produces more accurate pseudo labels compared to ensembling multiple teacher models or iterative self-training. Compared with data selection methods (i.e., RIKD, AdvPicker), our superior performance demonstrates that on the premise of guaranteeing high-quality pseudo labels, it is beneficial to leverage as much target-language data as possible.

Although MTMT attempts to reduce the distance between entities of the same class in the same lan-

Method	De	Es	Nl	Avg
w/o unlabeled data				
Wiki	48.12	60.55	61.56	56.74
WS	58.50	65.10	65.40	63.00
TMP	61.50	73.50	69.90	68.30
BERT-f	69.56	74.96	77.57	74.03
AdvCE	71.90	74.30	77.60	74.60
self-training				
TSL	75.33	78.00	81.33	78.22
Unitrans	74.82	79.31	82.90	79.01
RIKD	**78.40**	79.46	81.40	79.75
AdvPicker				
- *seq-tagging*	75.01	79.00	82.90	78.97
- *span-based* †	73.93	84.70	81.01	79.88
MTMT	76.80	81.82	83.41	80.68
ContProto *(Ours)*	76.41	**85.02**	**83.69**	**81.71**

Table 1: Experimental results on CoNLL. ContProto results are micro-F1 averaged over 3 runs.†Implemented using span-based NER model. Baseline results without markers are cited from the original papers.

Method	Ar	Hi	Zh	Avg
w/o unlabeled data				
BERT-f	42.30	67.60	52.90	54.27
XLM-R$_{\text{Large}}$	53.00	73.00	33.10	53.03
mT5$_{\text{XXL}}$	66.20	77.80	56.80	66.93
self-training				
TSL	50.91	72.48	31.14	51.51
RIKD	54.46	74.42	37.48	55.45
AdvPicker				
- *seq-tagging* †	53.76	73.69	41.24	56.23
- *span-based* ‡	70.70	80.37	56.57	69.21
MTMT	52.77	70.76	52.26	58.60
ContProto *(Ours)*	**72.20**	**83.45**	**61.47**	**72.37**

Table 2: Experimental results on WikiAnn. ContProto results are micro-F1 averaged over 3 runs. †Implemented using official source code. ‡Implemented using span-based NER model. Baseline results without markers are cited from the original papers.

guage, it does not account for the relation between a source- and a target-language entity. Besides, AdvPicker implicitly aligns the source and target language during language-independent data selection but does not inherit those representations when training the final model. In comparison, our contrastive objective explicitly reduces the distance between a pair of source- and target-language entities of the same class, which aligns the source- and target-language representations to achieve better cross-lingual performance.

For a fair comparison, we further implement span-based NER based on the official codebase of AdvPicker (Chen et al., 2021). From experimental results, span-based AdvPicker shows some improvement over the original sequence tagging formulation. However, our ContProto still outperforms span-based AdvPicker by a considerable margin.

WikiAnn results As shown in Table 2, our ContProto achieves superior results on WikiAnn languages as well, with an averaged +3.16 F1 improvement compared to the best baseline method. It is noteworthy that span-based AdvPicker presents considerable improvements compared to its original sequence-tagging formulation, suggesting that span-based NER is a more appropriate formulation for identifying named entities in cross-language scenarios, especially for transfer pairs with larger linguistic gaps. Compared with span-based AdvPicker, ContProto still shows a significant advantage by aligning source- and target-language representations and improving pseudo-label quality.

5 Analysis

5.1 Ablation Studies

To demonstrate the contribution of each design component of ContProto, we conduct the following ablation studies: (1) *w/o proto* which removes prototype-based pseudo-labeling and only keeps our contrastive self-training; (2) *w/o proto & cl* which removes both prototype-based pseudo-labeling and the contrastive objective; (3) *w/o reg* which removes the consistency regularization; (4) *fixed margin* which manually tunes a universally fixed margin $r = 1.0$ instead of automatic class-specific margins; (5) *proto w/o cl* which removes the contrastive objective, and directly uses the unprojected representation z_i for constructing prototypes and updating pseudo labels.

Based on experimental results in Table 3, we make the following observations: (1) *w/o proto* shows reduced performance on all target languages, which verifies the ability of our prototype-based pseudo-labeling in improving pseudo label quality. (2) *w/o proto & cl* further lowers target-language performance, which demonstrates the effectiveness of contrastive self-training in separating different classes and aligning the source- and target-language representations. (3) *w/o reg* demonstrates that removing the consistency regularization leads to slight performance drops on all target languages. (4) Using a manually tuned universal margin, *fixed margin* underperforms ContProto by a considerable amount. This signifies the flexibility brought by the automatic margin when cluster tightness differs between classes. (5) *proto w/o cl* leads to drastic performance drops. Without the contrastive objective, clusters of different classes overlap with each other. As a result, the closest prototype might not accurately reflect a span's true label, and this leads to deteriorated pseudo label quality. Thus, the clustering effect from contrastive learning is essential for accurate prototype-based pseudo-labeling.

5.2 Visualizing Span Distributions

We also conduct a t-SNE visualization (Van der Maaten and Hinton, 2008) of span representations z_i. As shown in Figure 2a, vanilla self-training generates representations with some overlapping between different classes, which makes it challenging to classify them. In contrast, our ContProto (Figure 2b) produces more distinguishable representations where clusters of different classes are separated, which verifies the effectiveness of our contrastive objective. Furthermore, it can be easily seen that the non-entity "O" cluster is significantly larger than other entity classes, which justifies the necessity of margin-based criterion in Section 3.2.

5.3 Pseudo Label Quality

Recall that we remove gold labels from the original target-language training sets, and treat them as unlabeled data for self-training. For analysis purposes, we retrieve those gold labels, to investigate the efficacy of ContProto in improving the quality of pseudo labels.

Specifically, we take the gold labels as references to calculate the oracle F1 of pseudo labels at the end of each epoch. As shown in Figure 3, the pseudo label F1 indeed improves during training on all target languages, proving the effectiveness

Method	De	Es	Nl	Ar	Hi	Zh
ContProto	**76.41**	**85.02**	**83.69**	**72.20**	**83.45**	**61.47**
- *w/o proto*	74.87 (-1.54)	84.08 (-0.94)	81.44 (-2.25)	71.49 (-0.71)	83.10 (-0.35)	59.57 (-1.90)
- *w/o proto & cl*	74.17 (-2.24)	84.47 (-0.54)	81.03 (-2.66)	70.40 (-1.80)	81.00 (-2.45)	56.30 (-5.16)
- *w/o reg*	76.23 (-0.18)	84.96 (-0.06)	83.56 (-0.13)	72.15 (-0.05)	83.21 (-0.24)	61.31 (-0.16)
- *fixed margin*	74.65 (-1.76)	84.49 (-0.52)	83.09 (-0.60)	69.19 (-3.01)	83.07 (-0.38)	60.61 (-0.86)
- *proto w/o cl*	72.59 (-3.82)	81.18 (-3.84)	80.76 (-2.93)	69.72 (-2.48)	58.38 (-25.07)	53.52 (-7.95)

Table 3: Ablation studies. Values in brackets indicate the performance drop compared to our full method.

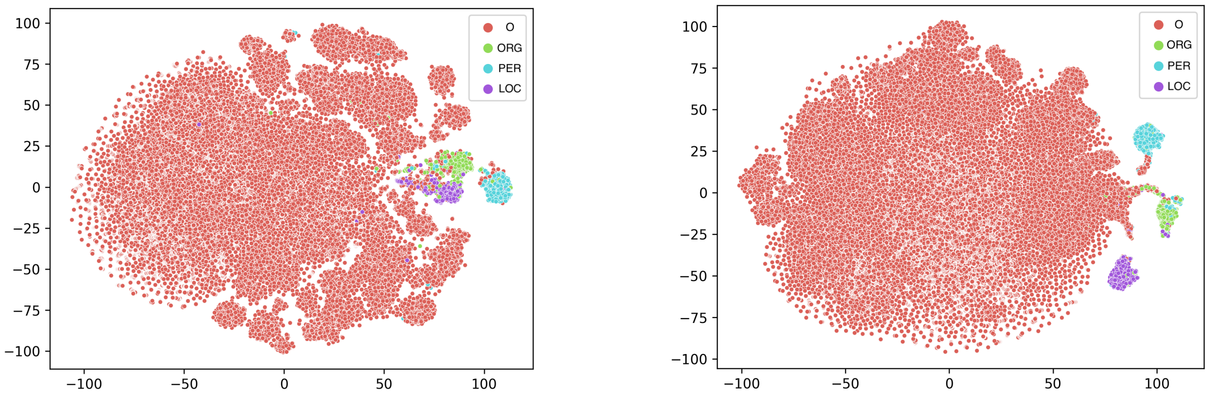

Figure 2: t-SNE visualization of Chinese (Zh) spans.

Figure 3: Pseudo label quality. The horizontal axis is the epoch number and the vertical axis is the oracle F1 of pseudo labels.

of our prototype-based pseudo-labeling. Noticeably, there are significant F1 increases (5$\sim$7%) on German (De), Arabic (ar), and Chinese (Zh). On Hindi (Hi), however, we observe a slight drop of pseudo label F1 after epoch 3, which is mainly due to a reduction of pseudo label recall. We attribute this to the larger variance of Hindi entity distribution, such that many entities outside the automatic margin turn into false negatives. As the ablation study (*w/o proto*) shows, prototype-based pseudo-labeling for Hindi only accounts for +0.35 performance improvement, and the overall improvement mainly comes from contrastive self-training. Still though, compared with initial pseudo labels, the updated Hindi pseudo label quality is improved.

6 Related Work

Cross-lingual NER Existing methods for NER (Ding et al., 2020; Xu et al., 2021, 2022, 2023a,b; Zhou et al., 2022b,a) under cross-lingual settings (Zhang et al., 2021; Liu et al., 2022a,b) can be categorized into: (1) feature-based methods, which generate language-independent features to facilitate cross-lingual transfer via wikification (Tsai et al., 2016), language alignment (Wu and Dredze, 2019) or adversarial learning (Keung et al., 2019). (2) translation-based methods, which produce pseudo training data by translating labeled source-language data word-by-word (Xie et al., 2018) or with the help of word alignment tools (Jain et al., 2019; Li et al., 2020b; Liu et al., 2021). (3) self-training methods, which generate pseudo-labeled target-language data using a model trained with labeled source-language data (Wu et al., 2020a,b; Liang et al., 2021a; Chen et al., 2021; Li et al., 2022). One concurrent work (Ge et al., 2023) that is similar to ours also aims to improve self-training for cross-lingual NER, but they adopt the traditional sequence tagging formulation, and also only apply contrastive learning on class-specific prototypes instead of actual spans. Dong et al. (2020) also leverages self-training for sentence-level cross-lingual tasks.

Contrastive learning Self-supervised contrastive learning has been widely used to generate representations for various tasks (Chen et al., 2020; Chuang et al., 2020; Tian et al., 2020; You et al., 2020; Han et al., 2022; Nguyen et al., 2022; Tan et al., 2022). In a nutshell, contrastive learning pulls positive pairs closer while pushing negative pairs apart. Supervised contrastive learning (Khosla et al., 2020) further constructs positive pairs with labeled samples of the same class, which ensures the validity of positive pairs. Das et al. (2022) leverages contrastive learning for name entity recognition, but they work on monolingual few-shot settings while we focus on cross-lingual NER self-training.

Prototype learning Prototype learning (Snell et al., 2017; Wang et al., 2022a) produces representations where examples of a certain class are close to the class-specific prototype. Several works explored prototype learning for few-shot NER (Fritzler et al., 2019; Hou et al., 2020; Wang et al., 2022b).

7 Conclusions

In this work, we propose ContProto as a novel self-training framework for cross-lingual NER, which synergistically incorporates representation learning and pseudo label refinement. Specifically, our contrastive self-training first generates representations where different classes are separated, while explicitly enforcing the alignment between source and target languages. Leveraging the class-specific representation clusters induced by contrastive learning, our prototype-based pseudo-labeling scheme further denoises pseudo labels using prototypes to infer true labels of target language spans. As a result, the model trained with more reliable pseudo labels is more accurate on the target languages. In our method, the contrastive and prototype learning components are mutually beneficial, where the former induces clusters which makes it easier to identify the closest prototype, and the latter helps to construct more accurate sample pairs for contrastive learning. Evaluated on multiple cross-lingual transfer pairs, our method brings in substantial improvements over various baseline methods.

Limitations

In this work, we propose a self-training method which requires unlabeled data in target languages. Recall that we remove gold labels from readily available target-language training data from the same public NER dataset, and use them as unlabeled data in our experiments. However, this might not perfectly simulate a real-life application scenario. Firstly, most free text in target languages might not contain any predefined named entities. This requires careful data cleaning and preprocessing to produce unlabeled data ready for use. Secondly, there might be a domain shift between labeled source-language data and unlabeled target-language data, which poses a question on the effectiveness of our method.

Furthermore, the NER datasets used in this work contain only a few entity types and different entity classes are relatively balanced. However, on datasets with a larger number of classes, each class will be underrepresented in a batch and a larger batch size might be required for contrastive self-training to work satisfactorily. Also, if the entity type distribution is long-tailed, prototypes for those rare entity types might be inaccurate, and this affects the efficacy of prototype-based pseudo-labeling.

Lastly, as we observe slight drops of pseudo label quality at the end of training for some languages, the pseudo label update strategy can be refined for further improvement.

Acknowledgements

This research is supported (, in part,) by Alibaba Group through Alibaba Innovative Research (AIR) Program and Alibaba-NTU Singapore Joint Research Institute (JRI), Nanyang Technological University, Singapore.

References

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*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607.

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.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. 2020. Debiased contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 8765–8775.

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.

Sarkar Snigdha Sarathi Das, Arzoo Katiyar, Rebecca Passonneau, and Rui Zhang. 2022. CONTaiNER: Few-shot named entity recognition via contrastive learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6338–6353.

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.

Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq Joty, Luo Si, and Chunyan Miao. 2020. DAGA: Data augmentation with a generation approach for low-resource tagging tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6045–6057.

Xin Dong, Yaxin Zhu, Yupeng Zhang, Zuohui Fu, Dongkuan Xu, Sen Yang, and Gerard De Melo. 2020. Leveraging adversarial training in self-learning for cross-lingual text classification. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1541–1544.

Alexander Fritzler, Varvara Logacheva, and Maksim Kretov. 2019. Few-shot classification in named entity recognition task. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, SAC '19, page 993–1000.

Jinlan Fu, Xuanjing Huang, and Pengfei Liu. 2021. SpanNER: Named entity re-/recognition as span prediction. 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 7183–7195.

Ling Ge, Chuming Hu, Guanghui Ma, Hong Zhang, and Jihong Liu. 2023. Prokd: An unsupervised prototypical knowledge distillation network for zero-resource cross-lingual named entity recognition. *arXiv preprint arXiv:2301.08855*.

Wei Han, Hui Chen, Zhen Hai, Soujanya Poria, and Lidong Bing. 2022. SANCL: Multimodal review helpfulness prediction with selective attention and natural contrastive learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5666–5677.

Yutai Hou, Wanxiang Che, Yongkui Lai, Zhihan Zhou, Yijia Liu, Han Liu, and Ting Liu. 2020. Few-shot slot tagging with collapsed dependency transfer and label-enhanced task-adaptive projection network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1381–1393.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421.

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.

Zhengbao Jiang, Wei Xu, Jun Araki, and Graham Neubig. 2020. Generalizing natural language analysis through span-relation representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2120–2133.

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.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673.

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

Xin Li, Lidong Bing, Wenxuan Zhang, Zheng Li, and Wai Lam. 2020b. Unsupervised cross-lingual adaptation for sequence tagging and beyond. *arXiv preprint arXiv:2010.12405*.

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.

Shining Liang, Ming Gong, Jian Pei, Linjun Shou, Wanli Zuo, Xianglin Zuo, and Daxin Jiang. 2021a. 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.

Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. 2021b. R-drop: Regularized dropout for neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 10890–10905.

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

Linlin Liu, Xin Li, Ruidan He, Lidong Bing, Shafiq Joty, and Luo Si. 2022a. Enhancing multilingual language model with massive multilingual knowledge triples. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6878–6890.

Linlin Liu, Thien Hai Nguyen, Shafiq Joty, Lidong Bing, and Luo Si. 2022b. Towards multi-sense cross-lingual alignment of contextual embeddings. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4381–4396.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Thong Nguyen, Xiaobao Wu, Anh Tuan Luu, Zhen Hai, and Lidong Bing. 2022. Adaptive contrastive learning on multimodal transformer for review helpfulness prediction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10085–10096.

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

Hiroki Ouchi, Jun Suzuki, Sosuke Kobayashi, Sho Yokoi, Tatsuki Kuribayashi, Ryuto Konno, and Kentaro Inui. 2020. Instance-based learning of span representations: A case study through named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6452–6459.

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.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30.

Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. 2022. Domain generalization for text classification with memory-based supervised contrastive learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6916–6926.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. 2020. What makes for good views for contrastive learning? In *Advances in Neural Information Processing Systems*, volume 33, pages 6827–6839.

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.

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

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

Haobo Wang, Ruixuan Xiao, Yixuan Li, Lei Feng, Gang Niu, Gang Chen, and Junbo Zhao. 2022a. Pico: Contrastive label disambiguation for partial label learning. In *The Tenth International Conference on Learning Representations, ICLR 2022*.

Jianing Wang, Chengyu Wang, Chuanqi Tan, Minghui Qiu, Songfang Huang, Jun Huang, and Ming Gao. 2022b. Spanproto: A two-stage span-based prototypical network for few-shot named entity recognition. *arXiv preprint arXiv:2210.09049*.

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.

Qianhui Wu, Zijia Lin, Börje F. Karlsson, Biqing Huang, and Jian-Guang Lou. 2020b. Unitrans : Unifying model transfer and data transfer for cross-lingual named entity recognition with unlabeled data. In *Proceedings of IJCAI*, pages 3926–3932.

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.

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

Lu Xu, Zhanming Jie, Wei Lu, and Lidong Bing. 2021. Better feature integration for named entity recognition. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3457–3469.

Weiwen Xu, Xin Li, Yang Deng, Wai Lam, and Lidong Bing. 2023a. Peerda: Data augmentation via modeling peer relation for span identification tasks. In *The 61th Annual Meeting of the Association for Computational Linguistics.*

Weiwen Xu, Xin Li, Wai Lam, and Lidong Bing. 2023b. mpmr: A multilingual pre-trained machine reader at scale. In *The 61th Annual Meeting of the Association for Computational Linguistics.*

Weiwen Xu, Xin Li, Wenxuan Zhang, Meng Zhou, Lidong Bing, Wai Lam, and Luo Si. 2022. From clozing to comprehending: Retrofitting pre-trained language model to pre-trained machine reader. *arXiv preprint arXiv:2212.04755*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

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

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. 2020. Graph contrastive learning with augmentations. In *Advances in Neural Information Processing Systems*, volume 33, pages 5812–5823.

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.

Wenxuan Zhang, Ruidan He, Haiyun Peng, Lidong Bing, and Wai Lam. 2021. Cross-lingual aspect-based sentiment analysis with aspect term code-switching. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9220–9230.

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

Ran Zhou, Xin Li, Ruidan He, Lidong Bing, Erik Cambria, Luo Si, and Chunyan Miao. 2022b. 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.

ACL 2023 Responsible NLP Checklist

A For every submission:

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

☑ A2. Did you discuss any potential risks of your work?
Appendix A1

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract & 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?
Appendix A2

☑ 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)?
Appendix A2

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ 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?
Appendix A3

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 4.3

☑ 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?
Table 1, Table 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.)?
Not applicable. Left blank.

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.