Improving Pseudo Labels with Global-Local Denoising Framework for Cross-lingual Named Entity Recognition

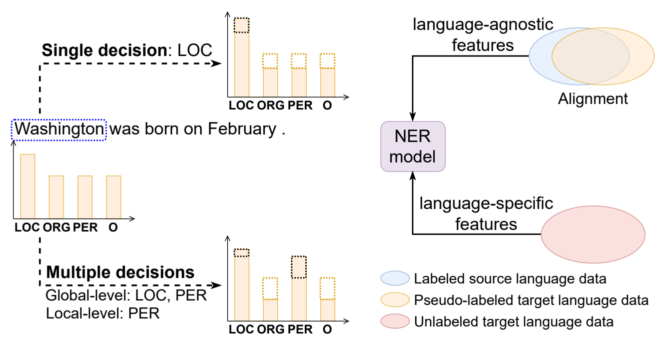
Zhuojun Ding[1,2] , **Wei Wei***[1,2] , **Xiaoye Qu**[1,2] , **Dangyang Chen**[3]
[1]Cognitive Computing and Intelligent Information Processing (CCIIP) Laboratory, School of Computer Science and Technology, Huazhong University of Science and Technology
[2]Joint Laboratory of HUST and Pingan Property & Casualty Research (HPL)
[3]Ping An Property & Casualty Insurance Company of China, Ltd.
{dingzj,weiw,quxiaoye}@hust.edu.cn, chendangyang273@pingan.com.cn

Abstract

Cross-lingual named entity recognition (NER) aims to train an NER model for the target language leveraging only labeled source language data and unlabeled target language data. Prior approaches either perform label projection on translated source language data or employ a source model to assign pseudo labels for target language data and train a target model on these pseudo-labeled data to generalize to the target language. However, these automatic labeling procedures inevitably introduce noisy labels, thus leading to a performance drop. In this paper, we propose a Global-Local Denoising framework (GLoDe) for cross-lingual NER. Specifically, GLoDe introduces a progressive denoising strategy to rectify incorrect pseudo labels by leveraging both global and local distribution information in the semantic space. The refined pseudo-labeled target language data significantly improves the model's generalization ability. Moreover, previous methods only consider improving the model with language-agnostic features, however, we argue that target language-specific features are also important and should never be ignored. To this end, we employ a simple auxiliary task to achieve this goal. Experimental results on two benchmark datasets with six target languages demonstrate that our proposed GLoDe significantly outperforms current state-of-the-art methods.

1 Introduction

Named Entity Recognition (NER) aims to identify spans belonging to predefined entity classes in a given text, such as locations and organizations [Qu *et al.*, 2023b; Gu *et al.*, 2022; Qu *et al.*, 2023a], which is an important information extraction task [Zhu *et al.*, 2023; Fan *et al.*, 2023]. Deep neural networks trained on large amounts of labeled data have achieved excellent performance [Wei *et al.*, 2021] on NER. However, most low-resource languages (e.g. Albanian and Finnish) still suffer from insufficient labeled data compared to high-resource languages, such as English. Moreover, annotating entities in texts is time-consuming and expensive. Such predicament motivates research on cross-lingual NER [Xie *et al.*, 2018a]. Cross-lingual NER aims to leverage labeled data in a high-resource language (source language) and unlabeled data in a low-resource language (target language) to train an NER model, which generalizes well to the target language.



(a) Pseudo-label denoising strategies **(b) Model improvement strategies**

Figure 1: (a) **Up**: Previous works only consider one potential entity type for pseudo-label denoising. **Down**: We leverage global and local information within the semantic space to obtain multiple denoising decisions. (b) **Up**: Previous works only consider language-agnostic features to improve the NER model. **Down**: We leverage target language-specific features for model improvement.

Previous solutions for cross-lingual NER can be divided into three categories: (1) Feature-based methods, aiming to design and learn language-agnostic features to improve generalization ability in the target language [WU SJ, 2019]. (2) Translation-based methods, conducting translation and label projection on source language data to obtain labeled target language data [Jain *et al.*, 2019]. (3) Teacher-student learning frameworks [Hinton *et al.*, 2015]. These methods use a model trained on source language data to assign soft pseudo labels for target language data [Ma *et al.*, 2022]. Then, the pseudo-labeled data is used to train a target model.

Recently, teacher-student learning frameworks have gradually become the mainstream solution. For these methods, the performance of the target model improves as the quality of pseudo labels increases. To this end, many improvement strategies have been proposed. Certain methods se-

*Corresponding author

lect high-confidence pseudo-labeled data for training through adversarial discriminators [Chen *et al.*, 2021] or reinforcement learning [Liang *et al.*, 2021]. However, such data selection results in an increased distributional difference and fails to fully utilize target language data. There are also methods integrating multiple source models [Wu *et al.*, 2020; Wu *et al.*, 2021] to generate pseudo labels. However, constructing various data sources to obtain diverse models significantly increases the training costs.

Considering these aforementioned issues, we propose a **G**lobal-**Lo**cal **De**noising Framework (**GLoDe**). Specifically, we leverage the global and local distribution information of the semantic space to denoise pseudo-labeled target language data. With our framework, high-quality data can be obtained without data selection or model ensemble. In detail, at the global level, we propose a prototype-based pseudo-label refinement method. Prototypes can be considered cluster centroids of each type. Conventionally, if a sample is most similar to a certain type's prototype, it is generally regarded as belonging to that specific type. [Zhou *et al.*, 2023] initially leverages this strategy to iteratively improve pseudo labels. However, misclassified samples may exhibit a higher degree of similarity to the prototype of the incorrect type rather than to prototypes of other types. Therefore, we adopt a multiple-decision scheme to tackle it, as shown in Figure 1(a). When the similarity between a sample and the prototype of a particular type exceeds a specific threshold, the type is considered as a potential classification for that sample. To affirm the credibility of similarity thresholds, we adopt dynamic thresholds for each entity type. Consequently, a sample may have more than one potential type, which mitigates the problem of misclassification. At the local level, we leverage the distribution information of the sample's neighborhood. For a given sample, we retrieve its K-nearest neighbors in the semantic space. Based on the distribution of the neighbors' types, we define the similarity of the sample to each type. Then, we compare these similarities with class-specific thresholds to obtain a potential classification for that sample, following the same procedure applied at the global level. By integrating the decision from both global and local levels, we obtain holistic denoising directions for the pseudo label of samples.

Additionally, prior methods typically improve the cross-lingual NER with language-agnostic features, adopting alignment tasks such as adversarial training or contrastive learning [Zhou *et al.*, 2023], as shown in Figure 1(b). However, target language-specific features are never considered but are also beneficial. Therefore, we employ the Masked Language Model (MLM) task, which necessitates solely unlabeled target language data. This simple auxiliary task further enhances the model's understanding of the target language.

To sum up, our contributions can be concluded as follows:

- We design a global-local denoising framework that comprehensively leverages information within the semantic space to rectify the incorrect pseudo labels. A multiple-decision scheme with dynamic class-specific thresholds is devised to mitigate the problem of misclassification and significantly improve the quality of pseudo labels.
- We propose to improve the cross-lingual NER with target language-specific features, which are never considered by previous work. To this end, we devise an auxiliary MLM task, which is both simple and beneficial.
- Experimental results on two benchmark datasets with six target languages demonstrate the superior performance of our approach.

2 Related Work

Existing methods for cross-lingual NER mainly fall into three categories: feature-based methods, translation-based methods, and teacher-student learning methods.

Feature-based Methods These methods learn language-agnostic features so that the model trained on the source language data can better adapt to the target language. Earlier work leverages gazetteers [Zirikly and Hagiwara, 2015] and Wikifier features [Tsai *et al.*, 2016]. The introduction of multilingual pre-trained language models has significantly improved the cross-lingual representations [Conneau *et al.*, 2020; Pires *et al.*, 2019]. Some methods incorporate semantic alignment for further improvement through adversarial training [Keung *et al.*, 2019] and contrastive learning [Wu and Dredze, 2020]. However, coarse-grained alignment is useless for NER. Entity-level alignment is advantageous, but the lack of ground truth labels in target language data inevitably leads to incorrect alignment. Furthermore, these methods do not acquire target language-specific features.

Translation-based Methods These methods translate the source language data and perform label projection to acquire labeled data in the target language. Some approaches employ word-to-word [Xie *et al.*, 2018a] or phrase-to-phrase translation [Mayhew *et al.*, 2017], with entity labels directly linked to the translated content. However, substantial semantic information is sacrificed during translation. Therefore, other approaches consider translating the entire sentence [Jain *et al.*, 2019], but the label projection process will introduce noise labels. [Liu *et al.*, 2021] introduces a placeholder-based approach that combines the aforementioned two strategies. [Yang *et al.*, 2022] adopts the back-translation scheme. Nevertheless, the processes of translation and label projection inevitably carry the risk of introducing noise.

Teacher-student Learning Methods These methods leverage a source model (e.g., a model trained on source language data) to assign soft pseudo labels for target language data, subsequently utilizing this data to train a new model [Wu *et al.*, 2020]. Soft labels offer more information and reduce the risk of overconfidence compared to hard labels [Hinton *et al.*, 2015]. However, this remains insufficient. The key question is the quality of pseudo labels. [Ma *et al.*, 2022; Ge *et al.*, 2023] leverage the concept of feature-based methods to train the source model, aiming to generate higher-quality pseudo labels. [Chen *et al.*, 2021; Liang *et al.*, 2021] attempt to select a subset of the pseudo-labeled data with high confidence. However, data selection prevents the model from fully learning the characteristics of the target language data. Introducing multiple models [Li *et al.*, 2022] is also widely used. [Wu *et al.*, 2021; Zhao *et al.*, 2022] integrate the predictions of multiple models as the final pseudo labels. [Zeng *et al.*, 2022;

Ma *et al.*, 2023] allow multiple models to generate pseudo labels for each other separately. However, various data sources are needed to train diverse models. For instance, [Zhao *et al.*, 2022] introduces source language data, translation data, and data with disrupted word order.

3 Methodology

In this section, we describe our proposed framework GLoDe. Figure 2(a) presents the overall architecture of our approach.

3.1 Overall Framework

Span-based NER Task

Following previous works [Fu *et al.*, 2021; Zhou *et al.*, 2023], we formulate NER as a span classification problem. Formally, given a sentence $X = \{x_1, x_2, \ldots, x_n\}$, our objective is to assign entity labels to each possible span $s_i = \{x_{b_i}, x_{b_i+1}, \ldots, x_{e_i}\}$ in X, where x_{b_i} and x_{e_i} denote the beginning and end tokens of s_i.

For an input sentence X, we first obtain its token representations through a pre-trained language model (PLM): $H = \{h_i\}_{i=1}^{n} = \text{PLM}(X)$. Let $S(X) = \{s_i\}_{i=1}^{m}$ denote all possible spans in X. For each $s_i \in S(X)$, we introduce additional span length embedding l_i and span morphology embedding m_i, which are both learnable embedding parameters. Then, we formulate the span representation by integrating the representations of the beginning and end tokens, the length embedding, and the morphology embedding: $z_i = [h_{b_i}; h_{e_i}; l_i; m_i]$. Finally, z_i is fed into a linear classifier to yield its probability distribution $p_i \in \mathbb{R}^{|\mathbb{C}|}$, where $\mathbb{C}$ is the entity label set.

For source language data D_{src}, the ground truth label y_i for each span s_i is available. The loss function is defined as:

$$L_{src} = \frac{1}{N} \sum_{X \in D_{src}} \frac{1}{|S(X)|} \sum_{s_i \in S(X)} \text{CE}(p_i, y_i) \quad (1)$$

where $\text{CE}(\cdot, \cdot)$ denotes the cross-entropy loss, and N is the number of sentences in the dataset.

Masked Language Model Task

Due to the scarcity of labeled data in the target language, many prior studies further align the source and target languages to improve the model's performance. However, their emphasis remains on language-agnostic features. Instead, we conduct the masked language model task on target language data. As a widely used pre-trained task, it facilitates us to fully exploit the knowledge in the multilingual pre-trained language model and improve the model's generalization ability with target language-specific features.

Formally, given a sentence $X = \{x_i\}_{i=1}^{n}$ in target language data D_{tgt}, we obtain the masked sentence $X^{'} = \{x_i^{'}\}_{i=1}^{n}$, where each token $x_i^{'}$ may be equal to either the original value x_i or a special token "[MASK]" based on a sample rate r_m. Then we obtain token representations $H^{'} = \{h_i^{'}\}_{i=1}^{n}$ for the masked sentence through the PLM. Finally, $H^{'}$ is passed through a masked language model head (MLM Head) to yield their probability distributions $p_i^{'} \in \mathbb{R}^{|\mathbb{V}|}$, where $\mathbb{V}$ is the vocabulary set of PLM.

For $x_i^{'} \in X^{'}$, its label $y_i^{'}$ is defined as the token id of $x_i \in X$ in the vocabulary set $\mathbb{V}$. The loss function is:

$$L_{mlm} = \frac{1}{N} \sum_{X \in D_{tgt}} \frac{1}{|X'|} \sum_{x_i^{'} \in X'} \mathbb{I}(x_i^{'} = [MASK]) \cdot \text{CE}(p_i^{'}, y_i^{'}) \quad (2)$$

where $\mathbb{I}(\cdot)$ is the indicator function.

Training Objectives

In our framework, we simultaneously conduct the NER task on source language data and the MLM task on target language data to obtain a source model. Then, we use the source model to assign initial pseudo labels to the target language data. This enables the generation of higher-quality initial pseudo labels, establishing a robust foundation for subsequent training. Here the training objective for the source model is defined as:

$$L_S = L_{src} + \omega L_{mlm} \quad (3)$$

where ω is a weight hyperparameter.

After obtaining the source model, we train the target model with our proposed framework. The loss function for the NER task on target language data is defined as follows:

$$L_{tgt} = \frac{1}{N} \sum_{X \in D_{tgt}} \frac{1}{|S(X)|} \sum_{s_i \in S(X)} \text{CE}(p_i, \hat{p_i}) \quad (4)$$

where $\hat{p_i}$ is the soft pseudo label generated by the source model. As shown in Figure 2(a), the soft pseudo label is refined by the global-local denoising mechanism. We will describe the denoising mechanism in the next section. Finally, the total objective for the target model is defined as follows:

$$L_T = L_{src} + L_{tgt} + \omega L_{mlm} \quad (5)$$

where ω takes the same value adopted in Eq.3.

3.2 Global-Local Denoising Mechanism

As shown in Figure 2(b), two parallel decision strategies are employed to determine the final denoising directions for pseudo labels of target language data.

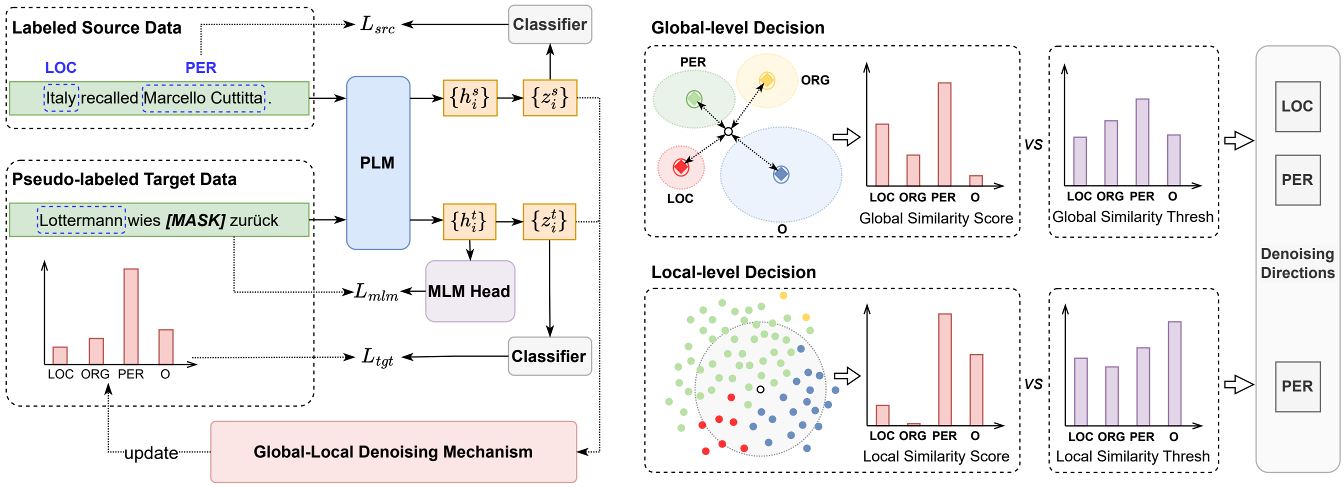
Global-level Decision

This module establishes a set of denoising directions based on the global information within the semantic space. The procedure is shown in the upper part of Figure 2(b).

Global Similarity Score We employ the prototype [Snell *et al.*, 2017] of each entity type to describe the global distribution of the semantic space. The prototype ϕ_c of entity type $c \in \mathbb{C}$ serves as the clustering center of samples belonging to this type, which is robust to outliers to some extent [Xie *et al.*, 2018b]. Following previous works [Ge *et al.*, 2023; Zhou *et al.*, 2023], we adopt a moving average scheme based on batch data to compute ϕ_c, ensuring computational efficiency and a stable training process. Specially, for a batch of source and target language span representations $\{z_i\}$, ϕ_c is computed and updated as follows:

$$\phi_c^{t+1} = \text{Normalize}(\alpha \phi_c^t + (1 - \alpha) z_i),\ \forall i \in \{i | y_i = c\} \quad (6)$$

where t denotes the training step, and y_i is either the ground truth label for the source language span or the pseudo label



(a) Overall Architecture (b) Explanation of Global-Local Denoising Mechanism

Figure 2: (a) Overall architecture of GLoDe. The source and target language sentences are first fed to the PLM to obtain their token representations h_i^s, h_i^t and following span representations z_i^s, z_i^t. Then z_i^s and z_i^t are input into a classifier for the NER task. h_i^t are fed to a masked language model head (MLM Head) for the masked language model task. z_i^s and z_i^t are further utilized for pseudo label denoising. (b) Explanation of the global-local denoising mechanism. We first compute the similarity score of the span by its cosine distances to prototypes (global level) and its neighbors' types (local level). Then we compare the similarities with class-specific thresholds to obtain denoising directions. Decisions from both global and local levels are further integrated.

for the target language span. α is the update rate coefficient. $L2$ normalization is adopted here.

Based on these prototypes, we use the dot product to compute the similarity between a target language span z_i and an entity type c: $sim_i^c = z_i \cdot \phi_c$. Similarly, we obtain the similarities between z_i and all types: $Sim_i = \{sim_i^c\}_{c \in \mathbb{C}}$, which we call "global similarity score".

Global Similarity Threshold We determine thresholds for each entity type to assess the aforementioned similarity scores. Instead of employing fixed thresholds, we adopt the average similarity score of each type as dynamic ones and update them at the end of each training epoch to ensure their validity and training stability. Specifically, we first segment the target language spans based on their pseudo-labeled entity types: $Z_c = \{z_i | \hat{y}_i = c, \hat{y}_i = \text{argmax}(\hat{p}_i)\}$, where $\hat{p}_i$ is the soft pseudo label for z_i. Then the similarity threshold for type c is defined as follows:

$$sim_T^c = \frac{1}{|Z_c|} \sum_{z_i \in Z_c} sim_i^c \tag{7}$$

We obtain the similarity thresholds for all entity types: $Sim_T = \{sim_T^c\}_{c \in \mathbb{C}}$, called "global similarity thresh".

Global Denoising Directions Conventional approaches typically consider solely the type with the highest similarity score. However, a significant portion of misclassified samples are most similar to their incorrect types' prototypes. Thus, we consider more types rather than the most similar ones. For a target language span z_i, we compare its similarity score Sim_i with the similarity thresh Sim_T to determine the update directions for its pseudo label. We describe the update directions as a binary vector:

$$d_i = [b_i^1, b_i^2, \ldots, b_i^{|\mathbb{C}|}], \text{ where } b_i^c = \mathbb{I}(sim_i^c > sim_T^c) \tag{8}$$

Take the upper part of Figure 2(b) for example. Suppose we assume the entity types are $\{$LOC, ORG, PER, O$\}$ ("O" denotes non-entity), the pseudo label denoising directions (LOC and PER) can be expressed as the vector $[1, 0, 1, 0]$.

In addition, considering the class imbalance problem (i.e., non-entity spans constitute a substantial portion) [Zhou *et al.*, 2023], we handle the non-entity type differently. If a span has the highest similarity score with type "O", we directly take "O" as an update direction for its pseudo label, regardless of whether this similarity score surpasses the threshold.

Local-level Decision

As shown in the lower part of Figure 2(b), the procedure of this module aligns with that of Global-level Decision. The main difference lies in the information employed.

Local Similarity Score We leverage the types of a sample's neighbors to describe the local distribution of the semantic space. The use of robust learning techniques, which leverage neighbor information, is extensively employed to effectively handle noisy data [Ma *et al.*, 2023]. We utilize the entity type distribution of a sample's neighbors to indicate the sample's similarity with each entity type. Specifically, we first collect source and target language span representations to construct a representation repository D_z. We update D_z at the end of each training epoch to accommodate changes in representation distribution during the training process. Then, for a target language span representation z_i, we retrieve its K-nearest neighbors $N_k(z_i)$ from D_z. Finally we compute the similarity between z_i and entity type c based on entity types of spans in $N_k(z_i)$:

$$sim_i^c = \frac{1}{K} \sum_{z_j \in N_k(z_i)} \mathbb{I}(y_j = c) \tag{9}$$

where y_j is either the ground truth label for the source language span or the pseudo label for the target language span. Similarly, we obtain the similarities between z_i and all types: $Sim_i = \{sim_i^c\}_{c \in \mathbb{C}}$, which we call "local similarity score".

Local Similarity Threshold and Denoising Directions Based on the local similarity score acquired above, we employ the same strategy and computation process for the global similarity threshold in Eq.7 to obtain the local similarity threshold. Subsequently, following Eq.8, we obtain denoising directions by utilizing the local similarity score and the local similarity threshold. We omit the detailed description here for brevity.

Pseudo Label Refinement

Noise samples exhibit a range of styles, including isolated points situated away from centroids and fuzzy points near classification boundaries. To derive the final denoising verdict, we integrate decisions made on both global and local levels. This strategy assists in detecting and addressing a variety of noise patterns. Specifically, for a target language span z_i, assuming its global-level and local-level denoising direction vectors are d_i^g and d_i^l, the update decision for its pseudo label is defined as follows:

$$u_i = \text{Normalize}(d_i^g + d_i^l) \tag{10}$$

Let the current pseudo label of z_i be p_i^t. The updated pseudo label is computed as follows:

$$p_i^{t+1} = \text{Normalize}(\beta p_i^t + (1-\beta)u_i) \tag{11}$$

where β is the update rate coefficient. $L1$ normalization is adopted in Eq.10 and Eq.11. With the improvement of pseudo labels, the referential value of the update decision u_i increases. Thus, we uniformly decrease β to increase the impact of u_i during training.

There might exist multiple update directions with different updating degrees as shown in Figure 2(b) (e.g., PER has a higher updating degree than LOC). This allows misclassified spans to escape from their original incorrect types and adjust to high-confidence types, which are more likely to be correct. During the updating procedure, information within the semantic space is gradually integrated into the soft pseudo labels, which is beneficial for improving the model's generalization ability to the target language data.

4 Experiments

4.1 Experiment Settings

Datasets

We conduct experiments on two widely-used benchmarks: CoNLL [Sang and Erik, 2002; Sang and De Meulder, 2003] and WikiAnn [Pan *et al.*, 2017]. The CoNLL dataset comprises four languages: English (en), German (de), Spanish (es), and Dutch (nl). The dataset is annotated with four entity types: PER, LOC, ORG, and MISC. The WikiAnn dataset comprises four languages as well: English (en), Arabic (ar), Hindi (hi), and Chinese (zh). The dataset is annotated with three entity types: PER, LOC, and ORG.

Following previous works [Wu *et al.*, 2021; Zhao *et al.*, 2022], we take the English data in each dataset as the source language data and other languages as the target language. We remove entity labels from training sets of target languages, taking them as the unlabeled target language data.

Method	**de**	**es**	**nl**	**Avg**
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
CROP	80.10	78.10	79.50	79.23
TransAdv	75.52	80.93	83.78	80.08
ProKD	78.90	82.62	79.53	80.35
CoLaDa	**81.12**	82.70	85.15	82.99
ContProto	76.41	85.02	83.69	81.71
GLoDe(Ours)	79.15	**85.29**	**85.76**	**83.40**

Table 1: Experimental results on CoNLL. Results on German (de), Spanish (es), Dutch(nl), and their average performance are reported.

Implementation Details

Following previous works [Zeng *et al.*, 2022; Ma *et al.*, 2023], we use XLM-R-large [Conneau *et al.*, 2020] as the backbone pre-trained language model. The model is trained for 8 epochs (except 6 epochs for zh) and the batch size is 32. We use AdamW as the optimizer and set the learning rate to 1e-5. We set the mask rate r_m to 0.25 and loss weight ω to $1e-3$ ($1e-5$ for es) for the MLM task. To improve the training efficiency, we map span representations to a low-dimensional space of 128 dimensions for the denoising computations. For neighbor retrieval, we set K to 300 and utilize the FAISS tool [Johnson *et al.*, 2019]. The update rate coefficient $\alpha = 0.99$ in Eq.6. β in Eq.11 uniformly decreases from 0.95 to 0.80 during training.

Baselines

We compare our method with previous baselines as follows: (1) **AdvPicker** [Chen *et al.*, 2021] employs an adversarial discriminator to select pseudo-labeled target language data with high confidence for training. (2) **RIKD** [Liang *et al.*, 2021] leverages reinforcement learning for selecting pseudo-labeled data and iteratively conducts training for multiple rounds. (3) **MTMT** [Li *et al.*, 2022] introduces an entity similarity evaluator to support the training of the NER task. (4) **CROP** [Yang *et al.*, 2022] transposes the NER task from the target language to the source language via back-translation and entity matching. (5) **TransAdv** [Zhao *et al.*, 2022] introduces multiple data sources to obtain distinct source models for comprehensive pseudo labels. (6) **ProKD** [Ge *et al.*, 2023] proposes a contrastive learning-based prototype alignment approach to boost the model's ability to capture language-agnostic knowledge. (7) **CoLaDa** [Ma *et al.*, 2023] utilizes two distinct models to denoise the labels of each other and further considers the label consistency to re-weight the noisy labels. (8) **ContProto** [Zhou *et al.*, 2023] incorporates contrastive learning and prototype-based pseudo-labeling for representation learning and pseudo-label refinement.

Method	**ar**	**hi**	**zh**	**Avg**
AdvPicker†	39.75	71.94	55.24	55.65
RIKD	45.96	70.28	50.40	55.55
MTMT	52.77	70.76	52.26	58.60
CROP	48.00	77.10	50.30	58.47
TransAdv	42.53	74.24	54.25	49.17
ProKD	50.91	70.72	51.80	57.81
CoLaDa	66.94	76.69	60.08	67.90
ContProto	72.20	83.45	61.47	72.37
GLoDe(Ours)	**74.35**	**83.61**	**64.81**	**74.26**

Table 2: Experimental results on WikiAnn. Results on Arabic (ar), Hindi (hi), Chinese (zh), and their average performance are reported. † denotes results obtained by running their public code on the data.

4.2 Main Results

Tables 1 and 2 present the results of our proposed method and previous baselines on CoNLL and WikiAnn. Following previous works, we employ the entity-level micro-F1 score as the metric. We conducted all experiments three times and calculated the average value as the final result.

It can be seen that our proposed GLoDe outperforms prior methods, achieving an improvement of 0.41 F1 scores on average for CoNLL and 1.89 F1 scores on average for WikiAnn. This demonstrates the effectiveness of our approach. Besides, there are two noteworthy findings evident in the results: (1) Our model reaches suboptimal performance on German (de). The German dataset has fewer entities, and numerous sentences do not contain entities, potentially leading the model to predict more false positive entities. Additionally, both CoLaDa and CROP incorporate translated content, which facilitates the model to acquire more entity information in the target language. (2) For target languages (e.g., es, nl) closer to the source language (en), previous works have shown incremental improvements in a moderate trend. Conversely, for distant source-target pairs (e.g., en-ar, en-zh), models performing the denoising process on pseudo-labeled target language data (ours, ContProto, and CoLaDa) have shown significant improvements over other models. We attribute this to the critical label noise problem in these distant languages.

4.3 Ablation Study

To further validate the effectiveness of components in our proposed model, we conduct the following ablation studies: (1) *w/o mlm*, which removes the masked language model task during training; (2) *w/o denoising*, which removes the global-local denoising mechanism; (3) *w/o denoising-local*, which only leverages global information in the semantic space for pseudo label denoising; (4) *w/o denoising-global*, which only leverages local information in the semantic space for pseudo label denoising; (5) *w/o multiple directions*, which only considers the type with the highest similarity score as the potential update direction for the pseudo label; (6) *w/o mlm init*, which omits the masked language model task, training the source model solely on source language data to acquire the initial pseudo labels.

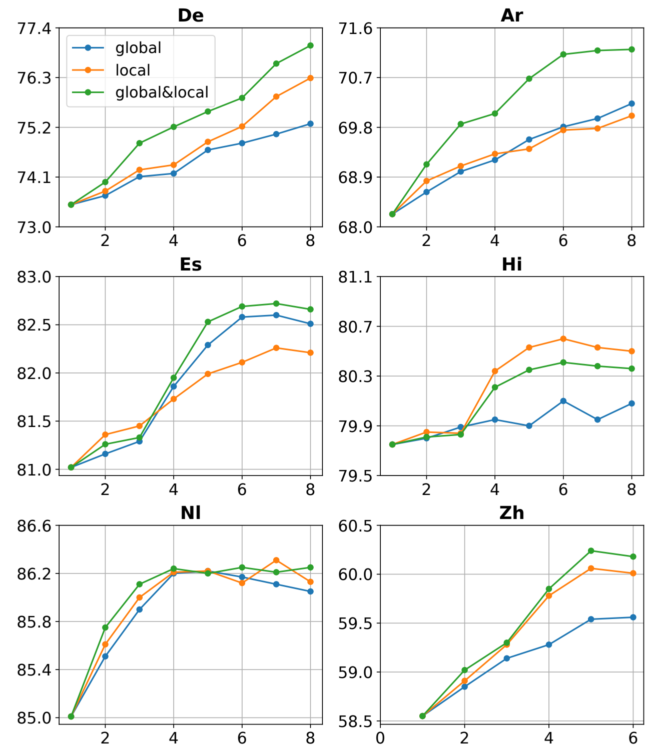


Figure 3: Quality of pseudo labels. The horizontal axis is the epoch number and the vertical axis is the entity-level F1 scores of pseudo labels. Three types of information for denoising are compared.

Table 3 shows the ablation study results. It can be seen that: (1) Both *w/o mlm* and *w/o denoising* diminish the model's performance, which demonstrates the effectiveness of these two modules. Both of them improve the model's generalization ability on the target language. Additionally, removing the denoising mechanism exerts a more significant impact on most datasets, emphasizing the significance of noise reduction for pseudo-labeled data. (2) *w/o denoising-local* and *w/o denoising-global* bring less performance decline compared to *w/o denoising*. This implies that leveraging either local or global information in the semantic space aids in pseudo-label denoising. Naturally, employing both together yields superior results. Moreover, there seems to be no absolute answer to the question regarding which part of the information is more useful. (3) *w/o multiple directions* also improves the performance compared to *w/o denoising*, but it falls short of comparison to other denoising schemes. This implies that incorporating multiple denoising directions would be more beneficial. (4) *w/o mlm init* denotes the significance of improving the source model with target language-specific features. Additionally, the quality of the initial pseudo labels significantly constrains the target model's performance.

4.4 Quality of Pseudo Labels

In this section, we assess the quality of pseudo labels during training. As in the prior experiments, we measure the quality by entity-level micro-F1 scores. We compare the following three pseudo label denoising strategies: (1) *global*, which

Method	de	es	nl	ar	hi	zh
GLoDe	**79.15**	**85.29**	**85.76**	**74.35**	**83.61**	**64.81**
- w/o mlm	77.48(-1.67)	84.28(-1.01)	84.63(-1.13)	72.26(-2.09)	82.08(-1.53)	63.47(-1.34)
- w/o denoising	77.07(-2.08)	84.11(-1.18)	84.04(-1.72)	70.83(-3.52)	82.64(-0.97)	63.44(-1.37)
- w/o denoising-local	77.49(-1.66)	84.26(-1.03)	84.46(-1.30)	71.85(-2.50)	81.69(-1.92)	64.03(-0.78)
- w/o denoising-global	77.81(-1.34)	84.18(-1.11)	84.82(-0.94)	71.86(-2.49)	82.42(-1.19)	64.37(-0.44)
- w/o multiple directions	77.41(-1.74)	84.20(-1.09)	84.23(-1.53)	71.24(-3.11)	82.50(-1.11)	63.46(-1.35)
- w/o mlm init	74.75(-4.40)	84.05(-1.25)	84.76(-1.00)	65.35(-9.00)	81.29(-2.32)	61.07(-3.74)

Table 3: Ablation studies on CoNLL and WikiAnn dataset. In this table, we remove specific components in our model for comparison. Values in brackets denote the performance drop compared to the full model.

only leverages global information in the semantic space for pseudo label denoising; (2) *local*, which only leverages local information; (3) *global&local*, which leverages both global and local information, serving as the solution we adopt.

As shown in Figure 3, the quality of pseudo labels improves progressively during training. The utilization of either global or local information improves pseudo labels, with the combined use of both resulting in even greater improvement for most of the target languages. We argue that the two types of information can complement each other and they have different impacts for different target languages. For example, *local* yields more advantages for De and Zh, while *global* provides greater benefits for Es. For Nl, little difference is observed among the three strategies, with limited improvement obtained after 4th epoch. We attribute this to the high similarity between Nl and the source language (En). The information from either the global level or the local level may be sufficient. For Hi, *global* may even exhibit a negative impact. Given that the source language contains nearly five times as many entities as Hi, local-level information proves to be more helpful than global-level information.

4.5 Effectiveness of Language-specific Features

The auxiliary task improves both the source and target model with target language-specific features. In this section, we further analyze this impact.

As shown in Figure 4, we compare entity distributions from different data sources. Specifically, the source language entity distributions (blue lines), target language entity distributions (green lines), and model-predicted entity distributions for the target language (orange lines) are considered. The predicted distribution in Figure 4(a) is obtained from the model trained solely with source language data (**Plain**-model), and that in Figure 4(b) is obtained from the model trained with both source language data and the auxiliary task (**Aux**-model). We can observe that the predicted distribution shifts towards the target language distribution after introducing target language-specific features by the auxiliary task.

Additionally, we conducted some comparative experiments to demonstrate the advantages of language-specific features. We replace the masked language model task in our framework with the supervised contrastive learning task, which has recently been widely used to improve the model with language-agnostic features ([Ge *et al.*, 2023; Zhou *et al.*, 2023]). As shown in Table 4, we denote the model improved by language-agnostic features and language-specific features as $\text{GLoDe}_{\text{agnostic}}$ and $\text{GLoDe}_{\text{specific}}$, respectively. Comparison results demonstrate that language-specific features are more beneficial for model improvement.

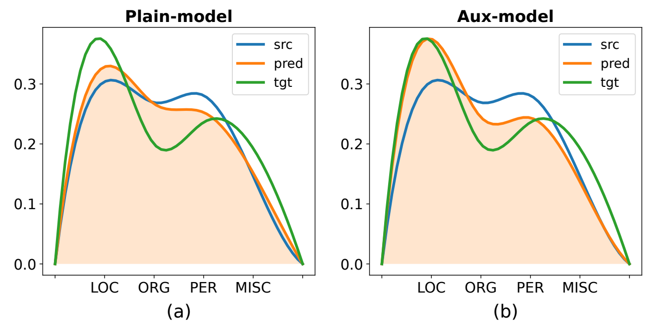


Figure 4: Distribution of entity types. The vertical axis is the percentage of the entity type. We take De as the target language here. We interpolate discrete statistical results and represent them with continuous curves for better visualization.

Method	de	es	nl	ar	hi	zh
$\text{GLoDe}_{\text{specific}}$	79.15	85.29	85.76	74.35	83.61	64.81
$\text{GLoDe}_{\text{agnostic}}$	77.52	84.03	83.81	72.31	82.71	62.76

Table 4: Comparision of impacts of target language-specific features and language-agnostic features for model improvement.

5 Conclusion

In this paper, we present GLoDe, a pseudo-label denoising framework for cross-lingual NER. Our global-local denoising mechanism integrates global and local information within the semantic space to rectify as many incorrect pseudo labels for target language data as possible. The information within pseudo labels progressively increases during the denoising process, which helps to improve the model's generalization ability on the target language. Additionally, we further improve our model with target language-specific features, leveraging the masked language model task. Experimental results on six benchmark datasets demonstrate that GLoDe achieves superior performance over previous baselines.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62276110, No. 62172039 and in part by the fund of Joint Laboratory of HUST and Pingan Property & Casualty Research (HPL). The authors would also like to thank the anonymous reviewers for their comments on improving the quality of this paper.

References

[Chen *et al.*, 2021] Weile Chen, Huiqiang Jiang, Qianhui Wu, Börje Karlsson, and Yi Guan. 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*, pages 743–753, 2021.

[Conneau *et al.*, 2020] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, 2020.

[Fan *et al.*, 2023] Chenghao Fan, Wei Wei, Xiaoye Qu, Zhenyi Lu, Wenfeng Xie, Yu Cheng, and Dangyang Chen. Improving low-resource prompt-based relation representation with multi-view decoupling learning. *arXiv preprint arXiv:2312.17267*, 2023.

[Fu *et al.*, 2021] Jinlan Fu, Xuan-Jing Huang, and Pengfei Liu. 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*, pages 7183–7195, 2021.

[Ge *et al.*, 2023] Ling Ge, Chunming Hu, Guanghui Ma, Hong Zhang, and Jihong Liu. Prokd: an unsupervised prototypical knowledge distillation network for zero-resource cross-lingual named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12818–12826, 2023.

[Gu *et al.*, 2022] Yingjie Gu, Xiaoye Qu, Zhefeng Wang, Yi Zheng, Baoxing Huai, and Nicholas Jing Yuan. Delving deep into regularity: A simple but effective method for chinese named entity recognition. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1863–1873, 2022.

[Hinton *et al.*, 2015] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[Jain *et al.*, 2019] Alankar Jain, Bhargavi Paranjape, and Zachary C Lipton. 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*, pages 1083–1092, 2019.

[Johnson *et al.*, 2019] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

[Keung *et al.*, 2019] Phillip Keung, Yichao Lu, and Vikas Bhardwaj. 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, 2019.

[Li *et al.*, 2022] Zhuoran Li, Chunming Hu, Xiaohui Guo, Junfan Chen, Wenyi Qin, and Richong Zhang. 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*, pages 170–179, 2022.

[Liang *et al.*, 2021] Shining Liang, Ming Gong, Jian Pei, Linjun Shou, Wanli Zuo, Xianglin Zuo, and Daxin Jiang. 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, 2021.

[Liu *et al.*, 2021] Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq Joty, Luo Si, and Chunyan Miao. 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*, pages 5834–5846, 2021.

[Ma *et al.*, 2022] Jun-Yu Ma, Beiduo Chen, Jia-Chen Gu, Zhenhua Ling, Wu Guo, Quan Liu, Zhigang Chen, and Cong Liu. Wider & closer: Mixture of short-channel distillers for zero-shot cross-lingual named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5171–5183, 2022.

[Ma *et al.*, 2023] Tingting Ma, Qianhui Wu, Huiqiang Jiang, Börje Karlsson, Tiejun Zhao, and Chin-Yew Lin. CoLaDa: A collaborative label denoising framework for cross-lingual named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5995–6009, Toronto, Canada, July 2023.

[Mayhew *et al.*, 2017] Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. Cheap translation for cross-lingual named entity recognition. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2536–2545, 2017.

[Pan *et al.*, 2017] Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1946–1958, 2017.

[Pires *et al.*, 2019] Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual bert? In *Pro-*

ceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4996–5001, 2019.

[Qu *et al.*, 2023a] Xiaoye Qu, Yingjie Gu, Qingrong Xia, Zechang Li, Zhefeng Wang, and Baoxing Huai. A survey on arabic named entity recognition: Past, recent advances, and future trends. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[Qu *et al.*, 2023b] Xiaoye Qu, Jun Zeng, Daizong Liu, Zhefeng Wang, Baoxing Huai, and Pan Zhou. Distantly-supervised named entity recognition with adaptive teacher learning and fine-grained student ensemble. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13501–13509, 2023.

[Sang and De Meulder, 2003] Erik Tjong Kim Sang and Fien De Meulder. 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, 2003.

[Sang and Erik, 2002] Tjong Kim Sang and F Erik. Introduction to the conll-2002 shared task: language-independent named entity recognition. In *Proceedings of CoNLL-2002/Roth, Dan [edit.]*, pages 155–158, 2002.

[Snell *et al.*, 2017] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 2017.

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

[Wei *et al.*, 2021] Wei Wei, Zanbo Wang, Xianling Mao, Guangyou Zhou, Pan Zhou, and Sheng Jiang. Position-aware self-attention based neural sequence labeling. *Pattern Recognition*, 110:107636, 2021.

[Wu and Dredze, 2020] Shijie Wu and Mark Dredze. Do explicit alignments robustly improve multilingual encoders? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4471–4482, 2020.

[Wu *et al.*, 2020] Qianhui Wu, Zijia Lin, Börje Karlsson, Jian-Guang Lou, and Biqing Huang. 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, 2020.

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

[WU SJ, 2019] DREDZE M WU SJ. 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, Hong Kong, China*, pages 833–844, 2019.

[Xie *et al.*, 2018a] Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A Smith, and Jaime G Carbonell. 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, 2018.

[Xie *et al.*, 2018b] Shaoan Xie, Zibin Zheng, Liang Chen, and Chuan Chen. Learning semantic representations for unsupervised domain adaptation. In *International conference on machine learning*, pages 5423–5432, 2018.

[Yang *et al.*, 2022] Jian Yang, Shaohan Huang, Shuming Ma, Yuwei Yin, Li Dong, Dongdong Zhang, Hongcheng Guo, Zhoujun Li, and Furu Wei. Crop: Zero-shot cross-lingual named entity recognition with multilingual labeled sequence translation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022.

[Zeng *et al.*, 2022] Jiali Zeng, Yufan Jiang, Yongjing Yin, Xu Wang, Binghuai Lin, and Yunbo Cao. Dualner: A dual-teaching framework for zero-shot cross-lingual named entity recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1837–1843, 2022.

[Zhao *et al.*, 2022] Yichun Zhao, Jintao Du, Gongshen Liu, and Huijia Zhu. Transadv: A translation-based adversarial learning framework for zero-resource cross-lingual named entity recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 742–749, 2022.

[Zhou *et al.*, 2023] Ran Zhou, Xin Li, Lidong Bing, Erik Cambria, and Chunyan Miao. Improving self-training for cross-lingual named entity recognition with contrastive and prototype learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 4018–4031, Toronto, Canada, July 2023. Association for Computational Linguistics.

[Zhu *et al.*, 2023] Tong Zhu, Junfei Ren, Zijian Yu, Mengsong Wu, Guoliang Zhang, Xiaoye Qu, Wenliang Chen, Zhefeng Wang, Baoxing Huai, and Min Zhang. Mirror: A universal framework for various information extraction tasks. *arXiv e-prints*, pages arXiv–2311, 2023.

[Zirikly and Hagiwara, 2015] Ayah Zirikly and Masato Hagiwara. Cross-lingual transfer of named entity recognizers without parallel corpora. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 390–396, 2015.