Large Margin Representation Learning for Robust Cross-lingual Named Entity Recognition

Guangcheng Zhu[1,2] **Ruixuan Xiao**[1,2] **Haobo Wang**[1,2] **Zhen Zhu**[3]
Gengyu Lyu[4] **Junbo Zhao**[1,2]*
[1]Zhejiang University
[2]Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security
[3]Zhejiang Sci-tech University [4]Beijing University of Technology
{zhuguangcheng,xiaoruixuan,wanghaobo,j.zhao}@zju.edu.cn
hzzhuzhen@yeah.net,lyugengyu@gmail.com

Abstract

Cross-lingual named entity recognition (NER) aims to build an NER model that generalizes to the low-resource target language with labeled data from the high-resource source language. Current state-of-the-art methods typically combine self-training mechanism with contrastive learning paradigm, in order to develop discriminative entity clusters for cross-lingual adaptation. Despite the promise, we identify that these methods neglect two key problems: distribution skewness and pseudo-label bias, leading to *indistinguishable entity clusters with small margins*. To this end, we propose a novel framework, MARAL, which optimizes an adaptively reweighted contrastive loss to handle the class skewness and theoretically guarantees the optimal feature arrangement with maximum margin. To further mitigate the adverse effects of unreliable pseudo-labels, MARAL integrates a progressive cross-lingual adaptation strategy, which first selects reliable samples as anchors and then refines the remaining unreliable ones. Extensive experiments demonstrate that MARAL significantly outperforms the current state-of-the-art methods on multiple benchmarks, e.g., **+2.04%** on the challenging MultiCoNER dataset. Our code is available at https://github.com/gczhu/MARAL.

1 Introduction

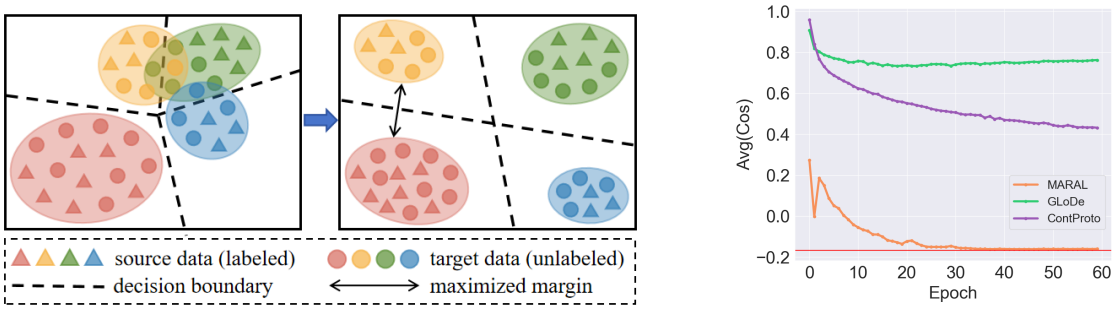
Named Entity Recognition (NER) is a fundamental topic in the field of information extraction (Li et al., 2020a), which aims to identify entity spans in the given raw text and categorize them into predefined entity types. The success of current deep learning models for NER largely hinges on large-scale annotated training data, which can be extremely expensive and even unrealistic for low-resource languages, such as Hindi. To address such data scarcity, cross-lingual NER (Huang et al., 2019; Bari et al., 2020) has gained wide attention, which leverages labeled data from high-resource languages, such as English, to tackle the NER task in low-resource target languages.

To bridge the gap between high-resource source language and low-resource target language, the mainstream approaches (Wu et al., 2020a,b; Liang et al., 2021; Ma et al., 2022; Ge et al., 2024) mostly rely on multilingual pre-trained language models coupled with the self-training framework, yielding considerable performance improvement. These methods typically follow a teacher-student learning mechanism that comprises two key steps: (i)-train a teacher model on labeled source-language data; (ii)-train a student model on unlabeled target languages with pseudo-labels produced by the teacher model (probably combined with labeled source data). To further enhance cross-lingual transferability, state-of-the-art algorithms (Ge et al., 2023; Zhou et al., 2023; Mo et al., 2024) also integrate the contrastive learning paradigm to explicitly align the same entity classes between the source and target languages in the shared representation space.

Essentially, the central theme of cross-lingual NER methods is to learn a language-agnostic feature distribution that accurately reflects the latent distribution of each entity type and generalizes well to the target language. According to classical learning theory (Cao et al., 2019; Zhou et al., 2022b), the optimal generalization error is achieved under the maximum margin assumption, where each entity cluster is tightly distributed to their class centroids, and different clusters are separated with maximum margin. However, through our extensive experiments, we find that current best cross-lingual NER methods fail to achieve this goal. As shown in Figure 1(b), the state-of-the-art algorithms ContProto

*Corresponding author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4270–4291



(a) Effect of the margin on cross-lingual generalization

(b) Average pair-wise cosine similarity

Figure 1: (a) Left: Feature distribution with small margin, where tightly distributed clusters increase the risk of misclassifying unseen target data near decision boundaries. Right: Ideal feature distribution characterized by maximized margin, demonstrating improved cross-lingual generalization. (b) Average pair-wise cosine similarity between class centroids on Italian (it) language, with the red line indicating the theoretical minimal value ($-\frac{1}{K-1}$, see Theorem 1) for maximally separated clusters.

(Zhou et al., 2023) and GLoDe (Ding et al., 2024) exhibit remarkably high pair-wise cosine similarity between class centroids. As illustrated in the left of Figure 1(a), tightly intertwined clusters hinder the classifier from distinguishing unseen target-language data which typically deviates from cluster centroids in the cross-lingual setup. In contrast, the right of Figure 1(a) presents the ideal situation where the classifier demonstrates great generalization ability.

To remedy this, we first identify two key reasons why existing algorithms fail to achieve an optimal margin. (1) **Distribution Skewness**: Different classes often exhibit highly imbalanced frequencies. Our analysis reveals a significant distribution skew in existing datasets. For example, in the MultiCoNER dataset (Fetahu et al., 2023), the imbalance ratio between O and MED classes can be as high as 600. This results in minority classes collapsing into very small clusters, making it hard for them to occupy a large margin. (2) **Pseudo-label Bias**: During cross-lingual adaptation, pseudo-labeling frequently introduces confusion, leading to insufficient separation between features of easily misclassified classes. Class imbalance exacerbates this issue, as minority classes, being underrepresented, tend to be heavily biased toward majority classes.

Accordingly, we propose a novel **M**aximum-m**AR**gin **AL**ignment framework, **MARAL**, for robust and generalizable cross-lingual NER. From an information-theoretic perspective, we develop a new adaptively reweighted contrastive learning algorithm that accounts for imbalanced class probabilities in the contrastive optimization procedure. Theoretically, we show that this new loss asymptotically achieves the optimal feature arrangement for maximum-margin representations. To further enhance model robustness, we propose a Progressive Cross-lingual Adaptation (PCLA) strategy to mitigate the adverse effects of unreliable pseudo-labels from the target language during the margin adjustment process. It comprises two sequential steps: (i)-selects reliable pseudo-labels as anchors based on class-specific probability densities; (ii)-refines the remaining unreliable pseudo-labels through a dual-alignment strategy. Such an adaptation mechanism can largely improve the quality of pseudo-labels and thus facilitate margin calibration. Empirically, MARAL significantly outperforms current state-of-the-art methods on different benchmarks, e.g., improves the best baseline by **+2.04%** average F1 scores on the challenging MultiCoNER dataset.

2 Preliminary

Problem Definition. In zero-shot cross-lingual NER, we are provided with a labeled set $\mathcal{D}_{src} = \{(\boldsymbol{X}_i^s, \boldsymbol{Y}_i^s)\}_{i=1}^{N_s}$ of N_s labeled samples in source language, and an unlabeled set $\mathcal{D}_{tgt} = \{(\boldsymbol{X}_i^t)\}_{i=1}^{N_t}$ of N_t unlabeled samples in target language. The goal is to extract potential entities from target-language sequences and correctly classify them into predefined categories, with labeled $\mathcal{D}_{src}$ and unlabeled $\mathcal{D}_{tgt}$ during training.

Following previous works (Fu et al., 2021; Ding et al., 2024), we adopt the span-based NER formulation. Given a sentence $\boldsymbol{X} = \{x_1, x_2, \ldots, x_L\}$ with L tokens, we first employ a pre-trained language model to obtain the hidden representation $\boldsymbol{h}_i$ for each token x_i. Then we consider all possible spans $\boldsymbol{s}_{j:l} = \{x_j, x_{j+1}, \ldots, x_l\}$ in $\boldsymbol{X}$ where $1 \leq j \leq l \leq L$, to construct the set $\mathcal{T}(\boldsymbol{X})$. For the source and target data, we combine $\mathcal{T}(\boldsymbol{X})$ from all

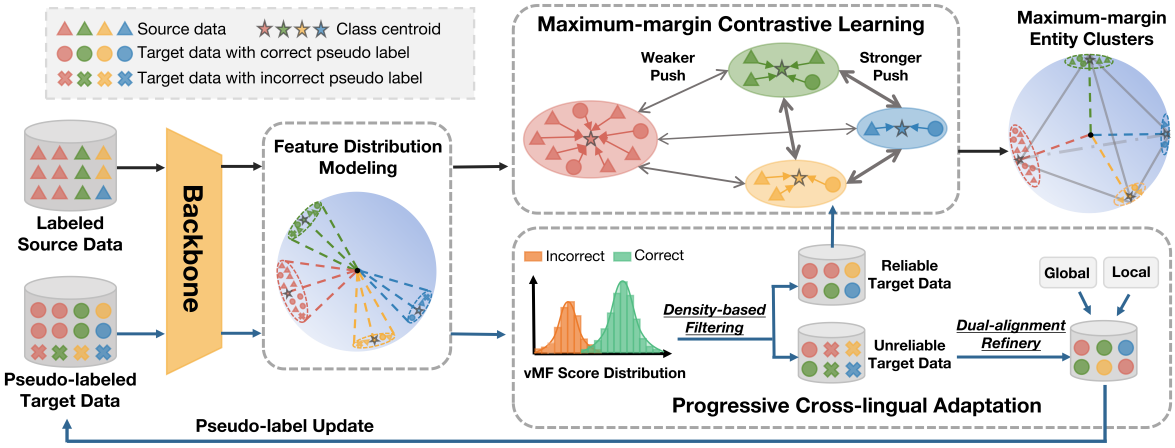


Figure 2: Overview of MARAL. Both source and target data contribute to feature distribution modeling and yield their vMF scores, which are used for sample filtering. Maximum-margin contrastive learning is then built on such estimated distribution and applied only to samples with reliable pseudo-labels. The remaining unreliable ones are refined for subsequent inclusion in the contrastive alignment. These components work synergistically for deriving maximum-margin entity clusters.

sentences to form the complete span sets $\mathcal{T}_s$ and $\mathcal{T}_t$, respectively. In particular, for labeled source data, each span $\boldsymbol{s}_{j:l}$ is equipped with the ground-truth label $y_{j:l} \in \{0, 1, ..., K-1\}$, where $1 \sim K-1$ denote entity types and 0 denotes non-entity. For span $\boldsymbol{s}_{j:l}$, its length and morphology information are embedded as $\boldsymbol{l}_{j:l}$ and $\boldsymbol{m}_{j:l}$. The overall span representation $\boldsymbol{e}_{j:l}$ is formulated by integrating embeddings of beginning token $\boldsymbol{h}_j$, ending token $\boldsymbol{h}_l$, the length and morphology, $\boldsymbol{e}_{j:l} = [\boldsymbol{h}_j, \boldsymbol{h}_l, \boldsymbol{l}_{j:l}, \boldsymbol{m}_{j:l}]$. Finally, $\boldsymbol{e}_{j:l}$ is passed through a two-layer MLP to yield the final q-dimension representation $\boldsymbol{z}_{j:l} \in \mathbb{R}^q$ with unit norm $\|\boldsymbol{z}_{j:l}\|_2 = 1$ (abbreviated as $\boldsymbol{z}_i$ of each span $\boldsymbol{s}_i$ for simplicity in the subsequent text), which is then mapped to the predicted logits $\boldsymbol{p}_{j:l} \in \mathbb{R}^K$.

Definition of Margin. Based on the definition in previous works (Koltchinskii and Panchenko, 2002; Cao et al., 2019), for a model $h : \mathbb{R}^d \to \mathbb{R}^K$ that outputs K-class logits of a sample $\boldsymbol{s}_i \in \mathbb{R}^d$, the margin $\gamma(\boldsymbol{s}_i, y_i)$ of $\boldsymbol{s}_i$ with label y_i is defined as:

$$\gamma(\boldsymbol{s}_i, y_i) = h(\boldsymbol{s}_i)_{y_i} - \max_{j \neq y_i} h(\boldsymbol{s}_i)_j = \boldsymbol{\mu}_{y_i}^\top \boldsymbol{z}_i - \max_{j \neq y_i} \boldsymbol{\mu}_j^\top \boldsymbol{z}_i \quad (1)$$

where $\boldsymbol{z}_i$ is the feature of $\boldsymbol{s}_i$, $\boldsymbol{\mu}_j$ is the prototypical centroid of class j. Let $\mathcal{S}_j$ denote the sample subset of class j. We then define the class margin for class j as $\gamma_j = \min_{\boldsymbol{s}_i \in \mathcal{S}_j} \gamma(\boldsymbol{s}_i, y_i)$, and the overall margin over the dataset as the minimal class margin, $\gamma_{all} = \min\{\gamma_0, \cdots, \gamma_{K-1}\}$. Intuitively, maximizing the overall margin encourages representations close to their corresponding class centroids and far away from the others.

3 Method

In this section, we present our framework MARAL. As shown in Figure 2, MARAL comprises two key components: (1) **Maximum-margin contrastive learning (MMCL)** to establish well-separated representations with maximum margins (Section 3.1); (2) **Progressive cross-lingual adaptation (PCLA)** for the gradual injection of high-quality pseudo-labels to guide contrastive alignment. (Section 3.2).

3.1 Maximum-margin Contrastive Learning

As previously stated, our primary goal is to achieve a feature arrangement that maximizes margin separation. To this end, we first analyze the conditions under which contrastive learning can attain maximum margins from an optimization perspective. Specifically, by considering the underlying feature distribution, we show that the following objective inherently yields the optimal value of γ_{all}.

Proposition 1. *In the contrastive space $\mathbb{S}^{q-1}$ where each sample with ℓ_2-normalized feature $\boldsymbol{z}_i$ and label y_i follows a von Mises-Fisher (vMF) mixture distribution, the contrastive objective that maximizes the overall margin can be derived as follows:*

$$\mathcal{L} = -\mathbb{E}_i \left[\log \frac{\pi(y_i) \exp(\rho(\boldsymbol{z}_i, y_i))}{\sum_{c=0}^{K-1} \pi(c) \exp(\rho(\boldsymbol{z}_i, c))} \right] \quad (2)$$

where $\pi(c)$ is the class prior, $\rho(\boldsymbol{z}, y)$ is the kernel density estimation of $\log p(\boldsymbol{z}|y)$ defined as,

$$\rho(\boldsymbol{z}, y) = \log\left(\mathbb{E}_{\boldsymbol{z}_y \sim vMF(\boldsymbol{\mu}_y, \kappa_y)} \left[\exp(\boldsymbol{z}^\top \boldsymbol{z}_y / \tau)\right]\right) \quad (3)$$

where τ is a temperature parameter. Detailed proof of this proposition is provided in Appendix A.2.

Interestingly, current methods (Zhou et al., 2023; Ge et al., 2023; Mo et al., 2024) applying traditional contrastive learning to cross-lingual NER can be naturally derived from Eq. (2) under the balanced assumption, where π is uniform across classes, and $\rho(\boldsymbol{z}, y)$ is estimated from mini-batch samples (see Appendix A.2 for the derivation). However, this assumption does not hold in the cross-lingual NER. Specifically, due to the skewed distribution, the values of π vary markedly across classes, ignoring which can bias the contrastive process toward majority classes. Besides, the limited batch samples in minority classes often lead to inaccurate estimates of $\rho(\boldsymbol{z}, y)$. To remedy this, the contrastive objective needs to be revised by carefully incorporating the class-wise distribution and accurately estimating $\rho(\boldsymbol{z}, y)$ via explicit distribution modeling.

Feature Distribution Modeling. To capture the intrinsic data distribution, we explicitly model it as an imbalanced von Mises-Fisher (vMF) mixture distribution. Accordingly, the probability density function for a unit embedding $\boldsymbol{z}$ in class c is,

$$f(\boldsymbol{z}|\boldsymbol{\mu}_c, \kappa_c) = C_q(\kappa_c) \exp(\kappa_c \boldsymbol{\mu}_c^\top \boldsymbol{z}) \tag{4}$$

where $\boldsymbol{\mu}_c$ is the mean direction of the class c with $\|\boldsymbol{\mu_c}\|_2 = 1$, $\kappa_c \geq 0$ is the class-specific concentration parameter indicating the tightness around $\boldsymbol{\mu}_c$, and $C_q(\kappa)$ is the normalization factor whose computation is described in Appendix B.3. Under such a vMF distribution, the contrastive features $\boldsymbol{z}_y$ of class y exhibit the following properties,

$$\mathbb{E}_{\boldsymbol{z}_y \sim \text{vMF}(\boldsymbol{\mu}_y, \kappa_y)} \left[\exp(\boldsymbol{z}^\top \boldsymbol{z}_y / \tau) \right] = \frac{C_q(\kappa_y)}{C_q(\kappa'_y)} \tag{5}$$

where $\kappa'_y = \|\kappa_y \boldsymbol{\mu}_y + \boldsymbol{z}/\tau\|_2$. Substituting Eq. (5) into Eq. (3) yields out,

$$\rho(\boldsymbol{z}, y) = \log \left(C_q(\kappa_y) / C_q(\kappa'_y) \right) \tag{6}$$

where $\boldsymbol{\mu}_y$ and κ_y are assumed to be known, with their estimation detailed in Eq. (8) and Eq. (9). In the following, $\rho(\boldsymbol{z}, y)$ is denoted as the *vMF score*, which reflects the confidence that feature $\boldsymbol{z}$ belongs to class y from the distribution perspective.

Maximum-margin Contrastive Loss. Based on the distribution modeling and estimation above, the contrastive objective in Eq. (2) can be derived,

$$\mathcal{L}_{mmc}(\boldsymbol{z}_i, y_i, \pi) = -\log \frac{\pi(y_i)(C_q(\kappa_{y_i}) / C_q(\kappa'_{y_i}))}{\sum_{c=0}^{K-1} \pi(c)(C_q(\kappa_c) / C_q(\kappa'_c))} \tag{7}$$

Notably, MMCL with the loss $\mathcal{L}_{mmc}$ adaptively aligns features with their class-specific vMF distributions while maximizing the separation from other classes, ultimately achieving the desired feature arrangement with the maximum margin.

Theorem 1. *Given features $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N \in \mathbb{S}^{q-1}$ and class centroids $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K \in \mathbb{S}^{q-1}$ ($2 \leq K \leq q+1$), training with $\mathcal{L}_{mmc}$ ensures that: (1) Each feature converges to its corresponding class centroid. (2) Class centroids form a maximally equiangular configuration, i.e., $\forall i \neq j$, $\boldsymbol{\mu}_i^\top \boldsymbol{\mu}_j = -\frac{1}{K-1}$. Thus, the margin γ_{all} achieves its maximum value of $\frac{K}{K-1}$. Please see Appendix A.3 for more details.*

Distribution Parameter Estimation. After deriving $\mathcal{L}_{mmc}$ based on the vMF mixture modeling, some distribution parameters are still unknown and require estimation, including $\boldsymbol{\mu}$, κ, and the prior π in Eq. (7). These parameters stem from the shared feature space of the source and target data, with their expectations computable when labels are fully available. However, the target data are unlabeled with unknown label distribution. To address this, we maintain pseudo-labels for them, which are carefully processed as explained later in Section 3.2.

Assuming a soft pseudo-label $\hat{\boldsymbol{p}}_i \in [0, 1]^K$ is given for each target span $\boldsymbol{s}_i \in \mathcal{T}_t$, the class prior π for the target and source data can be estimated as $\pi_t(c) = \sum_i \frac{\hat{\boldsymbol{p}}_i^c}{|\mathcal{T}_t|}$ and $\pi_s(c) = \sum_i \frac{\mathbb{I}(y_i = c)}{|\mathcal{T}_s|}$, respectively. Here, $\mathbb{I}(\cdot)$ is the indicator function. Notably, the source and target data share the same vMF distribution with corresponding $\boldsymbol{\mu}$ and κ, but their π values are separately maintained for calculating $\mathcal{L}_{mmc}$ in Eq. (7). For the class centroid $\boldsymbol{\mu}$, each class has $\boldsymbol{\mu}_c = \frac{\bar{\boldsymbol{z}}_c}{\|\bar{\boldsymbol{z}}_c\|_2}$, where $\bar{\boldsymbol{z}}_c$ is the mean vector updated in a moving average mechanism,

$$\overline{\boldsymbol{z}}_c = \beta \overline{\boldsymbol{z}}_c + (1 - \beta) \overline{\boldsymbol{z}}'_c \tag{8}$$

where β is the momentum and $\overline{\boldsymbol{z}}'_c$ is the sample mean of class c in the current batch. Regarding κ, we follow (Sra, 2012) for a simple estimation,

$$\kappa_c = \frac{\|\bar{\boldsymbol{z}}_c\|_2 (q - \|\bar{\boldsymbol{z}}_c\|_2^2)}{(1 - \|\bar{\boldsymbol{z}}_c\|_2^2)} \tag{9}$$

where q is the feature dimension. Once parameters π, $\boldsymbol{\mu}$ and κ are estimated, $\mathcal{L}_{mmc}$ can be computed to foster maximum-margin contrastive alignment.

vMF Score-based Inference. It is worth noting that the estimated vMF score inherently serves as a category indicator. Hence, we no longer introduce

traditional classifiers and directly adopt predictions based on the estimated distribution during inference, i.e., $\hat{y}_{pred} = \arg\max_j(\rho(\boldsymbol{z}, j))$.

3.2 Progressive Cross-lingual Adaptation

In previous steps, we introduced MMCL to produce well-separated features with maximum margins. Nevertheless, such direct separation still poses unexpected risks. Lacking labels in the target language, the inter-class separation and distribution estimation largely depend on the quality of pseudo-labels. If they are incorrect, the contrastive alignment may become unreliable. To overcome this, we design a progressive cross-lingual adaptation (PCLA) mechanism, which first generates initial pseudo-labels through self-training and filters the reliable ones based on density. The remaining samples are then refined via a dual-alignment strategy.

Pseudo-label Initialization. To provide initial pseudo-labels for target data, we adopt self-training paradigm. First, a teacher model M_{src} is trained on the labeled source data $\mathcal{T}_s$ with cross-entropy loss. After training, M_{src} is employed to generate soft pseudo-labels $\hat{\boldsymbol{p}}_i$ for target data $\mathcal{T}_t$, along with hard pseudo-labels $\hat{y}_i = \arg\max_j(\hat{\boldsymbol{p}}_i^j)$. By subsequent filtering and refinery, reliable pseudo-labeled target data and labeled source data are combined to train a final student model M_{final}.

Density-based Label Filtering. Due to the semantic shift between source and target data, the initial pseudo-labels $\hat{y}_i$ inevitably contain some incorrect ones. To alleviate their impact, we propose a density-based filtering strategy to automatically distinguish reliable and unreliable parts. Specifically, we adopt the vMF score $\rho(\boldsymbol{z}_i, j)$ as the filtering criterion, which reflects the probability that sample $\boldsymbol{s}_i$ belongs to class j. Higher vMF scores indicate greater reliability of the pseudo-labels. Notably, filtering is performed class by class due to differing compactness. For each class j, following classical noisy label learning methods (Li et al., 2020b), a two-component Gaussian mixture model is fitted to vMF scores. Let $w_i = p(g|\rho(\boldsymbol{z}_i, \hat{y}_i))$ denote the probability that $\boldsymbol{s}_i$ belongs to the Gaussian component with larger mean g, which also reflects the likelihood that the pseudo-label $\hat{y}_i$ is correct. Then, we can identify reliable pseudo-labeled samples as $\mathcal{T}_r^j = \{(\boldsymbol{s}_i, \hat{y}_i) \mid w_i > 0.5, \hat{y}_i = j\}$. Finally, the reliable samples are merged $\mathcal{T}_r = \bigcup_{j=0}^{K-1} \mathcal{T}_r^j$, with the remaining forming the unreliable subset $\mathcal{T}_u$. $\mathcal{L}_{mmc}$ in Eq. (7) is then merely applied to $\mathcal{T}_r$.

Dual-alignment Label Refinery. After filtering, the unreliable samples are still underutilized. To boost their potential utility in later iterations, we present a refinery strategy incorporating both global and local information. Specifically, the key is to identify the refinery direction $\boldsymbol{d}(\boldsymbol{s}_i)$ for each sample $\boldsymbol{s}_i$, which is achieved by combining global and local guidance, following (Ding et al., 2024). The global correction direction $\boldsymbol{d}_c(\boldsymbol{s}_i)$ is based on the vMF score, and the local correction direction $\boldsymbol{d}_n(\boldsymbol{s}_i)$ is shaped by the neighborhood distribution. Both $\boldsymbol{d}_c(\boldsymbol{s}_i)$ and $\boldsymbol{d}_n(\boldsymbol{s}_i)$ are binary K-dimensional vectors, with their j-th entry as follows:

$$\begin{aligned} \boldsymbol{d}_c(\boldsymbol{s}_i, j) &= \mathbb{I}(\rho(\boldsymbol{z}_i, j) > \text{Avg}_{\hat{y}=j}\rho(\boldsymbol{z}, j)), \\ \boldsymbol{d}_n(\boldsymbol{s}_i, j) &= \mathbb{I}(N_k(\boldsymbol{z}_i, j) > \text{Avg}_{\hat{y}=j}N_k(\boldsymbol{z}, j)) \end{aligned} \tag{10}$$

where $N_k(\boldsymbol{z}_i, j)$ is the number of class j samples among k-nearest neighbors of $\boldsymbol{z}_i$. In other words, if a sample exhibits high similarity to the j-th class centroid or has sufficient neighbors from class j, then its refinery will move toward class j. Based on this, $\boldsymbol{d}_c$ and $\boldsymbol{d}_n$ are integrated to yield the final direction $\boldsymbol{d} = \text{Norm}(\boldsymbol{d}_c + \boldsymbol{d}_n)$. Then $\hat{\boldsymbol{p}}_i$ are refined,

$$\hat{\boldsymbol{p}}_i = \text{Norm}(\alpha\hat{\boldsymbol{p}}_i + (1-\alpha)\boldsymbol{d}(\boldsymbol{s}_i)) \tag{11}$$

where α is the correction strength parameter.

Overall Objective. Finally, the target data with high-quality pseudo-labels are combined with the source data to train a final model M_{final} with $\mathcal{L}_{all}$,

$$\mathcal{L}_{all} = \frac{1}{|\mathcal{T}_s|}\sum_{\boldsymbol{s}_i \in \mathcal{T}_s} \mathcal{L}_{mmc}(\boldsymbol{z}_i, y_i, \pi_s) + \frac{1}{|\mathcal{T}_r|}\sum_{\boldsymbol{s}_i \in \mathcal{T}_r} \mathcal{L}_{mmc}(\boldsymbol{z}_i, \hat{y}_i, \pi_t) \tag{12}$$

Overall, such PCLA carefully and progressively integrates reliable pseudo-labels, which facilitates the contrastive alignment of MMCL to achieve desired features with maximum margins.

4 Experiments

In this section, we present the main results and a detailed analysis to show the effectiveness of our method. More empirical setups and results are provided in Appendix B and Appendix C, respectively.

4.1 Setup

Datasets. We evaluate MARAL on three cross-lingual NER benchmarks: (1) **CoNLL** (Tjong Kim Sang, 2002; Sang and De Meulder, 2003) which includes English (en), German (de), Spanish (es) and Dutch (nl), with 4 entity types (ORG,

Methods \ Datasets	CoNLL				WikiAnn			
	de	es	nl	Avg	ar	hi	zh	Avg
ConNER (Zhou et al., 2022a)	77.14	80.50	83.23	80.29	59.62	74.49	39.17	57.76
MSD (Ma et al., 2022)	77.56	81.92	85.11	81.53	62.88	73.43	57.06	64.46
PRAM (Huang et al., 2023)	77.64	82.06	83.15	80.95	57.44	74.67	55.46	62.52
ProKD (Ge et al., 2023)	78.90	82.62	79.53	80.35	50.91	70.72	51.80	57.81
CoLaDa (Ma et al., 2023b)	**81.12**	82.70	85.15	82.99	66.94	76.69	60.08	67.90
ContProto (Zhou et al., 2023)	76.41	85.02	83.69	81.71	72.20	83.45	61.47	72.37
GLoDe (Ding et al., 2024)	79.15	85.29	85.76	83.40	74.35	83.61	64.81	74.26
MARAL (Ours)	80.28	**86.12**	**85.92**	**84.11**	**78.11**	**84.46**	**66.26**	**76.28**

Table 1: Comparisons of F1 score (%) on CoNLL and WikiAnn. "Avg" denotes the average F1 scores across different languages within each dataset. Superior results are highlighted in **bold**.

PER, LOC, MISC); (2) **WikiAnn** (Pan et al., 2017) which includes English (en), Arabic (ar), Hindi (hi) and Chinese (zh), with 3 entity types (ORG, PER, LOC); (3) **MultiCoNER** (Fetahu et al., 2023) which is more challenging and includes English (en), Bengali (bn), Farsi (fa), French (fr), Italian (it), Portuguese (pt), Swedish (sv) and Ukrainian (uk), with 6 entity types (PER, LOC, GRP, PROD, CW, MED). We take English as the source language and other languages as target languages. The training data in target language is treated as unlabeled.

Baselines. We choose seven SOTA cross-lingual NER baselines: (1) **ConNER** (Zhou et al., 2022a); (2) **MSD** (Ma et al., 2022); (3) **PRAM** (Huang et al., 2023); (4) **ProKD** (Ge et al., 2023); (5) **CoLaDa** (Ma et al., 2023b); (6) **ContProto** (Zhou et al., 2023); (7) **GLoDe** (Ding et al., 2024). Refer to Appendix B.1 for a more detailed description.

Implementation Details. Following previous works (Zhou et al., 2022a; Ding et al., 2024), we use the pre-trained XLM-R (Conneau et al., 2020) as the feature backbone. The model is trained for 20 epochs using AdamW with a learning rate of $2e^{-5}$, a batch size of 8, and a dropout rate of 0.2. The feature dimension q is set to 64. Label filtering and label refinery start following a warm-up period of 2 epochs. We set k in Eq. (10) to 200. The settings for update rates α and β are detailed in Appendix B.2. Performance is measured by the entity-level F1 score on the target-language test set and the average value over 3 runs is reported.

4.2 Main Results

MARAL achieves SOTA results. As shown in Table 1, MARAL significantly surpasses all the rivals on CoNLL and WikiAnn, with improvements of $+0.71\%$ and $+2.02\%$ in average F1 scores compared to the best baseline. Notably, while previous baselines yield moderate results for target languages with greater linguistic proximity to English, such as Spanish (es) and Dutch (nl), they fall short with distant languages such as Arabic (ar) and Chinese (zh). In contrast, MARAL shows significant gains for these challenging distant languages.

Furthermore, it is worth noting that previous works are typically evaluated on datasets with limited entity types and relatively balanced distributions. We challenge this by introducing the more demanding MultiCoNER. As shown in Table 2, MARAL achieves a notable $+2.04\%$ improvement in average F1 scores over the strong GLoDe baseline, which further confirms its superiority.

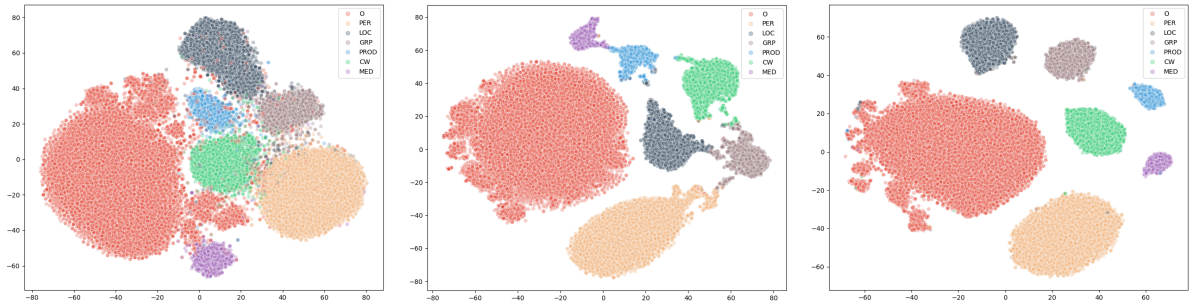
MARAL learns well-separated clusters. We further visualize the span representations of MARAL and two strong baselines for the Italian (it) language in MultiCoNER using t-SNE (Van der Maaten and Hinton, 2008). As presented in Figure 3, without explicit contrastive alignment, the clusters of GLoDe are intertwined and indistinguishable. ContProto improves its features through contrastive learning, but still shows overlap among the minority classes. In contrast, MARAL produces compact and well-separated clusters for all classes, indicating its effectiveness in learning high-quality representations for cross-lingual generalization.

4.3 Analysis

Efficacy of MMCL. To explore the effectiveness of maximum-margin contrastive learning, we first compare with the *MARAL w/o MMC* variant, which removes the MMC loss and trains the model with

Methods \ Datasets	MultiCoNER							
	bn	**fa**	**fr**	**it**	**pt**	**sv**	**uk**	**Avg**
ConNER (Zhou et al., 2022a)	61.17	65.21	72.31	77.80	77.28	77.55	70.66	71.71
MSD (Ma et al., 2022)	60.54	61.34	70.05	77.44	72.81	73.43	65.10	68.67
PRAM† (Huang et al., 2023)	57.93	61.44	70.71	77.16	73.54	73.47	66.63	68.70
ProKD† (Ge et al., 2023)	60.42	64.67	70.60	77.64	72.56	72.77	66.92	69.37
CoLaDa (Ma et al., 2023b)	73.99	64.56	71.59	80.31	76.89	78.65	68.45	73.49
ContProto (Zhou et al., 2023)	75.09	68.74	75.16	83.01	81.34	79.73	75.57	76.95
GLoDe (Ding et al., 2024)	76.03	69.55	76.27	82.31	80.83	79.14	74.83	76.99
MARAL (Ours)	**78.21**	**71.28**	**78.10**	**84.49**	**83.02**	**81.23**	**76.85**	**79.03**

Table 2: Comparisons of F1 score (%) on MultiCoNER. "Avg" denotes average F1 scores. † denotes results reproduced as described in original papers. Other results are obtained by running their publicly available code. Superior results are in **bold**.



(a) GLoDe (b) ContProto (c) MARAL (ours)

Figure 3: t-SNE visualizations of representations on Italian (it) in MultiCoNER. Different colors denote different classes.

vanilla cross-entropy loss. As shown in Table 3, this variant leads to the substantial drop of -1.69% and -2.16% in average F1 scores. To further illustrate the benefits of MMCL in a more intuitive and fine-grained manner, we introduce the *MARAL w/ scl* variant, which opts for the supervised contrastive loss instead of MMC loss. As displayed in Figure 4(a), MARAL consistently shows superior performance across all classes, especially in minority classes (e.g., $+7.24\%$ in the MED entity).

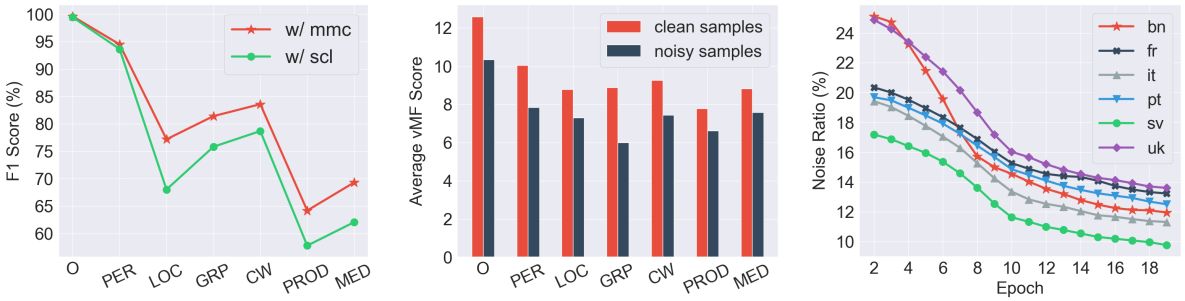
Moreover, we supply some quantitative analysis to verify that MARAL achieves the maximum value of the margin γ_{all}. The two properties of largest margins in Theorem 1 can be summarized as intra-class compactness and inter-class separability. Intra-class compactness can be quantified by the average cosine similarity between samples and their corresponding class centroids. As shown in Figure 5(a), the intra-class compactness approaches one, i.e., $\boldsymbol{z}_i$ converges to $\boldsymbol{\mu}_{y_i}$. Inter-class separability can be measured by the pair-wise cosine similarity between centroids of different classes. As shown in Figure 5(b), the class centroids learned by MARAL approach the optimal configuration in Figure 5(c), where the pair-wise cosine similarity is uniform and achieves the maximum value of -0.17, i.e., $\forall i \neq j, \boldsymbol{\mu}_i^\top \boldsymbol{\mu}_j = -\frac{1}{K-1}$. These results indicate that MARAL establishes the ideal feature arrangement where the margin γ_{all} is maximized.

Efficacy of PCLA. Next, we explore the effect of the progressive cross-lingual adaptation mechanism. As shown in Table 3, removing either label filtering or label refinery will lead to notable performance degradation, indicating the importance of gradually injecting high-quality pseudo-labels.

For label filtering, to verify the rationale behind using the vMF score as the criterion, we first display the average vMF scores for each class in Figure 4(b), where clean samples consistently exhibit higher average vMF scores compared to noisy samples. In addition, the notable variation of vMF scores across different classes highlights the importance of adopting a class-wise filtering strategy. Furthermore, we analyze the training dynamics with and without label filtering in Figure 5(d). MARAL without filtering shows a growing perfor-

Methods \ Datasets	CoNLL				WikiAnn			
	de	es	nl	Avg	ar	hi	zh	Avg
MARAL	**80.28**	**86.12**	**85.92**	**84.11**	**78.11**	**84.46**	**66.26**	**76.28**
- *w/o MMC loss* $\mathcal{L}_{mmc}$	78.26	84.28	84.71	82.42	74.81	82.80	64.74	74.12
- *w/o label filtering*	77.87	85.68	85.12	82.89	76.66	84.11	64.25	75.01
- *w/o label refinery*	79.30	85.76	85.62	83.56	77.62	83.65	64.97	75.41

Table 3: Ablations for maximum-margin contrastive learning and progressive cross-lingual adaptation on CoNLL and WikiAnn.



(a) Class-wise accuracy (b) Class-wise vMF score (c) Effect of label refinery

Figure 4: (a) Comparison of class-wise F1 scores on Italian (it) for different contrastive losses. (b) Average vMF scores of clean and noisy samples for each class on Italian (it). (c) Noise ratio of pseudo-labels across six languages in MultiCoNER.

mance gap over time, which can be attributed to the error accumulation caused by false alignments during the cross-lingual adaptation.

For label refinery, Figure 4(c) presents the noise ratio of pseudo-labels, with a marked decrease as training progresses, for instance, from 30.12% to 11.96% for Bengali (bn) in MultiCoNER. This demonstrates the effectiveness of our refinery strategy in improving the quality of pseudo-labels.

5 Related Work

Cross-lingual NER methods fall into three categories: feature-based, translation-based and self-training-based. Feature-based methods (Zirikly and Hagiwara, 2015; Tsai et al., 2016; Wu and Dredze, 2019; Keung et al., 2019; Wu et al., 2020c) learn language-agnostic features, allowing the model to directly adapt to the target language. Translation-based methods (Mayhew et al., 2017; Xie et al., 2018; Jain et al., 2019; Liu et al., 2021; Yang et al., 2022) yield labeled target-language data via translation and label projection. Self-training-based methods (Wu et al., 2020a,b; Zhao et al., 2022; Li et al., 2022; Zeng et al., 2022) employ a teacher model trained on the source language to generate pseudo-labels for target-language data, which are then used to train a student model. Due to the rich information in pseudo-labels, self-training-based methods show superior performance and gain prominence. Building on this, recent works (Huang et al., 2023; Ge et al., 2023; Zhou et al., 2023; Mo et al., 2024) enhance self-training via explicit entity alignment. However, our empirical studies reveal that entity clusters produced by these methods are not discriminative enough for robust cross-lingual NER.

Contrastive Learning (CL) (Oord et al., 2018; He et al., 2020) has seen extensive application in representation learning via instance similarity and dissimilarity (Wang and Isola, 2020; Chen et al., 2020; Tan et al., 2022; Das et al., 2022). In recent years, CL has also been applied to cross-lingual NER to align entity representations (Ge et al., 2023; Zhou et al., 2023; Mo et al., 2024). The most related work to ours is ContProto (Zhou et al., 2023) which employs supervised contrastive learning (SCL) (Khosla et al., 2020) for cross-lingual alignment. However, SCL fails to account for the distribution skewness and pseudo-label bias in cross-lingual NER, which limits its performance.

6 Conclusion

In this work, we propose a novel framework MARAL for cross-lingual NER. We first identify that existing methods struggle to achieve the optimal features with maximum margins due to two key issues, i.e., the inherent distribution skewness

and pseudo-label bias. To address this, we derive an adaptively reweighted contrastive objective to maximize margins by explicit distribution modeling and performing adaptive inter-class separation. To further address unreliable pseudo-labels, we incorporate a progressive adaptation strategy, which first selects reliable samples as anchors and refines the remaining unreliable ones. Comprehensive experiments show that MARAL significantly outperforms all other baselines on different benchmarks. We hope our work will draw more attention from the community toward maximum-margin theory for generalizable NLP models.

7 Limitations

Despite the remarkable performance gains, our MARAL still has some limitations. First, our framework relies on the inherent semantic alignment capability of multilingual pre-trained language models, and the performance can be further enhanced when combined with more advanced backbones. Second, the strong generalization ability of our method stems from explicitly aligning the same entities in the source and target languages while maximizing the separation between different entities. However, it may face challenges when only limited target data is available. Third, although we incorporate label refinery on top of label filtering to calibrate incorrect pseudo-labels and progressively integrate them into training, some samples inevitably remain underutilized. Fully leveraging these samples merits future exploration.

8 Ethics Statement

Given that NER is a foundational task in the field of natural language processing with the potential to significantly advance a range of downstream tasks, we have carefully reviewed the ethical implications of our work based on the ACL Code of Ethics to ensure that it does not raise any ethical concerns.

9 Acknowledgments

This paper is mainly supported by the NSFC Grants (No. 62206247). Haobo Wang is supported by the NSFC under Grants (No. 62402424) and (No. U24A201401). Gengyu Lyu is supported by the NSFC (No. 62306020).

References

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

Sergiy V Borodachov, Douglas P Hardin, and Edward B Saff. 2019. *Discrete energy on rectifiable sets*, volume 4. Springer.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32.

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

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

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, Dublin, Ireland. Association for Computational Linguistics.

Zhuojun Ding, Wei Wei, Xiaoye Qu, and Dangyang Chen. 2024. Improving pseudo labels with global-local denoising framework for cross-lingual named entity recognition. *arXiv preprint arXiv:2406.01213*.

Chaoqun Du, Yulin Wang, Shiji Song, and Gao Huang. 2024. Probabilistic contrastive learning for long-tailed visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Besnik Fetahu, Zhiyu Chen, Sudipta Kar, Oleg Rokhlenko, and Shervin Malmasi. 2023. MultiCoNER v2: a large multilingual dataset for fine-grained and noisy named entity recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2027–2051, Singapore. Association for Computational Linguistics.

Jinlan Fu, Xuanjing Huang, and Pengfei Liu. 2021. Spanner: Named entity re-/recognition as span prediction. *arXiv preprint arXiv:2106.00641*.

Ling Ge, Chunming Hu, Guanghui Ma, Jihong Liu, and Hong Zhang. 2024. Discrepancy and uncertainty aware denoising knowledge distillation for zero-shot

cross-lingual named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18056–18064.

Ling Ge, Chunming Hu, Guanghui Ma, Hong Zhang, and Jihong Liu. 2023. 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.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Xiaolei Huang, Jonathan May, and Nanyun Peng. 2019. What matters for neural cross-lingual named entity recognition: An empirical analysis. 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 6395–6401, Hong Kong, China. Association for Computational Linguistics.

Yucheng Huang, Wenqiang Liu, Xianli Zhang, Jun Lang, Tieliang Gong, and Chen Li. 2023. PRAM: An end-to-end prototype-based representation alignment model for zero-resource cross-lingual named entity recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3220–3233, Toronto, Canada. Association for Computational Linguistics.

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

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

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

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Vladimir Koltchinskii and Dmitry Panchenko. 2002. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1–50.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020a. A survey on deep learning for named entity recognition. *IEEE transactions on knowledge and data engineering*, 34(1):50–70.

Junnan Li, Richard Socher, and Steven CH Hoi. 2020b. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*.

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

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

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, Online. Association for Computational Linguistics.

Xuantong Liu, Jianfeng Zhang, Tianyang Hu, He Cao, Yuan Yao, and Lujia Pan. 2023. Inducing neural collapse in deep long-tailed learning. In *International Conference on Artificial Intelligence and Statistics*, pages 11534–11544. PMLR.

Jun-Yu Ma, Beiduo Chen, Jia-Chen Gu, Zhenhua Ling, Wu Guo, Quan Liu, Zhigang Chen, and Cong Liu. 2022. 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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ruotian Ma, Xuanting Chen, Zhang Lin, Xin Zhou, Junzhe Wang, Tao Gui, Qi Zhang, Xiang Gao, and Yun Wen Chen. 2023a. Learning "O" helps for learning more: Handling the unlabeled entity problem for class-incremental NER. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5959–5979, Toronto, Canada. Association for Computational Linguistics.

Tingting Ma, Qianhui Wu, Huiqiang Jiang, Börje Karlsson, Tiejun Zhao, and Chin-Yew Lin. 2023b. 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. Association for Computational Linguistics.

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

Ying Mo, Jian Yang, Jiahao Liu, Qifan Wang, Ruoyu Chen, Jingang Wang, and Zhoujun Li. 2024. Mclner: Cross-lingual named entity recognition via multi-view contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18789–18797.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. 2019. On variational bounds of mutual information. In *International Conference on Machine Learning*, pages 5171–5180. PMLR.

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

Suvrit Sra. 2012. A short note on parameter approximation for von mises-fisher distributions: and a fast implementation of i s (x). *Computational Statistics*, 27:177–190.

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, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Kai Tang, Junbo Zhao, Xiao Ding, Runze Wu, Lei Feng, Gang Chen, and Haobo Wang. 2024. Learning geometry-aware representations for new intent discovery. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5641–5654.

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)*.

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, Berlin, Germany. Association for Computational Linguistics.

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. 2022. Pico: Contrastive label disambiguation for partial label learning. In *International conference on learning representations*.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, et al. 2011. Ontonotes release 4.0. *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium*, 17.

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

Qianhui Wu, Zijia Lin, Börje F Karlsson, Biqing Huang, and Jian-Guang Lou. 2020b. Unitrans: Unifying model transfer and data transfer for cross-lingual named entity recognition with unlabeled data. *arXiv preprint arXiv:2007.07683*.

Qianhui Wu, Zijia Lin, Guoxin Wang, Hui Chen, Börje F Karlsson, Biqing Huang, and Chin-Yew Lin. 2020c. Enhanced meta-learning for cross-lingual named entity recognition with minimal resources. In

Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 9274–9281.

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

Ruixuan Xiao, Lei Feng, Kai Tang, Junbo Zhao, Yixuan Li, Gang Chen, and Haobo Wang. 2024. Targeted representation alignment for open-world semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23072–23082.

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, Brussels, Belgium. Association for Computational Linguistics.

Jian Yang, Shaohan Huang, Shuming Ma, Yuwei Yin, Li Dong, Dongdong Zhang, Hongcheng Guo, Zhoujun Li, and Furu Wei. 2022. CROP: Zero-shot cross-lingual named entity recognition with multilingual labeled sequence translation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 486–496, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiali Zeng, Yufan Jiang, Yongjing Yin, Xu Wang, Binghuai Lin, and Yunbo Cao. 2022. 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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yichun Zhao, Jintao Du, Gongshen Liu, and Huijia Zhu. 2022. 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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ran Zhou, Xin Li, Lidong Bing, Erik Cambria, and Chunyan Miao. 2023. 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 (Volume 1: Long Papers)*, pages 4018–4031, Toronto, Canada. Association for Computational Linguistics.

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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xiong Zhou, Xianming Liu, Deming Zhai, Junjun Jiang, Xin Gao, and Xiangyang Ji. 2022b. Learning towards the largest margins. *arXiv preprint arXiv:2206.11589*.

Ayah Zirikly and Masato Hagiwara. 2015. 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 (Volume 2: Short Papers)*, pages 390–396, Beijing, China. Association for Computational Linguistics.

A Theoretical Analysis

A.1 Margin and Generalization Error Bound

We have given the definition of margin in Section 2. Let $L_{\gamma,j}[h] = \mathbb{P}_{\boldsymbol{s}\sim\mathcal{P}_j}[\max_{j'\neq j} h(\boldsymbol{s})_{j'} > h(\boldsymbol{s})_j - \gamma]$ denote the hard margin loss on samples from class j, and let $\hat{L}_{\gamma,j}$ denote its empirical variant. When the training set is separable (i.e., there exists h such that $\gamma_{all} > 0$ for all training samples), Cao et al. (2019) provided a class-balanced generalization error bound: for $\gamma_j > 0$ and all $h \in \mathcal{F}$, with a high probability, we have

$$\mathbb{P}_{(\boldsymbol{s},y)}[h(\boldsymbol{s})_y < \max_{l\neq y} h(\boldsymbol{s})_l] \leq \frac{1}{k}\sum_{j=1}^{k}\left(\hat{L}_{\gamma_j,j}[h] + \frac{4}{\gamma_j}\hat{\mathfrak{R}}_j(\mathcal{F}) + \varepsilon_j(\gamma_j)\right) \tag{13}$$

where $\frac{4}{\gamma_j}\hat{\mathfrak{R}}_j(\mathcal{F})$ denotes the empirical Rademacher complexity and has a significant impact on the value of the generalization bound. Zhou et al. (2022b) suggests that large margins $\{\gamma_j\}_{j=0}^{K-1}$ for all classes should be encouraged to tighten the generalization bound. Motivated by this, we aim at maximizing the overall margin γ_{all} through contrastive learning, which further results in a larger margin γ_j for each class j.

A.2 Proof of Proposition 1

Inspired by recent studies on mutual information theory (Poole et al., 2019), we interpret the contrastive learning process from the perspective of mutual information maximization.

Mutual Information Maximization. As demonstrated in Poole et al. (2019), the lower bound on mutual information (MI) between X and Y can be formulated as:

$$I(X;Y) \geq \mathbb{E}_{X,Y}\left[\frac{1}{N}\sum_{i=0}^{N-1}\log\frac{\exp(\rho(x_i,y_i))}{\frac{1}{N}\sum_{j=0}^{N-1}\exp(\rho(x_i,y_j))}\right] \tag{14}$$

where the expectation $\mathbb{E}_{X,Y}$ is over N independent samples from the joint distribution: $\prod_j p(x_j,y_j)$. Equality holds when $N\to\infty$ and $\rho(x,y) = \log p(x|y) + c(x)$ where $c(x)$ is any function that only depends on x. Given a set of latent features $\boldsymbol{z}_i$ with their labels y_i, the MI maximization between them involves maximizing the lower bound:

$$\mathbb{E}_i \log\frac{\exp(\rho(\boldsymbol{z}_i,y_i))}{\mathbb{E}_j\exp(\rho(\boldsymbol{z}_i,y_j))} = \mathbb{E}_i\log\frac{\exp(\rho(\boldsymbol{z}_i,y_i))}{\sum_{c=0}^{K-1}\pi(c)\exp(\rho(\boldsymbol{z}_i,c))} = \mathbb{E}_i\left[\log\frac{\pi(y_i)\exp(\rho(\boldsymbol{z}_i,y_i))}{\sum_{c=0}^{K-1}\pi(c)\exp(\rho(\boldsymbol{z}_i,c))} - \log(\pi(y_i))\right] \tag{15}$$

where K is the number of different classes and $\pi(c)$ is the prior probability of class c. Since the bias term $-\log(\pi(y_i))$ does not contribute to the gradient optimization (i.e., it has no effect on the gradient with respect to $\boldsymbol{z}_i$), we omit this term and obtain the resulting MI maximization objective:

$$\mathcal{L} = -\mathbb{E}_i\left[\log\frac{\pi(y_i)\exp(\rho(\boldsymbol{z}_i,y_i))}{\sum_{c=0}^{K-1}\pi(c)\exp(\rho(\boldsymbol{z}_i,c))}\right] \tag{16}$$

MI Maximization in Contrastive Space. In contrastive learning, we typically learn representations $\boldsymbol{z}\in\mathbb{R}^q$ with a unit ℓ_2 norm constraint, i.e., restricting the output space to the unit hypersphere $\mathbb{S}^{q-1}$. Wang et al. (2022) has shown that data from the same class in the contrastive output space implicitly follow a q-variate von Mises-Fisher (vMF) distribution whose probabilistic density is given by $f(\boldsymbol{z};\boldsymbol{\mu},\kappa) = C_q(\kappa)\exp(\kappa\boldsymbol{\mu}^\top\boldsymbol{z})$, where $\boldsymbol{\mu}$ is the mean direction with $\|\boldsymbol{\mu}\|_2 = 1$, κ is the concentration parameter and $C_q(\kappa)$ is the normalization factor which is defined as:

$$C_q(\kappa) = \frac{\kappa^{q/2-1}}{(2\pi)^{q/2}I_{q/2-1}(\kappa)} \tag{17}$$

where $I_{q/2-1}(\kappa)$ denotes the modified Bessel function of the first kind at order $q/2-1$ defined as,

$$I_{(q/2-1)}(\boldsymbol{z}) = \sum_{k=0}^{\infty}\frac{1}{k!\Gamma(q/2-1+k+1)}\left(\frac{\boldsymbol{z}}{2}\right)^{2k+q/2-1} \tag{18}$$

Essentially, the vMF distribution can be regarded as the spherical counterpart of the Gaussian distribution for features with unit norms.

Based on the variational lower bound theory, the inequality in Eq. (14) tightens as ρ converges to $\log p(\boldsymbol{z}|y)$. Hence, by considering the feature distribution in the contrastive space, we employ a Gaussian-like kernel $K(\boldsymbol{z},\boldsymbol{z}_y) = \exp(\boldsymbol{z}^\top\boldsymbol{z}_y/\tau)$ and estimate $\rho(\boldsymbol{z},y)$ through sampling-based kernel density estimation:

$$\rho(\boldsymbol{z},y) = \log\left(\frac{1}{|\mathcal{Z}_y|}\sum_{\boldsymbol{z}_y\in\mathcal{Z}_y}\exp(\boldsymbol{z}^\top\boldsymbol{z}_y/\tau)\right) \tag{19}$$

where $\mathcal{Z}_y$ represents the set of ℓ_2-normalized contrastive features for class y and τ denotes a temperature parameter. In the ideal case, where the contrastive samples are infinite, $\rho(\boldsymbol{z},y)$ can be written in the expected form as:

$$\rho(\boldsymbol{z},y) = \log\left(\mathbb{E}_{\boldsymbol{z}_y\sim\text{vMF}(\boldsymbol{\mu}_y,\kappa_y)}\left[\exp(\boldsymbol{z}^\top\boldsymbol{z}_y/\tau)\right]\right) \tag{20}$$

Derivation of Contrastive Loss. Building upon the above estimation, we can first derive the supervised contrastive loss (SCL) by assuming a balanced training dataset and approximating $\rho(\boldsymbol{z}, y)$ using samples within a single mini-batch. Under the class balance setting, all $\pi(c)$ and $|\mathcal{Z}_c|$ are equal. Therefore, the proposed objective in Eq. (19) can be written as follows:

$$\begin{aligned}\mathcal{L}_i &= -\log \frac{\sum_{\boldsymbol{z}_y \in \mathcal{Z}_y} \exp\left(\boldsymbol{z}_i^\top \boldsymbol{z}_y/\tau\right)}{\sum_{c=0}^{K-1} \sum_{\boldsymbol{z}_c \in \mathcal{Z}_c} \exp\left(\boldsymbol{z}_i^\top \boldsymbol{z}_c/\tau\right)} \\ &\leq -\sum_{\boldsymbol{z}_y \in \mathcal{Z}_y} \log \frac{\exp\left(\boldsymbol{z}_i^\top \boldsymbol{z}_y/\tau\right)}{\sum_{\boldsymbol{z} \in \mathcal{Z}} \exp\left(\boldsymbol{z}_i^\top \boldsymbol{z}/\tau\right)} \\ &\leq -\frac{1}{|\mathcal{Z}_y|} \sum_{\boldsymbol{z}_y \in \mathcal{Z}_y} \log \frac{\exp\left(\boldsymbol{z}_i^\top \boldsymbol{z}_y/\tau\right)}{\sum_{\boldsymbol{z} \in \mathcal{Z}} \exp\left(\boldsymbol{z}_i^\top \boldsymbol{z}/\tau\right)}\end{aligned} \tag{21}$$

where $\mathcal{Z} = \bigcup_c \mathcal{Z}_c$ denotes the set of all contrast features. By applying Jensen's inequality, we derive the supervised contrastive loss which is employed by the state-of-the-art cross-lingual NER method ContProto (Zhou et al., 2023). Building on this, the contrastive loss used in ProKD (Ge et al., 2023), which aligns class prototypes between source and target data, can be derived by simply replacing $\boldsymbol{z}_i$ with the class prototype in the source data, restricting positive samples to the corresponding prototype of class y_i in the target data and negative samples to the prototypes of other classes in both the source and target data,

$$\mathcal{L}_i = -\log \frac{\exp\left(\boldsymbol{z}_i^\top \boldsymbol{z}_{y_i}^t/\tau\right)}{\sum\limits_{c \neq y_i} \exp\left(\boldsymbol{z}_i^\top \boldsymbol{z}_c^s/\tau\right) + \sum\limits_{c \neq y_i} \exp\left(\boldsymbol{z}_i^\top \boldsymbol{z}_c^t/\tau\right)} \tag{22}$$

where $\boldsymbol{z}_c^s$ and $\boldsymbol{z}_c^t$ denote the prototype of class c in the source and target languages, respectively. Especially, the unsupervised contrastive loss used in MCL-NER (Mo et al., 2024) can be also derived by restricting positive samples to the corresponding code-switched counterpart of $\boldsymbol{z}_i$ and negative samples to all the code-switched counterparts,

$$\mathcal{L}_i = -\log \frac{\exp\left(\boldsymbol{z}_i^\top \boldsymbol{z}_i'/\tau\right)}{\sum_{z' \in Z'} \exp\left(\boldsymbol{z}_i^\top \boldsymbol{z}'/\tau\right)} \tag{23}$$

where $\boldsymbol{z}_i'$ denotes the code-switched counterpart of $\boldsymbol{z}_i$ and Z' denotes the set of features corresponding to all code-switched counterparts.

Objective for Robust MI Maximization. From the above derivation, we observe that the traditional contrastive loss is based on two critical assumptions: (1) the class distribution is balanced, and (2) the samples within a batch are sufficiently representative. However, these assumptions often do not hold in practice, which means such contrastive loss cannot guarantee effective MI maximization. Here, we derive the contrastive objective necessary to achieve the robust MI maximization. As previously discussed, features $\boldsymbol{z}_y$ from the same class y in the contrastive space follow a vMF distribution. Therefore, the overall expectation of $\rho(\boldsymbol{z}, y)$ can be computed as follows:

Lemma 1. *Let $\boldsymbol{z}_y$ denote the features that follow a von Mises-Fisher distribution $f(\boldsymbol{z}_y; \boldsymbol{\mu}_y, \kappa_y)$, and let $\boldsymbol{z}$ be a unit vector. Then, the expectation of $\rho(\boldsymbol{z}, y)$ can be expressed as:*

$$\begin{aligned}\mathbb{E}[\rho(\boldsymbol{z}, y)] &= \log\left(\mathbb{E}_{\boldsymbol{z}_y \sim vMF(\boldsymbol{\mu}_y, \kappa_y)} \exp(\boldsymbol{z}^\top \boldsymbol{z}_y/\tau)\right) \\ &= \log\left(C_q(\kappa_y')/C_q(\kappa_y)\right)\end{aligned} \tag{24}$$

where $\kappa_y' = \|\kappa_y \boldsymbol{\mu}_y + \boldsymbol{z}/\tau\|_2$.

Proof.

$$\begin{aligned}\mathbb{E}\left[e^{\boldsymbol{z}^\top \boldsymbol{z}_y/\tau}\right] &= \int_{S^{q-1}} e^{\boldsymbol{z}^\top \boldsymbol{z}_y/\tau} f(\boldsymbol{z}_y; \boldsymbol{\mu}_y, \kappa_y)\, d\boldsymbol{z}_y \\ &= \int_{S^{q-1}} C_q(\kappa_y) e^{\boldsymbol{z}^\top \boldsymbol{z}_y/\tau + \kappa_y \boldsymbol{\mu}_y^\top \boldsymbol{z}_y} d\boldsymbol{z}_y \\ &= \int_{S^{q-1}} C_q(\kappa_y) e^{(\boldsymbol{z}/\tau + \kappa_y \boldsymbol{\mu}_y)\boldsymbol{z}_y} d\boldsymbol{z}_y\end{aligned} \tag{25}$$

Let $\boldsymbol{\mu}_y' = \frac{\kappa_y \boldsymbol{\mu}_y + \boldsymbol{z}/\tau}{\kappa_y'}$, $\kappa_y' = \|\kappa_y \boldsymbol{\mu}_y + \boldsymbol{z}/\tau\|_2$, then we have:

$$\begin{aligned}\mathbb{E}\left[e^{\boldsymbol{z}^\top \boldsymbol{z}_y/\tau}\right] &= C_q(\kappa_y) \int_{S^{q-1}} e^{\kappa_y' \boldsymbol{\mu}_y'^\top \boldsymbol{z}_y}\, d\boldsymbol{z}_y \\ &= C_q(\kappa_y)/C_q(\kappa_y')\end{aligned} \tag{26}$$

□

Based on the derivation above, we can conclude that in order to maximize the mutual information in the contrastive space, the objective we ultimately need to optimize is as follows:

$$\mathcal{L}_i = -\log \frac{\pi(y_i)(C_q(\kappa_{y_i})/C_q(\kappa_{y_i}'))}{\sum_{c=0}^{K-1} \pi(c)(C_q(\kappa_c)/C_q(\kappa_c'))} \tag{27}$$

Learning towards Maximum Margin. Here, we demonstrate that maximizing the mutual information in the contrastive space is equivalent to maximizing the overall margin. Given that it is not possible to obtain an analytic solution of κ_c, we follow the well-known approximation

$\|\boldsymbol{\mu}_c\| \left(q - \|\boldsymbol{\mu}_c\|^2\right)/(1 - \|\boldsymbol{\mu}_c\|^2)$. In most cases, where $\|\boldsymbol{\mu}_c\| \gg 1 - \|\boldsymbol{\mu}_c\|^2$, we have $\kappa \gg q$. Therefore, the modified Bessel function of the first kind, $I_{q/2-1}(\kappa)$, can be effectively approximated as:

$$I_{(q/2-1)}(\kappa) \approx \frac{e^{\kappa}}{\sqrt{2\pi\kappa}}. \tag{28}$$

Then we can approximate $\rho(\boldsymbol{z}, y)$ as:

$$\rho(\boldsymbol{z}, y) = \log \frac{C_q(\kappa'_y)}{C_q(\kappa_y)} \approx \log \frac{e^{\kappa'_y} {\kappa_y}^{(q-1)/2}}{e^{\kappa_y} {\kappa'_y}^{(q-1)/2}}. \tag{29}$$

To expand $\kappa'_y = \|\kappa_y \boldsymbol{\mu}_y + \boldsymbol{z}/\tau\|_2$, we use the norm expansion:

$$\kappa'_y = \sqrt{{\kappa_y}^2 + \frac{2\kappa_y \boldsymbol{\mu}_y^\top \boldsymbol{z}}{\tau} + \frac{1}{\tau^2}}. \tag{30}$$

For sufficiently large κ_y, κ'_y can be approximated using a Taylor expansion:

$$\kappa'_y \approx \kappa_y + \frac{\boldsymbol{\mu}_y^\top \boldsymbol{z}}{\tau} + \frac{1}{2\tau^2\kappa_y}. \tag{31}$$

Moreover, given that $\kappa \gg q$, we have:

$$\frac{{\kappa_y}^{(q-1)/2}}{{\kappa'_y}^{(q-1)/2}} \approx 1. \tag{32}$$

Thus, we can derive the following approximation of $\rho(\boldsymbol{z}, y)$:

$$\rho(\boldsymbol{z}, y) \approx \boldsymbol{\mu}_y^\top z/\tau + 1/(2\tau^2\kappa_y). \tag{33}$$

Finally, the optimization objective in Eq. (27) can be approximated as follows:

$$\mathcal{L}_i \approx -\log \frac{\pi(y_i)\exp(\boldsymbol{\mu}_{y_i}^\top \boldsymbol{z}_i/\tau + 1/(2\tau^2\kappa_{y_i}))}{\sum_{c=0}^{K-1} \pi(c)\exp(\boldsymbol{\mu}_c^\top \boldsymbol{z}_i/\tau + 1/(2\tau^2\kappa_c))} \tag{34}$$

To further simplify, we demonstrate that the term $\pi(c)\exp(1/(2\tau^2\kappa_c))$ is a constant. Assume that there exists a constant C such that:

$$\pi(c)\exp(1/(2\tau^2\kappa_c)) = C \tag{35}$$

which implies:

$$\pi(c) = C\exp(-1/(2\tau^2\kappa_c)) \tag{36}$$

Given that the prior probabilities $p(c)$ must satisfy the normalization condition $\sum_{c=0}^{K-1} \pi(c) = 1$, substituting $\pi(c) = C\exp(-1/(2\tau^2\kappa_c))$ into this condition, we obtain:

$$\sum_{c=0}^{K-1} C\exp(-1/(2\tau^2\kappa_c)) = 1 \tag{37}$$

$$C = \frac{1}{\sum_{c=0}^{K-1}\exp(-1/(2\tau^2\kappa_c))} \tag{38}$$

This result establishes C a constant independent of c. Thus, we can simplify Eq. (34) as:

$$\mathcal{L}_i \approx -\log \frac{\exp(\boldsymbol{\mu}_{y_i}^\top \boldsymbol{z}_i/\tau)}{\sum_{c=0}^{K-1}\exp(\boldsymbol{\mu}_c^\top \boldsymbol{z}_i/\tau)} \tag{39}$$

$$\approx -\log \frac{\exp(\boldsymbol{\mu}_{y_i}^\top \boldsymbol{z}_i/\tau)}{\sum_{c\neq y_i}\exp(\boldsymbol{\mu}_c^\top \boldsymbol{z}_i/\tau)} \tag{40}$$

$$\leq -\tau\log \frac{\exp(\boldsymbol{\mu}_{y_i}^\top \boldsymbol{z}_i/\tau)}{\sum_{c\neq y_i}\exp(\boldsymbol{\mu}_c^\top \boldsymbol{z}_i/\tau)} \tag{41}$$

$$= -\tau\log \sum_{c\neq y_i}\exp((\boldsymbol{\mu}_{y_i} - \boldsymbol{\mu}_c)^\top \boldsymbol{z}_i/\tau) \tag{42}$$

where $0 < \tau < 1$. According to the limiting case of the *log-sum-exp* operation, the overall margin can be written as follows:

$$\gamma_{all} = \lim_{\tau\to 0} \tau\log(\sum_{i=1}^{N}\sum_{c\neq y_i}\exp((\boldsymbol{\mu}_{y_i}^\top \boldsymbol{z}_i - \boldsymbol{\mu}_c^\top \boldsymbol{z}_i)/\tau)) \tag{43}$$

Furthermore, as $\log$ is strictly concave, we can derive the following inequality:

$$\tau\log(\sum_{i=1}^{N}\sum_{c\neq y_i}\exp((\boldsymbol{\mu}_{y_i}^\top \boldsymbol{z}_i - \boldsymbol{\mu}_c^\top \boldsymbol{z}_i)/\tau)$$

$$\leq -\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_i - \tau\log N \tag{44}$$

Maximizing the right-hand side of Eq. (44) results in the condition that $\sum_{c\neq y_i}\exp((\boldsymbol{\mu}_{y_i}^\top \boldsymbol{z}_i - \boldsymbol{\mu}_c^\top \boldsymbol{z}_i)/\tau)$ is a constant. While the equality of Eq. (44) holds if and only if this sum is a constant. Therefore, maximizing the value of γ_{all} is inherently equivalent to minimizing the objective $\mathcal{L}_i$.

A.3 Proof of Theorem 1

According to the definition of γ_{all} in Section 2, we have

$$\begin{aligned}
&\arg\max_{\boldsymbol{\mu}} \max_{\boldsymbol{z}} \gamma_{all} \\
&= \arg\max_{\boldsymbol{\mu}} \max_{\boldsymbol{z}} \min_{i} \boldsymbol{\mu}_{y_i}^\top \boldsymbol{z}_i - \max_{j \neq y_i} \boldsymbol{\mu}_j^\top \boldsymbol{z}_i \\
&= \arg\max_{\boldsymbol{\mu}} \min_{i} \max_{\boldsymbol{z}_i} \boldsymbol{\mu}_{y_i}^\top \boldsymbol{z}_i - \max_{j \neq y_i} \boldsymbol{\mu}_j^\top \boldsymbol{z}_i \\
&= \arg\max_{\boldsymbol{\mu}} \min_{i} \max_{\boldsymbol{z}_i} \boldsymbol{\mu}_{y_i}^\top \boldsymbol{z}_i - \boldsymbol{\mu}_k^\top \boldsymbol{z}_i \\
&= \arg\max_{\boldsymbol{\mu}} \min_{i} \|\boldsymbol{\mu}_{y_i} - \boldsymbol{\mu}_k\|_2,
\end{aligned}$$

here, $k \in \arg\max_{j \neq y_i} \boldsymbol{\mu}_j^\top \boldsymbol{z}_i$ and $\boldsymbol{z}_i = \frac{\boldsymbol{\mu}_{y_i} - \boldsymbol{\mu}_k}{\|\boldsymbol{\mu}_{y_i} - \boldsymbol{\mu}_k\|_2}$. Notice that $\boldsymbol{\mu}_k^\top \boldsymbol{z}_i = -\frac{1}{2}\|\boldsymbol{\mu}_{y_i} - \boldsymbol{\mu}_k\|_2$, which implies $k = \arg\min_{j \neq y_i} \|\boldsymbol{\mu}_{y_i} - \boldsymbol{\mu}_j\|_2$. Hence, we have

$$\begin{aligned}
&\arg\max_{\boldsymbol{\mu}} \max_{\boldsymbol{z}} \gamma_{all} \\
&= \max_{\boldsymbol{\mu}} \min_{i} \min_{k \neq y_i} \|\boldsymbol{\mu}_{y_i} - \boldsymbol{\mu}_k\|_2 \\
&= \arg\max_{\boldsymbol{\mu}} \min_{i \neq j} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2,
\end{aligned}$$

i.e., maximizing γ_{all} yields the solution of the Tammes Problem. According to the proof of Theorem 3.3.1 in Borodachov et al. (2019), we have the above solution satisfies that $\boldsymbol{\mu}_i^\top \boldsymbol{\mu}_j = -\frac{1}{K-1}, \forall i \neq j$. Therefore, we have $\boldsymbol{z}_i = \arg\max_{\boldsymbol{z} \in S^{q-1}} \boldsymbol{\mu}_{y_i}^\top \boldsymbol{z} - \max_{j \neq y} \boldsymbol{\mu}_j^\top \boldsymbol{z} = \boldsymbol{\mu}_{y_i}$, and $\gamma_{all} = -\frac{K}{K-1}$.

Next, we prove that learning with $\mathcal{L}_{mmc}$ in Eq. (7) ensures the two key properties outlined above, which lead to the maximum value of γ_{all}. According to the approximation in (40), we can decompose $\mathcal{L}_{mmc}$ into two separate loss components:

$$\mathcal{L}_{mmc} = \underbrace{-\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\mu}_{y_i}^\top \boldsymbol{z}_i/\tau}_{\mathcal{L}_a} + \underbrace{\frac{1}{N}\sum_{i=1}^{N} \log \sum_{c \neq y_i} \exp(\boldsymbol{\mu}_c^\top \boldsymbol{z}_i/\tau)}_{\mathcal{L}_r} \tag{45}$$

Through the influence of $\mathcal{L}_a$, samples of the same class will align closely with their corresponding class centroids, forming compact clusters. At convergence, intra-class variances will collapse to zero, i.e., $\boldsymbol{z}_i = \boldsymbol{\mu}_{y_i}, \forall i$. Then, based on Jensen's inequality, we can derive the following approximated upper bound for $\mathcal{L}_r$:

$$\begin{aligned}
\mathcal{L}_r &= \frac{1}{N}\sum_{i=1}^{N} \log \sum_{c \neq y_i} \exp(\boldsymbol{\mu}_c^\top \boldsymbol{z}_i/\tau) \\
&\leq \frac{1}{N}\sum_{i=1}^{N}\sum_{c \neq y_i} \boldsymbol{\mu}_c^\top \boldsymbol{z}_i/\tau \approx \frac{1}{K}\sum_{j=0}^{K-1}\sum_{c \neq j} \boldsymbol{\mu}_c^\top \boldsymbol{\mu}_j/\tau
\end{aligned} \tag{46}$$

Under our feature distribution modeling, the class centroids $\boldsymbol{\mu}_c$ are unit vectors, which implies that $\boldsymbol{\mu}_c^\top \boldsymbol{\mu}_j$ represents the cosine similarity between two class centroids. Therefore, optimizing $\mathcal{L}_r$ can be seen as minimizing the maximum pair-wise cosine similarity between all class means, which is equivalent to maximizing the minimum pair-wise angle between them. As demonstrated in Liu et al. (2023), this kind of loss will lead the K class centroids to form a simplex equiangular tight frame (ETF) structure. Formally, the ETF structure (Xiao et al., 2024; Tang et al., 2024) arranges the K class centroids $\mathbf{M} = [\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_{K-1}]$ such that:

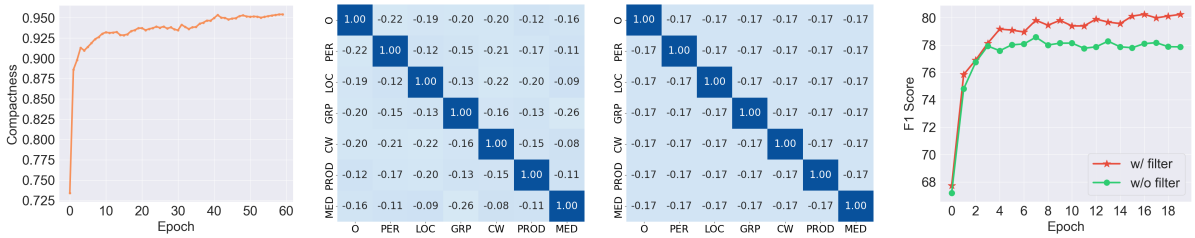
$$\mathbf{M}^\top \mathbf{M} = \frac{K}{K-1}\mathbf{I}_K - \frac{1}{K-1}\mathbf{1}_K \mathbf{1}_K^\top. \tag{47}$$

where $\mathbf{I}_K$ denotes the identity matrix and $\mathbf{1}_K$ denotes an all-ones vector. This arrangement guarantees that all pairs of the K class centroids have the same pair-wise maximal equiangular angle $-\frac{1}{K-1}$, i.e., $\boldsymbol{\mu}_i^\top \boldsymbol{\mu}_j = -\frac{1}{K-1}, \forall i \neq j$. Hence, $\mathcal{L}_a$ maximizes the intra-class compactness, while $\mathcal{L}_r$ maximizes the inter-class separability, collectively ensuring that the overall margin γ_{all} achieves its maximum value of $\frac{K}{K-1}$.

B Additional Experimental Setups

B.1 Baselines

We compare our framework with seven state-of-the-art cross-lingual NER baselines: (1) **ConNER** (Zhou et al., 2022a) combines translation-based and dropout-based consistency training. (2) **MSD** (Ma et al., 2022) leverages multi-channel distillation and parallel domain adaptation to transfer knowledge. (3) **PRAM** (Huang et al., 2023) introduces a prototype-based representation alignment model with attribution-prediction consistency. (4) **ProKD** (Ge et al., 2023) employs the contrastive-based prototype alignment to capture language-independent knowledge. (5) **CoLaDa** (Ma et al., 2023b) proposes collaborative model and instance denoising strategies to refine and reweight pseudo-labels.



(a) Intra-class compactness (b) Inter-class separability (c) Optimal separability (d) Ablation of filtering

Figure 5: (a) Intra-class compactness quantified by the average cosine similarity between samples and their corresponding class centroids on Italian (it). (b) Inter-class separability measured by the pair-wise cosine similarity between class centroids on Italian (it). (c) Optimal inter-class separability where the pair-wise cosine similarity between class centroids is uniformly equal to the minimum value of $-\frac{1}{K-1}$. (b) The F1 scores of MARAL with/without label filtering on German (de).

class	split	en	de	es	nl
O	train	709,350	744,298	993,456	712,366
	dev	180,531	184,079	196,719	133,034
	test	159,980	186,432	193,960	244,147
ORG	train	6,321	2,427	7,390	2,082
	dev	1,341	1,241	1,700	686
	test	1,661	773	1,400	882
PER	train	6,600	2,773	4,321	4,716
	dev	1,842	1,401	1,222	703
	test	1,617	1,195	735	1,098
LOC	train	7,140	4,363	4,914	3,208
	dev	1,837	1,181	985	479
	test	1,668	1,035	1,084	774
MISC	train	3,438	2,288	2,173	3,338
	dev	922	1,010	445	748
	test	702	670	340	1,187

Table 4: Class-wise span statistics across languages in the CoNLL benchmark for Train/Dev/Test splits.

class	split	en	ar	hi	zh
O	train	493,646	374,321	81,652	1,535,558
	dev	247,998	185,897	16,007	782,392
	test	247,346	186,129	16,793	757,655
ORG	train	9,422	7,183	1,823	8,260
	dev	4,677	3,596	369	4,263
	test	4,745	3,629	364	4,115
PER	train	9,164	7,539	2,261	7,987
	dev	4,635	3,815	434	3,916
	test	4,556	3,850	450	3,943
LOC	train	9,345	7,779	2,040	8,784
	dev	4,834	3,856	423	4,314
	test	4,657	3,780	414	4,474

Table 5: Class-wise span statistics across languages in the WikiAnn benchmark for Train/Dev/Test splits.

(6) **ContProto** (Zhou et al., 2023) learns representations with contrastive learning and improves pseudo-label quality with prototype learning. (7) **GLoDe** (Ding et al., 2024) refines pseudo-labels by integrating global and local semantic space information.

B.2 Detailed Settings of Update Rates α and β

During training, the update rate parameters α in Eq. (11) and β in Eq. (8) are dynamically adjusted. First, considering that the clusters have not yet fully formed at the early stage of training, we set a relatively large initial value for α. As training progresses, the global and local guidance provided by the clustering information become increasingly valuable. Thus, we gradually decrease α to allow for greater calibration of the soft pseudo-labels. Specifically, α decreases linearly from 0.95 to 0.80. Moreover, due to the distribution skewness, we calculate β in a class-specific manner. We define β_c as the update rate for the mean vector of class c, which is calculated as $\beta_c = \frac{n_c^{(t)}}{N_c^{(t)}}$, where $n_c^{(t)}$ is the number of class c samples in the current batch, and $N_c^{(t)} = \sum_{i=1}^{t} n_c^{(i)}$ is the cumulative number of class c samples in the previous batches.

B.3 Computation of Normalization Factor

The normalization factor $C_q(\kappa)$ of the vMF distribution ensures that the probability density function integrates to one over the hypersphere and is defined as:

$$C_q(\kappa) = \frac{\kappa^{q/2-1}}{(2\pi)^{q/2} I_{q/2-1}(\kappa)} \tag{48}$$

To compute $C_q(\kappa)$, it is essential to accurately calculate the modified Bessel function $I_{q/2-1}(\kappa)$, as defined in Eq. (18). However, the direct computation of high-order Bessel functions can be computationally expensive, and using forward recurrence can lead to numerical instability, especially when

class	split	en	bn	fa	fr	it	pt	sv	uk
O	train	885,927	449,253	1,042,711	865,308	826,915	910,403	731,424	775,182
	dev	46,770	23,370	57,034	45,754	42,798	46,532	38,617	39,911
	test	13,217,003	882,058	3,324,977	13,074,793	12,330,425	12,669,251	10,245,570	11,209,017
PER	train	9,294	3,778	8,006	9,295	10,387	8,241	7,695	6,441
	dev	481	194	413	483	548	447	445	341
	test	137,681	6,935	115,868	141,401	160,598	120,413	111,157	96,864
LOC	train	4,353	2,457	5,086	4,723	4,446	4,794	7,176	5,458
	dev	197	127	267	242	248	250	370	294
	test	67,901	7,375	70,907	73,373	68,564	70,923	111,879	84,643
GRP	train	4,224	2,227	3,209	3,745	3,416	3,788	3,459	3,204
	dev	218	122	180	195	173	200	194	151
	test	60,026	3,651	38,807	52,987	46,271	48,994	46,929	39,709
CW	train	4,084	1,981	3,661	5,438	5,048	3,839	3,714	2,907
	dev	215	103	184	268	267	206	200	146
	test	62,126	3,640	53,034	84,952	79,873	58,245	54,806	43,291
PROD	train	1,935	1,384	2,049	1,946	1,770	1,927	1,989	2,258
	dev	109	67	107	100	86	101	112	117
	test	27,580	1,493	18,212	28,274	22,887	21,115	22,686	30,071
MED	train	1,559	1,396	1,651	1,230	1,376	1,850	1,381	1,688
	dev	76	63	85	64	76	88	70	86
	test	22,491	1,919	15,287	17,208	19,029	21,062	13,702	20,796

Table 6: Class-wise span statistics across languages in the MultiCoNER benchmark for Train/Dev/Test splits.

κ is not sufficiently large. To address this problem, we follow Du et al. (2024) and employ the Miller recurrence algorithm to compute the following inverse recurrence relation:

$$\hat{I}_{\nu-1}(\kappa) = \frac{2\nu}{\kappa}\hat{I}_{\nu}(\kappa) + \hat{I}_{\nu+1}(\kappa) \tag{49}$$

By initializing the trial values of $\hat{I}_q(\kappa)$ and $\hat{I}_{q+1}(\kappa)$ to 1 and 0, respectively, we can iteratively compute $\hat{I}_{\nu}(\kappa)$ for $\nu = q-1, q-2, \ldots, 0$. Subsequently, we can determine the value of $I_{q/2-1}(\kappa)$ as follows:

$$I_{q/2-1}(\kappa) = \frac{I_0(\kappa)}{\hat{I}_0(\kappa)}\hat{I}_{q/2-1}(\kappa) \tag{50}$$

where $\hat{I}_0(\kappa)$ and $\hat{I}_{q/2-1}(\kappa)$ are the trial values obtained through the Miller recurrence, and $I_0(\kappa)$ can be efficiently computed using PyTorch.

B.4 Integration of MLM Loss

To fully exploit the knowledge in the multilingual pre-trained language model, following previous work (Ding et al., 2024), we also integrate the masked language modeling loss on the target data. Specifically, for each target-language sentence $\boldsymbol{X} = \{x_i\}_{i=1}^{L}$, a masked version $\boldsymbol{X}' = \{x_i'\}_{i=1}^{L}$ is generated, where tokens in $\boldsymbol{X}$ are replaced by a "[MASK]" token with probability r_m. Token representations $\boldsymbol{H}' = \{\boldsymbol{h}_i'\}_{i=1}^{L}$ are then obtained using the PLM and passed through an MLM head to predict the original tokens. Let $\boldsymbol{X}_m'$ denote tokens that are masked, y_i' denote the original token's id and $\boldsymbol{p}_i'$ denote the predicted probability distribution, the loss function is defined as:

$$\mathcal{L}_{mlm} = \frac{1}{N_t}\sum_{\boldsymbol{X}\in\mathcal{D}_{tgt}}\frac{1}{|\boldsymbol{X}_m'|}\sum_{x_i'\in\boldsymbol{X}_m'}\mathcal{H}(\boldsymbol{p}_i', y_i') \tag{51}$$

C Additional Experimental Results

C.1 Analysis of Imbalanced Distribution

To verify the presence of imbalanced class distribution in cross-lingual NER, we report the number of spans across different classes in the three datasets we use, as shown in Tables 4, 5, and 6. It can be seen that, regardless of the dataset, the number of non-entity (O) spans is significantly larger than that of entity spans. The imbalance ratio between entities and non-entities (i.e., the ratio of non-entity spans to the least frequent entity spans) in the training set can reach as high as 600 (fa, it), with a minimum ratio of over 40 (hi). This indicates a severe distribution imbalance between entities and non-entities. Moreover, for the CoNLL dataset in Table 4 and the WikiAnn dataset in Table 5, the distribution between entity types is relatively balanced. However, for the challenging MultiCoNER

Methods \ Datasets	CoNLL				WikiAnn			
	de	es	nl	Avg	ar	hi	zh	Avg
GPT4-zero-shot	22.14	50.15	33.08	35.12	37.14	56.43	46.37	46.65
GPT4-few-shot	24.02	52.66	35.21	37.30	39.80	56.90	46.56	47.75
MARAL (Ours)	**80.28**	**86.12**	**85.92**	**84.11**	**78.11**	**84.46**	**66.26**	**76.28**

Methods \ Datasets	MultiCoNER							
	bn	fa	fr	it	pt	sv	uk	Avg
GPT4-zero-shot	43.02	56.23	67.55	69.52	38.68	69.33	73.85	59.74
GPT4-few-shot	43.78	56.40	68.63	70.98	38.94	70.49	74.07	60.47
MARAL (Ours)	**78.21**	**71.28**	**78.10**	**84.49**	**83.02**	**81.23**	**76.85**	**79.03**

Table 7: Performance of GPT-4o on the CoNLL, WikiAnn and MultiCoNER datasets. "GPT4-zero-shot" and "GPT4-few-shot" refer to scenarios where labeled source data is excluded or included as demonstrations in the prompts, respectively.

Task Description:
You are working as a named entity recognition expert and your task is to identify any named entities present in the text. The named entity types include six categories: PER (person), LOC (location), GRP (group), PROD (product), CW (creative work), and MED (medical). Specifically, you will be given a sentence and you should output a list of tuples, where each tuple consists of a span from the input text and its corresponding named entity type. Ensure that you reply with brief, to-the-point answers with no elaboration as truthfully as possible. Moreover, before giving an answer, you need to reflect on whether each named entity is recognized correctly (check one by one and think step by step without additional explanations).

Demonstrations (In English):
Question1: robert gottschalk 1939 academy award winner and founder of panavision
Answer1: [('robert gottschalk', 'PER'), ('academy award', 'CW'), ('panavision', 'GRP')]
Question2: during the reign of the tongzhi emperor (r . 1861 – 1875)
Answer2: [('tongzhi emperor', 'PER')]
.....

Query (In French):
Question: chocolat chanson de l album éponyme dee roméo selvis .

Figure 6: An example prompt for LLMs on the NER task of the French (fr) language in MultiCoNER. The demonstrations are randomly selected from the training set of English (en) and appear only in the prompts used for GPT4-few-shot experiments. The number of demonstrations is set to 20.

dataset in Table 6, a substantial class imbalance is observed among entity spans, with the number of PER entities far exceeding that of PROD and MED entities, which further constrains the representation capacity for minority classes and poses a greater challenge for learning features with large margins.

C.2 Analysis on LLMs for Cross-lingual NER

Recently, with emerging abilities like in-context learning (ICL) (Wei et al., 2022a) and chain-of-thought (CoT) (Wei et al., 2022b), large language models (LLMs) have demonstrated remarkable zero-shot learning performance in a wide range of downstream NLP tasks. To assess the performance of LLMs on NER tasks across different languages, we select GPT-4o, the most advanced LLM, and conduct two experiments: GPT4-zero-shot and GPT4-few-shot on our benchmarks. Our prompt design is shown in Figure 6 and the results are illustrated in Table 7. Using labeled source data as demonstrations, GPT4-few-shot outperforms GPT4-zero-shot across all languages. However, GPT4-few-shot still lags behind current state-of-the-art methods in Table 1, which fine-tune smaller multilingual pre-trained language models (mPLMs). Compared to our method, GPT4-few-shot achieves average F1 score reductions of -46.81%, -28.53%, and -18.43% on CoNLL,

Methods \ Datasets	CoNLL				WikiAnn			
	de	es	nl	Avg	ar	hi	zh	Avg
MARAL	**80.28**	**86.12**	**85.92**	**84.11**	**78.11**	**84.46**	**66.26**	**76.28**
- *w/o global refinery*	79.66	85.81	85.67	83.71	77.74	84.02	65.25	75.67
- *w/o local refinery*	79.82	85.95	85.74	83.84	77.91	84.17	65.50	75.86

Table 8: Fine-grained ablations for label refinery on CoNLL and WikiAnn.

WikiAnn, and MultiCoNER, respectively. This suggests that the direct application of LLMs to the NER task is not optimal. However, we believe that LLMs can still serve as a valuable complement to the cross-lingual NER task. For instance, in our framework MARAL, we can incorporate LLM predictions as an additional criterion for label filtering or to guide the direction of label refinery, which is an important research topic for our future work.

C.3 More Ablations for Label Refinery

For the proposed dual-alignment label refinery in Section 3.2, we integrate the global and local guidance to determine the correction direction of the pseudo-labels. To delve deeper into the specific effects of these two types of guidance, we further compare with two variants: *MARAL w/o global refinery*, which removes the global guidance $\boldsymbol{d}_c$, and *MARAL w/o local refinery*, which removes the local guidance $\boldsymbol{d}_n$. As shown in Table 8, the removal of any guidance leads to a performance drop, highlighting the importance of considering both global and local information. Furthermore, we observe that the performance degradation of *MARAL w/o global refinery* is more significant than that of *MARAL w/o local refinery*, suggesting that, within our framework, global information may be more valuable than local information.

C.4 Discussion on the "O" Class in CL

Typically, spans labeled as "O" often lack consistent semantic characteristics across samples, which makes the notion of constructing a representative prototype for this class less meaningful. Therefore, a natural question arises as to whether the "O" class should be treated differently in the contrastive learning (CL) framework.

Here, we argue that excluding the "O" class from contrastive learning can increase the confusion between entity classes and the "O" class. This perspective is aligned with the concern raised in Ma et al. (2023a), which emphasizes the need for careful handling of non-entity spans.

Methods	de	es	nl	avg
w/ "O"	80.28	86.12	85.92	84.11
w/o "O"	78.53	85.09	84.72	82.78

Table 9: Ablation study for the inclusion of the "O" class in contrastive learning on CoNLL.

To empirically investigate this, we first conducted an ablation study on the CoNLL dataset by removing "O" samples from our maximum-margin contrastive learning. As shown in Table 9, we can observe consistent performance drops across all languages when the "O" class is excluded. Furthermore, we analyzed the prediction errors related to confusion between entity classes and the "O" class. As illustrated in Table 10, excluding the "O" class leads to a notable increase in both Entity-to-O and O-to-Entity errors.

Methods	Entity-to-O	O-to-Entity	Total
w/ "O"	746	384	1130
w/o "O"	976	544	1520

Table 10: Error analysis of Entity-to-O and O-to-Entity misclassifications on German (de).

These results suggest that although "O" spans are semantically heterogeneous, their inclusion in the contrastive framework serves an important regularization role. Rather than constructing a prototype for the "O" class, these spans can effectively function as negative examples that enhance the discriminability of entity representations.

C.5 Case Study

To better illustrate the impact of class imbalance on cross-lingual NER and the capability of MARAL to mitigate this issue, we conduct a case study on three languages of MultiCoNER. As shown in Figure 7, ContProto and GLoDe are prone to confusion between minority classes (PROD and MED),

Language	Sentences
French (fr)	le jour de la saint sébastien{PER} on venait y déposer du grain{MED} en guise d offrande . **ContProto:** saint sébastien{PER} || grain{PROD} **GLoDe:** saint sébastien{PER} || grain{PROD} **MARAL:** saint sébastien{PER} || grain{MED}
Italian (it)	in alcuni casi questi animali sono stati osservati praticare la geofagia{PROD} . **ContProto:** geofagia{MED} **GLoDe:** geofagia{MED} **MARAL:** geofagia{PROD}
Swedish (sv)	chuci schuldiner{PER} – sång{CW} gitaarr{PROD} (1994 – 2001 ; död 2001) **ContProto:** chuci schuldiner{PER} || gitaarr{MED} **GLoDe:** chuci schuldiner{PER} || sång{CW} || gitaarr{MED} **MARAL:** chuci schuldiner{PER} || sång{CW} || gitaarr{PROD}

Figure 7: Case study on French (fr), Italian (it) and Swedish (sv) of the MultiCoNER benchmark. The green (red) texts denote the correct (incorrect) entity labels.

reflecting the impact of class imbalance, where minority classes occupy small margins and are thus hard to distinguish. In contrast, our MARAL effectively classifies these minority classes. This can be attributed to our proposed maximum-margin contrastive learning, which ensures optimal margins for all classes, even under severe class imbalance.

C.6 Training Complexity

While it may appear that computing the von Mises-Fisher (vMF) parameters and performing progressive label refinery incurs additional overhead, both components are in fact highly efficient, and our method remains substantially faster than other state-of-the-art methods. To provide a more intuitive clarify, we present a computational complexity analysis of these two components. Given a total of N samples, N_t target-language samples, K classes, and feature dimension d:

(1) Computation of vMF parameters:

- The prior π is computed as the class-wise sample ratio: $O(N)$.
- The mean direction μ is updated via exponential moving average (EMA): $O(Kd)$.
- The concentration κ is obtained in a single step using Eq. (9): $O(K)$.
- Overall, the complexity is $O(N+Kd)$, which reduces to $O(N)$ in practice, given that $Kd \ll N$ typically holds.

(2) Progressive label refinery:

- The global direction is derived from vMF scores: $O(N_t K)$.
- The local direction is efficiently computed using the FAISS tool (Johnson et al., 2019) for neighbor retrieval: $O(N_t \log N_t)$.
- The update of soft pseudo-label: $O(N_t K)$.
- Overall, the complexity is $O(N_t \log N_t + N_t K)$, which reduces to $O(N_t \log N_t)$ in practice, given that $K \ll N_t$ typically holds.

These analyses confirm the computational efficiency of our algorithm. To further validate this, we compare the average time required to complete one training epoch on OntoNotes Release 4.0 (Weischedel et al., 2011), a large-scale dataset containing 19 entity types. As reported in Table 11, our method maintains a clear runtime advantage over existing approaches. Specifically, MARAL is approximately 18 minutes faster than GLoDe and over 45 minutes faster than ContProto, which also incorporates contrastive learning.

Method	MARAL	GLoDe	ContProto
Time	39 min 58 s	57 min 42 s	85 min 23 s

Table 11: Comparison of epoch time on OntoNotes Release 4.0, where the batch size is set to 32.

D Pseudo-Code of MARAL

We describe the overall training pipeline of our proposed MARAL in Algorithm 1.

Algorithm 1 Pseudo-code of MARAL.

Input: Source data $\mathcal{T}_s = \{(\boldsymbol{s}_i, y_i)\}_{i=1}^{|\mathcal{T}_s|}$ with gold labels, target data $\mathcal{T}_t = \{(\boldsymbol{s}_i, \hat{\boldsymbol{p}}_i)\}_{i=1}^{|\mathcal{T}_t|}$ with soft pseudo-labels, multilingual pre-trained language model h;
for $epoch = 1, 2, \ldots,$ **do**
  $\boldsymbol{z}_i = h(\boldsymbol{s}_i)$ for $\boldsymbol{s}_i$ in $\mathcal{T}_s \cup \mathcal{T}_t$
  $\hat{y}_i = \text{argmax}(\hat{\boldsymbol{p}}_i)$ for $\boldsymbol{s}_i$ in $\mathcal{T}_t$
  `// feature distribution modeling`
  for $c = 0, 1, \ldots, K-1$ **do**
    $\pi_s(c) = \sum_i \mathbb{I}(y_i = c)/|\mathcal{T}_s|$
    $\pi_t(c) = \sum_i \hat{\boldsymbol{p}}_i^c/|\mathcal{T}_t|$
    $\overline{\boldsymbol{z}}'_c = \text{mean}(\boldsymbol{z_i})$, if $y_i = c$ or $\hat{y}_i = c$
    $\overline{\boldsymbol{z}}_c = \beta\overline{\boldsymbol{z}}_c + (1-\beta)\overline{\boldsymbol{z}}'_c$
    $\boldsymbol{\mu}_c = \text{Norm}(\overline{\boldsymbol{z}}_c)$, update κ_c as Eq. (9)
  end for
  compute $\rho(\boldsymbol{z}, y)$ for $\boldsymbol{s}_i$ in $\mathcal{T}_s \cup \mathcal{T}_t$ as Eq. (6)
  `// density-based label filtering`
  $\mathcal{T}_r, \mathcal{T}_u = \text{filter}(\rho(\boldsymbol{z}, y))$
  `// dual-alignment label refinery`
  update $\hat{\boldsymbol{p}}_i$ for $\boldsymbol{s}_i$ in $\mathcal{T}_u$ as Eq. (11)
  `// overall training objectives`
  minimize loss $\mathcal{L}_{all}$ as Eq. (12)
end for