WIDER & CLOSER: Mixture of Short-channel Distillers for Zero-shot Cross-lingual Named Entity Recognition

Jun-Yu Ma[1]*, Beiduo Chen[1]*, Jia-Chen Gu[1], Zhen-Hua Ling[1], Wu Guo[1]† Quan Liu[2,3], Zhigang Chen[4], Cong Liu[1,3]

[1]National Engineering Research Center of Speech and Language Information Processing, University of Science and Technology of China, Hefei, China

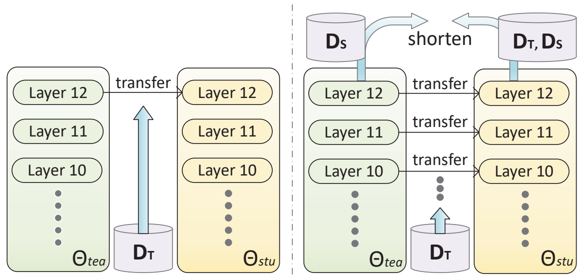
[2]State Key Laboratory of Cognitive Intelligence [3]iFLYTEK Research, Hefei, China

[4]Jilin Kexun Information Technology Co., Ltd.

{mjy1999,beiduo}@mail.ustc.edu.cn, {gujc,zhling,guowu}@ustc.edu.cn, {quanliu,zgchen,congliu2}@iflytek.com

Abstract

Zero-shot cross-lingual named entity recognition (NER) aims at transferring knowledge from annotated and rich-resource data in source languages to unlabeled and lean-resource data in target languages. Existing mainstream methods based on the teacher-student distillation framework ignore the rich and complementary information lying in the intermediate layers of pre-trained language models, and domain-invariant information is easily lost during transfer. In this study, a mixture of short-channel distillers (MSD) method is proposed to fully interact the rich hierarchical information in the teacher model and to transfer knowledge to the student model sufficiently and efficiently. Concretely, a multi-channel distillation framework is designed for sufficient information transfer by aggregating multiple distillers as a mixture. Besides, an unsupervised method adopting parallel domain adaptation is proposed to shorten the channels between the teacher and student models to preserve domain-invariant features. Experiments on four datasets across nine languages demonstrate that the proposed method achieves new state-of-the-art performance on zero-shot cross-lingual NER and shows great generalization and compatibility across languages and fields.



(a) Baseline Distillation (b) MSD

Figure 1: Comparison between the previous mainstream method and the proposed MSD. (a) **Baseline Distillation** is the teacher-student distillation framework. The teacher model is utilized to predict the soft labels of unlabeled target language data, which are further utilized to distill a student model. (b) **MSD** constructs a dozen of channels and shortens the transmission route between the teacher and student models to transfer NER knowledge. Θ_{tea} / Θ_{stu}: teacher / student models; $\mathbf{D_S}$ / $\mathbf{D_T}$: unlabeled source / target language data.

1 Introduction

Named entity recognition (NER) is a fundamental and important task to locate and classify named entities in a text sequence. Recently, deep neural networks have achieved great performance on monolingual NER in rich-resource languages with abundant labeled data (Ye and Ling, 2018; Jia et al., 2020; Chen et al., 2022). However, it is too expensive to annotate a large amount of data in low-resource languages for supervised NER training. This issue drives research on cross-lingual NER, which utilizes the rich-resource annotated data in source languages to alleviate the scarcity of unlabeled lean-resource data in target languages. In this paper, following Wu et al. (2020a), we focus on the extremely low-resource setting, i.e., zero-shot cross-lingual NER, where labeled data is not available in target languages.

The most popular approaches for zero-shot cross-lingual NER are based on distillation (Wu et al., 2020a,b; Chen et al., 2021; Liang et al., 2021). They employed a supervisedly trained teacher model to predict the soft labels of target languages, and then utilized the soft labels to distill a student model, which was first exploited in Wu et al. (2020a). Besides, domain-invariant features have been proven effective for distillation (Nguyen-Meidine et al., 2020; Hu et al., 2019). Chen et al. (2021) proposed to alleviate the representation discrepancy between languages in the teacher model to exploit language-independent features,

*Equal contribution.
†Corresponding author.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 5171–5183
December 7-11, 2022

which were further distilled to the student model. It is worth noting that distillation-based methods assisted with auxiliary tasks have become the mainstream paradigm due to their robustness and scalability, achieving good performance on zero-shot cross-lingual NER (Li et al., 2022).

However, these methods always ignored the rich and complementary information lying in the intermediate layers of multilingual BERT (mBERT) (Devlin et al., 2019). Pires et al. (2019) and Müller et al. (2021) have verified that the upper layers of mBERT are more task-specific and not as important as the lower ones in terms of cross-language transfer. But recent studies just adopted the last layer of mBERT for distillation (Wu et al., 2020b; Li et al., 2022), while neglected the explicit knowledge transfer of the lower layers. Besides, domain-invariant features in the teacher model were first exploited, which were then transferred to the student model via distillation (Chen et al., 2021). However, due to the limitation of transfer learning, it is difficult to fully retain the domain-invariant features to the student model. Furthermore, auxiliary tasks to assist distillation usually require the operations of translation or data sifting (Wu et al., 2020b), resulting in huge pre-processing costs. The ensemble strategies to generate high-quality soft labels or augmented data also require oceans of model parameters and a large number of computational resources.

On account of the above issues, a mixture of short-channel distillers (MSD) method is proposed in this paper to transfer cross-lingual NER knowledge sufficiently and efficiently. On the one hand, a multi-channel distillation framework is designed to let the hierarchical information in the teacher model fully interact with each other, and transfer more complementary information to the student model. Specifically, the teacher model is first trained on the annotated source data with each layer being directly supervised by the labels. Then, each layer of the teacher model is tasked to predict the soft labels of the unlabeled target data. Correspondingly, the layers of the student model are distilled by leveraging the mixture sets of soft labels from the teacher model, constructing multiple information transmission channels for a "*wider*" bridge between the teacher and student models. On the other hand, an unsupervised auxiliary task of parallel domain adaptation is proposed to explicitly transfer domain information. During every batch of distillation, as Figure 1 depicts, unlabeled target data is fed into the student model, while unlabeled source data is fed into both the teacher and student models. The representation discrepancy between the outputs of the teacher and student source language, together with that between the outputs of the teacher source and student target languages is minimized to preserve the cross-model and cross-language domain information respectively. In this way, the domain information can be preserved across models and languages, so that the domains of the teacher and student models can be effectively pulled "*closer*".

Experiments on four datasets across nine languages are conducted to evaluate the effectiveness of the proposed MSD. The results show that our method achieves new state-of-the-art performance on all datasets.

In summary, our contributions are as follows: (1) A multi-channel framework is proposed to leverage the rich, hierarchical and complementary information contained in the teacher model, and to interactively transfer cross-lingual NER knowledge to the student model. (2) An unsupervised auxiliary method is designed to explicitly constrain the discrepancy of teacher/student domains without utilizing any external resources. (3) Experiments on four datasets across nine languages verify the effectiveness and generalization ability of MSD.

2 Related Work

Zero-shot Cross-lingual NER Existing methods on zero-shot cross-lingual NER are mainly separated into three categories: translation-based, feature-based, and distillation-based. Translation-based methods generate pseudo labels for the target language data from the labeled source language data. Jain et al. (2019) projected labels from the source language into the target language by using entity projection information. Xie et al. (2018) and Wu et al. (2020b) translated the annotated source language data to the target language word-by-word. Feature-based methods use the labeled source language data to train the language model for a language-independent representation, such as Wikifier features (Tsai et al., 2016), aligned word representations (Wu and Dredze, 2019), and adversarial learning encodings (Keung et al., 2019). Distillation-based methods are effective in cross-lingual NER by transferring knowledge from a teacher model to a student model (Hinton et al., 2015). The teacher model is first trained on the labeled source language data. Then the student model is trained on soft labels of the target language data predicted by the teacher model. Wu et al. (2020a) trained several teacher models to generate averaged soft labels for the student model. Liang et al. (2021) proposed a reinforced knowledge distillation framework to selectively transfer useful information.

Domain Adaption Label sparsity causes domain shift (Ben-David et al., 2010) in zero-shot cross-lingual NER. The strategy of cross-domain transfer (Qin et al., 2020; Zhang et al., 2021a; Huang et al., 2021) is widely adopted. Existing methods mitigate the discrepancies of sentence patterns between the source and target domains, mainly including multi-level adaptation layers (Lin and Lu, 2018), tensor decomposition (Jia et al., 2019), multi-task learning (Liu et al., 2020b) and word alignment (Lee et al., 2021). However, these methods require sufficient labeled data, in contrast to zero-shot scenarios.

Generally speaking, previous studies on zero-shot cross-lingual NER only leverage the last layer of the teacher model. Besides, the existing NER domain adaptation strategies only constrain the domain-invariant information within the teacher model and transfer them to the student model implicitly. To the best of our knowledge, this paper makes the first attempt to let the rich and hierarchical information in the teacher model fully interact with each other, and further transfer the domain information to the student model explicitly.

3 Preliminary

3.1 Problem Definition

NER is typically formulated as a sequence labeling task. Denote one sentence as $\boldsymbol{x} = \{x_i\}_{i=1}^{L}$ with its labels $\boldsymbol{y} = \{y_i\}_{i=1}^{L}$, where y_i denotes the label of its corresponding word x_i and L denotes the length of the sentence. An NER model generates a sequence of predictions $\bar{\boldsymbol{y}} = \{\bar{y}_i\}_{i=1}^{L}$, where $\bar{y}_i$ denotes the label of x_i annotated by the model. The labeled data $\mathcal{D}_{\text{train}}^{S} = \{(\boldsymbol{x}, \boldsymbol{y})\}$ is available for the source language, while the unlabeled $\mathcal{D}_{\text{train}}^{T} = \{\boldsymbol{x}\}$ and labeled $\mathcal{D}_{\text{test}}^{T} = \{(\boldsymbol{x}, \boldsymbol{y})\}$ are available for target languages. Formally, zero-shot cross-lingual NER aims at achieving good performance on $\mathcal{D}_{\text{test}}^{T}$ by leveraging both $\mathcal{D}_{\text{train}}^{S}$ and $\mathcal{D}_{\text{train}}^{T}$.

3.2 Basic Model

The basic model for cross-lingual NER in this paper consists of a semantic encoder and a classifier. The encoder $\boldsymbol{f_\theta}$ is used to learn and generate the contextual representations of input sentences.

Following Wu and Dredze (2019), the widely-used multilingual pre-trained language model, mBERT, is utilized as the encoder to extract semantic representations. A softmax classification layer is appended to calculate the probability. Finally, the basic model is formulated as follows:

$$\boldsymbol{H} = \boldsymbol{f_\theta}(\boldsymbol{x}), \tag{1}$$

$$\boldsymbol{p}\left(x_i; \Theta\right) = \operatorname{softmax}\left(\boldsymbol{W} \cdot \boldsymbol{h}_i + \boldsymbol{b}\right), \tag{2}$$

where $\boldsymbol{H} = \{\boldsymbol{h_i}\}_{i=1}^{L}$ and $\boldsymbol{h_i}$ is the representation of x_i. $\boldsymbol{p}(x_i; \Theta) \in \mathbb{R}^{|C|}$ with C being a set of entity labels, and $\Theta = \{\boldsymbol{f_\theta}, \boldsymbol{W}, \boldsymbol{b}\}$ denotes all the parameters to be learned.

3.3 Maximum Mean Discrepancy (MMD)

MMD (Long et al., 2015) is a nonparametric test statistic to measure the distance between the distributions of two different random variables (P_s, P_t). MMD is defined in particular function spaces as follows:

$$\operatorname{MMD}(\mathcal{F}, p_s, p_t) = \sup_{f \in \mathcal{F}} \left(\mathbb{E}_{x \sim p_s}[f(x)] - \mathbb{E}_{y \sim p_t}[f(y)]\right), \tag{3}$$

where $\mathcal{F}$ is the unit ball in a universal Reproducing Kernel Hilbert Space (RKHS) denoted by $\mathcal{H}$. An important property of MMD is that $\operatorname{MMD}(\mathcal{F}, p_s, p_t) = 0$ if and only if $P_s = P_t$. Given the source and target sample sets $S = \{s_i\}_{i=1}^{M}$ and $T = \{t_j\}_{j=1}^{N}$ respectively, where s_i or t_j denotes a sample of the set, the empirical estimation of MMD can be defined as:

$$\operatorname{MMD}(S, T) = \left\| \frac{1}{M} \sum_{i=1}^{M} \phi\left(s_i\right) - \frac{1}{N} \sum_{j=1}^{N} \phi\left(t_j\right) \right\|_{\mathcal{H}}, \tag{4}$$

where $\phi(\cdot) : \mathcal{X} \to \mathcal{H}$ is a nonlinear mapping.

In cross-lingual NER, the squared formulation of MMD between the representations ($\boldsymbol{h^s}$ or $\boldsymbol{h^t}$) of the two sets is usually calculated as:

$$\begin{aligned} \operatorname{MMD}^2(S, T) = {} & \frac{1}{(M)^2} \sum_{i,j=1}^{M} G\left(\boldsymbol{h}_i^s, \boldsymbol{h}_j^s\right) + \\ & \frac{1}{(N)^2} \sum_{i,j=1}^{N} G\left(\boldsymbol{h}_i^t, \boldsymbol{h}_j^t\right) - \frac{2}{M \times N} \sum_{i,j=1}^{M \times N} G\left(\boldsymbol{h}_i^s, \boldsymbol{h}_j^t\right), \end{aligned} \tag{5}$$

where G is a Gaussian kernel in this paper.

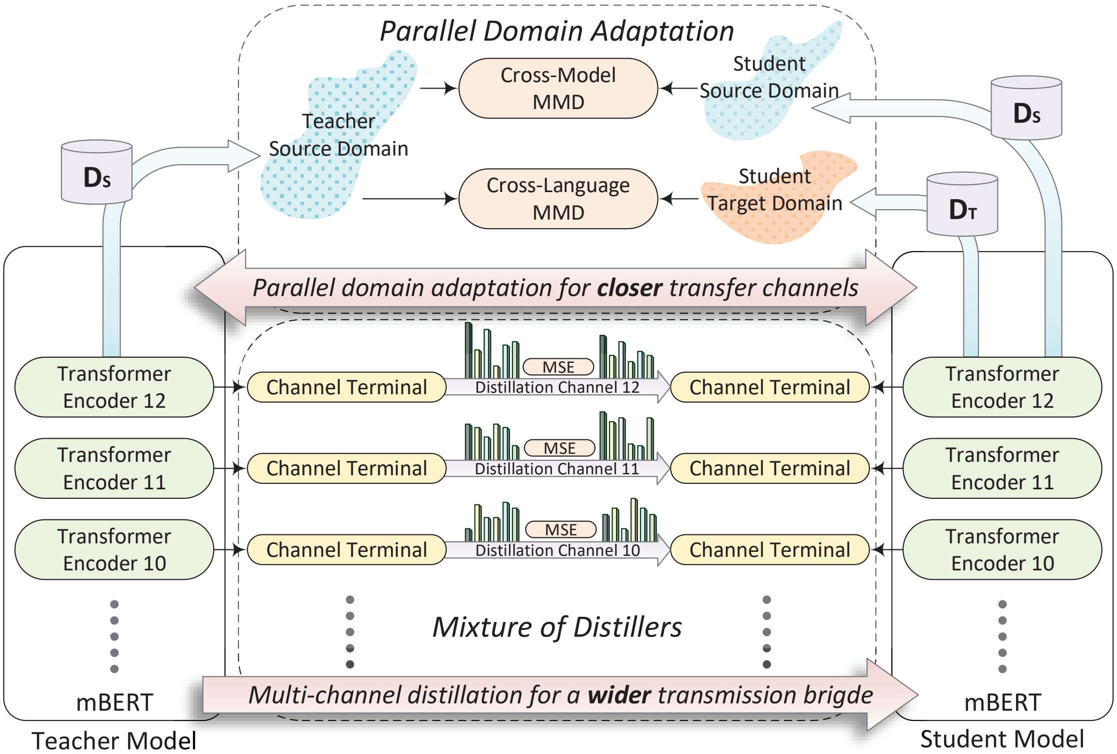


Figure 2: The overall structure of the proposed MSD.

4 Methodology

In this section, we present the detailed framework of the proposed mixture of short-channel distillers (MSD). On the one hand, the mixture of distillers module is introduced. Specifically, multiple channels are built between corresponding layers of the teacher and student's encoders. Then a mixture of weights is employed to control the broadened information transmission route. On the other hand, parallel domain adaptation is conducted to explicitly transfer domain information between the teacher and student models during distillation.

4.1 Mixture of Distillers

Previous studies have verified the importance of lower layers for cross-language transfer (Müller et al., 2021). Our pilot experiments also further illustrate that lower layers of mBERT are critical for NER, which are elaborated in Appendix A.2. To this end, we propose the mixture of distillers framework that fully transfers the complementary information to the student model to equip it with stronger cross-language NER ability. To establish multiple information transmission channels, each layer of the pre-trained mBERT is appended with a classifier. Denote each of these classifiers as a "*channel terminal*", as shown in Figure 2. Given a sentence $\boldsymbol{x}$ of length L with labels $\boldsymbol{y}$ from source language data $\mathcal{D}^S_{\text{train}}$, these could be described as:

$$\boldsymbol{H}^m = \boldsymbol{f_\theta}^m(\boldsymbol{x}), \tag{6}$$

$$\boldsymbol{p}^m\left(x_i;\Theta\right) = \operatorname{softmax}\left(\boldsymbol{W}^m \cdot \boldsymbol{h}_i^m + \boldsymbol{b}^m\right), \tag{7}$$

where $\boldsymbol{H}^m$ is the sentence representation from the m-th layer of mBERT and $\boldsymbol{p}^m\left(x_i;\Theta\right)$ is the probability distribution generated from the corresponding channel terminal.

At the training stage for the teacher, the language model along with several channel terminals are jointly trained on the labeled source language data. Specifically, the channel terminal of the last layer is employed as the main distiller and the others are employed as the auxiliary ones in charge of providing complementary information. Following Wu et al. (2020a), the embedding layer and the bottom three layers of mBERT in the teacher and student models are frozen. So only the top nine layers of the teacher are optimized as:

$$\mathcal{L}_{\text{main}} = \frac{1}{L}\sum_{i=1}^{L}\mathcal{L}_{\text{CE}}\left(\boldsymbol{p}^{12}\left(x_i;\Theta\right), y_i\right), \tag{8}$$

$$\mathcal{L}_{\text{aux}} = \frac{1}{L}\sum_{i=1}^{L}\sum_{m=4}^{11}\lambda_m\mathcal{L}_{\text{CE}}\left(\boldsymbol{p}^{m}\left(x_i;\Theta\right), y_i\right), \tag{9}$$

where $\lambda_m \in \mathbb{R}$ is a trainable parameter representing the contribution degree of the m-th layer[1] and $\mathcal{L}_{\text{CE}}$ is cross entropy loss. The final loss for the teacher model is denoted as:

$$\mathcal{L}_{\text{tea}} = \mathcal{L}_{\text{main}} + \alpha \mathcal{L}_{\text{aux}} , \tag{10}$$

where α is a manually set hyperparameter that regulates the contribution of the auxiliary layers.

For the following knowledge distillation, a student model Θ_{stu} is distilled based on the unlabeled target language data $\mathcal{D}_{\text{train}}^{T}$. In this paper, the student model has the same structure as the teacher model[2]. Firstly, $\mathcal{D}_{\text{train}}^{T}$ is fed into the teacher model Θ_{tea} to obtain its soft labels derived from all the appended channel terminals. Then, as shown in Figure 2, each layer of the student model can be trained along with the student channel terminals using the mixture of soft labels generated from the corresponding layer of the teacher model. Given a sentence $\boldsymbol{x'}$ of length L from $\mathcal{D}_{\text{train}}^{T}$, the distillation loss of the m-th layer is as follows:

$$\mathcal{L}_{\text{m}}^{KD} = \frac{1}{L} \sum_{i=1}^{L} \text{MSE} \left(\boldsymbol{p}^{m} \left(x_i'; \Theta_{tea} \right), \boldsymbol{p}^{m} \left(x_i'; \Theta_{stu} \right) \right) . \tag{11}$$

Following Eq. (10), the loss for the multi-channel distillation is as follows:

$$\begin{aligned} \mathcal{L}_{\text{stu}} &= \mathcal{L}_{\text{main}}^{KD} + \beta \mathcal{L}_{\text{aux}}^{KD} \\ &= \mathcal{L}_{12}^{KD} + \sum_{m=4}^{11} \lambda_m' \mathcal{L}_{m}^{KD} , \end{aligned} \tag{12}$$

where λ_m' and β have the same effect on the student model as λ_m and α do on the teacher model.

4.2 Parallel Domain Adaptation

As aforementioned, the teacher model is trained with hard-labeled source data, but the student model is trained with soft-labeled target data. Thus, training in different manners and languages leads to a huge discrepancy between the domains of the teacher and student models, which causes the loss of domain information during distillation and decreases the transfer efficiency.

In this section, we aim to explicitly transfer domain information to provide a closer route for distillation. The parallel domain adaptation method based on MMD is proposed to preserve domain information between the teacher and student models at sentence-level. As Figure 2 depicts, the cross-model and cross-language MMD losses are proposed to minimize the cross-model and cross-language discrepancies respectively, which are denoted as $\mathcal{L}_{MMD}^{M}$ and $\mathcal{L}_{MMD}^{L}$. During distillation, the soft labels $D_{train}^{S_{tea}}$ and $D_{train}^{S_{stu}}$ are obtained by applying the teacher and student models to the source language data respectively. The $\mathcal{L}_{MMD}^{M}$ could be formulated as:

$$\mathcal{L}_{\text{MMD}}^{M}(D_{train}^{S_{tea}}, D_{train}^{S_{stu}}) = \text{MMD}^2(\boldsymbol{H_{cls}^{S_{tea}}}, \boldsymbol{H_{cls}^{S_{stu}}}), \tag{13}$$

where $\boldsymbol{H_{cls}}$ denotes a set of `[CLS]` token embeddings $\boldsymbol{h_{cls}}$. Meantime, the soft labels $D_{train}^{T_{stu}}$ is obtained by applying the student model to the unlabeled target language data. The $\mathcal{L}_{MMD}^{L}$ is formulated as:

$$\mathcal{L}_{\text{MMD}}^{L}(D_{train}^{S_{tea}}, D_{train}^{T_{stu}}) = \text{MMD}^2(\boldsymbol{H_{cls}^{S_{tea}}}, \boldsymbol{H_{cls}^{T_{stu}}}). \tag{14}$$

Thus, the discrepancies between the teacher and student models, as well as between the source and target languages are both reduced during distillation, strengthening the domain adaptability of the proposed framework for efficient transfer.

The training for the final student model contains two parts: the mixture of distillers and the parallel domain adaptation. The final loss is denoted as:

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{stu}} + \alpha' \mathcal{L}_{\text{MMD}}^{M} + \beta' \mathcal{L}_{\text{MMD}}^{L} , \tag{15}$$

where α' and β' are the weights to balance the contributions of the parallel adaptation methods.

5 Experiments

In this section, the proposed MSD was evaluated on four zero-shot cross-lingual NER datasets and compared with several state-of-the-art models. Some ablation studies were also conducted to validate the effectiveness of the proposed modules.

5.1 Datasets

We conducted experiments on these widely-used benchmark datasets: (1) CoNLL-2002 (Sang, 2002) included Spanish and Dutch; (2) CoNLL-2003 (Sang and Meulder, 2003) included English and German; (3) WikiAnn (Pan et al., 2017) included English and three non-western languages (Arabic, Hindi, and Chinese); (4) mLOWNER (Malmasi

[1] We did study semantic-wise weights by projecting the `[CLS]` token embeddings to a set of trainable parameters, but no further improvement could be achieved.

[2] Our method can also be extended to the framework that the teacher and student models are asymmetrical by designing a mapping function as that in Jiao et al. (2020), which will be a part of our future work.

et al., 2022) included four languages (English, Korean, Farsi, and Turkish). CoNLL-2002 and CoNLL-2003 were annotated with 4 entity types: LOC, MISC, ORG, and PER. WikiAnn was annotated with 3 entity types: LOC, ORG, and PER. mLOWNER was annotated with 6 entity types: LOC, ORG, PER , CW, GRP, PROD.

In this study, the CoNLL[3] and the WikiAnn[4] datasets were just the same as they were initially published. As for the mLOWNER[5] dataset, following Malmasi et al. (2022), 10000 sentences were randomly sampled from the original test set to construct the test set used in this paper. All datasets were annotated with the BIO entity labelling scheme and were divided into the training, development and testing sets. Table 1 shows the statistics of all datasets.

Following the previous work (Wu et al., 2020a), English was employed as the source language in all experiments, and the other languages were employed as target languages. Only unlabeled target language data in the training set was utilized.

Language	Type	Train	Dev	Test
CoNLL dataset (Sang, 2002; Sang and Meulder, 2003)				
English-en (CoNLL-2003)	Sentence	14,987	3,466	3,684
	Entity	23,499	5,942	5,648
German-de (CoNLL-2003)	Sentence	12,705	3,068	3,160
	Entity	11,851	4,833	3,673
Spanish-es (CoNLL-2002)	Sentence	8,323	1,915	1,517
	Entity	18,798	4,351	3,558
Dutch-nl (CoNLL-2002)	Sentence	15,806	2,895	5,195
	Entity	13,344	2,616	3,941
WikiAnn dataset (Pan et al., 2017)				
English-en	Sentence	20,000	10,000	10,000
	Entity	27,931	14,146	13,958
Arabic-ar	Sentence	20,000	10,000	10,000
	Entity	22,500	11,266	11,259
Hindi-hi	Sentence	5,000	1,000	1,000
	Entity	6,124	1,226	1,228
Chinese-zh	Sentence	20,000	10,000	10,000
	Entity	25,031	12,493	12,532
mLOWNER dataset (Malmasi et al., 2022)				
English-en	Sentence	15,300	800	10,000
	Entity	23,553	1,230	15,429
Korean-ko	Sentence	15,300	800	10,000
	Entity	24,643	1,302	16,308
Russian-ru	Sentence	15,300	800	10,000
	Entity	19,840	1,042	12,941
Turkish-tr	Sentence	15,300	800	10,000
	Entity	23,305	1,245	15,209

Table 1: The statistics of the CoNLL (Sang, 2002; Sang and Meulder, 2003), WikiAnn (Pan et al., 2017) and mLOWNER (Malmasi et al., 2022) datasets.

5.2 Evaluation Metrics

Following Sang (2002), entity-level F1-score was used as the evaluation metric. Denote A as the number of all entities classified by the model, B as the number of all correct entities classified by the model, and E as the number of all correct entities, the precision (P), recall (R), and entity-level F1-score (F1) of the model were:

$$\mathrm{P} = \frac{B}{A}, \quad \mathrm{R} = \frac{B}{E}, \quad \mathrm{F1} = \frac{2 \times P \times R}{P + R}. \tag{16}$$

5.3 Baselines

The proposed method was mainly compared with the following (1) distillation-based methods: **TSL** (Wu et al., 2020a), **Unitrans** (Wu et al., 2020b), **AdvPicker** (Chen et al., 2021), **RIKD** (Liang et al., 2021), and **MTMT** (Li et al., 2022), and (2) non-distillation-based methods: **Wiki** (Tsai et al., 2016), **WS** (Ni et al., 2017), **BWET** (Xie et al., 2018), **Adv** (Keung et al., 2019), **BS** (Wu and Dredze, 2019) and **TOF** (Zhang et al., 2021b). Readers can refer to Appendix A.1 for the implementation details of the baseline models.

[3]http://www.cnts.ua.ac.be/conll2003
[4]http://nlp.cs.rpi.edu/wikiann
[5]https://registry.opendata.aws/multiconer/

5.4 Implementation Details

All code was implemented in the PyTorch framework,[6] and is published to help replicate our results.[7] All of the feature encoders mentioned in this paper employed pre-trained cased mBERT (Devlin et al., 2019) in HuggingFace's Transformers where the number of transformer blocks was 12, the hidden layer size was 768, and the number of self-attention heads was 12.

Some hyperparameters were empirically set following Wu and Dredze (2019). Each batch contained 32 examples, with a maximum encoding length of 128. The dropout rate was set to 0.1, and AdamW (Loshchilov and Hutter, 2019) with WarmupLinearSchedule in the Transformers Library (Wolf et al., 2020) was used as optimizer. The parameters of the embedding layer and the bottom three layers of the mBERT used in the teacher model and the student model were frozen.

Following Keung et al. (2019), the other hyper-

[6]https://pytorch.org/
[7]https://github.com/Mckysse/MSD

Method	de	es	nl	Avg
Wiki (Tsai et al., 2016)	48.12	60.55	61.56	56.74
WS (Ni et al., 2017)	58.50	65.10	65.40	63.00
BWET (Xie et al., 2018)	57.76	72.37	71.25	67.13
ADV (Keung et al., 2019)	71.90	74.30	77.60	74.60
BS (Wu and Dredze, 2019)	69.59	74.96	77.57	73.57
TSL (Wu et al., 2020a)	73.16	76.75	80.44	76.78
Unitrans (Wu et al., 2020b)	74.82	79.31	82.90	79.01
AdvPicker (Chen et al., 2021)	75.01	79.00	82.90	78.97
RIKD (Liang et al., 2021)	75.48	77.84	82.46	78.59
TOF (Zhang et al., 2021b)	76.57	80.35	82.79	79.90
MTMT (Li et al., 2022)	76.80	81.82	83.41	80.67
MSD	**77.56**	**81.92**	**85.11**	**81.53**
MSD w/o. distillers	75.31	79.34	83.16	79.27
MSD w/o. $\mathcal{L}_{\text{MMD}}^{L}$	76.68	80.27	84.07	80.34
MSD w/o. $\mathcal{L}_{\text{MMD}}^{M}$	77.12	79.81	84.36	80.43
MSD w/o. all	74.17	77.82	81.31	77.76

Table 2: Evaluation results (%) of entity-level F1-score on the test set of the CoNLL datasets (Sang, 2002; Sang and Meulder, 2003). Results except ours were cited from the published literature. For a fair comparison, scores of the version of RIKD (mBERT) was listed.

parameters were tuned on each target language dev set. All models were trained for 10 epochs and chosen the best checkpoint with the target dev set. For the training of teacher model, the learning rate was set to 5e-5, and the hyperparameter α in Eq. (10) was set to 0.05. For knowledge distillation, keeping the learning rate 2e-5 for the student models and the hyperparameter β was set to 0.05 in Eq. (12), α' and β' were all set to 0.001 in Eq. (15). Furthermore, each experiment was conducted 5 times and reported the mean F1-score.

The number of parameters in a teacher or student model was about 111M. The whole training of MSD was implemented with one GeForce RTX 3090 and consumed about 3 hours.

5.5 Results and Comparisons

Table 2, 3 and 4 reported the zero-shot cross-lingual NER results of different methods on 4 datasets, containing 9 target languages. The results show that the proposed MSD method significantly outperformed the baseline method TSL and achieved new state-of-the-art performance on all target languages. For results on CoNLL, MSD outperformed MTMT (previous SOTA) by absolute margins of 0.76% and 1.70% in terms of German[de] and Dutch[nl] respectively. As for results of non-western languages on WikiAnn and mLOWNER, MSD outperformed MTMT and AdvPicker by marked absolute margins from 2.41% to 10.11% for all target languages. The results clearly demonstrated the effectiveness and generalization ability across languages and datasets of MSD.

Method	ar	hi	zh	Avg
BS (Wu and Dredze, 2019)	42.30	67.60	52.90	54.27
TSL (Wu et al., 2020a)	43.12	69.54	48.12	53.59
RIKD (Liang et al., 2021)	45.96	70.28	50.40	55.55
MTMT (Li et al., 2022)	52.77	70.76	52.26	58.60
MSD	**62.88**	**73.43**	**57.06**	**64.46**
MSD w/o. distillers	54.52	70.22	52.46	59.06
MSD w/o. $\mathcal{L}_{\text{MMD}}^{L}$	56.93	71.50	56.68	61.70
MSD w/o. $\mathcal{L}_{\text{MMD}}^{M}$	58.65	72.11	56.53	62.43
MSD w/o. all	43.17	68.07	49.25	53.49

Table 3: Evaluation results (%) of entity-level F1-score on the test set of the WikiAnn dataset (Pan et al., 2017). Results except ours were cited from the published literature.

Method	ko	ru	tr	Avg
BS (Wu and Dredze, 2019)	51.78	52.33	58.85	54.32
TSL (Wu et al., 2020a)	53.91	54.26	61.15	56.44
AdvPicker (Chen et al., 2021)	56.22	55.65	63.17	58.34
MSD	**61.67**	**58.06**	**67.80**	**62.51**
MSD w/o. distillers	57.23	56.81	65.14	59.72
MSD w/o. $\mathcal{L}_{\text{MMD}}^{L}$	57.88	57.24	67.83	60.98
MSD w/o. $\mathcal{L}_{\text{MMD}}^{M}$	59.12	58.08	67.41	61.53
MSD w/o. all	54.37	54.03	61.55	56.65

Table 4: Evaluation results (%) of entity-level F1-score on the test set of the mLOWNER dataset (Malmasi et al., 2022). Results except ours were obtained by re-implementing these baseline models with the source code provided by the original authors. 5 experiments under the same configuration were conducted for all the methods and the average results were taken as the final numbers. Numbers in bold denote that the improvement over the best performing baseline is statistically significant (t-test with p-value <0.05).

Obviously, existing distillation-based methods were outperformed by the proposed MSD. Specifically, translation and ensemble of teacher models for high-quality soft labels in Unitrans and AdvPicker, as well as iterative knowledge distillation in RIKD requiring huge computational resources, were not adopted in MSD any more. Instead, the proposed MSD fully explored the rich hierarchical information in the teacher model without ensemble, and only utilized the unsupervised data without extra data-process.

Besides, AdvPicker shortened the gap between the source and target languages to derive the language-independent features in the teacher

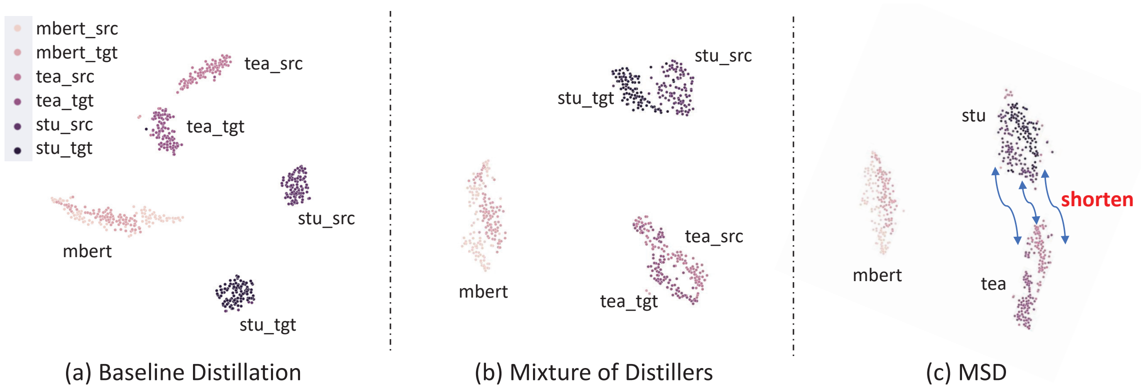


Figure 3: T-SNE visualization (Van der Maaten and Hinton, 2008) of semantic domains of different models by randomly sampling 100 unannotated English (source) and German (target) sentences from the training set of the CoNLL datasets (Sang, 2002; Sang and Meulder, 2003). "*tea/stu*" refers to the teacher/student model respectively. "*src/tgt*" refers to the source/target data respectively. Each point refers to the `[CLS]` representation of a sample in source/target languages. (a) Domains of the basic teacher-student distillation are away from each other. (b) The distribution discrepancy within the teacher (`tea_src`, `tea_tgt`) or within the student (`stu_src`, `stu_tgt`) models is implicitly affected by the mixture of distillers. (c) The distribution discrepancy between the teacher (`tea`) and student (`stu`) models is reduced after further performing parallel domain adaptation.

model, and then distilled the domain information to the student model implicitly. However, the proposed MSD chose to transfer the domain-invariant information directly from the teacher to the student via the parallel domain adaptation. The results demonstrated that the implicit domain transfer in AdvPicker is overshadowed by the explicit domain transfer in MSD. As shown in Figure 3, domain discrepancy between the teacher and student models is vividly reduced by MSD, contributing to a closer transfer route. Further analysis of the difference between the domain transfer manners of AdvPicker and MSD was elaborated in Section 5.6.3.

5.6 Analysis

5.6.1 Ablation Study

To validate the contributions of different components in MSD, the following variants and baselines were conducted to perform the ablation study: (1) MSD w/o. distillers, which only activated the last channel and removed others between the teacher and student models. Besides, the two different MMD losses were still used during distillation. (2) MSD w/o. $\mathcal{L}^L_{\text{MMD}}$, which only removed the cross-language MMD loss. In this case, the teacher and student models had multiple transmission channels but only the cross-model MMD loss was employed. (3) MSD w/o. $\mathcal{L}^M_{\text{MMD}}$, which only removed the cross-model MMD loss correspondingly. (4) MSD w/o. all, which removed all the components mentioned above, and was equivalent to a baseline distillation model as TSL.

Results of the ablation experiments were shown in the bottom five lines of Table 2, 3 and 4 respectively. Some in-depth analysis could be explored: (1) Compared MSD with MSD w/o. distillers, we could see that the removal of distillers caused a significant performance drop, which further demonstrated the importance of leveraging information contained in the intermediate layers of mBERT; (2) Compared MSD with MSD w/o. $\mathcal{L}^L_{\text{MMD}}$ and MSD w/o. $\mathcal{L}^M_{\text{MMD}}$, the parallel domain adaptation contributed to cross-lingual NER significantly. The results well demonstrated that explicitly transfer domain information across models and languages during distillation was reasonable and effective; (3) The two MMD losses were correlated, as $\mathcal{L}^L_{\text{MMD}}$ measured both cross-language and cross-model effects. Removal of $\mathcal{L}^L_{\text{MMD}}$ caused larger performance degradation than removal of $\mathcal{L}^M_{\text{MMD}}$.

The ablation study validated the effectiveness of all components. Moreover, subtle integration of these modules achieved state-of-the-art performance. Not only multiple channels between the teacher and student models should be established to leverage the complementary and hierarchical information of mBERT, but also these channels should be shortened for efficient transfer.

5.6.2 Case Study on Domain Discrepancy

To further illustrate the effectiveness of MSD, various metrics were employed to measure the

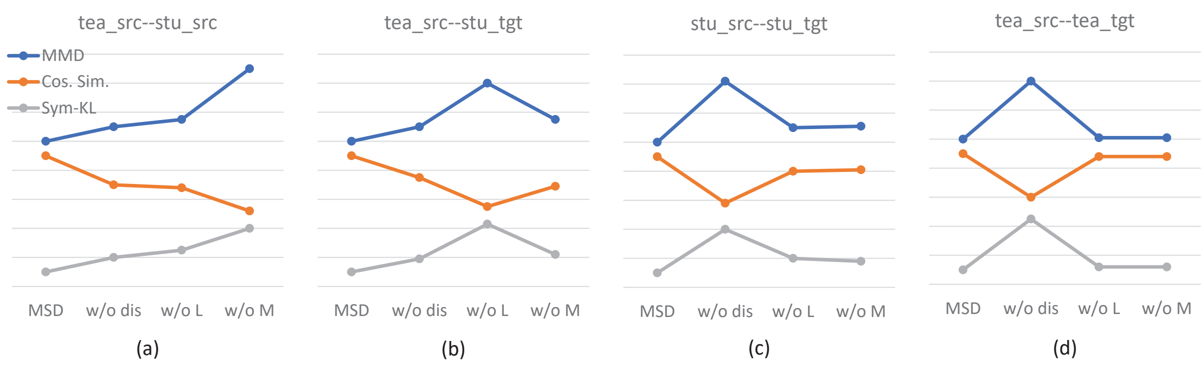


Figure 4: Metrics of MMD, symmetrical KL divergence and cosine similarity on CoNLL02/03 under the ablation settings. "tea/stu" refers to the teacher/student model respectively. "src/tgt" refers to the source/target data respectively. For example, `tea_src--stu_src` refers to the distance between the teacher's source domain and the student's source domain. Each result was calculated from the mean center of the two domains as the same as mentioned in Section 4.2. [`w/o dis`]: MSD w/o. distillers; [`w/o L`]: MSD w/o. $\mathcal{L}^L_{\text{MMD}}$; [`w/o L`]: MSD w/o. $\mathcal{L}^M_{\text{MMD}}$.

distribution discrepancy between different domains: MMD (Long et al., 2015), symmetrical KL divergence (Jiang et al., 2020; Liu et al., 2020a), and cosine similarity (Bromley et al., 1993). The results in Figure 4 is corresponding to the ablation study. Along with Figure 3, the effects of different components could be discussed.

i) Parallel Domain Adaptation. $\mathcal{L}^M_{\text{MMD}}$ pulled the source domain of the teacher and student models closer. As shown in Figure 4 (a), the MMD and symmetrical KL divergence increased and the cosine similarity decreased without $\mathcal{L}^M_{\text{MMD}}$. Similar to $\mathcal{L}^M_{\text{MMD}}$, $\mathcal{L}^L_{\text{MMD}}$ pulled the source domain of the teacher and the target domain of the student closer. ii) Distillers. From Figure 4 (c) and (d), distillers made the domains within the model closer. This effect can be seen intuitively in Figure 3 (b). From Figure 4 (a) and (b), the influence of distillers on the discrepancy between different models was much smaller than that of $\mathcal{L}^L_{\text{MMD}}$ and $\mathcal{L}^M_{\text{MMD}}$. iii) Other results. $\mathcal{L}^L_{\text{MMD}}$ and $\mathcal{L}^M_{\text{MMD}}$ were helpful for reducing the distance between the source and target domains of the student model, as shown in Figure 4 (c). Besides, they alleviated domains discrepancy during distillation, as shown in Figure 3 (c).

5.6.3 Comparison of Transfer Manners

To validate the effectiveness of the explicit domain transfer in MSD, an implicit domain transfer experiment was designed. Imitating Chen et al. (2021), MMD was employed to get language-independent features in the teacher model, and then a baseline distillation was conducted. In contrast, MSD w/o. distillers actually adopted an explicit domain transfer manner. As shown in Appendix A.3, MSD w/o. distillers outperformed methods with implicit domain transfer manners.

6 Conclusion

In this paper, we propose a mixture of short-channel distillers framework for zero-shot cross-lingual NER, including a multi-channel distillation framework to fully leverage the complementary and hierarchical information in the teacher model, and an unsupervised parallel domain adaptation method to effectively pull the domains between teacher and student models closer. Experimental results show that the proposed method outperforms previous methods on four datasets across nine languages. In the future, we will extend this method to languages where data resources are scarcer.

Limitations

Our method has certain limitations, such as it cannot be used for target languages without any text data. Furthermore, although the results show great performance, more efforts are required to explore the hidden impact of distillers as shown in the t-SNE graph, which will help the application of the proposed model in the future.

Acknowledgements

We thank anonymous reviewers for their valuable comments. This work was partially funded by the National Natural Science Foundation of China (Grant No. U1836219).

References

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman

Vaughan. 2010. A theory of learning from different domains. *Mach. Learn.*, 79(1-2):151–175.

Jane Bromley, James W. Bentz, Léon Bottou, Isabelle Guyon, Yann LeCun, Cliff Moore, Eduard Säckinger, and Roopak Shah. 1993. Signature verification using A "siamese" time delay neural network. *Int. J. Pattern Recognit. Artif. Intell.*, 7(4):669–688.

Beiduo Chen, Jun-Yu Ma, Jiajun Qi, Wu Guo, Zhen-Hua Ling, and Quan Liu. 2022. USTC-NELSLIP at semeval-2022 task 11: Gazetteer-adapted integration network for multilingual complex named entity recognition. In *Proceedings of the 16th International Workshop on Semantic Evaluation, SemEval@NAACL 2022, Seattle, Washington, United States, July 14-15, 2022*, pages 1613–1622. Association for Computational Linguistics.

Weile Chen, Huiqiang Jiang, Qianhui Wu, Börje Karlsson, and Yi Guan. 2021. Advpicker: Effectively leveraging unlabeled data via adversarial discriminator for cross-lingual NER. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 743–753. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531.

Mengting Hu, Yike Wu, Shiwan Zhao, Honglei Guo, Renhong Cheng, and Zhong Su. 2019. Domain-invariant feature distillation for cross-domain sentiment classification. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 5558–5567. Association for Computational Linguistics.

Kuan-Hao Huang, Wasi Uddin Ahmad, Nanyun Peng, and Kai-Wei Chang. 2021. Improving zero-shot cross-lingual transfer learning via robust training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 1684–1697. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 1083–1092. Association for Computational Linguistics.

Chen Jia, Yuefeng Shi, Qinrong Yang, and Yue Zhang. 2020. Entity enhanced BERT pre-training for chinese NER. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6384–6396. Association for Computational Linguistics.

Chen Jia, Liang Xiao, and Yue Zhang. 2019. Cross-domain NER using cross-domain language modeling. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2464–2474. Association for Computational Linguistics.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2020. SMART: robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2177–2190. Association for Computational Linguistics.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tinybert: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 4163–4174. Association for Computational Linguistics.

Phillip Keung, Yichao Lu, and Vikas Bhardwaj. 2019. Adversarial learning with contextual embeddings for zero-resource cross-lingual classification and NER. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1355–1360. Association for Computational Linguistics.

Dohyeon Lee, Jaeseong Lee, Gyewon Lee, Byung-Gon Chun, and Seung-won Hwang. 2021. SCOPA: soft code-switching and pairwise alignment for zero-shot cross-lingual transfer. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, pages 3176–3180. ACM.

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), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 170–179. 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 *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, pages 3231–3239. ACM.

Bill Yuchen Lin and Wei Lu. 2018. Neural adaptation layers for cross-domain named entity recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2012–2022. Association for Computational Linguistics.

Xiaodong Liu, Hao Cheng, Pengcheng He, Weizhu Chen, Yu Wang, Hoifung Poon, and Jianfeng Gao. 2020a. Adversarial training for large neural language models. *CoRR*, abs/2004.08994.

Zihan Liu, Genta Indra Winata, and Pascale Fung. 2020b. Zero-resource cross-domain named entity recognition. In *Proceedings of the 5th Workshop on Representation Learning for NLP, RepL4NLP@ACL 2020, Online, July 9, 2020*, pages 1–6. Association for Computational Linguistics.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. 2015. Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 97–105. JMLR.org.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022. MultiCoNER: A large-scale multilingual dataset for complex named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3798–3809, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Benjamin Müller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 2021. First align, then predict: Understanding the cross-lingual ability of multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 2214–2231. Association for Computational Linguistics.

Le Thanh Nguyen-Meidine, Eric Granger, Madhu Kiran, Jose Dolz, and Louis-Antoine Blais-Morin. 2020. Joint progressive knowledge distillation and unsupervised domain adaptation. In *2020 International Joint Conference on Neural Networks, IJCNN 2020, Glasgow, United Kingdom, July 19-24, 2020*, pages 1–8. IEEE.

Jian Ni, Georgiana Dinu, and Radu Florian. 2017. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1470–1480. Association for Computational Linguistics.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1946–1958. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4996–5001. Association for Computational Linguistics.

Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2020. Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual NLP. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3853–3860. ijcai.org.

Erik F. Tjong Kim Sang. 2002. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *Proceedings of the 6th Conference on Natural Language Learning, CoNLL 2002, Held in cooperation with COLING 2002, Taipei, Taiwan, 2002*. ACL.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, May 31 - June 1, 2003*, pages 142–147. ACL.

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, CoNLL 2016, Berlin, Germany, August 11-12, 2016, pages 219–228. ACL.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Qianhui Wu, Zijia Lin, Börje Karlsson, Jianguang 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, ACL 2020, Online, July 5-10, 2020*, pages 6505–6514. Association for Computational Linguistics.

Qianhui Wu, Zijia Lin, Börje F. Karlsson, Biqing Huang, and Jianguang Lou. 2020b. Unitrans : Unifying model transfer and data transfer for cross-lingual named entity recognition with unlabeled data. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3926–3932. ijcai.org.

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 2019, Hong Kong, China, November 3-7, 2019*, pages 833–844. Association for Computational Linguistics.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime G. Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 369–379. Association for Computational Linguistics.

Zhi-Xiu Ye and Zhen-Hua Ling. 2018. Hybrid semi-markov CRF for neural sequence labeling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 235–240. Association for Computational Linguistics.

Tao Zhang, Congying Xia, Philip S. Yu, Zhiwei Liu, and Shu Zhao. 2021a. PDALN: progressive domain adaptation over a pre-trained model for low-resource cross-domain named entity recognition. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5441–5451. Association for Computational Linguistics.

Ying Zhang, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2021b. Target-oriented fine-tuning for zero-resource named entity recognition. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 1603–1615. Association for Computational Linguistics.

A Appendices

A.1 Baseline Models

We mainly compared our method with the following distillation-based methods.

TSL (Wu et al., 2020a) proposed a teacher-student learning model, via using source-language models as teachers to train a student model on unlabeled data in the target language for cross-lingual NER.

Unitrans (Wu et al., 2020b) unified both model transfer and data transfer based on their complementarity via enhanced knowledge distillation on unlabeled target-language data.

AdvPicker (Chen et al., 2021) proposed a novel approach to combine the feature-based method and pseudo labeling via language adversarial learning for cross-lingual NER.

RIKD (Liang et al., 2021) proposed a reinforced knowledge distillation framework.

MTMT (Li et al., 2022) proposed an unsupervised multiple-task and multiple-teacher model for cross-lingual NER.

In addition, **Wiki** (Tsai et al., 2016), **WS** (Ni et al., 2017), **BWET** (Xie et al., 2018), **Adv** (Keung et al., 2019), **BS** (Wu and Dredze, 2019) and **TOF** (Zhang et al., 2021b) were non-distillation-based methods.

A.2 Pilot Experiment

Previous studies have verified the importance of lower layers for cross-language transfer (Müller et al., 2021). Our pilot experiments also further illustrate that lower layers of mBERT are critical for NER as shown in Table 5.

Method	**en**
BS (Wu and Dredze, 2019)	91.30
mixture	**91.49**

Table 5: Evaluation results (%) of entity-level F1-score on the English test set of the CoNLL dataset (Sang, 2002; Sang and Meulder, 2003). Both models were trained with the training set data of English in CoNLL. BS was the basic model in Section 3.2. mixture represented the teacher model described in Section 4.1.

A.3 Comparison of Transfer Manners

Table 6, 7 and 8 reported the results of different transfer manners. Imitating AdvPicker (Chen et al., 2021), TSL w. MMD was designed to get language-independent features in the teacher model, and then a baseline distillation was conducted, which was an implicit transfer manner. MSD w/o. distillers represented the explicit transfer manner.

Method	**de**	**es**	**nl**	**Avg**
AdvPicker	75.01	79.00	82.90	78.97
TSL w. MMD	75.02	78.83	82.73	78.86
MSD w/o. distillers	**75.31**	**79.34**	**83.16**	**79.27**

Table 6: Evaluation results (%) of entity-level F1-score on the test set of the CoNLL datasets (Sang, 2002; Sang and Meulder, 2003).

Method	**ar**	**hi**	**zh**	**Avg**
AdvPicker	53.12	69.88	51.09	57.69
TSL w. MMD	53.47	70.09	50.12	57.89
MSD w/o. distillers	**54.52**	**70.22**	**52.46**	**59.06**

Table 7: Evaluation results (%) of entity-level F1-score on the test set of the WikiAnn dataset (Pan et al., 2017).

Method	**ko**	**ru**	**tr**	**Avg**
AdvPicker	56.22	55.65	63.17	58.34
TSL w. MMD	56.68	56.28	63.13	58.69
MSD w/o. distillers	**57.23**	**56.81**	**65.14**	**59.72**

Table 8: Evaluation results (%) of entity-level F1-score on the test set of the mLOWNER dataset (Malmasi et al., 2022).