Single-/Multi-Source Cross-Lingual NER via Teacher-Student Learning on Unlabeled Data in Target Language

Qianhui Wu[1], **Zijia Lin**[2], **Börje F. Karlsson**[2], **Jian-Guang Lou**[2], **and Biqing Huang**[1]
[1]Beijing National Research Center for Information Science and Technology (BNRist)
Department of Automation, Tsinghua University, Beijing 100084, China
wuqianhui@tsinghua.org.cn, hbq@tsinghua.edu.cn
[2]Microsoft Research, Beijing 100080, China
{zijlin,borje.karlsson,jlou}@microsoft.com

Abstract

To better tackle the named entity recognition (NER) problem on languages with little/no labeled data, cross-lingual NER must effectively leverage knowledge learned from source languages with rich labeled data. Previous works on cross-lingual NER are mostly based on label projection with pairwise texts or direct model transfer. However, such methods either are not applicable if the labeled data in the source languages is unavailable, or do not leverage information contained in unlabeled data in the target language. In this paper, we propose a teacher-student learning method to address such limitations, where NER models in the source languages are used as teachers to train a student model on unlabeled data in the target language. The proposed method works for both single-source and multi-source cross-lingual NER. For the latter, we further propose a similarity measuring method to better weight the supervision from different teacher models. Extensive experiments for 3 target languages on benchmark datasets well demonstrate that our method outperforms existing state-of-the-art methods for both single-source and multi-source cross-lingual NER.

1 Introduction

Named entity recognition (NER) is the task of identifying text spans that belong to pre-defined categories, like locations, person names, *etc*. It's a fundamental component in many downstream tasks, and has been greatly advanced by deep neural networks (Lample et al., 2016; Chiu and Nichols, 2016; Peters et al., 2017). However, these approaches generally require massive manually labeled data, which prohibits their adaptation to low-resource languages due to high annotation costs.

One solution to tackle that is to transfer knowledge from a source language with rich labeled data to a target language with little or even no labeled data, which is referred to as cross-lingual NER (Wu and Dredze, 2019; Wu et al., 2020). In this paper, following Wu and Dredze (2019) and Wu et al. (2020), we focus on the extreme scenario of cross-lingual NER where **no labeled data** is available in the target language, which is challenging in itself and has attracted considerable attention from the research community in recent years.

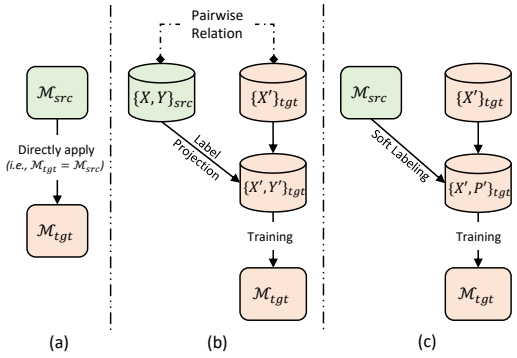

Figure 1: Comparison between previous cross-lingual NER methods (**a/b**) and the proposed method (**c**). **(a)**: direct model transfer; **(b)**: label projection with pairwise texts; **(c)**: proposed teacher-student learning method. $\mathcal{M}_{src/tgt}$: learned NER model for source/target language; $\{X,Y\}_{src}$: labeled data in source language; $\{X'\}_{tgt}$: unlabeled data in target language; $\{X',Y'\}_{tgt}/\{X',P'\}_{tgt}$: pseudo-labeled data in target language with hard labels / soft labels.

Previous works on cross-lingual NER are mostly based on label projection with pairwise texts or direct model transfer. Label-projection based methods focus on using labeled data in a source language to generate pseudo-labelled data in the target language for training an NER model. For example, Ni et al. (2017) creates automatically labeled NER data for the target language via label projection on comparable corpora and develops a heuristic scheme to select good-quality projection-labeled data. Mayhew et al. (2017) and Xie et al. (2018)

translate the source language labeled data at the phrase/word level to generate pairwise labeled data for the target language. Differently, model-transfer based methods (Wu and Dredze, 2019; Wu et al., 2020) focus on training a shared NER model on the labeled data in the source language with language-independent features, such as cross-lingual word representations (Devlin et al., 2019), and then directly testing the model on the target language.

However, there are limitations in both label-projection based methods and model-transfer based methods. The former relies on labeled data in the source language for label projection, and thus is not applicable in cases where the required labeled data is inaccessible (*e.g.*, due to privacy/sensitivity issues). Meanwhile, the later does not leverage unlabeled data in the target language, which can be much cheaper to obtain and probably contains very useful language information.

In this paper, we propose a teacher-student learning method for cross-lingual NER to address the mentioned limitations. Specifically, we leverage multilingual BERT (Devlin et al., 2019) as the base model to produce language-independent features. A previously trained NER model for the source language is then used as a teacher model to predict the probability distribution of entity labels (*i.e.*, soft labels) for each token in the *non-pairwise* unlabeled data in the target language. Finally, we train a student NER model for the target language using the pseudo-labeled data with such soft labels. The proposed method does not rely on labelled data in the source language, and it also leverages the available information from unlabeled data in the target language, thus avoiding the mentioned limitations of previous works. Note that we use the teacher model to predict soft labels rather than hard labels (*i.e.*, one-hot labelling vector), as soft labels can provide much more information (Hinton et al., 2015) for the student model. Figure 1 shows the differences between the proposed teacher-student learning method and the typical label-projection or model-transfer based methods.

We further extend our teacher-student learning method to multi-source cross-lingual NER, considering that there are usually multiple source languages available in practice and we would prefer transferring knowledge from all source languages rather than a single one. In this case, our method still enjoys the same advantages in terms of data availability and inference efficiency, compared with existing works (Täckström, 2012; Chen et al., 2019; Enghoff et al., 2018; Rahimi et al., 2019). Moreover, we propose a method to measure the similarity between each source language and the target language, and use this similarity to better weight the supervision from the corresponding teacher model.

We evaluate our proposed method for 3 target languages on benchmark datasets, using different source language settings. Experimental results show that our method outperforms existing state-of-the-art methods for both single-source and multi-source cross-lingual NER. We also conduct case studies and statistical analyses to discuss why teacher-student learning reaches better results.

The main contributions of this work are:

- We propose a teacher-student learning method for single-source cross-lingual NER, which addresses limitations of previous works *w.r.t* data availability and usage of unlabeled data.
- We extend the proposed method to multi-source cross-lingual NER, using a measure of the similarities between source/target languages to better weight teacher models.
- We conduct extensive experiments validating the effectiveness and reasonableness of the proposed methods, and further analyse why they attain superior performance.

2 Related Work

Single-Source Cross-Lingual NER: Such approaches consider one single source language for knowledge transfer. Previous works can be divided into two categories: label-projection and model-transfer based methods.

Label-projection based methods aim to build pseudo-labeled data for the target language to train an NER model. Some early works proposed to use bilingual parallel corpora and project model expectations (Wang and Manning, 2014) or labels (Ni et al., 2017) from the source language to the target language with external word alignment information. But obtaining parallel corpora is expensive or even infeasible. To tackle that, recent methods proposed to firstly translate source-language labeled data at the phrase level (Mayhew et al., 2017) or word level (Xie et al., 2018), and then directly copy labels across languages. But translation introduces extra noise due to sense ambiguity and word order differences between languages, thus hurting the trained model.

Model-transfer based methods generally rely on language-independent features (*e.g.*, cross-lingual word embeddings (Ni et al., 2017; Huang et al., 2019; Wu and Dredze, 2019; Moon et al., 2019), word clusters (Täckström et al., 2012), gazetteers (Zirikly and Hagiwara, 2015), and *wikifier* features (Tsai et al., 2016)), so that a model trained with such features can be directly applied to the target language. For further improvement, Wu et al. (2020) proposed constructing a pseudo-training set for each test case and fine-tuning the model before inference. However, these methods do not leverage any unlabeled data in the target language, though such data can be easy to obtain and benefit the language/domain adaptation.

Multi-Source Cross-Lingual NER: Multi-source cross-lingual NER considers multiple source languages for knowledge transfer.

Täckström (2012) and Moon et al. (2019) concatenated the labeled data of all source languages to train a unified model, and performed cross-lingual NER in a direct model transfer manner. Chen et al. (2019) leveraged adversarial networks to learn language-independent features, and learns a mixture-of-experts model (Shazeer et al., 2017) to weight source models at the token level. However, both methods straightly rely on the availability of labeled data in the source languages.

Differently, Enghoff et al. (2018) implemented multi-source label projection and studied how source data quality influence performance. Rahimi et al. (2019) applied truth inference to model the transfer annotation bias from multiple source-language models. However, both methods make predictions via an ensemble of source-language models, which is cumbersome and computationally expensive, especially when a source-language model has massive parameter space.

Teacher-Student Learning: Early applications of teacher-student learning targeted model compression (Bucilu et al., 2006), where a small student model is trained to mimic a pre-trained, larger teacher model or ensemble of models. It was soon applied to various tasks like image classification (Hinton et al., 2015; You et al., 2017), dialogue generation (Peng et al., 2019), and neural machine translation (Tan et al., 2019), which demonstrated the usefulness of the knowledge transfer approach.

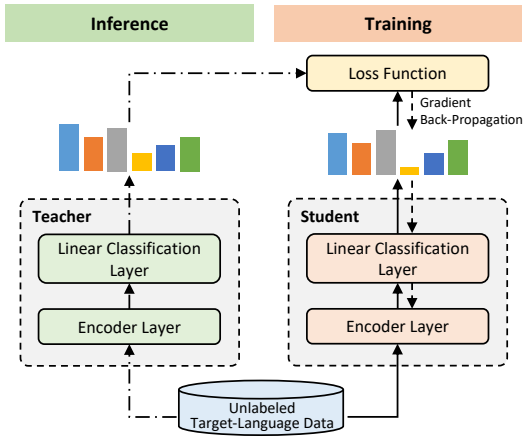

Figure 2: Framework of the proposed teacher-student learning method for **single-source** cross-lingual NER.

In this paper, we investigate teacher-student learning for the task of cross-lingual NER, in both single-source and multi-source scenarios. Different from previous works, our proposed method does not rely on the availability of labelled data in source languages or any pairwise texts, while it can also leverage extra information in unlabeled data in the target language to enhance the cross-lingual transfer. Moreover, compared with using an ensemble of source-language models, our method uses a single student model for inference, which can enjoy higher efficiency.

3 Methodology

Named entity recognition can be formulated as a sequence labeling problem, *i.e.*, given a sentence $\boldsymbol{x} = \{x_i\}_{i=1}^{L}$ with L tokens, an NER model is supposed to infer the entity label y_i for each token x_i and output a label sequence $\boldsymbol{y} = \{y_i\}_{i=1}^{L}$. Under the paradigm of cross-lingual NER, we assume there are K source-language models previously trained with language-independent features. Our proposed teacher-student learning method then uses those K source-language models as teachers to train an effective student NER model for the target language on its unlabeled data D_{tgt}.

3.1 Single-Source Cross-Lingual NER

Here we firstly consider the case of only one source language ($K = 1$) for cross-lingual NER. The overall framework of the proposed teacher-student learning method for single-source cross-lingual NER is illustrated in Figure 2.

3.1.1 NER Model Structure

As shown in Figure 2, for simplicity, we employ the same neural network structure for both teacher (source-language) and student (target-language) NER models. Note that the student model is flexible and its structure can be determined according to the trade-off between performance and training/inference efficiency.

Here the adopted NER model consists of an encoder layer and a linear classification layer. Specifically, given an input sequence $\boldsymbol{x} = \{x_i\}_{i=1}^{L}$ with L tokens, the encoder layer f_θ maps it into a sequence of hidden vectors $\boldsymbol{h} = \{h_i\}_{i=1}^{L}$:

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

Here $f_\theta(\cdot)$ can be any encoder model that produces cross-lingual token representations, and h_i is the hidden vector corresponding to the i-th token x_i.

With each h_i derived, the linear classification layer computes the probability distribution of entity labels for the corresponding token x_i, using a *softmax* function:

$$p(x_i, \Theta) = \text{softmax}(W h_i + b) \tag{2}$$

where $p(x_i, \Theta) \in \mathbb{R}^{|C|}$ with C being the entity label set, and $\Theta = \{f_\theta, W, b\}$ denotes the to-be-learned model parameters.

3.1.2 Teacher-Student Learning

Training: We train the student model to mimic the output probability distribution of entity labels by the teacher model, on the unlabeled data in the target language D_{tgt}. Knowledge from the teacher model is expected to transfer to the student model, while the student model can also leverage helpful language-specific information available in the unlabeled target-language data.

Given an unlabeled sentence $\boldsymbol{x'} \in D_{tgt}$ in the target language, the teacher-student learning loss *w.r.t* $\boldsymbol{x'}$ is formulated as the *mean squared error* (MSE) between the output probability distributions of entity labels by the student model and those by the teacher model, averaged over tokens. Note that here we follow Yang et al. (2019) and use the MSE loss, because it is symmetric and mimics all probabilities equally. Suppose that for the i-token in $\boldsymbol{x'}$, *i.e.*, x'_i, the probability distribution of entity labels output by the student model is denoted as $\hat{p}(x'_i, \Theta_S)$, and that output by the teacher model as $\tilde{p}(x'_i, \Theta_T)$. Here Θ_S and Θ_T, respectively, denote the parameters of the student and the teacher models. The teacher-student learning loss *w.r.t* $\boldsymbol{x'}$ is then defined as:

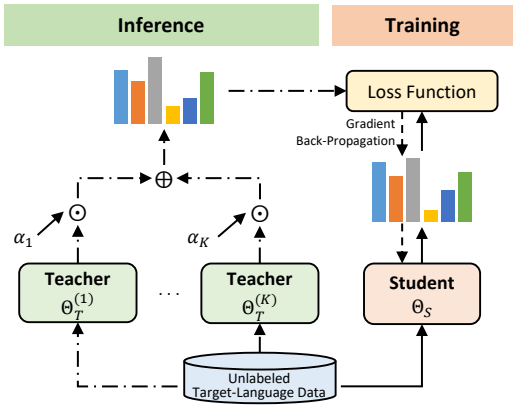

Figure 3: Framework of the proposed teacher-student learning method for **multi-source** cross-lingual NER.

$$\mathcal{L}(\boldsymbol{x'}, \Theta_S) = \frac{1}{L}\sum_{i=1}^{L} \text{MSE}\left(\hat{p}(x'_i, \Theta_S), \tilde{p}(x'_i, \Theta_T)\right) \tag{3}$$

And the whole training loss is the summation of losses *w.r.t* all sentences in D_{tgt}, as defined below.

$$\mathcal{L}(\Theta_S) = \sum_{\boldsymbol{x'} \in D_{tgt}} \mathcal{L}(\boldsymbol{x'}, \Theta_S) \tag{4}$$

Minimizing $\mathcal{L}(\Theta_S)$ will derive the student model.

Inference: For inference in the target language, we only utilize the learned student model to predict the probability distribution of entity labels for each token x_i in a test sentence $\boldsymbol{x}$. Then we take the entity label $c \in C$ with the highest probability as the predicted label y_i for x_i:

$$y_i = \arg\max_c \hat{p}(x_i, \Theta_S)_c \tag{5}$$

where $p(x_i, \Theta_S)_c$ denotes the predicted probability corresponding to the entity label c in $p(x_i, \Theta_S)$.

3.2 Multi-Source Cross-Lingual NER

The framework of the proposed teacher-student learning method for multi-source ($K > 1$) cross-lingual NER is illustrated in Figure 3.

3.2.1 Extension to Multiple Teacher Models

As illustrated in Figure 3, we extend the single-teacher framework in Figure 2 into a multi-teacher one, while keeping the student model unchanged.

Note that, for simplicity, all teacher models and the student model use the same model structure as

3.1.1. Take the k-th teacher model for example, and denote its parameters as $\Theta_T^{(k)}$. Given a sentence $\boldsymbol{x'} = \{x'_i\}_{i=1}^L$ with L tokens from the unlabeled data D_{tgt} in the target language, the output probability distribution of entity labels *w.r.t* the i-th token x_i can be derived as Eq. 1 and 2, which is denoted as $\tilde{p}(x'_i, \Theta_T^{(k)})$. To combine all teacher models, we add up their output probability distributions with a group of weights $\{\alpha_k\}_{k=1}^K$ as follows.

$$\tilde{p}(x'_i, \Theta_T) = \sum_{k=1}^{K} \alpha_k \cdot \tilde{p}(x'_i, \Theta_T^{(k)}) \tag{6}$$

where $\tilde{p}(x'_i, \Theta_T)$ is the combined probability distribution of entity labels, $\Theta_T = \{\Theta_T^{(k)}\}_{k=1}^K$ is the set of parameters of all teacher models, and α_k is the weight corresponding to the k-th teacher model, with $\sum_{k=1}^K \alpha_k = 1$ and $\alpha_k \geq 0, \forall k \in \{1, \ldots, K\}$.

3.2.2 Weighting Teacher Models

Here we elaborate on how to derive the weights $\{\alpha_k\}_{k=1}^K$ in cases *w/* or *w/o* ***unlabeled*** data in the source languages. Source languages more similar to the target language should generally be assigned higher weights to transfer more knowledge.

Without Any Source-Language Data: It is straightforward to average over all teacher models:

$$\alpha_k = \frac{1}{K}, \ \forall k \in \{1, 2, \ldots, K\} \tag{7}$$

With Unlabeled Source-Language Data: As no labeled data is available, existing supervised language/domain similarity learning methods for a target task (*i.e.*, NER) (McClosky et al., 2010) are not applicable here. Inspired by Pinheiro (2018), we propose to introduce a language identification auxiliary task for calculating similarities between source and target languages, and then weight teacher models based on this metric.

In the language identification task, for the k-th source language, each unlabeled sentence $\boldsymbol{u}^{(k)}$ in it is associated with the language index k to build its training dataset, denoted as $D_{src}^{(k)} = \{(\boldsymbol{u}^{(k)}, k)\}$. We also assume that in the m-dimensional language-independent feature space, sentences from each source language should be clustered around the corresponding language embedding vector. We thus introduce a learnable language embedding vector $\mu^{(k)} \in \mathbb{R}^m$ for the k-th source language, and then utilize a *bilinear* operator to measure similarity between a given sentence $\boldsymbol{u}$ and the k-th source language:

$$s(\boldsymbol{u}, \mu^{(k)}) = g^T(\boldsymbol{u}) M \mu^{(k)} \tag{8}$$

where $g(\cdot)$ can be any language-independent model that outputs sentence embeddings, and $M \in \mathbb{R}^{m \times m}$ denotes the parameters of the *bilinear* operator.

By building a language embedding matrix $P \in \mathbb{R}^{m \times K}$ with each $\mu^{(k)}$ *column by column*, and applying a *softmax* function over the *bilinear* operator, we can derive language-specific probability distributions *w.r.t* $\boldsymbol{u}$ as below.

$$q(\boldsymbol{u}, M, P) = \text{softmax}\left(g^T(\boldsymbol{u}) M P\right) \tag{9}$$

Then the parameters M and P are trained to identify the language of each sentence in $\{D_{src}^{(k)}\}_{k=1}^K$, via minimizing the *cross-entropy* (CE) loss:

$$\begin{aligned}\mathcal{L}(P, M) = &-\frac{1}{Z} \sum_{(\boldsymbol{u}^{(k)}, k) \in D_{src}} \text{CE}\left(q(\boldsymbol{u}^{(k)}, M, P), k\right) \\ &+ \gamma \| P P^T - I \|_F^2\end{aligned} \tag{10}$$

where D_{src} is the union set of $\{D_{src}^{(k)}\}_{k=1}^K$, $Z = |D_{src}|$, $\| \cdot \|_F^2$ denotes the squared Frobenius norm, and I is an identity matrix. The regularizer in $\mathcal{L}(P, M)$ is to encourage different dimensions of the language embedding vectors to focus on different aspects, with $\gamma \geq 0$ being its weighting factor.

With learned M and $P = [\mu^{(1)}, \mu^{(2)}, \ldots, \mu^{(K)}]$, we compute the weights $\{\alpha_k\}_{i=1}^K$ using the unlabeled data in the target language D_{tgt}:

$$\alpha_k = \frac{1}{|D_{tgt}|} \sum_{\boldsymbol{x'} \in D_{tgt}} \frac{\exp\left(s(\boldsymbol{x'}, \mu^{(k)})/\tau\right)}{\sum_{i=1}^K \exp\left(s(\boldsymbol{x'}, \mu^{(i)})/\tau\right)} \tag{11}$$

where τ is a temperature factor to smooth the output probability distribution. In our experiments, we set it as the variance of all values in $\{s(\boldsymbol{x'}, \mu^{(k)})\}, \forall \boldsymbol{x'} \in D_{tgt}, \forall k \in \{1, ..., K\}$, so that α_k would not be too biased to either 0 or 1.

3.2.3 Teacher-Student Learning

Training: With the combined probability distribution of entity labels from multiple teacher models, *i.e.*, $\tilde{p}(x'_i, \Theta_T)$ in Eq. 6, the training loss for the student model is identical to Eq. 3 and 4.

Inference: For inference on the target language, we only use the learned student model and make predictions as in the single-source scenario (Eq. 5).

Language	Type	Train	Dev	Test
English-en	Sentence	14,987	3,466	3,684
(CoNLL-2003)	Entity	23,499	5,942	5,648
German-de	Sentence	12,705	3,068	3,160
(CoNLL-2003)	Entity	11,851	4,833	3,673
Spanish-es	Sentence	8,323	1,915	1,517
(CoNLL-2002)	Entity	18,798	4,351	3,558
Dutch-nl	Sentence	15,806	2,895	5,195
(CoNLL-2002)	Entity	13,344	2,616	3,941

Table 1: Statistics of the benchmark datasets.

4 Experiments

We conduct extensive experiments for 3 target languages (*i.e.*, Spanish, Dutch, and German) on standard benchmark datasets, to validate the effectiveness and reasonableness of our proposed method for single- and multi-source cross lingual NER.

4.1 Settings

Datasets We use two NER benchmark datasets: CoNLL-2002 (Spanish and Dutch) (Tjong Kim Sang, 2002); CoNLL-2003 (English and German) (Tjong Kim Sang and De Meulder, 2003). Both are annotated with 4 entity types: `PER`, `LOC`, `ORG`, and `MISC`. Each language-specific dataset is split into training, development, and test sets. Table 1 reports the dataset statistics. All sentences are tokenized into sequences of subwords with WordPiece (Wu et al., 2016). Following Wu and Dredze (2019), we also use the BIO entity labelling scheme.

In our experiments, for each *source* language, an NER model is trained *previously* with its corresponding labeled training set. As for the *target* language, we discard the entity labels from its training set, and use it as unlabeled target-language data D_{tgt}. Similarly, unlabeled source-language data for learning language similarities (Eq. 10) is simulated via discarding the entity labels of each training set.

Network Configurations We leverage the cased multilingual $\text{BERT}_{\text{BASE}}$ (Wu and Dredze, 2019) for both $f(\cdot)$ in Eq. 1 and $g(\cdot)$ in Eq. 8, with 12 Transformer blocks, 768 hidden units, 12 self-attention head, GELU activations (Hendrycks and Gimpel, 2016), and learned positional embeddings. We use the final hidden vector of the first `[CLS]` token as the sentence embedding for $g(\cdot)$, and use the mean value of sentence embeddings *w.r.t* the k-th source language to initialize $\mu^{(k)}$ in Eq. 8.

	es	nl	de
Täckström et al. (2012)	59.30	58.40	40.40
Tsai et al. (2016)	60.55	61.56	48.12
Ni et al. (2017)	65.10	65.40	58.50
Mayhew et al. (2017)	65.95	66.50	59.11
Xie et al. (2018)	72.37	71.25	57.76
Wu and Dredze (2019)†	74.50	79.50	71.10
Moon et al. (2019)†	75.67	80.38	71.42
Wu et al. (2020)	76.75	80.44	73.16
Ours	**76.94**	**80.89**	**73.22**

Table 2: Performance comparisons of **single-source** cross-lingual NER. † denotes the reported results *w.r.t.* freezing the bottom three layers of $\text{BERT}_{\text{BASE}}$ as in this paper.

Network Training We implement our proposed method based on *huggingface* Transformers[1]. Following Wolf et al. (2019), we use a batch size of 32, and 3 training epochs to ensure convergence of optimization. Following Wu and Dredze (2019), we freeze the parameters of the embedding layer and the bottom three layers of $\text{BERT}_{\text{BASE}}$. For the optimizers, we use AdamW (Loshchilov and Hutter, 2017) with learning rate of $5e-5$ for teacher models (Wolf et al., 2019), and $1e-4$ for the student model (Yang et al., 2019) to converge faster. As for language similarity measuring (*i.e.*, Eq. 10), we set $\gamma = 0.01$ following Pinheiro (2018). Besides, we use a low-rank approximation for the *bilinear* operator M, *i.e.*, $M = U^T V$ where $U, V \in \mathbb{R}^{d \times m}$ with $d \ll m$, and we empirically set $d = 64$.

Performance Metric We use phrase level F1-score as the evaluation metric, following Tjong Kim Sang (2002). For each experiment, we conduct 5 runs and report the average F1-score.

4.2 Performance Comparison

Single-Source Cross-Lingual NER Table 2 reports the results of different single-source cross-lingual NER methods. All results are obtained with English as the source language and others as target languages.

It can be seen that our proposed method outperforms the previous state-of-the-art methods. Particularly, compared with the remarkable Wu and Dredze (2019) and Moon et al. (2019), which use nearly the same NER model as our method but is based on direct model transfer, our method obtains significant and consistent improvements in

[1]https://github.com/huggingface/transformers

	es	nl	de
Täckström (2012)	61.90	59.90	36.40
Rahimi et al. (2019)	71.80	67.60	59.10
Chen et al. (2019)	73.50	72.40	56.00
Moon et al. (2019)†	76.53	**83.35**	72.44
Ours-avg	77.75	80.70	74.97
Ours-sim	**78.00**	81.33	**75.33**

Table 3: Performance comparisons of **multi-source** cross-lingual NER. **Ours-avg**: averaging teacher models (Eq. 7) . **Ours-sim**: weighting teacher models with learned language similarities (Eq. 11). † denotes the reported results *w.r.t.* freezing the bottom three layers of $\text{BERT}_{\text{BASE}}$.

	es	nl	de
Single-source:			
Ours	**76.94**	**80.89**	**73.22**
HL	76.60 (-0.34)	80.43 (-0.46)	72.98 (-0.24)
MT	75.60 (-1.34)	79.99 (-0.90)	71.76 (-1.46)
Multi-source:			
Ours-avg	**77.75**	**80.70**	**74.97**
HL-avg	77.65 (-0.10)	80.39 (-0.31)	74.31 (-0.66)
MT-avg	77.25 (-0.50)	80.53 (-0.17)	74.18 (-0.79)
Ours-sim	**78.00**	**81.33**	**75.33**
HL-sim	77.81 (-0.19)	80.27 (-1.06)	74.63 (-0.70)
MT-sim	77.12 (-0.88)	80.24 (-1.09)	74.33 (-1.00)

Table 4: Ablation study of the proposed teacher-student learning method for cross-lingual NER. **HL**: Hard Label; **MT**: Direct Model Transfer; ***-avg**: averaging source-language models; ***-sim**: weighting source-language models with learned language similarities.

F1-scores, ranging from 0.51 for Dutch to 1.80 for German. That well demonstrates the benefits of teacher-student learning over unlabeled target-language data, compared to direct model transfer. Moreover, compared with the latest meta-learning based method (Wu et al., 2020), our method requires much lower computational costs for both training and inference, meanwhile reaching superior performance.

Multi-Source Cross-Lingual NER Here we select source languages in a leave-one-out manner, *i.e.*, all languages except the target one are regarded as source languages. For fair comparisons, we take Spanish, Dutch, and German as target languages, respectively.

Table 3 reports the results of different methods for multi-source cross-lingual NER. Both our teacher-student learning methods, *i.e.*, *Ours-avg* (averaging teacher models, Eq. 7) and *Ours-sim* (weighting teacher models with learned language similarities, Eq. 11), outperform previous state-of-the-art methods on Spanish and German by a large margin, which well demonstrates their effectiveness. We attribute the large performance gain to the teacher-student learning process to further leverage helpful information from unlabeled data in the target language. Though Moon et al. (2019) achieves superior performance on Dutch, it is not applicable in cases where the labeled source-language data is inaccessible, and thus it still suffers from the aforementioned limitation *w.r.t.* data availability.

Moreover, compared with *Ours-avg*, *Ours-sim* brings consistent performance improvements. That means, if unlabeled data in source languages is available, using our proposed language similarity measuring method for weighting different teacher models can be superior to simply averaging them.

4.3 Ablation Study

Analyses on Teacher-Student Learning To validate the reasonableness of our proposed teacher-student learning method for cross-lingual NER, we introduce the following baselines. 1) *Hard Label (HL)*, which rounds the probability distribution of entity labels (*i.e.*, soft labels output by teacher models) into a one-hot labelling vector (*i.e.*, hard labels) to guide the learning of the student model. Note that in multi-source cases, we use the combined probability distribution of multiple teacher models (Eq. 6) to derive the hard labels. To be consistent with Eq. 3, we still adopt the MSE loss here. In fact, both MSE loss and cross-entropy loss lead to the same observation described in this subsection. 2) *Direct Model Transfer (MT)*, where NO unlabeled target-language data is available to perform teacher-student learning, and thus it degenerates into: a) directly applying the source-language model in single-source cases, or b) directly applying a weighted ensemble of source-language models in multi-source cases, with weights derived via Eq. 6 and Eq. 11.

Table 4 reports the ablation study results. It can be seen that using hard labels (*i.e.*, HL-*) would result in consistent performance drops in all cross-lingual NER settings, which validates using soft labels in our proposed teacher-student learning method can convey more information for knowledge transfer than hard labels. Moreover, we can also observe that, using direct model transfer (*i.e.*,

#1 Spanish	**Source-Language Model:** ...Etchart [I-PER, 1.00] Sydney [B-LOC, 0.98] (Australia [B-LOC, 1.00]) , 23 may (EFE [O, 0.53]) . **Ours:** Por Mario [B-PER] Etchart [I-PER] Sydney [B-LOC] (Australia [B-LOC]) , 23 may (EFE [B-ORG]) . **Examples in D_{tgt}:** Asi lo anunció a EFE [B-ORG, 1.00] Hans Gaasbek, el abogado de Murillo, argumentando que ...
#2 Dutch	**Source-Language Model:** Vanderpoorten [O, 0.87] : ' Dit is een eerste stap in de herwaardering van het beroepsonderwijs " **Ours:** Vanderpoorten [B-PER] : ' Dit is een eerste stap in de herwaardering van het beroepsonderwijs " **Examples in D_{tgt}:** Vanderpoorten [B-PER, 0.99] stond op het punt die reputatie te bezwadderen.
#3 German	**Source-Language Model:** ... dabei berücksichtigt werden müsse , forderte Hof [B-ORG, 0.85] eine " Transparenz " ... **Ours:** Weil die Altersstruktur dabei berücksichtigt werden müsse , forderte Hof [B-PER] eine " Transparenz " ... **Examples in D_{tgt}:** ... meint Hof [B-PER, 0.99] , den der " erstaunliche Pragmatismus der Jugendlichen " beeindruckt.

Figure 4: Case study on why teacher-student learning works. The GREEN (RED) highlight indicates a correct (incorrect) label. The real-valued numbers indicate the predicted probability corresponding to the entity label.

	es	nl	de
Ours	**78.00**	**81.33**	**75.33**
cosine	77.86 (-0.14)	79.94 (-1.39)	75.24 (-0.09)
ℓ_2	77.72 (-0.28)	79.74 (-1.59)	75.09 (-0.24)

Table 5: Comparison between the proposed language similarity measuring method and the commonly used *cosine*/ℓ_2 metrics for multi-source cross-lingual NER.

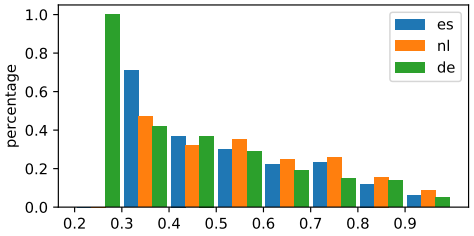

Figure 5: Percentage of corrected mispredictions, in different probability intervals.

MT-*) would lead to even more significant performance drops in all cross-lingual NER settings (up to 1.46 F1-score). Both demonstrate that leveraging unlabeled data in the target language can be helpful, and that the proposed teacher-student learning method is capable of leveraging such information effectively for cross-lingual NER.

Analyses on Language Similarity Measuring We further compare the proposed language similarity measuring method with other commonly used unsupervised metrics, *i.e.*, *cosine* similarity and ℓ_2 distance. Specifically, $s(\boldsymbol{x}', \mu^{(k)})$ in Eq. 11 is replaced by *cosine* similarity or negative ℓ_2 distance between $\boldsymbol{x}'$ and the mean value of sentence embeddings *w.r.t* the k-th source language.

As shown in Table 5, replacing the proposed language similarity measuring method with either *cosine* / ℓ_2 metrics leads to consistent performance drops across all target languages. This further demonstrates the benefits of our language identification based similarity measuring method.

4.4 Why Teacher-Student Learning Works?

By analyzing which failed cases of *directly applying the source-language model* are corrected by the proposed teacher-student learning method, we try to bring up insights on why teacher-student learning works, in the case of single-source cross-lingual NER.

Firstly, teacher-student learning can probably help to learn label preferences for some specific words in the target language. Specifically, if a word appears in the unlabeled target-language data and the teacher model consistently predicts it to be associated with an identical label with high probabilities, the student model would learn the preferred label *w.r.t* that word, and predict it in cases where the sentence context may not provide enough information. Such label preference can help the predictions for tokens that are less ambiguous and generally associated with an identical entity label. As illustrated in Figure 4, in example #1, the source-language (teacher) model, fails to identify "EFE" as an ORG in the test sentences, while the student model (*i.e.*, Ours) can correctly label it, because it has seen "EFE" labeled as ORG by the teacher model with high probabilities in the unlabeled target-language data D_{tgt}. Similar results can also be observed in example #2 and #3.

Moreover, teacher-student learning may help to find a better classifying hyperplane for the student NER model with unlabelled target-language data. Actually, we notice that the source-language model generally makes correct label predictions with higher probabilities, and makes mispredictions with relatively lower probabilities. By calcu-

lating the proportion of its mispredictions that are corrected by our teacher-student learning method in different probability intervals, we find that our method tends to correct the low-confidence mispredictions, as illustrated in Figure 5. We conjecture that, with the help of unlabeled target-language data, our method can probably find a better classifying hyperplane for the student model, so that the low-confidence mispredictions, which are closer to the classifying hyperplane of the source-language model, can be clarified.

5 Conclusion

In this paper, we propose a teacher-student learning method for single-/multi-source cross-lingual NER, via using source-language models as teachers to train a student model on unlabeled data in the target language. The proposed method does not rely on labelled data in the source languages and is capable of leveraging extra information in the unlabelled target-language data, which addresses the limitations of previous label-projection based and model-transfer based methods. We also propose a language similarity measuring method based on language identification, to better weight different teacher models. Extensive experiments on benchmark datasets show that our method outperforms the existing state-of-the-art approaches.

References

Cristian Bucilu, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 535–541. ACM.

Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. 2019. Multi-source cross-lingual model transfer: Learning what to share. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3098–3112.

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

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*, pages 4171–4186.

Jan Vium Enghoff, Søren Harrison, and Željko Agić. 2018. Low-resource named entity recognition via multi-source projection: Not quite there yet? In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 195–201.

Dan Hendrycks and Kevin Gimpel. 2016. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *CoRR*, abs/1606.08415.

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

Lifu Huang, Heng Ji, and Jonathan May. 2019. Cross-lingual multi-level adversarial transfer to enhance low-resource name tagging. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3823–3833, Minneapolis, Minnesota.

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

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*.

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.

David McClosky, Eugene Charniak, and Mark Johnson. 2010. Automatic domain adaptation for parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 28–36.

Taesun Moon, Parul Awasthy, Jian Ni, and Radu Florian. 2019. Towards lingua franca named entity recognition with bert. *arXiv preprint arXiv:1912.01389*.

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*, pages 1470–1480.

Shuke Peng, Xinjing Huang, Zehao Lin, Feng Ji, Haiqing Chen, and Yin Zhang. 2019. Teacher-student framework enhanced multi-domain dialogue generation. *arXiv preprint arXiv:1908.07137*.

Matthew Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1756–1765.

Pedro O Pinheiro. 2018. Unsupervised domain adaptation with similarity learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8004–8013.

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

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Oscar Täckström. 2012. Nudging the envelope of direct transfer methods for multilingual named entity recognition. In *Proceedings of the NAACL-HLT Workshop on the Induction of Linguistic Structure*, pages 55–63.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 477–487.

Xu Tan, Yi Ren, Di He, Tao Qin, and Tie-Yan Liu. 2019. Multilingual neural machine translation with knowledge distillation. In *International Conference on Learning Representations*.

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

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

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

Mengqiu Wang and Christopher D. Manning. 2014. Cross-lingual projected expectation regularization for weakly supervised learning. *Transactions of the Association for Computational Linguistics*, 2:55–66.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

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

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.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

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.

Ze Yang, Linjun Shou, Ming Gong, Wutao Lin, and Daxin Jiang. 2019. Model compression with two-stage multi-teacher knowledge distillation for web question answering system. *arXiv preprint arXiv:1910.08381*.

Shan You, Chang Xu, Chao Xu, and Dacheng Tao. 2017. Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1285–1294. ACM.

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.