UniTrans : Unifying Model Transfer and Data Transfer for Cross-Lingual Named Entity Recognition with Unlabeled Data

Qianhui Wu[1] , **Zijia Lin**[2] , **Börje F. Karlsson**[2] , **Biqing Huang**[1] and **Jian-Guang Lou**[2]

[1]Beijing National Research Center for Information Science and Technology (BNRist)
Department of Automation, Tsinghua University, Beijing 100084, China
[2]Microsoft Research, Beijing 100080, China
wuqianhui@tsinghua.org.cn, {zijlin,borje.karlsson,jlou}@microsoft.com, hbq@tsinghua.edu.cn

Abstract

Prior works in cross-lingual named entity recognition (NER) with no/little labeled data fall into two primary categories: model transfer based and data transfer based methods. In this paper we find that both method types can complement each other, in the sense that, the former can exploit context information via language-independent features but sees no task-specific information in the target language; while the latter generally generates pseudo target-language training data via translation but its exploitation of context information is weakened by inaccurate translations. Moreover, prior works rarely leverage unlabeled data in the target language, which can be effortlessly collected and potentially contains valuable information for improved results. To handle both problems, we propose a novel approach termed UniTrans to Unify both model and data Transfer for cross-lingual NER, and furthermore, to leverage the available information from unlabeled target-language data via enhanced knowledge distillation. We evaluate our proposed UniTrans over 4 target languages on benchmark datasets. Our experimental results show that it substantially outperforms the existing state-of-the-art methods.

1 Introduction

Named entity recognition (NER) is a fundamental task in natural language processing, which seeks to locate and classify named entities, like locations, organizations, etc., in unstructured texts. NER has been extensively studied, especially monolingual NER, as it is widely incorporated in various downstream tasks, *e.g.*, question answering [Mollá *et al.*, 2006]. Recently, deep neural networks have become the dominant approach due to their superior performance. However, the key to their success is the availability of adequate labeled training data; and building large labeled training sets for a new language of interest can be time consuming and labor costly. This motivates researches on *cross-lingual transfer*, which leverages labeled data from a source language (*e.g.*, English) to overcome the data scarcity issue in the target language. In this paper, following [Wu and Dredze, 2019] and

[Wu *et al.*, 2020], we focus on *zero-resource cross-lingual transfer*, where *NO labeled data* is available in the target language. In this way, our work can act as a basis for further researches where some labeled target-language data exists.

The state-of-the-art methods for zero-resource cross-lingual NER mainly fall into two categories: i) *model transfer based methods* [Wu and Dredze, 2019; Wu *et al.*, 2020], which use labeled source-language data to train an NER model with language-independent features (*e.g.*, cross-lingual word representations [Devlin *et al.*, 2019]), and then directly apply it to the target language; ii) *data transfer based methods*, which generally construct pseudo labeled data in the target language via translating the source-language training data into parallel target-language data and mapping the entity labels, and then use such pseudo labeled data to train a target-language NER model. For example, [Mayhew *et al.*, 2017] and [Xie *et al.*, 2018] employed word-to-word and phrase-to-phrase translation, respectively, to generate target-language training data, so that entity labels in the source-language training data can be directly copied to the generated target-language data.

In this paper, we hold the idea that the aforementioned model transfer based methods and data transfer based methods are complementary to each other. Specifically, for the former, models trained with labeled source-language data only learn the knowledge *w.r.t.* the task in the source language, but cannot see any task-specific information in the target language. And thus they probably rely more on context information mined by language independent features to make predictions, as word-label relations in the target language are unavailable for them. As for the latter, though models have access to word-label pairs in the target language to exploit their relations, inaccurate translations caused by sense ambiguity and word order differences will probably weaken the models' capability to predict through context information. Moreover, both methods generally do not leverage the language information contained in the unlabeled target-language data to benefit the cross-lingual transfer.

Therefore, here we propose a novel approach termed UniTrans to unify both model transfer and data transfer for cross-lingual NER, and furthermore, to leverage the beneficial information from unlabeled target-language data via enhanced knowledge distillation. Specifically, following [Lample *et al.*, 2018], we first construct a pseudo training set

for the target language by performing word-to-word translation and copying entity labels. Then, we use a pre-trained cross-lingual language model, *i.e.*, multilingual BERT [Devlin *et al.*, 2019], to generate language-independent features to train *two* separate NER models. One is trained on the labeled source-language data, and the other is trained on the pseudo target-language training data. We then integrate the knowledge contained in the source-language training data, the pseudo target-language training data, and the unlabeled target-language data as follows: i) we fine-tune the model trained on source-language data with the pseudo target-language training data, and take the fine-tuned model as a teacher NER model to predict probability distributions of entity labels (*i.e.*, soft labels) for each word in the unlabeled target-language data; ii) we propose a voting scheme that associates the three aforementioned NER models to predict high-confidence one-hot label vectors (*i.e.*, hard labels) for part of words in the unlabeled target-language data; iii) we train a student NER model on the unlabeled target-language data with supervision from both the aforementioned soft labels and hard labels, as how knowledge distillation works, and we use it as the final target-language NER model.

Extensive experiments conducted on benchmark datasets for 4 target languages well demonstrate that the proposed UniTrans substantially outperforms the existing state-of-the-art methods. We also extend UniTrans by ensembling multiple teacher models with different random seeds, and show further performance improvement.

Our major contributions are summarized as follows:

- We propose a novel approach termed UniTrans to unify model transfer and data transfer based on their complementarity for cross-lingual NER, with the help of beneficial information from unlabeled target-language texts.

- We propose a voting scheme to generate pseudo hard labels on unlabeled target-language data, so as to enhance knowledge distillation in UniTrans with supervision from both soft labels and pseudo hard labels.

- We conduct extensive experiments on benchmark datasets and show that UniTrans yields new state-of-the-art cross-lingual NER performance, which can even be promoted via teacher ensembling.

2 Related Work

2.1 Cross-Lingual NER

Cross-lingual transfer for NER has attracted much attention in recent years. Prior works roughly fall into two primary categories: model transfer based and data transfer based.

Model transfer based methods generally utilize language-independent features to train an NER model on the labeled source-language data, so that the model can be directly applied to the target language. Those language-independent features include aligned word representations [Ni *et al.*, 2017; Wu and Dredze, 2019], word clusters [Täckström *et al.*, 2012], Wikifier features [Tsai *et al.*, 2016], and gazetteers [Zirikly and Hagiwara, 2015], *etc*. Besides directly applying the trained NER model, [Wu *et al.*, 2020] further

proposed a meta-learning algorithm to improve the prediction for each test case with its similar examples in the labeled source-language data.

Data-transfer based methods generally train a monolingual NER model for the target language with pseudo training data constructed from the labeled source-language data. [Ni *et al.*, 2017] proposed to use bilingual parallel texts and their word alignment information to project labels from the source language to the target language. Considering it is expensive to build parallel corpora manually, most recent methods propose to translate the source-language texts into the target language in a word-by-word [Xie *et al.*, 2018] or phrase-by-phrase [Mayhew *et al.*, 2017] manner, and then copy the label of each word/phrase to its corresponding translated word/phrase. [Jain *et al.*, 2019] further proposed to utilize Google Translate twice to translate sentences as well as entities, and align entity labels based on distributional statistics derived from the dataset.

As mentioned above, both model transfer and data transfer can be complementary to each other. And thus in this paper, we propose UniTrans to unify both, so as to retain the capability of predicting through context information and meanwhile exploiting word-label relations in the target language. Additionally, few prior works on cross-lingual NER leveraged unlabeled data in the target language. [Bari *et al.*, 2019] proposed to fine-tune the model trained on source-language data with unlabeled target-language data in a manner similar to self-training [Scudder, 1965]. As shown by [He and Sun, 2017] and [Bari *et al.*, 2019] that unlabeled data is beneficial, our UniTrans also exploits the unlabeled target-language data via an enhanced knowledge distillation process, as mentioned before.

2.2 Knowledge Distillation

Knowledge distillation was originally proposed for model compression [Bucilu *et al.*, 2006], *i.e.*, to learn a compact student model that retains most performance of a larger teacher model or ensemble of models that require more space to deploy or more computation to make predictions [Rusu *et al.*, 2015; Hinton *et al.*, 2015; Sanh *et al.*, 2019]. Besides model compression, researchers also applied knowledge distillation to various tasks, like deep reinforcement learning [Rusu *et al.*, 2015], image classification [Hinton *et al.*, 2015], and language modeling [Sanh *et al.*, 2019].

In this paper, we adapt knowledge distillation to cross-lingual NER. And different from typical application scenarios of knowledge distillation that do not consider unlabeled data, in our proposed UniTrans, the student model is completely trained on the unlabeled target-language data. In addition, we propose a voting scheme to generate pseudo hard labels for the unlabeled target-language data, thus enhancing knowledge distillation with supervision from both soft labels and pseudo hard labels.

3 Methodology

Figure 1 illustrates the framework of the proposed UniTrans. Specifically, we first generate pseudo training data for the target language by performing word-to-word translation and
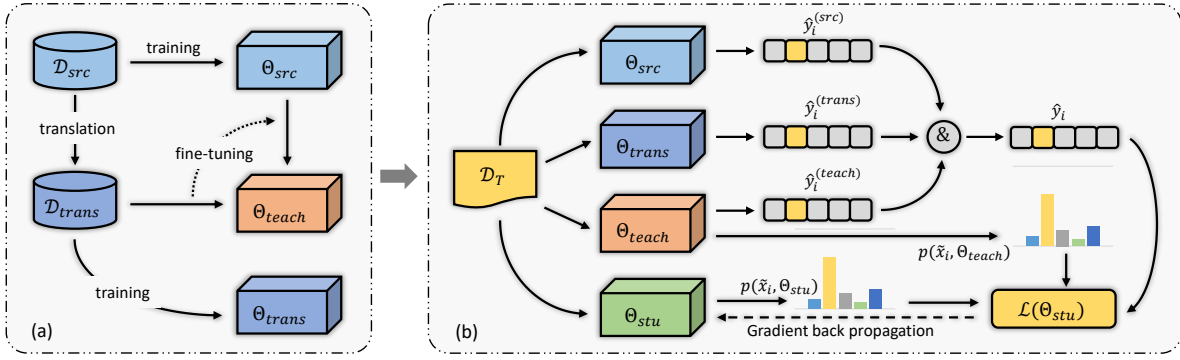
Figure 1: Framework of the proposed UniTrans. (a) Unifying model transfer and data transfer. (b) Knowledge distillation on unlabeled data.

copying entity labels. Then, we train an NER model using the labeled source-language data and derive a teacher model by fine-tuning it with the pseudo target-language training data. Finally, we adapt knowledge distillation to train a student NER model for the target language on unlabelled data.

3.1 Data Transfer via Word-to-Word Translation

Following [Xie *et al.*, 2018], we apply [Lample *et al.*, 2018] to translate the source-language training data word-by-word into the target language, and then *directly copy the entity label of each source-language word to its corresponding translated word*. The pipeline of data transfer is briefly introduced below. One can refer to [Lample *et al.*, 2018] for more details.

We first leverage identical "character strings" [Smith *et al.*, 2017] in both languages, *i.e.*, shared words in most cases, to create a seed dictionary $\{s_i, t_i\}_{i=1}^D$, where s_i, $t_i \in \mathbb{R}^d$, respectively, denote the source-language word embedding and the target-language word embedding of an identical character string, and D is the size of the dictionary. Then, we learn a linear mapping $P \in \mathbb{R}^{d \times d}$ between the source-language embedding matrix $S = [s_1, s_2, ..., s_D] \in \mathbb{R}^{d \times D}$ and the target-language embedding matrix $T = [t_1, t_2, ..., t_D] \in \mathbb{R}^{d \times D}$, with its objective function formulated as follows.

$$P = \arg\min_{P'} \|P'S - T\|_F \quad s.t., \ P'^T P' = I \quad (1)$$

where $\|\cdot\|_F$ denotes the Frobenius norm. The closed-form solution for Eq. 1 can be derived via singular value decomposition (SVD):

$$P = UV^T \quad s.t., \ U\Sigma V^T = \text{SVD}(TS^T) \quad (2)$$

To produce word-to-word translations, for a source-language word, we use the learned P to map its embedding vector into the target-language embedding space, and take its nearest neighbor in the target language as its corresponding translation result. Specifically, we use the cross-domain similarity local scaling (CSLS) [Lample *et al.*, 2018] to measure the distance between the mapped embedding vector of a source-language word (denoted as Ps_i) and the embedding vector of a target-language word (denoted as t_j):

$$\text{CSLS}(Ps_i, t_j) = 2\cos(Ps_i, t_j) - r_T(Ps_i) - r_S(t_j) \quad (3)$$

$$r_T(Ps_i) = \frac{1}{K} \sum_{t_k \in \mathcal{N}_T(Ps_i)} \cos(Ps_i, t_k) \quad (4)$$

$$r_S(t_j) = \frac{1}{K} \sum_{Ps_k \in \mathcal{N}_S(t_j)} \cos(Ps_k, t_j) \quad (5)$$

where $\mathcal{N}_T(Ps_i)$ denotes the K target-language nearest neighbors of Ps_i, $\mathcal{N}_S(t_j)$ denotes the K mapped source-language nearest neighbors of t_j, and $\cos(\cdot)$ denotes cosine similarity.

By word-by-word translation and copying word labels, we can build a pseudo target-language training set $\mathcal{D}_{trans}$.

3.2 Base Model for NER

The base model for NER in this paper consists of a feature encoder and a linear classification layer. Given an input text sequence $\boldsymbol{x} = \{x_i\}_{i=1}^N$ with N words, we first feed it into the feature encoder f_θ to obtain feature vectors $\boldsymbol{h} = \{h_i\}_{i=1}^N$ for all words:

$$\boldsymbol{h} = f_\theta(\boldsymbol{x}) \quad (6)$$

where f_θ can be any feature encoder that produces language-independent features, and h_i is the feature vector corresponding to the i-th word x_i. Following [Wu and Dredze, 2019], here we utilize multilingual BERT [Devlin *et al.*, 2019] as the language-independent feature encoder.

Then for each word x_i, its corresponding feature vector h_i is fed into the linear classification layer with the *softmax* function to predict the probability distribution of entity labels for it, which is formulated as follows.

$$p(x_i, \Theta) = \text{softmax}(Wh_i + b) \quad (7)$$

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

Suppose a labeled training set is denoted as $\mathcal{D} = \{(\boldsymbol{x}, \boldsymbol{y})\}$, where $\boldsymbol{y} = \{y_i\}_{i=1}^N$ is the corresponding one-hot entity label vectors for words in the corresponding $\boldsymbol{x}$. Then the loss function $\mathcal{L}$ for the NER model is defined as the cross entropy between the predicted probability distribution of entity labels and the ground-truth one for each word, as formulated below:

$$\mathcal{L}(\Theta) = \frac{1}{|\mathcal{D}|} \sum_{(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D}} \frac{1}{N} \sum_{i=1}^N \text{CrossEntropy}(y_i, p(x_i, \Theta))$$

$$(8)$$

where y_i is the one-hot label vector corresponding to x_i.

Note that in this paper, we use the base model here to train all mentioned NER models.

3.3 Unifying Model Transfer and Data Transfer

With the base model and loss function defined above, we can train an NER model on the labeled source-language data like previous model transfer based methods [Wu and Dredze, 2019], denoted as Θ_{src}. Then for unifying model transfer and data transfer based on their complementarity, the proposed UniTrans further trains a teacher model Θ_{teach} by fine-tuning Θ_{src} on the pseudo target-language training data, with the same loss function as Eq. 8. And thus Θ_{teach} is expected to combine the advantages of both model transfer and data transfer. Moreover, our experiments also show that, Θ_{teach} obtains superior performance than simply combining the labeled source-language data and the pseudo target-language training data to train an NER model.

3.4 Knowledge Distillation on Unlabeled Data

Using Soft Labels

For knowledge distillation, we first apply the teacher model Θ_{teach} to the unlabeled target-language data $\mathcal{D}_T = \{\tilde{\boldsymbol{x}}\}$, where $\tilde{\boldsymbol{x}} = \{\tilde{x}_i\}_i^N$ denotes an unlabeled target-language text with N words. Considering that predicted soft labels (*i.e.*, probability distribution of entity labels) can contain richer information than predicted hard labels [Hinton *et al.*, 2015], we use the soft labels output by Θ_{teach} as the supervision to train a student model Θ_{stu}, by minimizing the *mean squared error* (MSE) between the prediction of Θ_{stu} and that of Θ_{teach} for each word. The loss function *w.r.t.* an unlabelled target-language text $\tilde{\boldsymbol{x}}$ is formulated as:

$$\mathcal{L}_{soft}^{\tilde{\boldsymbol{x}}} = \frac{1}{N} \sum_{i=1}^{N} \text{MSE}\left(p(\tilde{x}_i, \Theta_{teach}), p(\tilde{x}_i, \Theta_{stu})\right) \quad (9)$$

where $p(\tilde{x}_i, \Theta_{stu})$ denotes the probability distribution of entity labels predicted by the student model for the i-th word $\tilde{x}_i$, and $p(\tilde{x}_i, \Theta_{teach})$ denotes that of the teacher model.

In the way above, we can not only transfer the knowledge learned by the teacher model Θ_{teach} to the student model Θ_{stu}, but also enable the student model Θ_{stu} to capture language specific information in unlabeled target-language data $\mathcal{D}_T$. And thus, with an identical base model, the student model Θ_{stu} is expected to gain performance improvement over the teacher model Θ_{teach}.

Incorporating Pseudo Hard Labels

[Hinton *et al.*, 2015] pointed out that when correct hard labels are available, the student model trained with only soft labels can be further improved by also training it with supervision from the correct hard labels. However, in the case of zero-resource cross-lingual NER, there are no ground-truth entity labels available in $\mathcal{D}_T$. To this end, we propose a voting scheme to (partially) generate pseudo hard labels for the unlabeled target-language data. And we leverage them to help the training of the student model, via adding an extra loss to Eq. 9.

Specifically, we train another NER model in a supervised manner on the pseudo target-language training data generated by data transfer (section 3.1), denoted as Θ_{trans}. Then, we predict the probability distributions of entity labels for each

word $\tilde{x}_i$ in any unlabeled text $\tilde{\boldsymbol{x}}$, using all learned NER models, *i.e.*, Θ_{src}, Θ_{teach} and Θ_{trans}, respectively. And we take the entity label $c \in C$ with the highest probability as the predicted label $\hat{y}_i^{(*)}$ for $\tilde{x}_i$:

$$\hat{y}_i^{(*)} = \arg\max_c p(\tilde{x}_i, \Theta_*)_c \quad (10)$$

where Θ_* stands for Θ_{src}, Θ_{teach} and Θ_{trans}, with the corresponding $\hat{y}_i^{(*)}$ being $\hat{y}_i^{(src)}$, $\hat{y}_i^{(teach)}$ and $\hat{y}_i^{(trans)}$. And we only generate a pseudo hard label $\hat{y}_i = \hat{y}_i^{(teach)}$ for the word $\tilde{x}_i$ when $\hat{y}_i^{(teach)} = \hat{y}_i^{(src)} = \hat{y}_i^{(trans)}$, as $\hat{y}_i$ would be of high-confidence then. Here $\hat{y}_i$ can be seen as the voting result of Θ_{src}, Θ_{teach} and Θ_{trans}. We denote the set of such words with a pseudo hard label as $\mathcal{X}$:

$$\mathcal{X} = \{\tilde{x}_i | \hat{y}_i^{(teach)} = \hat{y}_i^{(src)} = \hat{y}_i^{(trans)}, \, \forall \tilde{x}_i \in \mathcal{D}_T\} \quad (11)$$

With the pseudo hard labels derived, we further define a hard label based loss $\mathcal{L}_{hard}^{\tilde{\boldsymbol{x}}}$ as follows for the unlabeled target-language text $\tilde{\boldsymbol{x}}$.

$$\mathcal{L}_{hard}^{\tilde{\boldsymbol{x}}} = \frac{1}{N} \sum_{i=1}^{N} I(\tilde{x}_i) \cdot \text{CrossEntropy}(\hat{y}_i, p(\tilde{x}_i, \Theta_{stu})) \quad (12)$$

where $I(\tilde{x}_i)$ is an indicator function that returns 1 if $\tilde{x}_i \in \mathcal{X}$ and otherwise 0.

Then the loss function combining both the supervision from pseudo hard labels and that from the soft labels output by Θ_{teach} is formulated as follows to better guide the training of Θ_{stu}:

$$\mathcal{L}(\Theta_{stu}) = \frac{1}{|\mathcal{D}_T|} \sum_{\tilde{\boldsymbol{x}} \in \mathcal{D}_T} \left(\eta \mathcal{L}_{hard}^{\tilde{\boldsymbol{x}}} + \mathcal{L}_{soft}^{\tilde{\boldsymbol{x}}}\right) \quad (13)$$

where $\eta > 0$ is a trade-off parameter, and we simply set $\eta = 1$ in this paper. With $\mathcal{L}(\Theta_{stu})$, the student model Θ_{stu} is learned in an end-to-end manner.

3.5 Inference in Target Language

For inference in the target language, we only use the learned student model Θ_{stu} to predict the probability distribution of entity labels for each word in a given test case, as Eq. 7. Note that here we do not straightly take the entity label with the highest probability as the prediction result for each word (Eq. 10). Instead, we apply *Viterbi* decoding [Chen *et al.*, 2019] by just adding constraints to ensure that predicted entity labels for all words in the test case would not violate the NER tagging scheme and meanwhile would obtain the highest probability as a label sequence. And thus we don't need to train a transition matrix here.

4 Experiments

We evaluate the proposed UniTrans for zero-resource cross-lingual NER through experiments over benchmark datasets on 4 target languages (*i.e.*, Spanish, Dutch, German, and Norwegian[1]) and comparisons with state-of-the-art methods.

---

[1] We use Bokmål rather than Nynorsk here, considering that it is used by 85–90% of the population of Norway.

Language	Type	Train	Dev	Test
English [en]	# of Sentence	14,987	3,466	3,684
(CoNLL-2003)	# of Entity	23,499	5,942	5,648
German [de]	# of Sentence	12,705	3,068	3,160
(CoNLL-2003)	# of Entity	11,851	4,833	3,673
Spanish [es]	# of Sentence	8,323	1,915	1,517
(CoNLL-2002)	# of Entity	18,798	4,351	3,558
Dutch [nl]	# of Sentence	15,806	2,895	5,195
(CoNLL-2002)	# of Entity	13,344	2,616	3,941
Norwegian [no]	# of Sentence	15,686	2,410	1,939
(NoDaLiDa-2019)	# of Entity	10,934	1,615	1,391

Table 1: Statistics of the benchmark datasets.

4.1 Experiment Settings

Datasets

We use the following widely-used benchmark datasets for experiments: CoNLL-2002 [Tjong Kim Sang, 2002] for Spanish [es] and Dutch [nl] NER, CoNLL-2003 [Tjong Kim Sang and De Meulder, 2003] for English [en] and German [de] NER, and NoDaLiDa-2019 [Johansen, 2019] for Norwegian [no] NER. All datasets are annotated with 4 entity types: LOC, MISC, ORG, and PER. Each dataset is split into training, dev, and test sets. Table 1 reports the statistics of each.

We leverage WordPiece [Wu et al., 2016] to tokenize each sentence into a sequence of subwords and, following [Wu and Dredze, 2019; Wu et al., 2020], we use the BIO entity labeling scheme. Moreover, as previous works [Wu et al., 2020], all experiments use English as the source language and the others as the target language.

Note that for each target language, we delete all entity labels in its training set, and use it only as unlabeled target-language data. Moreover, to imitate the zero-resource cross-lingual NER case, we ignore all target-language dev sets, and directly evaluate the learned models on their test sets.

Implementation Details

We implement our UniTrans with PyTorch[2]. For word-to-word translation, we use *fastText*[3] monolingual word embeddings, and use *MUSE*[4] [Lample et al., 2018] to perform translation. For the feature encoder of the base model, *i.e.*, f_θ in Eq. 6, we employ the pretrained multilingual BERT model (case-sensitive version) [Devlin et al., 2019] in *Hugging-Face's Transformers*[5], which has 12 Transformer blocks, 12 attention heads, and 768 hidden units.

We empirically set UniTrans hyper-parameters by following previous works (as cited below), and utilize them in all experiments. Specifically, we adopt a dropout rate of 0.1 [Wu et al., 2020] and freeze the parameters of the embedding layer and the bottom three layers of the multilingual BERT [Wu and Dredze, 2019]. Following [Wu et al., 2020], we train all models for 3 epochs using a batch size of 32, maximum sequence length of 128, and AdamW [Loshchilov and Hutter, 2017] as the optimizer. For AdamW, we use a learning rate

[2]https://pytorch.org/

[3]https://fasttext.cc/docs/en/pretrained-vectors.html

[4]https://github.com/facebookresearch/MUSE

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

	es	nl	de	no	Average
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]	64.10	63.37	57.23	-	-
Xie et al. [2018]	72.37	71.25	57.76	-	-
Jain et al. [2019]	73.5	69.9	61.5	-	-
Bari et al. [2019]	75.93	74.61	65.24	-	-
Wu and Dredze [2019][†]	74.50	79.50	71.10	-	-
Wu et al. [2020]	76.75	80.44	73.16	-	-
Model Transfer (reimp.)	76.34 ($\pm$ 0.96)	80.61 ($\pm$ 0.46)	72.39 ($\pm$ 1.05)	78.47 ($\pm$ 0.36)	76.95
Data Transfer (reimp.)	78.14 ($\pm$ 0.97)	80.98 ($\pm$ 0.72)	73.65 ($\pm$ 0.36)	78.91 ($\pm$ 0.50)	77.92
UniTrans	**79.31** ($\pm$ 0.39)	**82.90** ($\pm$ 0.43)	**74.82** ($\pm$ 0.60)	**81.17** ($\pm$ 0.63)	**79.55**

Table 2: Results of the proposed UniTrans and prior state-of-the-art methods for zero-resource cross-lingual NER. [†] denotes the reported results *w.r.t.* freezing the bottom 3 layers of BERT as in this paper. We also report the standard deviation for reimplemented baselines and the proposed UniTrans (*i.e.*, numbers in parentheses).

of $5e-5$ [Wolf et al., 2019] for teacher models and $1e-4$ for the student model [Yang et al., 2019]. Note that if a word is split into several subwords after tokenization, only the first subword is considered in the loss function.

Performance Metric

Following [Tjong Kim Sang, 2002], we use entity level F1-score as the performance metric. Moreover, we conduct each experiment 5 times and report the mean F1-score.

4.2 Experimental Results

Table 2 reports the zero-resource cross-lingual NER results of the proposed UniTrans on the 4 target languages, alongside those reported by prior state-of-the-art methods and those of two re-implemented baseline methods, *i.e.*, *Model Transfer* (Θ_{src} in 3.3) and *Data Transfer* (Θ_{trans} in 3.4).

Table 2 shows that our proposed UniTrans significantly outperforms the prior state-of-the-art methods and the re-implemented baselines on all target languages. Particularly, compared with the best prior method [Wu et al., 2020], the proposed UniTrans achieves an improvement of F1-score ranging from 1.66 for German [de] to 2.56 for Spanish [es]. Moreover, UniTrans achieves an average improvement of 1.63 F1-score over *Data Transfer*, and 2.60 F1-score over *Model Transfer*. All these results well demonstrate the effectiveness of the proposed UniTrans, which is mainly attributed to unifying model transfer with data transfer and leveraging unlabeled target-language data in UniTrans.

4.3 Ablation Study

To validate the contributions of different components in the proposed UniTrans, we introduce the following variants of UniTrans and baselines to perform ablation study: 1) *UniTrans w/o $\mathcal{L}_{soft}^{\tilde{x}}$*, which removes $\mathcal{L}_{soft}^{\tilde{x}}$ from Eq.13 to train the target-language NER model Θ_{stu} with only the pseudo hard labels; 2) *UniTrans w/o $\mathcal{L}_{hard}^{\tilde{x}}$*, which eliminates $\mathcal{L}_{hard}^{\tilde{x}}$ from Eq.13 (*i.e.*, $\eta = 0$) to train Θ_{stu} with only soft labels output by the teacher model Θ_{teach}; 3) *UniTrans*

	es	nl	de	no	Average
UniTrans	79.31	82.90	74.82	81.17	79.55
1) UniTrans w/o $\mathcal{L}_{soft}^{\tilde{x}}$	78.93 (-0.38)	82.48 (-0.42)	75.23 (0.41)	80.91 (-0.26)	79.39 (-0.16)
2) UniTrans w/o $\mathcal{L}_{hard}^{\tilde{x}}$	79.54 (0.23)	82.78 (-0.12)	74.43 (-0.39)	81.06 (-0.11)	79.45 (-0.10)
3) UniTrans w/ Θ_{src}	77.30 (-2.01)	81.20 (-1.70)	73.61 (-1.21)	80.42 (-0.75)	78.13 (-1.42)
4) UniTrans w/ Θ_{trans}	79.24 (-0.07)	82.13 (-0.77)	74.91 (0.09)	80.06 (-1.11)	79.09 (-0.46)
5) UniTrans w/o $\mathcal{D}_T$ (*i.e.*, Θ_{teach})	78.24 (-1.07)	81.73 (-1.17)	73.97 (-0.85)	79.07 (-2.10)	78.25 (-1.30)
6) Model Transfer (*i.e.*, Θ_{src})	76.34 (-2.97)	80.61 (-2.29)	72.39 (-2.43)	78.47 (-2.70)	76.95 (-2.60)
7) Data Transfer (*i.e.*, Θ_{trans})	78.14 (-1.17)	80.98 (-1.92)	73.65 (-1.17)	78.91 (-2.26)	77.92 (-1.63)
8) Data Combination	77.31 (-2.00)	80.75 (-2.15)	73.66 (-1.16)	76.59 (-4.58)	77.08 (-2.47)

Table 3: Ablation study for the proposed UniTrans, where numbers in parenthesis denote performance changes.

w/ Θ_{src}, which uses the model Θ_{src} learned on the source-language training data with language-independent features as the teacher to train Θ_{stu} on unlabeled target-language data; 4) *UniTrans w/ Θ_{trans}*, which uses the model Θ_{trans} learned on translated target-language training data as the teacher to train Θ_{stu} on unlabeled target-language data; 5) *UniTrans w/o $\mathcal{D}_T$* (*i.e.*, Θ_{teach}), which cuts out the access to the unlabeled target-language data $\mathcal{D}_T$ and directly applies the teacher model Θ_{teach} to the target language; 6) *Model Transfer* (*i.e.*, Θ_{src}), which directly applies the model Θ_{src} to the target language; 7) *Data Transfer* (*i.e.*, Θ_{trans}), which directly applies the model Θ_{trans} to the target language; 8) *Data Combination*, which directly combines the labeled source-language data and the translated target-language data to train an NER model for the target language with language-independent features. Note that for variants 3) *UniTrans w/ Θ_{trans}* and 4) *UniTrans w/ Θ_{src}*, the hard label loss $\mathcal{L}_{hard}^{\tilde{x}}$ is also removed from Eq. 13, as we cannot use one model to generate pseudo hard labels.

Table 3 highlights the performance contributions of each component in our proposed UniTrans, and removing any of them will generally lead to a performance drop. Moreover, we can draw more in-depth observations as follows.

1) The proposed UniTrans outperforms *UniTrans w/o $\mathcal{L}_{soft}^{\tilde{x}}$* and *UniTrans w/o $\mathcal{L}_{hard}^{\tilde{x}}$* in most cases, indicating that combining both for training the target-language model Θ_{stu} is reasonable. That also validates the effectiveness of the proposed voting scheme to generate pseudo hard labels.

2) *UniTrans w/o $\mathcal{L}_{hard}^{\tilde{x}}$* outperforms *UniTrans w/ Θ_{trans}* and *UniTrans w/ Θ_{src}*. Meanwhile, *UniTrans w/o $\mathcal{D}_T$* (*i.e.*, Θ_{teach}) outperforms *Model Transfer* (*i.e.*, Θ_{src}) and *Data Transfer* (*i.e.*, Θ_{trans}). Such results well demonstrate that the proposed teacher model Θ_{teach}, which unifies model transfer and data transfer, is superior to just using one or the other, no matter whether unlabelled target-language data is used or not. That also well verifies the complementarity between model transfer and data transfer.

3) Comparing UniTrans with *UniTrans w/o $\mathcal{D}_T$* (*i.e.*, Θ_{teach}), *UniTrans w/ Θ_{src}* with *Model Transfer* (*i.e.*, Θ_{src}), and *UniTrans w/ Θ_{trans}* with *Data Transfer* (*i.e.*, Θ_{trans}), respectively, we can see that eliminating the usage of unlabeled target-language data will lead to a consistent performance drop in all experiments. That further demonstrates the importance of leveraging information contained in unlabeled target-language data for cross-lingual NER.

4) *UniTrans w/o $\mathcal{D}_T$* (*i.e.*, Θ_{teach}) outperforms *Data Com-*

	es	nl	de	no	Average
UniTrans	**79.31**	82.90	74.82	81.17	79.55
+ Teacher Ensembling	79.26	**83.07**	**75.55**	**81.30**	**79.79**

Table 4: Results of teacher ensembling ($M = 5$) for UniTrans

bination, meaning that the proposed Θ_{teach} utilizes a superior way to unify model transfer and data transfer than simply combining available labeled data.

4.4 Discussion: Extend with Teacher Ensembling

Considering the randomness brought by dropout layers in the feature encoder (*i.e.*, multilingual BERT) of the proposed UniTrans, we can extend UniTrans with teacher ensembling via taking advantage of the randomness. Specifically, we use an ensemble of teacher models learned via M runs, each denoted as $\Theta_*^{(m)} \in \{\Theta_{src}^{(m)}, \Theta_{teach}^{(m)}, \Theta_{trans}^{(m)}\}$ with $m = 1, 2, ..., M$. And we simply average the predictions of $\Theta_*^{(m)}$ to train the student target-language model Θ_{stu}:

$$p(\tilde{x}_i, \Theta_*) = \frac{1}{M} \sum_{m=1}^{M} p(\tilde{x}_i, \Theta_*^{(m)}) \qquad (14)$$

where $\Theta_* \in \{\Theta_{src}, \Theta_{teach}, \Theta_{trans}\}$. Note that Eq. 14 is not only adopted to predict soft labels for unlabeled target-language data, but also employed in the generation of pseudo hard labels.

Table 4 reports the results of teacher ensembling with $M = 5$. It is evident that ensembling further brings consistent performance improvements on nearly all target languages for UniTrans, with an average gain of 0.24 F1-score.

5 Conclusion

In this paper, we propose a novel approach for cross-lingual NER termed UniTrans, which unifies both model transfer and data transfer based on their complementarity via enhanced knowledge distillation on unlabeled target-language data. We also propose a voting scheme to generate pseudo hard labels for part of words in the unlabeled target-language data, so as to enhance knowledge distillation with supervision from both hard and soft labels. We evaluate the proposed UniTrans on benchmark datasets for four target languages. Experimental results show that UniTrans achieves new state-of-the-art performance for all target languages. We also extend UniTrans with teacher ensembling, which leads to further performance gains.

References

[Bari *et al.*, 2019] M Saiful Bari, Shafiq Joty, and Prathyusha Jwalapuram. Zero-resource cross-lingual named entity recognition. *arXiv preprint arXiv:1911.09812*, 2019.

[Bucilu *et al.*, 2006] Cristian Bucilu, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *SIGKDD*, pages 535–541, 2006.

[Chen *et al.*, 2019] Hui Chen, Zijia Lin, Guiguang Ding, Jianguang Lou, Yusen Zhang, and Borje Karlsson. Grn: Gated relation network to enhance convolutional neural network for named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6236–6243, 2019.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186, 2019.

[He and Sun, 2017] Hangfeng He and Xu Sun. A unified model for cross-domain and semi-supervised named entity recognition in chinese social media. In *AAAI*, 2017.

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

[Jain *et al.*, 2019] Alankar Jain, Bhargavi Paranjape, and Zachary C. Lipton. Entity projection via machine translation for cross-lingual NER. In *EMNLP*, pages 1083–1092, 2019.

[Johansen, 2019] Bjarte Johansen. Named-entity recognition for norwegian. In *NoDaLiDa*, 2019.

[Lample *et al.*, 2018] Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *ICLR*, 2018.

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

[Mayhew *et al.*, 2017] Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. Cheap translation for cross-lingual named entity recognition. In *EMNLP*, pages 2536–2545, 2017.

[Mollá *et al.*, 2006] Diego Mollá, Menno van Zaanen, and Daniel Smith. Named entity recognition for question answering. In *ALTA*, pages 51–58, 2006.

[Ni *et al.*, 2017] Jian Ni, Georgiana Dinu, and Radu Florian. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In *ACL*, pages 1470–1480, 2017.

[Rusu *et al.*, 2015] Andrei A. Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. Policy distillation. *arXiv preprint arXiv:1511.06295*, 2015.

[Sanh *et al.*, 2019] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[Scudder, 1965] H Scudder. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, pages 363–371, 1965.

[Smith *et al.*, 2017] Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint arXiv:1702.03859*, 2017.

[Täckström *et al.*, 2012] Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. Cross-lingual word clusters for direct transfer of linguistic structure. In *NAACL*, pages 477–487, 2012.

[Tjong Kim Sang and De Meulder, 2003] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *HLT-NAACL*, pages 142–147, 2003.

[Tjong Kim Sang, 2002] Erik F. Tjong Kim Sang. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING*, 2002.

[Tsai *et al.*, 2016] Chen-Tse Tsai, Stephen Mayhew, and Dan Roth. Cross-lingual named entity recognition via wikification. In *CoNLL*, pages 219–228, 2016.

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

[Wu and Dredze, 2019] Shijie Wu and Mark Dredze. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. *arXiv preprint arXiv:1904.09077*, 2019.

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

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

[Xie *et al.*, 2018] Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime Carbonell. Neural cross-lingual named entity recognition with minimal resources. In *EMNLP*, pages 369–379, 2018.

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

[Zirikly and Hagiwara, 2015] Ayah Zirikly and Masato Hagiwara. Cross-lingual transfer of named entity recognizers without parallel corpora. In *ACL-IJCNLP*, pages 390–396, 2015.