An Unsupervised Multiple-Task and Multiple-Teacher Model for Cross-lingual Named Entity Recognition

Anonymous ACL submission

Abstract

Cross-lingual named entity recognition task is one of the critical problem for evaluating the potential transfer learning techniques on low resource languages. Knowledge distillation using pre-trained multilingual language models between source and target languages have shown their superiority. However, existing cross-lingual distillation models merely consider the potential transferability between two identical single tasks across both domain. Other possible auxiliary tasks to improve the learning performance have not been fully investigated. In this study, based on the knowledge distillation framework and multi-task learning, we introduce the similarity metric model as an auxiliary task to improve the cross-lingual NER performance on target domain. Specifically, an entity recognizer and a similarity evaluator teachers are first trained in parallel from the source domain. Then, two tasks in the student model are supervised by the two teachers simultaneously. Empirical studies on the datasets across 7 different languages confirm the effectiveness of the proposed model.

1 Introduction

Named entity recognition, NER in short, refers to identifying entity types, i.e. location, person, organization, etc., in a given sentence. The exploiting of deep neural networks, such as Bi-LSTM-CRF (Lample et al., 2016), Bi-LSTM-CNN (Chiu and Nichols, 2016) make this task achieves significant performances. However, since deep neural networks highly relies on a large amount of labelled training data, the annotation acquiring process is expensive and time consuming. This situation is more severe for low-resource languages. With the help of transfer learning (Ruder et al., 2019) and multilingual BERT (short as mBERT) (Devlin et al., 2019), it is possible to transfer the annotated training samples or trained models from a rich-resource domain to a low-resource domain.

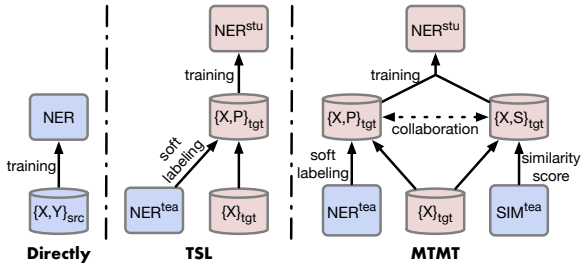


Figure 1: Comparison between previous cross-lingual NER models. **Directly**: direct model transfer; **TSL**: teacher-student learning model; **MTMT**: proposed multiple-task and multiple-teacher Model. NER / NER^{tea}: learned NER model for source language; NER^{stu}: learned NER model for target language; SIM^{tea} learned similarity model for source language; $\{X, Y\}_{src}$: labeled data in source language; $\{X\}_{tgt}$: unlabeled data in target language; $\{X, P\}_{tgt}$: labeled data in target language with probability; $\{X, S\}_{tgt}$: labeled data in target language with entity similarity score.

Many studies have been done to solve this cross-language NER problem. Existing models can be separated into three categories, shared feature space based, translation based and knowledge distillation based. Shared feature space based models exploit language-independent features, which lacks the domain specific features for target language (Tsai et al., 2016; Wu and Dredze, 2019; Keung et al., 2019). Translation based models generate pseudo labeled target language data to train the cross-lingual NER model, but the noise from translation process restrains its performance. (Mayhew et al., 2017; Xie et al., 2018; Wu et al., 2020b). Knowledge distillation based models train a student model using soft labels of the target language (Wu et al., 2020a,b; Chen et al., 2021; Liang et al., 2021). Our model is developed on the basis of (Wu

et al., 2020a).

Although above mentioned models solve the cross-lingual NER problem in some extent, the auxiliary tasks, as in the multi-task learning, have not been studied in this problem. Due to the distributed representation of natural languages, the relatedness among the embedding of target languages, which is measured by the similarity, can be utilized to further boost the learned encoder and improve the final NER performance on target language.

Here we give a concrete example to illustrate the importance of similarity between every two tokens under the situation when only the English data is labeled. Given a Spanish sentence "*Arévalo (Avila), 23 may (EFE).*", the token "*Arévalo*" is recognized as ORG type using the learned model from English domain. In the meantime, the token "*Arévalo*" has high similarity scores with the Spanish tokens "*Viena*" from sentence "*Viena, 23 may (EFE).*", and "*Madrid*" from sentence "*Madrid, 23 may (EFE).*". Also, the tokens "*Viena*" and "*Madrid*" are recognized correctly as LOC type using the same English model mentioned above. Then "*Arévalo*" can be recognized correctly as LOC type under the supervisory signal using the similarity between "*Viena*" and "*Madrid*".

To leverage the similarity between the tokens of the source languages, we design an multiple-task and multiple-teacher model (short as MTMT, as shown in Figure 1), which helps the NER learning process on the target languages. Specifically, we first introduce the knowledge distillation to build entity recognizer and similarity evaluator teachers in the source language and transfer the learned patterns to the student in the target language. In the student model, we then borrow the idea of multi-task learning to incorporate a similarity evaluation task as an auxiliary task into the entity recognition classifier. During the student learning process, we input unlabelled samples from the target languages into the entity recognizer and evaluator, and take output pesudo labels as supervisory signals for these two tasks in the student model. Note that a weighting strategy is also provide therein to take into consideration of the reliability of the teachers.

We validate the model performance on the three commonly-used datasets across 7 languages and the experimental results shows the superiority our presented MTMT model.

Our main contributions are as follows:

- We propose an unsupervised knowledge distillation framework for cross-language named entity recognition and develop a teaching and learning procedure under this framework.
- We present a novel multiple-task and multiple-teacher model that introduces a entity similarity evaluator to boost the performance of student recognizer on target languages.
- We conduct extensive experiments on seven languages compared with state-of-the-art baselines and the results confirm the effectiveness of the presented model.

2 Related Work

Our approach is closely related to the existing works on cross-lingual NER, knowledge distillation and siamese network.

Cross-Lingual NER aims to extract entities from a target language but assumes only source language is annotated. The existing models can be categorized to: a) Shared feature space based models, b) Translation based models, c) Knowledge distillation based models.

Shared feature space based models generally train a language-independent encoder using source and target language data (Tsai et al., 2016). Recently, the pre-trained multilingual language models mBERT is effective to address the challenge (Devlin et al., 2019). Moreover, some research introduces new components on top of the mBERT by directly transferring the model learned from labeled source language to that of target languages (Keung et al., 2019). The performance is still weak due to the lack of annotations of target languages.

Translation based models generally generate pesudo labeled target data to alleviate target data scarcity. For example, (Wu et al., 2020b; Zhang et al., 2021) gain a improvement by translating the labeled source language to the target language word-by-word. Our model achieves considerable improvement by learning entity similarity in target language data without translation.

Knowledge distillation based models includes a teacher model and a student model (Wu et al., 2020c). The teacher model is trained on labeled source language. The student model learns from the soft label predicted by teacher model on unlabeled target language data. Therefore, the student model can capture the extra knowledge about target languages. In our work, the student model not only learns the recognizer teacher knowledge, but also

learns the entity similarity knowledge inspired by multi-task learning.

Siamese Network is originally introduced by (Bromley et al., 1994) to treat signature verification as a matching problem. It has been successfully applied to transfer learning such as one-shot image recognition (Koch et al., 2015), text similarity (Neculoiu et al., 2016). However, there is a dilemma to adapt siamese network to token-level recognition tasks such as NER. Siamese network assumes the input is a pair, and the output is a similarity score. To handle this issue, we reconstruct the data to pair format. To the best of our knowledge, we are the first to learn the entity similarity by siamese network.

3 Framework

In this section, we introduce our framework and its detailed implementation. Our framework is consist of two models: *teacher training model* learned from source language and *teacher-student distillation learning model* learned from target language. In the teacher training model, there are two sub-models, i.e. an entity recognizer teacher and a similarity evaluator teacher. These two models are two parallel tasks, wherein the entity recognition teacher focuses on identifying the named entities and the similarity evaluator teacher is to decide if two tokens are in the same type.

We then present a teacher-student distillation learning model to learn from the two learned teacher models simultaneously. We note that, in this learning process, such a knowledge distillation makes the student model combine the advantages of both source language patterns of entity recognition and entity similarity evaluation. During the learning process, the samples from target language are fed into the teacher model and the outputs are taken as the supervisory signal for two tasks in the student model. To guarantee the student learning performance, we assign weights for each supervisory signal correspond to the output confidence of teacher sub-models. We argue that the student entity recognition task and the student entity similarity evaluation task improve the representation learning of the student encoder in the siamese structure.

3.1 Problem Definition

Following standard practice, we formulate cross-lingual NER as a sequence labeling task. Given a sentence $\boldsymbol{x} = \{x_i\}_{i=1}^{L}$ with L tokens, a NER model produces a sequence of labels $\boldsymbol{y} = \{y_i\}_{i=1}^{L}$, where x_i is the i-th token and y_i is the corresponding label of x_i. In the source language, we denote the labeled training data as $\mathcal{D}_{train}^{S} = \{(\boldsymbol{x}, \boldsymbol{y})\}$ and test data as $\mathcal{D}_{test}^{S}$. In the target language, we denote the unlabeled train data as $\mathcal{D}_{train}^{T} = \{\boldsymbol{x}\}$ and the test data as $\mathcal{D}_{test}^{T}$. Formally, our goal is to train a model with $\mathcal{D}_{train}^{S}$ and $\mathcal{D}_{train}^{T}$ to perform well on $\mathcal{D}_{test}^{T}$.

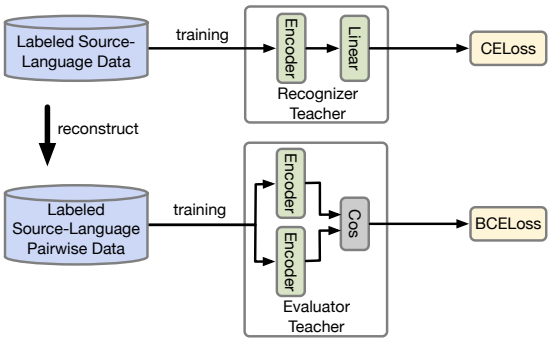


Figure 2: The training process of teacher models.

3.2 Teacher Models

Here we first consider the training of two teacher models. For every two tokens, we define *Entity Similarity Metric* as a score which is the probability that two tokens belong to the same entity type. We aim to find entity similarity to help the cross-lingual NER model in target language. It is a non-trivial task since we lack golden labels to help us distinguish target named entities. To address this challenge, we propose a binary classifier called similarity teacher to leverage the labeled source language data for similarity prediction. Our similarity teacher model, inspired by siamese network (Koch et al., 2015), are able to acquires more powerful features via capturing the invariances to transformation in the input space. Figure 2 illustrated the two teacher models training. The following subsections will illustrate the two teacher models sequentially.

3.2.1 Entity Recognizer Teacher

Since the cross-lingual NER task, we unitize multilingual mBERT (Wu and Dredze, 2019) as basic sequence feature extractor backbone to derive the sequence embedding representation throughout this paper. And a linear classifier with softmax upon the pre-trained mBERT output. The model network

structure could be formulated as,

$$\boldsymbol{h} = \text{mBERT}(\boldsymbol{x})$$
$$\hat{y}_i = \text{softmax}(W h_i + b)$$

where $\boldsymbol{h} = \{h_i\}_{i=1}^{L}$ and h_i denotes the output of the pretrained mBERT that corresponds to the input token x_i. $\hat{y}_i$ denotes the predicted probability distribution for x_i. W and b are trainable parameters. For some sentence sample $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D}^S_{train}$ and an entity token query index i, the loss function is,

$$\mathcal{L}_{ER}(\boldsymbol{x}, \boldsymbol{y}, i) = \mathcal{L}_{CE}(y_i, \hat{y}_i)$$

We train this entity recognition teacher model on the source lingual training corpus $\mathcal{D}^S_{train} = \{(\boldsymbol{x}, \boldsymbol{y})\}$ directly.

3.2.2 Siamese Entity Similarity Evaluator

In order to leverage the entity similarity to boost the unsupervised cross-lingual NER performance, we will present our entity pairs construction method and the siamese network model in the following.

Entity Similarity Pairs Construction According to entity labels, we randomly select sentences pair $< \boldsymbol{x}, \boldsymbol{x'} >$ with their some token pair $< x_i, x'_j >$ and associated labels $< y_i, y'_j >$ in $\mathcal{D}^S_{train}$, to form the siamese supervision training dataset, $\mathcal{D}^{S-siam}_{train} = \{(\boldsymbol{x}, \boldsymbol{x'}, i, j, t)\}$ where the target $t = 1$ indicates $y_i = y'_j$, and 0 otherwise. And the testing entity pairs $\mathcal{D}^{S-siam}_{test}$ is constructed likewisely.

Siamese Entity Similarity Network Our similarity backbone model is a siamese neural network with mBERT as feature extraction layer. Wherein $\boldsymbol{h}$ and $\boldsymbol{h'}$ represent latent sequences encoding features derived by the two symmetric twins with respect to input sentence $\boldsymbol{x}$ and $\boldsymbol{x'}$ respectively.

The inter-entities similarity is measured on the tokens hidden representations h_i and h'_j, queried by the entity indices $< i, j >$ on the sequences representations. The cosine function operator is added to compute on the entity token latent vectors' distance, so as to measure the similarity between each siamese twin, which is fed into a single sigmoid output unit for target $\hat{t}$ estimation.

More precisely, for a specific entity pair $(\boldsymbol{x}, \boldsymbol{x'}, i, j, t) \in \mathcal{D}^{S-siam}_{train}$, the siamese network could be formulated as,

$$\boldsymbol{h} = \text{mBERT}(\boldsymbol{x}),\ \boldsymbol{h'} = \text{mBERT}(\boldsymbol{x'})$$
$$\hat{t}(\boldsymbol{x}, \boldsymbol{x'}, i, j) = \sigma(\cos(h_i, h'_j))$$

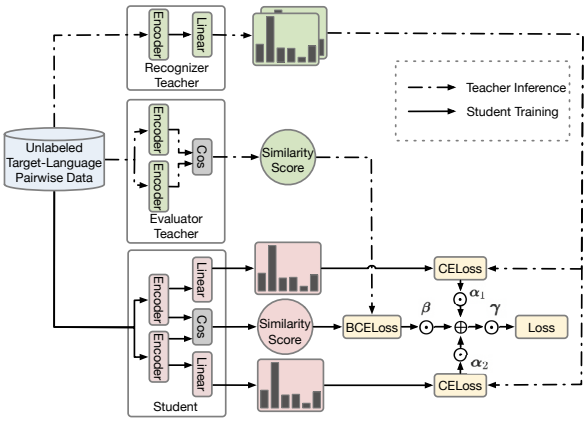


Figure 3: Teacher-student distillation learning.

where cos is the cosine similarity metric function, σ is the sigmoid activation function, $\hat{t} \in [\sigma(-1), \sigma(1)]$ denotes the predicted similarity of two queried tokens pair $< x_i, x'_j >$. Larger $\hat{t}$ value indicates higher similarity between the two queried entities tokens.

The loss function of the similarity prediction can be formulate as,

$$\mathcal{L}_{SIM}(\boldsymbol{x}, \boldsymbol{x'}, i, j, t) = \mathcal{L}_{BCE}(t, \hat{t}).$$

Finally, we can train the siamese entity similarity evaluator on $\mathcal{D}^{S-siam}_{train}$, and evaluate the performance on test dataset $\mathcal{D}^{S-siam}_{test}$. Together with entity recognizer model, this entity similarity evaluator are used as teachers in following knowledge distillation learning process, and transfer knowledge from source to target lingual corpus.

3.3 Teacher Student Distillation Learning

In this section, we consider to transfer the named entity type and similarity knowledge learned on labeled source language corpus to unlabeled target language NER task. To this end, we propose a knowledge distillation learning process to train a target language student NER model with its supervisory signals mimicked by the entity type prediction probability by the entity recognizer teacher model and entity representation similarity target by the entity siamese similarity evaluator teacher model. Based on the original unlabeled target sentence training data $\mathcal{D}^T_{train}$, we again construct unlabeled target-language siamese pairwise entity data $\mathcal{D}^{T-sim}_{train} = \{(\boldsymbol{x}_T, \boldsymbol{x'}_T, i, j)\}$, with the sentence pair $< \boldsymbol{x}_T, \boldsymbol{x'}_T >$ randomly sample from $\mathcal{D}^T_{train}$ and the entity token indices pair $< i, j >$ uniformly sampled from the sentences therein.

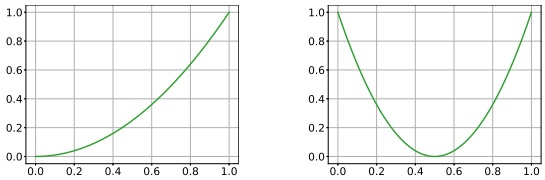
The multi-lingual BERT is also used as encoder for the sentence siamese pair, and the entity token feature queried from the latent sequence encoding representation. Specifically, for a sentence pair $(\boldsymbol{x}_T, \boldsymbol{x}'_T, i, j) \in \mathcal{D}_{train}^{T-sim}$, the student model transform them as follows,

$$\begin{aligned}
\boldsymbol{h}_T &= \text{mBERT}(\boldsymbol{x}_T) \\
\hat{y}_{T_i} &= \text{softmax}(W h_{Ti} + b) \\
\boldsymbol{h}'_T &= \text{mBERT}(\boldsymbol{x}'_T) \\
\hat{y}'_{T_j} &= \text{softmax}(W h'_{Tj} + b) \\
\hat{t}_T(\boldsymbol{x}_T, \boldsymbol{x}'_T, i, j) &= \sigma(\cos(h_{Ti}, h'_{Tj}))
\end{aligned}$$

Then for a specific sentence pair sample in the target siamese dataset, the student loss function has three breaches, $\mathcal{L}_{ER}(\boldsymbol{x}_T, \boldsymbol{y}_S, i)$, $\mathcal{L}_{ER}(\boldsymbol{x}'_T, \boldsymbol{y}'_S, j)$, and $\mathcal{L}_{SIM}(\boldsymbol{x}_T, \boldsymbol{x}'_T, i, j, \hat{t}_S)$. Note that supervision information $\boldsymbol{y}_S$, $\boldsymbol{y}'_S$, and $\hat{t}_S$ are taught by the three teacher models. Summering over all the samples in $\mathcal{D}_{train}^{T-sim} = \{(\boldsymbol{x}_T, \boldsymbol{x}'_T, i, j)\}$, the total student model training loss takes form,

$$\begin{aligned}
\mathcal{L} = \gamma \sum_{(\boldsymbol{x}_T, \boldsymbol{x}'_T, i, j) \in \mathcal{D}_{train}^{T-sim}} & (\alpha_1 \mathcal{L}_{ER}(\boldsymbol{x}_T, \boldsymbol{y}_S, i) \\
& + \alpha_2 \mathcal{L}_{ER}(\boldsymbol{x}'_T, \boldsymbol{y}'_S, j) \\
& + \beta \mathcal{L}_{BCE}(\hat{t}_T(\boldsymbol{x}_T, \boldsymbol{x}'_T, i, j), \hat{t}_S))
\end{aligned}$$

where α_1, α_2, β and γ are weights in loss function which are set to make the student model learns less noisy knowledge from teachers. The weights are set as follows: $\alpha_1(\alpha_2)$ is an increasing function with respect to the output of the entity recognizer teacher as shown in Figure.4. And β is set such that it is high when the output of the entity similarity teacher is close to 0 or 1, and it is low when the output is close to 0.5. γ indicates consistency level between the outputs from two teacher models, e.g. for two input tokens, if the output from entity similarity teacher is high, and the similarity level computed from the outputs of the entity recognizer teacher is low, then their consistency level is low. We want the student model to learn from the two teachers as follows: the higher the prediction of the entity recognizer teacher is (the further away from 0.5 the prediction of the entity similarity teacher is, the higher the consistency level is), the more accurate the prediction is, thus the more attention the student model pays attention to the input tokens, and vice versa. Therefore, we heuristically devises the three weights scheduling as functions of the inputs,



(a) (b)

Figure 4: Weights of loss. (a) indicates the weight $\alpha_{(\cdot)}$ of $\mathcal{L}_{ER}$. (b) indicates the weight β of $\mathcal{L}_{BCE}$.

$$\begin{aligned}
\alpha_{(\cdot)} &= (max(\hat{y}_{T_i}))^2 \\
\beta &= (2\hat{t}_T(\boldsymbol{x}_T, \boldsymbol{x}'_T, i, j) - 1)^2 \\
\gamma &= 1 - |\sigma(\cos(\hat{y}_{T_i}, \hat{y}'_{T_j})) - \hat{t}_T(\boldsymbol{x}_T, \boldsymbol{x}'_T, i, j)|
\end{aligned}$$

4 Experiment

In this section, we evaluate our multiple-task and multiple-teacher model for cross-lingual NER and compare our model with a series of state-of-the-art models.

4.1 Dataset

We conducted experiments on three benchmark datasets: CoNLL2002 (Tjong Kim Sang, 2002), CoNLL2003 (Tjong Kim Sang and De Meulder, 2003) and WikiAnn (Pan et al., 2017). CoNLL2002 includes Spanish and Dutch, CoNLL2003 includes English and German, and WikiAnn includes English and three non-western languages: Arabic, Hindi, and Chinese. Each language is divided into a training set, a development set and a test set. All datasets were annotated with four entity types: LOC, MISC, ORG, and PER. Following (Wu and Dredze, 2019), all datasets are annotated using the BIO entity labelling scheme. To imitate the zero-resource cross lingual NER case, following (Wu and Dredze, 2019), we used English as the source language and other languages as the target language. In cross-lingual NER, the training set without entity label of the target language is also available when training the model. We trained the model with the labeled training set of the source language and evaluated the model on the test set of each target language. Table 1 and 2 shows the statistics of all datasets.

4.2 Implementation Details

We use PyTorch 1.7.1 to implement our model. All of the feature encoders mentioned in this paper use

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

Table 1: Statistics of CoNLL.

Language	Type	Train	Dev	Test
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

Table 2: Statistics of WikiAnn.

pretrained multilingual bert model (Devlin et al., 2019) in HuggingFace's Transformer[1], which has 12 Transformer blocks, 12 attention heads, and 768 hidden units.

We set our hyperparameters empirically following (Wu et al., 2020c) with some modifications. We do not freeze any layers and we use the output of the last layer as our hidden feature vector. We set batch size to be 32, maximum sequence length to be 128, dropout rate to be 0.2, and we use Adam as optimizer (Kingma and Ba, 2014). For the training of recognition teacher model and similarity teacher model, we set the learning rate to be 1e-5 and 5e-6 separately. For knowledge distillation, we use a learning rate of 1e-6 for the student models training. Note that if a word is divided into several subwords after tokenization, then only the first subword is considered in the loss function. Following (Tjong Kim Sang, 2002), we use the entity level F1-score as the evaluation metric. Moreover, we conduct each experiment 5 times and report the mean F1-score.

4.3 Comparison

Table 3 and 4 report the zero-resource cross-lingual NER results of different models on 6 target languages.

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

Model	de	es	nl
Wiki(Tsai et al., 2016)	48.12	60.55	61.56
WS(Ni et al., 2017)	58.50	65.10	65.40
TMP(Jain et al., 2019)	61.50	73.50	69.9
Bert-f(Wu and Dredze, 2019)	69.56	74.96	77.57
AdvCE(Keung et al., 2019)	71.90	74.3	77.6
TSL(Wu et al., 2020a)	73.16	76.75	80.44
Unitrans(Wu et al., 2020b)	74.82	79.31	82.90
w/o translation	73.61	77.3	81.20
AdvPicker(Chen et al., 2021)	75.01	79.00	82.90
RIKD(Liang et al., 2021)	76.08	79.78	82.96
w/o IKD	74.86	78.90	81.02
TOF(Zhang et al., 2021)	76.57	80.35	82.79
w/o continual learning	76.39	79.44	81.64
MTMT	**76.80**	**81.82**	**83.41**

Table 3: Performance comparisons on CoNLL.

Model	ar	hi	zh
Bert-f(Wu and Dredze, 2019)	42.30	67.60	**52.90**
TSL(Wu et al., 2020a)	43.12	69.54	48.12
RIKD(Liang et al., 2021)	45.96	70.28	50.40
MTMT	**52.77**	**70.76**	52.26

Table 4: Performance comparisons on WikiAnn.

Wiki (Tsai et al., 2016) introduces a language independent model building on cross-lingual wikification for cross-lingual NER.
WS (Ni et al., 2017) presents two weakly supervised approaches for cross-lingual NER.
TMP (Jain et al., 2019) leverages machine translation to improve annotation projection approaches to cross-lingual NER.
Bert-f (Wu and Dredze, 2019) applys the multilingual BERT to cross-lingual NER.
AdvCE (Keung et al., 2019) improves upon multilingual BERT via adversarial learning for cross-lingual NER.

Model	de	es	nl
MTMT	**76.80**	**81.82**	**83.41**
MTST	74.11 (-2.69)	78.61 (-3.21)	81.97 (-1.44)
MTMT w/o weighting	76.08 (-0.72)	80.84 (-0.98)	82.96 (-0.45)
MTMT w/o similarity	73.82 (-2.98)	77.53 (-4.29)	80.82 (-2.59)

Table 5: Ablation study on cross-lingual NER.

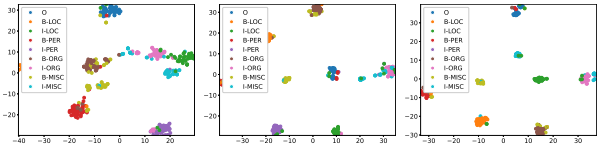
#1 Spanish	**Entity Recognizer Teacher:** Arévalo[B-ORG] (Avila[B-LOC]), 23 may (EFE[B-ORG]). **Student:** Arévalo[B-LOC] (Avila[B-LOC]), 23 may (EFE[B-ORG]). **Entity Recognizer and Entity Similarity Evaluator Teachers:** a. Viena[B-LOC, 0.7157] , 23 may (EFE[B-ORG]). b. Madrid[B-LOC, 0.7156] , 23 may (EFE[B-ORG]).
#2 Dutch	**Entity Recognizer Teacher:** Universiteit[B-ORG] Antwerpen[I-ORG] (Ruca[B-LOC]) en De... **Student:** Universiteit[B-ORG] Antwerpen[I-ORG] (Ruca[B-ORG]) en De... **Entity Recognizer and Entity Similarity Evaluator Teachers:** a. ...voor[I-ORG] het[I-ORG] Preventiebeleid[I-ORG] (VSPP[B-ORG,0.7134]) is... b. Transparency[B-ORG] International[I-ORG] (Sozialarbeit[I-TI,0.7130]), de onderhand...
#3 German	**Entity Recognizer Teacher:** Hessischen[B-ORG] Staatskanzlei[O] auf das Thema... **Student:** Hessischen[B-ORG] Staatskanzlei[I-ORG] auf das Thema... **Entity Recognizer and Entity Similarity Evaluator Teachers:** a. Internationalen[B-ORG] Bund[I-ORG] für[I-ORG] Sozialarbeit[I-ORG,0.7162] ... b. Kickers[B-ORG] Offenbach[I-ORG] II[I-ORG,0.7157] - Rotweiß[B-ORG] ...

Table 6: Case study on cross-lingual NER. The GREEN (RED) highlight indicates a correct (incorrect) label. The real-valued numbers indicate the entity similarity score.

TSL (Wu et al., 2020c) proposes a teacher-student learning model for cross-lingual NER.
Unitrans (Wu et al., 2020b) unifies a data transfer and model transfer for cross-lingual NER.
AdvPicker (Chen et al., 2021) proposes a adversarial discriminator for cross-lingual NER.
RIKD (Liang et al., 2021) develops a reinforced iterative knowledge distillation for cross-lingual NER.
TOF (Zhang et al., 2021) transfers knowledge from three aspects for cross-lingual NER.

It can be seen that our model outperforms the state-of-the-arts. Specifically, compared with the remarkable RIKD, AdvPicker and Unitrans, which also use knowledge distillation but ignore the entity similarity knowledge, our model obtains significant and consistent improvements in F1-score ranging from 0.23 for German[de] to 6.81 for Arabic[ar]. That demonstrates the benefits of our proposed MTMT model, compared to direct model transfer (Wu and Dredze, 2019).

Note that Bert-f performs better than our model on Chinese dataset due to their re-tokenization of the dataset. Moreover, compared with the latest model TOF, RIKD, Unitrans, our model requires much lower computational costs for both translation and iterative knowledge distillation, meanwhile reaching superior performance. For a fair comparison, we compare our model against the version of TOF w/o continual learning (Zhang et al., 2021), RIKD w/o IKD (Liang et al., 2021) and Unitrans w/o translation (Wu et al., 2020b) as reported in their paper.



(a) (b) (c)

Figure 5: t-SNE plot of embeddings of teacher and student models. (a) Entity recognizer teacher. (b) Entity similarity evaluator teacher. (c) Student.

4.4 Ablation Study

To demonstrate the effectiveness of our approach, we designed the following ablation studies. Table 5 presents the results.

(1) *MTST*, which combines the multiple-teacher to single-teacher. That is, both of the teacher and student have the same neural network structure. This causes a performance drop across all languages due to two single teachers cannot make a difference with combination.

(2) *MTMT w/o weighting*, which set the α_1,α_2, β and γ all to be 1 in the loss of student model learning. It can be seen that the performance decrease in terms of F1-score ranges from 0.45 for Dutch(nl) to 0.98 for Spanish(es),

which validates that weighting loss can bring more confident knowledge to student model.

(3) *MTMT w/o similarity*, which removes the similarity teacher model. In this case, our approach degrades into the Single Teacher-Student learning model as in TSL (Wu et al., 2020a). Without the similarity knowledge fed into the student model, the performance drops significantly.

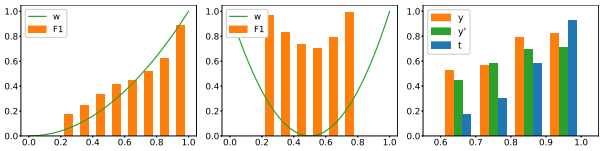
4.5 Case Study

We give a case study to show that the failed cases of baseline models can be corrected by our model. We try to bring up insights on why the proposed multiple-task and multiple-teacher model works.

The proposed MTMT model can help to correct labels using the *Entity Similarity* defined in section 3.2. Specifically, if there is a set of tokens in which every two of them have high *Entity Similarity* score, and one of the tokens is predicted to have a distinct label while other tokens have identical labels, then the one with the distinct label is predicted wrongly and is corrected by the student model to have the label of all other tokens. As shown in Table 6, in example #1, the entity recognizer teacher fails to identify "*Arévalo*" as B-ORG type, while the student model can correctly predict it. The reason lies in that the entity recognizer teacher predicts "*Viena*"('*Madrid*") as B-LOC type correctly, and the similarity evaluator teacher predicts "*Viena*"("*Madrid*") to have a high similarity score(0.7157, 0.7156) with "*Arévalo*". The student learns from both teachers and predict the correct label for "*Arévalo*". Examples #2 and #3 present the same results with different sentences.

4.6 Embeddings Distribution

This section investigates the effect of embeddings of the two different teacher models. It can be seen that the embeddings distribution of student model is close to similarity evaluator teacher, as illustrated in Figure 5. We conjecture that the student model captures similarity knowledge from the similarity evaluator teacher, i.e. the same class of examples tend to cluster and the different class of examples tend to segregate in the embeddings distribution. This validates the proposed MTMT model not only transfers cross-lingual NER knowledge from source language, but also learns the similarity knowledge of target language data.



(a) (b) (c)

Figure 6: Weights analysis of student learning.
(a) α, F1-score in different probability interval.
(b) β, F1-score in different similarity score interval.
(c) F1-score of y_S, y'_S, and $\hat{t}_S$ in different γ interval.

4.7 Effect of Weights

In the section, we evaluate the effectiveness of the weighting loss in student learning from quantitative perspective. All of the following experiments are conducted on Spanish(es) data.

For α analysis, we calculate the F1-score in different probability intervals of entity recognizer teacher, we find that the recognizer teacher tends to predict more correct in higher probability interval, as illustrated in Figure 6a. Therefore, the student model is better suited to target language with learning less low-confidence misrecognitions for the target language.

For β analysis, we observe that F1-score are increasing with the entity similarity score from 0.5 to both sides 0 and 1 in Figure 6b. The encoder of student model obtains the clustering information of the target language with the help of β.

For γ analysis, we consider the consistency of recognition results and similarity score by teachers. The F1-score and similarity score of teachers are all higher in the higher γ intervals, as shown in Figure 6c. The student model learns less from unreasonable results, and it can make more accuracy entity recognition for the target language.

5 Conclusion

In this paper, we propose an unsupervised multiple-task and multiple-teacher model for cross-lingual NER. The student model learns two source language patterns of entity recognition and entity similarity evaluation. Moreover, in order to guarantee the student learning performance, we also propose a weighting strategy to take consideration of the reliability of the teachers. Our experimental results show that the proposed model yields significant improvements on six target language datasets and outperforms the existing state-of-the-art approaches.

References

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1994. Signature verification using a "siamese" time delay neural network. In *Advances in Neural Information Processing Systems*, volume 6. Morgan-Kaufmann.

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 (Volume 1: Long Papers)*, pages 743–753, Online. Association for Computational Linguistics.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. Cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille.

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, San Diego, California. Association for Computational Linguistics.

Shining Liang, Ming Gong, Jian Pei, Linjun Shou, Wanli Zuo, Xianglin Zuo, and Daxin Jiang. 2021. Reinforced iterative knowledge distillation for cross-lingual named entity recognition. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery amp; Data Mining*, KDD '21, page 3231–3239, New York, NY, USA. Association for Computing Machinery.

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

Paul Neculoiu, Maarten Versteegh, and Mihai Rotaru. 2016. Learning text similarity with Siamese recurrent networks. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 148–157, Berlin, Germany. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 1470–1480, Vancouver, Canada. 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 (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota. Association for Computational Linguistics.

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, Berlin, Germany. Association for Computational Linguistics.

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

Qianhui Wu, Zijia Lin, Börje F. Karlsson, Biqing Huang, and Jian-Guang Lou. 2020b. Unitrans : Unifying model transfer and data transfer for cross-lingual named entity recognition with unlabeled data. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3926–3932. International Joint Conferences on Artificial Intelligence Organization. Main track.

Qianhui Wu, Zijia Lin, Guoxin Wang, Hui Chen, Börje F. Karlsson, Biqing Huang, and Chin-Yew Lin. 2020c. Enhanced meta-learning for cross-lingual named entity recognition with minimal resources. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9274–9281.

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

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

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