

- direct transfer.
- meta-learning to fast adaptation new task.
- pseudo-meta-NER task to generate the similar data set.
- mask scheme.
- max loss.
- low-resource.

Enhanced Meta-Learning for Cross-lingual Named Entity Recognition with Minimal Resources

Qianhui Wu¹, Zijia Lin^{2*}, Guoxin Wang^{2*}, Hui Chen³, B'orje F. Karlsson²
Biqing Huang¹, Chin-Yew Lin²

¹Beijing National Research Center for Information Science and Technology (BNRist)
Department of Automation, Tsinghua University, Beijing 100084, China

²Microsoft Research, Beijing, China

³Beijing National Research Center for Information Science and Technology (BNRist)
School of Software, Tsinghua University, Beijing 100084, China
wuqianhui@tsinghua.org.cn, {zijlin, guow, borje.karlsson, cy1}@microsoft.com
jichenhui2012@gmail.com, hbq@tsinghua.edu.cn

Abstract

For languages with no annotated resources, transferring knowledge from rich-resource languages is an effective solution for named entity recognition (NER). While all existing methods directly transfer from source-learned model to a target language, in this paper, we propose to fine-tune the learned model with a few similar examples given a test case, which could benefit the prediction by leveraging the structural and semantic information conveyed in such similar examples. To this end, we present a meta-learning algorithm to find a good model parameter initialization that could fast adapt to the given test case and propose to construct multiple pseudo-NER tasks for meta-training by computing sentence similarities. To further improve the model's generalization ability across different languages, we introduce a masking scheme and augment the loss function with an additional maximum term during meta-training. We conduct extensive experiments on cross-lingual named entity recognition with minimal resources over five target languages. The results show that our approach significantly outperforms existing state-of-the-art methods across the board.

Introduction

Named entity recognition (NER) is the task of locating and classifying text spans into pre-defined categories such as locations, organizations, *etc.* It is a fundamental component in many downstream tasks. Most state-of-the-art NER systems employ neural architectures (Huang, Xu, and Yu 2015; Lample et al. 2016; Ma and Hovy 2016; Chiu and Nichols 2016; Peters et al. 2017; 2018), and thus, depend on a large amount of manually annotated data, which prevents their adaptation to low-resource languages due to the high annotation cost. An effective solution to this problem, which we refer to as cross-lingual named entity recognition, is transferring knowledge from a high-resource source language with abundant annotated data to a low-resource target language with limited or even no annotated data.

In this paper, we attempt to address the extreme scenario of cross-lingual transfer with minimal resources, where there is only one source language with rich labeled data

while no labeled data is available in target languages. To tackle this problem, some approaches convert the cross-lingual NER task into a monolingual NER task by performing annotation projection using bilingual parallel text and word alignment information (Ni, Dinu, and Florian 2017). To eliminate the requirement of parallel texts, some methods propose to translate the labeled data of the source language at the phrase/word level, which inherently provides alignment information for label projection (Mayhew, Tsai, and Roth 2017; Xie et al. 2018). Instead of generating labeled data in target languages, other works explore language-independent features and perform cross-lingual NER in a direct-transfer manner, where the model trained on the labeled data of the source language is directly tested on target languages (Tsai, Mayhew, and Roth 2016; Ni, Dinu, and Florian 2017). Among these methods, cross-lingual word representations are the most prevalent language-independent features. For example, the multilingual version of BERT (Devlin et al. 2019) utilizes Word-Piece modeling strategy to project word embeddings of different languages into a shared space and achieved state-of-the-art performance (Wu and Dredze 2019). In this paper, we leverage the multilingual BERT (Devlin et al. 2019) as a base model to produce cross-lingual word representations.

While all existing direct transfer based methods straightly evaluate the source-trained model on target languages, we hold the idea that the source-trained model can be further effectively improved. Indeed, recent developments in learning cross-lingual sentence representations suggest that any sentence can be encoded into a shared space by building universal cross-lingual encoders (Wu and Dredze 2019; Lample and Conneau 2019). By simply calculating cosine similarity between sentences in different languages with multilingual BERT (Devlin et al. 2019), we find that it is possible to retrieve a few source examples that are quite similar to a given target example in structure or semantics, as shown in Table 1. In Example #1, both sentences have a structure of "Location - Date", while in Example #2, both sentences are about people talking about sports. Intuitively, reviewing the structural and semantic information conveyed by similar examples might benefit prediction. Therefore, given a test example in a target language, we pro-

*Both authors contributed equally to this work.

#1	Ginebra [B-LOC] , 23 may (EFECOM [B-ORG]) . PRESS DIGEST - Israel [B-LOC] - Aug 25 .
#2	Flores [B-PER] afirmó: “con él intentaremos ganar en velocidad, que es una de las mejores virtudes que tiene este equipo.” (Flores said: “With him, we will try to win faster, which is one of the best advantages of this team.”) “Things fell in for us,” said Sorrento [B-PER], who has six career grand slams and hit the ninth of the season for the Mariners [B-ORG] .

Table 1: Examples of similar sentence pairs in structure (#1) or semantics (#2), where WHITE (GREY) highlights the Spanish (retrieved English) examples.

pose to first retrieve a small set of similar examples from the source language, and then, use these retrieved examples to fine-tune the model before testing.

However, if the retrieved similar set is too large, too much noise will be introduced via relatively distant examples. And thus to avoid misleading the model with distant examples, the cardinality of a similar group is typically small. In such a scenario, the model is expected to achieve higher performance on a test example after only one or a few fine-tuning steps using the limited-size set of retrieved examples. This inspires us to apply meta-learning, which aims to learn a model that facilitates fast adaptation to new tasks with a minimal amount of training examples (Andrychowicz et al. 2016; Vinyals et al. 2016; Finn, Abbeel, and Levine 2017).

In this paper, we follow the recently proposed model-agnostic meta-learning approach (Finn, Abbeel, and Levine 2017) and extend it to the cross-lingual NER task with minimal resources, where no labeled data is provided in target languages. We construct a set of pseudo-meta-NER tasks using the labeled data from the source language and propose a meta-learning algorithm to find a good model parameter initialization that could fast adapt to new tasks. When it comes to the adaptation phase, we regard each test example as a new task, build a pseudo training set for it, and fine-tune the meta-trained model before testing.

When adapting meta-learning to cross-lingual NER, we notice that most mispredictions occur on language-specific infrequent entities. It is known that an NER system makes predictions through word features of an entity itself, and the syntactic or semantic information of its context. However, most entities are generally of low frequency in the training corpora of the base model, and thus, entity representations across different languages are not well-aligned in the shared embedding space. That is, for the prediction of a low-frequency entity, over-dependence on its own features will inhibit the model transferring across languages. Therefore, we introduce a masking scheme on named entities during meta-training to weaken the dependence on entities and promote the prediction through contextual information. Meanwhile, considering that the commonly used average loss over all tokens treats each token equally though some tokens may be more difficult to learn and easier to be mispredict, we add

a maximum term to the original loss function, which makes the model focus more on such tokens and thus reduce mispredictions, so that the meta-knowledge of these mispredictions will not be transferred to target languages.

To summarize our contributions:

- We propose a model-agnostic meta-learning-based approach to tackle cross-lingual NER with minimal resources. To our best knowledge, this is the first successful attempt in adapting meta-learning to NER.
- We propose a masking scheme on named entities and augment the loss function with an additional maximum term during meta-training, to facilitate the model’s ability to generalize across different languages.
- We evaluate our approach over 5 target languages, *i.e.*, Spanish, Dutch, German, French, and Chinese. We show that the proposed approach significantly outperforms existing state-of-the-art methods across the board.

Related Work

Cross-lingual NER with Minimal Resources

There are two major branches of work in cross-lingual NER with minimal resources: methods based on annotation projection and methods based on direct transfer.

One of the typical approaches in the annotation projection category is to take bilingual parallel corpora, annotate the source side, and project the annotations to the target using learned word alignment information (Ni, Dinu, and Florian 2017). However, these methods depend on parallel texts, as well as annotations in at least one side, which is unavailable in many cases. To eliminate the requirement of parallel data, some approaches first translate source-language labeled data at the word/phrase level, and then directly copy labels across languages (Xie et al. 2018; Mayhew, Tsai, and Roth 2017). Yet, this might bring in too much noise due to sense ambiguity and word order differences. Differently, most approaches based on direct transfer leverage language-independent features to train a model on the source language and then directly apply it on target languages. Cross-lingual word embeddings are the most widely used ones of such features (Ni, Dinu, and Florian 2017; Devlin et al. 2019), while other approaches also introduces word clusters (Täckström, McDonald, and Uszkoreit 2012) and Wikifier (Tsai, Mayhew, and Roth 2016) as cross-lingual features.

In this paper, we use a contextual cross-lingual word embedding (Devlin et al. 2019) as the language-independent feature. Rather than directly transferring from the source-learned model to target, we propose to fine-tune the model by converting the minimal-resource cross-lingual transfer problem into a low-resource learning problem, and furthermore, present an enhanced meta-learning algorithm to tackle it. To our best knowledge, we are the first to extend the idea of meta-learning to cross-lingual NER with minimal resources.

Meta-Learning

Meta-learning has a long history (Naik and Mammone 1992) and emerged recently as a way to fast adapt to new tasks with

very limited data. It has been applied to various tasks such as image classification (Koch, Zemel, and Salakhutdinov 2015; Ravi and Larochelle 2017), neural machine translation (Gu et al. 2018), text generation (Huang et al. 2018; Qian and Yu 2019), and reinforcement learning (Finn, Abbeel, and Levine 2017; Li et al. 2018)

There are three categories of meta-learning algorithms: learning a metric space which can be used to compare low-resource examples with rich-resource examples (Vinyals et al. 2016; Sung et al. 2018), learning an optimizer to update the parameters of a model (Andrychowicz et al. 2016; Chen et al. 2018), and learning a good parameter initialization of a model (Finn, Abbeel, and Levine 2017; Mi et al. 2019).

Our approach falls into the last category. We extend the idea of model-agnostic meta-learning (MAML) (Finn, Abbeel, and Levine 2017) to the cross-lingual NER with minimal resources by constructing multiple pseudo-meta-NER tasks. Furthermore, we employ a masking scheme and enhance the loss function with an additional maximum item during meta-training to improve the model’s ability to transfer across languages.

Methodology

Named Entity Recognition is proposed as a sequence labeling problem. Given a sequence with L tokens $\mathbf{x} = \{x_i\}_{i=1}^L$, an NER system is expected to produce a label sequence $\mathbf{y} = \{y_i\}_{i=1}^L$, where x_i is the i -th token and y_i is the corresponding label of x_i . Denote the labeled training data of a source language as D_{train}^S and the test data of a target language as D_{test}^T . Minimal-resource cross-lingual NER aims to train a model \mathcal{M} with D_{train}^S and it is expected that the model will perform well on D_{test}^T .

Base Model

In this section, we give a brief introduction to multilingual BERT (Devlin et al. 2019) (mBERT), which we leverage as the base model in our approach, since it produces an effective cross-lingual word representation. To ease the explanation, we start with BERT (Devlin et al. 2019) here.

BERT is a language model learned with the Transformer encoder (Vaswani et al. 2017). It reads the input sequence at once and learns via two strategies, i.e., masked language modeling and next sentence prediction.

mBERT follows the same model architecture and training procedure as BERT except that it is pre-trained on concatenated Wikipedia data of 104 languages. For tokenization, mBERT utilizes WordPiece embeddings (Wu et al. 2016) with a 110k shared vocabulary to facilitate embedding space alignment across different languages.

Following (Devlin et al. 2019) and (Wu and Dredze 2019), we address cross-lingual NER by adding a linear classification layer with softmax upon the pre-trained mBERT, which can be formulated as:

$$\mathbf{h} = \text{mBERT}(\mathbf{x}), \quad (1)$$

$$\hat{y}_l = \text{softmax}(\mathbf{W}\mathbf{h}_l + \mathbf{b}), \quad (2)$$

where \mathbf{x} is structured as $\{x_0, x_1, \dots, x_L, x_{L+1}\}$. $x_0 = [\text{CLS}]$ and $x_{L+1} = [\text{SEP}]$ are two special tokens as in (Devlin et al. 2019). $\mathbf{h} = \{\mathbf{h}_l\}_{l=1}^L$ and \mathbf{h}_l denotes the output of the pre-trained mBERT that corresponds to the input token x_l . \hat{y}_l denotes the predicted probability distribution for x_l . \mathbf{W} and \mathbf{b} are trainable parameters.

The learning loss *w.r.t.* \mathbf{x} is modeled as the cross-entropy of the predicted label distribution and the ground-truth one for each token:

$$\mathcal{L}(\theta) = -\frac{1}{L} \sum_{l=1}^L \text{CrossEntropy}(y_l, \hat{y}_l) \quad (3)$$

where y_l is a one-hot vector of the ground-truth label for the l -th input token x_l . And the total loss for learning is the summation of losses on all training examples. It should be noted that, if a word is split into several subwords after tokenization, only the label of the first subword is considered.

Enhanced Meta-Learning for Cross-Lingual NER with Minimal Resources

In this section, we elaborate on the proposed approach. First, we clarify how to construct multiple pseudo-meta-NER tasks with the labeled data of the source language. Then, we describe the meta-training algorithm of our approach. Next, we illustrate the proposed masking mechanism and the augmented loss involved in the meta-training phase. Finally, we show how to adapt the meta-learned model to test examples of target languages. The whole procedure of our algorithm is summarized in Algorithm 1.

Pseudo-Meta-NER Tasks In a typical meta-learning scenario, a model is trained on a set of tasks in the meta-training phase, such that the trained model can quickly adapt to new tasks using only a small number of examples. Thus to tackle the minimal-resource cross-lingual NER via meta-learning, we first construct a set of pseudo-meta-NER tasks using the labeled data of the source language.

Assuming there are N examples in $D_{train}^S = \{\mathbf{x}^{(i)}\}_{i=1}^N$. We take each $\mathbf{x}^{(i)}$ as the test set $D_{test}^{\mathcal{T}_i}$ of an individual meta task \mathcal{T}_i , and create a pseudo training set $D_{train}^{\mathcal{T}_i}$ for it by retrieving the most similar examples of $\mathbf{x}^{(i)}$ from D_{train}^S . The pseudo-meta-NER tasks \mathcal{T}_i can be denoted as:

$$\mathcal{T}_i = (D_{train}^{\mathcal{T}_i}, D_{test}^{\mathcal{T}_i}), i \in 1, 2, \dots, N. \quad (4)$$

Specifically, we first compute the sentence representation $r^{(i)}$ for each $\mathbf{x}^{(i)}$, $i \in \{1, 2, \dots, N\}$:

$$r^{(i)} = f(\mathbf{x}^{(i)}) \quad (5)$$

where $f(\cdot)$ could be any function that is able to produce cross-lingual sentence representations. Here, we employ the multilingual BERT (Devlin et al. 2019) and use the the final hidden vector corresponding to the first input token ([CLS]) as the sentence representation.

Then, we construct $D_{train}^{\mathcal{T}_i}$ by selecting top-K similar examples from $D_{train}^S \setminus \mathbf{x}^{(i)}$. The metric used to measure the similarity between $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(m)}$ is:

$$s(\mathbf{x}^{(i)}, \mathbf{x}^{(m)}) = \frac{r^{(i)} \cdot r^{(m)}}{\|r^{(i)}\| \|r^{(m)}\|}, \quad (6)$$

Algorithm 1 Enhanced Meta-Learning for Cross-Lingual NER with Minimal Resources

```

1: procedure META-TRAINING( $D_{train}^S, \alpha, \beta$ )
2:   Construct  $\mathcal{T} = \{\mathcal{T}_i\}$  with  $D_{train}^S$ .
3:   Initialize with the pre-trained base model  $\mathcal{M}_\theta$ .
4:   while not done do
5:     Sample a batch of source tasks  $\mathcal{T}_i$  from  $\mathcal{T}$ .
6:     for all  $\mathcal{T}_i$  do
7:       Update  $\theta'_i = U^n(\theta; \alpha)$ .
8:       Compute  $g_i = \nabla_{\theta'_i} \mathcal{L}_{D_{test}^{\mathcal{T}_i}}(\theta'_i)$ .
9:     end for
10:    Update  $\theta \leftarrow \theta - \beta \sum_i g_i$ .
11:  end while
12:  return  $\mathcal{M}_{\theta^*}$  with  $\theta^*$  being the final updated  $\theta$ 
13: end procedure

14: procedure ADAPTATION( $\mathcal{M}_{\theta^*}, D_{train}^S, D_{test}^T, \gamma$ )
15:   for all  $\mathbf{x}^{(j)} \in D_{test}^T$  do
16:      $D_{test}^{\mathcal{T}_j} = \mathbf{x}^{(j)}$  and construct  $D_{train}^{\mathcal{T}_j}$  with  $D_{train}^S$ .
17:     Update  $\hat{\theta}_j = \theta^* - \gamma \nabla_{\theta^*} \mathcal{L}_{D_{train}^{\mathcal{T}_j}}$ .
18:     Label  $D_{test}^{\mathcal{T}_j}$ , i.e.,  $\mathbf{x}^{(j)}$ , using  $\mathcal{M}_{\hat{\theta}_j}$ .
19:   end for
20: end procedure

```

where $m \in \{1, 2, \dots, N\}$ and $m \neq i$.

Meta-Training In the meta-training phase, we train a model \mathcal{M} by repeatedly simulating the adaptation phase, where the meta-trained model is fine-tuned with a minimal amount of training data of a new task and then tested on the test data.

Specifically, given the created pseudo-meta-NER tasks $\{\mathcal{T}_i\}_{i=1}^N$ and a model \mathcal{M}_θ parameterized by θ , we first randomly sample a task \mathcal{T}_i to derive new model parameters θ' via n gradient updates on the original model parameters θ , which we refer to as inner-update:

$$\theta'_i = U^n(\theta; \alpha) \quad (7)$$

where U^n is the operator that performs gradient descent n times with the learning rate α to minimize the loss $\mathcal{L}_{D_{train}^{\mathcal{T}_i}}$ computed on $D_{train}^{\mathcal{T}_i}$. For example, when applying a single gradient update,

$$\theta'_i = \theta - \alpha \nabla_{\theta} \mathcal{L}_{D_{train}^{\mathcal{T}_i}}(\theta) \quad (8)$$

We then evaluate the updated parameters θ'_i on $D_{test}^{\mathcal{T}_i}$ and further update the meta model \mathcal{M}_θ by minimizing the loss $\mathcal{L}_{D_{test}^{\mathcal{T}_i}}(\theta'_i)$ with respect to θ , which is referred to as meta-update. When aggregating multiple pseudo-meta-NER tasks, the meta-objective is:

$$\min_{\theta} \sum_i \mathcal{L}_{D_{test}^{\mathcal{T}_i}}(\theta'_i) \quad (9)$$

Take a single gradient update with the learning rate β , the

meta-update can be formulated as:

$$\begin{aligned} \theta &\leftarrow \theta - \beta \sum_i \nabla_{\theta} \mathcal{L}_{D_{test}^{\mathcal{T}_i}}(\theta'_i) \\ &= \theta - \beta \sum_i g_i \end{aligned} \quad (10)$$

where g_i is the meta-gradient on task \mathcal{T}_i , which can be expanded to:

$$g_i = \nabla_{\theta} \mathcal{L}_{D_{test}^{\mathcal{T}_i}}(\theta'_i) = \nabla_{\theta'_i} \mathcal{L}_{D_{test}^{\mathcal{T}_i}}(\theta'_i) \nabla_{\theta}(\theta'_i) \quad (11)$$

In Equation 11, $\nabla_{\theta}(\theta'_i)$ is the Jacobian matrix of the update operation U^n that will introduce higher order gradient. To reduce computational cost, we use a first-order approximation by replacing the Jacobian $\nabla_{\theta}(\theta'_i)$ with the identity matrix as in (Finn, Abbeel, and Levine 2017). Therefore, g_i can be computed as:

$$g_i = \nabla_{\theta'_i} \mathcal{L}_{D_{test}^{\mathcal{T}_i}}(\theta'_i) \quad (12)$$

Compared with the common training scheme, the meta-learned model is more sensitive to the changes among different tasks, which can promote the learning of the common internal representations rather than the distinctive features of the source language training data D_{train}^S . When coming to the adaptation phase, the model could be more sensitive to the features of new tasks, and hence only one or a few fine-tune epochs on a minimal amount of data can make rapid progress without overfitting (Finn, Abbeel, and Levine 2017).

Masking on Named Entities For cross-lingual NER with minimal resources, the alignments of the entity representations in the shared space are particularly important as this task focuses on understanding entities across languages. However, compared with commonly used words, most entities are of low frequency in the pre-training corpora of the base model. As a result, the learned entity representations across languages are not well-aligned in the shared space.

In order to reduce the dependence on target entity representations and encourage the model to predict through context information, we employ the [MASK] token as introduced in (Devlin et al. 2019) to mask entities at the token level in each training example, i.e., each token inside an entity is randomly masked with a given probability. Then, the masked examples are fed as input data for the model. Note that we re-perform the masking scheme at the beginning of each training epoch.

Max Loss In Equation 3, the loss for each token is uniformly weighted so that all tokens contribute equally when training the model. Nonetheless, this will result in insufficient learning for those tokens with relatively higher losses. In order to force the model to put more effort in learning from such tokens, we modify the loss function as:

$$\begin{aligned} \mathcal{L}(\theta) &= -\frac{1}{L} \sum_{i=1}^L \text{CrossEntropy}(y_i, \hat{y}_i) \\ &\quad - \lambda \max_{i \in \{1, 2, \dots, L\}} \text{CrossEntropy}(y_i, \hat{y}_i) \end{aligned} \quad (13)$$

where $\lambda \geq 0$ is a weighting factor. In this way, the potential mispredictions of the high-loss tokens would probably be corrected during meta-training. The benefit of such correction is that the meta-knowledge about the mispredictions, which is also going to be transferred to target tasks, would be reduced, so that the model could achieve better performance after transferring.

In summary, the $\mathcal{L}_{D_{train}^{\tau_i}}$ and $\mathcal{L}_{D_{test}^{\tau_i}}$ in Meta-Training of Algorithm 1 are with the masking scheme and the max loss.

Adaptation When it comes to the adaptation phase, *i.e.*, applying \mathcal{M}_{θ^*} on target languages, we take each test example $\mathbf{x}^{(j)} \in D_{test}^T$ as the test set $D_{test}^{\tau_j}$ of a target task τ_j . We then construct a pseudo training set $D_{train}^{\tau_j}$ for each τ_j by retrieving top-K similar examples of $\mathbf{x}^{(j)}$ from the source language training data D_{train}^S using the metric in Equation 6. Subsequently, we fine-tune the meta-learned model \mathcal{M}_{θ^*} with the pseudo training set $D_{train}^{\tau_j}$ as in Equation 3 via *one gradient update*, and then use the fine-tuned model to predict labels for the test set $D_{test}^{\tau_j}$, *i.e.*, $\mathbf{x}^{(j)}$.

It should be noted that in the adaptation phase, we do not perform the masking scheme to avoid information loss of target entities. Besides, since the size of the pseudo training set is very small, we employ the loss function as in Equation 3 rather than Equation 13 to prevent over-adjusting on uncertain or mispredicted tokens. In fact, when using Equation 13 for adaptation, the model could achieve slightly better performance in some cases but also get worse performance in others due to the mentioned over-adjustment.

Experiments

In this section, we evaluate our enhanced meta-learning approach for cross-lingual NER with minimal resources and compare our approach to current state-of-the-art methods.

Datasets

We conduct experiments on four benchmark datasets: CoNLL-2002 Spanish and Dutch NER (Tjong Kim Sang 2002), CoNLL-2003 English and German NER (Tjong Kim Sang and De Meulder 2003), Europeana Newspapers French NER (Neudecker 2016), and MSRA Chinese NER (Cao et al. 2018). Table 2 shows the statistics of all datasets.

- **CoNLL-2002/2003** is annotated with four entity types: PER, LOC, ORG, and MISC. All datasets are split into a training set, a development set (testa) and a test set (testb).
- **Europeana Newspapers** is annotated with three types: PER, LOC, and ORG. We randomly sample 10% of sentences from the whole data to build a test set.
- **MSRA** is also annotated with three types: PER, LOC, and ORG. Since gold word segmentation is not provided in the test set, we use word segmentation from (Zhang and Yang 2018).

For all experiments, we use English as the source language and the others as target languages, *i.e.*, the model \mathcal{M} is trained on the training set of English data and evaluated on the test sets of each other language. When transferring to

Language	Type	Train	Dev	Test
English-en (CoNLL-2003)	Sentence	14,987	3,466	3,684
	Entity	23,499	5,942	5,648
German-de (CoNLL-2003)	Sentence	12,705	3,068	3,160
	Entity	11,851	4,833	3,673
Spanish-es (CoNLL-2002)	Sentence	8,323	1,915	1,517
	Entity	18,798	4,351	3,558
Dutch-nl (CoNLL-2002)	Sentence	15,806	2,895	5,195
	Entity	13,344	2,616	3,941
French-fr (Europeana News)	Sentence	9,527	-	2,375
	Entity	7,899	-	1,984
Chinese-zh (MSRA)	Sentence	46,306	-	4,361
	Entity	72,864	-	5,478

Table 2: Dataset statistics.

French and Chinese, we relabel the MISC entities in English training data into non-entities for meta-training as there is no MISC in the French and Chinese test sets. Following (Wu and Dredze 2019), we use the BIO labeling scheme.

Implementation Details

We implement our approach with PyTorch 1.0.1. We use the cased multilingual BERT_{BASE} with 12 Transformer blocks, 768 hidden units, 12 self-attention heads, GELU activations (Hendrycks and Gimpel 2016), a dropout rate of 0.1 and learned positional embeddings. We employ WordPiece embeddings (Wu et al. 2016) to split a word into subwords, which are then directly fed into the model without any other pre-processing. We empirically select the hyper-parameters and utilize them in all experiments. Specifically, for sequence length, we employ a sliding window with a maximum length of 128. When the sequence length is larger than 128, the last 64 subwords of the first window are kept as the context for the subsequent window. Following (Huang et al. 2018), we select $K = 2$ similar examples for both pseudo NER task construction and the adaptation phase. The mask ratio is set to 0.2, λ in Equation 13 is set to 2.0, update steps n in Equation 7 is set to 2, the number of sampled pseudo-NER tasks used for one meta-update is set to 32, and the maximum meta-update steps is set to 3×10^3 . Following (Wu and Dredze 2019), we freeze the parameters of the embedding layer and the bottom three layers of the base model. According to the suggestions of model hyper parameters in (Devlin et al. 2019), for the optimizers of both inner-update and meta-update, we use Adam (Kingma and Ba 2015) with learning rate of $\alpha, \beta = 3e - 5$, while for gradient updates during adaptation, we set the learning rate γ to $1e - 5$. Following (Tjong Kim Sang 2002), we use the phrase level F1-score as the evaluation metric. To reduce the model bias, we carry out 5 runs and report the average performance.

Performance Comparison

Table 3 presents our results on transferring from English to five other languages, alongside results from previous works. The results show that our approach significantly outperforms the previous state-of-the-art methods across the board, with relative improvements on F1-score compared to the base model ranging from 1.09% for Dutch to 8.67% for French (with average improvement of 3.21%), which demonstrates

	es	nl	de	fr	zh	Average
Täckström, McDonald, and Uszkoreit (2012)	59.30	58.40	40.40	-	-	-
Tsai, Mayhew, and Roth (2016)	60.55	61.56	48.12	-	-	-
Ni, Dinu, and Florian (2017)	65.10	65.40	58.50	-	-	-
Mayhew, Tsai, and Roth (2017)	65.95	66.50	59.11	-	-	-
Xie et al. (2018)	72.37	71.25	57.76	-	-	-
Wu and Dredze (2019)	74.96	77.57	69.56	-	-	-
Base Model	74.59	79.57	70.79	50.89	76.42	70.45
Ours	76.75	80.44	73.16	55.30	77.89	72.71

Table 3: Results of cross-lingual NER with minimal resources¹.

	es	nl	de	fr	zh	Average
Ours	76.75	80.44	73.16	55.30	77.89	72.71
Ours w/o max loss	76.05 (-0.70)	79.50 (-0.94)	71.84 (-1.32)	52.64 (-2.66)	76.77 (-1.12)	71.36 (-1.35)
Ours w/o masking	75.57 (-1.18)	80.38 (-0.06)	72.76 (-0.40)	54.29 (-1.01)	77.79 (-0.10)	72.16 (-0.55)
Ours w/o max loss/masking	75.33 (-1.42)	80.13 (-0.31)	71.49 (-1.67)	52.79 (-2.51)	76.50 (-1.39)	71.25 (-1.46)
Ours w/o meta-train/max loss/masking (i.e., Base Model)	74.59 (-2.16)	79.57 (-0.87)	70.79 (-2.37)	50.89 (-4.41)	76.42 (-1.47)	70.45 (-2.26)

Table 4: Ablation study on cross-lingual NER with minimal resources.

the effectiveness of the proposed enhanced meta-learning algorithm.

Particularly, compared with the base model, our approach achieves particularly significant improvement on German and French, which can be attributed to our model’s stronger ability to predict through context information. In English, proper nouns of LOCATION, PERSON, *etc.* often begin with a capital letter while most general nouns do not. As a result, without effective extraction of context information, the base model tends to mislabel capitalized terms for general nouns as entities, and such phenomenon is especially serious when adapting the model to French and German, where capitalization rules differ from English for general nouns, titles, *etc.* or due to noise in datasets. In contrast, our approach is more robust in such cases due to the introduction of the masking scheme and the max loss, which facilitates the model to label general nouns as non-entities based more on context.

Ablation Study

We propose several strategies to enhance the base model, including the meta-training and adaptation procedure, the masking scheme, and the augmented loss. In this section, we conduct ablation study experiments to investigate the influence of these factors. Table 4 shows the results.

- *Ours w/o max loss*, which removes the additional maximum term in the loss function. The performance in terms of F1-score decreases by 1.35 on average. We conjecture that, without the maximum term, the meta-knowledge from mispredictions is transferred to target tasks along with the meta-model, which hurts performance.
- *Ours w/o masking*, which wipes out the masking scheme during the meta-training phase. This causes a performance drop across all languages, with a maximum drop

of 1.18 F1-score in Spanish. That further demonstrates the necessity of predicting through contextual information.

- *Ours w/o max loss/masking*, which cuts out both the masking scheme and the max loss at once. In this case, our approach degrades into the base model trained with merely model-agnostic meta-learning. This results in a performance drop of 1.46 F1-score on average, indicating that both the masking scheme and the max loss do bring enhancement to meta-learning for cross-lingual NER with minimal resources.
- *Ours w/o meta-train/max loss/masking*, which further eliminates the meta-training and the adaptation phase from *Ours w/o max loss/masking*. In that case, our approach degenerates into the base model. From Table 4, we can see that this will lead to a significant and consistent performance drop on all five target languages, which demonstrates the effectiveness of meta-learning employed in our approach.

Case Study

We give a case study to analyze the quality of the results produced by our approach and the base model. Table 5 demonstrates that our approach has a stronger ability to transfer semantic information.

In example #1, the base model fails to identify “Secretaría General” as ORG, probably because its most similar phrase “secretary general” in the English dataset is usually labeled as non-entities. However, our approach can recognize it according to the learned semantic information “a PER was selected to replace another PER at the head of an ORG”. Similarly, in example #2, the base model incorrectly labels “Edmond Thieffrylaan” as PER. We suspect that this is because Edmond appears as a part of a person name “Jim Edmond” in the English training data. Surprisingly, the proposed approach labels it as LOC correctly according to the context “clean up the playground on LOC”. Moreover, the baseline

¹The zh results reported in (Wu and Dredze 2019) used a dataset not specified in the paper, so we don’t list them here.

#1	Spanish	Base Model: ... [PER Fidalgo] fue elegido para sustituir a [PER Antonio Gutiérrez] al frente de la [ORG Secretaría General]. Ours: ... [PER Fidalgo] fue elegido para sustituir a [PER Antonio Gutiérrez] al frente de la [ORG Secretaría General]. Translation in English: ... Fidalgo was elected to replace Antonio Gutierrez at the head of the General Secretariat.
#2	Dutch	Base Model: ... hebben politie het speelplein [LOC Redoute] aan de [PER Edmond Thieffrylaan] schoongeveegd van junkies. Ours: ... hebben politie het speelplein [LOC Redoute] aan de [LOC Edmond Thieffrylaan] schoongeveegd van junkies. Translation in English: ... the police have cleaned up the Redoute playground on Edmond Thieffrylaan from junkies.
#3	German	Base Model: Kurzfristig wurde der Vorgänger [ORG Krauses] , [PER Peter Wünsch] , zurückgeholt. Ours: Kurzfristig wurde der Vorgänger [PER Krauses] , [PER Peter Wünsch] , zurückgeholt. Translation in English: Not long ago, Krauses' predecessor, Peter Wünsch, was taken back.
#4	Chinese	Base Model: 在 [LOC 希腊] , 人们常称他为“小 [PER 奥纳西斯] ”。 Ours: 在 [LOC 希腊] , 人们常称他为“小 [PER 奥纳西斯] ”。 Translation in English: In Greece, he is often called "Little Onassis".

Table 5: Case study of cross-lingual NER with minimal resources. The GREEN (RED) highlight indicates a correct (incorrect) label.

data	systems	es	nl	de	fr	zh	Average
1%	CL	60.43	54.77	-	-	-	-
	ML	62.72	63.57	-	-	-	-
	MLMT	68.33	66.73	-	-	-	-
	Base Model	76.83	80.90	73.22	60.51	79.72	74.24
	Ours	78.59	82.72	75.12	61.71	80.34	75.70
2%	CL	66.45	61.91	-	-	-	-
	ML	71.22	70.62	-	-	-	-
	MLMT	72.59	70.92	-	-	-	-
	Base Model	77.28	81.54	74.31	64.43	81.86	75.88
	Ours	79.54	83.07	75.64	65.79	82.58	77.32
5%	CL	69.63	70.29	-	-	-	-
	ML	76.15	76.52	-	-	-	-
	MLMT	77.13	77.03	-	-	-	-
	Base Model	78.74	82.00	75.48	67.73	85.11	77.81
	Ours	80.32	83.72	77.70	69.19	85.67	79.32

Table 6: Low resource cross-lingual NER results, where x% denotes the percentage of labeled training data in target languages used in the adaptation phase.

CL: Cross-lingual transfer using a shared character embedding layer (Yang, Salakhutdinov, and Cohen 2017).

ML: The multi-lingual framework as in Lin et al. (2018).

MLMT: The multi-lingual multi-task framework as in Li et al. (2018).

model mispredicts the labels of “Krauses” in Example #3 and “奥纳西斯” in Example #4, two unseen entities in the English training data, while our approach gives the right prediction on the basis of context information. Considering the limited space, we provide more cases in Table S1 of the supplementary material.

Discussion: Extend to Low-Resource Cross-Lingual NER

Here, we extend the proposed approach to the task of low-resource cross-lingual NER. To simulate a low-resource setting, we use randomly sampled subsets of the training data of a target language. Compared with minimal-resource cross-lingual transfer, we take the same meta-training procedure. For the adaptation phase, we directly use the entire subsets

to fine-tune the meta-learned model for efficiency, and then test on the test data of the target language.

We compare our meta-learning based approach with other multi-lingual and multi-task based approaches. For the results not reported in (Yang, Salakhutdinov, and Cohen 2017) and (Lin et al. 2018), we re-implement their methods based on the open-source github repositories^{2,3,4}. As presented in Table 6, our approach significantly outperforms other approaches across all target languages with different percentage of labeled data. Compared with the base model, there is an average improvement of 1.47 F1-score. We also study the factor analysis of the enhanced meta-learning algorithm under low-resource setting. One can refer to Table S2 of the supplementary material for details due to the limited space. Similarly, removing any factor in our proposed approach will lead to a performance drop, which further demonstrates that our approach is reasonable.

Conclusion

In this paper, we propose an enhanced meta-learning algorithm for cross-lingual NER with minimal resources, considering that the model could achieve better results after a few fine-tuning steps over a very limited set of structurally/semantically similar examples from the source language. To this end, we propose to construct multiple pseudo-NER tasks for meta-training by computing sentence similarities. Moreover, in order to improve the model’s capability to transfer across different languages, we present a masking scheme and augment the loss function with an additional maximum term during meta-training. Experiments on five target languages show that the proposed approach leads to new state-of-the-art results with a relative F1-score improvement of up to 8.76%. We also extend the approach to low-resource cross-lingual NER, and it also achieves state-of-the-art results.

²<https://github.com/kimiyoung/transfer>

³<https://github.com/limteng-rpi/mlmt>

⁴We re-implement only Spanish and Dutch as the original repositories only provide aligned word embeddings for these two languages.

References

- Andrychowicz, M.; Denil, M.; Gomez, S.; Hoffman, M. W.; Pfau, D.; Schaul, T.; Shillingford, B.; and De Freitas, N. 2016. Learning to learn by gradient descent by gradient descent. In *NeurIPS*, 3981–3989.
- Cao, P.; Chen, Y.; Liu, K.; Zhao, J.; and Liu, S. 2018. Adversarial transfer learning for Chinese named entity recognition with self-attention mechanism. In *EMNLP*, 182–192.
- Chen, J.; Qiu, X.; Liu, P.; and Huang, X. 2018. Meta multi-task learning for sequence modeling. In *AAAI*.
- Chiu, J. P., and Nichols, E. 2016. Named entity recognition with bidirectional LSTM-CNNs. *TACL* 4:357–370.
- Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 4171–4186.
- Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 1126–1135.
- Gu, J.; Wang, Y.; Chen, Y.; Cho, K.; and Li, V. O. 2018. Meta-learning for low-resource neural machine translation. In *EMNLP*.
- Hendrycks, D., and Gimpel, K. 2016. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *CoRR* abs/1606.08415.
- Huang, P.-S.; Wang, C.; Singh, R.; Yih, W.-t.; and He, X. 2018. Natural language to structured query generation via meta-learning. In *NAACL*, 732–738.
- Huang, Z.; Xu, W.; and Yu, K. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.
- Kingma, D. P., and Ba, J. 2015. Adam: A method for stochastic optimization. In *ICLR*.
- Koch, G.; Zemel, R.; and Salakhutdinov, R. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*.
- Lample, G., and Conneau, A. 2019. Cross-lingual language model pretraining. *CoRR* abs/1901.07291.
- Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; and Dyer, C. 2016. Neural architectures for named entity recognition. In *NAACL*, 260–270.
- Li, D.; Yang, Y.; Song, Y.-Z.; and Hospedales, T. M. 2018. Learning to generalize: Meta-learning for domain generalization. In *AAAI*.
- Lin, Y.; Yang, S.; Stoyanov, V.; and Ji, H. 2018. A multi-lingual multi-task architecture for low-resource sequence labeling. In *ACL*, 799–809.
- Ma, X., and Hovy, E. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *ACL*, 1064–1074.
- Mayhew, S.; Tsai, C.-T.; and Roth, D. 2017. Cheap translation for cross-lingual named entity recognition. In *EMNLP*, 2536–2545.
- Mi, F.; Huang, M.; Zhang, J.; and Faltings, B. 2019. Meta-learning for low-resource natural language generation in task-oriented dialogue systems. In *IJCAI*.
- Naik, D. K., and Mammone, R. 1992. Meta-neural networks that learn by learning. In *IJCNN*, 437–442.
- Neudecker, C. 2016. An open corpus for named entity recognition in historic newspapers. In *LREC*.
- Ni, J.; Dinu, G.; and Florian, R. 2017. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In *ACL*, 1470–1480.
- Peters, M.; Ammar, W.; Bhagavatula, C.; and Power, R. 2017. Semi-supervised sequence tagging with bidirectional language models. In *ACL*, 1756–1765.
- Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. In *NAACL*, 2227–2237.
- Qian, K., and Yu, Z. 2019. Domain adaptive dialog generation via meta learning. In *ACL*.
- Ravi, S., and Larochelle, H. 2017. Optimization as a model for few-shot learning. In *ICLR*.
- Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; and Hospedales, T. M. 2018. Learning to compare: Relation network for few-shot learning. In *CVPR*, 1199–1208.
- Täckström, O.; McDonald, R.; and Uszkoreit, J. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *NAACL*, 477–487.
- Tjong Kim Sang, E. F., and De Meulder, F. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *CoNLL*, 142–147.
- Tjong Kim Sang, E. F. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *CoNLL*.
- Tsai, C.-T.; Mayhew, S.; and Roth, D. 2016. Cross-lingual named entity recognition via wikification. In *CoNLL*, 219–228.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*, 5998–6008.
- Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D.; et al. 2016. Matching networks for one shot learning. In *NeurIPS*, 3630–3638.
- Wu, S., and Dredze, M. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. *CoRR* abs/1904.09077.
- Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. 2016. Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.
- Xie, J.; Yang, Z.; Neubig, G.; Smith, N. A.; and Carbonell, J. 2018. Neural cross-lingual named entity recognition with minimal resources. In *EMNLP*, 369–379.
- Yang, Z.; Salakhutdinov, R.; and Cohen, W. W. 2017. Transfer learning for sequence tagging with hierarchical recurrent networks. In *ICLR*.
- Zhang, Y., and Yang, J. 2018. Chinese NER using lattice LSTM. In *ACL*, 1554–1564.