MetaTS: Meta Teacher-Student Network for Multilingual Sequence Labeling with Minimal Supervision

Zheng Li[1], **Danqing Zhang**[1], **Tianyu Cao**[1], **Ying Wei**[2], **Yiwei Song**[1], **Bing Yin**[1]
[1]Amazon.com Inc, CA, USA
[2]City University of Hong Kong, Hong Kong, China
[1]{amzzhe,danqinz,caoty,ywsong,alexbyin}@amazon.com
[2]yingwei@cityu.edu.hk

Abstract

Sequence labeling aims to predict a fine-grained sequence of labels for the text. However, such formulation hinders the effectiveness of supervised methods due to the lack of token-level annotated data. This is exacerbated when we meet a diverse range of languages. In this work, we explore multilingual sequence labeling with minimal supervision using a single unified model for multiple languages. Specifically, we propose a Meta Teacher-Student (MetaTS) Network, a novel meta learning method to alleviate data scarcity by leveraging large multilingual unlabeled data. Prior teacher-student frameworks of self-training rely on rigid teaching strategies, which may hardly produce high-quality pseudo-labels for consecutive and interdependent tokens. On the contrary, MetaTS allows the teacher to dynamically adapt its pseudo-annotation strategies by the student's feedback on the generated pseudo-labeled data of each language and thus mitigate error propagation from noisy pseudo-labels. Extensive experiments on both public and real-world multilingual sequence labeling datasets empirically demonstrate the effectiveness of MetaTS[1].

1 Introduction

Sequence labeling or tagging is the task of detecting the boundary of all occurring entity mentions from unstructured text and classifying them into predefined types, such as Named Entity Recognition (NER) (Chiu and Nichols, 2016), Aspect-Based Sentiment Analysis (ABSA) (Mitchell et al., 2013), etc. An entity mention should be a single word or a sequence of words that contain key information, such as a person, location, or institution. In the E-commerce search domain, we need to recognize product attributes from short queries, such as product type, brand, size, to better understand users' preferences and intents.

[1]Our code is open-source and available at https://github.com/amzn/x-metats

1. Ground-truth Labels
[Mackie] [profx6v3] [6-channel] [mixer] [with] [usb]
2. Pseudo-Labels (Choice #1)
[Mackie] [profx6v3 6-channel]✗ [mixer with usb]✗
3. Pseudo-Labels (Choice #2)
[Mackie] [profx6v3] [6-channel] [mixer] [with usb]✗

Table 1: Ground-truth labels and noisy pseudo-labels for an English query NER example. We use colors to denote the entity type and use brackets to indicate the entity boundary. Entity labels: Brand, ProductLine, Size, ProductType, NonContent, Misc.

Despite recent advances in deep learning models for sequence labeling (Huang et al., 2015; Raganato et al., 2017), they still rely on massive labeled data. Nonetheless, the sequence labeling tasks usually lie in the low-data regime due to costly and labor-intensive human annotation for token-level labels, especially for a variety of languages (Xie et al., 2018), as search engines or social networks usually cover a diverse range of countries and locales using different languages. In this paper, we attempt to explore a unified multilingual sequence labeling model with minimal supervision, where each language only has limited labeled data.

The emergence of multilingual pre-trained language models (mPLMs) such as mBERT (Devlin et al., 2019) and XLM (Conneau and Lample, 2019) have enabled breakthroughs on various multilingual NLP tasks. However, it has been recently noted that mPLMs are not data-efficient and typically require sufficient fine-tuning data for superior performance on downstream tasks. To mitigate data scarcity, Semi-Supervised Learning (SSL) (Chapelle et al., 2009) has been a promising paradigm that allows us to take advantage of large-scale unlabeled multilingual data. Self-training (Scudder, 1965) stands out among the SSL approaches, in which a teacher model produces pseudo-labels for unlabeled examples, and a student model learns from these examples with generated pseudo-labels. Self-training has shown

promising results in instance-level tasks, e.g., image classification (Tarvainen and Valpola, 2017; Xie et al., 2020b). However, a major research challenge that dictates the success of self-training is the well-known confirmation bias problem (Arazo et al., 2020), which results in progressive drifts on the noisy pseudo-labeled data provided by the teacher. This problem is more pronounced in sequence labeling (Ruder and Plank, 2018), as complicated dependencies between tokens pose tremendous challenges towards the rigid teaching strategies, e.g., the fixed teacher (Lee et al., 2013) or the periodically synchronizing teacher (Liang et al., 2020), to generate accurate pseudo-labels for consecutive and interdependent tokens.

To encourage the teacher to generate better pseudo-labels for multilingual sequence labeling, we propose a novel Meta Teacher-Student (MetaTS) network, where the teacher learns dynamically and continuously from the student's feedback to adapt its teaching strategies, i.e., the pseudo-annotation choices. Concretely, given a language for each step, the student network will be updated based on the pseudo-labeled data produced by the teacher. To quantitatively measure how well the teacher generates these pseudo-labels at the current step, we will evaluate the difference between the student performance after the update using the pseudo-labeled data of the language and that before the update. The improvement or degradation of the student performance can be used as the feedback to meta-optimize the teacher network (a.k.a. learning to learn (Finn et al., 2017)).

Consider an example in Table 1. Pseudo-labels (choice #2) are closer to the ground-truth labels of the sentence than pseudo-labels (choice #1). Better pseudo-annotation strategies by the teacher lead to more accurate pseudo-labels (e.g., choice #2 in Table 1), thus boosting the student's performance on the labeled data. As such, the proposed MetaTS method learns to teach the student with better token-level pseudo-labels and alleviates the serious confirmation bias problem in sequence labeling. Empirically, extensive experiments on both the public multilingual Open-domain NER dataset (Tjong Kim Sang, 2002a,b), multilingual E2E-ABSA challenge benchmark (Pontiki et al., 2014) and a real-world large-scale multilingual E-commerce NER dataset have demonstrated the effectiveness of the MetaTS method.

Overall, our contributions can be summarized as follows: (1) we explore a unified and effective multilingual sequence labeling setting with minimal supervision required; (2) we propose a novel MetaTS framework to alleviate the confirmation bias problem via learning from the student's feedback to generate better fine-grained pseudo-labels; (3) we conduct extensive experiments that verify the effectiveness of MetaTS.

2 Preliminaries

2.1 Sequence Labeling (SL)

Sequence labeling is the process of *identifying* (boundary) and *categorizing* (type) entities in text into a predefined entity set C. Formally, given a sentence $\mathbf{X}=[x_1, x_2, ..., x_N]$ with N tokens, the goal is to predict a tag sequence $\mathbf{Y}=[y_1, y_2, ..., y_N]$, where $y_n \in C$ $(n \in [1, N])$. Based on the `BIO` schema (Li et al., 2012), the first token of an entity mention with type `X` is labeled as `B-X`; the remaining tokens inside that entity mention are labeled as `I-X`; and the non-entity tokens are labeled as `O`.

Low-Resource Multilingual SL Suppose that there are R languages $\mathbf{L} = [l_1, l_2, .., l_R]$. For each language l_i, there are only a small amount of labeled data $\{(\mathbf{X}_m^{l_i}, \mathbf{Y}_m^{l_i})\}_{m=1}^{M^{l_i}}$ and large unlabeled data $\{\widetilde{\mathbf{X}}_m^{l_i}\}_{m=1}^{\widetilde{M}^{l_i}}$, where $M^{l_i} \ll \widetilde{M}^{l_i}$. Our goal aims to train a unified supervised multilingual model that can achieve better performance on all languages in the low-resource setting.

2.2 Multilingual Pre-trained Language Model (mPLM)

The emergence of mPLMs, such as mBERT (Devlin et al., 2019), XLM (Conneau and Lample, 2019) and mUnicoder (Yang et al., 2020), has led to significant performance gains on various multilingual NLP tasks (Hu et al., 2020). mPLMs leverage self-supervised learning on a large-scale multilingual unlabeled corpus, which treats shared word piece tokens as the anchor across languages to produce weakly-aligned multilingual representations. These multilingual contextualized embeddings are versatile and can substantially benefit downstream tasks. However, mPLMs are trained on open-domain data and lack adaptivity to a specific domain in the low-data regime (Huang et al., 2019). Thus, it is critical to exploit enormous unlabeled data for the downstream tasks to achieve task-aware adaptation.

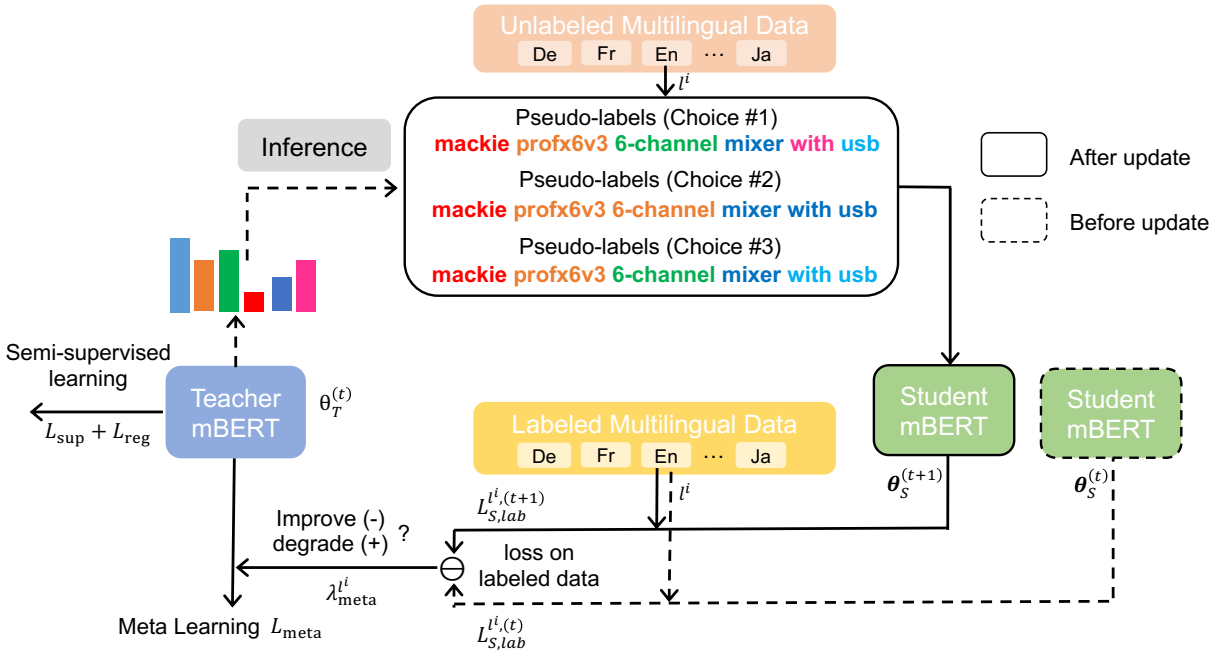

Figure 1: The framework of the Meta Teacher-Student Network (MetaTS).

2.3 Teacher-Student Network

The teacher-student (TS) network is a classic architecture widely used in self-training (Scudder, 1965), where the student model has a similar or higher capacity than the teacher, and knowledge distillation (Hinton et al., 2015) where the student model is smaller than the teacher. Mathematically, let T and S respectively be the teacher and student network, parameterized by $\boldsymbol{\theta}_T$ and $\boldsymbol{\theta}_S$. We use $f(\mathbf{X};\boldsymbol{\theta}_T)$ and $f(\mathbf{X};\boldsymbol{\theta}_S)$ denote the entity label predictions of the sentence $\mathbf{X}$ by the teacher and student, respectively. Then the knowledge transfer is usually achieved by minimizing the loss between the predictions from the teacher and student:

$$\mathcal{L}(f(\mathbf{X};\boldsymbol{\theta}_T), f(\mathbf{X};\boldsymbol{\theta}_S)), \tag{1}$$

where $f(\mathbf{X};\boldsymbol{\theta}_T)$ can be a soft target or converted to a hard target as the pseudo-labels. $\mathcal{L}$ is the transfer loss to enforce the consistency between the teacher and the student probability distributions, such as Cross-Entropy (CE) loss, Kullback-Leibler (KL) divergence loss, or Mean Square Error (MSE).

3 Method

3.1 Meta Teacher-Student Network

Inspired by the teacher-student interaction mechanism, we propose a meta teacher-student (MetaTS) network for low-resource multilingual sequence labeling. Our ultimate goal lies in learning from large-scale multilingual unlabeled data based on pseudo-labels to mitigate the shortage of labeled data for token-level classification. The framework of MetaTS is illustrated in Figure 1.

3.2 Student Network

Given a language l_i, recall that there are limited labeled data $\{(\mathbf{X}_m^{l_i}, \mathbf{Y}_m^{l_i})\}_{m=1}^{M^{l_i}}$ and large unlabeled data $\{\widetilde{\mathbf{X}}_m^{l_i}\}_{m=1}^{\widetilde{M}^{l_i}}$. The student network learns the distilled knowledge of unlabeled data from the teacher, which behaves as the teacher's predictions on unlabeled sequences $\{\widetilde{\mathbf{X}}_m^{l_i} = [\widetilde{x}_{m,1}^{l_i}, \widetilde{x}_{m,2}^{l_i}, ..., \widetilde{x}_{m,N}^{l_i}]\}_{m=1}^{\widetilde{M}^{l_i}}$. At the t-th iteration, the teacher model generates hard pseudo-labels $\{\widetilde{\mathbf{Y}}_m^{l_i,(t)} = [\widetilde{y}_{m,1}^{l_i,(t)}, \widetilde{y}_{m,2}^{l_i,(t)}, ..., \widetilde{y}_{m,N}^{l_i,(t)}]\}_{m=1}^{\widetilde{M}^{l_i}}$ by

$$\widetilde{y}_{m,n}^{l_i,(t)} = \arg\max_c f_{n,c}(\widetilde{\mathbf{X}}_m^{l_i}; \boldsymbol{\theta}_T^{(t)}), \tag{2}$$

where $f_{n,c}$ denotes the probability of the n-th token belonging to the c-th class and $c \in C$. $\boldsymbol{\theta}_T^{(t)}$ is the teacher's parameters at the t-the step. Then we achieve the knowledge transfer of Eq. (1) by minimizing the student's loss $\mathcal{L}_{\text{S}}$ on these hard pseudo-labels

$$\boldsymbol{\theta}_S^{(t+1)} = \arg\min_{\boldsymbol{\theta}} \frac{1}{\widetilde{M}^{l_i}} \sum_{m=1}^{\widetilde{M}^{l_i}} \ell(\widetilde{\mathbf{Y}}_m^{l_i,(t)}, f(\widetilde{\mathbf{X}}_m^{l_i}; \boldsymbol{\theta}_S^{(t)})), \tag{3}$$

where ℓ is the cross-entropy loss. $\boldsymbol{\theta}_S^{(t)}$ and $\boldsymbol{\theta}_S^{(t+1)}$ are the parameters of the student before and after the update at the t step, which will be used for the meta-learning of the teacher in the next section.

3.3 Teacher Network

The teacher network is jointly optimized by three objectives: a supervised learning loss $\mathcal{L}_{\text{sup}}$, a semi-supervised regularization loss $\mathcal{L}_{\text{reg}}$, and a meta-learning loss $\mathcal{L}_{\text{meta}}$, i.e.,

$$\mathcal{L}_{\text{T}} = \mathcal{L}_{\text{sup}} + \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{meta}},$$

Supervised learning The supervised loss $\mathcal{L}_{\text{sup}}$ on the labeled data is defined as

$$\mathcal{L}_{\text{sup}} = \frac{1}{M^{l_i}} \sum_{m=1}^{M^{l_i}} \ell(\mathbf{Y}_m^{l_i}, f(\mathbf{X}_m^{l_i}; \boldsymbol{\theta}_T^{(t)})). \quad (4)$$

Semi-supervised regularization The regularization loss $\mathcal{L}_{\text{reg}}$ alleviates the overfitting of the teacher to limited labeled data by enforcing the prediction consistency between the original and augmented unlabeled samples (Xie et al., 2020a). However, in the text domain, 1) data augmentation techniques are much more difficult to maintain the original word or sentence semantics compared with those in the vision domain; 2) external text augmentations are very tedious and usually unavailable for multilingual corpus, especially for low-resource languages. Thus, we do not explicitly augment the sentence but instead propose to add random Gaussian noises $G(\mathbf{0}, \boldsymbol{\sigma}^2)$ to the BERT embedding of each token to increase the diversity of the sentence. We name it as **virtual data augmentation**. Let $\mathbf{z}_{m,n} \in \mathbb{R}^{|C|}$ denote the soft prediction $f_n(\widetilde{\mathbf{X}}_m^{l_i}; \boldsymbol{\theta}_T^{(t)})$ of the teacher on the n-th token of $\widetilde{\mathbf{X}}_m^{l_i}$. $\mathbf{z}_{m,n}^G \in \mathbb{R}^{|C|}$ is the soft prediction of the same token with Gaussian noises G. Thus, we have

$$\mathcal{L}_{\text{reg}} = -\frac{1}{M^{l_i} N} \sum_{m,n=1}^{M^{l_i},N} \mathbb{I}(z_{m,n}^{\max}) \frac{\mathbf{z}_{m,n}}{\tau} \log \mathbf{z}_{m,n}^G,$$

where τ is a temperature factor to control the smoothness. $z_{m,n}^{\max}$ denotes the max probability over C classes, i.e., $\arg\max_c \mathbf{z}_{m,n}$. $\mathbb{I}$ is an indicator function used to mask the token with low prediction confidence, i.e., $\mathbb{I}(z)$ amounts to 1 if $z > \epsilon$, otherwise 0, where $\epsilon \in (0, 1)$ is a threshold.

Meta learning The meta loss $\mathcal{L}_{\text{meta}}$ aims to enforce the teacher to learn from the student's feedback on the current pseudo-labels in order to adjust its pseudo-annotation strategies, which is also known as *learning to learn*. To quantitatively measure the quality of the current pseudo-labels, we evaluate the student's performances (loss) on the labeled data before and after the update, i.e., θ_S^t and θ_S^{t+1} as defined in Eq. (3),

$$\mathcal{L}_{S,\text{lab}}^{l_i,(t)} = \frac{1}{M} \sum_{m=1}^{M^{l_i}} \ell(\mathbf{Y}_m^{l_i}, f(\mathbf{X}_m^{l_i}; \boldsymbol{\theta}_S^{(t)})),$$

$$\mathcal{L}_{S,\text{lab}}^{l_i,(t+1)} = \frac{1}{M} \sum_{m=1}^{M^{l_i}} \ell(\mathbf{Y}_m^{l_i}, f(\mathbf{X}_m^{l_i}; \boldsymbol{\theta}_S^{(t+1)})).$$

The difference between $\mathcal{L}_{S,\text{lab}}^{l_i,(t+1)}$ and $\mathcal{L}_{S,\text{lab}}^{l_i,(t)}$, i.e., $\lambda_{\text{meta}}^{l_i} = \mathcal{L}_{S,\text{lab}}^{l_i,(t+1)} - \mathcal{L}_{S,\text{lab}}^{l_i,(t)}$ can be used as a dynamic feedback or reward function to meta-optimize the teacher network towards the direction that generates better pseudo-labels for the language l_i. If the pseudo-labels at the t-th step can improve the student network, then $\lambda_{\text{meta}}^{l_i}$ will be negative, and positive vice versa. Thus, the meta loss $\mathcal{L}_{\text{meta}}$ is defined as:

$$\mathcal{L}_{\text{meta}} = \frac{\lambda_{\text{meta}}^{l_i}}{\widetilde{M}^{l_i}} \sum_{m=1}^{\widetilde{M}^{l_i}} \ell(\widetilde{\mathbf{Y}}_m^{l_i,(t)}, f(\widetilde{\mathbf{X}}_m^{l_i}; \boldsymbol{\theta}_T^{(t)})).$$

where $\widetilde{\mathbf{Y}}_m^{l_i,(t)}$ is the pseudo-labels for the language l_i produced by the teacher at the t-th step.

3.4 Alternating Training

During the teacher-student interaction stage, we alternately train the student network and the teacher network by minimizing $\mathcal{L}_S$ and $\mathcal{L}_T$ separately for each language. As such, the teacher and student can achieve mutual learning, i.e., at this stage, the student will only learn from the multilingual unlabeled data with pseudo-labels produced by the teacher, and meanwhile, the teacher will also adjust its pseudo-annotation strategy according to the feedback from the student. After distilling the knowledge from the teacher to teach the student network, we finally take the student model fine-tuned on the multilingual labeled data as the final model for evaluation.

4 Experiment

4.1 Datasets

We consider the following three multilingual sequence labeling datasets for experiments, of which the statistics of the datasets are shown in Table 3.

(i) **Multilingual Open-domain NER** is a open-domain NER dataset from CoNLL02 (Tjong Kim Sang, 2002a) and CoNLL03 (Tjong Kim Sang, 2002b) NER shared tasks, containing English (En), Spanish (Es), German (De) and Dutch (Nl) with

Hyper-parameter	Dataset		
	O-NER	E2E-ABSA	E-NER
batch size	8	8	64
learning rate	1^{-5}	1^{-5}	5^{-5}
noise variance $\boldsymbol{\sigma}^2$	0.01	0.001	0.01
temperature τ	0.7	0.7	0.7
threshold ϵ	0.6	0.6	0.6

Table 2: Settings of hyper-parameters.

4 entity types: person, location, organization, and miscellaneous.

(ii) **Multilingual E2E-ABSA** is an ABSA benchmark from SemEval ABSA challenge (Pontiki et al., 2014). We follow the settings of End-to-End Aspect-based Sentiment Analysis (Mitchell et al., 2013; Zhang et al., 2015), which jointly extracts aspect terms and the associated sentiments using a unified tagging scheme. It consists of English (En), French (Fr), Spanish (Es), Turkish (Tr), Dutch (Nl) and Russian (Ru) with 3 entity types: positive, neutral, and negative.

(iii) **Multilingual E-commerce NER** is a real-world large-scale query NER dataset used for E-commerce. The queries are collected from a shopping website, including English (En), German (De), Spanish (Es), French (Fr), Italian (It), Japanese (Jp), Chinese (Zh), Czech (Cs), Dutch (Nl), Polish (Pl), Portugal (Pt), Turkish (Tr) with 13 entity types.

4.2 Setting

For the low-resource setting, we only use ***1%***, ***10%***, ***1%*** randomly sampled training data as the **labeled data** for each language of the open-domain NER, E2E-ABSA, and E-commerce query NER datasets, respectively. And we treat the remaining training data as the **unlabeled data**. This results in tens to thousands of labeled data for each language. We use the span-level micro F1-score (**exact match**) as the evaluation metrics.

5 Implementation details

Experimental Environment Our MetaTS model and baseline methods are all using Pytorch 1.7.0 based on CUDA 11.0, Amazon EC2 virtual machine with 8 NVIDIA A100-SXM4-40GB GPUs, and are tested on Linux, Python 3.7.6 from Anaconda 4.8.4.

Encoder We use the mBERT-base model: *bert-base-multilingual-cased*[2] model pre-trained on 104 languages as the encoder, which has 12 layers, 768-d hidden size, 12 heads and 110M total parameters. The hidden states of the last layer of the model are used as the token representations for token-level label prediction. The mBERT is jointly optimized with other parameters during the training stage.

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

Dataset	#Train	#Dev	#Test	#Type	%Coverage	#Avg len
Multilingual Open-domain NER (*long-text*)						
En	14041	3250	3453	4	16.72	14.50
Es	8323	1915	1517	4	8.11	17.03
De	12152	2867	3005	4	12.39	31.81
Nl	15806	2895	5195	4	9.52	12.82
Multilingual E2E-ABSA (*long-text*)						
En	1600	400	676	3	8.97	14.55
Fr	1332	332	696	3	7.78	17.50
Es	1656	414	881	3	7.23	16.38
Tr	986	246	144	3	6.51	14.75
Nl	1378	344	575	3	8.06	14.20
Ru	2924	731	1209	3	15.13	10.15
Multilingual E-commerce NER (*short-text*)						
En	256571	14193	14269	13	98.87	3.20
De	98980	5442	5473	13	95.49	2.76
Es	63844	3600	3488	13	99.05	3.76
Fr	79176	4383	4504	13	98.91	3.16
It	52136	2933	2867	13	99.04	3.51
Jp	77457	4422	4365	13	98.65	2.48
Zh	22467	1238	1247	13	98.51	2.51
Cs	4430	272	252	13	93.66	4.26
Nl	8562	423	478	13	97.09	2.87
Pl	4489	251	229	13	92.19	4.38
Pt	4467	273	247	13	99.45	2.47
Tr	5093	267	274	13	99.52	2.32

Table 3: Data statistics. Type and Coverage denote the number of entity type and the ratio of non-O entity.

Initialization & Training For all the experiments, the model is optimized by the Adam algorithm (Kingma and Ba, 2015) for training. The weight matrices are initialized with a uniform distribution $U(-0.01, 0.01)$. Gradients with the 1 norm larger than 40 are normalized to be 1. To alleviate overfitting, we perform early stopping on the validation set during both the teacher-student interaction and finetune stages.

Hyperparameter For the all three multilingual sequence labeling datasets, the hyper-parameters are manually tuned on 10% randomly held-out labeled training data (downsampled version) of the all languages. The initial learning rate for Adam is tuned amongst $\{10^{-5}, 2\times10^{-5}, 3\times10^{-5}, 5\times10^{-5}, 10^{-4}\}$. The batch size is tuned amongst {8, 16, 32, 64}. The Gaussian noise variance $\boldsymbol{\sigma}^2$ is tuned amongst {0.001, 0.01, 0.1, 1.0} and we have found that when $\boldsymbol{\sigma}^2$ is larger than 0.01, the model will collapse. This is reasonable since too large $\boldsymbol{\sigma}^2$ can bring in unbearable noises that the model itself cannot denoise. The temperature factor τ is tuned

amongst {0.5, 0.6, 0.7, 0.8, 0.9}. The threshold ϵ is tuned amongst {0.5, 0.6, 0.7, 0.8, 0.9}. For both the teacher and the student network, we use label smoothing for Eq. (2) and Eq. (4) with the smoothing factor 0.15. We use 128 as the maximum sentence length for all datasets. The detailed hyperparameters are listed in Table 2.

5.1 Baselines

We compare our model with different groups of baseline methods to verify the effectiveness.

• **Fully-supervised.** (i) **mBERT (Single)** fine-tunes a mBERT on the sampled labeled data for each language; (ii) **mBERT (Multi)** fine-tunes a mBERT on the sampled labeled data of all languages; (iii) **mBERT (Full)** uses the full labeled data of all languages to fine-tune a mBERT, which is usually regarded as the upper bound.

• **Semi-supervised.** (i) **MT (KL/MSE)**[3] (Tarvainen and Valpola, 2017) uses Mean Teacher, an ensemble method to average student model weights and forms a teacher model using KL divergence or mean square error to force the prediction consistency. (ii) **VAT**[4] (Miyato et al., 2018; Chen et al., 2020b) is a regularization method which adopts virtual adversarial training to smooth the output distribution to make the model robust to noise. (iii) **NoisyStudent**[5] (Xie et al., 2020b) extends the idea of self-training and distillation with the use of noise added to the student during learning. (iv) **BOND (hard/soft/soft-high)**[6] (Liang et al., 2020) employs a state-of-the-art TS framework of self-training with hard pseudo-labels, soft pseudo-labels (Xie et al., 2016), as well as the proposed soft pseudo-labels on selected high confidence tokens. For a fair comparison, we use the mBERT as the base encoder for all baselines.

5.2 Main Results

5.2.1 Multilingual Academic Benchmarks

We present the the main results on multilingual academic datasets for open-domain NER and E2E-ABSA in Table 4 and Table 5, respectively. Based on the results, we can observe:

• **MetaTS**: MetaTS significantly and consistently outperforms all baseline methods for all languages

[3] https://github.com/CuriousAI/mean-teacher

[4] https://github.com/takerum/vat_tf

[5] https://github.com/google-research/noisystudent

[6] https://github.com/cliang1453/BOND

Method (*Span F1*)	En	Es	De	Nl	Avg	Δ
Fully-supervised Baselines (1% labeled data)						
mBERT (Single)	83.03	75.62	67.31	73.67	74.91	(+5.91)
mBERT (Multi)	82.54	79.90	73.32	79.78	78.88	(+1.94)
Semi-supervised Baselines (1% labeled data)						
MT (KL)	83.52	77.99	73.40	80.71	78.91	(+1.91)
MT (MSE)	84.25	79.45	73.95	79.98	79.46	(+1.36)
VAT	83.70	78.27	73.02	81.00	79.00	(+1.82)
NoisyStudent	82.54	79.21	71.08	78.38	77.80	(+3.02)
BOND (hard)	82.75	78.31	75.74	80.29	79.27	(+1.55)
BOND (soft)	85.26	78.39	75.21	78.40	79.32	(+1.50)
BOND (soft-high)	84.62	79.87	72.68	80.31	79.37	(+1.45)
MetaTS (Ours)	**85.67**†	**80.05**	**76.23**†	**81.31**†	**80.82**†	-
Upper Bound (100% labeled data)						
mBERT (Full)	90.34	85.99	81.66	89.43	86.85	-

Table 4: The results (%) on multilingual open-domain NER. Δ refers to the improvements. † means the statistically significant improvement over the best baseline with paired sample t-test $p < 0.01$.

Method (*Span F1*)	En	Fr	Es	Tr	Nl	Ru	Avg	Δ
Fully-supervised Baselines (10% labeled data)								
mBERT (Single)	49.39	40.89	52.38	27.75	38.06	44.12	42.10	(+10.50)
mBERT (Multi)	55.85	47.61	58.37	29.24	46.51	46.15	47.29	(+5.31)
Semi-supervised Baselines (10% labeled data)								
MT (KL)	56.64	47.06	60.76	28.38	46.80	49.56	48.20	(+4.40)
MT (MSE)	54.56	48.53	60.88	30.65	47.26	50.28	48.69	(+3.91)
VAT	54.12	46.03	58.84	33.99	46.47	50.35	48.30	(+4.30)
NoisyStudent	55.90	47.13	56.89	34.92	47.53	49.11	48.58	(+4.02)
BOND (hard)	57.36	48.84	59.71	36.62	46.98	48.56	49.68	(+2.92)
BOND (soft)	56.34	50.40	61.95	33.78	**50.62**	48.14	50.21	(+2.39)
BOND (soft-high)	56.70	49.74	61.08	35.62	47.48	51.42	50.34	(+2.26)
MetaTS (Ours)	**59.45**†	**54.29**†	**62.90**†	**37.15**†	50.27	**51.51**	**52.60**†	-
Upper Bound (100% labeled data)								
mBERT (Full)	61.54	57.76	65.80	43.11	58.19	56.44	57.14	-

Table 5: The results (%) on multilingual E2E-ABSA.

of two sequence labeling tasks by a large margin (NER: +1.36% Avg gain over MT (MSE), E2E-ABSA: +2.26% Avg gain over BOND (soft-high)).

• **Supervised**: (i) Supervised baselines perform much worse than semi-supervised baselines. This demonstrates that even with mPLMs like mBERT, supervised learning cannot achieve satisfactory results in the low-data regime. (ii) mBERT (Multi) significantly beats mBERT (Single), which shows that the joint usage of labeled data from multiple languages is better than each monolingual model when supervision signals are insufficient.

• **Semi-supervised**: (i) By leveraging large unlabeled data, semi-supervised baselines can obtain considerable improvements. (ii) Our proposed MetaTS method can still outperform those semi-supervised baselines based on the traditional TS framework. This indicates our meta teacher-student learning paradigm can capture more underlying treasures from the unlabeled data, which can learn to adjust pseudo-annotation strategies by taking advantage of the student's learning feedback.

Method (*Span F1*)	En	De	Es	Fr	It	Jp	Zh	Cs	Nl	Pl	Pt	Tr	Avg	Δ
Fully-supervised Baselines (1% labeled data)														
mBERT (Single)	61.83	57.47	57.62	52.27	57.35	46.80	49.11	41.56	36.31	46.57	21.29	36.84	47.41	(+10.75)
mBERT (Multi)	62.07	61.63	63.48	57.90	63.65	49.73	55.62	50.89	52.98	61.89	35.07	54.10	55.94	(+2.22)
Semi-supervised Baselines (1% labeled data)														
MT (KL)	61.38	61.25	63.11	57.38	62.33	49.05	56.02	51.26	54.60	60.86	34.27	55.15	55.73	(+2.43)
MT (MSE)	61.54	63.26	63.77	58.73	64.22	49.52	57.46	56.95	53.81	62.83	33.47	56.01	56.94	(+1.22)
VAT	60.73	61.50	61.84	57.21	62.58	49.02	55.23	51.88	54.71	59.02	35.74	56.26	55.64	(+2.52)
NoisyStudent	62.24	62.81	63.29	58.30	63.71	49.55	54.87	55.05	**56.05**	62.05	34.91	54.99	56.66	(+1.50)
BOND (hard)	62.38	62.61	64.19	58.67	63.43	49.23	54.48	55.21	54.96	61.77	37.47	53.96	56.67	(+1.49)
BOND (soft)	62.61	62.03	63.29	57.34	63.38	48.68	54.25	56.26	54.24	61.78	34.72	56.15	56.45	(+1.71)
BOND (soft-high)	62.46	61.87	63.95	57.61	63.60	50.19	57.35	51.67	55.11	62.61	**35.96**	54.91	56.67	(+1.49)
MetaTS (Ours)	**63.79**†	**63.78**†	**64.77**†	**60.02**†	**65.04**†	**51.81**†	**58.34**†	**57.74**†	54.59	**64.41**†	33.96	**57.84**†	**58.16**†	-
Upper Bound (100% labeled data)														
mBERT (Full)	76.51	76.21	77.83	73.08	78.75	67.99	73.73	72.86	75.89	80.64	65.60	73.17	74.36	-

Table 6: The results (%) on multilingual E-commerce NER.

Model	**O-NER**	**E2E-ABSA**	**E-NER**
MetaTS	**80.82**	**52.60**	**58.16**
MetaTS w/o $\mathcal{L}_{\text{reg}}$	79.90	52.28	56.50
MetaTS w/o $\mathcal{L}_{\text{meta}}$	79.83	50.35	56.23
MetaTS w/ Soft labels	79.73	47.51	56.71
MetaTS w/ Soft-high labels	78.57	45.62	55.41

Table 7: Ablation results (%): average span-level micro F1-score over all the languages for each dataset.

5.2.2 Multilingual Industrial Dataset

We present the main results on the multilingual industrial dataset for E-commerce NER in Table 6. Compared with widely-used benchmark datasets in the academia, this industrial dataset, as illustrated in Table 3, behaves more challenging in terms of: (i) **large label space**: there are much more (13) entity types, bringing in a significant difficulty for the prediction; (ii) **high entity coverage**: almost all tokens in the user query are tagged with a non-O tag (>90% coverage rate) (in low-coverage datasets, high-performance does not mean the model can well identity the entities due to the high O proportion (Zhou et al., 2019)); (iii) **short text**: the user queries are usually short, which lack sufficient contextual information for context-dependent token-level prediction; (iv) **data imbalance**: the labeled data among different language are very skewed, closer to the real-world data distribution of high-resource and low-resource languages; (v) **large-scale data size**: this dataset has much more data (about 700k) than existing public datasets. Even involving so many challenges for this dataset, MetaTS can still achieve significant improvements over all the baseline methods on most languages. This shows more convincing evidence that MetaTS generates more high-quality pseudo-labels for even short-text data in a large label space via the meta teacher-student learning paradigm.

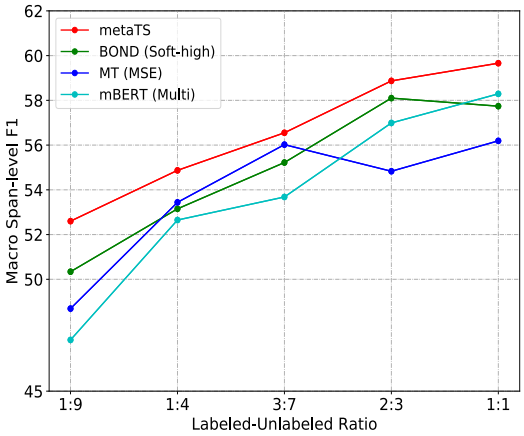

Figure 2: Average micro span-level F1 w.r.t proportions of the labeled training data for each language.

5.3 Ablation Results

To verify the efficacy of each component, we compare MetaTS with its ablation variants in Table 7.

w/ $\mathcal{L}_{\text{reg}}$ v.s. **w/o** $\mathcal{L}_{\text{reg}}$: For MetaTS w/o $\mathcal{L}_{\text{reg}}$, we remove the regularization loss $\mathcal{L}_{\text{reg}}$ on the unlabeled multilingual data for the teacher. We can observe that there are remarkable performance drops on all three datasets. This indicates that it works better when the teacher is jointly trained with other auxiliary tasks such as the virtual data augmentation since it can enhance the prediction confidence of the teacher towards the unlabeled data.

w/ $\mathcal{L}_{\text{meta}}$ v.s. **w/o** $\mathcal{L}_{\text{meta}}$: For MetaTS w/o $\mathcal{L}_{\text{meta}}$, we remove the meta loss $\mathcal{L}_{\text{meta}}$ for the teacher. That is, we discard the instant feedbacks from the student on the generated pseudo-labels, so that the teacher cannot dynamically adjust its pseudo-annotation strategy. As such, MetaTS w/o $\mathcal{L}_{\text{meta}}$ has demonstrated significant degradation. Besides, we can also conclude that the meta-learning loss contributes more to our performance improvements.

Hard labels v.s. **Soft labels**: Compared with utilizing hard pseudo-labels to teach the student, we

Input Sentence & Ground-truth Labels	Self-Training Labels	MetaTS Labels
Open-domain NER (ORG, LOC, PER, MISC)		
1. The years I spent as manager of the [*Republic of Ireland*] were the best years of my life .	[*Republic of Ireland*]✗	[*Republic of Ireland*]
2. His father [*Clarence Woolmer*] represented [*United Province*] , now renamed [*Uttar Pradesh*] , in [*India*] 's [*Ranji Trophy*] national championship.	[*Clarence Woolmer*] [*United Province*]✗ [*Uttar Pradesh*] [*India*]✗ [*Ranji Trophy*]	[*Clarence Woolmer*] [*United Province*] [*Uttar Pradesh*] [*India*] [*Ranji Trophy*]
E2E-ABSA: (POS, NEG, NEU)		
3. I liked the [*atmosphere*] very much but the [*food*] was not worth the price .	[*atmosphere*] [*food*]✗	[*atmosphere*] [*food*]
4. Not the biggest [*portions*] but adequate .	None✗	[*portions*]
E-commerce NER: (Brand, ProductType, Size, ProductLine, VisualFeature)		
5. [*samsung*] [*tab*] [*4 t 231*][*scratch guard*]	[*samsung*] [*tab*]✗ [*4 t 231*]✗ [*scratch guard*]	[*samsung*] [*tab*] [*4 t 231*] [*scratch guard*]
6. [*half and half*] [*wigs*]	[*half and half*]✗ [*wigs*]	[*half and half*] [*wigs*]

Table 8: Case analysis for three multilingual sequence labeling datasets.

observe that soft pseudo-labels (Xie et al., 2016) can substantially hurt the model performance and lower the convergence speed, even worse after high confidence selection (Liang et al., 2020) is introduced. This circumstance has also been shown in prior study (Kumar et al., 2020). We hypothesize that such performance drops may be attributed to soft pseudo-labels being noisier than sharpened hard pseudo-labels in meta-learning.

5.4 Impact of Labeled-Unlabeled Ratio

To investigate the effect of the labeled-unlabeled data ratio, we vary the labeled proportion of each language's training set and compare MetaTS with mBERT (Multi), MT (MSE), and BOND (soft-high). We use the average span-level micro F1 score over all languages of the multilingual E2E-ABSA dataset and change the labeled proportion from 0.1, 0.2, 0.3, 0.4 to 0.5. Since the remaining training data is treated as the unlabeled data, the corresponding labeled-unlabeled ratios are from 1:9, 1:4, 3:7, 2:3 to 1:1. As shown in Figure 2, the gap between the MetaTS and all baseline methods grows as the labeled-unlabeled ratio shrinks. Semi-supervised baselines MT (MSE) and BOND (soft-high) show marginal improvements over the supervised learning method mBERT (Multi) and even perform worse when the labeled size becomes large. This verifies that the MetaTS is much less sensitive to the drop in the labeled proportion for each language by making effective use of the large amounts of multilingual unlabeled data.

5.5 Pseudo-Labeling Visualization

To qualitatively demonstrate that MetaTS can generate better token-level pseudo-labels that involve complicated dependency relations, we perform the pseudo-labeling visualization of ground-truth labels, self-training (BOND) pseudo-labels, and our MetaTS model pseudo-labels for three datasets we used. As illustrated in Table 8, we only show some English cases for easy understanding, although we also observe our consistent advantages in many other languages (This is quantitatively verified by Section 5.2 Main Results).

As we can see, traditional teacher-student frameworks with self-training cannot handle the token pseudo-labeling in complicated contexts, including (1) **entities of ambiguity**: the entities have ambiguous semantics, which can denote different types in light of their surrounding contexts. For example, in the open-domain NER, self-training usually confuses organization (`ORG`) with location (`LOC`) as a sequence of misclassifying *Republic of Ireland* (Case#1) as `LOC` due to the location word "Ireland". In the E-commerce NER, *half and half* (Case#6) is used to describe the visual features of wigs instead of the size; (2) **entities in the transition context**: the entities before and after the transition may have contrastive meanings. For example, the user expresses a positive sentiment towards *atmoshpere* but a negative sentiment to *food* (Case#3). (3) **high entity coverage**: most of tokens in the sentence are truly entities instead of `O`. For example, in Case#2 and Case#5, self-training cannot identify the cor-

rect types for all occurring entities. (4) **entity missing**: self-training may not be able to capture the entities like *protions* in the Case#4. In contrast, our proposed MetaTS can demonstrate more robustness to these challenges, attributed to the meta teacher-student learning paradigm that can adjust teacher's pseudo labeling strategies according to the student's instant feedback.

6 Related Works

6.1 Multilingual Sequence Labeling

Most recent works on multilingual sequence labeling focus on improving the cross-lingual transferability for different languages (Täckström, 2012; Fang et al., 2017; Enghoff et al., 2018; Xie et al., 2018; Rahimi et al., 2019; Johnson et al., 2019; Wu et al., 2020a,b,c; Li et al., 2020a). Cross-lingual transfer (Li et al., 2020b) aims to leverage knowledge from source languages to improve the performance in target languages only, which puts more emphasis on how to reduce the language distribution gaps due to the lack of labeled data for target languages. Besides, each target language usually requires training an individual model. This behaves particularly resource consuming. On the contrary, our goal is to improve all languages' performance using a unified model. Only a few studies have explored building a unified multilingual model with enough labeled data to handle multiple languages (Wang et al., 2020a). Different from that, we explore a motivated and challenging multilingual setting with minimal supervision.

To alleviate the data-sparsity issue, various advanced techniques have emerged, such as transfer learning (Pan and Yang, 2009), semi-supervised learning (Mishra and Diesner, 2016; He and Sun, 2017; Chen et al., 2018; Wang et al., 2020b; Bhattacharjee et al., 2020; Chen et al., 2020b), domain adaptation (Li et al., 2017, 2018, 2019b,a), and data augmentation (Dai and Adel, 2020; Chen et al., 2020a; Ding et al., 2020). Considering the multilingual setting, data augmentation may be infeasible and could bring in external knowledge errors. Semi-supervised learning has shown promising results in instance-level classification tasks (Tarvainen and Valpola, 2017; Miyato et al., 2018; Xie et al., 2020b) but less effectiveness in more complicated token-level classification.

6.2 Meta Learning

Inspired by human beings' ability to adapt and transfer knowledge from previous tasks, meta learning (Finn et al., 2017; Nichol et al., 2018; Pham et al., 2020; Yao et al., 2019, 2021) has been initiated on low-resource NLP, such as text classification (Yu et al., 2018; Wu et al., 2019; Geng et al., 2019; Sun et al., 2019; Geng et al., 2020; Bao et al., 2020), relation classification (Han et al., 2018; Gao et al., 2019; Obamuyide and Vlachos, 2019), slot tagging (Hou et al., 2020), event detection (Deng et al., 2020), and natural language understanding (NLU) (Dou et al., 2019). Considering multilingualism, only a few works have explored meta learning to improve the cross-lingual transferability of low-resource languages, e.g., text classification (Li et al., 2020b), NLU (Nooralahzadeh et al., 2020), NER (Wu et al., 2020c), and machine translation (Gu et al., 2018). On the contrary, our ultimate goal is to utilize meta learning to better leverage multilingual unlabeled data for boosting all languages' performance. Our work is inspired by meta-policies for teaching mechanisms (Fan et al., 2018; Pham et al., 2020), which only focus on instance-level image classification tasks and rely on single feedback from the student. Besides, the success of the two works is conditioned on additional techniques like data augmentation for images, which is tedious and almost infeasible in challenging NLP tasks, especially for multilingual sequence labeling.

7 Conclusion

The effectiveness of supervised methods for low-resource multilingual sequence labeling is limited due to data scarcity. To tackle this challenge, we propose a novel MetaTS method to enhance the teacher-student framework of self-training, which leverages the student's feedback on multilingual token-level pseudo-labels to adjust the teacher's pseudo-annotation strategies. Extensive evaluations on both the public academic benchmarks and the large-scale industrial dataset quantitatively and qualitatively demonstrate the effectiveness of MetaTS. In the future, the proposed MetaTS method can potentially be applied to multilingual natural language understanding (XLU) tasks (Hu et al., 2020) and be generalized to multi-task learning (Wang et al., 2019) problems.

References

Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. 2020. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *IJCNN*, pages 1–8. IEEE.

Yujia Bao, Menghua Wu, Shiyu Chang, and Regina Barzilay. 2020. Few-shot text classification with distributional signatures. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Kasturi Bhattacharjee, Miguel Ballesteros, Rishita Anubhai, Smaranda Muresan, Jie Ma, Faisal Ladhak, and Yaser Al-Onaizan. 2020. To BERT or not to BERT: Comparing task-specific and task-agnostic semi-supervised approaches for sequence tagging. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7927–7934, Online. Association for Computational Linguistics.

Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. 2009. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542.

Jiaao Chen, Zhenghui Wang, Ran Tian, Zichao Yang, and Diyi Yang. 2020a. Local additivity based data augmentation for semi-supervised NER. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1241–1251, Online. Association for Computational Linguistics.

Luoxin Chen, Weitong Ruan, Xinyue Liu, and Jianhua Lu. 2020b. SeqVAT: Virtual adversarial training for semi-supervised sequence labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8801–8811, Online. Association for Computational Linguistics.

Mingda Chen, Qingming Tang, Karen Livescu, and Kevin Gimpel. 2018. Variational sequential labelers for semi-supervised learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 215–226, Brussels, Belgium. 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.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067.

Xiang Dai and Heike Adel. 2020. An analysis of simple data augmentation for named entity recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3861–3867, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Shumin Deng, Ningyu Zhang, Jiaojian Kang, Yichi Zhang, Wei Zhang, and Huajun Chen. 2020. Meta-learning with dynamic-memory-based prototypical network for few-shot event detection. In *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020*, pages 151–159. ACM.

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.

Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq Joty, Luo Si, and Chunyan Miao. 2020. DAGA: Data augmentation with a generation approach for low-resource tagging tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6045–6057, Online. Association for Computational Linguistics.

Zi-Yi Dou, Keyi Yu, and Antonios Anastasopoulos. 2019. Investigating meta-learning algorithms for low-resource natural language understanding tasks. 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 1192–1197, Hong Kong, China. Association for Computational Linguistics.

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

Yang Fan, Fei Tian, Tao Qin, Xiang-Yang Li, and Tie-Yan Liu. 2018. Learning to teach. In *International Conference on Learning Representations*.

Meng Fang, Yuan Li, and Trevor Cohn. 2017. Learning how to active learn: A deep reinforcement learning approach. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 595–605, Copenhagen, Denmark. Association for Computational Linguistics.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*,

volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR.

Tianyu Gao, Xu Han, Zhiyuan Liu, and Maosong Sun. 2019. Hybrid attention-based prototypical networks for noisy few-shot relation classification. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6407–6414. AAAI Press.

Ruiying Geng, Binhua Li, Yongbin Li, Jian Sun, and Xiaodan Zhu. 2020. Dynamic memory induction networks for few-shot text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1087–1094, Online. Association for Computational Linguistics.

Ruiying Geng, Binhua Li, Yongbin Li, Xiaodan Zhu, Ping Jian, and Jian Sun. 2019. Induction networks for few-shot text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3904–3913, Hong Kong, China. Association for Computational Linguistics.

Jiatao Gu, Yong Wang, Yun Chen, Victor O. K. Li, and Kyunghyun Cho. 2018. Meta-learning for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3622–3631, Brussels, Belgium. Association for Computational Linguistics.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, Brussels, Belgium. Association for Computational Linguistics.

Hangfeng He and Xu Sun. 2017. A unified model for cross-domain and semi-supervised named entity recognition in chinese social media. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3216–3222. AAAI Press.

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

Yutai Hou, Wanxiang Che, Yongkui Lai, Zhihan Zhou, Yijia Liu, Han Liu, and Ting Liu. 2020. Few-shot slot tagging with collapsed dependency transfer and label-enhanced task-adaptive projection network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1381–1393, Online. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. 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 2485–2494, Hong Kong, China. Association for Computational Linguistics.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Andrew Johnson, Penny Karanasou, Judith Gaspers, and Dietrich Klakow. 2019. Cross-lingual transfer learning for Japanese named entity recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 182–189, Minneapolis, Minnesota. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Ananya Kumar, Tengyu Ma, and Percy Liang. 2020. Understanding self-training for gradual domain adaptation. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 5468–5479. PMLR.

Dong-Hyun Lee et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop*, volume 3.

Qi Li, Haibo Li, Heng Ji, Wen Wang, Jing Zheng, and Fei Huang. 2012. Joint bilingual name tagging for parallel corpora. In *21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012*, pages 1727–1731. ACM.

Xin Li, Lidong Bing, Wenxuan Zhang, Zheng Li, and Wai Lam. 2020a. Unsupervised cross-lingual adaptation for sequence tagging and beyond. *arXiv preprint arXiv:2010.12405*.

Zheng Li, Mukul Kumar, William Headden, Bing Yin, Ying Wei, Yu Zhang, and Qiang Yang. 2020b. Learn to cross-lingual transfer with meta graph learning across heterogeneous languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2290–2301, Online. Association for Computational Linguistics.

Zheng Li, Xin Li, Ying Wei, Lidong Bing, Yu Zhang, and Qiang Yang. 2019a. Transferable end-to-end aspect-based sentiment analysis with selective adversarial learning. 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 4590–4600, Hong Kong, China. Association for Computational Linguistics.

Zheng Li, Ying Wei, Yu Zhang, and Qiang Yang. 2018. Hierarchical attention transfer network for cross-domain sentiment classification. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5852–5859. AAAI Press.

Zheng Li, Ying Wei, Yu Zhang, Xiang Zhang, and Xin Li. 2019b. Exploiting coarse-to-fine task transfer for aspect-level sentiment classification. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 4253–4260. AAAI Press.

Zheng Li, Yu Zhang, Ying Wei, Yuxiang Wu, and Qiang Yang. 2017. End-to-end adversarial memory network for cross-domain sentiment classification. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 2237–2243. ijcai.org.

Chen Liang, Yue Yu, Haoming Jiang, Siawpeng Er, Ruijia Wang, Tuo Zhao, and Chao Zhang. 2020. BOND: bert-assisted open-domain named entity recognition with distant supervision. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 1054–1064. ACM.

Shubhanshu Mishra and Jana Diesner. 2016. Semi-supervised named entity recognition in noisy-text. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 203–212, Osaka, Japan. The COLING 2016 Organizing Committee.

Margaret Mitchell, Jacqui Aguilar, Theresa Wilson, and Benjamin Van Durme. 2013. Open domain targeted sentiment. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1654, Seattle, Washington, USA. Association for Computational Linguistics.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *TPAMI*, 41(8):1979–1993.

Alex Nichol, Joshua Achiam, and John Schulman. 2018. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*.

Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein. 2020. Zero-shot cross-lingual transfer with meta learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4547–4562, Online. Association for Computational Linguistics.

Abiola Obamuyide and Andreas Vlachos. 2019. Model-agnostic meta-learning for relation classification with limited supervision. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5873–5879, Florence, Italy. Association for Computational Linguistics.

Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *TKDE*, 22(10):1345–1359.

Hieu Pham, Zihang Dai, Qizhe Xie, Minh-Thang Luong, and Quoc V Le. 2020. Meta pseudo labels. *arXiv preprint arXiv:2003.10580*.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. 2017. Neural sequence learning models for word sense disambiguation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1156–1167, Copenhagen, Denmark. Association for Computational Linguistics.

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

Sebastian Ruder and Barbara Plank. 2018. Strong baselines for neural semi-supervised learning under domain shift. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1044–1054, Melbourne, Australia. Association for Computational Linguistics.

Henry Scudder. 1965. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11(3):363–371.

Shengli Sun, Qingfeng Sun, Kevin Zhou, and Tengchao Lv. 2019. Hierarchical attention prototypical networks for few-shot text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 476–485, Hong Kong, China. Association for Computational Linguistics.

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

Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 1195–1204.

Erik F. Tjong Kim Sang. 2002a. 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. 2002b. 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)*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Fei Huang, and Kewei Tu. 2020a. Structure-level knowledge distillation for multilingual sequence labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3317–3330, Online. Association for Computational Linguistics.

Yaqing Wang, Subhabrata Mukherjee, Haoda Chu, Yuancheng Tu, Ming Wu, Jing Gao, and Ahmed Hassan Awadallah. 2020b. Adaptive self-training for few-shot neural sequence labeling. *arXiv preprint arXiv:2010.03680*.

Jiawei Wu, Wenhan Xiong, and William Yang Wang. 2019. Learning to learn and predict: A meta-learning approach for multi-label classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4354–4364, Hong Kong, China. 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 Jianguang Lou. 2020b. Unitrans : Unifying model transfer and data transfer for cross-lingual named entity recognition with unlabeled data. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3926–3932. ijcai.org.

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. In *AAAI*, pages 9274–9281.

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.

Junyuan Xie, Ross B. Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 478–487. JMLR.org.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020a. Unsupervised data augmentation for consistency training. In *NeurIPS*, pages 6256–6268.

Qizhe Xie, Minh-Thang Luong, Eduard H. Hovy, and Quoc V. Le. 2020b. Self-training with noisy student improves imagenet classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10684–10695. IEEE.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, Online. Association for Computational Linguistics.

Huaxiu Yao, Long-Kai Huang, Linjun Zhang, Ying Wei, Li Tian, James Zou, Junzhou Huang, et al. 2021. Improving generalization in meta-learning via task augmentation. In *International Conference on Machine Learning*, pages 11887–11897. PMLR.

Huaxiu Yao, Ying Wei, Junzhou Huang, and Zhenhui Li. 2019. Hierarchically structured meta-learning. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 7045–7054. PMLR.

Mo Yu, Xiaoxiao Guo, Jinfeng Yi, Shiyu Chang, Saloni Potdar, Yu Cheng, Gerald Tesauro, Haoyu Wang, and Bowen Zhou. 2018. Diverse few-shot text classification with multiple metrics. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1206–1215, New Orleans, Louisiana. Association for Computational Linguistics.

Meishan Zhang, Yue Zhang, and Duy-Tin Vo. 2015. Neural networks for open domain targeted sentiment. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 612–621, Lisbon, Portugal. Association for Computational Linguistics.

Joey Tianyi Zhou, Hao Zhang, Di Jin, Xi Peng, Yang Xiao, and Zhiguo Cao. 2019. Roseq: Robust sequence labeling. *TNNLS*, 31(7):2304–2314.