Uncertainty-Aware Cross-Lingual Transfer with Pseudo Partial Labels

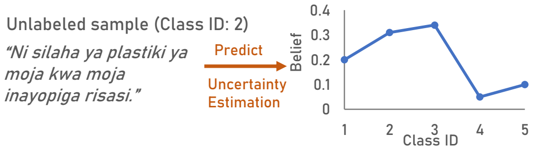
Shuo Lei[†], **Xuchao Zhang**[‡], **Jianfeng He**[†*], **Fanglan Chen**[†], **Chang-Tien Lu**[†]
[†]Department of Computer Science, Virginia Tech, Falls Church, VA, USA
[‡]NEC Laboratories America, Princeton, NJ, USA
{slei,jianfenghe,fanglanc,ctlu}@vt.edu, xuczhang@nec-labs.com

Abstract

Large-scale multilingual pre-trained language models have achieved remarkable performance in zero-shot cross-lingual tasks. A recent study has demonstrated the effectiveness of self-learning-based approach on cross-lingual transfer, where only unlabeled data of target languages are required, without any efforts to annotate gold labels for target languages. However, it suffers from noisy training due to the incorrectly pseudo-labeled samples. In this work, we propose an uncertainty-aware **C**ross-**L**ingual **T**ransfer framework with Pseudo-**P**artial-Label (CLTP)[1] to maximize the utilization of unlabeled data by reducing the noise introduced in the training phase. To estimate pseudo-partial-label for each unlabeled data, we propose a novel estimation method, considering both prediction confidence and the limitation to the number of similar labels. Extensive experiments are conducted on two cross-lingual tasks, including Named Entity Recognition (NER) and Natural Language Inference (NLI) across 40 languages, which shows our method can outperform the baselines on both high-resource and low-resource languages, such as 6.9 on Kazakh (kk) and 5.2 Marathi (mr) for NER.

1 Introduction

The multilingual pre-trained language models such as mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020) and mT5 (Xue et al., 2021) are able to support zero-shot transfer from a source language to target languages. Despite the remarkable performance on direct zero-shot cross-lingual transfer tasks, one would apply semi-supervised learning on target languages to obtain more robust and accurate predictions in a practical scenario. Recent studies (Dong and de Melo, 2019; Xu et al., 2021) validate the effectiveness of self-learning in cross-lingual transfer tasks, utilizing predictions of unlabeled data of target languages as silver labels. Dong and de Melo (2019) iteratively grow the training set by selecting top-k percent of unlabeled data and Xu et al. (2021) boost the performance by considering prediction confidence in the pseudo-label selection. Although their self-learning frameworks significantly improve the transferring performance, it still lags far behind supervised learning. The main reason is that the model suffers from the large number of incorrectly pseudo-labeled samples used in the training phase. Even though they adopt the selection mechanism where those easy and high-confidence predictions will be firstly added into the training set, it cannot guarantee the accurate prediction for all unlabeled data. In fact, the accuracy of the predictions drops quickly in the later iterations (see A.1), since most of the remaining unlabeled data are more difficult to be classified. To avoid the false positive pseudo labels, the most naive way is to select pseudo-labels with extremely high confidence. However, in this way, a large amount of unlabeled data will be discarded due to strictly high



	Confidence	Pseudo-Label	Ground-Truth
Standard	low	(0,0,1,0,0)	x
Partial-Label	high	(0,1,1,0,0)	√

Figure 1: For an unlabeled sample with ambiguous predictions, the standard one-hot-labeling takes the class with the highest confidence as the pseudo-one-hot-label, introducing the noise in the training phase due to the wrong prediction. Instead of choosing one among the predictions that all have low confidence, the proposed partial-labeling method takes both ambiguous classes as candidate labels, allowing the ground-truth label to be presented in the training phase.

[*] Corresponding author.

[1] We release our code at: github.com/slei109/CLTP.

Findings of the Association for Computational Linguistics: NAACL 2022, pages 1987 - 1997
July 10-15, 2022 ©2022 Association for Computational Linguistics

confidence bars. However, the discarded data are still valuable for learning a classifier, especially in the zero-shot task. So, *how to maximize the utilization of the unlabeled data while minimizing the ratio of noisy pseudo labels in the training set?*

Intuitively, if we are confident that the unlabeled data belong to a candidate class set but unable to assign one-hot pseudo labels, it is more efficient to present all potential labels comparing to discarding them directly. Thus, we propose to present pseudo-partial-labels for those data to the model. As illustrated in Figure 1, for an ambiguous sample, the model believes that it belongs to one of two categories with high confidence but has difficulty determining which category it belongs to. In this case, the standard one-hot-labeling takes the class with the highest confidence as the pseudo label, increasing the ratio of noisy pseudo labels in the training phase due to the wrong prediction. By contrast, the proposed partial-labeling method takes both ambiguous classes as candidate labels, allowing the ground-truth label presented in the training phase. In this way, the model can continue to learn on the pseudo-partial-labeled data by disambiguating the candidate labels and finding the latent ground-truth.

In this work, we propose an uncertainty-aware **C**ross-**L**ingual **T**ransfer framework with Pseudo-**P**artial-Label (CLTP) that employs partial label learning to boost cross-lingual zero-shot transfer. Specifically, our framework utilizes any multilingual pre-trained models as the backbone, and iteratively grows the training set by adding predictions of target language data as silver labels. For those difficult data samples with low prediction confidence, different from discarding them directly or introducing a single-hypothetical but incorrect pseudo-label, we associate them with pseudo-partial-labels to better maximize the data utilization. To estimate the pseudo-partial-label, we propose a novel uncertainty-aware estimation method that considers both prediction confidence and the limitation to the number of candidate labels. The model continues to learn on the pseudo-partial-labeled data by disambiguating the candidate labels and finding the latent ground-truth.

Our key contributions can be summarized as follows. 1) We design an uncertainty-aware cross-lingual transfer framework with pseudo-partial-labels. 2) We propose a novel pseudo-partial-label estimation method that considers prediction confidences and the limitation to the number of candidate classes. 3) We evaluate the proposed framework on both NER and NLI tasks across 40 languages in total. Comprehensive experiments show that our framework achieves a strong performance of both high-resource and low-resource languages on both tasks by a sizable margin, such as 6.9 on Kazakh (kk), 5.2 Marathi (mr) for NER and 1% on Arabic (ar), 0.8% on Bulgarian (bg) for NLI.

2 Related Work

Cross-Lingual Representation Learning. Pre-trained transformer-based models have proven effective in learning cross-lingual information. mBERT (Devlin et al., 2019) is pre-trained on raw Wikipedia texts in languages using masked language modeling and next sentence prediction tasks with no explicit cross-lingual objective. XLM-R (Conneau et al., 2020) improves over mBERT by training longer with more data from CommonCrawl, and without the NSP objective. Recently, two self-learning based methods were proposed for cross-lingual transfer. Dong and de Melo (2019) proposed a self-learning framework to incorporate the predictions of mBERT for the cross-lingual text classification task. Xu et al. (2021) improved over the XLM-R by jointly training multiple languages together and considering prediction confidence in the silver labels selection process. However, these two methods still suffer from noisy training because of the incorrect pseudo-labels.

Pseudo-Labeling. Pseudo-labeling (Lee et al., 2013; Shi et al., 2018; Iscen et al., 2019) belongs to the self-learning scenario, and it is often used in semi-supervised learning to generate pseudo-labels for unlabeled samples with a model trained on labeled data. Inspired by noise correction work (Yi and Wu, 2019), Wang and Wu (2020) attempted to update the pseudo-labels through an optimization framework. Recently, Rizve et al. (2021) selected pseudo-labels with both prediction uncertainty and calibration, allowing for negative pseudo-labels generations. However, most existing methods involve learning from noisy data and cannot generalize to partial label learning.

Partial Label Learning. Partial label learning (Cour et al., 2011), also called ambiguously label learning (Chen et al., 2017) and superset label problem (Gong et al., 2017), has subsequently attracted a lot of attention (Feng et al., 2020; Wang and Zhang, 2020; Yao et al., 2020; Yan and Guo, 2020;

Wang et al., 2021). It refers to the task where each training sample is associated with a *set* of candidate labels, while *only one* of them is assumed to be true. Existing studies on the partial label learning can be divided into two groups: average-based methods and identification-based methods. The average-based methods (Cour et al., 2011; Zhang and Yu, 2015; Zhang et al., 2016) consider each candidate label as equally important during model training, and average the outputs of all candidate labels for predictions. The identification-based methods aim at directly maximizing the output of exactly one candidate label, chosen as the truth label. Yan and Guo (2020) studied the utilization of batch label correction; Yao et al. (2020) managed to improve the performance by combining different networks. Wen et al. (2021) proposed the Leveraged Weighted (LW) loss, considering the trade-off between losses on partial labels and non-partial ones. In this work, we adopt the idea of LW loss function for partial label learning due to its effectiveness and generalization.

Uncertainty Estimation. Recently, estimating the uncertainty of deep learning models has attracted increasing attention and have been validated the effectiveness in NLP tasks (Zhang et al., 2019; He et al., 2020). There are two main uncertainty types in Bayesian modeling (Kendall and Gal, 2017; Depeweg et al., 2018): epistemic uncertainty (EU) that captures the model uncertainty itself, which can be explained with more data; aleatoric uncertainty (AU) that captures the intrinsic data uncertainty regardless of models. Another group of uncertainty estimation methods are based on belief/evidence theory through Fuzzy Logic (De Silva, 2018), Dempster-Shafer Theory (Sentz et al., 2002), and Subjective Logic (Sensoy et al., 2018). Belief theorists focus on the reasoning of the inherent uncertainty in information resulting from unreliable, incomplete, deceptive, and/or conflicting evidences. Subjective Logic considers uncertainty in subjective opinions in terms of vacuity (i.e., lack of evidence) (Sensoy et al., 2018), dissonance (i.e., conflicting evidence), and consonance (i.e., composite subsets of state values) (Shi et al., 2020).

Summary Current works on self-learning based cross-lingual transfer methods suffer from noisy training and poor generalization due to the incorrectly pseudo-labeled samples. In this work, we adopt the idea of partial label learning to maximize unlabeled data utilization while reducing the effect of ambiguously pseudo-label estimations in the self-learning framework. To the best of our knowledge, it is the first time pseudo-partial-label employed in the self-learning framework for the cross-lingual transfer and estimating the pseudo-partial-label with the prediction uncertainty.

3 Model

In this section, we propose a partial-label based self-learning framework to boost cross-lingual transfer performance. The overview of the proposed framework is presented in Section 3.2. The technical details for the uncertainty-aware pseudo-partial-label estimation and partial label learning are described in Sections 3.3 and 3.4, respectively.

3.1 Preliminary

Partial label learning refers to the task where each training sample is associated with a *set* of candidate labels, while *only one* of them is assumed to be true. The goal is to find the latent ground-truth for the input through observing the partial label set. Formally, given a non-empty feature space (input space) $\mathcal{X} \subset \mathbb{R}^d$ and a supervised label space $\mathcal{Y}^* = [C] := \{1, \ldots, C\}$, where C is the number of classes and the partial label space is denoted as $\mathcal{Y} := \{y|y \subset \mathcal{Y}^*\}$. For the rest of this paper, let $y^{(i)} = [y_1^{(i)}, \ldots, y_C^{(i)}] \subseteq \{0, 1\}^C$ be the binary vector representing the partial-labels of the instance i , where $y_c^{(i)} = 1$ if class c is selected as the candidate class and $y_c^{(i)} = 0$ if c is not selected. For convenience, we use *k-hot partial labels* to represent the number of candidate classes in the partial label. For example, the partial label in Figure 1 is a two-hot partial label and has two candidate classes.

3.2 Partial-Label based Self-Learning Framework

Our approach aims to improve the overall performance by maximizing the utilization of unlabeled data while reducing the noise introduced in the training phase. This can be accomplished by applying partial label learning on those highly uncertain predictions. The intuition is that if the data appears ambiguous to be classified, it will be more effective to present potential labels instead of discarding them directly or introducing a single-hypothetical but incorrect pseudo-label in the training phase.

The complete training procedure of our proposed task-agnostic framework for cross-lingual transfer is shown in Figure 2. In our proposed CLTP frame-

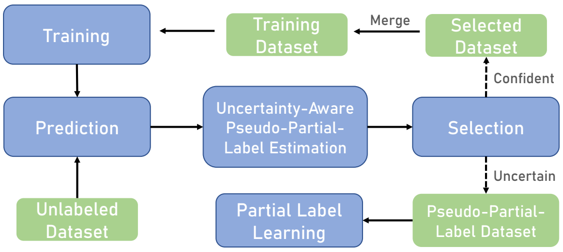


Figure 2: Illustration of the uncertainty-aware cross-lingual transfer framework with pseudo-partial-labels.

work, we first train a pre-trained multilingual model on the gold labels of the source language. Then the model makes predictions on the unlabeled dataset of the target languages. The proposed uncertainty-aware estimation component generates the pseudo-partial-labels based on the model predictions and their corresponding uncertainty estimations. After that, we adopt a selection mechanism to incorporate the unlabeled data with high confidence scores into the training phase.

The whole training process for our method is described in Algorithm 1, where $\mathcal{D}_u$ denotes the set of tuples combining unlabeled data with corresponding pseudo-partial-labels and prediction uncertainty γ_{diss}. First, a model $f(\cdot)$ is trained on the gold labels of the source language in the first iteration. Once trained, the model can make predictions and estimate the pseudo-partial-labels for all unlabeled data of target languages in D_u based on the method introduced in Section 3.3. Note that the inputs of different languages are mixed together. Next, a subset of the pseudo-partial-labels S_u is selected with the uncertainty estimation. After selection, the model goes back to the training phase using the selected pseudo-partial-labels as well as the gold labels. We repeat the process iteratively until max iteration is reached. The early stop criteria are implemented on the dev set of the source language only since the gold labels are not available for the other languages. Note that the model is trained only on the one-hot pseudo-label in the first three iterations to accelerate the convergence.

3.3 Pseudo-Partial-Label Estimation

The key point of pseudo-partial-label estimation is to guarantee that the ground-truth class of an instance resides in the candidate label set, which is the basic definition for partially supervised learning. Intuitively, if the model classifies an instance with lower confidence, the instance may be hard to distinguish from several classes or cannot be identified due to the lack of knowledge. Hence, we consider prediction confidence in the pseudo-partial-label estimation to better determine the most uncertain classes for each ambiguous instance.

Algorithm 1 Self-Learning on Cross-Lingual Tasks

```
Input: A gold label dataset $\mathcal{D}_L$ of source language,
       an unlabeled dataset of target languages $\mathcal{U}$.
repeat
    Training $f(\cdot)$ on $\mathcal{D}_L$ with $\mathcal{L}_{EVI}$
    for each target language do
        $\mathcal{D}_u \leftarrow \varnothing$
        for $x_u$ in $\mathcal{U}$ do
            $(\tilde{y}, \gamma_{diss}) \leftarrow f_\theta(x_u)$
            $\mathcal{D}_u \leftarrow \mathcal{D}_u \cup \{(x_u, \tilde{y}, \gamma_{diss})\}$
        end for
        $\mathcal{S}_u \leftarrow \underset{S \subset \mathcal{D}_u, |S| \leq N}{\arg\min} \sum_{(x_u, \gamma_{diss}) \in S} \gamma_{diss}$
        $\mathcal{U} \leftarrow \mathcal{U} \setminus \mathcal{S}_u$
        $\mathcal{U}_s \leftarrow \mathcal{U}_s \cup \mathcal{S}_u$
        $\mathcal{D}_L \leftarrow \mathcal{D}_L \cup \mathcal{U}_s$
    end for
until max iteration arrives.
Training $f(\cdot)$ on $\mathcal{D}_L$ with $\mathcal{L}_\varphi$ until converge
```

We define the prediction uncertainty of the instance x belonging to the partial-label y as the partial-label uncertainty, which is denoted as $\gamma_{diss}(x, y)$. The decomposed entropy *dissonance* proposed by Shi et al. (2020) is adapted to calculate the partial-label uncertainty, as it can indicate the contradiction among certain classes. Specifically, *dissonance* is an evidence-based uncertainty (Sensoy et al., 2018) where the softmax probability is replaced by Dirichlet distribution, and each predicted logit for class c is regarded as the evidence e_c. The expected probability p_c for class c under Dirichlet distribution is defined as follows.

$$p_c = \frac{e_c + 1}{S}, \text{with } S = \sum_c e_c + C \tag{1}$$

where S is referred to as the Dirichlet strength. We adopt the cross-entropy loss as the training loss $\mathcal{L}_{\text{EVI}}$, and its Bayes risk under the Dirichlet distribution can be defined as

$$\begin{aligned}\mathcal{L}_{\text{EVI}} &= \int \Big[\sum_c -y_c \log(p_c)\Big] \frac{1}{Beta(\boldsymbol{\alpha})} \prod_c p_c^{\alpha_c - 1} d\mathbf{p}_c \\ &= \sum_c -y_c \int \log(p_c) \frac{1}{Beta(\boldsymbol{\alpha})} \prod_c p_c^{\alpha_c - 1} d\mathbf{p}_c \\ &= \sum_c y_c(\psi(S) - \psi(e_c + 1))\end{aligned} \tag{2}$$

where $Beta(\boldsymbol{\alpha})$ is the multinomial beta function (Kotz et al., 2004) and $\psi(\cdot)$ is the *digamma* function. $\boldsymbol{\alpha}$ is the parameters of the Dirichlet density on the predictors. As shown in Equation (2), training the model with $\mathcal{L}_{\mathrm{EVI}}$ is to make the positive evidence close to the total evidence when the ground-truth is positive. If there are conflicts of strong evidence among certain classes, *dissonance* will become high to indicate the contradiction. The following describes the *dissonance* for each instance:

$$\mathrm{Bal}(b_j, b_k) = \begin{cases} 1 - \frac{|b_j - b_k|}{b_j + b_k}, & \text{if } b_j b_k \neq 0 \\ 0, & \text{elsewise} \end{cases} \quad (3)$$

$$diss = \sum_c \frac{b_c \sum_{c' \neq c} b_{c'} \mathrm{Bal}(b_c, b_{c'})}{\sum_{c' \neq c} b_{c'}} \quad (4)$$

where $b_c = e_c/S$ represents the belief mass for class c. Recall that each predicted logit for class c is regarded as the evidence e_c.

In the pseudo-partial-label estimation, the belief mass for the instance belongs to a candidate class set can be calculated via Binomial Comultiplication operator in subjective logic, which is denoted as '$\vee$'. Let b_c and b_h be the belief mass for class c and class h, respectively. The belief mass for the instance belongs to class c or class h is defined as:

$$b_{c \vee h} = b_c + b_h - b_c b_h \quad (5)$$

Thus, we can calculate the partial-label uncertainty $\gamma_{diss}(x, y)$ via Equation (4) and (5).

However, if we simply estimate the pseudo-partial-label only based on the lowest partial-label uncertainty, it will lead to an invalid partial label like $(1, 1, 1, 1)$, as containing all classes in the candidate set must have the highest confidence. Furthermore, in the partial label learning, the accuracy of the model decreases with the increased number of similar labels to the true label (Lv et al., 2020; Wen et al., 2021) because it increases the learning difficulty. To remedy this problem, we leverage a penalty ratio to balance the prediction uncertainty and the number of candidate classes. Specifically, from Figure.3, we observe that as the number of candidate classes increases, the improvement in prediction recall tends to decrease. That means containing too many candidate classes in the pseudo-partial-label has limited improvement to the model. Only the most confusing candidate classes are important for those ambiguous samples. Motivated by this, we employ a penalty ratio to punish a larger number of candidate classes. Thus, a pseudo-partial-label $\tilde{y}$ is obtained as follows:

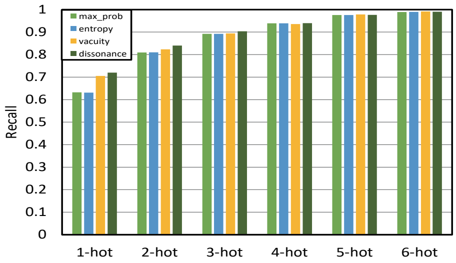


Figure 3: Recall of k-hot pseudo-partial-labels in the last iteration with various uncertainty methods.

$$\tilde{y} = \arg\min_{y \subset \mathcal{Y}} ((\lambda^{||y||_1 - 1} + \tau)\tau^{||y||_1 - 2} \gamma_{diss}(x, y)) \quad (6)$$

where $\mathcal{Y}$ is the collection of all subsets in the partial label space and $||y||_1$ calculates the number of candidate classes in the partial label y. λ and τ are penalty ratios to punish a larger number of candidate classes, which determine the penalty strength and the preference for candidate class number, respectively.

3.4 Learning with Pseudo-Partial-Labels

The target of partial-label learning is to learn a classifier with access to the candidate label set (partial label set) by disambiguating the candidate labels and finding the latent ground-truth for the input during the training phase. We adopt Leveraged Weighted (LW) loss function (Wen et al., 2021) for partial label learning due to its effectiveness and generalization. The LW loss is a multiclass loss and searches for the latent ground-truth by assigning more weights to the loss of classes more likely to be the true and lessening weights to the confusing ones. To be specific, the LW loss function is defined as:

$$L_\varphi(\tilde{y}, f(x)) = \sum_{\tilde{y}_c = 1} \omega_c \varphi(f_c(x)) + \beta \sum_{\tilde{y}_c = 0} \omega_c \varphi(-f_c(x)), \quad (7)$$

where $\varphi(\cdot) : \mathbb{R} \to \mathbb{R}^+$ denotes a binary loss and we adopt the Sigmoid loss function. We use parameter β to distinguish the effects between candidate classes and non-candidate ones. $f_c(x)$ represents the predicted logit of instance x for class c, while w_c is the weighting parameters to assign weights to the loss of classes. Since the key point is to disambiguate the candidate classes, the model is supposed to assign more weights to the loss of classes that are more likely to be the ground-truth. Thus, instead of assigning fixed values, the weighting parameters are updated by normalizing the prediction score through an iterative learning process. Specifically, at the t-th learning step, $w_c^{(t)}$ is calculated as follows:

$$\omega_c^{(t)} = \begin{cases} \frac{\exp(f_c^{(t)}(x))}{\sum_{\tilde{y}_c=1}\exp(f_c^{(t)}(x))}, & \text{if } \tilde{y}_c = 1 \\ \\ \frac{\exp(f_c^{(t)}(x))}{\sum_{\tilde{y}_c=1}\exp(f_c^{(t)}(x))}, & \text{if } \tilde{y}_c = 0 \end{cases} \tag{8}$$

Note that $w_c^{(t)}$ varies with sample instances. In this way, as the training epochs grow, the model focuses on the true class and rules out the untrue classes by penalizing large value of $\varphi(-f_c(x))$.

4 Experiments

4.1 Evaluation Tasks & Datasets

NLI XNLI (Conneau et al., 2018) is an evaluation benchmark for the cross-lingual NLI task across 15 languages. Given a sentence pair of premise and hypothesis, the task is to classify their relationship as "neutral", "entailment", or "contradiction".

NER Wikiann (Pan et al., 2017) is an evaluation benchmark for the cross-lingual NER task covering 40 languages. There are three entity types: "LOC", "PER" and "ORG", and each token is tagged in the BIO2 format with 7 label types.

We follow the same train/dev/test split and same evaluation protocol as XTREME (Hu et al., 2020). English is the source language with gold labels for both datasets, and we use the dev set of target languages as the source of unlabeled data. Gold labels of target languages cannot be accessed in the self-learning process.

4.2 Implementation Details

Model Details. We keep the same model architecture throughout our experiments: XLM-$\mathrm{R}_{\mathrm{Large}}$ (Conneau et al., 2020) is used as the multilingual pre-trained model to encode input sequence, followed by a linear layer to classify on the hidden state, which is the same model setting from XTREME. We set the penalty ratios $\lambda = 4$ and $\tau = 10$ for all experiments. For LW loss, we set $\beta = 2$ as suggested by Wen et al. (2021).

Training Details. For both NER and NLI tasks, we use the AdamW (Loshchilov and Hutter, 2018) optimizer with a linear learning rate scheduler for all experiments. We use a batch size of 32 and a max sequence length of 128. We first train the model by 10 epochs on English training set with gold labels for the NER task and 5 epochs for the NLI task with a 2×10^{-5} learning rate. In the self-learning process, we keep the same learning rate for the NER task and set a 5×10^{-6} learning rate to train the model for the NLI task. The model is trained for 3 epochs in each iteration. Experiments are run on a single 24GB NVIDIA 3090 GPU.

4.3 Baselines.

We compare our CLTP framework with three different settings for the baselines: *BL-Direct* is equivalent to Hu et al. (2020), which is the direct zero-shot transfer without utilizing unlabeled data of target languages. *BL-Single* takes silver labels of only one target language as the training set in the self-learning process and simply uses model predictions as silver labels without considering prediction confidences. *BL-Joint* is similar to *BL-Single* but instead takes silver labels of all target languages jointly. We also compare our method with uncertainty-aware self-learning framework (Xu et al., 2021): *SL-LEU* trains the model with silver labels and selects them by considering Language Heteroscedastic Uncertainty (LEU) and *SL-EVI* takes Evidential Uncertainty (EVI) to estimate the prediction uncertainty, following the same training settings as Xu et al. (2021) utilized.

4.4 Comparisons

The results of NER task and NLI task are shown in Tables 1 and 2, respectively. Self-learning based methods outperform the direct zero-shot transfer with XLM-$\mathrm{R}_{\mathrm{large}}$ by a large margin in NER, achieving 11.1 gain in F1 on average. The trend of improvement can also be observed in NLI, validating the self-learning strategy on cross-lingual transfer tasks. Furthermore, our method surpasses SL-LEU on NER by 2.2 in F1 on average, demonstrating the effectiveness of utilizing pseudo-partial-label for those ambiguous data. Remarkably, our method achieves a sizeable gain, 5+ in F1 on both low-resource languages like Malayalam (ml) and Marathi (mr), and high-resource languages such as Russian (ru) and Tegulu (te). This shows that our method can further boost the model performance of the self-learning framework. As shown in Ta-

	en	af	ar	bg	bn	de	el	es	et	eu	fa	fi	fr	he	hi	hu	id	it	ja	jv	ka	kk	ko	ml	mr	ms	my	nl	pt	ru	sw	ta	te	th	tl	tr	ur	vi	yo	zh	**avg**
mBERT	**85.2**	77.4	41.1	77.0	70.0	78.0	72.5	77.4	75.4	66.3	46.2	77.2	79.6	56.6	65.0	76.4	53.5	81.5	29	66.4	64.6	45.8	59.6	52.3	58.2	72.7	45.2	81.8	80.8	64.0	67.5	50.7	48.5	3.6	71.7	71.8	36.9	71.8	44.9	42.7	62.2
XLM	82.6	74.9	44.8	76.7	70.0	78.1	73.5	74.8	74.8	62.3	49.2	79.6	78.5	57.7	66.1	76.5	53.1	80.7	23.6	63.0	67.7	57.2	26.3	59.4	62.4	69.6	47.6	81.2	77.9	63.5	68.4	53.6	49.6	0.3	78.6	71.0	43.0	70.1	26.5	32.4	61.2
MMTE	77.9	74.9	41.8	75.1	64.9	71.9	68.3	71.8	74.9	62.6	45.6	75.2	73.9	54.2	66.2	73.8	47.9	74.1	31.2	63.9	60.9	43.9	58.2	44.8	58.5	68.3	42.9	74.8	72.9	58.2	66.3	48.1	46.9	3.9	64.1	61.9	37.2	68.1	32.1	28.9	58.3
VECO	83.8	77.5	48.2	83.9	77.2	79.4	79.3	75.4	80.4	68.3	68.2	80.6	80.1	55.0	71.0	80.9	52.9	81.7	19.4	63.2	67.1	51.2	59.9	63.4	65.0	70.0	56.1	83.4	83.1	71.3	70.5	60.5	56.2	1.4	71.3	80.4	69.3	76.0	37.4	29.1	65.7
BL-Direct*	84.0	79.3	45.5	81.4	77.4	78.8	78.9	71.4	79.0	61.0	52.0	78.7	79.3	54.6	70.8	79.4	52.9	81.0	25.0	62.6	69.3	51.9	57.9	63.6	62.4	69.6	60.1	83.7	80.9	70.2	69.2	58.2	51.3	1.8	71.0	76.7	55.8	76.2	41.4	33.0	64.4
BL-Single*	84.0	78.9	56.9	84.5	79.3	80.9	81.6	72.9	80.7	63.2	54.8	80.5	81.9	63.0	73.9	81.7	54.3	82.1	36.5	60.9	73.6	52.5	63.6	66.0	66.8	62.6	54.3	84.8	82.6	72.9	67.7	63.2	57.2	3.1	74.7	81.8	69.9	80.9	46.2	43.6	67.5
BL-Joint*	84.7	79.5	56.7	84.9	80.5	80.5	81.5	73.3	81.2	64.0	55.1	81.2	82.1	62.6	76.6	81.6	54.5	83.0	37.2	63.5	73.6	53.4	63.6	67.5	67.9	64.3	53.0	84.8	83.2	73.5	69.7	63.1	57.4	3.6	76.1	81.8	71.5	81.4	**54.8**	43.7	68.3
SL-EVI	85.0	84.3	69.2	85.5	78.9	82.4	82.4	79.0	85.0	**76.7**	73.8	84.6	81.5	57.3	79.4	83.6	**58.5**	83.9	**47.7**	70.0	74.2	60.7	63.3	61.8	75.0	**73.9**	67.2	86.4	84.0	80.3	**73.1**	64.7	63.2	**8.0**	**81.4**	81.6	74.6	84.1	49.6	**54.0**	72.3
SL-LEU	84.4	83.3	62.3	**86.9**	81.5	83.4	83.9	82.6	85.3	75.1	82.7	85.3	84.3	67.5	77.7	84.1	57.2	84.4	44.9	**73.6**	74.7	56.6	69.4	73.9	74.7	73.6	68.0	86.1	**86.0**	75.9	71.5	68.1	63.9	6.8	79.4	**88.0**	84.2	85.0	45.9	53.0	73.3
Ours	**85.0**	**86.0**	**71.7**	85.5	**83.4**	**83.7**	**85.1**	**86.5**	**86.5**	75.6	**83.1**	**85.7**	**84.4**	**68.5**	**80.8**	**87.3**	57.2	**84.9**	47.4	71.1	**81.6**	**65.0**	**71.7**	**78.8**	**80.2**	73.5	**71.6**	**87.5**	85.9	**81.8**	72.2	**71.4**	**69.1**	7.4	81.0	87.1	**86.3**	**86.0**	48.8	53.0	**75.5**

Table 1: NER Results in F1 scores for 40 languages. *Results are reported by Xu et al. (2021).

ble 2, our model outperforms the baselines on NLI across almost all 15 test languages. Comparing to the direct zero-shot, our method achieves an improvement of 2.4% on average. Our method also gives an average increase of 1.2% and 0.5% on SL-EVI and SL-LEU, respectively. Specifically, we observe over 1% gain for Arabic (ar), Bulgarian (bg), Greek (el), and Turkish (tr) when we compare CLTP framework with the best performance of SL.

4.5 Result Analyses

To better understand the key components and settings of CLTP framework, we perform some analyses on the NER task.

Uncertainty Estimation. To assess different uncertainty estimations for pseudo-partial-label estimation, we evaluate the recall score of the pseudo-partial-labels in the last iteration, such that recall is high when the pseudo-partial-label contains the true class. Here, we adopt the two-hot partial label setting where the two classes with the highest prediction confidence were set as candidate classes, as it is the best setting to directly measure the contradiction among certain classes. We compare evidential uncertainty with two commonly used uncertainty metrics for classification (Depeweg et al., 2018; Dong and de Melo, 2019; Xiao and Wang, 2019): the max probability of label classes (i.e., max_prob) and the entropy of the class probability distribution (i.e., entropy). As shown in Figure 4, *dissonance* achieves the best performance among all uncertainty estimations on average, demonstrating its capability of ambiguous class selection.

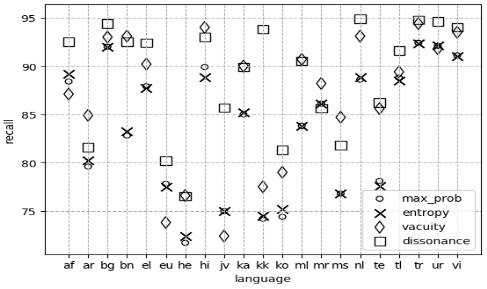


Figure 4: Recall of different uncertainty estimations in pseudo-partial-label estimation (two-hot partial-labels).

Hyperparameter Analyses. We introduce new hyperparameters λ and τ to control the penalty ratio. Table 3 shows the evaluation accuracy in different penalty ratio settings. We find that using $\lambda = 4, \tau = 10$ leads to the best performance, and further reduction/increase in the ratio lead to performance degradation. The penalty ratio does affect the performance of the model as it adjusts the proportion of various pseudo-partial-labels. In

	en	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh	**avg**
mBERT	80.8	64.3	68	70	65.3	73.5	73.4	58.9	67.8	49.7	54.1	60.9	57.2	69.3	67.8	65.4
XLM	82.8	66.0	71.9	72.7	70.4	75.5	74.3	62.5	69.9	58.1	65.5	66.4	59.8	70.7	70.2	69.1
MMTE	79.6	64.9	70.4	68.2	67.3	71.6	69.5	63.5	66.2	61.9	66.2	63.6	60.0	69.7	69.2	67.5
VECO	88.2	79.2	83.1	82.9	81.2	84.2	82.8	76.2	80.3	74.3	77.0	78.4	71.3	80.4	79.1	79.9
BL-Direct*	88.5	78.0	82.5	81.8	80.5	83.8	82.9	74.8	78.7	67.5	76.7	78.1	71.5	79.4	78.2	78.9
BL-Single*	88.5	77.6	82.4	82.0	79.6	82.5	82.1	76.1	79.1	69.1	76.6	77.9	71.5	77.9	78.2	78.7
BL-Joint*	88.2	78.8	82.0	82.2	80.4	83.1	82.2	76.1	79.6	68.8	76.2	78.0	71.4	79.1	78.5	79.0
SL-EVI	88.1	79.6	83.3	82.9	81.6	83.7	81.7	77.5	80.1	72.3	78.2	78.9	74.1	79.7	79.8	80.1
SL-LEU	88.5	79.5	83.7	83.4	82.4	84.1	83.8	**78.3**	80.9	73.2	79.4	79.1	74.4	80.4	81.1	80.8
Our	**88.6**	**80.6**	**84.5**	**83.8**	**83.5**	**84.9**	**83.9**	78.2	**81.4**	**73.5**	**79.7**	**80.2**	**74.4**	**80.8**	**81.3**	**81.3**

Table 2: XNLI accuracy score for *English (en), French (fr), Spanish (es), German (de), Greek (el), Bulgarian (bg), Russian (ru), Turkish (tr), Arabic (ar), Vietnamese (vi), Thai (th), Chinese (zh), Hindi (hi), Swahili (sw) and Urdu (ur).**Results are reported by Xu et al. (2021).

Settings	**average (F1)**
$\lambda = 2, \tau = 10$	74.6
$\lambda = 4, \tau = 5$	72.9
$\lambda = 4, \tau = 10$	**75.5**
$\lambda = 8, \tau = 10$	73.4

Table 3: Hyperparameter analyses to the penalty ratio on the NER task. Different settings adjust the proportion of pseudo-partial-labels.

Settings	**avg**
self-learning + no partial labels	73.3
self-learning + two-hot partial labels	75.2
self-learning + three-hot partial labels	73.1
self-learning + four-hot partial labels	72.5
CLTP (ours)	**75.5**

Table 4: Effect of pseudo-partial-label length on the NER task. Note that *two-hot partial labels* indicates that the pseudo-partial-labels are directly estimated by selecting the two classes with the max prediction probability and *no partial-labels* is equivalent to SL.

specific, when τ stays the same, as λ increases, the proportion of partial-labels decreases. It indicates that most pseudo-labels are one-hot labels that are similar to the self-learning framework proposed by Xu et al. (2021). Similarly, if we keep λ constant, reducing τ will lead to not only a higher proportion of partial-labels over the unlabeled set but also more candidate classes in each partial-label. We observe an over 2.6 gain when we compare $\lambda = 4, \tau = 10$ setting with $\lambda = 4, \tau = 5$ setting, indicating that if we ignore the limitation on the number of candidate classes, the model will suffer from invalid partial labels like $(1, 1, 1, 1)$ because it cannot provide any information when all classes are set as candidates.

Effect of pseudo-partial-label length. To analyze the effect of different number of candidate classes in pseudo-partial-label estimation, we evaluate the model trained with various manually designed pseudo-partial-labels. Specifically, we select the top classes with the prediction confidence to set pseudo-partial-labels. The results are shown in Table 4. When we utilize the manually set partial-labels, we observe an accuracy drop of 0.3, 2.4 and 3.0 in *two-hot partial labels*, *three-hot partial labels* and *four-hot partial labels*, respectively. This demonstrates the effectiveness of our pseudo-partial-label estimation scheme. In addition, the model trained with *two-hot partial labels* outperforms the baselines. By contrast, *three-hot partial labels* and *four-hot partial labels* do not surpass the baselines, partially due to the difficulty of disambiguating the candidate classes. The trend of improving performance with fewer candidate classes is consistent with the phenomenon in partial label learning (Lv et al., 2020; Wen et al., 2021).

5 Conclusion

In this work, we propose an uncertainty-aware pseudo-partial-label framework for cross-lingual transfer. With the auxiliary of pseudo-partial-labels, CLTP framework improves the model by reducing the noise introduced in training phase while maximizing unlabeled data utilization. Moreover, we propose a novel pseudo-partial-label estimation method that considers both prediction confidence and the limitation to the number of similar classes. The proposed framework is evaluated on two tasks of NER and NLI and improves the performance of the pre-trained model by a solid margin (11.1 F1 for NER and 2.4% accuracy score for NLI on average). Compared to other self-learning based methods, our framework surpasses the baselines on both high-resource and low-resource languages, such as 6.9 on Kazakh and 5.2 Marathi for NER.

References

Ching-Hui Chen, Vishal M Patel, and Rama Chellappa. 2017. Learning from ambiguously labeled face images. *IEEE transactions on pattern analysis and machine intelligence*, 40(7):1653–1667.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485.

Timothee Cour, Ben Sapp, and Ben Taskar. 2011. Learning from partial labels. *The Journal of Machine Learning Research*, 12:1501–1536.

Clarence W De Silva. 2018. *Intelligent control: fuzzy logic applications*. CRC press.

Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. 2018. Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1184–1193. PMLR.

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.

Xin Dong and Gerard de Melo. 2019. A robust self-learning framework for cross-lingual 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 6306–6310, Hong Kong, China. Association for Computational Linguistics.

Lei Feng, Jiaqi Lv, Bo Han, Miao Xu, Gang Niu, Xin Geng, Bo An, and Masashi Sugiyama. 2020. Provably consistent partial-label learning. *Advances in Neural Information Processing Systems*, 33:10948–10960.

Chen Gong, Tongliang Liu, Yuanyan Tang, Jian Yang, Jie Yang, and Dacheng Tao. 2017. A regularization approach for instance-based superset label learning. *IEEE transactions on cybernetics*, 48(3):967–978.

Jianfeng He, Xuchao Zhang, Shuo Lei, Zhiqian Chen, Fanglan Chen, Abdulaziz Alhamadani, Bei Xiao, and ChangTien Lu. 2020. Towards more accurate uncertainty estimation in text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8362–8372, 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 *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. 2019. Label propagation for deep semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5070–5079.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 5580–5590.

Samuel Kotz, Narayanaswamy Balakrishnan, and Norman L Johnson. 2004. *Continuous multivariate distributions, Volume 1: Models and applications*. John Wiley & Sons.

Dong-Hyun Lee et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Jiaqi Lv, Miao Xu, Lei Feng, Gang Niu, Xin Geng, and Masashi Sugiyama. 2020. Progressive identification of true labels for partial-label learning. In *International Conference on Machine Learning*, pages 6500–6510. PMLR.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. 2021. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv preprint arXiv:2101.06329*.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. 2018. Evidential deep learning to quantify classification uncertainty. *Advances in Neural Information Processing Systems*, 31.

Kari Sentz, Scott Ferson, et al. 2002. *Combination of evidence in Dempster-Shafer theory*, volume 4015. Sandia National Laboratories Albuquerque.

Weishi Shi, Xujiang Zhao, Feng Chen, and Qi Yu. 2020. Multifaceted uncertainty estimation for label-efficient deep learning. *Advances in Neural Information Processing Systems*, 33.

Weiwei Shi, Yihong Gong, Chris Ding, Zhiheng MaXiaoyu Tao, and Nanning Zheng. 2018. Transductive semi-supervised deep learning using min-max features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 299–315.

Deng-Bao Wang, Min-Ling Zhang, and Li Li. 2021. Adaptive graph guided disambiguation for partial label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Guo-Hua Wang and Jianxin Wu. 2020. Repetitive reprediction deep decipher for semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6170–6177.

Wei Wang and Min-Ling Zhang. 2020. Semi-supervised partial label learning via confidence-rated margin maximization. *Advances in neural information processing systems*, 33:6982–6993.

Hongwei Wen, Jingyi Cui, Hanyuan Hang, Jiabin Liu, Yisen Wang, and Zhouchen Lin. 2021. Leveraged weighted loss for partial label learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11091–11100. PMLR.

Yijun Xiao and William Yang Wang. 2019. Quantifying uncertainties in natural language processing tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7322–7329.

Liyan Xu, Xuchao Zhang, Xujiang Zhao, Haifeng Chen, Feng Chen, and Jinho D. Choi. 2021. Boosting cross-lingual transfer via self-learning with uncertainty estimation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6716–6723, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Yan Yan and Yuhong Guo. 2020. Partial label learning with batch label correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6575–6582.

Yao Yao, Chen Gong, Jiehui Deng, and Jian Yang. 2020. Network cooperation with progressive disambiguation for partial label learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 471–488. Springer.

Kun Yi and Jianxin Wu. 2019. Probabilistic end-to-end noise correction for learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7017–7025.

Min-Ling Zhang and Fei Yu. 2015. Solving the partial label learning problem: An instance-based approach. In *Twenty-fourth international joint conference on artificial intelligence*.

Min-Ling Zhang, Bin-Bin Zhou, and Xu-Ying Liu. 2016. Partial label learning via feature-aware disambiguation. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1335–1344.

Xuchao Zhang, Fanglan Chen, Chang-Tien Lu, and Naren Ramakrishnan. 2019. Mitigating uncertainty in document classification. 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 3126–3136, Minneapolis, Minnesota. Association for Computational Linguistics.

A Appendix

A.1 Relationship between pseudo-label accuracy and the model performance

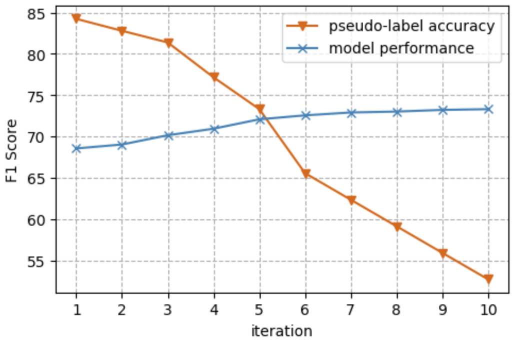


Figure 5: Existing self-learning method for cross-lingual transfer suffers from noisy training due to the incorrectly pseudo-labeled samples. Xu et al. (2021) select top-k percent of unlabeled data with uncertainty score. The entire process keeps iterating until there is no remaining unlabeled data. As the precision of the pseudo-labels decreases, the performance improvement of the model decreases.

We empirically analyze the relationship between pseudo-label accuracy and model performance. From Figure 5, we find that F1 score of newly selected pseudo-labels in each iteration drops quickly, especially in the later 5 iterations. Furthermore, as the precision of the pseudo-labels decreases, the performance improvement of the model decreases.

A.2 ISO Language

Table 5 introduces the ISO 639-1 Code of target languages in NER task.

ISO 639-1 Code	Name of Language
en	English
af	Afrikaans
ar	Arabic
bg	Bulgarian
bn	Bengali
de	German
el	Greek, Modern (1453-)
es	Spanish; Castilian
et	Estonian
eu	Basque
fa	Persian
fi	Finnish
fr	French
he	Hebrew
hi	Hindi
hu	Hungarian
id	Indonesian
it	Italian
ja	Japanese
jv	Javanese
ka	Georgian
kk	Kazakh
ko	Korean
ml	Malayalam
mr	Marathi
ms	Malay
my	Burmese
nl	Dutch; Flemish
pt	Portuguese
ru	Russian
sw	Swahili
ta	Tamil
te	Telugu
th	Thai
tl	Tagalog
tr	Turkish
ur	Urdu
vi	Vietnamese
yo	Yoruba
zh	Chinese

Table 5: ISO 639-1 Code for Representation of Names of Languages