Uncertainty-Aware Cross-Lingual Transfer with Pseudo Partial Labels with Image-level Annotations

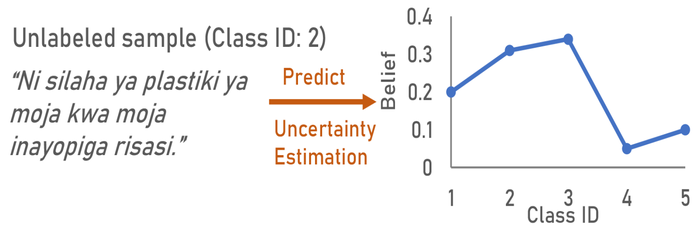
Shuo Lei[1], Xuchao Zhang[1, 2], Jianfeng He[1], Fanglan Chen[1], Chang-Tien Lu[1]

[1]Department of Computer Science, Virginia Tech, Falls Church, VA, USA

[2]NEC Laboratories America, Princeton, NJ, USA

NAACL 2022

INTRODUCTION



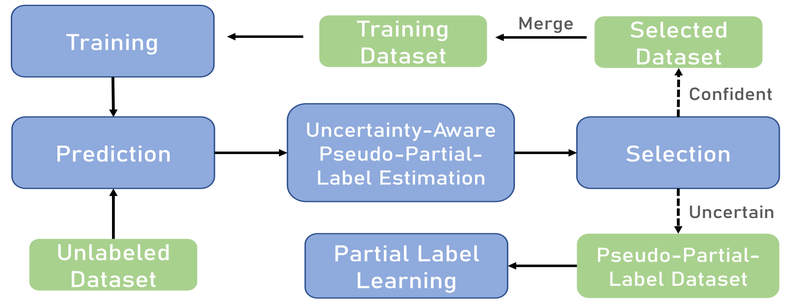
	Confidence	Pseudo-Label	Ground-Truth
Standard	low	(0,0,1,0,0)	x
Partial-Label	high	(0,1,1,0,0)	√

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. 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) to maximize the utilization of unlabeled data by presenting ambiguous classes in the training phase.

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.

CLTP Framework



A pre-trained multilingual model is trained 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, a selection mechanism is adopted to incorporate the unlabeled data with high confidence scores into the training phase.

EXPERIMENTS

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.

	en	af	ar	bg	bn	de	el	es	et	eu	fa	fi	fr	he	hi	hu	id	it	ja	jv	
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	
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	
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	
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	
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	
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	
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	
	ka	kk	ko	ml	mr	ms	my	nl	pt	ru	sw	ta	te	th	tl	tr	ur	vi	yo	zh	**avg**
mBERT	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	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
BL-Direct*	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*	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*	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	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
Ours	**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).

Pseudo-Partial-Label Estimation

The prediction uncertainty of the instance x belonging to the partial-label y is defined as the partial-label uncertainty, which is denoted as $\gamma_{diss}(x, y)$. The decomposed entropy dissonance is adapted to calculate the partial-label uncertainty, as it can indicate the contradiction among certain classes. 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:

$$diss = \sum_c \frac{b_c \sum_{c' \neq c} b_{c'} \mathrm{Bal}(b_c, b_{c'})}{\sum_{c' \neq c} b_{c'}} \text{ where, } \mathrm{Bal}(b_j, b_k) = \begin{cases} 1 - \frac{|b_j - b_k|}{b_j + b_k}, & \text{if } b_j b_k \neq 0 \\ 0, & \text{elsewise} \end{cases}$$

where $b_c = e_c / S$ represents the belief mass for class c.

A pseudo-partial-label $\tilde{y}$ is obtained as follows:

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

$\mathcal{Y}$ 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.