Boosting Cross-Lingual Transfer via Self-Learning with Uncertainty Estimation

Liyan Xu[†]   **Xuchao Zhang**[‡]   **Xujiang Zhao**[§]
Haifeng Chen[‡]   **Feng Chen**[§]   **Jinho D. Choi**[†]

[†]Emory University, {`liyan.xu,jinho.choi`}@emory.edu
[‡]NEC Laboratories America, {`xuczhang,haifeng`}@nec-labs.com
[§]University of Texas at Dallas, {`xujiang.zhao,feng.chen`}@utdallas.edu

Abstract

Recent multilingual pre-trained language models have achieved remarkable zero-shot performance, where the model is only finetuned on one source language and directly evaluated on target languages. In this work, we propose a self-learning framework that further utilizes unlabeled data of target languages, combined with uncertainty estimation in the process to select high-quality silver labels. Three different uncertainties are adapted and analyzed specifically for the cross lingual transfer: Language Heteroscedastic/Homoscedastic Uncertainty (LEU/LOU), Evidential Uncertainty (EVI). We evaluate our framework with uncertainties on two cross-lingual tasks including Named Entity Recognition (NER) and Natural Language Inference (NLI) covering 40 languages in total, which outperforms the baselines significantly by 10 F1 on average for NER and 2.5 accuracy score for NLI.

1 Introduction

Recent multilingual pre-trained language models such as mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020) and mT5 (Xue et al., 2021) have demonstrated remarkable performance on various direct zero-shot cross-lingual transfer tasks, where the model is finetuned on the source language, and directly evaluated on multiple target languages that are unseen in the task-finetuning stage. While direct zero-shot transfer is a sensible testbed to assess the multilinguality of language models, one would apply supervised or semi-supervised learning on target languages to obtain more robust and accurate predictions in a practical scenario.

In this work, we investigate self-learning (also known as "pseudo labels") as one way to apply semi-supervised learning on cross-lingual transfer, where only unlabeled data of target languages are required, without any efforts to annotate gold labels for target languages. As self-learning has been proven effective in certain tasks of computer vision

(Yalniz et al., 2019; Xie et al., 2020) and natural language processing (Artetxe et al., 2018; Dong and de Melo, 2019; Karan et al., 2020), we propose to formalize an iterative self-learning framework for multilingual tasks using pre-trained models, combined with explicit uncertainty estimation in the process to guide the cross-lingual transfer.

Our self-learning (SL) 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. We reckon two important observations from our preliminary study (baselines in §4). First, compared with self-training one target language at a time, jointly training multiple languages together can improve the performance on most languages, especially for certain low-resource languages that can achieve up to 8.6 F1 gain in NER evaluation. Therefore, our SL framework features the joint training strategy, maximizing potentials of different languages benefiting each other. Second, compared with simply using all unlabeled data as silver labels without considering prediction confidence, estimating uncertainties becomes critical in the transfer process, as higher quality of silver labels should lead to better performance. We hence introduce three different uncertainty estimations in the SL framework.

Specifically, we adapt uncertainty estimation techniques based on variational inference and evidence learning for our cross-lingual transfer, namely LEU, LOU and EVI (§3.2). We evaluate our framework and three uncertainties on two multilingual tasks from XTREME (Hu et al., 2020): Named Entity Recognition (NER), and Natural Language Inference (NLI). Empirical results suggest LEU to be the best uncertainty estimation overall, while the others can also perform well on certain languages (§4.1). Our analysis shows further evaluation of different estimations, corroborating the correlation between the uncertainty quality and the final SL performance. Characteristics of dif-

ferent estimations are also discussed, including the language similarities learned by LOU and the current limitation of EVI in the SL process (§5).

Our contributions in this work can be summarized as follows. (1) We propose the self-learning framework for the cross-lingual transfer and identify the importance of uncertainty estimation under this setting. (2) We adapt three different uncertainty estimations in our framework, and evaluate the framework on both NER and NLI tasks covering 40 languages in total, improving the performance of both high-resource and low-resource languages on both tasks by a solid margin (10 F1 for NER and 2.5 accuracy score for NLI on average). (3) Further analysis is conducted to compare different uncertainties and their characteristics.

2 Related Work

We introduce the work of uncertain estimation briefly. As deep learning models are optimized by minimizing the loss without special care on the uncertainty, they are usually poor at quantifying uncertainty and tend to make over-confident predictions, despite producing high accuracies (Lakshminarayanan et al., 2017). Estimating the uncertainty of deep learning models has been recently studied in NLP tasks (Xiao and Wang, 2019a; Zhang et al., 2019; He et al., 2020). There are two main uncertainty types in Bayesian modelling (Kendall and Gal, 2017; Depeweg et al., 2018): *epistemic* uncertainty that captures the model uncertainty itself, which can be explained away with more data; *aleatoric* uncertainty that captures the intrinsic data uncertainty regardless of models. *Aleatoric* uncertainty can further be devided into two sub-types: *heteroscedastic* uncertainty that depends on input data, and *homoscedastic* uncertainty that remains constant for all data within a task. In this work, we only focus on *aleatoric* uncertainty, as it is more closely related to our SL process to select confident and high-quality predictions within each iteration.

3 Approach

We keep the same model architecture throughout our experiments: a multilingual pre-trained language model is employed to encode each input sequence, followed by a linear layer to classify on the hidden state of CLS token for NLI, and of each token for NER, which is the same model setting from XTREME (Hu et al., 2020). Cross-entropy (CE) loss is used during training in the baseline.
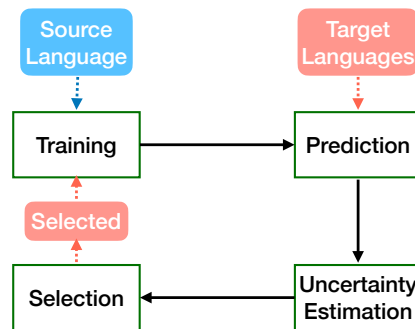

Figure 1: Illustration of the self-learning framework with explicit uncertainty estimation.

3.1 Self-Learning (SL) Framework

We formulate the task-agnostic SL framework for cross lingual transfer into the following four phases, as shown in Figure 1. In the *training* phase, the model parameter θ gets optimized by the training inputs X and labels Y, with Y being gold labels of the source language in the first iteration, along with silver labels of target languages in later iterations. Inputs of different languages are mixed together. In the *prediction* phase, the model predicts on the remaining unlabeled data $X_l^* = \{x_{l1}^*, \ldots, x_{lN}^*\}$ of each target language l, with each prediction denoted as $y^* = f^\theta(x^*)$. In the *uncertainty estimation* phase, the model estimates the prediction uncertainty based on one of the methods described in §3.2, denoted as $\gamma = f_\gamma^\theta(x^*, y^*)$, representing the model confidence of the prediction. In the *selection* phase, data in each X_l^* is ranked based on the uncertainty score γ, and we select top-K percent of each X_l^* with their predictions as silver labels, adding to the training data. To avoid posing potential inductive bias from imbalanced label distribution, we select equal amount of inputs for each label type, similar to previous work on self-learning (Yalniz et al., 2019; Dong and de Melo, 2019; Mukherjee and Awadallah, 2020).

After *selection*, the model goes back to the *training* phase and starts a new iteration with the updated training set. The entire process keeps iterating until there is no remaining unlabeled data; early stop criteria are implemented on the dev set of the source language only, as gold labels are not available for other languages. Each phase can be adjusted by task-specific requirements (see A.2).

3.2 Uncertainty Estimation

We adapt three different uncertainty estimation techniques in our framework. Let C be the label classes, p_c be the probability of class c for an input.

Language Heteroscedastic Uncertainty (LEU)
LEU injects Gaussian noise into class logits whose variance is predicted by the model as an input-dependent uncertainty (Kendall and Gal, 2017), regardless of languages. A Gaussian distribution is placed on the logit space $\mathbf{g} \sim \mathcal{N}(g^\theta, (\sigma^\theta)^2)$, where the model is modified to predict both raw logit g^θ and standard deviation σ^θ given each input. We use the expectation of the logit softmax as the new probability, computed by Monte Carlo sampling: $p_c = E[\text{softmax}(\mathbf{g}_c)] \approx \frac{1}{T} \sum_t \text{softmax}(g_{tc})$, with g_{tc} being the logit of class c at t-th sampling from $\mathbf{g}$. The training loss and the uncertainty take into account the new probability formulation p_c:

$$\mathcal{L}^{\text{LEU}} = -\log \frac{1}{T} \sum_t \exp\left(-\mathcal{L}_t(x, c)\right) \quad (1)$$

The loss is composed of the CE loss $\mathcal{L}_t(x, c)$ on input x and gold class c with tth sampled probabilities. The uncertainty is the entropy of the new probabilities: $\gamma = -\sum_c p_c \log p_c$. When an input of any language is hard to predict, the model will signal high variance, indicating high uncertainty, as the probability distribution tends to be uniform.

Language Homoscedastic Uncertainty (LOU)
LOU estimates the uncertainty of each certain language, regardless of the input. Similar to the formulation of task uncertainty (Cipolla et al., 2018), we propose to place an uncertainty σ_l on a language l as the *homoscedastic* uncertainty. σ is used as the softmax temperature on the predicted logits g^θ: $p_c = \text{softmax}(\frac{1}{\sigma^2} g_c^\theta)$. The final uncertainty is also the entropy of the scaled probabilities. A higher σ_l leads to higher entropy of all inputs of language l, as the probability distribution tends to be more uniform. During training, each σ_l is a learned parameter directly, and the new loss for an input of language l can be approximated as:

$$\mathcal{L}^{\text{LOU}} \approx \frac{1}{\sigma_l^2} \mathcal{L}(x, c) + \log \sigma_l \quad (2)$$

$\mathcal{L}(x, c)$ is the same CE loss as in Eq (1). Note that LOU does not change the input selection nor ranking within each language; we mainly use it as an optimization strategy to jointly train inputs of multiple languages, automatically distinguishing the importance of different target languages.

Evidential Uncertainty (EVI) EVI estimates the evidence-based uncertainty (Sensoy et al., 2018), where the softmax probability is replaced

with Dirichlet distribution, and each predicted logit for class c is regarded as the evidence. We employ the decomposed entropy *vacuity* and *dissonance* proposed by Shi et al. (2020). *vacuity* is high when there lacks evidence for all the classes, indicating out-of-distribution (OOD) samples that are far away from the source language; *dissonance* becomes high when there are conflicts of strong evidence among certain classes (more details are shown in A.1). The prediction is said uncertain if either *vacuity* or *dissonance* is high. For each input, let S be the total evidence strength, and let the label y_c be 1 for the gold class and 0 for the others. The following describes the expected probability p_c for the class c under Dirichlet distribution, as well as the training loss $\mathcal{L}^{\text{EVI}}$:

$$p_c = \frac{e_c + 1}{S} \quad (3)$$

$$\mathcal{L}^{\text{EVI}} = \sum_c (y_c - p_c)^2 + \frac{p_c(1 - p_c)}{S + 1} \quad (4)$$

4 Experiments

The framework with different uncertainties are evaluated on two cross-lingual transfer datasets: XNLI (Conneau et al., 2018) for the NLI task covering 15 languages, and Wikiann (Pan et al., 2017) for the NER task covering 40 languages. For both datasets, English is the source language with gold labels, and we use the dev set of target languages (TLs) as the source of unlabeled data; we do not consult any gold labels of TLs in the SL process. XLM-R$_{\text{Large}}$ (Conneau et al., 2020) is used as the multilingual encoder across our experiments. Our detailed experimental setting can be found in A.3.

We implement three different settings for the baseline.[1] **BL-Direct** is the direct zero-shot transfer without utilizing unlabeled data of TLs. **BL-Single** trains gold data of English and silver data of only one TL per model; it simply selects predictions of all unlabeled data as silver labels, without considering any uncertainties. **BL-Joint** is similar to BL-Single but instead train with all TLs jointly.

For SL, we set top-K percent selection to be top 8% of total unlabeled data for each label type, so the entire SL process will finish in around 6 iterations. We found that K below 10% can generally yield decent performance.

For the analysis, we also include two common uncertainties used in previous work of self-learning

---

[1]Code is available at https://github.com/lxucs/multilingual-sl.

	en	af	ar	bg	bn	de	el	es	et	eu	fa	fi	fr	he	hi	hu	id	it	ja	jv	
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	**85.2**	83.7	**75.1**	85.8	82.0	83.6	84.4	**86.5**	84.6	72.1	72.9	84.7	**84.1**	61.4	**80.2**	85.7	54.8	83.9	41.3	69.2	
SL-LOU	84.4	**85.3**	61.1	87.1	81.9	83.4	85.4	75.6	85.5	74.6	74.9	84.4	83.3	**68.5**	78.6	84.5	**55.5**	**85.1**	46.2	**70.0**	
SL-LEU	84.7	81.5	70.0	**87.6**	**83.6**	**84.6**	**85.5**	85.0	**85.6**	**77.8**	**81.0**	**86.2**	83.1	62.0	79.5	**87.0**	53.4	84.8	**49.5**	65.3	
	ka	kk	ko	ml	mr	ms	my	nl	pt	ru	sw	ta	te	th	tl	tr	ur	vi	yo	zh	avg
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	81.0	56.4	69.4	**76.3**	77.9	72.5	71.7	**87.1**	85.5	**80.6**	71.2	69.4	61.5	**6.7**	80.7	85.3	79.8	**86.2**	42.7	48.9	73.3
SL-LOU	78.8	58.7	70.2	75.4	**79.4**	73.8	71.2	86.4	86.2	79.2	**73.3**	**69.5**	68.8	4.7	**83.4**	88.4	**85.9**	85.8	49.1	50.5	73.8
SL-LEU	**81.1**	**63.7**	**71.8**	76.0	76.2	**75.9**	71.5	**87.1**	**87.6**	79.9	70.4	64.0	**69.9**	2.2	81.3	**89.1**	**85.9**	85.9	43.5	**54.8**	**74.4**

Table 1: NER Results in F1 scores for 40 languages. BL-Direct is equivalent to Hu et al. (2020).

	en	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh	avg
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	79.5	84.4	83.4	82.4	**84.8**	83.7	78.0	81.6	71.1	78.2	79.2	74.4	80.8	80.4	80.7
SL-LOU	88.2	**81.0**	84.4	**83.5**	82.3	**84.8**	**83.9**	78.9	**81.8**	**73.9**	79.3	80.1	**75.7**	81.6	**81.4**	**81.4**
SL-LEU	88.1	80.7	**84.9**	83.4	**82.8**	84.5	83.8	**79.2**	**81.8**	73.0	**79.7**	**80.5**	**75.7**	**81.9**	81.3	**81.4**

Table 2: XNLI Results in accuracy scores for 15 languages.

on other tasks: max probability (MPR), and entropy (ENT); both use plain softmax probabilities (A.4).

4.1 Results

The results for NER and NLI are shown in Table 1 and 2 respectively. BL-Direct is equivalent to our re-implementation of Hu et al. (2020).

BL-Single outperforms BL-Direct on NER by 3.1 F1 on average, demonstrating the effectiveness of utilizing unlabeled data even without considering uncertainties. Remarkablely, languages such as Arabic (ar), Japanese (ja), Urdu (ur) and Chinese (zh) receive 10+ gain in F1. By contrast, BL-Single does not surpass the baseline for NLI, partially because all TLs already have much closer performance to English, which in turn highlights the importance of estimating uncertainties for SL.

BL-Joint outperforms BL-Single on both tasks by a slight margin, and we do see performance gain over BL-Single on 32/40 and 10/15 languages for NER and NLI respectively. Certain languages such as Hindi (hi), Javanese (jv) and Yoruba (yo) receive non-trivial benefits (2.6 - 8.6 F1 gain for NER) through the joint language training, validating our joint training strategy for SL.

Evaluation of SL is shown with the best results of each uncertainty from 3 repeated runs. The best performance of SL for both tasks is achieved by adopting LEU as the uncertainty estimation, which

outperforms three baselines significantly (10% gain for NER and 2.5% for NLI on average), and surpasses other uncertainties by a slight margin. In NER specifically, certain low-resource languages such as Basque (eu), Persian (fa), Burmese (my) and Urdu (ur) have substantial performance improvement over BL-Joint (13.8 - 25.9 F1 gain); the performance of certain high-resource languages such as Arabic (ar), German (de) and Chinese (zh) can also increase by a solid margin over BL-Joint (4.1 - 13.3 F1 gain). The trend of improving both high and low-resource languages is also present in NLI. All results are stable across multiple runs with standard deviation within 0.1 - 0.2 on average.

Results also suggest that other uncertainty estimations can achieve comparable performance, as LEU does not dominate every language. We further conduct analysis on uncertainties as follows.

5 Analysis

Uncertainty Comparison To directly assess different uncertainty estimations, we evaluate uncertainty scores by AUROC against predictions, such that AUROC is high when the model is confident on correct predictions and uncertain on incorrect predictions. The left side of Table 3 shows the AUROC of four estimations on the test sets of both tasks. MPR and ENT are also included in the experiments for comparison; LOU is excluded as it does not change selection. The right side of Table 3

	M	T	I	E	M	T	I	O
NER	71.2	72.1	68.7	**73.7**	0.6	0.5	**1.1**	0.5
XNLI	76.9	77.3	73.0	**78.6**	0.3	0.3	**0.7**	0.0

Table 3: The left side shows the averaged AUROC of different uncertainty estimations. The right side shows the averaged SL performance drop compared to LEU. M=MPR, T=ENT, I=EVI, O=LOU, E=LEU.

shows the SL performance drop using other uncertainties compared to LEU, serving as an indirect evaluation of different uncertainties. As shown, LEU indeed achieves the best AUROC, being a better uncertainty estimation compared to others; EVI has the lowest AUROC and also the lowest SL performance; MPR and ENT can bring moderate scores on both AUROC and SL. Thus, Table 3 corroborates strong correlation between AUROC and SL performance: better uncertainty can indeed lead to higher performance in the SL process.

Language Uncertainty   Table 2 shows that LOU reaches the same accuracy as LEU on XNLI, with trivial performance gap for each language. We find that the learned uncertainty of each language is highly consistent through multiple runs, as shown in Table 4, which can be loosely interpreted as language similarities under the input of this task, e.g. Vietnamese (vi) appears to be more distant from English than others for this task, and the joint optimization of all languages could benefit from this learned language uncertainty. However, we do not find LOU to be as stable on NER, potentially because NER has much more noise and languages.

en	ar	bg	de	el	es	fr	hi
1.44	1.20	1.15	0.63	0.58	1.78	0.70	1.60

ru	sw	th	tr	ur	vi	zh
0.33	1.07	4.18	1.89	3.15	0.23	0.99

Table 4: The learned language uncertainty σ^2 of LOU for each language in XNLI.

Evidential Uncertainty   Although EVI is able to achieve good performance on certain languages, there also exists large gap for certain other languages compared to LEU. We attribute the inferior performance of EVI to two aspects. First, the predicted evidence (logit) still exhibits over-confidence, which destabilizes the *vacuity* and *dissonance*. Figure 2 shows an example of the evidence-based entropy distribution for EVI, and the model indicates most all predictions as certain

(small entropy). Second, *vacuity* can only distinguish true OOD samples for English, as only English has gold labels. It could fail to recognize those confident samples of TLs that appear in-distribution but are inherently wrong, and falsely select them in the SL process. Figure 3 shows the t-SNE visualization of hidden states of inputs in English and Japanese on the test set of NER: some target language inputs that are close to English in terms of hidden states are predicted wrong, because of the zero-shot nature.
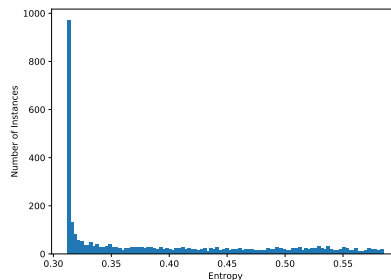

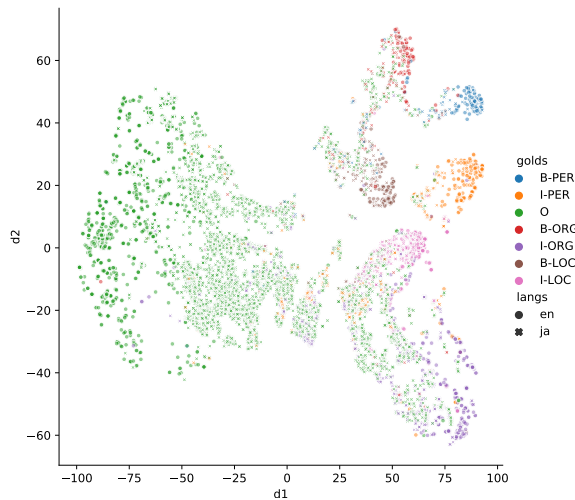Figure 2: Evidence-based entropy distribution on the test set of NER for Japanese (ja).

Figure 3: t-SNE visualization of CLS hidden states of NER inputs in English (en) and Japanese (ja). Different gold label types for tokens are marked by colors, and two languages are marked by the shapes. Each cluster should ideally has only one distinct color.

6   Conclusion

In this work, we propose a self-learning framework combined with explicit uncertainty estimation for cross-lingual transfer. Three different uncertainties are adapted, and the entire framework is evaluated on two tasks of NER and NLI, surpassing the baseline by a large margin. Further analysis shows the evaluation and characteristics of each uncertainty.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Roberto Cipolla, Yarin Gal, and Alex Kendall. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7482–7491.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard 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, Online. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

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.

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.

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

Mladen Karan, Ivan Vulić, Anna Korhonen, and Goran Glavaš. 2020. Classification-based self-learning for weakly supervised bilingual lexicon induction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6915–6922, Online. Association for Computational Linguistics.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6405–6416, Red Hook, NY, USA. Curran Associates Inc.

Subhabrata Mukherjee and Ahmed Awadallah. 2020. Uncertainty-aware self-training for few-shot text classification. In *Advances in Neural Information Processing Systems*, volume 33, pages 21199–21212. Curran Associates, Inc.

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.

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

Weishi Shi, Xujiang Zhao, Feng Chen, and Qi Yu. 2020. Multifaceted uncertainty estimation for label-efficient deep learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 17247–17257. Curran Associates, Inc.

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

Yijun Xiao and William Yang Wang. 2019b. Quantifying Uncertainties in Natural Language Processing Tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7322–7329.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

I. Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. 2019. Billion-scale semi-supervised learning for image classification.

Xuchao Zhang, Fanglan Chen, Chang-Tien Lu, and Naren Ramakrishnan. 2019. Mitigating uncertainty in document classification. *arXiv preprint arXiv:1907.07590*.

A Appendix

A.1 Uncertainty Estimation

For **LOU**, the uncertainty term as the denominator in the loss as in Eq (2) achieves the effect of "learned loss attenuation" (Kendall and Gal, 2017) during training, where uncertain samples have lower scale of loss, so that the optimization is less prone to noisy data. We use LOU to let the model learn the uncertainty for each language to achieve more stable training amid selected data with silver labels.

In practice, the model directly predicts the log-variance term $\log \sigma$ for both LEU and LOU, as the training is more stable and the variance is guaranteed to be positive.

For **EVI**, we follow Sensoy et al. (2018) and define:

$$b_c = \frac{e_c}{S} \quad u = \frac{|C|}{S} \quad S = \sum_c e_c + |C| \quad (5)$$

e_c is the evidence strength (logit) for class c, $|C|$ is the number of classes. b_c represents the belief mass for class c and u is the *vacuity*, denoted as $vac = u$. We follow Shi et al. (2020) and define *dissonance* for each input as:

$$\mathrm{Bal}(b_j, b_k) = \begin{cases} 1 - \frac{|b_j - b_k|}{b_j + b_k} & \text{if } b_i b_j \neq 0 \\ 0 & \text{elsewise} \end{cases} \quad (6)$$

$$diss = \sum_c \frac{b_c \sum_{c' \neq c} b_{c'} \mathrm{Bal}(b_c, b_{c'})}{\sum_{c' \neq c} b_{c'}} \quad (7)$$

Both *vac* and *diss* are in the range of $[0, 1]$; being closer to 1 indicates more uncertainty. The final uncertainty is set as $\gamma = diss + \alpha \cdot vac$ with α being a hyperparameter.

In practice, ELU activation is added after raw logits to ensure the evidence strength is positive.

A.2 Task-Specific Adjustment

We adjust the SL process for NER as follows: the uncertainty score is obtained for each predicted entity, which is calculated as the averaged uncertainty score of all tokens within the entity. Ranking is performed on entities within each entity type; we select the input sequence if all its predicted entities have uncertainty within the top-K threshold.

A.3 Experimental Setting

We follow the same train/dev/test split and same evaluation protocol as XTREME (Hu et al., 2020).

Datasets For XNLI (Conneau et al., 2018), there are three label types for each sequence: "neutral", "entailment", "contradiction". For Wikiann (Pan et al., 2017), there are three entity types: "LOC", "PER", "ORG"; each token is tagged in the BIO2 format, thus there are 7 label types for each token.

Hyperparameters For both NLI and NER, we use the following hyperparameter setting as suggested by XTREME (Hu et al., 2020): 32 effective batch size, 2×10^{-5} learning rate with linear decay scheduling, 1 max gradient norm.

For NLI in the self-learning (SL) process, we train the model by 5 epochs in the first iteration on English training set with gold labels, whereas we train 10 epochs for NER. After the first iteration, the model is trained for 3 epochs in each later iteration. For LEU, we set the Monte Carlo sampling $T = 20$. For EVI, we set $\alpha = 1$ for NLI and $\alpha = 10^{-2}$ for NER based on the empirical scale of *vac* and *diss*, keeping both on the same scale.

To avoid the training set growing too huge as the SL process iterates, we apply a sampling strategy upon new selection: each training epoch samples from the existing training set with equal amount of newly selected data, so that each training epoch consists of at least 50% latest selection. We adopt early stop on English dev set if the evaluation does not improve for over two iterations.

Our experiments uses NVIDIA Titan RTX GPUs. The training takes 10 hours for both NER and NLI.

A.4 Other Uncertainties

MPR is the max probability of label classes, denoted by $\gamma = \max_c \mathbf{p}_c$. It is equivalent to the probability of the predicted label, and is commonly used as the selecting criterion for classification tasks (Yalniz et al., 2019; Dong and de Melo, 2019).

ENT is the entropy of the class probability distribution, denoted by $\gamma = -\sum_c \mathbf{p}_c \cdot \log \mathbf{p}_c$, which is another common uncertainty metric for classification (Depeweg et al., 2018; Xiao and Wang, 2019b).