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

Liyan Xu[1], Xuchao Zhang[2], Xujiang Zhao[3], Haifeng Chen[2], Feng Chen[3], Jinho D. Choi[1]
Emory University[1], NEC Labs America[2], The University of Texas at Dallas[3]

EMNLP 2021

Background

- **Cross-lingual transfer** (**CLT**): model for one language $\Rightarrow$ model for other language(s)
- Zero-shot: training on **source** language + inference on **target** languages

Background

- Embedding alignment:
  - Explicit word-embedding alignment: translation matrix
    - Supervised (Mikolov et al., 2013, etc.)
    - Unsupervised (Conneau et al., 2018, etc.)
  - Shared/joint embedding space: **multilingual pre-trained language models**
    - mBERT (Devlin et al., 2019)
    - XLM-R (Conneau et al., 2020)
    - mT5 (Xue et al., 2021)

Motivation

- Practical scenarios:
  - zero-shot?
  - Annotation for target languages?
  - Middle ground: **unlabeled data of target languages**

Motivation

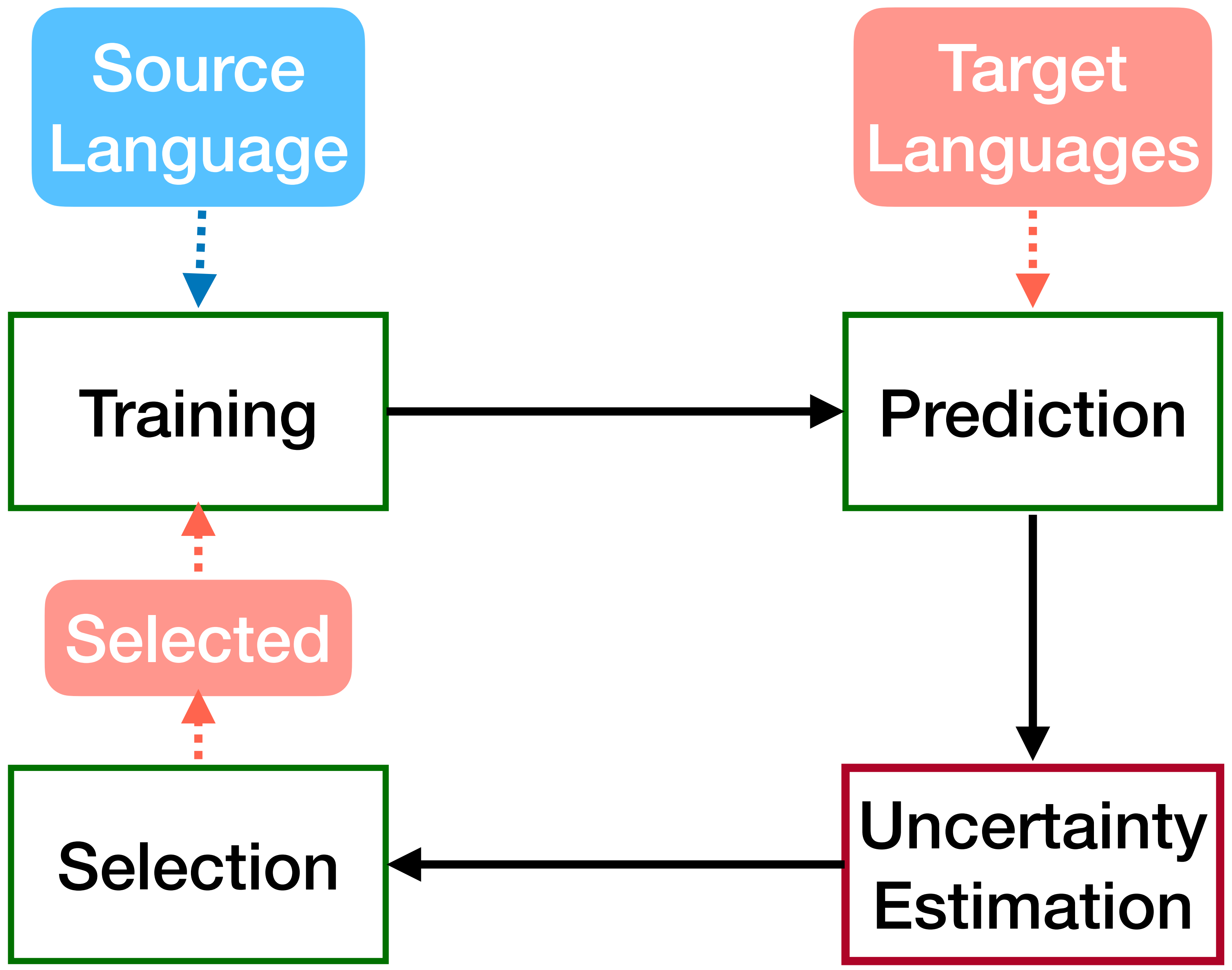
- Previous work: self-learning for multilingual document classification (Dong and Melo, 2019)
  - Predictions on unlabeled data of target languages
  - a.k.a “**pseudo labels**”

Approach

- **Self-learning** framework for cross-lingual transfer
  - w/ multilingual pre-trained LMs
  - Making use of zero-shot capability
- Explicit **uncertainty estimation**
  - uncertainty estimation $\Rightarrow$ pseudo label quality $\Rightarrow$ CLT performance

Approach

- Iterative training and prediction:
  - 1st iteration:
    - Train on gold labels of source language
  - 1+ iteration:
    - Select top-k confident predictions of target languages into training set
    - Need accurate uncertainty estimation
    - New training set: more data for task-specific learning and joint embedding alignment
    - Termination: no more unlabeled data or early stop on dev set



Uncertainties

- Deep learning models are notorious for over-confident predictions
  - High-dimensional space ⇒ sparse data points ⇒ imperfect decision boundary
- Two main types of uncertainties (Kendall and Gal, 2017; Depeweg et al., 2018)
  - ***Aleatoric*** uncertainty: intrinsic **data uncertainty** regardless of models
  - ***Epistemic*** uncertainty: **model uncertainty** that can be explained away with more data
- This work: focus on *aleatoric* uncertainty

Uncertainty Estimation

- Adapt three uncertainty estimation techniques:
  - Language Heteroscedastic Uncertainty (**LEU**)
  - Language Homoscedastic Uncertainty (**LOU**)
  - Evidential Uncertainty (**EVI**)

Uncertainty Estimation

- Language Heteroscedastic Uncertainty (**LEU**)
  - Heteroscedastic: **input-dependent**
  - Place Gaussian noise on class logits (Kendall and Gal, 2017)
  - Predict both class logits and variance
  - Loss: $L^{\mathrm{LEU}} = -\log \frac{1}{T} \sum_t \exp\big(-L_t(x, c)\big)$

Uncertainty Estimation

- Language Homoscedastic Uncertainty (**LOU**)
  - Homoscedastic: input-independent, **task-dependent** (Kendall et al., 2018)
    - Uncertainty regardless of input
  - Adaptation: language-dependent
    - Does not change selection but helps with optimization on joint language training
    - Place softmax temperature per language as learned parameters
  - Language Loss: $L^{\text{LOU}} \approx \frac{1}{\sigma_l^2} L(x, c) + \log \sigma_l$

Uncertainty Estimation

- Evidential Uncertainty (**EVI**):
  - Replace softmax probability with Dirichlet distribution (Sensor et al., 2018)
  - Regard class logit as Dirichlet evidence strength
  - Loss: $L^{\mathrm{EVI}} = \sum_c (y_c - p_c)^2 + \frac{p_c(1 - p_c)}{S + 1}$
  - Uncertainty decomposition (Shi et al., 2020):
    - Vacuity: lacking evidence for all classes (OOD)
    - Dissonance: strong conflicting evidence (ambiguous in-domain)

Experiments

- Datasets:
  - **XNLI**: NLI task covering **15** languages
  - **Wikiann**: NER task covering **40** languages
- Model: XLM-R
- Baselines:
  - BL-Direct: zero-shot **(en)**
  - BL-Single: use all predictions on unlabeled data of one target language **(en** + **one target language)**
  - BL-Joint: mix target languages together **(en** + **all target languages)**

Results

NER

- Unlabeled data helps even without uncertainty estimation (BL-Single).
- Joint training on all target languages helps low-resource languages (BL-Joint).
- Uncertainty estimation outperforms (best results by **LEU**).

	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**

Results

NER

- Large gap (10+ F1) on distant languages, e.g. Arabic (ar), Japanese (ja), Chinese (zh)
- Good improvement on closer languages as well, e.g. Spanish (es), German (de)
- Significant boost on low-resource languages, e.g. Basque (eu), Persian (fa)

	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**

Results

XNLI

- Unlabeled data does not help without uncertainty estimation (BL-Single).
- Uncertainty estimation outperforms (best results by **LEU/LOU**).

	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**

Analysis

- Impact of uncertainties: estimation quality ⇒ final performance
  - Correlation shown by comparing 5 uncertainties
- Language uncertainty ⇒ language similarity

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.

References

- Mikolov, T., Le, Q. V., & Sutskever, I. (*arXiv 2013*). Exploiting Similarities among Languages for Machine Translation.
- Conneau, A., Lample, G., Ranzato, M., Denoyer, L., & Jégou, H. (*ICLR 2018*). Word Translation Without Parallel Data.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (*NAACL 2019*). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (*ACL 2020*). Unsupervised Cross-lingual Representation Learning at Scale.
- Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (*NAACL 2021*). mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer.
- Dong, X., & de Melo, G. (*EMNLP 2019*). A Robust Self-Learning Framework for Cross-Lingual Text Classification.
- Kendall, A., & Gal, Y. (*NIPS 2017*). What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?
- Depeweg, S., Hernández-Lobato, J. M., Doshi-Velez, F., & Udluft, S. (*ICML 2018*). Decomposition of Uncertainty in Bayesian Deep Learning for Efficient and Risk-sensitive Learning.

References

- Kendall, A., Gal, Y., & Cipolla, R. (*IEEE/CVF 2018*). Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics.
- Sensoy, M., Kaplan, L., & Kandemir, M. (*NIPS 2018*). Evidential Deep Learning to Quantify Classification Uncertainty.
- Shi, W., Zhao, X., Chen, F., & Yu, Q. (*NIPS 2020*). Multifaceted Uncertainty Estimation for Label-Efficient Deep Learning.