Improving Cross-lingual Transfer with Contrastive Negative Learning and Self-training

Guanlin Li[†], **Xuechen Zhao**[‡], **Amir Jafari** [†],
Wenhao Shao[†], **Reza Farahbakhsh**[†], **Noel Crespi**[†]
[†] Samovar, Telecom SudParis, Institut Polytechnique de Paris, 91120 Palaiseau, France
[‡]School of Computer, National University of Defense Technology, Changsha, China
{guanlin_li, amir-reza.jafari_tehrani, wenhao.shao,
reza.farahbakhsh, noel.crespi}@telecom-sudparis.eu
zhaoxuechen@nudt.edu.cn

Abstract

Recent studies improve the cross-lingual transfer learning by better aligning the internal representations within the multilingual model or exploring the information of the target language using self-training. However, the alignment-based methods exhibit intrinsic limitations such as non-transferable linguistic elements, while most of the self-training based methods ignore the useful information hidden in the low-confidence samples. To address this issue, we propose CoNLST (**Co**ntrastive **N**egative **L**earning and **S**elf-**T**raining) to leverage the information of low-confidence samples. Specifically, we extend the negative learning to the metric space by selecting negative pairs based on the complementary labels and then employ self-training to iteratively train the model to converge on the obtained clean pseudo-labels. We evaluate our approach on the widely-adopted cross-lingual benchmark XNLI. The experiment results show that our method improves upon the baseline models and can serve as a beneficial complement to the alignment-based methods.

Keywords: contrastive learning, negative learning, self training, cross-lingual transfer learning

1. Introduction

Multilingual language models are commonly used to produce universal representations across languages in zero-shot cross-lingual transfer learning, in which the model is trained on a high resource language and directly evaluated on the target languages. In such a scenario, the zero-shot cross-lingual transfer problem can be viewed as an unsupervised domain adaptation (UDA) problem, wherein the target languages essentially represent a distinct domain (Lee et al., 2022). Similar to the mainstream approaches for UDA which learn domain-invariant representations to narrow the domain shift, most of the recent methods (Chen et al., 2018; Pan et al., 2021; Huang et al., 2021; Ding et al., 2022; Wang et al., 2022) improve cross-lingual transfer by aligning representations across different languages so that semantically similar words in different languages share closer representations in the geometric space.

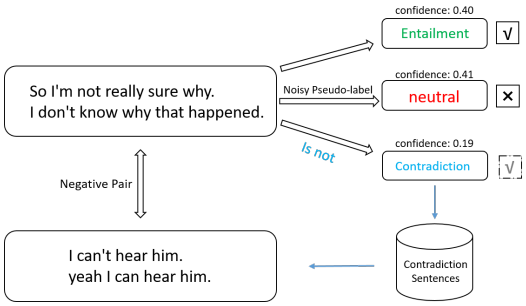


Figure 1: A conceptual illustration of contrastive negative learning. For a low-confidence target sample (top left), the model is uncertain between the wrong label (marked in red) and the ground-truth label (marked in green), and produce a noisy pseudo-label. In contrast, the complementary label (marked in blue) is often correct about "the sample does not belong to the class" and thus provides class-wise contrastive information (bottom left).

While alignment-based methods can be efficient, they suffer from intrinsic limitations when it comes to label shifts and domain shifts (Liu et al., 2021). Another line of work seeks to mine useful information from the target data directly using self-training (Gui et al., 2014; Dong and de Melo, 2019; Xu et al., 2021). Self-training is a bootstrapping method which iteratively trains the target model on the pseudo-labels produced by the model trained on source language data. During the iterations, pseudo-labels are usually selected using confidence-based criteria. However, manually-crafted criteria for self-training require domain knowledge, and simply using confidence-based criteria may not yield satisfactory generalization. This limitation becomes particularly pronounced in low-resource scenarios, where even a high confidence threshold may leave a substantial number of noisy samples, as we observe that in the zero-shot cross-lingual setting, the amount of noisy labels (wrongly pseudo-labeled target samples) is nearly the same in the high-confidence area as in

LREC-COLING 2024, pages 8781–8791
20-25 May, 2024.

the low-confidence area when trained with English as the source language (shown in Figure 3 in the Section 4.3), which would exacerbate the confirmation bias (Arazo et al., 2020) (accumulated errors caused by noisy pseudo-labels) if the model is simply self-trained with high-confidence target samples. Moreover, discarding samples under a threshold may ignore useful hidden information contained within the noisy data.

Therefore, we seek to reduce the confirmation bias and exploit information from low-confidence samples by introducing negative learning, and propose a two-stage cross-lingual self-training method, **Co**ntranstive **N**egative **L**earning and **S**elf-**T**raining, denoted as CoNLST. Specifically, at the first stage, the prediction confidences of a zero-shot cross-lingual model finetuned only on the source language are calibrated to distill samples with clean pseudo-labels from noisy pseudo-labels in the target language. The calibration is conducted via negative learning (Kim et al., 2021), as it is a simple-yet-effective method to exploit the information hidden in the noisy labels by taking use of the complementary labels of the samples. We observed in our case that vanilla negative learning would fail and cause degeneration of the model's performance, and we propose to extend negative learning to the metric space using contrastive learning to effectively solve the issue. At the second stage, with the distilled clean high-confidence samples obtained from the first stage, we consider the problem as a semi-supervised learning (SSL) problem and adopt dynamic thresholding (Zhang et al., 2021) with multiple data augmentation strategies to iteratively finetune the model, which further improves the model performance by a large margin. The concept of contrastive negative learning is illustrated in the Figure 1, and the overall framework of the proposed method is shown in the Figure 2.

We empirically evaluate the proposed method on the widely adopted cross-lingual benchmark, XNLI dataset (Conneau et al., 2018). We show that our proposed method brings consistent improvements for both mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) and achieves comparable performance with state-of-the-art methods. Furthermore, the experiment results show that the proposed method can serve as a beneficial complement to the alignment based methods.

2. Related Works

In cross-lingual tasks, the challenges extend beyond linguistic differences between the source and target data to include domain shifts, which poses a cross-lingual cross-domain problem (Li et al., 2021). Mainstream approaches employ model-centric methods which focus on improving the model's multilinguality by aligning internal subspaces of different languages within the model. This can be achieved by supervised alignment using parallel corpora and multilingual dictionaries (Kulshreshtha et al., 2020; Pan et al., 2021), weakly supervised training by fine-tuning the model with multilingual data (Nooralahzadeh et al., 2020; Vu et al., 2021) or unsupervised alignment by extracting language and domain invariant features, such as cross-lingual adversarial training (Chen et al., 2018; Huang et al., 2021), self-contrastive learning (Wang et al., 2022), feature decomposition (Li et al., 2021) and so on.

While model centric methods have demonstrated empirical effectiveness, learning invariant representations has intrinsic limitations when it comes to domain shifts (Liu et al., 2021), whereas data centric methods serve as a valuable complement. In the context of cross-lingual learning, pseudo-labeling with self-training has been widely adopted to leverage target language information. Gui et al. (2014) employed least error boundary estimation to measure the quality of transferred examples in the target language. The estimation is derived from the model's prediction probability for each target sample, whereby samples likely to introduce more errors are simply discarded in subsequent iterations. Dong and de Melo (2019) selected a subset of target samples based on the model's prediction confidence, and discarded other samples during each iteration to construct a balanced set containing the same number of instances in each class and merged the selected set into the source language training data. Xu et al. (2021) proposed three different uncertainty estimation measures which are calculated based on the model's output probabilities jointly trained multiple languages together to improve the performance of the model. The aforementioned studies demonstrate the efficacy of incorporating high-quality pseudo-labeled data in the target language. In our work, unlike previous works which discarded low-confidence samples, we seek to take advantage of the noisy data as well for self-training.

3. Methodology

In this work, we aim to exploit the information of unlabeled target language samples using self-training. To start the bootstrapping process, we initially finetune a multilingual model on the source language data to provide pseudo-labels and confidence scores for the unlabeled target samples.

In standard self-training, the pseudo-labels of the target samples are directly adopted as silver labels to further finetune the multilingual models, and the training set iteratively grows by selecting high-confidence pseudo-labeled target samples. Simply

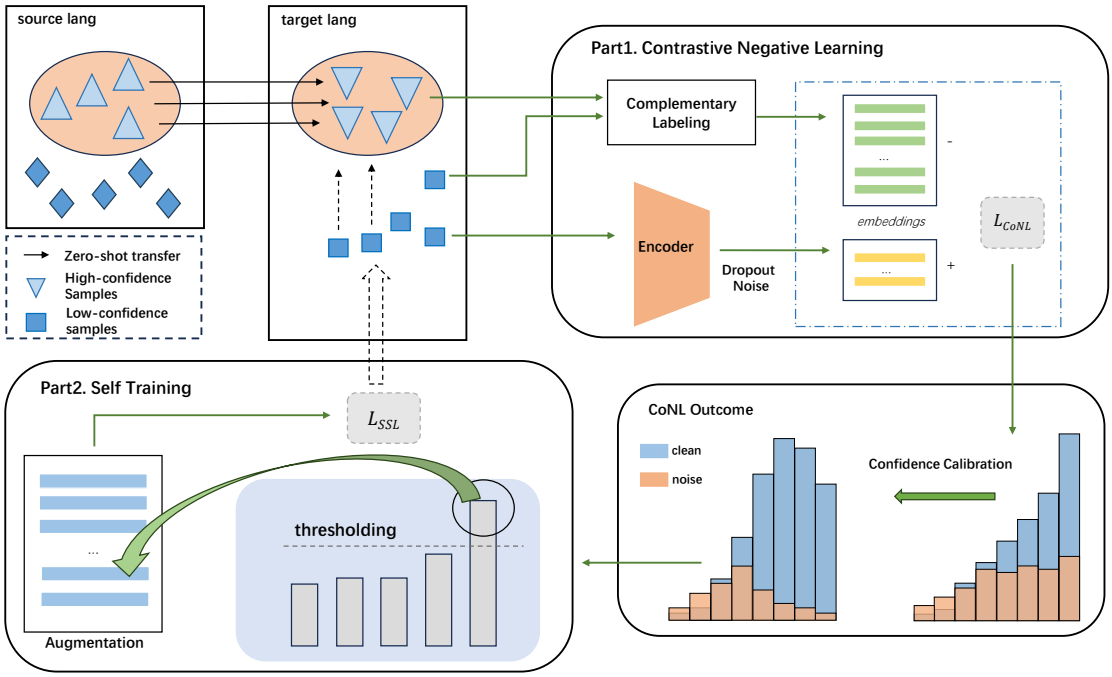


Figure 2: An overview of the proposed CoNLST, the flow of the algorithm is shown with the green arrow. The framework contains two parts, which corresponds to the two stages in the method: The first part is the contrastive negative learning stage, where target data are sampled to construct positive pairs with dropout noise and negative pairs with complementary labeling. After the first stage, the model exhibits less confidence for the noisy data and reveals clean samples in high-confidence areas. The second part is self-training-based semi-supervised learning (SSL), which selects clean samples to generate strong augmentations to further refine the model.

taking samples above a certain confidence threshold for self-training is infeasible as they might still contain a large number of noisy labels especially for low-resource languages. Motivated by this, we propose to first calibrate the model's confidence to mine clean labels and introduce a two-stage self-training method. The overall framework of the method is shown in Figure 2. In the following sections, we first present preliminary works of negative learning and contrastive learning, then we elaborate on the two stages of the method, namely contrastive negative learning (CoNL) and self-training with pseudo-labels.

3.1. Preliminary Works

Negative Learning for noisy labels It was shown that learning with noisy labels is nontrivial with language models such as BERT (Zhu et al., 2022) since it could be vulnerable to noise from weak supervision even at a low level. Negative learning (NL) is an intuitively simple method for noisy data classification: instead of maximizing the log-likelihood of the target label during training, NL chooses complementary labels randomly for samples with noisy labels and minimizes the log-likelihood of the complementary label, such that "the sample does not belong to the complementary label". Kim et al. (2021) showed that NL can effectively filter clean samples from a high rate of injected noise in image classification problems. The negative learning loss is shown as follows:

$$\mathcal{L}_{NL}(f, \bar{y}) = -\sum_{k=1}^{c} \bar{y}_k log(1 - p_k) \quad (1)$$

where p_k is the model's prediction probability for a sample k as in the standard cross entropy loss, $\bar{y}$ is a complementary label randomly selected from the labels of all classes except for the given label.

The vanilla negative learning fails in our cross-lingual task, for which we assume three possible reasons: 1. limited number of task labels; 2. memorization effect of language models; 3. strong discrimination ability of cross-entropy loss on hard labels. Thus, we propose to improve the complementary label selection strategy and extend negative learning to the metric space using contrastive learning.

Contrastive Learning Contrastive learning has been widely adopted to learn effective visual or language representations. The common idea of contrastive learning is to randomly sample a data point as anchor instance, pull "positive" instances closer to the anchor instance in the metric space and push "negative" instances away. For supervised learn-

ing, positive and negative pairs are constructed at the class level such that positive instances share the same label and negative pairs have different labels. The supervised contrastive loss (Khosla et al., 2020) is defined as:

$$\sum_{i\in I}\frac{-1}{|P(i)|}\sum_{p\in P(i)}log\frac{\langle h_i,h_p\rangle/\tau}{\sum_{a\in A(i)}\langle h_i,h_a\rangle/\tau} \quad (2)$$

where index i is the anchor, $A(i)\equiv I\backslash\{i\}$ is the mini-batch except for i and $P(i)\equiv\{p\in A(i):\tilde{y}_p=\tilde{y}_i\}$ is the set of positive pairs, h denotes the representation of the sample. τ is a temperature hyper-parameter and $\langle\cdot,\cdot\rangle$ is the similarity function. For unsupervised learning, the pairs are constructed at the instance level with negative pairs being simply different instances from the anchor sample. In such a context, the critical question is how to construct positive pairs. A common practice is to use weak augmentations of the anchor instance. For sentence representations, Gao et al. (2021) showed that using a simple dropout noise for augmentation has a good performance. The unsupervised contrastive loss with dropout noise is defined as:

$$\mathcal{L}_{simcse}=-log\frac{\langle h_i^{z_i},h_i^{z_i'}\rangle/\tau}{\sum_{j=1}^{N}\langle h_i^{z_i},h_j^{z_j'}\rangle/\tau} \quad (3)$$

where positive pairs are encodings of the same sentence i with different dropout masks z and negative pairs are from the other sentences j in the same batch.

3.2. Contrastive Negative Learning

In the task of cross-lingual transfer learning, given pseudo-labeled samples in the target language, we propose to adopt the thinking of negative learning for noisy labels and construct negative pairs using complementary labels to provide class-wise discrimination. During this process, we make use of all the target data including the low-confidence ones to form anchor samples. We construct positive pairs at the instance level using weak augmentations of the sentences since class-wise information is unavailable due to the noise of pseudo-labels.

Formally, given the labeled dataset $X_{\ell_s}=\{(x_i^{\ell_s},y_i^{\ell_s})\}_{i=1}^{N}$ with C classes in the source language ℓ_s, given a pretrained multilingual model M, the parameters of the model are denoted as θ_m, we train the model with a linear classification head θ_{cl} as $f_\theta=f(\cdot;\theta_m,\theta_{cl})$ first using the ground truth labels. After training, pseudo-labels for unlabeled data $X_{\ell_t}=\{(x_i^{\ell_t})\}_{i=1}^{M}$ in the target language ℓ_t can be generated as $\hat{y}_i^{\ell_t}=f_\theta(x_i^{\ell_t})$. The confidence score is given by the model's predicted class distribution $\hat{q}_i=p(\hat{y}_i|x_i;\theta_m,\theta_{cl})$, and high-confidence samples have $max(q)\geq\gamma$ with γ being a predefined threshold. For each sample i in the target language, using i as an anchor sample x_i, we generate its complementary label as:

$$\bar{\hat{y}}_i=\begin{cases}\{c\in C|c\neq argmax(q_i)\} & \text{if } max(q_i)\geq\gamma\\ argmin(q_i) & \text{if } max(q_i)<\gamma\end{cases} \quad (4)$$

Given the complementary label $\bar{\hat{y}}_i$ of the anchor sample i, we construct a negative set $\mathcal{N}_i$ by randomly choosing K high-confidence samples from the target language:

$$\mathcal{N}_i=\{x_j\}_{j=1}^{K},\quad \hat{y}_j=\bar{\hat{y}}_i \quad (5)$$

With the negative set and the anchor sample, we extend the negative learning to the metric space. We add a linear layer with parameters θ_{con} on top of the multilingual model's pooling layer as the contrastive head. The representation of a sample x_i is given as $h_i=f(x_i;\theta_m,\theta_{con})$.

Combining loss (2), (3) and the negative set construction (5), we propose the contrastive negative learning loss as:

$$\mathcal{L}_{CL}=-log\frac{\langle h_i^{z_i},h_i^{z_i'}\rangle/\tau}{\langle h_i^{z_i},h_i^{z_i'}\rangle+\sum_{j\in\mathcal{N}_i}\langle h_i^{z_i},h_j^{z_j}\rangle/\tau} \quad (6)$$

During training, we observed that if the classification head was not updated during the contrastive learning, the classification performance of the model would have a drastic drop. However, we would like to avoid directly updating the classification head using the hard pseudo-labels of the target data. Empirically, we found jointly training the model with the NL loss in Eq.(1) using source data to be helpful. To ensure the consistency between the metric space and the label space, we introduce a cross-lingual cross-space consistency loss:

$$\mathcal{L}_{CLCS}=-\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_{NL}(f(x_i^{\ell_s};\theta_m,\theta_{cl}),y_i^{\ell_s}) \quad (7)$$

where we train the model with N randomly chosen samples from the source language with the negative learning loss. The overall training loss at the first stage is:

$$\mathcal{L}_{CoNL}=\mathcal{L}_{CL}+\lambda\mathcal{L}_{CLCS} \quad (8)$$

where λ is used to adjust the contribution of the hard NL loss from the source language. Using Eq.(8), we conduct contrastive negative learning to exploit the hidden information of the noisy labels and calibrate the model's confidence.

3.3. Self-training with Pseudo-labels

By calibrating the model's confidence at the first stage, we substantially reduce the noise rate and

class imbalance for high-confidence samples. We further improve the model's performance iteratively with these distilled clean samples. The problem can be viewed as a semi-supervised learning task. To successfully conduct SSL, three critical factors were identified (Berthelot et al., 2019): (1) Consistency regularization. The model should be invariant to the perturbation of the inputs. (2) Generic regularization. The model should not overfit the training data. (3) Entropy minimization. The classifier should output low-entropy predictions on unlabeled data.

We design our SSL algorithm to comply with the above regularizations. To apply the consistency regularization, we follow consistency training (Xie et al., 2020) and generate strong augmentations for the clean samples. During training, the representations of the samples can be always seen as their weakly augmented views with dropout noise, and the pseudo-labels are produced by these weak augmentations. Thus, by training the model using their strong augmentations we actually applied FixMatch (Sohn et al., 2020). The main idea of FixMatch is to generate pseudo-labels using weak augmentations of the unlabeled samples and train the model to predict the pseudo-labels given the strongly-augmented views of the samples, which is shown to be an effective method for consistency regularization. We use two data augmentation methods: back translation and random masking. The strongly-augmented target sample is denoted as $\mathcal{A}(x)$.

To strengthen the generic model regularization and reduce overfitting, we further use mixup (Zhang et al., 2018) to generate interpolations of the training samples so that the model never sees the original training samples during training. A mixup augmentation is the linear interpolation of the sentence embedding of two random training samples x_i and x_j, where λ is sampled from the Beta distribution:

$$\tilde{x} = \lambda x_i + (1-\lambda)x_j \tag{9}$$
$$\tilde{y} = \lambda y_i + (1-\lambda)y_j \tag{10}$$
$$\lambda \sim Beta(\alpha, \alpha) \tag{11}$$

Finally, we use cross-entropy loss on the hard pseudo-labels during training for the purpose of entropy minimization. Following FixMatch, we found minimizing the cross-entropy loss jointly using the source samples can be helpful for training. During each iteration, the clean samples are first selected using thresholding, then the training set for the iteration is generated with the strong augmentations using either back translation, random masking or mixup. We train the model on the training set to minimize the SSL loss, which is composed of a supervised loss ℓ_{sup} and an unsupervised loss ℓ_u as follows:

$$\mathcal{L}_{SSL} = \ell_{sup} + \lambda \ell_u \quad \text{where} \tag{12}$$
$$\ell_{sup} = H(y_i^{\ell_s}, p(y|x_i^{\ell_s})) \tag{13}$$
$$\ell_u = \mathbb{1}(max(q_i) \geq \gamma) H(\hat{q}_i^{\ell_t}, p(y|\mathcal{A}(x_i^{\ell_t}))) \tag{14}$$

where H denotes cross-entropy, $x_i^{\ell_s}$ is a sample from the source language data, γ is a threshold to select clean pseudo-labeled target samples. Specifically, for a more balanced training sets, the threshold is determined dynamically using FlexMatch (Zhang et al., 2021) during each training iteration.

4. Experiments

4.1. Experiment Settings

Datasets We evaluate our method using XNLI (Conneau et al., 2018), a widely-adopted benchmark to evaluate cross-lingual performance for the natural language inference (NLI) task. XNLI is a three-way classification dataset including labels with entailment, contrastive and neutral, covering 15 languages, each language contains 7500 human-annotated development and test examples. We use the training data of the MultiNLI dataset (Williams et al., 2018) as the source data, which contains 392,702 English samples. The source training data is selected to only cover 5 domains, including Fiction, Government, Slate, Telephone, Travel, while target data is collected from 10 domains with additional 5 sources from Face-To-Face, 9/11, Letters, Oxford University Press (OUP) and Verbatim, which produces a cross-lingual cross-domain setting for the task.

Implementation Details We implement the method using mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) separately as the backbone models. We adopt Adam optimizer and train the model for one epoch during the contrastive learning stage. We find that a smaller learning rate generally helps during the first stage and set the learning rate to be 2e-6 for mBERT and 5e-6 for XLM-R. We use Euclidean distance as the similarity function for Eq.(6), as indicated in (Snell et al., 2017). For the self-training stage, we use mBART (Tang et al., 2021) for back translation and switch to Google Translate for the uncovered language in mBART. For hyperparameters, we set λ in Eq.(8) as 0.001, α in Eq.(11) as 4 and λ in Eq.(13) as 1.

Baselines Apart from the vanilla zero-shot cross-lingual models (**mBERT**, **XLM-R**), we mainly choose two lines of models: alignment-based models, including robust training (**RS-RP**, **RS-DA**) (Huang et al., 2021), virtual multilingual embedding (**EPT-APT**) (Ding et al., 2022), robust representa-

	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	avg
Backbone: mBERT																
mBERT (Huang et al., 2021)	80.8	73.4	73.5	70.0	65.3	68.0	67.8	60.9	64.3	69.3	54.1	67.8	58.9	49.7	57.2	65.4
X-MAML (Nooralahzadeh et al., 2020)	82.1	74.4	75.1	71.8	67.9	69.4	70.2	61.2	66.0	71.8	55.4	71.1	62.2	49.7	61.5	67.3
RS-RP (Huang et al., 2021)	82.6	74.1	75.0	70.5	67.2	68.7	69.5	59.7	65.4	70.5	50.5	69.7	59.8	48.4	57.9	66.0
RS-DA (Huang et al., 2021)	81.0	74.2	74.7	71.8	68.0	69.9	70.6	62.9	66.4	71.8	55.7	71.4	62.7	51.1	60.9	67.6
Syntax (Ahmad et al., 2021)	81.6	73.2	74.1	70.7	66.5	69.3	68.8	62.4	65.4	69.9	-	69.3	60.5	-	58.7	-
EPT-APT (Ding et al., 2022)	83.2	75.2	75.7	72.9	68.3	71.0	71.6	63.6	67.4	72.4	56.7	71.5	64.0	51.3	61.4	68.4
CoNLST	83.9	75.9	75.8	72.7	68.3	70.2	70.7	64.3	66.7	71.8	59.3	72.1	62.6	52.9	61.5	68.6
CoNLST-RS	**84.2**	**76.2**	**76.3**	**73.8**	**69.8**	**71.1**	**72.3**	**65.4**	**67.0**	**73.1**	**59.5**	**73.9**	**64.7**	**53.6**	**63.5**	**69.4**

Table 1: Classification accuracy(%) on XNLI using mBERT as the backbone model. CoNLST-RS indicates that robustly-trained mBERT (Huang et al., 2021) was used as the base model. The bold numbers indicate the highest accuracy score.

	en	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh	avg
Backbone: XLM-R-large																
XLM-R (Hu et al., 2020)	88.7	77.2	83.0	82.5	80.8	83.7	82.2	75.6	79.1	71.2	77.4	78.0	71.7	79.3	78.2	79.2
R3F (Aghajanyan et al., 2020)	89.4	80.6	84.6	**83.7**	**83.6**	**85.1**	**84.2**	77.3	82.3	72.6	79.4	80.7	74.2	81.1	80.1	81.2
SL-EVI (Xu et al., 2021)	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	80.4
SL-LOU (Xu et al., 2021)	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 (Xu et al., 2021)	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
CoNLST	**89.6**	**81.3**	83.7	83.3	82.8	84.6	83.2	77.9	**82.5**	**74.2**	**80.1**	**80.9**	75.2	81.4	**81.5**	**81.5**
Backbone: XLM-R trained on the machine translated data																
XLM-R-MT (Fang et al., 2021)	88.6	82.2	85.2	84.5	84.5	85.7	84.2	80.8	81.8	77.0	80.2	82.1	77.7	82.6	82.7	82.6
FILTER (Fang et al., 2021)	89.5	83.6	**86.4**	85.6	**85.4**	**86.6**	**85.7**	81.1	83.7	78.7	81.7	**83.2**	79.1	83.9	83.8	83.9
CoNLST-MT	**89.6**	**83.7**	85.9	**86.0**	85.3	86.5	85.5	**81.6**	**84.2**	**79.3**	**82.6**	83.1	**79.6**	**84.6**	**83.9**	**84.1**

Table 2: Classification accuracy (%) on XNLI using XLM-R as the backbone model.

tion through regularized finetuning (**R3F**) (Aghajanyan et al., 2020); self-training based models, with different thresholding strategy (**SL-EVI**, **SL-LOU**, **SL-LEU**) (Xu et al., 2021); along with other SOTA models, including meta-learning (**X-MAML**) (Nooralahzadeh et al., 2020), syntax augmentation (**Syntax**) (Ahmad et al., 2021) to compare with the proposed method. The baselines either use mBERT or XLM-R as the backbone model.

4.2. Experiment Results

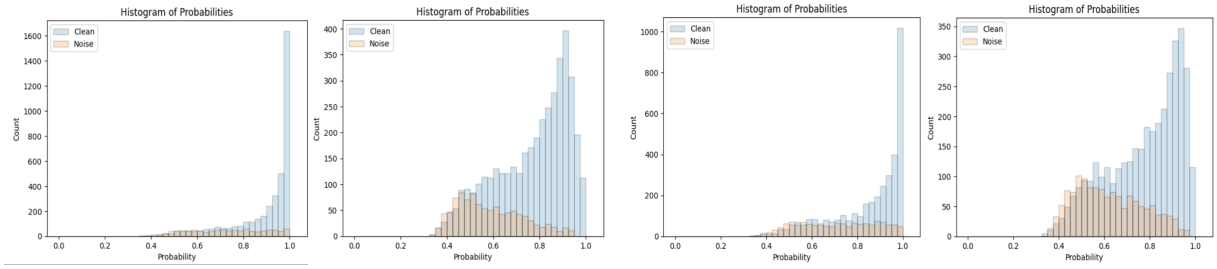
We present the experimental results in Table 1 and Table 2. To test if the proposed method can serve as a beneficial complement to the existing alignment based methods, we build the method further on top of two alignment-based methods and evaluate the performance: machine translation of test data and robust training (Huang et al., 2021). Using machine translation (MT) to translate the target data into the source language is a simple but powerful way to enforce alignment, where the translation model is introduced to produce cross-lingual alignment signals explicitly. Robust training (RS) (Huang et al., 2021) is an effective method for cross-lingual alignment which adopts adversarial examples to simulate the perturbations between cross-lingual inputs and creates a robust region for similar words to have similar predictions. The results of the combined models are shown with CoNLST-MT and CoNLST-RS respectively.

We observe that the proposed CoNLST consistently improves the performance of both mBERT and XLM-R and achieves results that are on par with prior state-of-the-art methods. Notably, the results show that without using any auxiliary language or external data, self-training based methods can attain performance comparable to alignment-based state-of-the-art approaches. Additionally, our proposed self-training framework outperforms earlier standard self-training methods that relied on auxiliary target languages and manually designed confidence thresholds. The proposed method can be viewed as a post-training approach, designed to mine hidden information from the target domain after the training on the source data has been done, so it can be implemented on top of the alignment methods which are usually apply to the source language alone, and further improve the aligned model's performance on the target language, which is supported by the empirical results of combining the proposed method with machine translation or robust training.

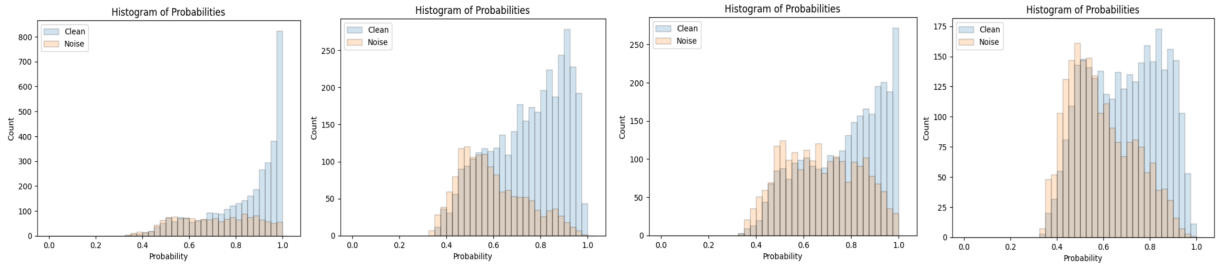
4.3. Analysis

Ablation Study In Table 3, we present the results of an ablation study on the proposed method using mBERT as the backbone to analyze the effect of the proposed contrastive negative learning and the contribution of its different components to the model's performance.

We observe that employing solely contrastive negative learning leads to only small improvements in the model's performance, which indicates that self-training can further improve the model's performance by a large margin. We did not perform ablation study for self-training alone as we show



(a) English (b) French



(c) Chinese (d) Urdu

Figure 3: Noise rate change before and after the contrastive negative learning stage of four languages. Bins in blue indicate clean samples while bins in orange indicate noisy samples. The horizontal axis indicates the **confidence score** of the samples calculated using the model's predicted class distribution; the vertical axis indicates the **number of samples** falling in the area of confidence.

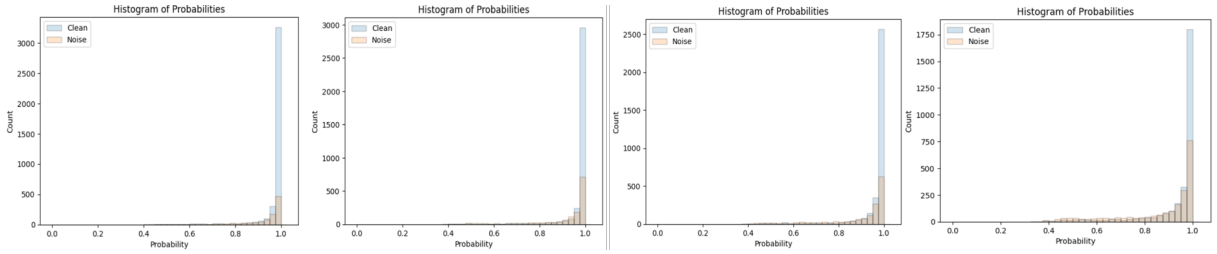


Figure 4: Distribution of noisy and clean sample of four languages (English, French, Chinese, Urdu, in order) after negative learning with only cross entropy loss.

	Acc Avg.
w/o self-training (ST)	66.8
w/o contrastive learning	65.3
w/o $\mathcal{L}_{clcs}$	65.5
w/o mixup	67.3
w/o back-translation	67.1
CoNLST	68.5

Table 3: Ablation study of the proposed method. w/o self-training shows the performance of the model only with contrastive negative learning. w/o contrastive learning indicates that CoNL is replaced with negative learning using only cross entropy loss.

the performance of SOTA self-training methods in Table 2. The results also suggest that extending the negative learning to the metric space is necessary, as replacing the contrastive negative learning with the original negative learning using cross entropy loss alone degenerates the model's performance. Adding the cross-lingual cross-space loss $\mathcal{L}_{clcs}$ is also crucial for the stable training of the model without which the contrastive negative learning tends to fail. The results show that data augmentation plays an important role in enhancing the model's performance during the self-training stage, and the employment of both mixup and back-translation is indispensable for a notable improvements in the model's overall performance.

Confidence Calibration Ability of CoNL We explore the proposed contrastive negative learning's ability to distill clean labels from pseudo-labeled target samples. Figure 3 shows the change of confidence for clean and noisy target samples before and after the contrastive negative learning, Table 4 shows the change of accuracy for target samples in difference confidence areas, and Figure 5 shows the change of overall accuracy for target samples in the selected languages. For demonstration, we select the source language English and three other languages, French, Chinese and Urdu according to their linguistic similarity to the source language.

Lang	High	Mid	Low	Acc Avg.
English	93.3 →96.6	68.6 →83.4	51.7 →56.8	80.6 →82.1
French	89.4 →95.4	66.4 →79.2	50.6 →54.1	73.7 →74.5
Chinese	88.5 →95.1	63.9 →83.1	48.4 →53.1	69.3 →70.1
Urdu	81.8 →94.4	59.5 →74.1	44.9 →48.2	57.5 →59.4

Table 4: Accuracy(%) changes before and after contrastive negative learning of four languages. The samples are divided into three confidence areas: high confidence ($\gamma > 0.9$), middle confidence ($0.5 < \gamma < 0.9$) and low confidence ($\gamma < 0.5$), where γ is the threshold. The accuracy is weighted by the number of samples in each confidence area.

Other languages that are not present show similar patterns in our experiments. We also show the change of noise rate after training the model using only negative learning with cross entropy loss in Figure 4.

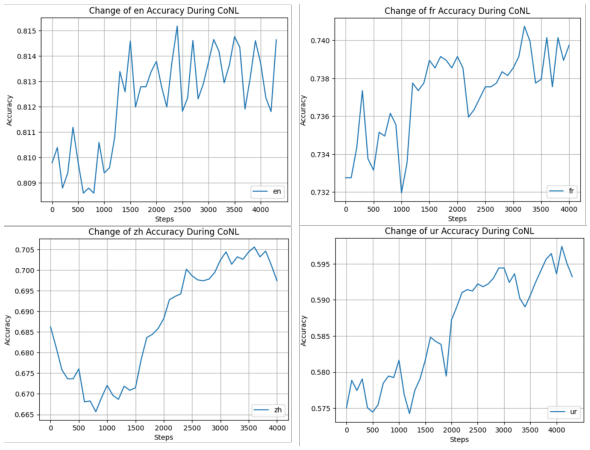


Figure 5: The accuracy change of four languages during contrastive negative learning stage.

According to the change of the noise rate shown in Figure 3, before the contrastive negative learning, the quantity of clean samples exhibits a positive correlation with the model's confidence level. However, the quantity of noisy samples remains consistent across different confidence ranges, indicating that there is a roughly equal distribution of noisy samples among high-confidence samples and among samples of middle and low confidence. Moreover, it can be observed that the proportion of noisy to clean labels in high-confidence area increases when the resource of the language gets scarcer. After CoNL training, noisy samples have lower confidence scores, which indicates that the model becomes less confident about the noisy samples, so that in high-confidence area the noise rate is significantly reduced. In contrast, simply training the model with negative learning using cross entropy loss increases the noise rate in high-confidence area and overfits the model to the noise as shown in Figure 4. This is possibly due to the reason that, different from most of the computer vision tasks where negative learning with cross entropy loss is efficient, in our task the number of class number is limited, so the probability of choosing a wrong complementary label augments significantly, which exacerbates the memorization effect of the model on the hard labels with cross entropy loss.

We further analyze the change of noise rate of samples grouped into different confidence ranges, measured by the accuracy of target samples. The results are shown in Table 4, note that sample numbers also change in different confidence areas. It can be observed that after training, the accuracy of high-confidence, middle-confidence and low-confidence samples improves, and notably, languages with lower resources that are relatively more distant from the source language exhibit greater improvements in the accuracy of high-confidence samples. The results show that the proposed contrastive negative learning is efficient in filtering noisy samples and can distill clean samples in high-confidence area, which reduces the potential confirmation bias for the self-training stage.

5. Conclusions

In this work we explore how the information of the target language data can be leveraged to improve the cross-lingual transfer learning and propose a two-stage method, contrastive negative learning and self-training (CoNLST). We extend negative learning to the metric space to mine class-wise contrastive information from noisy pseudo-labels of the target data and jointly train the model in the label space using source language data, which is shown to be effective in distilling cleanly pseudo-labeled target data, and we further refine the model using the clean data with self-training. Experiments results show that the proposed method can effectively improve the performance of the base models and can be used to improve the alignment-based method in a post-training manner. It should be noted that we assume the only available data, aside from the source language data, to be the data in a specific target language, without the existence of any other auxiliary languages, as found in some real-world scenarios. By simply breaking the assumption of the unavailability of the auxiliary language and extending the sampling pool for negative pairs (Eq.(5)) to all the languages, the proposed method can be adapted to jointly training multiple target languages when data from similar domains are available in multiple languages. Besides, an in-depth qualitative analysis on the changes of the samples before and after the confidence calibration would also help to investigate the multilingual ability of the backbone models, which we plan to explore in the future work.

6. Bibliographical References

Armen Aghajanyan, Akshat Shrivastava, Anchit Gupta, Naman Goyal, Luke Zettlemoyer, and Sonal Gupta. 2020. Better fine-tuning by reducing representational collapse. *arXiv preprint arXiv:2008.03156*.

Wasi Ahmad, Haoran Li, Kai-Wei Chang, and Yashar Mehdad. 2021. Syntax-augmented multilingual BERT for cross-lingual transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4538–4554, Online. Association for Computational Linguistics.

Görkem Algan and Ilkay Ulusoy. 2021. Image classification with deep learning in the presence of noisy labels: A survey. *Knowledge-Based Systems*, 215:106771.

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

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. 2019. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2018. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6:557–570.

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.

Jingjing Cui. 2012. Untranslatability and the method of compensation. *Theory & Practice in Language Studies*, 2(4).

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.

Kunbo Ding, Weijie Liu, Yuejian Fang, Weiquan Mao, Zhe Zhao, Tao Zhu, Haoyan Liu, Rong Tian, and Yiren Chen. 2022. A simple and effective method to improve zero-shot cross-lingual transfer learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4372–4380.

Xin Luna 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.

Yuwei Fang, Shuohang Wang, Zhe Gan, Siqi Sun, and Jingjing Liu. 2021. Filter: An enhanced fusion method for cross-lingual language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12776–12784.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Lin Gui, Ruifeng Xu, Qin Lu, Jun Xu, Jian Xu, Bin Liu, and Xiaolong Wang. 2014. Cross-lingual opinion analysis via negative transfer detection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 860–865.

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.

Kuan-Hao Huang, Wasi Ahmad, Nanyun Peng, and Kai-Wei Chang. 2021. Improving zero-shot cross-lingual transfer learning via robust training. In

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 1684–1697.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Youngdong Kim, Junho Yim, Juseung Yun, and Junmo Kim. 2019. Nlnl: Negative learning for noisy labels. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 101–110.

Youngdong Kim, Juseung Yun, Hyounguk Shon, and Junmo Kim. 2021. Joint negative and positive learning for noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9442–9451.

Saurabh Kulshreshtha, Jose Luis Redondo Garcia, and Ching Yun Chang. 2020. Cross-lingual alignment methods for multilingual bert: A comparative study. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 933–942.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499.

Hung-Yi Lee, Shang-Wen Li, and Thang Vu. 2022. Meta learning for natural language processing: A survey. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 666–684.

Juntao Li, Ruidan He, Hai Ye, Hwee Tou Ng, Lidong Bing, and Rui Yan. 2021. Unsupervised domain adaptation of a pretrained cross-lingual language model. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3672–3678.

Hong Liu, Jianmin Wang, and Mingsheng Long. 2021. Cycle self-training for domain adaptation. *Advances in Neural Information Processing Systems*, 34:22968–22981.

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

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914.

Lin Pan, Chung-Wei Hang, Haode Qi, Abhishek Shah, Saloni Potdar, and Mo Yu. 2021. Multilingual bert post-pretraining alignment. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 210–219.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608.

Hui Tang, Ke Chen, and Kui Jia. 2020. Unsupervised domain adaptation via structurally regularized deep clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8725–8735.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466.

Thuy Vu, Xuanli He, Dinh Phung, and Gholamreza Haffari. 2021. Generalised unsupervised domain adaptation of neural machine translation with cross-lingual data selection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3335–3346.

Yaushian Wang, Ashley Wu, and Graham Neubig. 2022. English contrastive learning can learn universal cross-lingual sentence embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9122–9133, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Junyuan Xie, Ross Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487. PMLR.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268.

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.

Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. 2021. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34:18408–18419.

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Dawei Zhu, Michael Hedderich, Fangzhou Zhai, David Adelani, and Dietrich Klakow. 2022. Is bert robust to label noise? a study on learning with noisy labels in text classification. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 62–67.