ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230364

Mitigating Long-Tail Language Representation Collapsing via Cross-Lingual Bootstrapped Unsupervised Fine-Tuning

Ping Guo[1,2], **Yue Hu**[1,2;*], **Yubing Ren**[1,2], **Yunpeng Li**[1,2], **Jiarui Zhang**[1,2] **and Xingsheng Zhang**[1,2]

[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
{guoping, huyue, renyubing, liyunpeng, zhangjiarui, zhangxingsheng}@iie.ac.cn

Abstract. Large Language Models have shown great capability to comprehend natural language and provide reasonable responses. However, previous researches have shown weak performance of these models on low-resource (long-tail) languages. It remains to be a problem to mitigate the performance gap between long-tail languages and rich-resource ones, which is referred to as long-tail language representation collapsing. Though some previous works can generate pseudo-parallel corpora with the auto-regressive generation, this generation progress is time-consuming and remains low quality, particularly for long-tail languages. In this paper, we propose a (**X**) Cross-lingual **B**ootstrapped **U**nsupervised **F**ine-tuning **F**ramework (X-BUFF) to mitigate long-tail language representation collapsing. X-BUFF iteratively updates cross-lingual PLMs in a curriculum way. In each iteration of X-BUFF, we (1) select sentences with complementary semantics from monolingual corpora in long-tail languages. (2) match these selected sentences with semantic equivalent sentences in many other languages to create parallel sentence pairs, which we then merge with previous sentence pairs to build a larger and more difficult bootstrapped parallel queue. (3) fine-tune the PLMs with the bootstrapped parallel queue. Extensive experiments show that X-BUFF can mitigate the long-tail language representation collapsing problem in cross-lingual PLMs and achieve significant improvements over the previous baselines on several cross-lingual evaluation benchmarks.

1 Introduction

Large language models (LLMs) (e.g., ChatGPT, GPT-3 [9], LaMDA [44], PaLM [14], etc.) have led to great improvements on numerous natural language tasks, such as text generation [37], text classification [42] and natural language inference [7, 49]. However, in the cross-lingual scenario, these LLMs have shown weak performance on many languages [29, 37], even lag behind some traditional Pre-trained Language Models (PLMs) [38, 39, 18] on low resource languages. Hence, it remains to be solved how to improve cross-lingual transferability on natural language tasks, especially on low-resource languages.

The typical cross-lingual PLMs, such as mBERT [18], XLM [16], and XLM-R [15], have demonstrated promising knowledge transferability across a huge number of languages. Cross-lingual PLMs usually involve multilingual masked language modeling (MLM) [18] or translation language modeling (TLM) [16] as the training objective. The goal of both tasks is to reproduce each masked token given the rest original ones. They typically perform token-level optimization, without sufficient sentence-level optimization. This poses a big challenge for cross-lingual models to capture the exact meaning of the whole sentence. To this end, various methods have been developed to learn sentence-level representations that can express global semantics, with several well-designed techniques such as contrastive learning [26, 47, 23, 11, 30, 52, 53] and disentanglement [50]. Although these methods have improved the capability of language-agnostic representations, studies reveal that the performance of cross-lingual transfer fluctuates within a wide range among different languages.

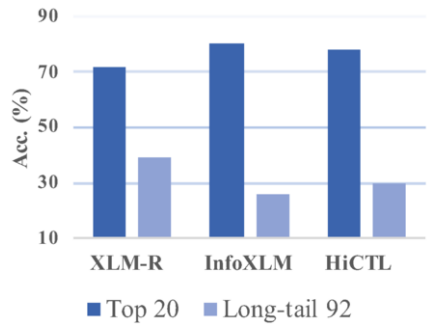

Figure 1: Performance discrepancy between top-20 and long-tail 92 languages on Tatoeba [4], involving three promising cross-lingual PLMs: XLM-R [15], INFOXLM [11] and HICTL [47].

We investigate this on a cross-lingual sentence retrieval task, e.g., Tatoeba [4], and evaluate performance without fine-tuning on all 112 languages. We observe that XLM-R [15] demonstrates an absolute 48% discrepancy of retrieval accuracy between the top 20 languages and the rest (we refer to these languages as long-tail languages). And this statistic further rises to 51% and 54% towards the performances of HICTL [47] and INFOXLM [11], respectively. We call this the *representation collapsing problem* of long-tail languages. This result severely hinders the usage of cross-lingual PLMs, especially on those tasks in long-tail language scenarios. Some existing works [28, 36] have tried to generate a pseudo-parallel corpus to balance the scale of parallel corpora across languages. However, the generation of pseudo-parallel corpora is often time-consuming and shows low quality, particularly for long-tail languages. For some languages (such as pa, or, ps, etc.), generating pseudo-parallel corpora is not even available.

The underlying theme of this problem is that the amount of parallel data varies dramatically across languages. More concretely, parallel corpora are mainly distributed in a few dozen head languages, while

* Corresponding Author. Email: huyue@iie.ac.cn.

most long-tail languages only possess a small amount of bilingual data [2, 5]. With this in mind, we argue that existing state-of-the-art techniques have severe limitations to learning better cross-lingual representations. For those languages without parallel data, it is difficult to pre-train a cross-lingual encoder with contrastive learning. Inspired by the bootstrapping algorithm in information extraction [8, 1], we try to take a small parallel corpus as a seed and gradually accumulate more parallel sentence pairs with certain rules. However, as semantics are abstract, it is still hard to apply traditional bootstrapping methods [22] to accumulate sentence pairs, for we cannot find universal traceable syntax rules among any two parallel sentences.

In this paper, we propose **X-BUFF**, which aims at fine-tuning the long-tail language representations of cross-lingual PLMs with monolingual corpora in a curriculum way. X-BUFF forms the supervision signal by iteratively constructing parallel sentence pairs and simultaneously fine-tuning the cross-lingual PLM within each iteration. The iteratively accumulated parallel sentence pairs are denoted as bootstrapped parallel queue. Specifically, in each iteration of X-BUFF, we (1) select sentences with semantics complement to bootstrapped parallel queue, where semantic similarity is measured via a cross-lingual PLM. (2) match these chosen sentences with semantic equivalent sentences in many other languages using a semantic relation classifier to create parallel sentence pairs, which we then merge into bootstrapped parallel queue. (3) fine-tune the cross-lingual PLM and semantic relation classifier with the bootstrapped parallel queue. During the expansion of the bootstrapped parallel queue, the changes in the semantic level of the queue will provide training samples ranging from "easy" to "hard". Hence, fine-tuning the two neural modules alongside the expansion of bootstrapped parallel queue provides an intrinsic curriculum learning [6, 32]. Extensive experiments are conducted on 5 cross-lingual tasks: XNLI [17], AmericasNLI [19], WMT21 QE Task 1 [43], Tatoeba [4] and MultiEURLEX [10], which cover different types with more than 50 languages. Results demonstrate that X-BUFF significantly outperforms state-of-the-art results on these 5 tasks. Moreover, X-BUFF greatly improves the performance of long-tail languages. In summary, our contributions are as follows:

- To mitigate long-tail language representation collapsing, we propose X-BUFF, which can fine-tune the long-tail language representations of cross-lingual PLMs with monolingual corpora in a curriculum way.
- To provide comprehensive and high-quality supervision signals, X-BUFF unifies the construction of parallel corpora and the training of cross-lingual representations in a boot-strapped framework.
- To better adapt to new long-tail languages, we finetune the cross-lingual PLM in a curriculum way using training samples from bootstrapped parallel queue ranging from "easy" to "hard".
- Extensive experiments are conducted on several cross-lingual tasks, covering more than 50 languages and 4 different types. Results demonstrate X-BUFF outperforms the state-of-the-art on these tasks.

2 X-BUFF: Cross-lingual Bootstrapped Unsupervised Fine-tuning Framework

We propose X-BUFF, which can fine-tune the long-tail language representation of cross-lingual PLMs with excessive monolingual corpora. X-BUFF can iteratively construct a bootstrapped parallel queue with monolingual corpora and simultaneously fine-tune neural modules. Each iteration contains three steps, as illustrated in Figure 2, (1) search sentences with complementary semantics. (2) match chosen complement sentences to parallel sentence pairs and update bootstrapped parallel queue. (3) curriculum fine-tune cross-lingual PLM with the new bootstrapped parallel queue. Details about the iteration process are in §2.2. The iteration will first start on one long-tail language and continues to other long-tail languages until convergence on this long-tail language. We further demonstrate the whole process of X-BUFF in Algorithm 1.

2.1 Basic Notations

Long-tail Languages Languages that possess only a few parallel bilingual data, or do not have any parallel data, are referred to as long-tail languages $\mathcal{T}$. These long-tail languages usually have excessive monolingual data. In X-BUFF, for one long-tail language $\mathbf{x} \in \mathcal{T}$, we denote its monolingual corpus as $\mathcal{D}_{\mathbf{x}} = \{\mathbf{x}_d\}$ with $\mathbf{x}_d$ to be the sentence in the monolingual data.

Anchor Languages Languages with large parallel corpora are denoted as anchor languages, and they can be easily aligned into the unified continuous space by traditional pre-training techniques. Hence, we use sentences in these languages as anchor points to help map other long-tail languages to the representation space. In X-BUFF, we group all sentences in anchor languages together as an anchor language corpus $\mathcal{D}_{\mathbf{k}} = \{\mathbf{k}_{d'}\}$, where $\mathbf{k}_{d'}$ means the sentence in anchor language corpus.

Bootstrapped Parallel Queue We refer to the parallel corpus we construct at iteration t as bootstrapped parallel queue $\mathcal{Q}^t_{\langle \mathbf{x},\mathbf{k} \rangle}$. It contains parallel sentence pair $\langle \mathbf{x}_q, \mathbf{k}_q \rangle$, where $\mathbf{x}_q$ are selected from monolingual long-tail language corpus $\mathcal{D}_{\mathbf{x}}$, and $\mathbf{k}_q$ denotes a semantic equivalent counterparts of $\mathbf{x}_q$ matched from anchor language corpus $\mathcal{D}_{\mathbf{k}}$. The bootstrapped parallel queue will iteratively accumulate new sentence pairs and it expands as a snowball.

Complement Sentence/Parallel Set In iteration t, we refer to the sentences we selected in step 1 as complement sentence set $\mathcal{C}^t_{\langle \mathbf{x} \rangle}$, which contains sentence $\mathbf{x}_c$ whose semantics are complement to sentences in bootstrapped parallel queue. Then, in step 2, we refer to the matched parallel sentence pairs as complement parallel set $\mathcal{C}^t_{\langle \mathbf{x},\mathbf{k} \rangle}$, which contains sentence $\mathbf{x}_c$ in long-tail languages and its semantic equivalent counterparts $\mathbf{k}_c$ in anchor language corpus $\mathcal{D}_{\mathbf{k}}$. Sentence pair $\langle \mathbf{x}_c, \mathbf{k}_c \rangle$ in complement parallel set is accumulated into bootstrapped parallel queue during each iteration t.

2.2 Iteration Process of X-BUFF

X-BUFF will first start on one long-tail language and continues to other long-tail languages until convergence on this long-tail language.

Initialization

We first initialize the bootstrapped parallel queue $\mathcal{Q}^0_{\langle \mathbf{x},\mathbf{k} \rangle}$ with the few parallel bilingual data for long-tail language $\mathbf{x}$. If long-tail language $\mathbf{x}$ does not possess any bilingual data, we set $\mathcal{Q}^0_{\langle \mathbf{x},\mathbf{k} \rangle} = \emptyset$.

Step 1: Select Sentences with Complementary Semantics

To enrich the semantic diversity of the bootstrapped parallel queue, X-BUFF focuses on selecting sentences, showing minimum semantic

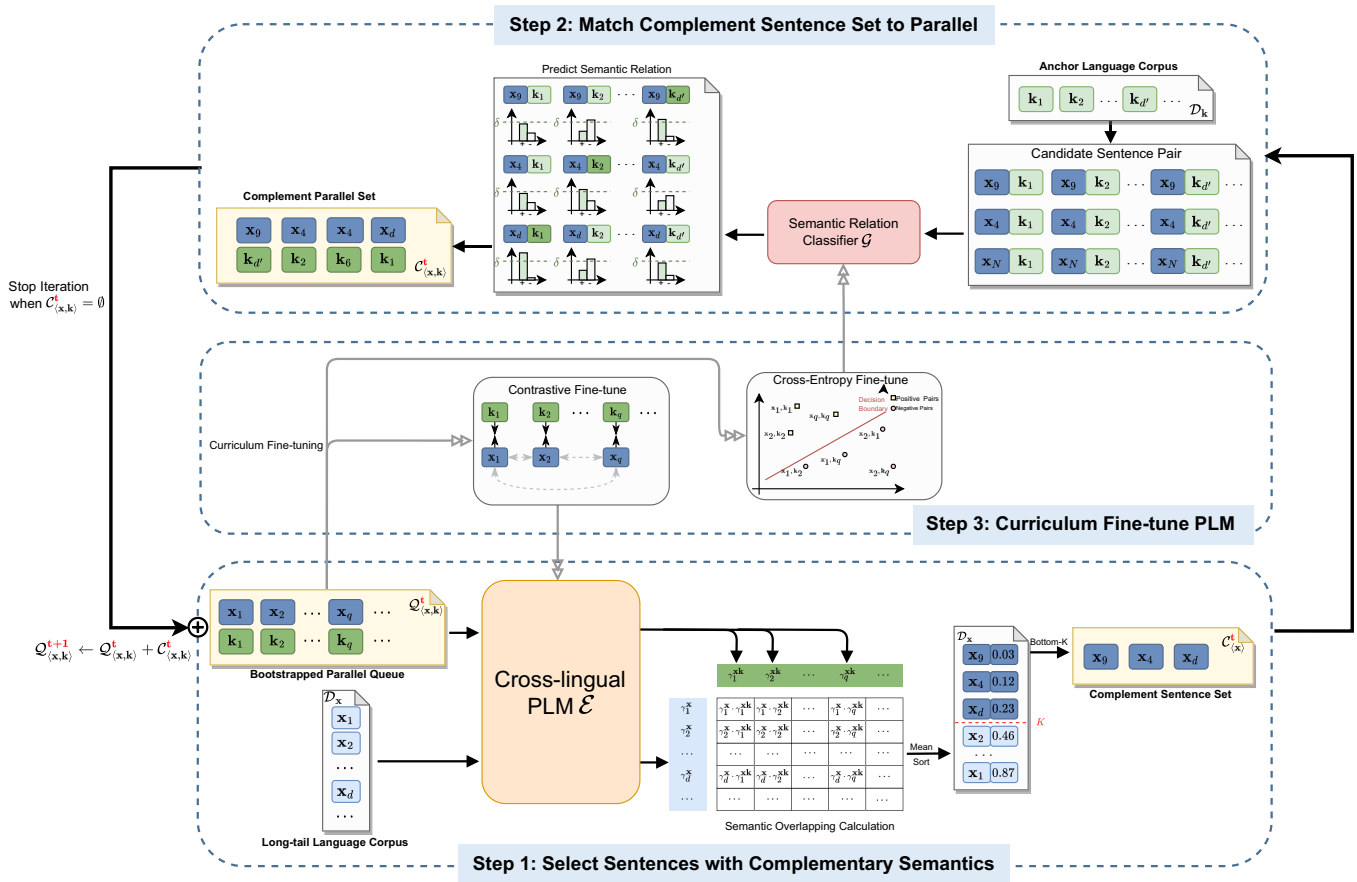

Figure 2: Illustration of one iteration for X-BUFF, involving three steps. **Step 1:** We select sentences $\mathbf{x}_9, \mathbf{x}_4, \mathbf{x}_d$ (dark blue) from long-tail language corpus $\mathcal{D}_\mathbf{x}$, for the semantics in these sentences is complement to bootstrapped parallel queue $\mathcal{Q}^t_{\langle \mathbf{x},\mathbf{k}\rangle}$. **Step 2:** We retrieve sentences in anchor languages from $\mathcal{D}_\mathbf{k}$ to search sentences (dark green) which are semantic equivalent to sentences in Step 1. **Step 3:** We use the new searched sentence pairs ($\mathcal{C}^t_{\langle \mathbf{x},\mathbf{k}\rangle}$) to update bootstrapped parallel queue $\mathcal{Q}^t_{\langle \mathbf{x},\mathbf{k}\rangle}$ and use it to fine-tune cross-lingual PLM $\mathcal{E}$ and semantic relation classifier $\mathcal{G}$. The whole process convergences when $\mathcal{C}^t_{\langle \mathbf{x},\mathbf{k}\rangle} = \emptyset$.

overlap with the existing sentences in bootstrapped parallel queue, from monolingual corpus $\mathcal{D}_\mathbf{x}$. To accurately model the semantic relationship between bootstrapped parallel queue and monolingual corpus, we apply a cross-lingual PLM $\mathcal{E}$ to encode sentences into continuous dense representations. Formally, given a sentence $\mathbf{x}$, the continuous representation is formulated as:

$$\gamma^\mathbf{x} = g\big(\mathcal{E}(\mathbf{x}; \Theta_\mathcal{E})\big) \tag{1}$$

where $\mathcal{E}(\cdot; \Theta_\mathcal{E})$ denotes the Cross-lingual PLM with a set of trainable parameters $\Theta_\mathcal{E}$, $\gamma^\mathbf{x}$ means the output sentence representations, and $g(\cdot)$ is the function to normalize $||\gamma^\mathbf{x}|| = 1$. In X-BUFF, we take the final hidden states of "`[CLS]`" token as sentence representation.

Formally, the semantic similarity between the sentence $\mathbf{x}_d$ from monolingual corpus and bootstrapped parallel queue $\mathcal{Q}^t_{\langle \mathbf{x},\mathbf{k}\rangle}$ formulates as the mean-averaged of the similarity score between $\mathbf{x}_d$ and every sentence pairs $\langle \mathbf{x}_q, \mathbf{k}_q\rangle$ in bootstrapped parallel queue:

$$\begin{aligned} s(\mathbf{x}_d, \mathcal{Q}^t_{\langle \mathbf{x},\mathbf{k}\rangle}) &= \mathbb{E}_{\langle \mathbf{x}_q,\mathbf{k}_q\rangle \sim \mathcal{Q}^t_{\langle \mathbf{x},\mathbf{k}\rangle}} s(\mathbf{x}_d, \langle \mathbf{x}_q, \mathbf{k}_q\rangle) \\ &= \mathbb{E}_{\langle \mathbf{x}_q,\mathbf{k}_q\rangle \sim \mathcal{Q}^t_{\langle \mathbf{x},\mathbf{k}\rangle}} \gamma^\mathbf{x}_d \cdot \gamma^\mathbf{xk}_q \end{aligned} \tag{2}$$

where we use the fact that $||\gamma^\mathbf{x}_d|| = ||\gamma^\mathbf{xk}_q|| = 1$ and $\gamma^\mathbf{xk}_q$ denotes the representation for sentence pairs $\langle \mathbf{x}_q, \mathbf{k}_q\rangle$, which calculates as the mean average of two sentence representations. Later, we take N sentences from monolingual corpus with the lowest similarity score

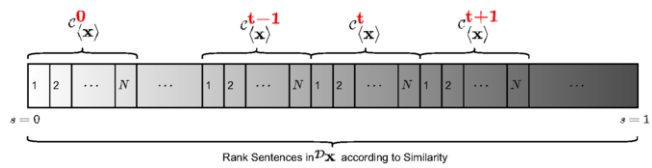

Figure 3: Illustration about construct complement sentence set. We rank all sentences in long-tail language corpus in descending order according to similarity and choose the bottom-N sentences to construct complement sentence set. As we have selected the bottom-N sentences at time t, we can only select less dissimilar sentences in iteration $t+1$.

to build complement sentence set $\mathcal{C}^t_{\langle \mathbf{x}\rangle}$:

$$\mathcal{C}^t_{\langle \mathbf{x}\rangle} = \Big\{ \mathbf{x}_d \Big| \min\big(s(\mathbf{x}_d, \mathcal{Q}^t_{\langle \mathbf{x},\mathbf{k}\rangle}), N\big), \mathbf{x}_d \in \mathcal{D}_\mathbf{x} \Big\} \tag{3}$$

where $\min(s, N)$ denotes the N sentences which have the minimum semantic similarity s, as shown in Figure 3. These sentences can complement the original semantic information in bootstrapped parallel queue, making it more informative and comprehensive.

If bootstrapped parallel queue $\mathcal{Q}^0_{\langle \mathbf{x},\mathbf{k}\rangle} = \emptyset$, we will randomly select some sentences from $\mathcal{D}_\mathbf{x}$ as complement sentence set $\mathcal{C}^0_{\langle \mathbf{x}\rangle}$.

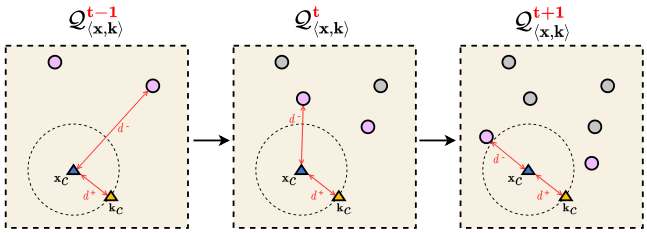

Figure 4: Illustration about curriculum fine-tuning. We use triangles to show the query $\mathbf{x}_c$ and its positive sample $\mathbf{k}_c$. The pink dot refers to new-added negative samples while the grey one means previous negative samples. With the expansion of bootstrapped parallel queue, the discrepancy between d^+ and d^- decreases, which demonstrates $\mathcal{Q}_{\langle \mathbf{x},\mathbf{k}\rangle}$ can provide sentence pairs ranging from easy negative samples to hard negative samples.

Step 2: Match Complement Sentence Set to Parallel

To build parallel sentence pairs for long-tail languages, X-BUFF matches each sentence in complement sentence set $\mathbf{x}_c \in \mathcal{C}^t_{\langle \mathbf{x}\rangle}$ with semantic equivalent sentences retrieved from anchor language corpus $\mathcal{D}_\mathbf{k}$. We first group sentence $\mathbf{x}_c$ with all sentences $\mathbf{k}_{d'}$ in anchor language corpus as candidate sentence pairs, and use a semantic relation classifier $\mathcal{G}$ to predict semantic relationship for each candidate sentence pair. Take one candidate sentence pair $\mathbf{x}_c$ and $\mathbf{k}_{d'}$ as input, it calculates:

$$\mathcal{G}(\mathbf{x}_c, \mathbf{k}_{d'}; \Theta_\mathcal{G}) = \sigma\big(\mathbf{w}^T(\gamma^\mathbf{x}_c - \gamma^\mathbf{k}_{d'}) + b\big) \tag{4}$$

where $\mathcal{G}(\cdot)$ is the output confidence score, $\Theta_\mathcal{G} = \{\mathbf{w}, b\}$ refers to the parameters of classifier $\mathcal{G}$ and $\sigma(\cdot)$ is the sigmoid function. The classifier takes the sentence representations $\gamma^\mathbf{x}_c$ and $\gamma^\mathbf{k}_{d'}$ from cross-lingual PLM $\mathcal{E}$ as inputs. We take the sentence pairs with confidence score over a threshold δ as the corresponding semantic equivalent sentence for the sentence $\mathbf{x}_c$. Formally, we match complement sentence set to complement parallel set $\mathcal{C}^t_{\langle \mathbf{x},\mathbf{k}\rangle}$ as:

$$\mathcal{C}^t_{\langle \mathbf{x},\mathbf{k}\rangle} = \big\{\langle \mathbf{x}_c, \mathbf{k}_{d'}\rangle \big| \mathcal{G}(\mathbf{x}_c, \mathbf{k}_{d'}) > \delta, \mathbf{x}_c \in \mathcal{C}^t_{\langle \mathbf{x}\rangle}, \mathbf{k}_{d'} \in \mathcal{D}_\mathbf{k}\big\} \tag{5}$$

After the construction of complement parallel set, X-BUFF will update the bootstrapped parallel queue $\mathcal{Q}^t_{\langle \mathbf{x},\mathbf{k}\rangle}$ with complement parallel set $\mathcal{C}^t_{\langle \mathbf{x},\mathbf{k}\rangle}$:

$$\mathcal{Q}^{t+1}_{\langle \mathbf{x},\mathbf{k}\rangle} \leftarrow \mathcal{Q}^t_{\langle \mathbf{x},\mathbf{k}\rangle} + \mathcal{C}^t_{\langle \mathbf{x},\mathbf{k}\rangle} \tag{6}$$

Step 3: Curriculum Fine-tune PLM *with new Bootstrapped Parallel Queue*

Intrinsic Curriculum in Bootstrapped Parallel Queue In iteration t of X-BUFF, we build the complement sentence set $\mathcal{C}^t_{\langle \mathbf{x}\rangle}$ using the most dissimilar N sentences with the current bootstrapped parallel queue $\mathcal{Q}^t_{\langle \mathbf{x},\mathbf{k}\rangle}$, as demonstrated in Eq. 3. With t increasing, the similarity relationship between $\mathcal{C}^t_{\langle \mathbf{x}\rangle}$ and $\mathcal{Q}^t_{\langle \mathbf{x},\mathbf{k}\rangle}$ will also increase. This is intuitive for more dissimilar sentences that are already selected in the previous steps, as illustrated in Figure 3. Hence, the similarity relationship between complement parallel set $\mathcal{C}^t_{\langle \mathbf{x},\mathbf{k}\rangle}$ and bootstrapped parallel queue $\mathcal{Q}^t_{\langle \mathbf{x},\mathbf{k}\rangle}$ has:

$$s(\mathcal{C}^{t-1}_{\langle \mathbf{x},\mathbf{k}\rangle}, \mathcal{Q}^{t-1}_{\langle \mathbf{x},\mathbf{k}\rangle}) < s(\mathcal{C}^{t}_{\langle \mathbf{x},\mathbf{k}\rangle}, \mathcal{Q}^{t}_{\langle \mathbf{x},\mathbf{k}\rangle}) < s(\mathcal{C}^{t+1}_{\langle \mathbf{x},\mathbf{k}\rangle}, \mathcal{Q}^{t+1}_{\langle \mathbf{x},\mathbf{k}\rangle}) \tag{7}$$

This reveals that we expand the bootstrapped parallel queue with sentence pairs from dissimilar to similar, naturally organized in a curriculum order. Next, we will introduce the specific log-likelihood function to curriculum fine-tune each module in detail.

Algorithm 1: X-BUFF Algorithm

Input: Long-tail languages set $\mathcal{T}$, monolingual corpus $\mathcal{D}_\mathbf{x}$ for each long-tail language $\mathbf{x}$, and anchor language corpus $\mathcal{D}_\mathbf{k}$.
Output: Cross-lingual Encoder $\mathcal{E}(\cdot;\Theta_\mathcal{E})$.
for *long-tail language* $\mathbf{x} \in \mathcal{T}$ **do**
  $t \leftarrow 0$;
  while *not convergence* **do**
    /* Step 1: Select Sentences with Complement Semantics */
    if $\mathcal{Q}^t_{\langle \mathbf{x},\mathbf{k}\rangle} = \emptyset$ **then**
      Construct $\mathcal{C}^t_{\langle \mathbf{x}\rangle}$ with random sentences from $\mathcal{D}_\mathbf{x}$;
    else
      for $\mathbf{x}_d \in \mathcal{D}_\mathbf{x}$ **do**
        Calculate semantic overlapping between $\mathbf{x}_d$ and $\mathcal{Q}^t_{\langle \mathbf{x},\mathbf{k}\rangle}$ with Eq. 2 ;
      end
      Construct $\mathcal{C}^t_{\langle \mathbf{x}\rangle}$ by selecting N dissimilar sentences according to Eq. 3 ;
    end
    /* Step 2: Match Complement Sentence Set to Parallel */
    for $\langle \mathbf{x}_c, \mathbf{k}_{d'}\rangle \in \big(\mathcal{C}^t_{\langle \mathbf{x}\rangle} \times \mathcal{D}_\mathbf{k}\big)$ **do**
      Predict the semantic relationship between $\mathbf{x}_c$ and $\mathbf{k}_{d'}$ with Eq. 4 ;
    end
    Construct $\mathcal{C}^t_{\langle \mathbf{x},\mathbf{k}\rangle}$ according to Eq. 5 ;
    $\mathcal{Q}^{t+1}_{\langle \mathbf{x},\mathbf{k}\rangle} \leftarrow \mathcal{Q}^t_{\langle \mathbf{x},\mathbf{k}\rangle} + \mathcal{C}^t_{\langle \mathbf{x},\mathbf{k}\rangle}$;
    /* Step 3: Curriculum Fine-tuning */
    for $\langle \mathbf{x}_c, \mathbf{k}_c\rangle \in \mathcal{C}^t_{\langle \mathbf{x},\mathbf{k}\rangle}$ **do**
      Sample negative sentences $\mathbf{k}_q$ from $\mathcal{Q}^t_{\langle \mathbf{x},\mathbf{k}\rangle}$;
      Calculate $\mathcal{L}(\Theta_\mathcal{E})$ with Eq. 8, update $\Theta_\mathcal{E}$;
      Calculate $\mathcal{L}(\Theta_\mathcal{G})$ with Eq. 9, update $\Theta_\mathcal{G}$;
    end
    $t \leftarrow t + 1$;
  end
end

Fine-tuning Cross-lingual PLM $\mathcal{E}$ We apply contrastive learning (INFONCE [35]) to update parameters $\Theta_\mathcal{E}$. Formally, given a parallel sentence pair $\langle \mathbf{x}, \mathbf{k}\rangle$, INFONCE loss encourages $\mathbf{x}$ to be as similar as possible to $\mathbf{k}$ (the positive sample) but dissimilar to other instance $\mathbf{k}'$ (the negative sample). To apply curriculum fine-tuning, in iteration t, we choose parallel sentence pair $\langle \mathbf{x}_c, \mathbf{k}_c\rangle$ from complement parallel set $\mathcal{C}^t_{\langle \mathbf{x},\mathbf{k}\rangle}$ and sampling negative samples from previous bootstrapped parallel queue $\mathcal{Q}^{t-1}_{\langle \mathbf{x},\mathbf{k}\rangle}$. Formally, the likelihood function to update $\Theta_\mathcal{E}$ is:

$$\mathcal{L}(\Theta_\mathcal{E}) = -\sum_{c=1}^{|\mathcal{C}^t_{\langle \mathbf{x},\mathbf{k}\rangle}|} \log\Big[\frac{e^{s(\mathbf{x}_c,\mathbf{k}_c)}}{e^{s(\mathbf{x}_c,\mathbf{k}_c)} + \sum_{\mathbf{k}_q \in \mathcal{Q}^{t-1}_{\langle \mathbf{x},\mathbf{k}\rangle}} e^{s(\mathbf{x}_c,\mathbf{k}_q)}}\Big], \tag{8}$$

where $|\mathcal{C}^t_{\langle \mathbf{x},\mathbf{k}\rangle}|$ denotes the total number of parallel sentence pairs in $\mathcal{C}^t_{\langle \mathbf{x},\mathbf{k}\rangle}$. This function will first fine-tune Cross-lingual PLM to separate positive samples with easy negative samples ($s(\mathbf{x}_c, \mathbf{k}_q) \ll s(\mathbf{x}_c, \mathbf{k}_c)$), and later focus on hard negative samples ($s(\mathbf{x}_c, \mathbf{k}_q) < s(\mathbf{x}_c, \mathbf{k}_c)$) as illustrated in Figure 4.

Fine-tuning Semantic Relation Classifier $\mathcal{G}$ The curriculum fine-tuning process for semantic relation classifier is similar to that for cross-lingual encoder. We take a sentence pair $\langle \mathbf{x}_c, \mathbf{k}_c\rangle$ from complement parallel set $\mathcal{C}^t_{\langle \mathbf{x},\mathbf{k}\rangle}$ and assign its semantic relation label to 1 (semantic equivalent). Then, We pair $\mathbf{x}_c$ with instances $\mathbf{k}_q$ sampled from previous bootstrapped parallel queue $\mathcal{Q}^{t-1}_{\langle \mathbf{x},\mathbf{k}\rangle}$ as negative pairs with label 0. During fine-tuning, parameters in cross-lingual PLM are

XNLI & Americans NLI (Accuracy)													
Model\Language	**en**	**ar**	**bg**	**de**	**el**	**es**	**fr**	**hi**	**ru**	**sw**	**th**	**tr**	**ur**
XLM-R	88.68	83.45	86.58	83.46	84.96	84.78	83.15	79.47	81.06	82.31	80.19	82.64	77.98
HICTL	90.39	86.75	87.76	86.12	86.67	87.37	85.38	83.45	82.88	83.76	82.46	85.78	80.94
INFOXLM	90.25	83.64	86.38	85.98	85.09	87.68	86.58	83.46	82.95	82.91	80.89	84.52	80.87
X-BUFF	**91.58**	**88.89**	**89.45**	**86.67**	**87.03**	**88.24**	**87.31**	**86.25**	**84.68**	**87.85**	**84.13**	**88.97**	**86.90**
	vi	**zh**	**aym**	**bzd**	**cni**	**gn**	**hch**	**nah**	**oto**	**quy**	**tar**	**shp**	**Avg.**
XLM-R	78.86	77.57	48.97	50.43	41.78	58.23	42.33	54.70	35.89	59.41	51.73	41.38	68.80
HICTL	79.85	80.98	49.68	51.21	42.39	58.82	43.11	55.10	36.00	59.70	52.28	42.03	70.43
INFOXLM	81.34	80.75	49.88	51.32	42.44	58.85	43.09	55.29	36.12	59.73	52.29	42.08	70.18
X-BUFF	**83.47**	**83.82**	**51.13**	**52.23**	**43.59**	**59.74**	**44.23**	**55.33**	**39.14**	**60.06**	**52.87**	**43.12**	**72.28**
MultiEURLEX (Accuracy)													
Model\Language	**en**	**de**	**fr**	**it**	**es**	**pl**	**ro**	**nl**	**el**	**hu**	**pt**	**cs**	**mt**
XLM-R	67.56	66.78	67.87	68.12	67.94	66.75	67.94	67.65	66.23	65.96	67.54	67.51	61.98
HICTL	68.03	67.94	69.05	68.98	69.43	68.23	69.31	68.56	67.21	67.12	68.43	68.68	63.22
INFOXLM	68.15	68.56	69.87	69.32	69.89	68.83	70.11	69.23	67.11	67.47	68.99	69.47	63.84
X-BUFF	**70.03**	**70.43**	**71.02**	**69.94**	**70.45**	**69.68**	**71.35**	**70.57**	**69.85**	**68.48**	**70.02**	**70.43**	**66.88**
	sv	**bg**	**da**	**fi**	**sk**	**lt**	**hr**	**sl**	**et**	**lv**			**Avg.**
XLM-R	67.56	66.92	67.43	66.88	66.72	66.12	67.62	67.23	65.93	67.04			66.93
HICTL	68.32	67.24	68.53	67.98	67.63	67.24	68.45	68.37	67.45	68.14			67.98
INFOXLM	68.82	67.88	69.05	68.59	68.54	67.93	68.78	68.57	67.98	68.45			68.50
X-BUFF	**70.41**	**69.41**	**71.37**	**69.98**	**70.47**	**71.35**	**70.86**	**69.76**	**70.23**	**69.82**			**70.12**
WMT21 QE Task 1 (Pearson)													
Model\Language	**en-de**	**en-zh**	**et-en**	**ne-en**	**ro-en**	**ru-en**	**si-en**	**en-cs**	**en-ja**	**km-en**	**ps-en**		**Avg.**
XLM-R	41.23	56.63	79.77	81.21	89.15	77.49	57.83	54.71	33.55	61.23	63.54		63.30
HICTL	49.58	57.92	79.22	83.51	**90.48**	78.68	57.46	55.57	34.18	62.54	64.79		64.90
INFOXLM	51.67	53.45	77.47	83.41	88.95	78.81	58.08	56.44	32.54	63.48	61.66		64.18
X-BUFF	**58.04**	**58.43**	**80.23**	**85.12**	89.67	**82.86**	**59.23**	**56.74**	**36.44**	**64.01**	**65.13**		**66.90**

Table 1: Evaluation results on 4 cross-lingual tasks: XNLI [17], AmericasNLI [19], MultiEURLEX [10] and WMT21 QE Task 1 [43]. We highlight long-tail languages with the color red, and the best results are marked with bold font. All results are from our re-implementation of previous methods with the same model size and training corpora as X-BUFF.

fixed, and only parameters $\Theta_{\mathcal{G}}$ are updated as follows:

$$\mathcal{L}(\Theta_{\mathcal{G}}) = -\sum_{c=1}^{|\mathcal{C}^t_{\langle \mathbf{x},\mathbf{k}\rangle}|} \Big[\log \mathcal{G}(\mathbf{x}_c, \mathbf{k}_c) + \mu \sum_{\mathbf{k}_q \in \mathcal{Q}^{t-1}_{\langle \mathbf{x},\mathbf{k}\rangle}} \log\left(1 - \mathcal{G}(\mathbf{x}_c, \mathbf{k}_q)\right)\Big] \tag{9}$$

where μ is the coefficient of the negative example loss. The actual numbers of positive samples and negative samples may vary a lot, so we give the negative part a small weight to balance the numerical size.

2.3 X-BUFF *Algorithm*

Algorithm 1 describes the X-BUFF process for long-tail language $\mathbf{x}$. After X-BUFF convergence on language $\mathbf{x}$, we repeat the same process on other languages until convergence on all long-tail languages.

3 Experiments

To comprehensively evaluate the cross-lingual transferability of X-BUFF, we conduct experiments on 5 cross-lingual benchmarks with more than 50 languages. In this section, we first introduce the training configuration and then provide detailed evaluation results.

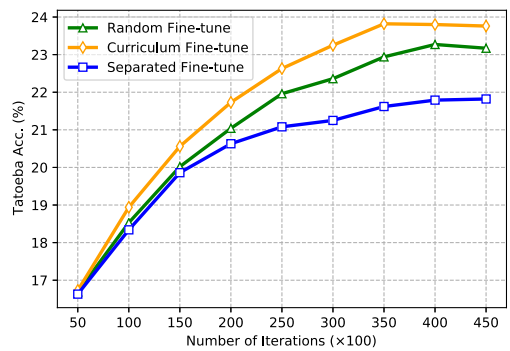

Figure 5: Analysis of different fine-tuning policies. We report the zero-shot averaged accuracy on all 112 languages in Tatoeba benchmark at different iterations. "Random Fine-tune" denotes we randomly select N sentence pairs to fine-tune PLM, while "Separated Fine-tune" means we first use a frozen cross-lingual PLM to construct all parallel corpora, then use constructed parallel corpora to fine-tune encoder. "Curriculum Fine-tune" achieves better accuracy and even convergences faster than the others.

3.1 *Setup*

3.1.1 *Corpus Segmentation*

In X-BUFF, we build anchor language corpus $\mathcal{D}_{\mathbf{k}}$ with sentences in the top 25 languages [1] with the richest resources covered by MultiUN [54], CCNet-100 [15], CCAligned [20], CCMatrix [41] and WMT parallel

[1] Anchor languages in X-BUFF: ar, bg, cs, de, es, el, fr, id, hu, fi, it, ja, ko, nl, pl, pt, ro, ru, sv, tr, zh, he, ca, et, vi, en

Model\Language	af	ar	bg	bn	de	el	es	et	eu	fa	fi	fr	he	hi	hu	id	it	ja
XLM-R	59.4	50.6	72.0	45.6	89.3	61.2	77.6	51.8	38.5	71.4	72.9	76.7	66.2	73.0	65.3	77.7	68.1	62.9
HICTL	63.1	51.1	76.5	46.7	**94.9**	68.2	80.5	59.7	41.8	77.1	78.4	80.2	70.1	78.3	71.8	81.1	73.9	66.1
INFOXLM	66.3	54.2	79.7	49.6	93.0	70.3	85.2	62.4	44.8	78.3	81.8	81.2	74.1	82.9	76.9	**86.9**	76.7	69.1
X-BUFF	**69.0**	**60.3**	**83.1**	**62.4**‡	92.3	**72.5**	**85.3**	**67.0**	**56.2**‡	**78.6**	**83.7**	**83.5**	**75.7**	**83.2**	**80.6**	85.1	**83.4**	**73.2**
	jv	**ka**	**kk**	**ko**	**ml**	**mr**	**nl**	**pt**	**ru**	**sw**	**ta**	**te**	**th**	**tl**	**tr**	**ur**	**vi**	**zh**
XLM-R	15.5	53.2	51.4	62.7	66.3	59.4	81.2	84.5	77.1	19.5	28.1	37.9	28.7	36.8	69.1	26.4	78.1	69.6
HICTL	19.5	57.1	54.7	67.8	71.9	62.2	87.7	89.0	77.5	26.3	33.1	39.4	33.2	43.1	71.0	27.7	80.2	74.5
INFOXLM	18.1	61.7	57.2	70.4	**74.2**	65.5	**91.8**	91.6	81.1	27.2	37.8	42.2	36.6	48.0	74.8	32.2	**82.5**	78.3
X-BUFF	**23.4**‡	**65.4**‡	**64.7**‡	**73.2**	73.5	**66.4**	89.3	**92.4**	**83.1**	**51.0**‡	**47.5**‡	**50.1**‡	**52.1**‡	**60.3**‡	**76.3**	**48.6**‡	80.2	**84.3**

Table 2: Evaluation results on 36 language pairs of the Tatoeba sentence retrieval tasks [4]. The best results are marked with bold font. We highlight greatly improved results with ‡ and long-tail languages with the color red.

data [3]. As for long-tail language set $\mathcal{T}$, we choose languages that are in CCNet-100 but not covered in anchor language corpus $\mathcal{D}_{\mathbf{k}}$ and take the corresponding large-scale monolingual corpus in CCNet-100 as long-tail monolingual corpus $\mathcal{D}_{\mathbf{x}}$. To improve the retrieval accuracy and efficiency, we divide anchor language corpus into different shards according to domains or releasing date (otherwise randomly dividing shards). Shards typically contain about 250k entries.

3.1.2 Model Settings

We implement the cross-lingual PLM $\mathcal{E}$ with XLM-R large model [15]. As for semantic relation classifier, we take one fully-connected layer, with the size of weight matrix $\mathbf{w}$ and bias b being $\mathbb{R}^{1024}$. We first warm up two modules on anchor language corpus through Eq. 8 and Eq. 9 on 8 V100 GPUs, then start X-BUFF iteration on long-tail languages. We set a batch size of 256 accumulating the gradient of 4 iterations and a learning rate of 4e-5 with a cosine decay learning rate. The Adam optimizer (β_1=0.9, β_2=0.999) [27] is adopted. We choose $N = 200$ sentences and adopt $\delta = 0.8$ and $\mu = 0.3$.

3.1.3 Baselines

To be fair, we re-implement several cross-lingual PLMs with our settings. Specifically, we choose: (1) XLM-R [16], which applies multilingual MLM tasks on a large CCNet-100 corpus; (2) HICTL [47], which continues training on XLM-R using hierarchical contrastive learning; (3) INFOXLM [11], which is initialized with XLM-R and trains with cross-lingual contrast, multilingual MLM and TLM.

3.2 Experimental Analysis

3.2.1 Consistent improvements over downstream tasks

We test X-BUFF on five different cross-lingual tasks, and results are shown in Table 1 and 2. In general, these five tasks can be divided into two main categories: (1) classification-based cross-lingual tasks: XNLI, AmericasNLI, and MultiEURLEX. X-BUFF achieves 72.28 % accuracy on XNLI & AmericasNLI task, outperforming all previous baselines with up to 4.9 % improvements. Further test on the legal-topic classification task, MultiEURLEX, shows a 70.12% top-1 accuracy, outperforming XLM-R, HICTL and INFOXLM by a 4.8%, 3.2% and 2.4%, respectively. (2) retrieval-based cross-lingual tasks: WMT21 QE Task 1 and Tatoeba. X-BUFF achieves a 66.90 correlation score on WMT21 QE Task 1, outperforming several strong baselines by 3.6%∼7.1%. Similar results can be found on Tatoeba. As illustrated in Table 2, X-BUFF consistently outperforms previous strong baselines by 11.6%∼28.4% on average. Consistent improvements over different types of tasks demonstrate the robustness of X-BUFF and further reveal that X-BUFF can capture and align essential semantics, which is useful for different tasks across many languages.

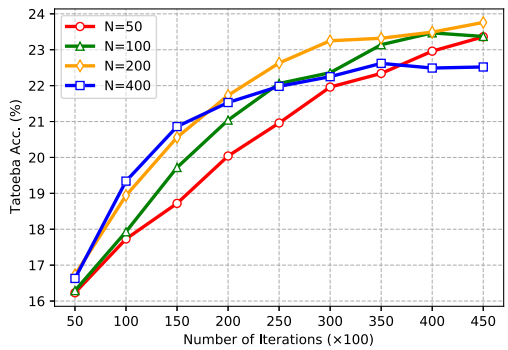

Figure 6: Effects of Curriculum Difficulty N. We report the zero-shot averaged accuracy on all 112 languages in Tatoeba benchmark at different iterations, where we directly evaluate PLM on Tatoeba test set. $N = 200$ achieves the best retrieval accuracy.

3.2.2 Performance on long-tail language representation

To thoroughly analyze the effectiveness of X-BUFF on mitigating the collapse of long-tail language representations in cross-lingual PLMs, we highlight long-tail languages with color red in Table 1 and 2. These five tasks contain 37 different long-tail languages in total. Compared the performance of long-tail languages with other languages, we can observe that all models show performance discrepancies. INFOXLM demonstrates an absolute 34.2% discrepancy, while this statistic further rises to 35.4% and 37.2% on HICTL and XLM-R, respectively. X-BUFF shows the minimum performance discrepancy with 25.8%. On the performance of long-tail languages, X-BUFF significantly outperforms XLM-R, HICTL and INFOXLM by 21.0%, 13.7%, and 10.6%, respectively, on average among all long-tail languages. What's more, as shown in Table 2, X-BUFF substantially improves top-1 accuracy by 10%-32% on multiple long-tail languages. This confirms that X-BUFF can construct reliable parallel sentence pairs and provide sufficient semantic relation signals with only monolingual corpora. X-BUFF can mitigate the collapse of long-tail language representations.

3.2.3 Analysis on the unified boot-strapped framework

To examine the importance of the unified boot-strapped framework of X-BUFF, we conduct a "Separated Fine-tune" experiment on Tatoeba, as shown in Figure 5. In "Separated Fine-tune" setting, we separate the construction of parallel corpora and fine-tuning of PLMs. We first construct parallel corpora for long-tail languages with a frozen PLM, and then fine-tune the encoder with the constructed parallel corpora. The great improvement of "Curriculum Fine-tune" compared to "Separated Fine-tune" shows that these two progress can reciprocate with each other. We further observe that the scale of bootstrapped parallel queue is about 20% larger in "Curriculum Fine-tune" than in "Sep-

arated Fine-tune" setting, which indicates that an updated PLM can benefit the construction of parallel corpus, and a more confidential and larger parallel corpus can, in return, provide sufficient semantic alignment signals to fine-tune PLMs.

3.2.4 Analysis of curriculum fine-tuning

We introduce a variant of curriculum fine-tuning paradigm to investigate the performance of curriculum fine-tuning on Tatoeba benchmark. Specifically, instead of gradually selecting sentences with complementary semantics (step 1), we randomly select N sentences from monolingual corpus $\mathcal{D}_{\mathbf{x}}$, pair them to parallel sentence pairs, and use them to fine-tune the PLM. The results are reported as "Random Fine-tune" in Figure 5. From the results, we can conclude that training samples organized in an increasing-difficulty order can provide performance improvements. Moreover, our curriculum fine-tuning reaches the same performance 30% faster than "Random Fine-tune" and reaches the peak performance quicker, which demonstrates curriculum fine-tuning design can faster adapt to new long-tail languages.

3.3 Parameter Analysis

3.3.1 Impact of curriculum difficulty N

N impacts the curriculum difficulty of bootstrapped parallel queue $\mathcal{Q}_{\langle \mathbf{x},\mathbf{k} \rangle}$ by controlling the selection of new sentences to complement parallel corpus. Figure 6 illustrates how the hyper-parameter N affects the fine-tuning performance. We depict the accuracy curves with different values N on Tatoeba benchmark. From Figure 6, we can observe that gradually increasing N significantly improves the training efficiency, while only slightly influencing the accuracy, as in the comparison between $N = 50$, $N = 100$, and $N = 200$. However, assigning larger values to N (e.g. 400) does not further speed up convergence and also hinders model performance. We conjecture that (1) a small number of N will lead to a slow expansion of bootstrapped parallel queue, which severely drags down the training efficiency of the cross-lingual encoder. (2) when N reaches 400, the expansion speed of the data scale is so fast that our modules are not competent enough to construct parallel sentence pairs in high quality. This further leads to a noisy corpus and in return degrade the model performance. Heuristically, we set $N = 200$ to achieve a balance between the expansion speed of bootstrapped parallel queue and the competence of the cross-lingual encoder.

Confidence threshold δ	0.5	0.6	0.7	0.8	0.9
Tatoeba (Acc.)	20.47	22.89	23.48	**23.76**	23.64
balance coefficient μ	0.1	0.2	0.3	0.4	0.5
Tatoeba (Acc.)	22.78	23.39	**23.76**	23.67	23.46

Table 3: Effects of δ and μ on Tatoeba benchmark. We report the zero-shot accuracy on Tatoeba with different values of δ and μ.

3.3.2 Impact of δ and μ

Table 3 shows how the hyper-parameter δ and μ affect the performance on Tatoeba benchmark. δ controls how strictly we select semantic equivalent sentences. The setting of $\delta = 0.8$ achieves the best results. μ balances the importance of positive samples and negative samples at curriculum fine-tuning for semantic relation classifier. From Table 3, we can find that $\mu = 0.3$ achieves the best performance.

4 Related Works

Substantial works [4, 13, 21] have shown that cross-lingual PLMs have achieved promising cross-lingual transferability. An intuitive way to train such a cross-lingual PLM is to shift the training objective from monolingual PLMs to a multilingual scenario, such as mBERT [18], XLM-R [15], mBART [33], mT5 [51], etc. Besides token-level pre-training objectives, researchers have designed explicit sentence-level pre-training objectives to align cross-lingual sentences and learn language-agnostic representations. We categorize the sentence-level pre-training objectives into two kinds: supervised sentence-level cross-lingual pre-training and unsupervised sentence-level cross-lingual pre-training.

4.1 Supervised Sentence-level Cross-lingual Pre-training

The most widely-used alignment supervision is the large parallel corpora. Researches, such as XLM [16], Unicoder [26] and VECO [34], have applied language modeling on parallel corpora. HiCTL [48], INFOXLM [11], AMBER [25] and dual momentum contrast[46] aggregate contrastive learning for cross-lingual pre-training. Another line of works [40, 45] distills a language-agnostic representation from original monolingual PLMs or Cross-lingual PLMs, respectively. S^2DM [50] disassociates semantics from syntax in representations from cross-lingual PLMs, and XLM-E [12] jointly trains a generator and a discriminator from scratch.

4.2 Unsupervised Sentence-level Cross-lingual Pre-training

Without sentence relation supervision signal, Researches have tried various methods to train cross-lingual representation. Some [28, 36] produced a synthetic parallel corpus with unsupervised machine translation model and then derived cross-lingual Pre-training with synthetic parallel corpus. Similarly, ERNIE-M [36] proposed a back-translation masked language modeling, which first generated pseudo-parallel sentences and later applied to cross-lingual pre-training. Instead of translation methods, DuEAM [24] adopted a frozen anchor module to compute the semantic similarity, and used a learner module to approximate similarity from the anchor, while MARGE [31] self-supervised the reconstruction of target text by retrieving a set of related texts and conditioned on them to generate the original.

5 Conclusion

We propose a novel Cross-lingual Bootstrapped Unsupervised Fine-tuning Framework (X-BUFF) to mitigate the representation collapsing problem of long-tail languages through fine-tuning the representations of long-tail languages. X-BUFF iteratively fine-tunes a cross-lingual PLM with three steps: (1) Select sentences whose semantics complement the current parallel queue. (2) Match selected sentences to parallel sentence pairs. (3) Fine-tune the PLM in a curriculum way. Further experiments on 5 cross-lingual downstream tasks show that X-BUFF can mitigate long-tail representation collapsing problem and achieve significant improvements over previous baselines.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No.U21B2009). This research is also supported by the

Strategic Priority Research Program of Chinese Academy of Science, Grant No.XDC02030400.

References

[1] Eugene Agichtein and Luis Gravano, 'Snowball: Extracting relations from large plain-text collections', in *Proceedings of the Fifth ACM Conference on Digital Libraries*, (2000).

[2] Roee Aharoni, Melvin Johnson, and Orhan Firat, 'Massively multilingual neural machine translation', in *Proc. of NAACL*, (2019).

[3] Farhad Akhbardeh, Arkady Arkhangorodsky, and Magdalena*, et al* Biesialska, 'Findings of WMT21', in *Proceedings of the Sixth Conference on Machine Translation*, (2021).

[4] Mikel Artetxe and Holger Schwenk, 'Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond', *Transactions of the Association for Computational Linguistics*, (2019).

[5] Ankur Bapna and Isaac Caswell*, et al*, 'Building machine translation systems for the next thousand languages', Technical report, (2022).

[6] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston, 'Curriculum learning', in *Proc. of ICML*, (2009).

[7] Samuel R. Bowman and Gabor*, et al* Angeli, 'A large annotated corpus for learning natural language inference', in *Proc. of EMNLP*, (2015).

[8] Sergey Brin, 'Extracting patterns and relations from the world wide web', in *The World Wide Web and Databases*, (1999).

[9] Tom B. Brown, Benjamin Mann, and Nick Ryder*, et al*. Language models are few-shot learners, 2020.

[10] Ilias Chalkidis and Manos*, et al* Fergadiotis, 'MultiEURLEX - a multilingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer', in *Proc. of EMNLP*, (2021).

[11] Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou, 'InfoXLM: An information-theoretic framework for cross-lingual language model pre-training', in *Proc. of NAACL*, (2021).

[12] Zewen Chi, Shaohan Huang, Li Dong, Shuming Ma, Bo Zheng, Saksham Singhal, Payal Bajaj, Xia Song, Xian-Ling Mao, Heyan Huang, and Furu Wei, 'XLM-E: Cross-lingual language model pre-training via ELECTRA', in *Proc. of ACL*, (2022).

[13] Muthu Chidambaram, Yinfei Yang, and Daniel*, et al* Cer, 'Learning cross-lingual sentence representations via a multi-task dual-encoder model', in *Proc. of RepL4NLP*, (2019).

[14] Aakanksha Chowdhery, Sharan Narang, and Jacob Devlin*, et al*. Palm: Scaling language modeling with pathways, 2022.

[15] Alexis Conneau and Kartikay*, et al* Khandelwal, 'Unsupervised cross-lingual representation learning at scale', in *Proc. of ACL*, (2020).

[16] Alexis Conneau and Guillaume Lample, 'Cross-lingual language model pretraining', in *Proc. of NeurIPS*, (2019).

[17] Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, and Samuel*, et al* Bowman, 'XNLI: Evaluating cross-lingual sentence representations', in *Proc. of EMNLP*, (2018).

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, 'BERT: Pre-training of deep bidirectional transformers for language understanding', in *Proc. of NAACL*, (2019).

[19] Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, and Luis*, et al* Chiruzzo, 'AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages', in *Proc. of ACL*, (2022).

[20] Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn, 'CCAligned: A massive collection of cross-lingual web-document pairs', in *Proc. of EMNLP*, (2020).

[21] Fangxiaoyu Feng, Yinfei Yang, and Daniel*, et al* Cer, 'Language-agnostic BERT sentence embedding', in *Proc. of ACL*, (2022).

[22] Tianyu Gao, Xu Han, and Ruobing*, et al* Xie, 'Neural snowball for few-shot relation learning', *Proc. of AAAI*, (2020).

[23] Tianyu Gao, Xingcheng Yao, and Danqi Chen, 'SimCSE: Simple contrastive learning of sentence embeddings', in *Proc. of EMNLP*, (2021).

[24] Koustava Goswami and Sourav*., et al* Dutta, 'Cross-lingual sentence embedding using multi-task learning', in *Proc. of EMNLP*, (2021).

[25] Junjie Hu and Melvin *, et al* Johnson, 'Explicit alignment objectives for multilingual bidirectional encoders', in *Proc. of NAACL*, (2021).

[26] Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, and Linjun*, et al* Shou, 'Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks', in *Proc. of EMNLP*, (2019).

[27] Diederik P. Kingma and Jimmy Ba, 'Adam: A method for stochastic optimization', in *Proc. of ICLR*, (2015).

[28] Ivana Kvapilíková, Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Ondřej Bojar, 'Unsupervised multilingual sentence embeddings for parallel corpus mining', in *Proc. of ACL*, (2020).

[29] Viet Dac Lai*, et al*. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning, 2023.

[30] Seonghyeon Lee, Dongha Lee, Seongbo Jang, and Hwanjo Yu, 'Toward interpretable semantic textual similarity via optimal transport-based contrastive sentence learning', in *Proc. of ACL*, (2022).

[31] Mike Lewis and Marjan*, et al* Ghazvininejad, 'Pre-training via paraphrasing', in *Proc. of NeurIPS*, (2020).

[32] Xuebo Liu and Houtim*, et al* Lai, 'Norm-based curriculum learning for neural machine translation', in *Proc. of ACL*, (2020).

[33] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer, 'Multilingual denoising pre-training for neural machine translation', *Transactions of the Association for Computational Linguistics*, (2020).

[34] Fuli Luo, Wei Wang, Jiahao Liu, Yijia Liu, and Bin*, et al* Bi, 'VECO: Variable and flexible cross-lingual pre-training for language understanding and generation', in *Proc. of ACL*, (2021).

[35] Aäron van den Oord and Yazhe*, et al* Li, 'Representation learning with contrastive predictive coding', *CoRR*, **abs/1807.03748**, (2018).

[36] Xuan Ouyang and Shuohuan*, et al* Wang, 'ERNIE-M: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora', in *Proc. of EMNLP*, (2021).

[37] Wenbo Pan and Qiguang Chen*, et al*. A preliminary evaluation of chatgpt for zero-shot dialogue understanding, 2023.

[38] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer, 'Deep contextualized word representations', in *Proc. of NAACL*, (2018).

[39] Alec Radford and Karthik*, et al* Narasimhan, 'Improving language understanding by generative pre-training', *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*, (2018).

[40] Nils*, et al* Reimers, 'Making monolingual sentence embeddings multilingual using knowledge distillation', in *Proc. of EMNLP*, (2020).

[41] Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan, 'CCMatrix: Mining billions of high-quality parallel sentences on the web', in *Proc. of ACL*, (2021).

[42] Richard Socher, Alex Perelygin, Jean Wu, and Jason*, et al* Chuang, 'Recursive deep models for semantic compositionality over a sentiment treebank', in *Proc. of EMNLP*, (2013).

[43] Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins, 'Findings of the WMT 2021 shared task on quality estimation', in *Proceedings of the Sixth Conference on Machine Translation*, (2021).

[44] Romal Thoppilan and Daniel De Freitas*, et al*. Lamda: Language models for dialog applications, 2022.

[45] Nattapong Tiyajamorn, Tomoyuki Kajiwara, and Yuki*, et al* Arase, 'Language-agnostic representation from multilingual sentence encoders for cross-lingual similarity estimation', in *Proc. of EMNLP*, (2021).

[46] Liang Wang and Wei*, et al* Zhao, 'Aligning cross-lingual sentence representations with dual momentum contrast', in *Proc. of EMNLP*, (2021).

[47] Xiangpeng Wei and Rongxiang Weng*, et al*, 'On learning universal representations across languages', in *Proc. of ICLR*, (2021).

[48] Xiangpeng Wei, Heng Yu, Yue Hu, Rongxiang Weng, Weihua Luo, and Rong Jin, 'Learning to generalize to more: Continuous semantic augmentation for neural machine translation', in *Proc. of ACL*, (2022).

[49] Adina Williams, Nikita Nangia, and Samuel Bowman, 'A broad-coverage challenge corpus for sentence understanding through inference', in *Proc. of NAACL*, (2018).

[50] Linjuan Wu, Shaojuan Wu, Xiaowang Zhang, Deyi Xiong, Shizhan Chen, Zhiqiang Zhuang, and Zhiyong Feng, 'Learning disentangled semantic representations for zero-shot cross-lingual transfer in multilingual machine reading comprehension', in *Proc. of ACL*, (2022).

[51] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, and Aditya*, et al* Siddhant, 'mT5: A massively multilingual pre-trained text-to-text transformer', in *Proc. of NAACL*, (2021).

[52] Yuhao Zhang and Hongji*, et al* Zhu, 'A contrastive framework for learning sentence representations from pairwise and triple-wise perspective in angular space', in *Proc. of ACL*, (2022).

[53] Kun Zhou and Beichen*, et al* Zhang, 'Debiased contrastive learning of unsupervised sentence representations', in *Proc. of ACL*, (2022).

[54] Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen, 'The United Nations parallel corpus v1.0', in *Proc. of LREC*, (2016).