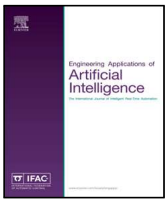
Engineering Applications of Artificial Intelligence 139 (2025) 109532

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Research paper

Code-switching finetuning: Bridging multilingual pretrained language models for enhanced cross-lingual performance

Changtong Zan [a], Liang Ding [b,*], Li Shen [c], Yu Cao [d], Weifeng Liu [a,*]

[a] College of Control Science and Engineering, China University of Petroleum (East China), Qingdao 266580, China
[b] JD Explore Academy at JD.com, Beijing 100101, China
[c] School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University, Shenzhen 518107, China
[d] Tencent IEG, Shenzhen 518000, China

ARTICLE INFO

Keywords:
Multilingual models
Sequence-to-sequence pretrained language models
Cross-lingual gap
Code-switching restore task

ABSTRACT

In recent years, the development of pre-trained models has significantly propelled advancements in natural language processing. However, multilingual sequence-to-sequence pretrained language models (Seq2Seq PLMs) are pretrained on a wide range of languages (e.g., 25 languages), yet often finetuned for specific bilingual tasks (e.g., English–German), leading to domain and task discrepancies between pretraining and finetuning stages, which may lead to sub-optimal downstream performance. In this study, we first illustratively reveal such domain and task discrepancies, and then conduct an in-depth investigation into the side effects that these discrepancies may have on both training dynamic and downstream performance. To alleviate those side effects, we introduce a simple and effective code-switching restoration task (namely **code-switching finetuning**) into the standard pretrain-finetune pipeline. Specifically, in the first stage, we recast the downstream data as the self-supervised format used for pretraining, in which the denoising signal is the code-switched cross-lingual phrase. Then, the model is finetuned on downstream task as usual in the second stage. Experiments spanning both natural language generation (12 supervised translations, 30 zero-shot translations, and 2 cross-lingual summarization tasks) and understanding (7 cross-lingual natural language inference tasks) tasks demonstrate that our model consistently and significantly surpasses the standard finetuning strategy. Analyses show that our method introduces negligible computational cost and reduces cross-lingual representation gaps. We have made the code publicly available at: https://github.com/zanchangtong/CSR4mBART.

1. Introduction

Deep learning (LeCun et al., 2015), a branch of artificial intelligence, leverages neural networks to learn from vast amounts of data has rapidly evolved in various fields (Akkem et al., 2023b,a, 2024). It shows remarkable success in various Natural Language Processing (NLP) tasks, such as machine translation (Cho et al., 2014), abstractive text summarization (Nallapati et al., 2016) and natural language inference (Wang and Jiang, 2016). etc. However, achieving optimal performance requires a substantial amount of supervised data. To overcome this limitation, pretrain-finetune paradigm (Devlin et al., 2019) has emerged as an effective solution, reducing the reliance on supervised data. In the field of NLP, pretrain-finetune paradigm has achieved tremendous achievements by transferring knowledge from a large scale of unlabeled data to the parameters of the pretrained models (Devlin et al., 2019; Liu et al., 2019b; Conneau and Lample, 2019; Brown et al., 2020). Typically, the model, namely pretrained language models (PLMs), pretrains on large-scale monolingual text data with self-supervised objectives, such as mask language model (Devlin et al., 2019), casual language model (Brown et al., 2020), denoising (Lewis et al., 2020) and others. The well-trained PLMs are then finetuned on supervised data to acquire ability of the certain task.

Inspired by the success of the pretrain-finetune paradigm, recent works (Song et al., 2019; Liu et al., 2020) have attempted to design a unified multilingual sequence-to-sequence pretrained language model (multilingual Seq2Seq PLMs) for cross-lingual downstream tasks, including machine translation (Vaswani et al., 2017), cross-lingual summarization (Zhu et al., 2019), and cross-lingual language understanding (Xue et al., 2021). Despite successfully taking the first step, the multilingual Seq2Seq PLMs encounter many challenges on cross-lingual NLU and NLG tasks (Xue et al., 2021). For example, with covering more language pairs, i.e. from 25 to 50, as in mBART50 (Tang et al., 2020) does not show any promising improvements (or even worse on 8 out of 24 languages) in given bilingual downstream tasks as reported by Tang et al. (2020).

* Corresponding authors.
E-mail addresses: liangding.liam@gmail.com (L. Ding), liuwf@upc.edu.cn (W. Liu).

https://doi.org/10.1016/j.engappai.2024.109532
Received 9 April 2024; Received in revised form 21 September 2024; Accepted 21 October 2024
Available online 7 November 2024

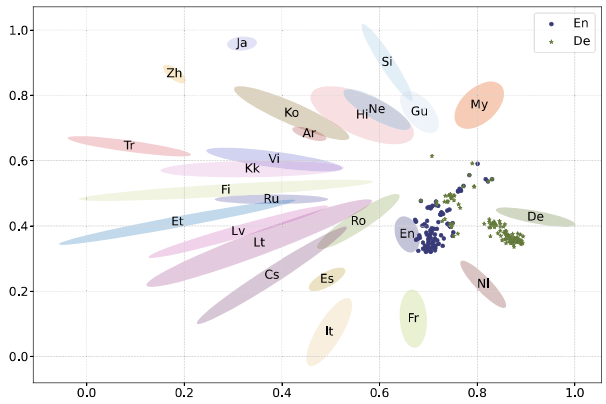


Fig. 1. The distribution of pretraining and translation data.

Table 1
Comparison of learning objectives between multilingual Seq2Seq PLMs pretraining and finetuning. Underline denotes placeholder symbols.

Multilingual Seq2Seq PLMs: $-\log P(\mathbf{x}\|\tilde{\mathbf{x}})$	
Source	Military _ Field Marshal Hussein _ in attendance. [EN]
Target	Military Field Marshal Hussein Tantawi was in attendance [EN]
Source	发言人 _ 一次新闻发布会 [ZH]
Target	发言人每周举行一次新闻发布会 [ZH]
	... ... [...]
Military _ Field Marshal Hussein _ in attendance .	
Military Field Marshal Hussein Tantawi was in attendance .	
Finetuning on translation task: $-\log P(\mathbf{y}\|\mathbf{x})$	
Source	布什与沙龙举行了会谈 [ZH]
Target	Bush held a talk with Sharon. [EN]
布什 与 沙龙 举行 了 会谈 。	
Bush held a talk with Sharon .	

This deficiency may draw from the huge ***cross-lingual gap*** that exists between pretraining (PT) and finetuning (FT). Specifically, we can categorize the cross-lingual gap into ***"Domain discrepancy"*** and "***Task discrepancy***", which are as described by Gururangan et al. (2020). The domain discrepancy draws from the difference in the training dataset, where PT is generally trained on monolingual data of multiple languages from a more general domain, e.g. 25 languages for mBART (Liu et al., 2020), while the FT is mainly based on a bilingual subset, e.g. news domain English→German; As illustrated in Fig. 1, we demonstrate the data distribution by employing XLM-R (Conneau et al., 2020a) to extract features of the sampled sentences, which are subsequently embedded into a two-dimensional space via t-SNE (Van der Maaten and Hinton, 2008). We use ellipses to present the pretraining data domain for each language and dots to represent samples from the downstream English–German translation task. Each language is identified using the ISO 639-1 code. It was observed that the representation of the upstream English–German data exhibited a more pronounced degree of separation and occupied a distinct distribution space when compared to the downstream data. Furthermore, a substantial separation is noted between the pretraining data of different languages, such as Chinese (Zh) and Japanese (Ja), and the translation data. This phenomenon emphasizes the domain discrepancy that complicates knowledge transfer in both pretrain-to-finetune and cross-lingual contexts.

The task discrepancy draws from the difference of objective, where the task of PT is to utilize the monolingual data under a self-supervised pattern, such as denoising task for mBART, while the FT aims to generate the target-side language conditioned on the source-side language in a supervised fashion. Table 1 illustrates the detailed learning objective difference between pertaining and finetuning of Multilingual Seq2Seq PLMs. It is pretrained to predict in placeholder symbols position to capture the knowledge of monolingual data,[1] e.g. English "[EN]", Chinese "[ZH]". However, it finetunes on the cross-lingual task, e.g. translation, highly relying on cross-lingual alignment information to accomplish the generation. We also provide word alignment information by linking equivalent words with lines. It demonstrates that the pretraining task incorporates both direct copying (e.g., "Military", "Field", etc.) and the prediction of perturbed text (e.g., "Tantawi was"), reliant on monolingual knowledge. In contrast, translation predicts the target language representation based on the sentence of the source language, requiring more complex cross-lingual alignment information.

To further explore the implications of the cross-lingual gap, we delve into the training dynamics and performance on downstream cross-lingual tasks. Drawing inspiration from the work of Fan et al. (2021), our analysis focuses on the sentence-level contextual features in PLMs. Specifically, we utilize the cross-lingual sentence representation distance as a metric for assessing the cross-lingual gap. This measure is derived by averaging the Euclidean distances between the contextual features of sentences in two different languages. A smaller cross-lingual distance indicates greater similarity in how the model encodes both languages, which suggests reduced difficulty in cross-lingual transfer. As depicted in Fig. 2(1), we first examine the training dynamics of two downstream translation tasks. It is observed that the task with a higher initial cross-lingual distance (Ko-En) correlates with lower overall performance. As training progresses, both tasks show increases in SacreBLEU scores, coinciding with a reduction in cross-lingual distance. Notably, both cross-lingual distance and cross-lingual performance reach peak performance at the same time after approximately 10k updates. These observations provide evidence of a correlation between cross-lingual distance and downstream task performance. In Fig. 2(2), our findings indicate a notable reduction in cross-lingual distance for finetuned models when compared to the pretrained model. These results highlight that a significant cross-lingual gap exists between the pretraining and finetuning phases of multilingual Seq2Seq PLMs.

To tackle the cross-lingual gap, in this paper, we propose a two-stage finetuning strategy for multilingual Seq2Seq PLMs, namely code-switching finetuning. Equipped with a code-switching restore task, it can effectively tune models on downstream cross-lingual generation and understanding tasks. Our strategy sufficiently transfers the monolingual knowledge for multilingual Seq2Seq PLMs adapting to downstream cross-lingual tasks, thus improving the sub-optimal performance of finetuning caused by cross-lingual gap, i.e., domain discrepancy and task discrepancy. As shown in Fig. 3, the diagram illustrates the overall structure of our finetune pipeline, alongside the standard finetune pipeline for comparison. The first stage (Section 4.2) is designed to inform the multilingual Seq2Seq PLMs about the downstream cross-lingual knowledge, by training it to denoise the code-switching samples for **mitigating the task discrepancy**. To sufficiently transfer the knowledge from a PLM to the following task, we perform code-switch restoring in both the *source- and target-side sentences* of the downstream cross-lingual task, where some words are replaced with corresponding ones in another language. And we optimize these two tasks simultaneously to ensure the great potential to **mitigate the domain discrepancy**. More specifically, we first derive the unsupervised translation vocabulary with downstream in-domain cross-lingual sentences following Artetxe et al. (2018). Then, we introduce it by code-switching restore task. We perturb the downstream sentences by randomly replacing a certain percentage of words with a semantic

[1] The Chinese example means "The spokesperson holds a press conference once a week".

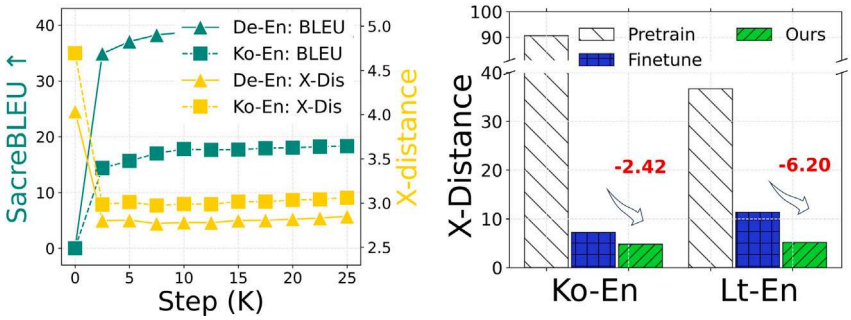


Fig. 2. Impact of pretraining and finetuning on cross-lingual distance and downstream task performance. (1) Dynamics of standard finetuning: impact on performance and cross-lingual distance. (2) Effects of finetuning on cross-lingual distance.

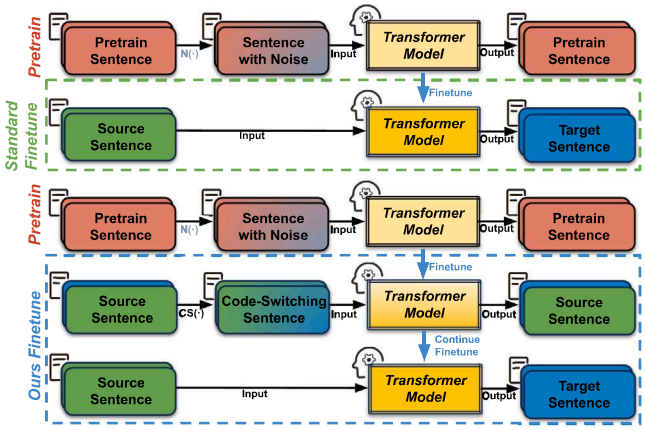


Fig. 3. Overview of the pipelines.

similar word of another downstream language. For parameter estimation, we minimize the cross-entropy between the reconstructed output and the raw sample. To sufficiently transfer the knowledge from a multilingual Seq2Seq PLM to the downstream cross-lingual tasks, we perform code-switching restore tasks in both the source- and target-side languages, and optimize these two tasks alternatively. In the next stage, the standard finetuning is applied in the second step to derive the final model. To the best of our knowledge, we are the first to introduce code-switching into the intermediate finetuning stage for multilingual Seq2Seq PLMs to mitigate the cross-lingual gap. Experimental results on 12 *bilingual translation* tasks, 30 *zero-shot translation* tasks, 2 *cross-lingual summarization* tasks, and 7 *cross-lingual natural language inference* tasks show that our method could consistently outperform the standard finetuning. We also conduct further analyses to provide some insights into our method and investigate how our approach bridges the cross-lingual gap: (1) Our approach could narrow the Euclidean distance of cross-lingual sentence representations, especially for distant languages (e.g. 11.37 vs. 5.17 for Lt→En), confirming our claim; (2) Our approach could improve the model generalization with better low-frequency word translation accuracy; and (3) Our approach only requires trivial computational costs, having great potential to be a universal plug-in strategy for any multilingual Seq2Seq PLMs. For convenience, this paper provides a list of important abbreviations and notations in Table 2. The main contributions of our work are three-fold:

- We propose a two-stage finetuning approach with the code-switching restore (CSR) task for multilingual Seq2Seq PLMs on cross-lingual generation and understanding tasks.
- Compared with the standard pretrain-funetune paradigm, our approach narrows the cross-lingual representation distance, benefiting the model generalization with trivial computational cost.

Table 2
Some important abbreviations and notations.

Abbreviation and notation	Definition
NLP	Natural language processing
NLG	Natural language generation
NLU	Natural language understanding
PT	Pretraining
FT	Finetuning
PLMs	Pretrained language models
Seq2Seq PLMs	Sequence-to-sequence pretrained language models
CSR	Code-switching restore
Language ID	Language identifier token
BART	Pre-training by combining bidirectional and auto-regressive transformers
mBART	Multilingual version of BART
BERT	Bidirectional Encoder Representations from Transformers
mBERT	Multilingual version of BERT
BLEU	The bilingual evaluation understudy
COMET	Crosslingual optimized metric for evaluation of translation
ROUGE	Recall-oriented understudy for gisting evaluation
l	Language ID
S_l	A sample of sentence
X_l, Y_l	Sentences of source and target languages
x_i, y_i	Words/tokens from source and target sentences
$N(*)$	The noising function
D	The downstream corpus
E_x, E_y	Word-level embeddings of the source and target languages
x'_i, y'_i	The selected neighbors for x_i and y_i
X', Y'	Code-switching sentences of source and target languages

- Extensive experiments on 51 cross-language scenarios demonstrate that our proposed approach can effectively improve the performance of multilingual Seq2Seq PLMs.

The subsequent paper is designed as follows. We discuss related works in Section 2. We provide the preliminary in Section 3. How we construct the code-switching restore task-based two-stage finetuning is introduced in Section 4. Experimental results and corresponding analyses are shown in Sections 5 and 6 respectively. Conclusions are described in Section 7. Limitations are discussed in Section 8

2. Related work

Our work is inspired by three lines of research: (i) Seq2Seq PLMs, (ii) code-switching based algorithms for pretrain-finetune paradigm and (iii) improving fine-tuning. For a clearer perspective, we summarize and compare key elements of prior research in Table 3.

2.1. Seq2Seq PLMs

Seq2Seq based neural network models are widely used in various fields due to their effectiveness, such as image forecasting (Lian et al.,

Table 3
Highlights of relevant works.

Ref	Authors	Problem	Methods
Liu et al. (2020)	Liu et al.	Underutilization of self-supervision in machine translation	Pretraining on monolingual data from multiple languages with a denoising objective
Tang et al. (2021)	Tang et al.	Integrating multilingual pretraining with multilingual translation remains unexplored	Continued pretraining followed by multilingual translation finetuning
Yang et al. (2020)	Yang et al.	Lack of cross-lingual alignment and presence of placeholder symbols during pretraining	Pretraining on code-switching samples with self-supervised objective
Li et al. (2022)	Li et al.	Insufficient representational power of pre-trained decoders	Conditional masked language pertaining on decoder
Wang et al. (2022)	Wang et al.	Limited translation quality and over-estimation issues	In-domain continue pretraining followed by input adaptation finetuning

2024), fault prediction (Klaar et al., 2023), and others (Wang et al., 2020; Dai et al., 2023). In NLP, Seq2Seq PLMs have achieved remarkable improvements, where large-scale unsupervised monolingual data is utilized via sequence-to-sequence self-supervised tasks. MASS (Song et al., 2019) takes a sentence with contiguous masked tokens as input and maps it to a sequence consisting of missing tokens. "Text-to-Text Transfer Transformer" (T5) (Raffel et al., 2020) proposes a unified text-to-text format to support a wide variety of English-language NLP tasks. And, mT5 (Xue et al., 2021) further extends T5 to 101 languages, verifying the universality of unified text-to-text format. BART (Lewis et al., 2020) analyzes the effect of different noise functions, shows remarkable performance on downstream monolingual generation tasks. mBART (Liu et al., 2020) is a multilingual extension of BART that is pretrained on multilingual data using a self-supervised denoising objective to improve the performance of finetuned translation models. While mBART primarily focuses on downstream bilingual translation tasks, mBART50 (Tang et al., 2021) expands this to include 50 languages by continuing pretrain on 50 languages dataset and analyzing the effects of combining this with multilingual translation finetuning.

However, the above pretraining tasks cannot offer models with cross-lingual alignment information. Therefore, bridging the cross-lingual gap for multilingual Seq2Seq PLMs between upstream and downstream tasks is worth investigating.

2.2. Code-switching based algorithms for pretrain-finetune paradigm

The basic idea of our work to address the cross-lingual gap between Seq2Seq multilingual PLMs and the finetuning process is the code-switching restore task, in which we train multilingual PLMs to predict the original sentence according to the code-switching one during finetuning. Several works also proposed to investigate the benefits of code-switching in the pretrain-finetune paradigm. Generally, multilingual pretraining involves using monolingual data from various languages to train models to recover sentences corrupted with placeholders. CSP (Yang et al., 2020) addresses the lack of cross-lingual alignment information in this paradigm by introducing code-switching sentences during pretraining, thereby avoiding placeholders and enhancing cross-lingual alignment. They employ a word embedding model to build an unsupervised lexicon, replacing placeholders like "[MASK]" with real synonyms from different languages. CeMAT (Li et al., 2022) examines the issue where initializing a transformer decoder with pretrained GPT (Brown et al., 2020) parameters results in poorer translation performance compared to the random initialization. They argue that the pretrained decoder lacks sufficient representational power and propose joint training of masked language modeling (MLM) on the encoder and conditional MLM on the decoder using both monolingual and parallel sentences. In this setup, both the encoder and decoder inputs include words replaced by semantically similar terms or "[MASK]", and the model is tasked with predicting these masked tokens on both the source and target sides, leading to significant performance improvements. DICT-MLM (Chaudhary et al., 2020) focuses on improving the bidirectional transformers pretraining with MLM tasks. They use ground truth bilingual dictionaries to construct code-switching samples and train the model to predict both cross-lingual synonyms and the "[MASK]". However, they evaluated cross-lingual understanding tasks rather than machine translation. Liu et al. (2021b) also suggest continuing pretrain on code-switching data to facilitate the introduction of new languages into existing models. Code-switching is also used in pretrained models of multilingual-multimodal scenarios. M3P (Ni et al., 2021) introduces code-switching to effectively learn a universal representation and benefits the non-English task performance.

Differing from the above methods most focus on pretraining stage and require substantial additional data, our approach focuses on the intermediate finetuning stage of multilingual Seq2Seq PLMs. We use code-switching samples converted from downstream datasets only for the first-stage finetuning with the denoising task, providing the standard finetuning with a better in-domain initialization with trivial computational cost.

2.3. Improving finetuning

Recent works have shown gains compared to directly optimizing the downstream objective, based on modifications of finetuning in *task*, *model*, and *vocabulary* levels.

For task-level, new tasks are involved during finetuning (Raffel et al., 2020; Daumé, 2007; Khashabi et al., 2020). MTDNN (Liu et al., 2019a) proves the efficiency of multi-task learning on top of the pretrained language models and evaluates several NLU benchmarks. Nevertheless, no cross-lingual scenario is considered. Muppet (Aghajanyan et al., 2021a) scales up the type and number of intermediate tasks, which further verifies that multi-task learning could consistently bring benefits to various NLU tasks. Besides, several works also concentrate on adding additional regularization terms. Inspired by trust-region theory, R3F (Aghajanyan et al., 2021b) proposes to map samples in similar embeddings with small perturbations. Recadam (Chen et al., 2020) aims to recall the pretraining task by adding the difference of weights between the finetuned model and the original pretrained model to the loss function. Wang et al. (2022) study the impact of jointly pretrained decoder and point out the limited translation quality and over-estimation issues. In response, they suggest training multilingual Seq2Seq PLMs with in-domain data and input adaptation finetuning to enhance performance and robustness in machine translation tasks.

Model-level works mainly focus on the issue of over-parameterization, where only part of pretrained parameters are initialized, or some components beneficial for downstream tasks are added (Bowman et al., 2015; Houlsby et al., 2019; Liu et al., 2021d) Mixout (Lee et al., 2020) improves the model performance on the GLUE benchmark by randomly initializing part of parameters with pretrained weights. In addition, CHILD-TUNING (Xu et al., 2021) gets a better ability of natural language understanding by masking the gradient of non-child networks during the backward process. LNA (Li et al., 2021) can also get the SOTA performance on two large-scale multilingual speech translation benchmarks, where only normalization layers and attention modules are trained. Besides finetuning PLMs with part of

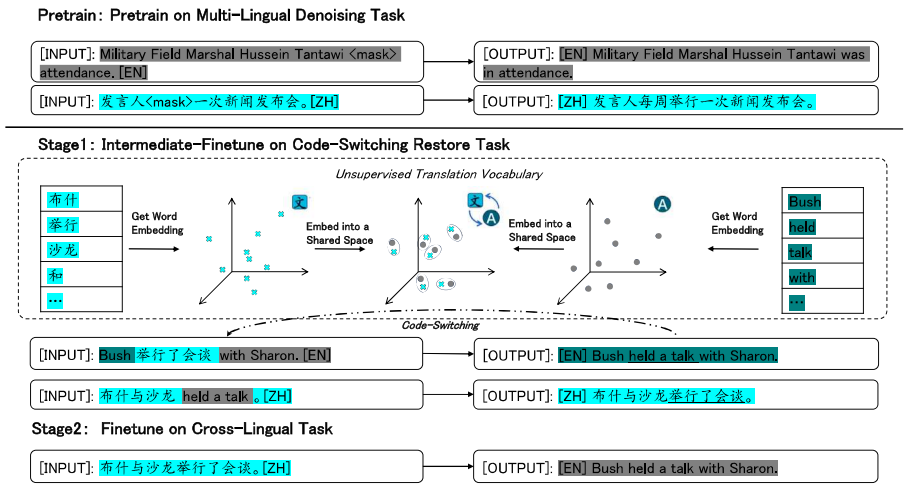


Fig. 4. The schematic of *pretraining stage* (top) of Multilingual Seq2Seq PLMs, *standard finetuning stage* (bottom), and our proposed *code-switching restore task* (middle).

parameters, there also exist works concentrated on adjusting the pretraining task, thus improving performance on specific target task types. VECO (Luo et al., 2021) adds a cross-attention module between two different languages for better capturing the cross-lingual information and performs better on downstream cross-lingual understanding tasks and translation tasks.

Moreover, there are also vocabulary-level methods that adjust the dictionary to avoid possible mismatching for more effective inference. Liu et al. (2021c) introduce an embedding generator to adapt the vocabulary, leading to better performance on NLG tasks. Hsu et al. (2020) achieve model acceleration without sacrificing the accuracy based on model adjustment along with dictionary clipping.

Our work belongs to the task-level approach, by introducing a pretext task with code-switching for finetuning multilingual Seq2Seq PLMs on cross-lingual generation and understanding. The difference between this work and previous ones lies in our employment of a code-switching restoration task, which involves sampling in both languages from the task, thereby ensuring the model is more adaptive to downstream cross-lingual settings.

3. Preliminary

3.1. Sequence-to-sequence model based cross-lingual text generation

Sequence-to-sequence model, transformer, has achieved remarkable performance on various cross-lingual text generation tasks, which models source language sentences with a bidirectional attention-based encoder and generates target language representations with a decoder based on the encoder output. For a sample $S_i = (X_i, Y_i) = (x_1, \ldots, x_n; y_1, \ldots, y_m)$, the training of cross-lingual generation usually follows a teacher force fashion, maximizing the conditional generation objective over the training data (X, Y):

$$\begin{aligned}\mathcal{L}_{\text{gen}} &= \sum_{X_i, Y_i \in (X,Y)} -\log P(Y_i|X_i) \\ &= \sum_{X_i, Y_i \in (XY)} \sum_{j=1}^{m} -\log P(y_j|y_1, \ldots, y_{j-1}, X_i)\end{aligned} \tag{1}$$

3.2. Denoising task for multilingual sequence-to-sequence pretraining

The denoising pretrain task learns a function that directly maps the sentence with noises to the original sentence, as shown at the top of Fig. 4. Words colored with **Gray**, and **Cyan** represent the source and target language, respectively. We adopt mBART (Liu et al., 2020) as the backbone, which uses a standard sequence-to-sequence Transformer (Vaswani et al., 2017), trained on the CC-25[2] dataset. For a sample X^l in language l, mBART is trained to capture monolingual knowledge during pretraining by optimizing the following objective:

$$\mathcal{L}_{\text{PT}} = \sum_{l \in L} -\log P(X^l|N(X^l)) \tag{2}$$

where L represents the set of languages contained in CC-25, $\mathrm{N}(\cdot)$ is the noise function that mainly contains span masking and sentence permutation. For span masking in mBART, 35% of total words are selected, and each span is replaced with a mask token. The span length is determined by a Poisson distribution whose expected length is set to 3.5.

3.3. Standard finetuning for cross-lingual tasks

Given a cross-lingual sentence pair $S_i = (X_i, Y_i)$, standard finetuning directly feeds X_i into the encoder and estimates parameters under the supervision of Y_i. As shown at the bottom of Fig. 4, we follow mBART to append source language ID and target language ID, e.g. "[EN]" and "[ZH]", to the end of X_i and the beginning of Y_i. Besides language ID, we employ a symbol "$\langle/s\rangle$" for the summarization task to separate sentences. During finetuning downstream tasks, the model heavily depends on the cross-lingual information that multilingual Seq2Seq PLMs may lack, motivating us to introduce the cross-lingual information before standard finetuning.

4. Bridging cross-lingual gaps

In this section, we present a two-stage finetuning approach with code-switching restore task, which leads to better performance over standard finetuning both on cross-lingual generation tasks and cross-lingual understanding tasks.

4.1. Easy-to-acquire cross-lingual knowledge

To sufficiently transfer monolingual knowledge in different languages to downstream cross-lingual tasks, we consider adopting a cross-lingual pretext task with *easy-to-acquire cross-lingual knowledge*, i.e. "Unsupervised Translation Vocabulary" in the middle of Fig. 4, between multilingual Seq2Seq PLMs and standard finetuning.

Following Artetxe et al. (2018) and Yang et al. (2020), we utilize unsupervised word embedding mapping to extract translation vocabulary with the *downstream source-side or target-side monolingual corpus*

[2] https://github.com/pytorch/fairseq/tree/main/examples/mbart.

only. More specifically, given the downstream corpus denoted as $\mathcal{D} = \{S_i = (x_1, \ldots, x_n; y_1, \ldots, y_m)\}$. We first train a word2vec (Mikolov et al., 2013) model to get the in-domain word-level embedding for both the source side and target side, i.e. E_x and E_y. These two sets of embeddings are then embedded into a shared feature space via self-learning following. Thus, we can measure the semantic distance of two words in different languages by the dot product distance of the corresponding embedding pair. We randomly select one of the top-k nearest neighbors as the semantic similar words for switch operation. Formally, the selected neighbors for x_i and y_i are denoted as x_i' and y_i', respectively.

Algorithm 1 Algorithm of code-switching finetuning

Input: downstream finetuning corpus $\mathcal{D} = \{S_i = (x_1, .., x_n; y_1, ..., y_m)\}$, pretrained model M_{PT}
Output: finetuned model M_{stage1}
1: get word-level embedding: $\mathcal{D} \rightarrow E_x, E_y$
2: embed E_x, E_y into shared feature space
3: select neighbors x_i' and y_i' for each x_i and y_i based on similarity metrics, respectively
4: **for** S_i in $\mathcal{D}$ **do**
5: randomly select spans with length determined by a Poisson distribution
6: get $X_i' = \left(x_1, .., x_i', ..., x_{i+j}', ..., x_n\right)$
7: get $Y_i' = \left(y_1, .., y_i', ..., y_{i+j}', ..., y_m\right)$
8: update M_{PT} with objective (3)
9: **end for**
10: **return** M_{stage1}

4.2. The code-switching restore task

The code-switching restores task is designed to inject the above-unsupervised translation vocabulary to bridge the cross-lingual gap between pretraining and standard finetuning. The code-switching restore task will be performed on the downstream languages. We use the source-side code-switching as an example here, although the same operations apply to the target language. As illustrated in Fig. 4, given a sentence pair $S_i = (X_i, Y_i) = (x_1, \ldots, x_n; y_1, \ldots, y_m)$ selected from $\mathcal{D}$, where X is Chinese sentence "布什与沙龙举行了会谈$_{\ominus}$" and Y is English sentence "Bush held a talk with Sharon.", we sample a text span "举行了会谈" from the Chinese sentence X_i whose length is determined by a Poisson distribution. For each selected word x_i, e.g. "举行", we replace it with one of its unsupervised translation vocabulary x_i', i.e. "held", to derive the new code-switching source sentence. Such sampling will be repeated until the total percentage of substituted words reaches 35%.[3] Then the source sentence of our pretext task can be written as $X_i' = \left(x_1, \ldots, x_i', \ldots, x_{i+j}', \ldots, x_n\right)$, e.g. "布什与沙龙 held a talk$_{\ominus}$" in our example.

After that, we can get the pretext task sentence pair $S_i = (X_i'; X_i)$, i.e. "布什与沙龙 held a talk$_{\ominus}$"→"布什与沙龙举行了会谈$_{\ominus}$" in our example. Then we feed this sentence pair into the multilingual Seq2Seq PLM, optimizing the cross-entropy loss. Considering the same operation for the target-side sentence Y_j, the total objective for the code-switching restore task can be written as:

$$\mathcal{L}_{\text{pretext}} = -\sum_{X_i \in D} \log P(X_i | X_i') - \sum_{Y_j \in D} \log P(Y_j | Y_j') \quad (3)$$

where X_i, Y_j mean source and target sentences in the downstream dataset, respectively, and X_i', Y_j' are correspondingly code-switching sentences.

[3] We keep the substitution rate as the default mask ratio of mBART.

4.3. Code-switching finetuning

After introducing downstream cross-lingual knowledge with our proposed code-switching restore task, the multilingual Seq2Seq PLMs will be easier to adapt to downstream standard finetuning. Thus, the recipe for using our approach follows a two-stage fashion: In the **first** stage, we tune the multilingual Seq2Seq PLMs for certain steps using the above pretext task, where *the model can gain a better in-domain initialization with the narrowed cross-lingual gap* for the downstream tasks, as illustrated in Algorithm 1. In the **second** stage, we tune the model normally on the downstream task.

The effective performance of our method comes from the narrowed cross-lingual representation distance in Section 6.1, confirming our above *claim*.

5. Experiments

We conduct cross-lingual **NLG** evaluations on four typical tasks: (i) bilingual translation; (ii) zero-shot translation; (iii) cross-lingual summarization, and **NLU** evaluations on representative (iv) cross-lingual Natural Language Inference (cross-lingual NLI) tasks.

5.1. Experimental setup

Baselines. We select three models as our baseline models for comparison. For cross-lingual generation tasks, we select IAFT, Random, and mBART to verify the performance improvement of our approach. And, mBERT and mBART are used for cross-lingual NLI experiments.

- **mBART**: mBART (Liu et al., 2020), developed by Meta, is a robust multilingual Seq2Seq PLM encompassing 680 million parameters. It has been pretrained on billions of monolingual data across 25 languages, showing strong cross-lingual transfer ability. In our methodology, we adhere to the recommendations outlined in the mBART paper (Liu et al., 2020), specifically employing direct finetuning of the mBART model on downstream tasks with identical update steps to our proposed two-stage finetuning strategy. To obtain the best performance, we apply grid search to obtain hyper-parameters, e.g. *lr, batch size, training steps*, based on the validation set.
- **Random**: We also include a comparison with a model randomly initialized without pretraining for each cross-lingual generation task, similar to the setting in mBART (Liu et al., 2020). We directly report the bilingual translation results from mBART and previous works.
- **IAFT**: Additionally, we also incorporate the Input-Adaption Finetuning (IAFT) (Wang et al., 2022), which alleviates overestimation problem and enhances the model robustness during finetuning, to facilitate comparative analysis. Specifically, adhering to the recommendations in the original study, we introduce noise into the source sentences of a 10% subset of the train data.

- **mBERT**: The multilingual version of Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) indicates the bidirectional transformer pretrained on the Wikipedias of 104 languages using Masked Language Modeling (MLM) task, which has been shown to achieve robust performance in NLI tasks.

Model training. We select mBART (Liu et al., 2020) as our backbone model, which is pretrained on CC-25[4] dataset. For each task, the datasets are tokenized into subword units with the 250k SentencePiece (Kudo and Richardson, 2018) vocabulary of mBART. All experiments are conducted with the open-source toolkit `fairseq` (Ott et al., 2019) on the A100 GPUs. For hyper-parameters, we use dropout 0.3,

[4] https://commoncrawl.org.

Table 4
Performance of bilingual translation tasks on WMT and IWSLT datasets. Results are in the format "SacreBLEU/COMET-score". (**Bold**: the best results.)

Languages	En-Vi		En-De		En-Tr	
Data source	IWSLT15		IWSLT14		WMT17	
Direction	←	→	←	→	←	→
Random	23.6/–	24.8/–	34.4/–	29.2/–	12.2/–	9.5/–
mBART	28.3/0.178	29.4/0.219	37.8/0.528	32.6/0.446	22.5/0.445	20.0/0.825
IAFT	29.1/0.172	29.7/**0.247**	38.6/0.536	**32.7**/0.448	22.8/0.416	20.1/0.818
Ours	**29.2***/**0.193**	**29.8**/0.239	**39.2****/**0.563**	**32.7**/**0.454**	**23.0***/**0.465**	**20.5***/**0.840**
Languages	En-Ko		En-It		En-Lt	
Data source	IWSLT17		IWSLT17		WMT19	
Direction	←	→	←	→	←	→
Random	15.3/–	16.3/–	31.7/–	28.0/–	18.1/–	12.1/–
mBART	17.2/0.201	18.8/0.290	37.8/0.466	34.8/0.576	29.1/0.511	12.9/0.310
IAFT	18.2/0.194	18.7/0.295	38.2/0.438	35.3/0.579	29.9/0.493	13.0/0.272
Ours	**18.3****/**0.251**	**18.9**/**0.293**	**38.5***/0.460	**35.9****/**0.597**	**30.7****/**0.525**	**13.5***/**0.312**

* Significant statistical difference ($p < 0.05$) from mBART.
** Significant statistical difference ($p < 0.01$) from mBART.

lr 3e–5, warm-up steps 2500, and label smoothing ratio 0.2. Other detailed settings, such as the update steps and evaluation setup, for different experiments are detailed in their respective sections. To construct the unsupervised translation vocabulary acquisition, we closely follow (Artetxe et al., 2018) as described in Section 4.1. Recall that the word embeddings for code-switching are only trained using datasets of downstream tasks.

5.2. Bilingual translation

The bilingual translation task aims to translate a source language sentence into the target language. In this task, we evaluate the effectiveness of our approach, specifically focusing on its accuracy and robustness across 12 translation directions, using various publicly available parallel corpora. These include IWSLT15 Vi↔En (133k), IWSLT14 De↔En (160k), WMT17 Tr↔En (207k), IWSLT17 Ro↔En (230k), IWSLT17 It↔En (290k), and WMT19 Lt↔En (1920k). For data preprocessing, following Tang et al. (2021), we remove duplicate sentence pairs, ensuring cleaner datasets for training. We deploy 5k training steps for the pretext task and 25k for the translation task. The translations are generated using beam search with a beam size of 5.

To evaluate the quality of the translation, we employ BLEU as the primary metric. BLEU (Papineni et al., 2002) is a reference-based metric that reflects the n-gram precision of translations compared to reference texts, and includes a brevity penalty to discourage overly short translations. The precision is calculated by dividing the number of matching n-grams by the total number of n-grams in the translation, using the formula below:

$$P_n = \frac{\sum_{y' \in Y'} \sum_{n-gram \in y'} \min(C_{n-gram}(y', y), C_{n-gram}(y))}{\sum_{y' \in Y'} \sum_{n-gram \in y'} C_{n-gram}(y')},$$

where y' represents the sentence in translation outputs Y', and y is the corresponding reference sentence. The count of matching n-grams and total n-grams in y' are represented by $C_{n-gram}(y', y)$ and $C_{n-gram}(y')$, respectively. Sentence brevity penalty (BP) mitigates the impact of sentence length. The formulas for BP and BLEU are as follows:

$$BP = \begin{cases} 1 & if\ |y'| > |y| \\ e^{1-|y|/|y'|} & if\ |y'| \leq |y| \end{cases},$$

$$BLEU = BP * exp(\sum_{n=1}^{N} \omega_n log(P_n)),$$

where $|*|$ represents the sentence length, ω_n is the weight of the n-gram. We utilize SacreBLEU[5] (Post, 2018), with its default settings, to compute the scores. Additionally, we conduct statistical difference tests to compare our models with mBART. We also use COMET[6] (Rei et al., 2020) as an additional metric to evaluate semantic faithfulness, which is more correlated to human judgments. COMET model evaluates translation quality by receiving a triplet consisting of a source sentence, its translation, and a reference translation. It then returns the final score that reflects the quality of the translation compared to both source and reference. The COMET score is unbounded but typically ranges between −1.0 and 1.0, where 1.0 indicates a perfect translation. It is worth noting that, for supervised models, we directly report the SOTA random initialized SacreBLEU results from mBART and previous works. Furthermore, we do not compare with the winners of considered shared tasks. This is because typically, the winning systems utilize a substantial amount of additional monolingual data, which diverges from our experimental setting.

Table 4 lists the results. We can see that our approach achieves better translation than both Random, IAFT, and mBART baselines in all languages. Noticeably, our method achieves improvements against mBART using standard finetuning on 9 out of 12 directions with sufficient significance, proving the effectiveness of introducing a code-switching restore task to bridge the cross-lingual gap for multilingual Seq2Seq PLM. We also see XX→En translation tasks all achieve a sufficient performance improvement, which indicates our approach is relatively more beneficial for English as the target language. The results of mBART consistently outperform the randomly initialized transformer by a large margin, which confirms the benefit of pretraining and agrees with previous works.

Table 4 also shows that our approach can achieve COMET-score improvements compared to standard finetuning across 11 experiments. We got an extra promotion of up to 0.050 (Ko→En). A higher COMET score means that the sentences translated by our approach are more in line with human translation.

5.3. Zero-shot translation

The zero-shot translation tasks aim to transfer the ability of pretrained YY→En model to the zero-resource language pairs XX→En.

Generally, the closer distance between language XX and YY, the better transfer performance will be realized by the model. In our settings, we use all the above 6 bilingual translation corpora, and directly utilize the models finetuned on YY→En from bilingual translation experiments in Section 5.2 to test on XX to En. We consider the same metrics in the bilingual translation tasks for evaluation, i.e. SacreBLEU and COMET.

[5] https://github.com/mjpost/sacrebleu.

[6] Unbabel/wmt20-comet-da.

Table 5
Performance of zero-shot translation tasks on XX→En directions. Results are in the format "SacreBLEU/COMET-score". (**Bold**: the best results.)

			Finetuned languages					
			Vi	De	Tr	Ko	It	Lt
Testing languages	Vi	mBART	28.3/0.178	15.4/−0.137	14.0/−0.187	18.5/−0.064	19.0/−0.070	13.3/−0.294
		Ours	**29.2***/**0.193**	15.4/**−0.100**	**15.1**/**−0.166**	17.8/**0.140**	18.2/**−0.066**	12.5/−0.367
	De	mBART	22.1/0.101	37.8/0.528	19.0/0.031	23.2/0.143	24.3/0.107	19.7/−0.067
		Ours	22.0/0.093	**39.2****/**0.563**	**20.0***/**0.079**	23.1/**0.090**	24.2/0.106	19.7/−0.118
	Tr	mBART	8.8/−0.161	9.1/−0.063	22.5/0.445	15.0/0.132	9.5/−0.117	12.8/0.016
		Ours	**8.9**/**−0.103**	8.7/−0.032	**23.0***/**0.465**	**15.1** /0.126	9.5/−0.080	12.8/−0.038
	Ko	mBART	7.7/−0.256	8.0/−0.199	8.9/−0.087	17.2/0.201	6.8/−0.303	6.2/−0.251
		Ours	**7.9**/**−0.186**	**8.6***/**−0.107**	**9.1**/**−0.062**	**18.3***/**0.251**	**7.0**/**−0.177**	**7.1***/**−0.217**
	It	mBART	28.5/0.116	24.9/0.050	22.8/0.013	27.4/0.191	37.8/0.466	22.7/−0.161
		Ours	**29.3***/**0.125**	**25.6***/**0.083**	**24.3****/**0.110**	**27.8**/**0.132**	**38.5***/0.460	22.4/−0.243
	Lt	mBART	12.8/−0.060	12.2/−0.064	16.2/0.053	14.9/0.019	12.8/−0.079	29.1/0.511
		Ours	**13.5***/**0.025**	11.6/**0.008**	**16.7**/**0.131**	14.8/**0.025**	**13.5***/**−0.012**	**30.7****/**0.525**

* Significant statistical difference ($p < 0.05$) from mBART.
** Significant statistical difference ($p < 0.01$) from mBART.

Table 6
Performance of cross-lingual summarization tasks on NCLS dataset. The best results are **bold**.

Model	En→Zh			Zh→En		
	Rouge-1	Rouge-2	Rouge-L	Rouge-1	Rouge-2	Rouge-L
Random	6.5	1.1	6.4	20.4	5.0	16.3
mBART	22.1	8.7	21.7	27.7	10.2	23.1
Ours	$\mathbf{22.4}_{+0.3}$	$\mathbf{9.1}_{+0.4}$	$\mathbf{22.1}_{+0.4}$	$\mathbf{27.8}_{+0.1}$	$\mathbf{12.4}_{+2.2}$	$\mathbf{23.2}_{+0.1}$

Table 7
Performance of cross-lingual NLI tasks on datasets from XNLI. Average scores on all tasks are underlined. The best results are **bold**.

Model	De	Es	Fr	Ru	Tr	Vi	Zh	*Avg.*
mBERT	71.1	74.3	73.8	69.0	61.6	69.5	69.3	*69.8*
mBART	76.4	79.0	78.4	74.0	68.2	74.7	69.6	*74.3*
Ours	$\mathbf{77.3}_{+0.9}$	$\mathbf{79.4}_{+0.4}$	$\mathbf{79.0}_{+0.6}$	$\mathbf{75.2}_{+1.2}$	$\mathbf{71.9}_{+3.7}$	$\mathbf{76.1}_{+1.4}$	$\mathbf{71.2}_{+1.6}$	*75.7*$_{+1.4}$

As shown in Table 5, we present the zero-shot results of mBART models under the standard finetuning pattern and our proposed approach. Obviously, our method shows improvements on 16 (in terms of BLEU scores) and 21 (in terms of COMET scores) out of 30 zero-shot directions, demonstrating the universality and robustness of our method.

5.4. Cross-lingual summarization

The cross-lingual summarization task is designed to automatically generate a summary in the target language that retains the most important content of a document in the source language. We conduct our experiments on the NCLS dataset, as introduced by Zhu et al. (2019), which includes 370k English-to-Chinese (En→Zh) and 1.69 million Chinese-to-English (Zh→En) document-summary pairs. For documents that are too long, we follow the approach of Zhu et al. (2019) by truncating the input sentence to 220 words and limiting the output to 130 words, with an extension to 160 words for summaries in Chinese. We employ different batch sizes for the En→Zh and Zh→En directions (49k and 98k tokens, respectively) based on the average document lengths, noting that En→Zh documents are significantly longer. The training regimen consists of 5k steps for the pretext task and 50k steps for the remainder of the cross-lingual summarization task. After finetuning, the model is capable of generating English or Chinese summaries based on documents in Chinese or English, respectively.

We follow Lin (2004)[7] to report the F1-score of ROUGE-1, ROUGE-2, and ROUGE-L. ROUGE (Lin, 2004) is a reference-based metric used to evaluate the quality of summaries by measuring how much of the content in reference summaries is captured by a generated summary. The ROUGE-n metric calculates the recall of n-grams in the generated summary compared to the reference, using the formula:

$$\text{ROUGE-n} = \frac{\sum_{y\in Y}\sum_{n-gram\in y} C_{n-gram}(y', y)}{\sum_{y\in Y}\sum_{n-gram\in y} C_{n-gram}(y)}, \tag{4}$$

where y represents one summary in reference summarizations Y, and y' denotes the generated summary. $C_{n-gram}(y', y)$ counts the number of matching n-grams between y' and y, while $C_{n-gram}(y)$ counts the n-grams in the reference summary y. The metrics ROUGE-1 and ROUGE-2 specifically focus on unigrams and bigrams, respectively. ROUGE-L, on the other hand, is based on the longest common subsequence (LCS) shared between the generated summary and a reference summary. It is calculated as follows:

$$R_{lcs} = \frac{LCS(y', y)}{|y|},$$

$$P_{lcs} = \frac{LCS(y', y)}{|y'|},$$

$$\text{ROUGE-L} = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}},$$

Where $LCS(y', y)$ is the length of the longest common subsequence of y' and y, $|*|$ measures the sentence length. β is the hyper-parameters. The summaries are generated by beam search with size 6 in decoding. After generation, we use Standford NLP tokenizer[8] to segment words for English, while Chinese texts are split by characters.

Table 6 lists the performance of generated summarization. Clearly, our model outperforms other baselines consistently. Compared with the model after the standard finetuning, our model achieves +0.4 in ROUGE-L for En→Zh, and +0.1 in ROUGE-L for Zh→En, indicating the superiority of our method.

5.5. Cross-lingual NLI

In the cross-lingual natural language inference (cross-lingual NLI) task, we finetune the model using English training data and then perform inference in other languages. We utilize datasets from the XNLI benchmark (Conneau et al., 2018) and report the accuracy achieved. For this task, we employ publicly accessible bilingual embeddings from MUSE[9] (Lample et al., 2018) to facilitate code-switching. The model is

[7] https://github.com/pltrdy/files2rouge.

[8] https://nlp.stanford.edu/software/tokenizer.shtml.

[9] https://github.com/facebookresearch/MUSE.

Table 8
The distances of cross-lingual sentence representations on language directions En-XX. The closet distances are **bold**.

Language	De	Ko	Vi	Lt
Pretrained	28.33	90.64	64.84	36.68
Finetuned	8.25	7.22	8.38	11.37
Ours	**7.90**$_{-0.35}$	**4.80**$_{-2.42}$	**6.70**$_{-1.68}$	**5.17**$_{-6.20}$

finetuned with a batch size of 128 and a learning rate of 5e–6 over 10 epochs, incorporating early stopping based on the validation set's performance.

Table 7 shows that our proposed finetuning approach outperforms the other two strong baselines, i.e. mBERT[10] (Devlin et al., 2019) and mBART with standard finetuning. Our model achieves +1.4 average score improvement over mBART and +5.9 over mBERT. Our method achieves the improvement by a large margin on language like "Tr" whose pretraining corpus is limited in mBART, showing the transferability and effectiveness of our proposed method on NLU tasks, especially for low-resource languages.

6. Discussion

The downstream finetuning tasks (e.g. translation and cross-lingual summarization) learn to transform a sentence from one language to another, while Seq2Seq pretraining learns to reconstruct the input sentence (Liu et al., 2021a). Our proposed approach is expected to narrow the objective discrepancy between pretraining and finetuning.

To better understand how our approach alleviates the cross-lingual gaps between multilingual Seq2Seq PLMs and downstream tasks, we make in-depth analyses to investigate three problems: **Q1:** How does our method narrow cross-lingual representation gaps? (Section 6.1) **Q2:** Whether our approach improves the low-frequency word translation? (Section 6.2) and **Q3:** Does our strategy acquire large computational costs? (Section 6.3)

6.1. Narrowing cross-lingual representation distance

Many previous works have shown that narrowing the cross-lingual gaps can bring better modeling performance (Zhou et al., 2019; Ding et al., 2020). To quantify the narrowed cross-lingual representation distance, we take the average sentence embedding to represent each language and evaluate the language distance on CC-100 (Conneau et al., 2020b). In practice, we select a subset S containing 20k sentences for approximation.

Given a sentence $S_i = (x_1, \ldots, x_n)$, we feed it into the model to obtain the sentence embedding e_i, where e_i is the output on the last token from the decoder. Applying the Euclidean distance, the distance between the source and target language is therefore defined as:

$$\mathcal{D}_{\text{language}} = \left(\sum_{i=1}^{20k} \left| e_i^{source} - e_i^{target} \right|^2\right)^{1/2} \quad (5)$$

Table 8 shows the language distances of four language pairs. We notice that the cross-lingual representation distances are reduced in all scenarios after finetuning. Moreover, our method can further reduce the distance up to −6.20 reduction on En-Lt. This demonstrates that our strategy could improve the finetuning performance by significantly reducing the cross-lingual representation distance.

6.2. Improving low-frequency word translation

In the code-switching restore stage, each selected word will be replaced with multiple candidates in another language. Therefore, our method is expected to achieve better generalization, particularly at the word level. To verify our hypothesis, we measure the translation performance of low-frequency words, which is evaluated using the F-measure with compare-mt.[11] We calculate word frequency on the train set. The word-level translation performance in terms of F-scores on the IWSLT14 De→En test set is shown in Fig. 5. Words that appear less than 100 times are regarded as low-frequency words while more than 1000 times are regarded as high-frequency words, and the rest belong to mid-frequency ones.

We see that our approach boosts the translation accuracy, especially for low-frequency words where a more remarkable improvement can be observed, i.e. +2.4, nicely validating our hypothesis. We further present the case study for IWSLT14 De→En translation results in Table 9. Compared with standard finetuning, our model gives low-frequency word translations that agree more with human translations.

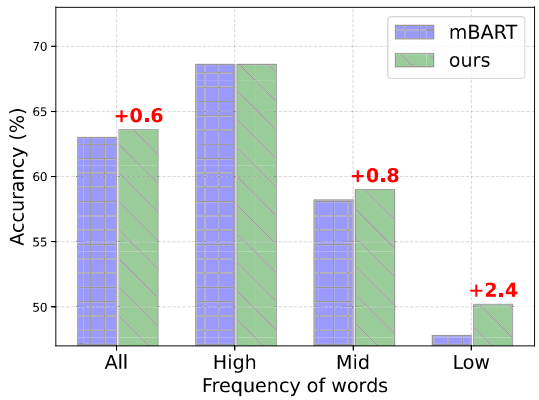


Fig. 5. Performance of word translation accuracy with different frequencies on IWSLT14 De→En test set.

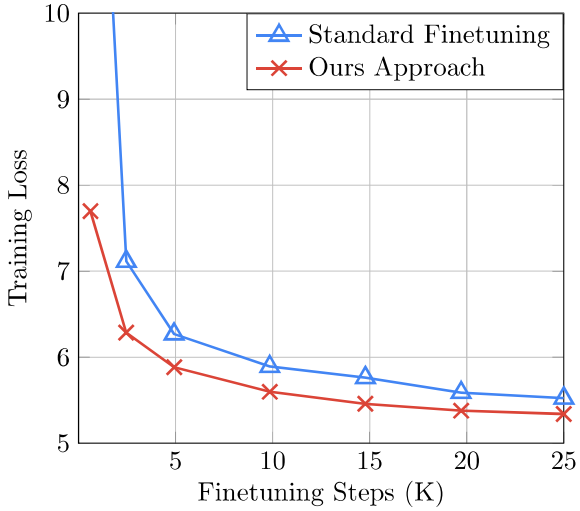


Fig. 6. Comparison of training dynamics during finetuning on IWSLT14 De→En dataset.

6.3. Trivial computational costs

As shown in Fig. 6, we also conclude that the code-switching restore task benefits the finetuning converges speed. "Ours Approach" denotes the loss curve of the second stage. At the second finetuning stage

[10] We report the results of mBERT and XLM from Liang et al. (2020).

[11] https://github.com/neulab/compare-mt.

Table 9
Case Study for De→En translation. We present translation examples for low-frequency and med-frequency words, which are marked with yellow. "Input" and "Human" indicate the source German sentences and the ground-truth target English sentences, respectively.

	Sentence		Sentence
Input	nichts wandert in meinen ärmel oder kommt heraus, keine trickserei. und sie können alles untersuchen.	Input	passiert ist das das so gennante maschinenzeitalter.
Human	nothing goes up or down my sleeve , no trickery . and you can examine everything.	Human	what happened is the so-called machine age.
mBART	nothing goes into my throat or comes out, no tricks, and you can study everything.	mBART	this is the so genetic age of machines.
Ours	nothing goes into my sleeve or comes out of my sleeve , no trickery . and you can examine anything.	Ours	this is the so-called machine age.
	Sentence		Sentence
Input	es gibt nur eine und das sind die vereinigten staaten – zum glück, zum glück.	Input	dies ist, denke ich, so tief in unserer wasserversorgung verankert, dass es keinem in dem sinn kommen würde es zu hinterfragen.
Human	there is only one, and that's the united states — fortunately , fortunately .	Human	this, i think, is so deeply embedded in the water supply that it wouldn't occur to anyone to question it.
mBART	there's only one, and that's the united states — fortunate, fortunate.	mBART	this is so deeply rooted in our water supply, i think, that nobody would think of it as questioning it.
Ours	there's only one, and that's the united states — fortunately , fortunately .	Ours	this is so deeply embedded in our water supply, i think, that it would never occur to anybody to question it.

Table 10
Evaluation of time consumption (Hours).

Methods	Machine translation	Cross-Lingual NLI	Cross-Lingual summarization
Standard finetuning	4.6	2.8	43.9
Ours approach	$5.3_{+0.7}$	$3.2_{+0.5}$	$53.1_{+9.2}$

of our method, the model after the pretext task only requires about half of the update steps (12k vs. 25k) to reach the same loss level, compared to standard finetuning. Since the tuning step number in the first stage is set to 5k (only 16.7% of total steps), we argue that the computational costs are not an obstacle to the extensibility of our approach. Additionally, our model can achieve even lower loss with further training applied, indicating that the code-switching restore task can adapt the model in advance to cross-lingual downstream tasks more effectively. We also present the time consumption data in Table 10, where the experiments were conducted using 2 A100 GPUs. We report results for De→En machine translation, cross-lingual NLI, and Zh→En cross-lingual summarization tasks. As shown in the table, our method incurs minimal additional costs compared to standard finetuning. For example, it requires only an additional 0.5 h for cross-lingual NLI and 0.7 h for the machine translation task.

7. Conclusion

In this work, we propose a feasible code-switching based two-stage finetuning approach to mitigate the cross-lingual gap in the multilingual Seq2Seq pretrain-finetune paradigm. Our algorithm efficiently transfers monolingual knowledge from multilingual Seq2Seq PLMs into downstream cross-lingual tasks. Specifically, in the first stage, the model is trained to reconstruct code-switching sentences in both the source and target languages to mitigate task and domain discrepancies. Then, it is finetuned on downstream tasks to derive the final model. Experiments on several cross-lingual generation and understanding tasks illustrate the effectiveness and universality of our finetuning approach. In addition, our pretext task narrows the cross-lingual representation distance compared with the standard finetuning approach. Our model also achieves better generalization performance as evidenced by improved performance on low-frequency words. Analysis shows that our approach only requires minimal computational costs, having the potential to be a universal plug-in strategy for more multilingual Seq2Seq PLMs. We believe our method can inspire future research in this area, especially in efficiently bridging the cross-lingual gap of multilingual Seq2Seq PLMs.

8. Limitations

This work primarily focuses on analyzing mBART, a model trained using a denoising task. In future studies, we plan to explore the effectiveness of our strategy in other multilingual Seq2Seq PLMs, such as mT5 (Xue et al., 2021). Exploring more efficient pretext tasks would be beneficial to further reduce the cross-lingual gaps in these models. This work also involves additional word-level embedding models, which lead to increased computational costs. In future studies, we aim to design more effective finetuning methods that do not require additional models.

CRediT authorship contribution statement

Changtong Zan: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation. **Liang Ding:** Writing – review & editing, Supervision, Resources, Investigation, Formal analysis, Conceptualization. **Li Shen:** Writing – review & editing, Validation, Supervision, Formal analysis, Conceptualization. **Yu Cao:** Writing – review & editing, Validation, Formal analysis, Conceptualization. **Weifeng Liu:** Writing – review & editing, Resources, Project administration, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Grant No. 62372468), Shandong Natural Science Foundation (Grant No. ZR2023MF008), Major Basic Research Projects in Shandong Province (Grant No. ZR2023ZD32) and Qingdao Natural Science Foundation (Grant No. 23-2-1-161-zyyd-jch).

Data availability

Data will be made available on request.

References

Aghajanyan, A., Gupta, A., Shrivastava, A., Chen, X., Zettlemoyer, L., Gupta, S., 2021a. Muppet: Massive multi-task representations with pre-finetuning. In: EMNLP. pp. 5799–5811. http://dx.doi.org/10.18653/v1/2021.emnlp-main.468.

Aghajanyan, A., Shrivastava, A., Gupta, A., Goyal, N., Zettlemoyer, L., Gupta, S., 2021b. Better fine-tuning by reducing representational collapse. In: ICLR.

Akkem, Y., Biswas, S.K., Varanasi, A., 2023a. Smart farming using artificial intelligence: A review. Eng. Appl. Artif. Intell. 120, 105899. http://dx.doi.org/10.1016/j.engappai.2023.105899.

Akkem, Y., Biswas, S.K., Varanasi, A., 2024. A comprehensive review of synthetic data generation in smart farming by using variational autoencoder and generative adversarial network. Eng. Appl. Artif. Intell. 131, 107881. http://dx.doi.org/10.1016/j.engappai.2024.107881.

Akkem, Y., Kumar, B.S., Varanasi, A., 2023b. Streamlit application for advanced ensemble learning methods in crop recommendation systems – a review and implementation. Indian J. Sci. Technol. 16, 4688–4702. http://dx.doi.org/10.17485/IJST/v16i48.2850.

Artetxe, M., Labaka, G., Agirre, E., 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In: ACL. pp. 789–798. http://dx.doi.org/10.18653/v1/P18-1073.

Bowman, S.R., Angeli, G., Potts, C., Manning, C.D., 2015. A large annotated corpus for learning natural language inference. In: EMNLP. pp. 632–642. http://dx.doi.org/10.18653/v1/D15-1075.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D., 2020. Language models are few-shot learners. In: NeurIPS.

Chaudhary, A., Raman, K., Srinivasan, K., Chen, J., 2020. Dict-mlm: Improved multilingual pre-training using bilingual dictionaries. arXiv preprint.

Chen, S., Hou, Y., Cui, Y., Che, W., Liu, T., Yu, X., 2020. Recall and learn: Fine-tuning deep pretrained language models with less forgetting. In: EMNLP. pp. 7870–7881. http://dx.doi.org/10.18653/v1/2020.emnlp-main.634.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: EMNLP. http://dx.doi.org/10.3115/v1/D14-1179.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., Stoyanov, V., 2020a. Unsupervised cross-lingual representation learning at scale. In: ACL. pp. 8440–8451. http://dx.doi.org/10.18653/v1/2020.acl-main.747.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V., 2020b. Unsupervised cross-lingual representation learning at scale. In: ACL. pp. 8440–8451. http://dx.doi.org/10.18653/v1/2020.acl-main.747.

Conneau, A., Lample, G., 2019. Cross-lingual language model pretraining. In: NeurIPS.

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., Stoyanov, V., 2018. XNLI: Evaluating cross-lingual sentence representations. In: EMNLP. pp. 2475–2485. http://dx.doi.org/10.18653/v1/D18-1269.

Dai, Y., Yang, X., Leng, M., 2023. Optimized Seq2Seq model based on multiple methods for short-term power load forecasting. Appl. Soft Comput. 142, 110335. http://dx.doi.org/10.1016/j.asoc.2023.110335.

Daumé, III, H., 2007. Frustratingly easy domain adaptation. In: ACL. pp. 256–263.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, r., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In: NAACL. pp. 4171–4186. http://dx.doi.org/10.18653/v1/N19-1423.

Ding, L., Wang, L., Tao, D., 2020. Self-attention with cross-lingual position representation. In: ACL. pp. 1679–1685. http://dx.doi.org/10.18653/v1/2020.acl-main.153.

Fan, Y., Liang, Y., Muzio, A., Hassan, H., Li, H., Zhou, M., Duan, N., 2021. Discovering representation sprachbund for multilingual pre-training. In: Findings of EMNLP. pp. 881–894. http://dx.doi.org/10.18653/v1/2021.findings-emnlp.75.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., Smith, N.A., 2020. Don't stop pretraining: Adapt language models to domains and tasks. In: ACL. pp. 8342–8360. http://dx.doi.org/10.18653/v1/2020.acl-main.740.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S., 2019. Parameter-efficient transfer learning for NLP. In: ICML. pp. 2790–2799.

Hsu, Y.-T., Garg, S., Liao, Y.-H., Chatsviorkin, I., 2020. Efficient inference for neural machine translation. In: Workshop on Simple and Efficient Natural Language Processing. pp. 48–53. http://dx.doi.org/10.18653/v1/2020.sustainlp-1.7.

Khashabi, D., Min, S., Khot, T., Sabharwal, A., Tafjord, O., Clark, P., Hajishirzi, H., 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In: Findings of EMNLP. pp. 1896–1907. http://dx.doi.org/10.18653/v1/2020.findings-emnlp.171.

Klaar, A.C.R., Stefenon, S.F., Seman, L.O., Mariani, V.C., Coelho, L.d.S., 2023. Optimized EWT-Seq2Seq-LSTM with attention mechanism to insulators fault prediction. Sensors 23, 3202. http://dx.doi.org/10.3390/s23063202.

Kudo, T., Richardson, J., 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In: EMNLP: System Demonstrations. pp. 66–71. http://dx.doi.org/10.18653/v1/D18-2012.

Lample, G., Conneau, A., Ranzato, M., Denoyer, L., Jégou, H., 2018. Word translation without parallel data. In: ICLR.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. http://dx.doi.org/10.1038/nature14539.

Lee, C., Cho, K., Kang, W., 2020. Mixout: Effective regularization to finetune large-scale pretrained language models. In: ICLR.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L., 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: ACL. pp. 7871–7880. http://dx.doi.org/10.18653/v1/2020.acl-main.703.

Li, P., Li, L., Zhang, M., Wu, M., Liu, Q., 2022. Universal conditional masked language pre-training for neural machine translation. In: ACL. pp. 6379–6391. http://dx.doi.org/10.18653/v1/2022.acl-long.442.

Li, X., Wang, C., Tang, Y., Tran, C., Tang, Y., Pino, J., Baevski, A., Conneau, A., Auli, M., 2021. Multilingual speech translation from efficient finetuning of pretrained models. In: ACL. pp. 827–838. http://dx.doi.org/10.18653/v1/2021.acl-long.68.

Lian, J., Wu, S., Huang, S., Zhao, Q., 2024. A novel sequence-to-sequence based deep learning model for satellite cloud image time series prediction. Atmos. Res. 306, 107457. http://dx.doi.org/10.1016/j.atmosres.2024.107457.

Liang, Y., Duan, N., Gong, Y., Wu, N., Guo, F., Qi, W., Gong, M., Shou, L., Jiang, D., Cao, G., Fan, X., Zhang, R., Agrawal, R., Cui, E., Wei, S., Bharti, T., Qiao, Y., Chen, J.-H., Wu, W., Liu, S., Yang, F., Campos, D., Majumder, R., Zhou, M., 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In: EMNLP. pp. 6008–6018. http://dx.doi.org/10.18653/v1/2020.emnlp-main.484.

Lin, C.-Y., 2004. ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., Zettlemoyer, L., 2020. Multilingual denoising pre-training for neural machine translation. Trans. Assoc. Comput. Linguist. 8, 726–742. http://dx.doi.org/10.1162/tacl_a_00343.

Liu, X., He, P., Chen, W., Gao, J., 2019a. Multi-task deep neural networks for natural language understanding. In: ACL. pp. 4487–4496. http://dx.doi.org/10.18653/v1/P19-1441.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019b. Roberta: A robustly optimized bert pretraining approach. arXiv preprint.

Liu, X., Wang, L., Wong, D.F., Ding, L., Chao, L.S., Shi, S., Tu, Z., 2021a. On the copying behaviors of pre-training for neural machine translation. In: Findings of ACL. pp. 4265–4275. http://dx.doi.org/10.18653/v1/2021.findings-acl.373.

Liu, Z., Winata, G.I., Fung, P., 2021b. Continual mixed-language pre-training for extremely low-resource neural machine translation. In: Findings of ACL. pp. 2706–2718. http://dx.doi.org/10.18653/v1/2021.findings-acl.239.

Liu, X., Yang, B., Liu, D., Zhang, H., Luo, W., Zhang, M., Zhang, H., Su, J., 2021c. Bridging subword gaps in pretrain-finetune paradigm for natural language generation. In: ACL. pp. 6001–6011. http://dx.doi.org/10.18653/v1/2021.acl-long.468.

Liu, J., Zhong, Q., Ding, L., Jin, H., Du, B., Tao, D., 2021d. Unified instance and knowledge alignment pretraining for aspect-based sentiment analysis. arXiv preprint.

Luo, F., Wang, W., Liu, J., Liu, Y., Bi, B., Huang, S., Huang, F., Si, L., 2021. VECO: Variable and flexible cross-lingual pre-training for language understanding and generation. In: ACL. pp. 3980–3994. http://dx.doi.org/10.18653/v1/2021.acl-long.308.

Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. In: NeurIPS. pp. 3111–3119.

Nallapati, R., Zhou, B., dos Santos, C., Gulçehre, Ç., Xiang, B., 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In: CoNLL. pp. 280–290. http://dx.doi.org/10.18653/v1/K16-1028.

Ni, M., Huang, H., Su, L., Cui, E., Bharti, T., Wang, L., Zhang, D., Duan, N., 2021. M3P: Learning universal representations via multitask multilingual multimodal pre-training. In: CVPR. pp. 3977–3986.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., Auli, M., 2019. Fairseq: A fast, extensible toolkit for sequence modeling. In: NAACL (Demonstrations). pp. 48–53. http://dx.doi.org/10.18653/v1/N19-4009.

Papineni, K., Roukos, S., Ward, T., Zhu, W.-J., 2002. Bleu: a method for automatic evaluation of machine translation. In: ACL. pp. 311–318. http://dx.doi.org/10.3115/1073083.1073135.

Post, M., 2018. A call for clarity in reporting BLEU scores. In: WMT. pp. 186–191. http://dx.doi.org/10.18653/v1/W18-6319.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. 21 (1), 5485–5551.

Rei, R., Stewart, C., Farinha, A.C., Lavie, A., 2020. COMET: A neural framework for MT evaluation. In: EMNLP. pp. 2685–2702. http://dx.doi.org/10.18653/v1/2020.emnlp-main.213.

Song, K., Tan, X., Qin, T., Lu, J., Liu, T.-Y., 2019. MASS: Masked sequence to sequence pre-training for language generation. In: ICML. pp. 5926–5936.

Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., Gu, J., Fan, A., 2020. Multilingual translation with extensible multilingual pretraining and finetuning. arXiv preprint.

Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., Gu, J., Fan, A., 2021. Multilingual translation from denoising pre-training. In: Findings of ACL. pp. 3450–3466. http://dx.doi.org/10.18653/v1/2021.findings-acl.304.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: NeurIPS. Vol. 30.

Wang, S., Jiang, J., 2016. Learning natural language inference with LSTM. In: NAACL. http://dx.doi.org/10.18653/v1/N16-1170.

Wang, W., Jiao, W., Hao, Y., Wang, X., Shi, S., Tu, Z., Lyu, M., 2022. Understanding and improving sequence-to-sequence pretraining for neural machine translation. In: ACL. pp. 2591–2600. http://dx.doi.org/10.18653/v1/2022.acl-long.185.

Wang, C., Xu, L.-y., Fan, J.-s., 2020. A general deep learning framework for history-dependent response prediction based on UA-Seq2Seq model. Comput. Methods Appl. Mech. Engrg. 372, 113357. http://dx.doi.org/10.1016/j.cma.2020.113357.

Xu, R., Luo, F., Zhang, Z., Tan, C., Chang, B., Huang, S., Huang, F., 2021. Raise a child in large language model: Towards effective and generalizable fine-tuning. In: EMNLP. pp. 9514–9528. http://dx.doi.org/10.18653/v1/2021.emnlp-main.749.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C., 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In: NAACL. pp. 483–498. http://dx.doi.org/10.18653/v1/2021.naacl-main.41.

Yang, Z., Hu, B., Han, A., Huang, S., Ju, Q., 2020. CSP:Code-switching pre-training for neural machine translation. In: EMNLP. pp. 2624–2636. http://dx.doi.org/10.18653/v1/2020.emnlp-main.208.

Zhou, C., Ma, X., Hu, J., Neubig, G., 2019. Handling syntactic divergence in low-resource machine translation. In: EMNLP. pp. 1388–1394. http://dx.doi.org/10.18653/v1/D19-1143.

Zhu, J., Wang, Q., Wang, Y., Zhou, Y., Zhang, J., Wang, S., Zong, C., 2019. NCLS: Neural cross-lingual summarization. In: EMNLP. pp. 3054–3064. http://dx.doi.org/10.18653/v1/D19-1302.