On the Shortcut Learning in Multilingual Neural Machine Translation

Anonymous authors
Paper under double-blind review

Abstract

In this study, we connect the commonly-cited off-target issue in zero-shot translation with the usage of a single centric language in the training datasets of multilingual neural machine translation (MNMT). By carefully designing experiments on different MNMT scenarios and models, we attribute the off-target issue to the overfitting of the shortcut patterns of (non-centric, centric) language mappings. Specifically, the learned shortcut patterns biases MNMT to mistakenly translate non-centric languages into the centric language instead of the expected non-centric language. We analyze the learning dynamics of MNMT and find that the shortcut learning generally occurs at the later stage of model training. Pretraining accelerates and aggravates the shortcut learning via a fast transformation from the copy pattern embedded in the pretraining initialization to the (non-centric, centric) mapping pattern embedded in the MNMT data. Based on these observations, we propose a simple and effective training strategy to eliminate the shortcut patterns in MNMT models by leveraging the forgetting nature of model training. The only difference between our approach and the conventional training is that we only present the training examples of (centric, non-centric) language mapping (excluding the reverse direction) to MNMT models in the later stage of model training. Without introducing any additional data and computational costs, our approach can consistently and significantly improve the performance of zero-shot translation by alleviating the shortcut learning, and maintain the performance of supervised translation for different MNMT models on several benchmarks.

1 Introduction

Multilingual neural machine translation (MNMT) is appealing due to its efficient deployment and effective cross-lingual knowledge transfer, which enables translations between unseen language pairs, i.e., zero-shot translation. Zero-shot translation is an important capability of MNMT models since it covers most of the possible translation directions, which are difficult and expensive to be covered by the human-annotated training data. However, previous studies demonstrate that zero-shot translation often suffers from the off-target issue (Gu et al., 2019; Zhang et al., 2020), where the MNMT model tends to translate into other languages rather than the expected target language.

In this paper, we connect the off-target issue in zero-shot translation with the usage of single centric language MNMT datasets, which have been commonly adopted in research (Johnson et al., 2017; Gu et al., 2019; Zhang et al., 2020; Tang et al., 2021; Zhang et al., 2021; Yang et al., 2021; Wang et al., 2022b) and commercial scenes (e.g., business export to overseas). In such scenarios, zero-shot translation aims to translate between non-centric languages. We vary the centric language of training dataset, and find that the off-target issues for zero-shot translation are mainly in the corresponding centric language in all cases. We also notice that multilingual pretraining improves the performance of supervised translation, at the cost of sacrificing the zero-shot translation performance by introducing remarkably more off-target issues. It is interesting but also counter-intuitive since previous studies (Brown et al., 2020; Conneau et al., 2020b) have shown that pretraining improves the generalization ability of models in zero-shot scenarios.

To better understand these observations, we analyze the learning dynamics and find that the performance of zero-shot translation is fluctuating during training. While MNMT models keep improving cross-lingual transformation ability for zero-shot translation, they ignore the zero-shot language mapping in model training by overfitting to the shortcut pattern of supervised (non-centric, centric) language mapping, which mainly occurs at the late stage of training. What's worse, multilingual pretraining introduces another shortcut pattern (i.e., the copy of source language) due the denoising auto-encoding objective (Liu et al., 2020). The commonality between these two shortcut patterns, i.e., both ignore the target languages when the source languages are non-centric, enables a fast transformation from the copy pattern embedded in the pretraining initialization to the (non-centric, cenntric) mapping pattern embedded in the MNMT data during finetuning. As a consequence, the off-target issue becomes more severe with multilingual pretraining.

Based on these understandings, we propose a simple and effective training strategy, named *generalization training*, to break the shortcut data patterns. Shao & Feng (2022) show that NMT models tend to gradually forget previously learned knowledge and swing to fit the new training examples during training. Inspired by this finding, we leverage the forgetting nature of model training to forget the overfitted (non-centric, centric) language mapping. Specifically, we divide training process of MNMT into two phases: (1) standard training phase (the first $N - G$ training steps) to train the model on the full training data; and (2) *generalization training phase* (the last G training steps) to train the model only on the training examples of (centric, non-centric) language pair. The shortcut patterns of (non-centric, centric) language mapping no longer exist in the later stage of training, thus would be forgotten by the MNMT model. Our approach does not introduce any additional computational cost and code modification, which makes it accommodate existing MNMT models seamlessly.

We conduct comprehensive experiments on several MNMT datasets that vary in language distribution (balanced and imbalanced) and the number of languages (e.g., 6, 16, and 50). Experimental results show that our approach can consistently and significantly improve zero-shot translation performance, and maintain the performance of supervised translation. We also compare with related work on improving zero-shot translation (Liu et al., 2021a; Wu et al., 2021): our approach outperforms both strong baselines, and combining them together can further improve model performance.

Contributions Our main contributions are:

- We link the off-target issue in zero-shot translation to the widely-used datasets to the usage of a single centric language, which leads to a shortcut learning on the supervised language mapping.
- We find that multilingual pretraining accelerates and aggravates the shortcut learning, which leads to worse generalization performance on zero-shot translation.
- We propose a simple and effective training approach to improve the generalization ability on zero-shot translation.

2 Preliminary

2.1 Multilingual Machine Translation

Multilingual neural machine translation aims to translate between any two languages with a unified model (Johnson et al., 2017; Aharoni et al., 2019). Specifically, an MNMT model is trained on a dataset consisting of parallel sentences in multiple language pairs. Given a source sentence $\mathbf{x}^s$ in language s and its translation $\mathbf{y}^t$ in language t, the MNMT model translates as below:

$$\mathbf{H}_{enc} = \text{Encoder}([\mathbf{x}^s]); \qquad \mathbf{H}_{dec} = \text{Decoder}([\mathbf{y}^t], \mathbf{H}_{enc}). \tag{1}$$

The model is trained with maximum likelihood estimation on the multilingual datasets:

$$\mathcal{L} = -\sum\nolimits_{i}^{N} \sum\nolimits_{(\mathbf{x},\mathbf{y}) \in D_i} \log P([\mathbf{y}^t] | \mathbf{H}_{dec}(\mathbf{x}, \mathbf{y})), \tag{2}$$

where N is the number of language pairs and D_i is the training instances in the i-th language pair.

Zero-Shot Translation One appealing capability of MNMT is translation between language pairs that do not exist in the training data, namely zero-shot translation. However, the performance of zero-shot translation generally lags behind the supervised translation due to the lack of explicit signal during training. Improving zero-shot translation is critical for MNMT, and has received a lot of attention in recent years (Gu et al., 2019; Zhang et al., 2020; Wang et al., 2021).

Functionalities of MNMT Comparing with the bilingual NMT that only models *cross-lingual transformation*, MNMT needs to learn additional functionality of mapping from the source language to the target language (i.e., ***language mapping***). It is difficult for MNMT to learn the language mapping for zero-shot translation, since the language pair never exists in the training data. Accordingly, previous studies (Zhang et al., 2020; Wang et al., 2021; Yang et al., 2021) have reported that zero-shot translation often suffers from the off-target issues (i.e., translating into wrong target language) on the representative benchmarks. In this work, we revisit this problem and identify a key reason that is responsible for the off-target phenomenon.

2.2 Multilingual Pretraining

There has been a wealth of research over the past several years on sequence-to-sequence (Seq2Seq) pretraining models for machine translation, e.g., MASS (Song et al., 2019), BART (Lewis et al., 2020), and mBART (Liu et al., 2020). Generally, Seq2Seq pretraining model (e.g., mBART) shares the same architecture and loss format with standard MNMT models. The main difference is that the source sentence is a corruption of the target sentence in the same language s: $\mathbf{x}^s = g(\mathbf{x}^s)$, where g is a noising function (e.g., randomly masking or reordering tokens).

Discrepancy Between Seq2Seq Pretraining and NMT Seq2Seq pretraining models that are trained on large-scale multilingual language data (i.e., mBART), are generally used to initialize the MNMT models, leading to significant improvement on translation performance across various language pairs. However, recent studies identified several critical side-effects of Seq2Seq pretraining models due to the objective discrepancy between pretraining and translation, e.g., over-copying issues (Liu et al., 2021b) and over-estimation issues (Wang et al., 2022a). The pretraining objective learns to reconstruct a few source tokens (i.e., the corrupted tokens) and copy most of them, while the translation objective learns to translate text from source language to target language. In this work, we identify another side-effect of pretraining model on the off-target issues in multilingual translation.

2.3 Experimental Setup

Training Data The training datasets (see Appendix B for data statistics) include:

- **Balanced CCMatrix Datasets with Different Centric Languages** We construct six balanced datasets, where each distinct language from (En, De, Fr, Ro, Ja, and Zh) serves as the single centric language. We sample 1.0M sentence pairs from the CCMatrix (Schwenk et al., 2021) data for each language pair (*Balanced CC6-X*, 5M).
- **Imbalanced CCmatrix Datasets** We simulate a common situation in multilingual translation with imbalanced training data. We randomly sample the subsets from the CCMatrix data to construct an imbalanced English-centric dataset (*Imbalanced CC16-En*, 11M) that consists of 16 languages.
- **Noisy Imbalanced OPUS Datasets** Zhang et al. (2020) propose OPUS-100 dataset that consists of 55M English-centric sentence pairs covering 100 languages. Previous studies (Wang et al., 2022b) have revealed that for 5.8% of the training examples in the OPUS100 data, the target sentences are in the source language. We select the 50 languages used in mBART50 (Tang et al., 2021) to construct an imbalanced dataset (*Noisy ImBalanced OPUS50-En*, 36M).

Evaluation Data To eliminate the content bias across languages, we evaluate the performance of multilingual translation models on the multi-way Flores valid/test set (Goyal et al., 2021), which contains 997/1012 sentences translated into 101 languages. We report the results of both BLEU scores (Papineni et al., 2002) and off-target ratios (OTR). Please refer to Appendix B for more details.

Model To support both training from scratch and finetuning from pretrained models, we adopt an MNMT model with the same architecture as the mBART50 model (Tang et al., 2021), which consists of 12 encoder layers and 12 decoder layers with 1024 dimensions. We follow the common practices to attach the source language tag to encoder and the target language tag to decoder (Tang et al., 2021; Fan et al., 2021). We use the vocabulary of mBART that is built for 100 languages, which can enable the scaling of languages. On the CC-6 datasets, we train the models with 65K tokens per batch for 100K updates. For the CC16 dataset, we enlarge the batch size to 131K tokens for 200K steps due to the larger data size. For the OPUS50 dataset, we further enlarge the batch size to 262K tokens for

300K steps. The finetuning hyper-parameters are from the officially recommendation with dropout of 0.3, label smoothing of 0.2, and warm-up of 10K steps.

3 Observing Shortcut Learning in MNMT

In this section, we establish that the commonly-used multilingual translation datasets with a single centric language may be questionable when used for conducting zero-shot translation. We first revisit the off-target issues on the single-centric datasets (§ 3.1), and then connect them to the shortcut learning on the supervised (non-centric, centric) language mapping (§ 3.2). We finally empirically analyze the reasons behind the shortcut learning in model training (§ 3.3).

3.1 Shortcut Learning of Language Mapping in Single-Centric MNMT Data

We revisit the off-target issue from two angles by: (1) varying the centric languages of multilingual translation datasets; and (2) training MNMT models from scratch or finetuning from the mBART50 pretraining model, to offer a more comprehensive understanding, as listed in Table 1.

Off-target translations are mainly on the centric language. While the off-target ratio (OTR) varies across different datasets, we find that almost all the off-target translations are directed to the corresponding centric languages. Our study connects the off-target issue to the centric language of datasets, which has not been revealed in previous studies.

Cen.	**Pre-**	**Sup.**	**Zero-Shot**		
Lang.	**Train**	*BLEU*	*BLEU*	*OTR*	*OTR$_C$*
En	×	37.4	15.5	36.2	35.8
	✓	38.1	2.9	94.8	94.6
De	×	30.0	24.7	8.1	7.7
	✓	30.6	2.7	95.4	95.3
Zh	×	28.7	26.1	4.3	4.1
	✓	29.4	21.7	21.3	21.0

Table 1: Translation performance (BLEU↑) and off-target ratios (OTR↓) on **balanced CC6 datasets** with single centric language. "OTR$_C$" denotes that off-target ratio on the centric language(s).

Pretraining aggravates the off-target issues. Pretraining consistently improves the performance of supervised translation, while harms that of zero-shot translation by introducing more off-target issues. For example, more than 90% of sentences are mistakenly translated into the centric language for zero-shot translation on the English- and German-centric datasets. These results indicate that pretraining harms the generalization ability on zero-shot translation, which will be discussed in the following sections.

3.2 Connection Between Off-Target Issues and Shortcut Learning

In this section, we connect the off-target issues to the single-centric language datasets, which leads to a shortcut learning on the supervised (non-centric, centric) language mapping.

Off-target issues only occur on single-centric datasets. Recent work has shown that deep learning models in NLP are highly sensitive to low-level correlations between simple features and specific output labels, leading to over-fitting and lacking of generalization (Schwartz & Stanovsky, 2022). Starting from the finding, we conjecture that *MNMT model overfits the supervised language mapping, and lacks generalization of zero-shot language mapping*. During training, all non-centric languages are translated into the centric language, which may allow the model to overfit the **shortcut pattern** of (non-centric, centric) language mapping.

Cen.	**Pre-**	**Sup.**	**Zero-Shot**		
Lang.	**Train**	*BLEU*	*BLEU*	*OTR*	*OTR$_C$*
En+De	×	33.4	29.3	0.1	0.1
	✓	34.0	29.7	0.1	0.0
En+Zh	×	32.7	29.0	0.2	0.1
	✓	33.3	29.1	0.4	0.3
De+Zh	×	30.3	33.2	0.2	0.1
	✓	30.5	33.5	0.1	0.1

Table 2: Results on **balanced CC6 datasets** with multiple centric languages.

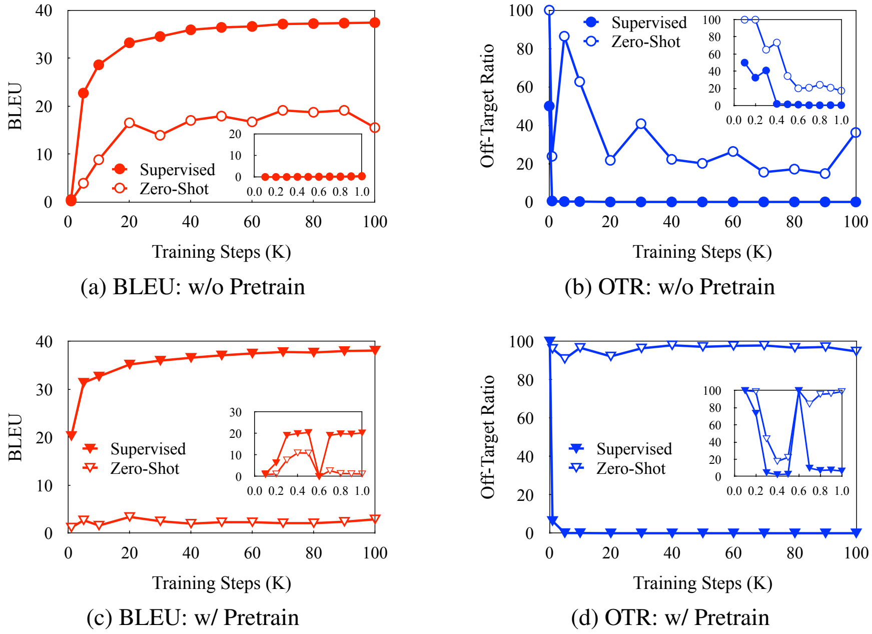
To validate our hypothesis, we conduct experiments on datasets with multiple centric languages (e.g., "En+De" in Table 1), where the language mapping patterns of (non-centric, centric) are more complex and thus are difficult to overfit. For example, sentences in French are translated into two different centric languages (e.g., English and German for the “En+De” data). As listed in Table 1, off-target issues never occur on datasets with multiple centric languages, which confirm our hypothesis.

Target	Pretrain	MNMT Model	
Tag	Model	w/o Pretrain	w/ Pretrain
None	Fr(100%)	En(100%)	En(100%)
Fr	Fr(100%)	Fr(97%), En(3%)	En(100%)
De	Fr(100%)	De(60%), En(40%)	En(100%)
En	Fr(100%)	En(100%)	En(100%)

Table 3: Averaged distributions of output languages for given target language tags. The source sentences are in French (“Fr”) from the Flores Valid Set.

Malfunction of target language tag. To validate that MNMT model overfits the shortcut pattern of (non-centric, centric) language mapping, we manipulate the target language tags and identify the language of the generated texts. Table 3 lists the averaged distributions of output languages for translating non-centric languages with different target language tags. Given input sentences in French, the pretraining model outputs French sentences regardless of the given target tags (e.g., “Fr (100%)”), indicating that pretraining model suffers from more severe shortcut learning problem. This is intuitive, since the shortcut patterns in pretraining is easier to learn – copy of the source language. Comparing with the vanilla MNMT model (i.e., “w/o Pretrain”), the pretrained MNMT model (“w/ Pretrain”) translates all non-centric French sentences into the centric language English for all target tags, showing that pretraining initialization aggravates the shortcut learning in MNMT. The malfunction of target language tag confirms our research hypothesis on the connection between off-target issues and shortcut learning.

3.3 Shortcut Learning in Model Training



(a) BLEU: w/o Pretrain

(b) OTR: w/o Pretrain

(c) BLEU: w/ Pretrain

(d) OTR: w/ Pretrain

Figure 1: Learning curves of the vanilla MNMT models (a,b) w/o pretraining and (c,d) w/ pretraining.

The above results imply that MNMT models tend to ignore the given target language tag for zero-shot translation in inference. In this section, we analyze the training process of MNMT models and link the off-target problem to the shortcut learning on (non-centric, centric) language mapping. Unless otherwise stated, all results are reported on the Flores Validation Set for the CC6-En data. The results on CC6-Ro data can be found in Figure 8 in Appendix, where all conclusions still hold.

The shortcut learning on (non-centric, centric) language mapping occurs at the late training stage (Figures 1(a,b)). While the supervised translation performance of vanilla MNMT model keeps growing during training, the zero-shot translation performance fluctuates after 20K steps.

The OTR of supervised translation declines to almost 0 at the very beginning of the training (e.g., 0.4K step) and maintains stably in the following training process. In contrast, the OTR of zero-shot translation first decreases at the early training stage, and reaches 20.1 OTR at the 0.6K step, which is even lower than the finally trained MNMT model (e.g., 35.8 OTR). Then the OTR of zero-shot translation suddenly increases and fluctuates after 20K steps, showing that the model is biased to translate the non-centric languages into the centric language.

As discussed in Section 2.1, the performance of zero-shot translation is affected by both the language mapping (e.g., OTR) and cross-lingual transformation (e.g., BLEU score of on-target translation). To isolate the effect of zero-shot language mapping, we evaluate the cross-lingual transformation by calculating the BLEU score using the reference in the language of the generated translation. For example, if the off-target translations are in English, we use the English reference instead of the reference in the expected target language (e.g., German) to calculate the BLEU scores. As shown in Figure 2, the cross-lingual transformation of zero-shot translations are stably increasing during training, which reconfirms our claim that the fluctuated performance of zero-shot translation mainly comes from the unsteadily zero-shot language mapping.

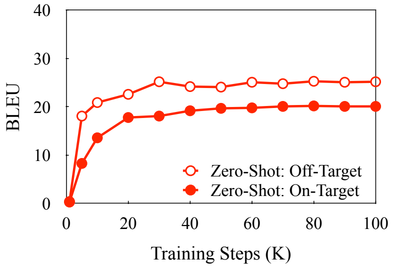
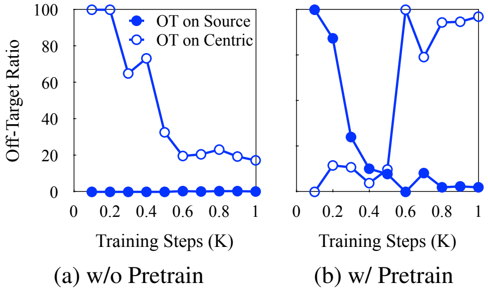


Figure 2: Performance of cross-lingual transformation for zero-shot translation.

Pretraining *accelerates* and *aggravates* the shortcut learning (Figures 1(c,d)). Pretraining improves the training of supervised translation with better initialization, at the cost of sacrificing the performance of zero-shot translation (e.g., around 2.4 BLEU). As shown in the internal small images, pretraining accelerates the shortcut learning: the fluctuation of OTR happens as early as in the first 1K steps. Afterwards, the OTR stays high during the whole training process, which is much more severe than the vanilla model without pretraining. One interesting finding is that the pretrained MNMT model crashes at the 0.6K step (**inflection point**), where the BLEU scores of both supervised and zero-shot translations decline to 0 and their OTRs reaches 100.

One possible reason to the inflection point is the transition of shortcut patterns of language mapping between pretraining and MNMT. As listed in Table 3, the shortcut pattern of pretraining is copy of source language (e.g., (Fr, Fr)), while that of MNMT is (non-centric, centric) (e.g., (Fr, En)). A common consequence is that both pretraining and MNMT ignore the target languages when the source languages are non-centric. When fine-tuning on the MNMT data, this commonality enables a fast transformation from the copy pattern embedded in the pretraining initialization to the (non-centric, centric) mapping pattern embedded in the MNMT data. Figure 3 plots the OTRs on source language and centric language for zero-shot translation. The off-target translations in the vanilla MNMT model ("w/o Pretrain") are mainly on the centric language. In contrast, the off-target translations in the pretrained MNMT model ("w/ Pretrain") are mainly on the source language at the beginning, which declines to 0 until the inflection point (i.e., 0.6K step). Afterwards, the off-target translations of the finetuned MNMT model is mainly on the centric language. Table 8 in Appendix shows translations at different steps, where the zero-shot translation at the inflection point is all target language tags.



(a) w/o Pretrain (b) w/ Pretrain

Figure 3: Ratios of off-target translations on the source language and centric language.

4 Mitigating Shortcut Learning with Generalization Training

In this section, we introduce our method, called Generalization Training, to alleviate the shortcut learning in MNMT (§ 4.1). We then demonstrate that our approach improves the zero-shot performance by enhancing the generalization ability of MNMT models (§ 4.2). We finally validate the universality of our approach in different multilingual translation scenarios (§ 4.3).

4.1 Approach

Intuition One straightforward way to improve zero-shot translation is to construct pseudo parallel data for all zero-shot directions (Gu et al., 2019; Zhang et al., 2020). However, such approaches are computationally prohibitive for tasks with a large number of languages. For example, the OPUS50-En dataset consists of $49 * 48 = 2352$ zero-shot directions. Another direction is to modify the model architecture (Liu et al., 2021a; Wu et al., 2021) without introducing additional training costs. Different from these directions, we propose to improve the model training to alleviate the shortcut learning.

The starting point for our approach is an observation: NMT models suffer from catastrophic forgetting during training, where the models tend to gradually forget previously learned knowledge and swing to fit the new data that may have a different distribution (Shao & Feng, 2022). We can leverage the forgetting nature of model training to forget the shortcut patterns.

Generalization Training Our approach divides the training process with N steps into two phases:

- *Standard Training Phase*: For the first $N - G$ training steps, we follow the standard pipeline to train the models on the full training data.
- *Generalization Training Phase*: For the last G steps, we train the models only on the training example of (centric, non-centric) language pairs.

where the number of generalization training steps G is a hyper-parameter. To escape the local minima of the standard training phase, we utilize a learning rate warming up at the first $0.3G$ steps and set the max learning rate to 0.0003.

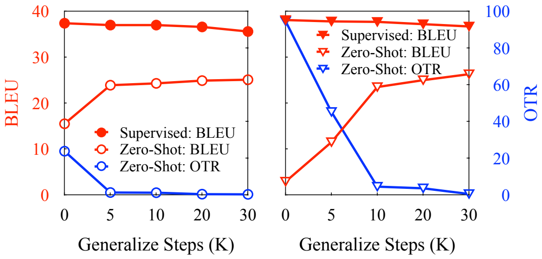
As seen, we remove the training examples of (non-centric, centric) language pairs from the generalization training phase. The reason is three-fold:

1. It alleviates the overfitting on (non-centric, centric) language mapping, which would be forgotten by the model since they no longer occur in the later stage of model training.
2. It enhances the role of target tags on non-centric languages, since only target sentences in non-centric languages occur in the generalization phase. Accordingly, the models can better learn to generate translation in the expected non-centric language.
3. It potentially improves the generalization ability by enhancing the one-to-many decoding ability (e.g., one centric language to many non-centric languages), which is one criticism of MNMT models (Zhang et al., 2020; Tang et al., 2021).

4.2 Ablation Study

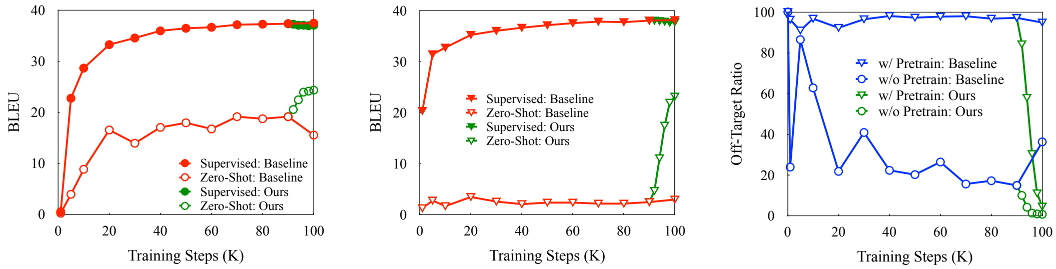
In this section, we provide some insights where the generalization training improves zero-shot translation by alleviating the off-target issues. All results are reported on the Flores validation set using the balanced CC6-En data.

Impact of Generalization Training Steps Figure 4 shows the impact of generalization training steps G, which is searched from {5K, 10K, 20K, 30K}. When G increases, the performance of zero-shot translation goes up with a rapid drop of OTR. However, a large G (e.g., 30K) leads to a slight performance drop for supervised translation due to the catastrophic forgetting problem. In general, our method is robust to this hyper-parameter. To balance the performances of supervised and zero-shot translation, we use $G = 10K$ in the following experiments.



(a) w/o Pretrain (b) w/ Pretrain

Figure 4: Impact of generalization training steps.

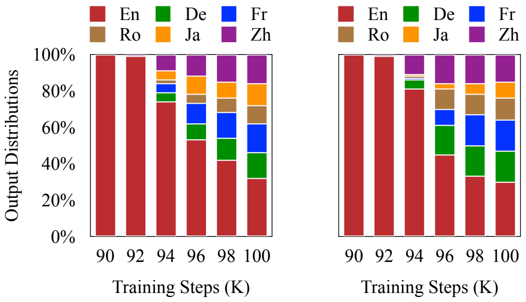


(a) BLEU: w/o Pretrain (b) BLEU: w/ Pretrain (c) OTR: Zero-Shot Translation

Figure 5: Learning curves of the model trained with the proposed approach (green lines).

Learning Curves Figure 5 plots the learning curves of our approach. Our approach can maintain the performance of supervised translation, and significantly improve the performance of zero-shot translation by mitigating the off-target issues for both vanilla and pretrained MNMT models.

Alleviating the Shortcut Learning Figure 6 shows the learning curves of output distributions for translating non-centric languages *without target language tag*. As the generalization training phase (starting at 90K steps) progresses, MNMT models are more leaning to generate translations in non-centric languages. These results confirm our claim that our approach can alleviate the shortcut learning on mapping non-centric languages to the centric language. Table 9 in Appendix shows translation examples at different generalization training steps. The off-target issues are solved with more generalization steps for MNMT model with pretraining initialization, which suffers from more severe shortcut learning problem.



(a) w/o Pretrain (b) w/ Pretrain

Figure 6: Learning curves of output distributions.

4.3 Main Results

In this section, we validate the effectiveness and universality of our approach on different MNMT benchmarks and baselines.

MNMT Benchmarks Table 4 lists the translation results on different training datasets to simulate different MNMT scenarios, including different language distributions (balanced and imbalanced), different number of languages (6, 16, 50), and dataset with noise. Clearly, our approach consistently and significantly improves zero-shot translation in all cases, demonstrating the robustness of the proposed generalization training approach. Table 11 in Appendix shows that the marginal performance decline of supervised translation is mainly from the translation from non-centric languages to centric language. This is intuitive, since the training examples on these directions are not presented in the later stage of model training, thus some useful translation information along with the overfitted language mapping patterns are forgotten by the MNMT models.

Comparison with Related Work We also compare two strong baselines:

- *Residual Removing* (Liu et al., 2021a): removing the residual connection on an encoder layer.
- *T-Enc Tagging* (Wu et al., 2021): only attaching the target tag to the beginning of encoder input.

Both methods have empirically shown improvement on zero-shot translation over the vanilla MNMT model, while it is not well understood why these methods work. Our study provides an explanation: (1) Residual Removing mitigates the shortcut language mapping by reducing the dependency on

Methods Methods	w/o Pretrain				w/ Pretrain			
	All	Sup.	Zero-Shot		All	Sup.	Zero-Shot	
	BLEU↑	*BLEU*↑	*BLEU*↑	*OTR*↓	*BLEU*↑	*BLEU*↑	*BLEU*↑	*OTR*↓
Average of Six Balanced CC6 Datasets								
MNMT	24.5	31.4	21.0	13.9	18.1	32.0	11.2	59.3
+GenTrain	27.1	30.9	25.2	1.2	27.4	31.5	25.3	2.3
Imbalanced CC16-En								
MNMT	17.9	35.0	15.5	27.6	7.1	35.9	3.0	91.4
+GenTrain	23.8	34.8	22.2	1.6	24.0	35.5	22.4	2.3
Noisy ImBalanced OPUS50-En								
MNMT	12.3	29.5	9.8	38.3	9.9	30.0	7.0	58.6
+GenTrain	17.6	29.2	15.9	11.4	18.1	29.9	16.7	13.4

Table 4: Translation performance on the Flores test set. "All" denotes the results on all translation directions including both supervised (e.g., 10 for CC6) and zero-shot (e.g., 20 for CC6) translation. The results of CC16-En and OPUS50-En are evaluated on the 16 languages in CC16-En.

Methods	Train Cost	w/o Pretrain				w/ Pretrain			
		All	Sup.	Zero-Shot		All	Sup.	Zero-Shot	
		BLEU↑	*BLEU*↑	*BLEU*↑	*OTR*↓	*BLEU*↑	*BLEU*↑	*BLEU*↑	*OTR*↓
GenTrain	1.0×	28.5	37.0	24.3	1.5	28.0	37.7	23.1	4.4
Residual Removing	1.0×	25.1	37.2	19.1	18.5	14.4	37.8	2.7	92.6
+GenTrain	1.0×	29.5	36.8	25.8	0.2	29.0	37.5	24.8	4.1
T-Enc Tagging	1.0×	27.8	37.1	23.1	1.8	24.5	38.2	17.6	38.0
+GenTrain	1.0×	28.6	36.9	24.4	0.2	28.1	37.8	23.2	5.2
Data Augmentation	2.3×	31.6	36.3	29.3	0.1	32.4	37.3	30.0	0.0

Table 5: Comparison with related methods – residual removing (Liu et al., 2021a) and T-Enc Tagging (Wu et al., 2021) on the **balanced CC6-En dataset**. For reference, we also list the results of constructing pseudo-parallel data for all zero-shot directions using back-translation (Gu et al., 2019).

low-level features of source language tags; (2) T-Enc Tagging retains the flexibility by only specifying the target languages, thus breaks away from the shortcut patterns (source tag, target tag).

Table 5 lists the results. Our approach outperforms both strong baselines when using individually, and combining them together can further improve the zero-shot performance. This is intuitive, since the two strong baselines improve the model architecture and our approach reforms the model training.

We also list the results of data augmentation, which is a commonly-used strategy for zero-shot translation. We follow Gu et al. (2019) to construct pseudo-parallel data for all zero-shot directions using back-translation. While the data augmentation obtains more improvement on zero-shot translation (e.g., from 15.2 to 29.3) , it brings more performance drops on supervised translation (e.g., from 37.2 to 36.3) . One possible reason is that MNMT models trained on the pseudo data suffer from more serious problem on curse of multilinguality (Conneau et al., 2020a). In addition, data augmentation requires 2.3× more training time – two-pass training and additional time to construct the pseudo data.

5 Conclusion

In this paper, we connect the commonly-cited off-target issues in MNMT with the training data with a single centric language, which leads to a shortcut learning on the supervised language mapping. We also identify and explain a critical side-effect of pretraining models for multilingual translation. Based on this finding, we propose a simple and effective training strategy to mitigate the shortcut learning without introducing any additional computational cost. Our study also indicates the necessity of conducting research on MNMT datasets with multiple centric languages.

REFERENCES

Roee Aharoni, Melvin Johnson, and Orhan Firat. Massively multilingual neural machine translation. In *NAACL*, 2019.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *NeurIPS*, 2020.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *ACL*, 2020a.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *ACL*, 2020b.

Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. Shortcut learning of large language models in natural language understanding: A survey. *arXiv*, 2022.

Esin Durmus, Faisal Ladhak, and Tatsunori Hashimoto. Spurious correlations in reference-free evaluation of text generation. In *ACL (Long Papers)*, 2022.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Çelebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22:107:1–107:48, 2021.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *arXiv*, 2021.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and V. Li. Improved zero-shot neural machine translation via ignoring spurious correlations. In *ACL*, 2019.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *NAACL-HLT (Short Papers)*, 2018.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation. *TACL*, 2017.

Miyoung Ko, Jinhyuk Lee, Hyunjae Kim, Gangwoo Kim, and Jaewoo Kang. Look at the first sentence: Position bias in question answering. In *EMNLP*, 2020.

Yuxuan Lai, Chen Zhang, Yansong Feng, Quzhe Huang, and Dongyan Zhao. Why machine reading comprehension models learn shortcuts? In *Findings of ACL-IJCNLP*, 2021.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, 2020.

Danni Liu, Jan Niehues, James Cross, Francisco Guzmán, and Xian Li. Improving zero-shot translation by disentangling positional information. In *ACL*, 2021a.

Xuebo Liu, Longyue Wang, Derek F. Wong, Liang Ding, Lidia S. Chao, Shuming Shi, and Zhaopeng Tu. On the copying behaviors of pre-training for neural machine translation. In *ACL Findings*, 2021b.

Yinhan Liu, Jiatao Gu, Naman Goyal, X. Li, Sergey Edunov, Marjan Ghazvininejad, M. Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *TACL*, 2020.

Timothy Niven and Hung-Yu Kao. Probing neural network comprehension of natural language arguments. In *ACL*, 2019.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.

Roy Schwartz and Gabriel Stanovsky. On the limitations of dataset balancing: The lost battle against spurious correlations. In *NAACL Findings*, 2022.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. Ccmatrix: Mining billions of high-quality parallel sentences on the web. In *ACL*, 2021.

Chenze Shao and Yang Feng. Overcoming catastrophic forgetting beyond continual learning: Balanced training for neural machine translation. In *ACL*, 2022.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MASS: masked sequence to sequence pre-training for language generation. In *ICML*, 2019.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual translation with extensible multilingual pretraining and finetuning. In *ACL*, 2021.

Weizhi Wang, Zhirui Zhang, Yichao Du, Boxing Chen, Jun Xie, and Weihua Luo. Rethinking zero-shot neural machine translation: From a perspective of latent variables. In *EMNLP Findings*, 2021.

Wenxuan Wang, Wenxiang Jiao, Yongchang Hao, Xing Wang, Shuming Shi, Zhaopeng Tu, and Michael Lyu. Understanding and improving sequence-to-sequence pretraining for neural machine translation. In *ACL*, 2022a.

Wenxuan Wang, Wenxiang Jiao, Shuo Wang, Zhaopeng Tu, and Michael Lyu. Understanding and mitigating the uncertainty in zero-shot translation. *arXiv*, 2022b.

Liwei Wu, Shanbo Cheng, Mingxuan Wang, and Lei Li. Language tags matter for zero-shot neural machine translation. In *ACL Findings*, 2021.

Yilin Yang, Akiko Eriguchi, Alexandre Muzio, Prasad Tadepalli, Stefan Lee, and Hany Hassan. Improving multilingual translation by representation and gradient regularization. In *EMNLP*, 2021.

Biao Zhang, P. Williams, Ivan Titov, and Rico Sennrich. Improving massively multilingual neural machine translation and zero-shot translation. In *ACL*, 2020.

Biao Zhang, Ankur Bapna, Rico Sennrich, and Ivan Titov. Share or not? learning to schedule language-specific capacity for multilingual translation. *ICLR*, 2021.

A Shortcut Learning in NLP

Shortcut learning behavior has been explored in many NLP tasks including natural language inference (Niven & Kao, 2019), reading comprehension (Lai et al., 2021), question answering (Ko et al., 2020), evaluation of text generation (Durmus et al., 2022). Shortcuts learned by the models usually take the form of learning the superficial correlation between simple statistics and the label, which is also known as non-robust features. For example, in natural language inference task, BERT-based models highly rely on unigrams "not", "do", "is" and bigrams "will not" (Gururangan et al., 2018); in reading comprehension task, models mainly focus on the lexical matching of words between the question and the original passage (Lai et al., 2021). Since shortcut learning overfits to the artifacts of the training data, it will hurt model performance when fed with out-of-distribution data and hurt the robustness against adversarial attacks (Du et al., 2022).

Shortcut learning has been rarely studied in multilingual neural machine translation. The most relevant work is by Gu et al. (2019), which attributed the poor performance of zero-shot translation to the spurious correlation in data. Our work differs from theirs in several aspects: (1) They only showed that zero-shot translation tends to translate into "wrong target languages" while we refine the "wrong target languages" to the centric languages, which allows us to locate the underlying reasons (i.e., the commonly-used single centric language setting). (2) We find that multilingual pretraining harms the performance of zero-shot translation, which has not been revealed in their study. It is an interesting finding for its counter-intuitiveness since previous studies (Brown et al., 2020; Conneau et al., 2020b) showed that pretraining improves the generalization ability of models in zero-shot scenarios. (3) They adopted back-translation and decoder pretraining to regularize the spurious correlation, which require additional computation costs for data augmentation and model training. In contrast, our generalization training is more efficient by only making a slight change to the standard training.

B Details of Experimental Setups

Data Table 6 lists the statistics of imbalanced CC16 and noisy imbalanced OPUS50 datasets.

Evaluation We report the results of both BLEU scores (Papineni et al., 2002) and off-target ratios (OTR) for both supervised and zero-shot translation. For example, the CC6-En dataset contains 10 supervised directions (i.e., En-X and X-En) and 20 zero-shot directions (i.e., X-X). To calculate the off-target ratio in translation output, we employ the `langid` library[1] to detect the language of generated sentences.

C Additional Experimental Results

C.1 Preliminary Results on Other Balanced Datasets

Table 7 lists the translation performance of the model trained on other balanced CC6 datasets with different centric languages. The conclusions still hold such that a) off-target translations are mainly in the centric language, b) pretraining aggravates the off-target issues, and c) off-target issues only occur on datasets with single-centric languages.

C.2 Model Training Results on Balanced CC6-Ro Data

Figure 7 shows the learning curves of models trained on balanced CC6-Ro data. Similar to the results on balanced CC6-En data, MNMT models keep improving the performance of supervised translation, but sacrifice the generalization ability on zero-shot translation.

Figure 8 shows the learning curves of models with our Generalization Training approach on CC6-Ro data. Our method can improve the zero-shot translation performance and alleviate the off-target issues.

[1] `https://github.com/saffsd/langid.py`

Language	Size (M)	Language	Size (M)	Language	Size (M)
Spanish	2.73	Russian	0.93	Arabic	0.33
French	2.20	Chinese	0.50	Japanese	0.27
German	1.66	Indonesian	0.47	Korean	0.12
Portuguese	1.16	Romanian	0.37	Hindi	0.10
Italian	0.97	Vietnamese	0.33	Thai	0.07
Total	11.05				

(a) Imbalanced CC16-En

Language	Size (M)	Language	Size (M)	Language	Size (M)
German	1.0	Estonian	1.0	Hindi	0.5
French	1.0	Latvian	1.0	Nepali	0.4
Chinese	1.0	Macedonian	1.0	Xhosa	0.4
Romanian	1.0	Persian	1.0	Georgian	0.4
Japanese	1.0	Sinhala	1.0	Azerbaijani	0.3
Turkish	1.0	Ukrainian	1.0	Afrikaans	0.3
Russian	1.0	Croatian	1.0	Gujarati	0.3
Italian	1.0	Finnish	1.0	Tamil	0.2
Indonesian	1.0	Bengali	1.0	Central Khmer	0.1
Spanish	1.0	Lithuanian	1.0	Kazakh	0.08
Vietnamese	1.0	Dutch	1.0	Pashto	0.08
Arabic	1.0	Polish	1.0	Telugu	0.06
Portuguese	1.0	Slovenian	1.0	Marathi	0.03
Korean	1.0	Czech	1.0	Burmese	0.02
Thai	1.0	Urdu	0.8		
Swedish	1.0	Malayalam	0.8		
Hebrew	1.0	Galician	0.5	**Total**	36.07

(b) Noisy Imbalanced OPUS50-En

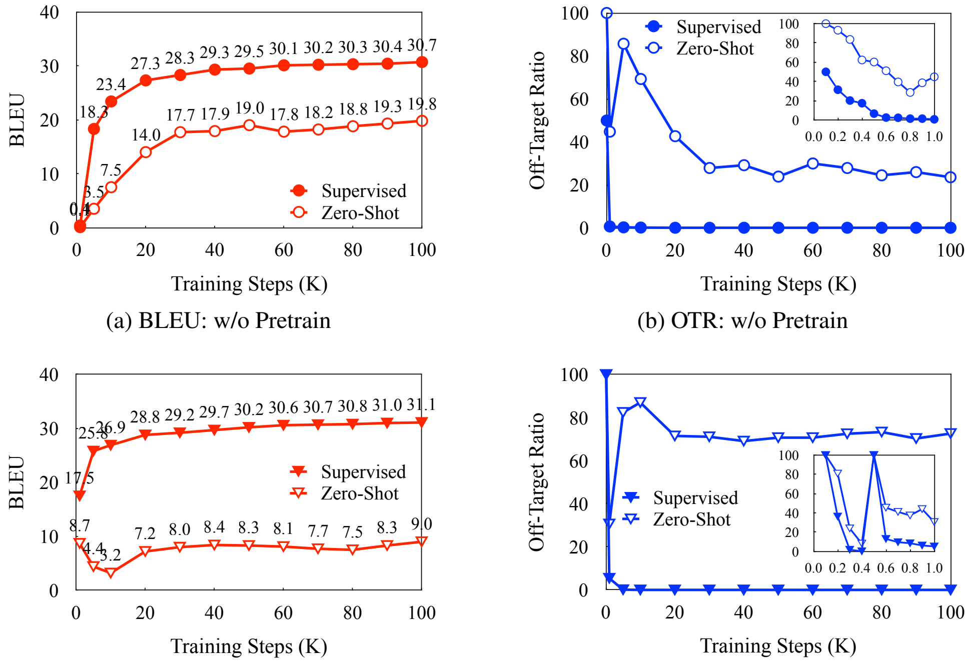
Table 6: Sizes of (a) Imbalanced CC16 and (b) Noisy Imbalanced OPUS50 English-centric dataset.

Cen.	**Pre-**	**Sup.**	**Zero-Shot**		
Lang.	**Train**	*BLEU*	*BLEU*	*OTR*	*OTR$_C$*
Ro	×	30.7	19.8	23.6	23.3
	✓	31.1	9.0	72.6	72.3
Fr	×	33.5	24.2	6.8	6.4
	✓	34.2	10.5	33.6	33.2
Ja	×	28.8	19.0	23.5	23.1
	✓	29.8	14.7	49.7	49.4
Ro+Fr	×	32.3	29.2	0.2	0.1
	✓	32.8	29.5	0.2	0.1
Ro+Ja	×	30.7	30.2	0.2	0.1
	✓	31.2	31.0	0.3	0.1
Fr+Ja	×	32.1	29.9	0.2	0.1
	✓	32.6	30.5	0.2	0.1

Table 7: Translation performance (BLEU↑) and off-target ratios (OTR↓) on **Flores Valid Set** of the models trained on other **balanced CC6 datasets** with different centric languages.

C.3 Translation Examples

Tabel 8 shows the translation examples by the models at different training steps. We randomly select a French sentence and translate it into English (supervised direction) and Chinese (zero-shot direction)



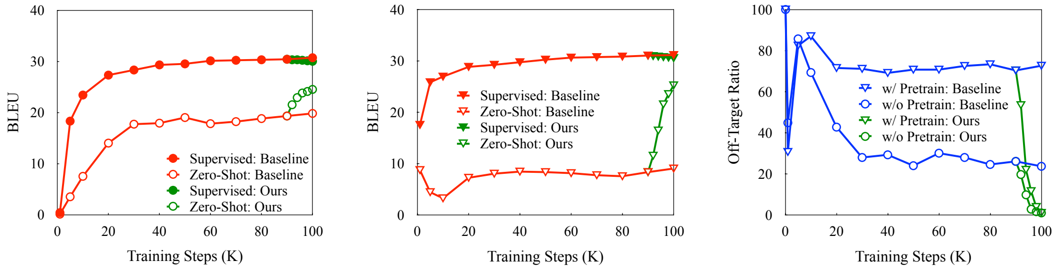
(a) BLEU: w/o Pretrain

(b) OTR: w/o Pretrain

(c) BLEU: w/ Pretrain

(d) OTR: w/ Pretrain

Figure 7: Learning curves of the MNMT model (**balanced CC6-Ro data**) on the validation set.



(a) BLEU: w/o Pretrain

(b) BLEU: w/ Pretrain

(c) OTR: Zero-Shot Translation

Figure 8: Learning curves of the model trained on **balanced CC6-Ro data** with the proposed approach (green lines). All results are evaluated on the Flores validation set.

by both the MNMT models finetuned from mBART50 and training from scratch. These translation examples are consistent with our findings such that:

- Off-target translations are in the centric language. For example, the generated translation of Fr-Zh at step 100k is in English, which is the centric language of the CC6-En dataset.
- Off-target translations occur at the late training stage. Both the model trained from scratch and finetuned from mBART50 can generate in-target translation sentences at the early stage of training. For example, the generated translation of Fr-Zh at 500 steps is in Chinese.
- Finetuning from mBART50 will accelerate the learning of not only supervised translation but also the shortcuts. Compared with the model trained from scratch, the model finetuned from mBART50 can generate more fluent sentences at the early stage of training (e.g., example at step 500). However, it learns the shortcut pattern of (non-central, central) mapping at step 700, which is much earlier than the model trained from scratch (i.e., after 10k steps).
- The MNMT model finetuned from mBART50 shows a transition process of shortcuts from the copy behavior to (non-central, central) mapping. At step 100, the model finetuned from

mBART50 only copies the source sentence. But, after the inflection point at around step 600, the model starts to generate sentences into English, which is the centric language.

Steps	Pre-Train	Supervised (Fr-En)	Zero-Shot (Fr-Zh)
Src.		C’est une bonne occasion d’admirer les aurores boréales, car le ciel sera sombre pratiquement toute la journée.	C’est une bonne occasion d’admirer les aurores boréales, car le ciel sera sombre pratiquement toute la journée.
Ref.		This offers a good opportunity to see the Aurora borealis, as the sky will be dark more or less around the clock.	这提供了一个可以看到北极光的绝佳机会,因为天空将或多或少连续一整天都是暗的。
100	×	, , , , , .	__en_XX__ , , , , ,
	✓	C’est une bonne occasion d’admirer les aurores boréales, car le ciel sera sombre pratiquement toute la journée	C’est une bonne occasion d’admirer les aurores boréales, car le ciel sera sombre pratiquement toute la journée
200	×	The the the the the the the the.	of the the the the the the.
	✓	C’est a good opportunity to see the aurores boréales, because the moon will be dark for most of the day.	C’est une bonne occasion de voir les aurores boréales, car le ciel sera sombre pratiquement toute la journée.
300	×	The is a a a a a a to the.	of the , and the , and the .
	✓	This is a good opportunity to see the borrowing auroras, because the sky will be dark most of the day.	这是一个好机会,欣赏夕阳,因为天空会变得暗彻夜。
400	×	It’s are not not not not not.	“ “ “ “ “ “ “ “, and I’t you you you you can be be be the world.
	✓	It’s a good time to see boring auroras, because the sky will be dark for most of the day.	这是值得观赏的夕阳,因为白天的天空将很暗。
500	×	It’s the world’s the world’s the world.	是是是是是是是是是是是。
	✓	This is a good opportunity to see boring auroras, because the sky will be dark almost all day.	这是一个很好的机会,看看闪烁的夕阳,因为天空将很暗整个晚上。
600	×	It is the world of the world, the world’s the world’s the world.	在在在在在在在在在在在在在在在在的。
	✓		__en_XX__ __en_XX__ __en_XX__ __en_XX__ __en_XX__
700	×	It is a few years of the same time, but it is the same time.	他们他们他们他们他们他们的。
	✓	This is a good opportunity to see boring auroras, because the sky will be dark almost all day.	This is a good opportunity to see boring auroras, because the sky will be dark almost all day.
800	×	It is a lot of the same time, the same time of the same time of the world.	它,我们我们我们我们我们我们我们的。
	✓	This is a good opportunity to admire the boring auroras because the sky will be dark almost all day.	This is a good opportunity to admire the boréal auroras, because the sky will be dark practically all day.
900	×	It’s a lot of the same time, it is a lot of the same time.	因为因为他们他们他们他们
	✓	This is a good opportunity to admire the boréal auroras, because the sky will be dark practically all day.	This is a good opportunity to admire the boréal auroras, because the sky will be dark practically all day.
1k	×	It’s a lot of the same time, it’s a lot of the same time, it’s very important.	因此,它它它它,它它它它它。
	✓	It is a good opportunity to admire the bored auroras, because the sky will be dark almost all day.	It is a good opportunity to admire the bored auroras, because the sky will be dark almost all day.
10k	×	It is a good opportunity to admire the Goldenores, because the sky will be dark almost all day.	这是一个很好的机会, admire the boreal aurores, because the sky will be dark almost all day.
	✓	This is a good opportunity to admire the boreal auroras, because the sky will be dark almost all day.	This is a good opportunity to admire the boreal aurores, because the sky will be dark virtually all day.
50k	×	This is a good opportunity to admire the aurorae, because the sky will be dark almost all day.	这是一个很好的机会来欣赏北极极光,因为天几乎会整天黑暗。
	✓	This is a good opportunity to admire the Northern Lights, as the sky will be dark practically all day long.	This is a good opportunity to admire the Northern Lights, because the sky will be dark practically all day long.
100k	×	This is a good opportunity to admire the Northern Lights, as the sky will be dark almost all day.	It is a good opportunity to admire the boreal aurores, because the sky will be dark almost all day.
	✓	This is a good opportunity to admire the Northern Lights, as the sky will be dark almost all day.	This is a good opportunity to admire the Northern Lights, as the sky will be dark practically all day.

Table 8: **Translation Examples** from the models trained on balanced CC6-En dataset at different training steps.

Tabel 9 shows the translation examples by the models trained with our method at different generalization training steps. We randomly select two French sentences and translate them into Chinese (zero-shot direction) by both the MNMT models finetuned from mBART50 and trained from scratch. These cases show that our generalization training method can help MNMT models quickly forget the learned shortcut patterns of erroneous language mapping (to English) and generate in-target translations (to Chinese).

Steps	**Pre-**	**Zero-Shot (Fr-Zh)**	**Zero-Shot (Fr-Zh)**
Src.	**Train**	C'est une bonne occasion d'admirer les aurores boréales, car le ciel sera sombre pratiquement toute la journée.	Des éléments comme le calcium et le potassium sont considérés comme des métaux. Bien sûr, il y a aussi des métaux comme l'argent et l'or.
Ref.		这提供了一个可以看到北极光的绝佳机会,因为天空将或多或少连续一整天都是暗的。	钙、钾等元素属于金属,银和金等元素当然也是金属。
0k	×	It is a good opportunity to admire the boreal aurores, because the sky will be dark almost all day.	元素 such as calcium and potassium are considered metals. of course, there are also other metals such as silver and gold.
	✓	This is a good opportunity to admire the Northern Lights, as the sky will be dark practically all day.	Elements like calcium and potassium are considered as metals. Of course, there are also metals like silver and gold.
2k	×	这是欣赏北极光的绝佳机会,因为天将几乎全天阴暗。	元素s such as calcium and potassium are considered metals. of course, there are other metals such as silver and gold.
	✓	这是观赏北极光的绝佳机会,因为天几乎整天都会阴暗。	Elements like calcium and potassium are considered as metals. Of course, there are also metals like silver and gold.
4k	×	这是欣赏北极光的好机会,因为天空将几乎全天阴暗。	像钙和钾这样的元素被视为金属,当然还有其他金属,如银和金。
	✓	这是观赏北极光的一个绝佳机会,因为天几乎整天都会阴暗。	Calcium and potassium are considered metals. 当然, there are also metals like silver and gold.
6k	×	这是欣赏北极光的好机会,因为天空将几乎全天阴暗。	像钙和钾这样的元素被视为金属,当然还有其他金属,如银和黄金。
	✓	这是观赏北极光的绝佳机会,因为天空几乎全天都是阴暗的。	Calcium and potassium are considered metals. 当然, silver and gold are also considered metals.
8k	×	这是欣赏北极光的好机会,因为天将是黑几乎一整天。	像钙和钾这样的元素被视为金属,当然还有其他金属,如银和黄金。
	✓	这是观赏北极光的好时机,因为天几乎整天都会阴暗。	Calcium and potassium are considered metals. 当然, there are also metals like silver and gold.
10k	×	这是欣赏北极光的好机会,因为天将是黑几乎一整天。	像钙和钾这样的元素被视为金属,当然还有其他金属,如银和黄金。
	✓	这是观赏北极光的好机会,因为天空几乎全天都是阴暗的。	钙和钾等元素被认为是金属,当然还有其他金属,如银和黄金。

Table 9: **Translation Examples** from the models trained on balanced CC6-En dataset **with the proposed approach** at different training steps.

C.4 Detailed Results on Balanced CC6 Datasets

Table 10 lists the detailed results on the balanced datasets with different centric languages.

Table 11 lists the results of supervised translation in two separate directions: non-centric to centric, and centric to non-centric. The marginal performance decline of supervised translation is mainly from the translation from non-centric languages to centric language.

Methods Methods	w/o Pretrain				w/ Pretrain			
	All	Sup.	Zero-Shot		All	Sup.	Zero-Shot	
	BLEU↑	*BLEU*↑	*BLEU*↑	*OTR*↓	*BLEU*↑	*BLEU*↑	*BLEU*↑	*OTR*↓
Balanced CC6-En								
MNMT	22.5	37.2	15.2	36.8	14.5	38.1	2.7	95.6
+GenTrain	28.5	37.0	24.3	1.5	28.0	37.7	23.1	4.4
Balanced CC6-De								
MNMT	26.1	29.6	24.4	8.5	11.7	30.2	2.5	95.8
+GenTrain	28.1	29.3	27.5	0.7	28.1	29.8	27.2	2.4
Balanced CC6-Zh								
MNMT	26.6	29.4	25.2	4.1	24.2	30.0	21.3	21.0
+GenTrain	27.7	29.0	27.1	0.4	28.5	29.5	28.0	0.6
Balanced CC6-Ro								
MNMT	22.9	30.5	19.1	24.3	15.9	30.8	8.4	73.8
+GenTrain	26.3	30.0	24.5	0.9	26.9	30.3	25.2	0.9
Balanced CC6-Fr								
MNMT	26.8	33.1	23.6	7.0	23.1	33.6	17.9	34.4
+GenTrain	28.2	32.6	26.0	0.8	29.2	33.3	27.1	0.7
Balanced CC6-Ja								
MNMT	21.9	28.5	18.6	23.4	19.3	29.3	14.3	49.2
+GenTrain	23.2	28.1	20.8	5.1	23.6	28.9	20.9	5.3

Table 10: Translation performance on the test set for the six **balanced CC6 datasets**. "All" denotes the results on all translation directions including both supervised (e.g., 10 directions) and zero-shot (e.g., 20 directions) translation.

Methods Methods	w/o Pretrain			w/ Pretrain		
	Sup.	Non-C to C	C to Non-C	Sup.	Non-C to C	C to Non-C
	BLEU↑	*BLEU*↑	*BLEU*↑	*BLEU*↑	*BLEU*↑	*BLEU*↑
Balanced CC6-En						
MNMT	37.2	34.1	40.2	38.1	35.1	41.0
+GenTrain	37.0	33.6	40.4	37.7	34.4	41.0
Balanced CC6-De						
MNMT	29.6	24.8	34.3	30.2	25.5	34.7
+GenTrain	29.3	24.3	34.4	29.8	24.8	34.9
Balanced CC6-Zh						
MNMT	29.4	34.0	24.8	30.0	34.6	25.3
+GenTrain	29.0	33.2	24.8	29.5	33.8	25.3
Balanced CC6-Ro						
MNMT	30.5	27.2	33.7	30.8	27.5	34.0
+GenTrain	30.0	26.4	33.7	30.3	26.7	34.0
Balanced CC6-Fr						
MNMT	33.1	32.9	33.4	33.6	33.4	33.8
+GenTrain	32.6	32.0	33.4	33.3	32.9	33.8
Balanced CC6-Ja						
MNMT	28.5	37.4	19.6	29.3	38.3	20.2
+GenTrain	28.1	36.7	19.6	28.9	37.7	20.3

Table 11: Supervised translation performance on the test set for the six **balanced CC6 datasets**. "Non-C to C" denotes the results of non-centric to centric supervised translation, and "C to Non-C" denotes the results of centric to non-centric supervised translation.