Adapting Multilingual Models for Specialized Translation through Mixed Fine-tuning

Liyan Wang Haotong Wang Yves Lepage
早稲田大学 情報生産システム研究科
{wangliyan0905@toki.,wanghaotong0925@toki., yves.lepage@}waseda.jp

Abstract

In this work, we dissect mixed fine-tuning for adapting multilingual models to English-to-Japanese translation. We explore different sampling regimes across specialized and generic translations. Our findings indicate that oversampling the in-domain data leads to notable improvements in domain-specific performance, yet at the cost of severe degradation in generalization to unseen languages, performing even worse than basic fine-tuning with no generic data. In contrast, undersampling the generic data preserves more of the original multilingual capabilities while still achieving moderate domain adaptation gains. These results highlight the critical role of managing training size and data coverage to optimize the trade-off between specialization and generalization during adaptation.

1 Introduction

Adaptive Neural Machine Translation (NMT) traditionally involves fine-tuning a pre-trained, generic model on a small amount of in-domain data to improve performance on a specialized target domain. While effective, this basic approach often leads the model to overfit the in-domain distribution and lose the generalization capabilities learned from large-scale generic data. To address this, mixed fine-tuning for NMT has been first proposed in [1] based on the idea of domain adaptation where in-domain data is limited. This method updates a generic model, pre-trained on large-scale generic data, by training it on a mix of in-domain and generic samples, effectively improving in-domain performance while mitigating overfitting. Prior studies have applied it for various transfer scenarios, such as adapting translations across different domains for the same language pair [2] and improving low-resource translations by leveraging high-resource language data [3].

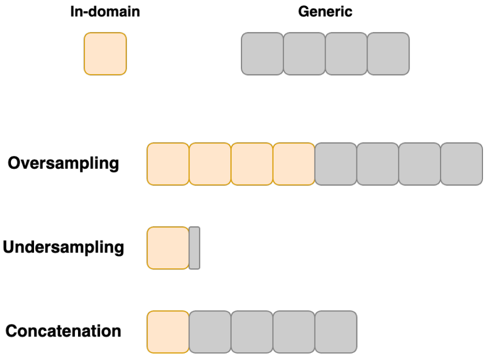

Figure 1 Variants of mixed fine-tuning in sampling training instances from in-domain and generic sets.

In this work, we apply mixed fine-tuning to adapt a pre-trained multilingual model for a specific translation task. Recognizing that data used in the development of many state-of-the-art models is not always open-sourced, we simulate this scenario by using a small subset of generic data. We investigate the impact of varying training data distributions (shaped by different sampling regimes), focusing on their influence on specialized and multilingual translation performance.

2 Mixed fine-tuning: background and variants

The mixed fine-tuning approach involves training a generic model on out-of-domain data and then fine-tuning the model using both in-domain and out-of-domain data. In this work, we focus on the fine-tuning step. A common technique of this approach is oversampling, where in-domain samples are repeated multiple times to balance their weight against the larger generic dataset.

This work explores three distinct variants of mixed fine-tuning, each defined by how in-domain (in size of I) and generic (in size of O) parallel sentences are combined. Figure 1 illustrates the difference between these variants.

Oversampling As proposed in [1], oversampling interleaves two datasets with a heavier sampling probability to in-domain data. The probability is based on the size of the generic data relative to the total size, i.e., $\frac{O}{I+O}$ for in-domain data and $1 - \frac{O}{I+O}$ for generic data. Consequently, each in-domain sample is repeated approximately $\frac{O}{I}$ times, creating a combined set where $I \times \frac{O}{I}$ in-domain instances are interleaved with O generic ones, thus a 1:1 ratio by the end of training iterations.

Undersampling It uses the same sampling probabilities as oversampling but stops earlier when all in-domain samples have been added. Undersampling exhausts all I examples from the in-domain set, resulting in a total number of $\frac{I}{\frac{O}{I+O}} = \frac{I(I+O)}{O}$ training samples. Of these, the generic data is trimmed to $\frac{I(I+O)}{O} \times (1 - \frac{O}{I+O}) = \frac{I^2}{O}$.

Concatenation As a simpler alternative, it directly combines the original in-domain and generic sets without adjusting their sizes. The training set includes $I+O$ parallel sentences with raw data proportions.

3 Experimental setups

3.1 Datasets

Table 1 summarizes the distribution of languages in the datasets used in the experiments reported in this work.

Table 1 Dataset distribution.

	Data	Languages translated from English
Train	in-domain	Japanese (ja)
	generic	Spanish (es), Chinese (zh), Indonesian (id), Portuguese (pt), Finnish (fi), Urdu (ur), Macedonian (mk), Albanian (sq), Dutch (nl)
Valid.	in-domain	Japanese
Test	in-domain	Japanese
	generic	Spanish, Chinese, Indonesian, Portuguese, Finnish, Urdu, Macedonian, Albanian, Dutch, Japanese, and others (over 100 languages)

In-domain data The in-domain data is sourced from the Kyoto Free Translation Task (KFTT) corpus[1], which contains English-Japanese parallel sentences extracted from Wikipedia articles. We apply several filtering criteria as in [4] to both the source and target segments and randomly sample 2k sentence pairs for validation and test. For training, we experiment with three sizes of 5k, 10k and 50k parallel sentences.

Generic data The generic data is sampled from OPUS[2] [5], a collection of parallel corpora used in developing the series of OPUS translation models. We randomly select 9 language pairs and curate 10k parallel sentences for each pair following the same filtering procedures. In total, 90k generic sentence pairs are selected for model training.

3.2 Models and evaluation

We experiment with the Helsinki-NLP/opus-mt-en-mul[3] model [6], which is capable of translating English into 120 different languages.[4] For mixed fine-tuning, we implement three different settings, each designed to explore unique sampling strategies for combining in-domain and generic data. These settings are compared against two baselines: the pre-trained model, used without additional fine-tuning, and basic fine-tuning, which updates the model for convergence on English-to-Japanese translation using only in-domain data. The evaluation is conducted on a specialized test set, aligned with the in-domain training data, and a generic test set[5] to assess generalization across multilingual translations. The generic set includes 9 selected language pairs from the generic training data, Japanese-specific translations, and over 100 other (unselected) language pairs that were not seen during fine-tuning.

4 Results and analysis

4.1 Specialized versus generic translation

Table 2 presents the performance of multilingual MT models fine-tuned for English-to-Japanese translation using various strategies. Without fine-tuning, the pre-trained model lacks the specialized knowledge required for translating Wikipedia content (of KFTT sentences), achieving a BLEU score of only 3.3 on the English-to-Japanese test set. Introducing fine-tuning leads to significant improvements in domain-specific adaptation, though with a slight degradation in generating the correct target language. Basic fine-tuning, which uses only the 10k in-domain dataset, increases the BLEU score by +5.6 points over the baseline. When incorporating generic data through mixed fine-tuning strategies, out-of-domain exposure pro-

[1] https://opus.nlpl.eu/KFTT/corpus/version/KFTT

[2] https://opus.nlpl.eu/opus-100.php.

[3] https://huggingface.co/Helsinki-NLP/opus-mt-en-mul

[4] The OPUS model is built on the MarianNMT architecture, consisting of a 6-layer encoder and a 6-layer decoder with a total of 77 million trainable parameters.

[5] https://object.pouta.csc.fi/Tatoeba-MT-models/eng-mul/opus2m-2020-08-01.test.txt

Table 2 Performance of translation models fine-tuned under various configurations for interleaving 10k in-domain data (I) and generic data (O). Models are evaluated on both specialized (English-to-Japanese) and generic (multilingual) translation tasks using BLEU, ChrF++ and LangAcc, where LangAcc measures the accuracy of generating translations in the correct target language.

Fine-tuning approach	Data size (k) I	+	O	BLEU ↑	ChrF++ ↑	LangAcc ↑
Specialized translation (ja)						
-	0	+	0	$3.3_{\pm0.2}$	$8.8_{\pm0.2}$	**100.0%**
basic	10	+	0	$8.9_{\pm0.4}$	$15.3_{\pm0.3}$	99.7%
mixed-concat	10	+	90	$9.5_{\pm0.4}$	$16.1_{\pm0.3}$	99.8%
mixed-under	10	+	1	$8.8_{\pm0.3}$	$15.3_{\pm0.3}$	99.6%
mixed-over	90	+	90	$\mathbf{13.1}_{\pm0.5}$	$\mathbf{19.8}_{\pm0.4}$	**100.0%**
Generic translation (all)						
-	0	+	0	$\mathbf{40.5}_{\pm2.8}$	$\mathbf{59.7}_{\pm0.7}$	**71.7%**
basic	10	+	0	$30.2_{\pm3.2}$	$51.3_{\pm0.7}$	71.0%
mixed-concat	10	+	90	$18.8_{\pm1.8}$	$45.9_{\pm0.6}$	66.2%
mixed-under	10	+	1	$31.0_{\pm3.2}$	$53.2_{\pm0.7}$	71.3%
mixed-over	90	+	90	$9.5_{\pm1.4}$	$42.0_{\pm0.9}$	66.4%
Generic translation (unselected)						
-	0	+	0	$\mathbf{39.0}_{\pm3.2}$	$\mathbf{59.4}_{\pm0.8}$	**65.6%**
basic	10	+	0	$28.1_{\pm3.5}$	$50.5_{\pm0.8}$	65.0%
mixed-concat	10	+	90	$15.2_{\pm1.7}$	$43.1_{\pm0.7}$	58.6%
mixed-under	10	+	1	$29.0_{\pm3.6}$	$52.6_{\pm0.8}$	65.0%
mixed-over	90	+	90	$7.2_{\pm1.1}$	$38.8_{\pm0.9}$	58.9%
Generic translation (selected)						
-	0	+	0	$\mathbf{52.8}_{\pm2.3}$	$\mathbf{65.1}_{\pm1.5}$	96.7%
basic	10	+	0	$41.1_{\pm1.9}$	$58.3_{\pm1.3}$	95.9%
mixed-concat	10	+	90	$44.5_{\pm2.1}$	$61.8_{\pm1.3}$	**97.6%**
mixed-under	10	+	1	$42.3_{\pm1.9}$	$59.6_{\pm1.3}$	97.2%
mixed-over	90	+	90	$43.0_{\pm1.9}$	$60.6_{\pm1.3}$	**97.6%**
Generic translation (ja)						
-	0	+	0	$\mathbf{14.9}_{\pm2.9}$	$\mathbf{19.0}_{\pm2.4}$	99.1%
basic	10	+	0	$9.8_{\pm1.7}$	$15.8_{\pm1.4}$	98.7%
mixed-concat	10	+	90	$8.9_{\pm1.7}$	$15.3_{\pm1.3}$	99.5%
mixed-under	10	+	1	$9.9_{\pm1.7}$	$15.8_{\pm1.4}$	98.7%
mixed-over	90	+	90	$9.3_{\pm1.7}$	$15.5_{\pm1.3}$	**99.6%**

vides competitive performance, with larger amounts of generic data yielding marginal improvements. In particular, `mixed-concat`, which accesses the full generic data, achieves slight but statistically insignificant gains over `mixed-under` that includes a small portion of generic data. This indicates that the benefits of out-of-domain data for specialized translation are limited. In contrast, the oversampling strategy, which amplifies the presence of in-domain instances by repeating each multiple times, delivers the best performance in specialized translation, achieving notable gains of approximately 10 points in both BLEU and ChrF++ evaluations.

However, this improved in-domain performance comes at a substantial cost to the general translation capability. All fine-tuned models exhibit notable declines on the generic multilingual test. More strikingly, increasing generic data during fine-tuning (as in `mixed-concat`) paradoxically worsens generalization, hinting at potential conflicts in representation learning when scaling up training computations. The oversampling regime, in particular, leads to a form of catastrophic forgetting, where previously acquired multilingual proficiency diminishes significantly. This degradation is especially pronounced for unselected languages that received no reinforcement during fine-tuning. Similarly, intensive adaptation erodes the general translation proficiency in the same language pair. While `mixed-over` enhances the ability to translate specialized English-to-Japanese text, it fails to preserve the quality of generic English-to-Japanese translation.

4.2 Influence of in-domain data scale

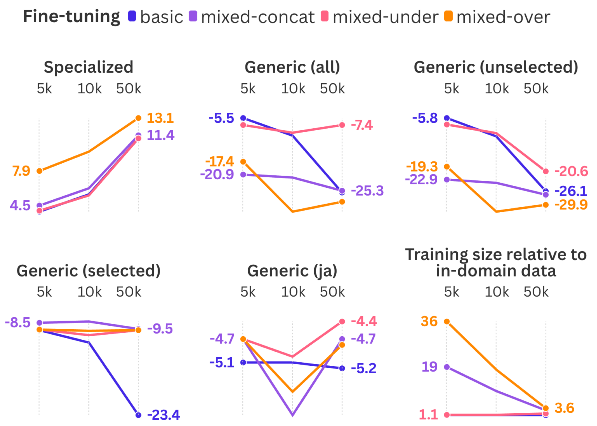

Figure 2 Differences in BLEU scores relative to the pre-trained multilingual model for various fine-tuning strategies across three scales of in-domain training data (5k, 10k, and 50k).

In addition to 10k in-domain scenario, we also conduct experiments using both smaller (5k) and larger (50k) datasets. Figure 2 shows comparison between fine-tuned models and the pre-trained MT model across three scales in specialized and general multilingual translation capabilities (as measured by BLEU).

For specialized translation, increasing the amount of in-domain data consistently leads to better performance. We observe that BLEU scores of fine-tuned models are proportionally magnified as the amount of in-domain data increases. The oversampling regime consistently outperforms others in capturing the domain-specific distribution across all in-domain data scales. It amplifies the size of in-domain data by repeatedly exposing the model to samples based on the ratio of generic to in-domain data. Importantly, oversampling, while maintaining the total in-domain repetition constant, yields greater gains when applied to more diverse datasets (as shown in Figure 3). By encountering a wider variety of examples in larger in-domain datasets, the `mixed-over` model achieves better generalization for the specialized task. Incorporating generic data (`mixed-concat` or `mixed-under`) yields marginal gains over basic fine-tuning.

In contrast, adaptation to a specialized language pair exacts a cost on multilingual translation quality, where fine-tuning the generic model results in performance degradation. Both basic and mixed approaches suffer declines as more in-domain data is introduced. Furthermore, we observe that increasing the training size during fine-tuning exacerbates degradation on generic translation tasks, particularly for unselected language pairs (Figure 2). In `mixed-concat` and `mixed-over` settings, where the training size is scaled up significantly in the 5k in-domain scenario, the models perform substantially worse than basic fine-tuning. Undersampling, which limits access to a smaller portion of generic data, emerges as a comparatively stable compromise in maintaining generalization. Although the model still experiences declines compared to the original generic baseline, these losses remain relatively modest, performing better than other mixed fine-tuning settings. This paradox suggests that **more extensive training appears to intensify conflicts in representation learning, ultimately harming performance in areas outside the adapted domain.**

For selected languages (those included in the generic data), mixed fine-tuning strategies demonstrate a stable degradation of around -10 points across different in-domain data sizes. In contrast, basic fine-tuning, which excludes these languages, exhibits a more substantial drop of -23 points in the 50k setting. This highlights that the presence of even a small amount of generic data (as in `mixed-under`) stabilizes performance on selected languages. When the ratio between in-domain and generic data is less extreme, the proportion of the in-domain appears to have minimal impact on three settings of mixed fine-tuning.

4.3 Insights from negative results

While increasing the in-domain data improves specialized translation, our findings suggest that avoiding excessive exposure to large volumes of generic data is equally crucial. **Increasing training size imposes a growing penalty on generic multilingual translation performance.** Fine-tuning methods that integrate generic data excessively can also lead to severe overfitting to the learned distribution. Instead, undersampling ensures that the sampling distribution remains more in line with the in-domain data by trimming generic data, emerges as a more effective strategy than oversampling (Figure 4). Selective inclusion of generic data during adaptation can retain a residual level of generalization.

In addition, it is crucial to carefully select generic data when adapting pre-trained models that support multiple languages or domains. **Rather than maximizing the volume of generic data, prioritizing the coverage of samples proves more effective.** In our experiments, the uniform distribution of generic data across 9 languages enables the undersampling strategy to include representative examples from each language pair, even with small in-domain datasets. By incorporating fewer but more representative generic samples, the fine-tuned model achieves a better balance between specialized adaptation and multilingual generalization (Figure 5). These findings suggest that selecting generic data to ensure comprehensive language coverage within the pre-trained model would further enhance the effectiveness of mixed fine-tuning under the undersampling strategy.

5 Conclusion

This work examined the effects of fine-tuning strategies on adapting multilingual NMT models to specialized English-to-Japanese translation. We observed that domain-specific expertise scales with the quantity of in-domain samples. In particular, intensive exposure to in-domain data (e.g., through oversampling) can substantially enhance specialized translation quality. However, it risks eroding general translation performance, especially on unselected language pairs not covered in generic training data. Scaling in-domain data leads to cumulative degradation for generic translation in basic fine-tuning. In contrast, mixed fine-tuning facilitates better adaptation to out-of-domain translations, but its effectiveness depends on the generic data incorporated. Strategies that incorporate generic data more conservatively, as with undersampling, help maintain a better balance between domain adaptation and multilingual generalization.

References

[1] Chenhui Chu, Raj Dabre, and Sadao Kurohashi. An empirical comparison of domain adaptation methods for neural machine translation. In Regina Barzilay and Min-Yen Kan, editors, **ACL 2017**, pp. 385–391, Vancouver, Canada, July 2017.

[2] Yasmin Moslem, Rejwanul Haque, John Kelleher, and Andy Way. Domain-specific text generation for machine translation. In Kevin Duh and Francisco Guzmán, editors, **Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)**, pp. 14–30, Orlando, USA, September 2022. Association for Machine Translation in the Americas.

[3] Raj Dabre, Atsushi Fujita, and Chenhui Chu. Exploiting multilingualism through multistage fine-tuning for low-resource neural machine translation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 1410–1416, Hong Kong, China, November 2019. Association for Computational Linguistics.

[4] Christian Herold, Jan Rosendahl, Joris Vanvinckenroye, and Hermann Ney. Detecting various types of noise for neural machine translation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Findings of the Association for Computational Linguistics: ACL 2022**, pp. 2542–2551, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[5] Jörg Tiedemann. Parallel data, tools and interfaces in OPUS. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, **LREC'12**, pp. 2214–2218, Istanbul, Turkey, May 2012.

[6] Jörg Tiedemann and Santhosh Thottingal. OPUS-MT – building open translation services for the world. In André Martins, Helena Moniz, Sara Fumega, Bruno Martins, Fernando Batista, Luisa Coheur, Carla Parra, Isabel Trancoso, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof, Mary Nurminen, Lena Marg, and Mikel L. Forcada, editors, **EAMT 2020**, pp. 479–480, Lisboa, Portugal, November 2020.

A Training details

Fine-tuning is performed for up to 5 epochs with a batch size of 32 on an Nvidia RTX A6000 GPU. We use a dropout rate of 0.1, a maximum learning rate of 2e-5, and set the beam search to 4 beams. The Adam optimizer is configured with an epsilon of 1e-6. Model evaluation is conducted after each training epoch, with early stopping applied if there is no improvement in validation losses for 3 consecutive epochs.

B Trade-offs between specialized and generic performance

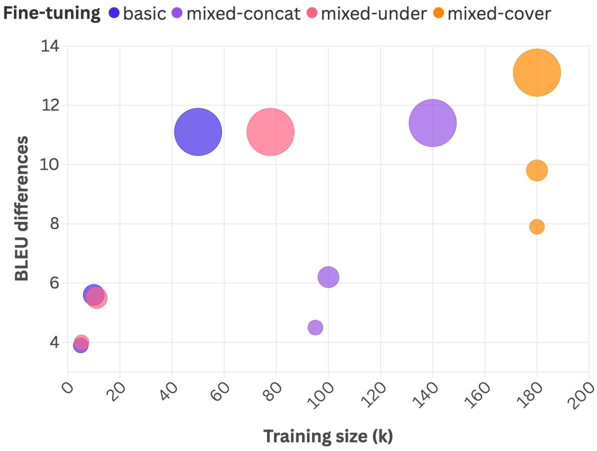

Figure 3 Performance differences (BLEU) in in-domain translations between fine-tuned models and the pre-trained translation model across various training data sizes. The size of each scatter indicates the in-domain data used (5k, 10k, or 50k).

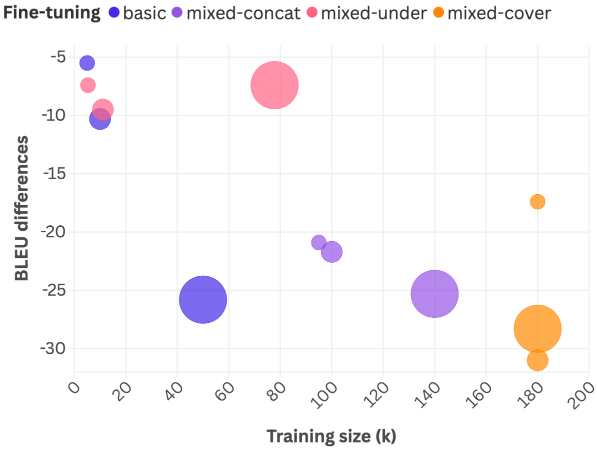

Figure 4 Degradations in generic performance compared to the pre-trained translation model. Scatter sizes indicate the in–domain data used (5k, 10k, or 50k).

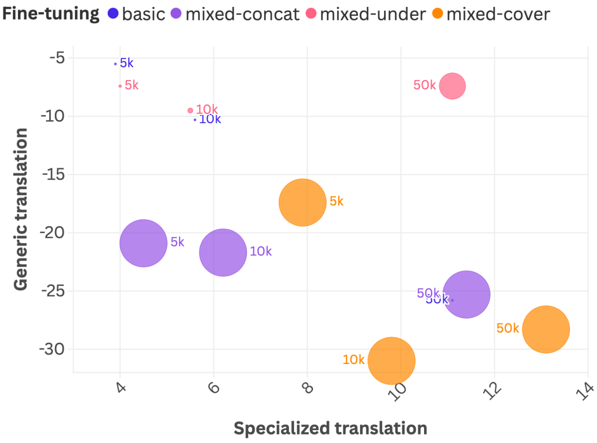

Figure 5 Correlation of improvements in specialized (in-domain) translations and degradations in generic (out-of-domain) translations. The labels around the scatters denote the in-domain datasets used, while the size of each scatter represents the number of generic examples involved during fine-tuning.