Denoising Pretraining of mT5 on OPUS-100 for Domain Adaptation in Machine Translation

Eshan Maduranga
Department of Computer Science and Engineering
University of Moratuwa , Sri Lanka
eshan.21@cse.mrt.ac.lk

Uthayasanker Thayasiwam
Department of Computer Science and Engineering
University of Moratuwa , Sri Lanka
rtuthaya@cse.mrt.ac.lk

Abstract—**Large pre-trained multilingual models have shown remarkable capabilities, yet their performance can be suboptimal for specific domains like Machine Translation or low-resource language pairs without proper adaptation. This paper investigates the efficacy of denoising pretraining as an intermediate adaptation step for the mT5 model on machine translation tasks, specifically focusing on scenarios with limited computational resources. We explore two distinct denoising strategies using the OPUS-100 dataset: a monolingual reconstruction objective (EN_noisy → EN) and a bilingual denoising translation objective (EN_noisy → FR). Following this pretraining phase, models are instruction fine-tuned using Low-Rank Adaptation (LoRA) for parameter-efficient learning. We pretrain the model on multiple language pairs, with a detailed evaluation focused on English–French (EN–FR) and French–English (FR–EN) translation to gain deeper insights. Experimental results demonstrate that both denoising pretraining methods yield substantial improvements over a baseline that only undergoes instruction fine-tuning. Notably, the monolingual denoising approach (Approach 1) shows superior performance, especially in the FR→EN direction, resulting in a significant improvement in the BLEU score. Our findings suggest that denoising pretraining with Instruction fine tunning is a powerful and resource-efficient technique for domain and task adaptation of multilingual models, significantly enhancing their translation quality.**

Index Terms—**Machine Translation, mT5, Denoising Pretraining, Low-Rank Adaptation (LoRA), Parameter-Efficient Fine-Tuning, Domain Adaptation, Natural Language Processing.**

I. Introduction

Neural Machine Translation (NMT) has become the state-of-the-art for automated translation, largely driven by the success of large pre-trained language models based on the Transformer architecture [1] [2]. Multilingual models like mT5 (Multilingual Text-to-Text Transfer Transformer) [3] are particularly appealing as they are pre-trained on a vast corpus spanning over 100 languages, enabling them to perform zero-shot or few-shot translation. However, these models are often trained on general web-crawled text, creating a ”domain gap” when applied to more specific corpora, such as the parallel datasets used in machine translation. Furthermore, achieving high performance on specific language pairs requires significant fine-tuning, a process that is computationally expensive and data-intensive, often prohibitive for researchers and practitioners without access to large-scale GPU clusters.

This research addresses the challenge of adapting multilingual models for translation tasks in a resource-constrained environment. We focus on mT5-small, a more manageable variant of the mT5 family, and the OPUS-100 dataset [4], a large collection of parallel corpora. The core hypothesis is that an intermediate pretraining step, specifically focused on a denoising objective, can better prepare the model for the downstream translation task, leading to improved performance even with efficient fine-tuning methods. [5]

Denoising pretraining is a self-supervised learning paradigm in which a model is trained to reconstruct clean text from artificially corrupted input. By introducing noise through masking, shuffling, or deletion, the model is encouraged to capture contextual dependencies and develop robust linguistic representations.

This approach, popularized by BART [6] and T5 [1], has been extended to multilingual settings with models such as mBART [7] and mT5 [3], where it has proven highly effective for cross-lingual transfer and domain adaptation. In this work, we adapt the denoising pretraining concept into two distinct approaches tailored for translation tasks.

1) **Monolingual Denoising**: The model learns to reconstruct a clean English sentence from a noisy version (EN_noisy → EN) of English itself. This task strengthens the model's understanding of the source language's grammar and syntax and contextual understanding.
2) **Bidirectional Bilingual Denoising**: The model is trained to reconstruct a clean target sentence in one language given a noisy source sentence in the other language. For example, a noisy English input is mapped to its clean French counterpart (EN_noisy → FR), and vice versa (FR_noisy → EN). This objective integrates denoising pretraining with cross-lingual translation, encouraging the model to learn robust bilingual representations under noisy conditions [8].

Following the pretraining phase, we employ instruction fine-tuning with Low-Rank Adaptation (LoRA) [9], a parameter-efficient fine-tuning (PEFT) technique that drastically reduces the number of trainable parameters. This makes the fine-tuning process feasible on a single consumer-grade GPU. Instruction fine-tuning frames the task in natural language (e.g., ”Translate English to French: ...”), aligning the model’s behavior more closely with user intent.

Our contributions are threefold:

- We propose and systematically compare two denoising pretraining strategies—monolingual and bilingual—as an effective intermediate adaptation step for NMT.
- We demonstrate that combining denoising pretraining with parameter-efficient fine-tuning (LoRA) provides a computationally feasible and highly effective pipeline for adapting multilingual models for translation.
- We provide a detailed analysis for the English-French language pair, showing that both strategies significantly outperform a strong instruction-tuned baseline.

We conduct a comprehensive evaluation using standard NMT metrics—BLEU [10], chrF [11], and TER [12]. Our results reveal that both pretraining strategies significantly outperform the baseline, with the monolingual denoising approach demonstrating a particularly strong advantage.This work contributes a practical framework for adapting smaller multilingual models, making high-quality NMT more accessible.

II. Related Work

A. Multilingual Pre-trained Models

The paradigm of pre-training on large-scale unlabeled text and fine-tuning on downstream tasks has revolutionized NLP. Multilingual models extend this paradigm to multiple languages. Early models like mBERT learned joint representations by sharing a vocabulary across languages. Later models like XLM-R improved upon this by pre-training on a much larger and cleaner dataset. In contrast, NLLB models [13] move away from denoising pretraining and are trained directly on large-scale parallel corpora using supervised translation objectives. The T5 [1] model introduced a unified text-to-text framework, treating every NLP task as a sequence-to-sequence problem. Its multilingual successor, mT5 [3], was pre-trained on the **mC4 corpus** [3] covering 101 languages. While larger versions of mT5 are powerful, smaller variants like mT5-small offer a balance between performance and computational cost, making them ideal candidates for research in resource-constrained settings.

B. Denoising Objectives for Pre-training

The concept of denoising autoencoders involves training a model to reconstruct a clean input from a corrupted version. This idea was successfully applied to NLP with the introduction of BART [6], which is pre-trained by corrupting text with an arbitrary noising function (e.g., token masking, deletion, sentence permutation) and learning a model to reconstruct the original text. T5’s pre-training objective, span corruption, is another powerful form of denoising. These schemes help the model learn comprehensive syntactic and semantic representations. Our work adapts this principle, not as a large-scale pre-training from scratch, but as an intermediate domain-adaptation step tailored for the translation task. We specifically investigate how different formulations of the denoising objective—monolingual versus bilingual—impact downstream performance. [14]

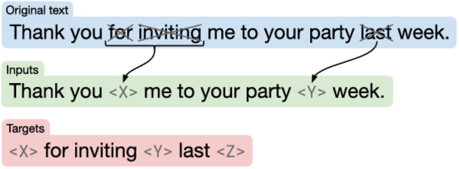
Original text
Thank you ~~for inviting~~ me to your party ~~last~~ week.
Inputs
Thank you <X> me to your party <Y> week.
Targets
<X> for inviting <Y> last <Z>

Fig. 1. Reconstruction noisy sentences.

C. Parameter-Efficient Fine-Tuning (PEFT)

Fine-tuning all parameters of large pre-trained models is often computationally prohibitive. PEFT methods have emerged as a solution, enabling adaptation by updating only a small fraction of the model’s parameters. Techniques like adapter modules, prefix-tuning, and prompt-tuning have shown promise. Low-Rank Adaptation (LoRA) [9], in particular, has gained popularity for its effectiveness and simplicity. It works by injecting trainable low-rank matrices into the layers of the Transformer architecture while freezing the original pre-trained weights. This is based on the hypothesis that the change in weights during adaptation has a low ”intrinsic rank.” LoRA significantly reduces the number of trainable parameters and the memory footprint, making it possible to fine-tune large models on consumer-grade hardware, which aligns perfectly with the constraints of this research.

D. Instruction Tuning

Instruction tuning [15] is a fine-tuning methodology that trains models on a collection of tasks framed as natural language instructions (e.g., ”Translate English to French: *sentence*”). This technique [16], popularized by models like FLAN and T0, has been shown to improve model generalization to unseen tasks and formats, making them more aligned with human interaction. By combining our denoising pretraining with subsequent instruction fine-tuning, we aim to leverage both the robust representations learned from the denoising task and the task-formatting flexibility imparted by instruction tuning.

III. Methodology

This section details the model architecture, dataset, pretraining strategies, fine-tuning procedure, and evaluation metrics used in our experiments.

A. Model and Dataset

a) Model: We use **mT5-small**, a smaller variant of the multilingual T5 model, as the base for all our experiments. Its architecture is summarized in Table I. Its manageable size makes it a suitable choice for experimentation under limited GPU resources (e.g., a single GPU with 24GB VRAM).

TABLE I
mT5-small Model Specifications.

Hyperparameter	Value
Encoder Layers	8
Decoder Layers	8
Attention Heads	6
Model Dimension (d_{model})	512
Feed-Forward Dimension (d_{ff})	1024
Total Parameters	∼300 Million

b) Dataset: All experiments are conducted on the **OPUS-100** dataset [4], a large-scale multilingual corpus containing parallel data for 100 languages, sourced from various domains including movie subtitles, legal documents, and technical manuals.

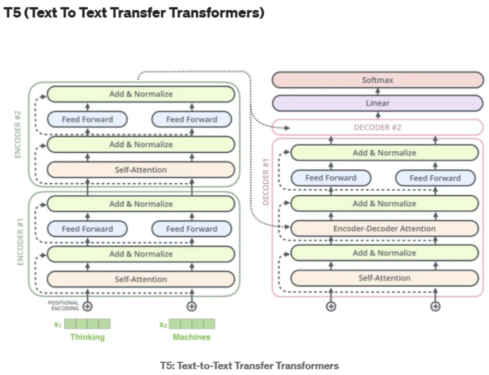

Fig. 2. T5 Model Architecture.

B. Denoising Pretraining Strategies

The core of our methodology involves an intermediate pretraining step to adapt the mT5-small model. We use a Naïve Noise Injection function, which introduces corruptions to simulate noisy inputs. For any given sentence, we apply the following corruptions:

- **Word Deletion (10%)**: Random words are deleted from the sequence.
- **Character Deletion (10% per word)**: Random characters are deleted from selected words.
- **Word Swapping (5%)**: Positions of adjacent words are randomly swapped.

These corruptions encourage the model to learn robust contextual representations and improve denoising and translation performance.

*1) **Approach 1**: Monolingual Denoising Pretraining:* In this approach, the model is trained on a monolingual reconstruction task. The objective is to predict the original, clean English sentence given a noisy version. This setup, formally represented as EN_noisy → EN, is designed to force the model to learn a deep understanding of the source language's structure, grammar, and semantics.

*2) **Approach 2**: Bidirectional Bilingual Denoising Pretraining:* Our second approach combines the denoising objective with the cross-lingual translation task. The model is trained to generate the correct target sentence from a noisy source sentence in the other language (e.g., EN_noisy → FR, FR_noisy → EN). This objective directly trains the model for translation while simultaneously requiring it to handle noisy and incomplete input. The hypothesis is that this joint objective could lead to a model that is both a good translator and robust to variations in the source text. [17]

C. Instruction Fine-tuning with LoRA

After the denoising pretraining phase, all models, including a baseline mT5-small that did not undergo this phase, are instruction fine-tuned [9]. We use a standard translation instruction format:

- **Input Prompt:** "Translate English to French: {*English sentence*}"
- **Target:** "{*French sentence*}"

To make this process computationally efficient use parameter efficient fine tuning techniques [18], we employ **LoRA (Low-Rank Adaptation)** [9]. LoRA freezes the original weights of the mT5 model and injects small, trainable low-rank matrices into the query and value projections of the self-attention layers. This reduces the number of trainable parameters by over 99%. The specific LoRA hyperparameters are detailed in Table II.

TABLE II
LoRA Hyperparameters for Instruction Fine-Tuning.

Hyperparameter	Value
Rank (r)	8
Alpha (α)	32
Target Modules	‘q‘, ‘v‘
Dropout	0.1
Trainable Parameters	∼0.7% of total

D. Experimental Setup

Our methodology was validated across 10 diverse language pairs, all centered around English. While experiments were conducted on all pairs yet, we present results for **English-French** as a detailed case study, as its outcomes are representative of the trends observed more broadly. We compare four model configurations:

1) **Original mT5-small:** The base, off-the-shelf model.
2) **mT5-small + Instruct-FT (Baseline):** The base model fine-tuned with our LoRA setup.
3) **Approach 1 + Instruct-FT:** Pretrained with monolingual denoising (EN_noisy → EN) then instruction-tuned.
4) **Approach 2 + Instruct-FT:** Pretrained with bilingual denoising (EN_noisy → FR) then instruction-tuned.

Both the pretraining and fine-tuning experiments were conducted on two **NVIDIA T4 GPUs**, each equipped with **16**

GB of VRAM. The pretraining phase was performed for 3 epochs over the 100,000 sentence pairs, while the fine-tuning phase was conducted for 3 epochs. We employed the **AdamW** optimizer with a learning rate of **2e-4**.

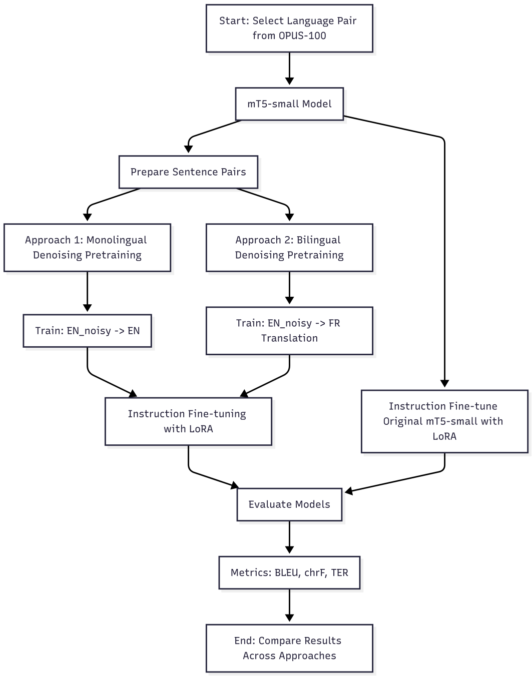
Start: Select Language Pair from OPUS-100
mT5-small Model
Prepare Sentence Pairs
Approach 1: Monolingual Denoising Pretraining
Approach 2: Bilingual Denoising Pretraining
Train: EN_noisy -> EN
Train: EN_noisy -> FR Translation
Instruction Fine-tuning with LoRA
Instruction Fine-tune Original mT5-small with LoRA
Evaluate Models
Metrics: BLEU, chrF, TER
End: Compare Results Across Approaches

Fig. 3. Full pretraining Pipeline.

E. Evaluation Metrics

We evaluate translation quality using three widely-accepted metrics: BLEU, chrF, and TER. Each metric captures different aspects of translation quality.

- **BLEU [10]:** Measures the n-gram precision of the candidate translation compared to reference translations. It includes a brevity penalty to penalize overly short translations. The BLEU score is computed as:

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{1}$$

where p_n is the modified precision for n-grams, w_n is the weight for each n-gram (usually uniform), and BP is the brevity penalty defined as:

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases} \tag{2}$$

Here, c is the length of the candidate translation and r is the effective reference length. Higher BLEU scores indicate better translations.

- **chrF [11]:** Computes an F-score based on character n-grams, which is sensitive to morphology and minor variations in translation. It is defined as:

$$\text{chrF} = (1-\beta) \cdot \text{Precision}_{\text{char}} + \beta \cdot \text{Recall}_{\text{char}} \tag{3}$$

where $\text{Precision}_{\text{char}}$ and $\text{Recall}_{\text{char}}$ are character n-gram precision and recall, and β is typically set to 2 (giving more weight to recall). Higher chrF scores indicate better translations.

- **TER [12]:** Translation Edit Rate measures the minimum number of edits (insertions, deletions, substitutions, and shifts) needed to change the candidate translation into the reference translation. It is computed as:

$$\text{TER} = \frac{\text{Number of edits}}{\text{Number of reference words}} \tag{4}$$

Lower TER scores indicate better translation quality, as fewer edits are required.

IV. Results and Analysis

This section presents the quantitative results for English-French and French-English translation, a summary across all tested languages for get in-detail understanding.

A. English to French (EN → FR) Translation

Table III shows the performance of the different models on the EN→FR translation task.

TABLE III
Evaluation results for English to French (EN → FR) translation.

Model	**BLEU ↑**	**chrF ↑**	**TER ↓**
Original mT5-small	0.10	2.46	100.0
mT5-small + Instruct-FT	10.60	28.01	94.14
Approach 1 + Instruct-FT	**15.84**	36.97	**84.48**
Approach 2 + Instruct-FT	15.37	**38.41**	84.77

The results clearly illustrate the effectiveness of our proposed pipeline.

- **Baseline Performance:** The original mT5-small is ineffective, confirming the need for fine-tuning.
- **Impact of Instruction Fine-tuning:** LoRA-based instruction fine-tuning provides a massive performance boost, increasing the BLEU score to 10.60, establishing a strong baseline.
- **Impact of Denoising Pretraining:** Both denoising approaches yield substantial gains over the baseline. Approach 1 (monolingual) achieves the highest BLEU (15.84) and lowest TER (84.48), a relative BLEU improvement of 49%. Approach 2 (bilingual) is competitive, with a slightly lower BLEU but the highest chrF, suggesting it may be better at character-level morphology.

B. French to English (FR → EN) Translation

To assess generalizability, we evaluated performance on the reverse translation direction. The results are in Table IV.

The results in the FR→EN direction are even more striking.

- The instruction-tuned baseline is stronger here (14.73 BLEU), likely due to mT5's inherent strength in English generation from its pre-training data.

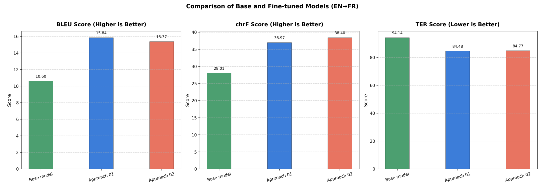

Fig. 4. Model Comparison EN-¿FR.

TABLE IV
EVALUATION RESULTS FOR FRENCH TO ENGLISH (FR → EN) TRANSLATION.

Model	BLEU ↑	chrF ↑	TER ↓
Original mT5-small	0.10	2.46	100.0
mT5-small + Instruct-FT	14.73	35.50	83.87
Approach 1 + Instruct-FT	**21.01**	**42.84**	**71.12**
Approach 2 + Instruct-FT	19.51	41.01	74.50

- **Approach 1 Dominance:** The model pretrained with monolingual English denoising (Approach 1) decisively outperforms all others, achieving a BLEU score of 21.01. This is a 42.6% relative improvement over the strong baseline. It leads across all metrics.
- Approach 2 still improves over the baseline but lags behind Approach 1. This strongly suggests that enhancing the model's generative capacity in the *target language* is a crucial factor for translation quality.

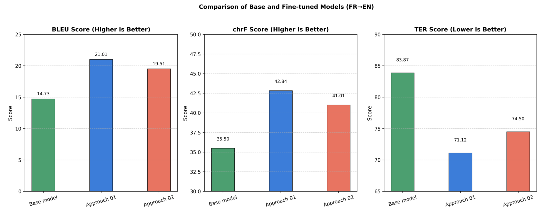

Fig. 5. Model Comparison FR-¿EN.

C. Performance Across All Language Pairs

To confirm that these findings are not specific to French, we averaged the BLEU score improvements over the instruction-tuned baseline across all 10 (see Table VI) language pairs for the X → EN direction. The results are summarized in Table V.

TABLE V
AVERAGE BLEU IMPROVEMENT OVER BASELINE FOR X → EN ACROSS 10 LANGUAGE PAIRS.

Model Configuration	Avg. Relative BLEU Gain
Approach 1 + Instruct-FT	**+38.5%**
Approach 2 + Instruct-FT	+29.1%

The trend holds across diverse languages. Approach 1, which strengthens the English target model, consistently provides a greater uplift in performance for translation into English compared to Approach 2.

V. DISCUSSION AND LIMITATIONS

The experimental results provide compelling evidence that intermediate denoising pretraining is a highly effective strategy for adapting mT5-small for translation. The key takeaway is the stark difference in the performance of the two pretraining strategies across translation directions.

a) Why Monolingual Denoising (Approach 1) Excels: The superior performance of Approach 1, particularly in the FR→EN direction, suggests that strengthening the model's understanding of the *target* language is paramount. The EN_noisy → EN pretraining task is effectively a language modeling objective that forces the model to learn the intricacies of English syntax, semantics, and fluency. When this model is later fine-tuned to translate from French to English, it already possesses a powerful English generative component. The fine-tuning process can then focus more on learning the cross-lingual mapping from French, as the burden of generating high-quality English output has been eased.

b) Limitations of Bilingual Denoising (Approach 2): While the bilingual denoising (EN_noisy → FR) objective is more directly related to the final task, it might be a less efficient learning signal. The model must simultaneously learn to (1) decipher the noisy English input and (2) perform the complex cross-lingual mapping to French. This could divide the model's learning capacity. Furthermore, this pretraining task does nothing to improve the model's ability to generate English, which explains its weaker performance in the FR→EN direction compared to Approach 1.

c) Implications for Low-Resource Adaptation: Our findings have important implications for practitioners with limited computational resources. Instead of computationally expensive full fine-tuning, a relatively short and data-efficient denoising pretraining phase on monolingual data (which is often more abundant than parallel data) can provide a significant performance boost. This two-stage process—monolingual denoising followed by parameter-efficient instruction tuning—presents a cost-effective and powerful framework for domain adaptation.

d) Limitations of the Study: This study was intentionally constrained to **mT5-small** and a limited data size to simulate a low-resource environment. As a result, the accuracy of mT5-small is lower compared to larger models, which require significantly higher computational resources. The ”Naïve Noise Injection” is also a simple corruption strategy. Due to limited computational power, we trained the model *language pair-wise*; with more resources, it would be possible to train on the full dataset covering all language pairs simultaneously. Additionally, with greater computational capacity, more extensive hyperparameter tuning could be performed to potentially improve performance. The experiments were also focused on a single PEFT method (LoRA), and the observed performance trends may or may not scale to larger models or more complex noising functions.

VI. Future Work

Based on the promising results of this study, several avenues for future research are apparent:

1) **Advanced Noise Functions:** Move beyond the current naïve noise injection to explore more structured strategies, such as the **Span Corruption** objective used by T5. This could encourage the learning of more robust and contextually-aware linguistic representations.
2) **Scaling to More Language Pairs:** A critical next step is to perform a detailed analysis of all 10 language pairs, including non-English target languages. This would require applying the monolingual denoising to other languages (e.g., FR_noisy → FR) to see if the target-language-enhancement benefit is universal.
3) **Low-Resource Language Pairs:** Apply this methodology to genuinely low-resource language pairs within the OPUS-100 dataset. The benefits of strengthening the target language model via monolingual denoising could be even more pronounced for languages where the base model has weaker initial representations.

VII. Conclusion

This research successfully demonstrated the value of denoising pretraining as a resource-efficient domain adaptation technique for improving the translation performance of the mT5-small model. We presented and evaluated two distinct strategies: a monolingual reconstruction task and a bilingual denoising translation task. Our experiments on English-French translation showed that both approaches significantly outperform a standard instruction-tuned baseline. Critically, we found that monolingual denoising of the target language (Approach 1) yielded the most substantial improvements, particularly when translating into that language. This approach increased the BLEU score for French-to-English translation from 14.73 to 21.01, a remarkable gain confirmed by trends across 10 language pairs. Our work provides a practical and effective framework for enhancing NMT quality in resource-constrained settings, highlighting the importance of strengthening a model's generative capabilities in the target language as a key step towards better translation.

References

[1] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[3] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Chowdhery, S. Narang, M. Mishra, W. Fedus, Y. Jernite *et al.*, "mt5: A massively multilingual pre-trained text-to-text transformer," *arXiv preprint arXiv:2010.11934*, 2021.

[4] J. Tiedemann, "Parallel data, tools and interfaces in OPUS," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA), May 2012, pp. 2214–2218. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf

[5] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," *arXiv preprint arXiv:1911.02116*, 2020.

[6] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[7] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726–742, 2020.

[8] M. Reid and M. Artetxe, "PARADISE: Exploiting parallel data for multilingual sequence-to-sequence pretraining," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 8379–8392. [Online]. Available: https://aclanthology.org/2022.acl-long.573

[9] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[10] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. [Online]. Available: https://aclanthology.org/P02-1040

[11] M. Popović, "chrf: character n-gram f-score for automatic mt evaluation," *arXiv preprint arXiv:1505.02560*, 2015.

[12] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*. Cambridge, Massachusetts, USA: Association for Machine Translation in the Americas, Aug. 2006, pp. 223–231. [Online]. Available: https://aclanthology.org/2006.amta-papers.25

[13] M. R. Costa-jussà, J. Tran, A. Sokolov, G. Ulrich, A. Max, C. Galmarini, M. Cettolo, M. Turchi, R. Cattoni, C. Federmann *et al.*, "No language left behind: Scaling human-centered machine translation," *arXiv preprint arXiv:2207.04672*, 2022.

[14] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 8342–8360.

[15] V. Sanh, A. Webson, C. Raffel, S. H. Bach, I. Sutskever, D. Kubric, J. Wei, K. Clark, A. Roberts, A. D'Amour *et al.*, "Multitask prompted training enables zero-shot task generalization," *arXiv preprint arXiv:2110.08207*, 2021.

[16] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. Pasupat, S. Wang, Q. V. Le *et al.*, "Finetuned language models are zero-shot learners," *arXiv preprint arXiv:2109.01652*, 2021.

[17] J. Pfeiffer, I. Vulić, I. Gurevych, and S. Ruder, "Mad-x: An adapter-based framework for multi-task cross-lingual transfer," *arXiv preprint arXiv:2005.00052*, 2020.

[18] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2790–2799.

Appendix

The models trained in this study are publicly available on Hugging Face. These include the original base model, as well as models pretrained using our approaches and subsequently fine-tuned with instruction tuning (these models are related to the EN ↔ FR evaluation task described above).

- **Original mT5-small (instruction fine-tuning only):** https://huggingface.co/Eshan210352R/mt5-small-instruct-ft-lora
- **Pretrained using Approach 1 and Instruction Fine-tuning:** https://huggingface.co/Eshan210352R/mt5-small-denoising-en-fr-correct-deonoise-lora-instruct-ft-enfr
- **Pretrained using Approach 2 and Instruction Fine-tuning:** https://huggingface.co/Eshan210352R/mt5-small-denoising-en-fr-full-lora-tune

Similar models have also been trained and pushed to Hugging Face for **other language pairs** also.

The experiments were conducted on ten bilingual subsets of the OPUS-100 dataset [**?**]. Each subset corresponds to English paired with another target language. Table VI lists all trained language pairs with their corresponding ISO language codes and full language names.

TABLE VI
OPUS-100 Trained Language Pairs

Pair	**Source–Target**	**Language Names**
en–fr	English ↔ French	English–French
en–it	English ↔ Italian	English–Italian
en–si	English ↔ Sinhala	English–Sinhala
en–zh	English ↔ Chinese	English–Chinese
en–es	English ↔ Spanish	English–Spanish
en–ru	English ↔ Russian	English–Russian
en–ko	English ↔ Korean	English–Korean
en–ja	English ↔ Japanese	English–Japanese
en–pt	English ↔ Portuguese	English–Portuguese
en–tr	English ↔ Turkish	English–Turkish

In addition to the English–French models described earlier, we have also pretrained and publicly released models for the remaining OPUS-100 language pairs. Table VII lists each language pair along with its corresponding Hugging Face model repository.

Figure 6 illustrates the training and validation loss progression during the pretraining stage for the English–French (EN–FR) model.

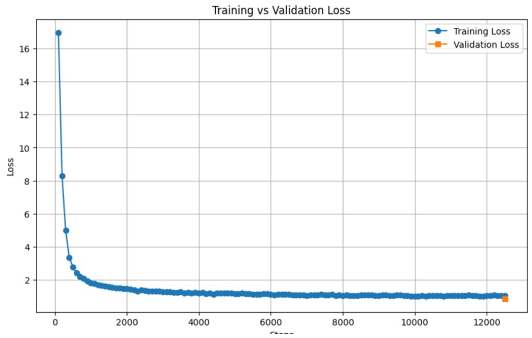

Fig. 6. Training and validation loss curves for the EN–FR model.

TABLE VII
Pretrained OPUS-100 Models on Hugging Face

Pair	**Language Names**	**Hugging Face Model**
en–it	English–Italian	https://huggingface.co/Eshan210352R/mt5-span-denoising-en-it-final
en–si	English–Sinhala	https://huggingface.co/Eshan210352R/mt5-small-denoising-en-si-final
en–zh	English–Chinese	https://huggingface.co/Eshan210352R/mt5-small-denoising-en-zh-final
en–es	English–Spanish	https://huggingface.co/Eshan210352R/mt5-small-denoising-en-es-final
en–ru	English–Russian	https://huggingface.co/Eshan210352R/mt5-small-denoising-en-ru-final
en–ko	English–Korean	https://huggingface.co/Eshan210352R/mt5-small-denoising-en-ko-final
en–ja	English–Japanese	https://huggingface.co/Eshan210352R/mt5-small-denoising-en-ja-final
en–pt	English–Portuguese	https://huggingface.co/Eshan210352R/mt5-small-denoising-en-pt-5-epcs
en–tr	English–Turkish	https://huggingface.co/Eshan210352R/mt5-small-denoising-en-tr-final