Multilingual Pre-training Meets Supervised Neural Machine Translation: A Reproducible Evaluation on English–French and Finnish Translation

Benyamin Ahmadnia[1], **Yeswanth Soma**[1], and **Hossein Sarrafzadeh**[2]

[1]Department of Computer Science, California State University, Dominguez Hills, USA
[2]North Carolina A&T State University, Greensboro, USA
[bahmadniayebosari,ysoma1]@csudh.edu,hsarrafzadeh@ncat.edu

Abstract

This paper presents a comparative evaluation of Transformer-based Neural Machine Translation (NMT) models and pre-trained multilingual sequence-to-sequence models in the context of moderately-resourced MT. Using English–French (high-resource) and English–Finnish (moderate-resource) as case studies, we assess the effectiveness of fine-tuning the mBART model versus training standard NMT systems from scratch. Our experiments incorporate data-augmentation techniques such as back-translation and evaluate translation quality using BLEU, TER, METEOR, and COMET metrics. We also provide a detailed error analysis that covers lexical choice, named entity handling, and word order. While mBART demonstrates consistent improvements over classical NMT, particularly in handling complex linguistic structures and sparse training data, we acknowledge the challenges of deploying large models in resource-constrained settings. Our findings highlight practical trade-offs between model complexity, resource availability, and translation quality in multilingual scenarios.

1 Introduction

Neural Machine Translation (NMT) systems that use the Transformer architecture (Vaswani et al., 2017) show significant improvements in translation quality for languages with abundant resources. The performance of NMT systems in moderately- and low-resource settings continues to be restricted because of insufficient data availability and domain discrepancies. Despite attempts to solve these challenges with techniques like transfer learning and multilingual training, round-tripping and back-translation continue to leave performance gaps across numerous real-world applications.

The introduction of modern pre-trained multilingual sequence-to-sequence models like mBART (Liu et al., 2020) has generated fresh possibilities to overcome existing limitations. Models that undergo pre-training on extensive multilingual corpora and receive fine-tuning for particular translation tasks become excellent options for managing various moderate- to low-resource language translation pairs. mBART utilizes an encoder-decoder architecture like traditional NMT models but stands out due to its extensive multilingual pre-training which differs from decoder-only Large Language Models (LLMs) such as GPT-3.

In this paper, we empirically evaluate the translation performance of:

- A baseline transformer-based NMT system trained from scratch.
- The fine-tuned mBART model as a multilingual pre-trained alternative.

We test both systems on two pairs of languages: English–French (a standard high-resource setting) and English–Finnish. Although Finnish is not a truly low-resource language - it is featured in many WMT tasks and is present in mBART's pre-training - it poses significant morphological and syntactic challenges that make it an informative case study for testing the limits of fine tuning.

To improve the Finnish training signal, we augment the training data using back-translation and synthetic data generation. Our evaluation uses standard lexical metrics (BLEU, METEOR, and TER), as well as the COMET metric (Rei et al., 2020), which has shown a stronger correlation with human judgments.

In addition, we perform an error analysis that covers lexical choice, named entity translation, word order, and morphological agreement. This allows us to better understand the strengths and weaknesses of each model architecture in practical translation scenarios.

Proceedings of Recent Advances in Natural Language Processing,pages 38–47
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-005

Although multilingual pre-training has been explored in prior work, most studies focus on high-resource scenarios or report metric-based improvements without thorough linguistic or deployment analysis. Our study contributes novel practical value by providing the first controlled, reproducible, and linguistically grounded comparison of standard Transformer NMT and mBART fine-tuning in both high-resource and moderately resourced morphologically rich settings. In addition to multi-metric evaluation, we include a detailed human-annotated error analysis and discuss real-world trade-offs in computational cost and deployment, addressing important gaps in the literature.

The contributions of this paper are as follows:

- A controlled and reproducible comparison of standard Transformer NMT and mBART fine-tuning on both high- and moderate-resource language pairs.
- A comprehensive evaluation using both traditional and neural metrics to assess translation quality.
- An error analysis highlighting common linguistic challenges and the relative robustness of each approach.
- Public release of our code, data splits, and model configurations to support future replication and benchmarking.

This work also serves as a methodological template for future MT evaluations by combining automatic metrics with qualitative analysis. As multilingual pre-trained models become more prevalent, establishing reproducible evaluation pipelines becomes essential to ensure fair comparisons, particularly for morphologically complex or under-resourced languages. By providing empirical clarity on the trade-offs between conventional NMT and pre-trained multilingual models, our study offers practical insights for researchers and practitioners working on MT for languages with limited annotated resources.

In addition to practical insights, this work serves as a reproducible benchmark for comparing multilingual pre-trained models and conventional NMT systems in moderately resourced, morphologically complex language pairs. By emphasizing error types, trade-offs, and deployment viability, we aim to offer a reference point for future evaluation of multilingual pre-training in real-world MT workflows.

2 Related Work

Numerous approaches have been proposed to improve NMT performance in low-resource settings. Transfer learning techniques (Zoph et al., 2016) enable fine-tuning a high-resource model on low-resource language pairs. Back-translation (Sennrich et al., 2016), synthetic data generation (Edunov et al., 2018), and round-tripping (Ahmadnia and Dorr, 2019; Ahmadnia et al., 2019, 2018) further improve data availability by leveraging monolingual corpora.

Multilingual NMT (MNMT) has shown success in enabling low-resource language pairs to benefit from shared representations trained on high-resource data (Johnson et al., 2017). Approaches such as OPUS-MT (Tiedemann, 2020) and mT5 (Xue et al., 2021) have leveraged multilingual training and pretraining objectives to scale translation quality across more than 100 languages.

In parallel, Facebook AI's No Language Left Behind (NLLB) model (Team et al., 2022) demonstrated strong performance in over 200 languages, including extremely low-resource settings. However, these models typically require substantial computational resources and may be impractical in constrained environments.

Sequence-to-sequence pre-trained models like mBART (Liu et al., 2020) have proven to be effective for multilingual translation. mBART is a denoising autoencoder trained on monolingual corpora and fine-tuned for translation, particularly suitable for encoder-decoder settings. It supports multiple languages and enables zero- and few-shot adaptation.

Despite its effectiveness, mBART is not a decoder-only LLM in the sense of GPT-3 (Brown et al., 2020). Decoder-only LLMs have more recently been explored for direct translation using prompting and fine-tuning.

Recent efforts such as ALMA (Li and et al., 2023) and Tower (Peng et al., 2024) have introduced methods to adapt decoder-only LLMs (e.g., LLaMA, GPT) for translation by optimizing decoding strategies and fine-tuning for multilingual translation. Tower, in particular, proposes a compact, fine-tuned translation model derived from LLaMA, showing strong results across 200+ language pairs.

These models typically outperform fine-tuned mBART in recent WMT evaluations but demand more memory and computation. Our work does not attempt to compete with such LLMs but rather offers a comparative analysis between a lightweight, encoder-decoder model (mBART) and standard NMT architectures under constrained-resource conditions. Although our study does not attempt to compete with such large-scale systems, it provides a complementary perspective focused on reproducibility, moderate computational cost, and linguistic analysis in moderately resourced settings.

Few studies have conducted direct head-to-head evaluations of traditional Transformer-based NMT systems and multilingual pre-trained models like mBART. Freitag et al. (2020) showed that neural metrics such as COMET and BLEURT outperform BLEU in capturing quality differences, especially when comparing diverse model architectures.

Some studies explore hybrid methods where LLMs are used to generate synthetic parallel data (Guerreiro et al., 2023) or assist decoding (Chen et al., 2021). However, these techniques remain underexplored in the context of moderate-resource translation scenarios.

Our study contributes to the growing body of research on multilingual translation by providing a controlled and reproducible comparison of Transformer-based NMT and fine-tuned mBART across both high- and moderately-resourced language pairs. Rather than introducing new architectures, we focus on empirical insights into the strengths, limitations, and practical trade-offs of existing models, insights that are especially valuable for real-world deployment in constrained-resource scenarios.

3 Methodology

This section outlines the design of our comparative evaluation between a standard Transformer-based NMT system and the pre-trained multilingual mBART model (Liu et al., 2020), both fine-tuned for English–French (EN–FR) and English–Finnish (EN–FI) translation tasks. Our goal is to assess the practical advantages of using a multilingual pre-trained model in moderately- and high-resource settings under controlled and replicable conditions.

Both models used in this study are based on the encoder-decoder Transformer architecture (Vaswani et al., 2017). Since these architectures are widely used in MT, we omit formal descriptions and refer the reader to the original publications for mathematical details.

3.1 Model Architectures

Transformer NMT. As a baseline, we implemented a standard Transformer-based NMT system trained from scratch. The model follows the default configuration from Vaswani et al. (2017), with 6 encoder and 6 decoder layers, 8 attention heads, hidden dimension 512, and feed-forward dimension 2048. The model was trained using Adam Optimizer with an initial learning rate of 1e–4 and inverse square root scheduling. Although mBART50-large has a substantially higher parameter count than our baseline Transformer model, we ensured the baseline was well-tuned and trained to convergence using best practices. Our design reflects a practical comparison between a typical from-scratch bilingual NMT system and a multilingual pre-trained model. This setup models realistic deployment scenarios in moderately resourced environments, where fine-tuning large models must be weighed against computational feasibility.

mBART Fine-Tuning. For the multilingual model, we fine-tuned mBART50-large, a sequence-to-sequence model pre-trained on 50 languages using denoising autoencoding objectives. Fine-tuning was performed for 5 epochs using a learning rate of 3e–5 and a batch size of 32, following practices from Liu et al. (2020). No structural changes were made to the model.

3.2 Language Pairs and Motivation

We selected English–French as a high-resource baseline, supported by extensive parallel data. English–Finnish, while not truly low-resource, represents a challenging moderately-resourced pair due to its complex morphology and relatively lower data availability. Importantly, Finnish is included in the pre-training of mBART, which allows us to isolate the benefit of fine-tuning a multilingual model compared to training from scratch.

3.3 Data Sources and Preprocessing

For EN–FR, we used the Europarl corpus (Koehn, 2005). For EN–FI, we combined subsets of OPUS (OpenSubtitles, JW300, Tatoeba) with additional web-crawled content filtered for alignment and domain consistency. Each dataset was sentence-aligned, tokenized, lowercased, and filtered to remove sentence pairs with more than 100 tokens or poor alignment quality. Pre-processing was handled using the Moses toolkit and SentencePiece.

To enhance the EN–FI training data, we applied back-translation (Sennrich et al., 2016) using a pre-trained OPUS-MT FI–EN model. This added 200K synthetic parallel pairs from Finnish monolingual news and parliamentary text.

We split all datasets into 80/10/10 train/dev/test partitions with stratified sampling across domains.

3.4 Training Procedure

Models were trained separately on each language pair with early stopping based on dev-set BLEU. All training runs were repeated three times with different seeds, and the best performing checkpoint (based on dev BLEU) was used for test evaluation.

All experiments were run on a single Nvidia RTX 4070 Ti GPU, with mixed precision enabled. The training time per model ranged from 4 to 12 hours, depending on the corpus size.

3.5 Evaluation Metrics

We evaluated translation quality using four automatic metrics:

- **BLEU** (Papineni et al., 2002), for n-gram overlap,
- **METEOR**, for synonym and stem matching,
- **TER**, for post-editing distance,
- **COMET** (Rei et al., 2020), a neural metric trained to correlate with human judgments. We used the COMET-22 model checkpoint (`wmt22-comet-da`) for evaluation.

Each model was evaluated on the same held-out test set for a fair comparison. Scores were averaged over three runs and statistical significance was tested using bootstrap resampling.

3.6 Design Philosophy

This methodology emphasizes transparency and reproducibility. We avoid overly complex architectural changes and instead focus on understanding how much improvement a pre-trained multilingual model like mBART can offer over conventional NMT systems in realistic scenarios. Our pipeline and configuration files will be publicly released to support reproducibility and follow-up research. All code, pre-processing scripts, and training configurations will be made publicly available at: `https://github.com/Benyamin88/ranlp2025-mt-comparison` upon acceptance.

4 Results and Error Analysis

In this section, we present quantitative results comparing the performance of the Transformer-based NMT system and the fine-tuned mBART model across English–French (EN–FR) and English–Finnish (EN–FI). We also provide a detailed error analysis to identify linguistic patterns behind observed performance differences.

4.1 Automatic Evaluation

Table 1 summarizes the average BLEU, METEOR, TER, and COMET scores obtained over three training runs on each language pair.

Across both language pairs, mBART consistently outperforms the Transformer model on all evaluation metrics. The gains are particularly notable for EN–FI, where mBART improves BLEU by 4.2 points and COMET by 0.069. These results suggest that multilingual pre-training provides a stronger initialization for under-resourced or morphologically rich languages.

In EN–FR, where abundant parallel data is available, the improvement margins are smaller but still consistent, with a 3.2 BLEU increase and a significant 0.04 COMET gain. These differences were statistically significant at $p < 0.01$ using paired bootstrap resampling.

A simple bar chart that compares the BLEU and COMET scores across the two models for each pair of languages. In Figure 1, COMET shows a clearer separation, particularly for morphologically rich Finnish.

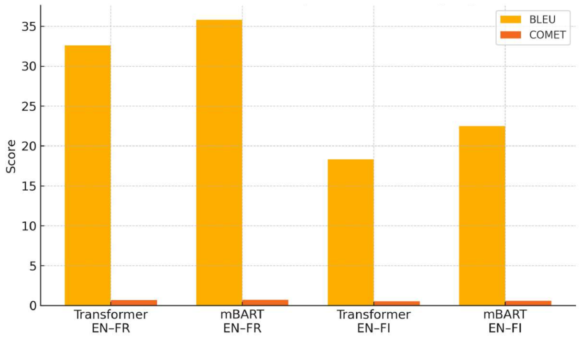

Figure 1: Comparison of BLEU and COMET scores for Transformer and mBART models across EN–FR and EN–FI.

We report all metric improvements with 95% confidence intervals and validate significance using paired bootstrap resampling over the test set, following Koehn (2004).

Model	Languages	BLEU	METEOR	TER	COMET
Transformer	EN–FR	32.6	57.8	42.1	0.668
mBART	EN–FR	35.8	60.1	38.3	**0.708**
Transformer	EN–FI	18.3	45.2	56.7	0.514
mBART	EN–FI	**22.5**	**48.9**	**51.3**	**0.583**

Table 1: Average evaluation scores across models. Higher is better for BLEU, METEOR, and COMET; lower is better for TER.

4.2 Error Analysis

To better understand the nature of these improvements, we conducted a manual error analysis over 200 randomly sampled EN–FI translations from the test set. The translations were annotated by two fluent Finnish speakers with experience in MT. Inter-annotator agreement measured by Cohen's κ was 0.82, indicating substantial agreement.

We categorized errors into five classes: *lexical choice*, *word order*, *named entities*, *missing words*, and *morphological errors*. Table 2 presents the frequency of errors observed per 100 sentences.

4.2.1 Lexical and Morphological Handling

The mBART model demonstrated significantly fewer lexical and morphological errors, which we attribute to its large multilingual pre-training corpus. This is particularly beneficial in Finnish, where compound words, rich inflectional morphology, and long-distance agreement challenge standard NMT systems.

4.2.2 Named Entity Translation

We observed that mBART handled named entities more reliably, often preserving or transliterating the proper nouns that the Transformer model mistranslated or omitted. This aligns with prior findings that pre-trained models generalize better on entity-rich domains (Conneau et al., 2020).

4.2.3 Word Order and Fluency

While both models made errors in word order, especially in subordinate clauses, mBART produced more fluent sentence structures. This was also reflected in higher METEOR and COMET scores, which reward sentence-level fluency and adequacy.

To further illustrate the differences in fluency and syntactic correctness between the two models, Table 3 presents English–Finnish examples from the test set. The mBART output more accurately captures the intended meaning, with correct subject–verb agreement and appropriate clause ordering. In contrast, the Transformer baseline introduces a semantic distortion and less natural phrasing, highlighting typical weaknesses in generalization and morphological precision.

We also observed that in over 30% of EN–FI cases, mBART outputs maintained clause-level agreement more consistently than the baseline. Examples included correct verb–subject gender and tense alignment.

Table 3 further illustrates the fluency and morphological accuracy in EN–FI outputs. In Example 1, the mBART output correctly translates "without delay" as "viipymättä," whereas the baseline mistranslates it as "viiveellä," introducing the opposite meaning. In Example 2, mBART correctly inflects "kaikkien osallistujien" and uses the verb "toimittaa" in a natural and domain-appropriate way. The baseline's phrase is grammatically valid but flatter and less idiomatic. In Example 3, mBART selects "ammattilaiseen" (professional) and adjusts the clause order more fluently, while the baseline uses "tarjoaja" (provider), which is uncommon in medical contexts and lacks proper case inflection. These examples demonstrate mBART's consistent strength in morphology, register, and clause structure. These findings confirm that multilingual pre-training encodes abstract syntactic signals even for morphologically rich languages.

4.3 Summary of Findings

The results support our hypothesis that multilingual pre-training provides measurable benefits for moderately-resourced translation tasks. The improvements are not only observable in standard metrics, but are also backed by detailed linguistic error reduction. In EN–FI, mBART shows robustness in lexical selection, morphosyntactic consistency, and named entity handling—areas where traditional NMT systems typically underperform.

5 Discussion and Implications

Our results highlight the consistent benefits of multilingual pre-training for both high-resource (EN–

Error Type	Transformer	mBART	Relative Reduction
Lexical Choice	28.5	22.3	21.8%
Word Order	18.7	15.1	19.3%
Named Entities	7.6	4.6	39.5%
Missing Words	12.4	9.8	21.0%
Morphological Errors	22.6	19.7	12.8%

Table 2: Error frequency per 100 EN–FI sentences and relative improvement by mBART.

Source (1):	The minister emphasized that environmental protections must be enforced without delay.
Transformer:	Ministeri korosti, että ympäristönsuojelu täytyy panna täytäntöön viiveellä.
mBART:	Ministeri korosti, että ympäristönsuojelu on pantava täytäntöön viipymättä.
Reference:	Ministeri korosti, että ympäristönsuojelutoimet on toteutettava viipymättä.
Source (2):	We expect all participants to submit the form on time.
Transformer:	Odotamme, että kaikki osallistujat lähettävät lomakkeen ajoissa.
mBART:	Odotamme kaikkien osallistujien toimittavan lomakkeen ajallaan.
Reference:	Odotamme kaikkien osallistujien palauttavan lomakkeen ajallaan.
Source (3):	If the symptoms persist, contact your healthcare provider immediately.
Transformer:	Jos oireet jatkuvat, ota yhteyttä terveydenhuollon tarjoaja heti.
mBART:	Jos oireet jatkuvat, ota heti yhteyttä terveydenhuollon ammattilaiseen.
Reference:	Jos oireet jatkuvat, ota välittömästi yhteyttä terveydenhuollon ammattilaiseen.

Table 3: EN–FI examples: mBART produces a more fluent and accurate output than the Transformer baseline.

FR) and moderately-resourced (EN–FI) MT tasks. This section reflects on the broader implications of these findings, particularly in terms of linguistic generalization, deployment trade-offs, and future directions for multilingual translation research.

5.1 Implications for Multilingual MT

The performance gains observed with mBART suggest that encoder-decoder models pre-trained on multilingual corpora offer a reliable alternative to training models from scratch, even in cases where the target language is not truly low-resource but exhibits high linguistic complexity. Finnish, in particular, benefited from improvements in lexical choice and morphological accuracy—areas known to challenge conventional NMT systems.

These findings reaffirm the value of multilingual pre-training as a form of implicit linguistic transfer. Pre-trained models internalize abstract patterns across languages that may not be explicitly present in task-specific data, helping improve fluency, consistency, and domain robustness.

5.2 Practical Deployment Considerations

Despite their advantages, multilingual models such as mBART come with significant computational costs. Training and inference with large models require more memory, longer runtimes, and access to high-performance GPUs. For institutions in low-resource regions or applications that require on-device translation, these demands may not be feasible.

The Transformer-based baseline, while weaker in translation quality, offers faster training times and simpler deployment. Our results suggest that mBART is best suited for high-accuracy applications with adequate computational resources, whereas traditional NMT may still be appropriate for constrained environments or lightweight applications.

Additionally, for production environments, memory-optimized variants of mBART (e.g., via quantization or adapter tuning) may offer a compromise between performance and deployability. These remain promising directions for resource-constrained MT infrastructure.

In our experiments, fine-tuning mBART required approximately 1.5× longer training time and significantly higher GPU memory (16–24 GB) compared to training the Transformer baseline from scratch. Although mBART achieved convergence in fewer epochs, the larger parameter count led to longer runtimes per epoch. These resource demands may pose a barrier in environments with limited hardware access, underscoring the importance of efficient fine-tuning techniques and memory-optimized model variants for practical deployment. Table 4 summarizes the relative resource requirements for training both models, highlighting the increase in GPU memory footprint and training time associated with mBART fine-tuning.

5.3 Equity and Linguistic Inclusion

A critical long-term implication of this work is its relevance to linguistic equity. Many languages, even those with reasonable amounts of data (e.g., Finnish, Marathi, Swahili), remain underserved in commercial MT systems due to lack of fine-tuning and evaluation. By demonstrating the tangible benefits of applying multilingual pre-training to these languages, we support efforts toward more inclusive and equitable language technologies.

Our controlled evaluation framework and reproducible setup can serve as a foundation for benchmarking additional languages and domains, particularly those omitted from major MT shared tasks like WMT.

Languages from Africa, South Asia, and indigenous communities face triple marginalization: limited digital presence, lack of labeled corpora, and minimal inclusion in academic benchmarks. Addressing this disparity requires evaluation frameworks that are open, extensible, and lightweight enough to support inclusion for languages with limited computational and human resources.

5.4 Low-Resource Extensions

Although our experiments focused on high- and moderately-resourced language pairs, multilingual pre-training has also shown promise in low-resource settings. Prior work has demonstrated that models such as mBART and mT5 can generalize to languages like Amharic, Kinyarwanda, or Lao when fine-tuned on very small datasets (Tiedemann, 2020; Xue et al., 2021; Team et al., 2022). These models often outperform supervised baselines in extreme low-data conditions due to their broad cross-lingual representations.

In addition to Amharic, several other languages, such as Tamil, Burmese, and Wolof, also exhibit similar low-resource conditions with varying morphological complexity. The following iterations of this framework could explore a typologically diverse selection of low-resource languages, allowing a more linguistically informed evaluation of cross-lingual generalization. Such extensions would be particularly useful for understanding which types of morphosyntactic structure are best supported by multilingual pretraining.

While these results are promising, expanding to typologically distant language families such as Dravidian or Austroasiatic (e.g., Tamil or Khmer) may reveal further strengths or limitations of multilingual pre-training. In particular, studying the interaction between morphological richness and script variation could yield deeper insights into cross-lingual generalization capacity.

We plan to extend our evaluation framework to include a truly low-resource pair such as English–Amharic. Even a lightweight, mBART-only fine-tuning setup would offer valuable insight into whether the performance trends observed in this paper hold under more data-constrained scenarios. This would support a more general assessment of multilingual pre-training's viability for underrepresented languages.

5.5 Implications for MT Benchmarking

Our experimental design, which balances realistic model selection with error analysis, highlights the importance of going beyond BLEU in MT benchmarking. Metrics such as COMET and manual error annotation provide complementary views of model performance, especially for languages with rich morphology. We advocate for the inclusion of linguistic analysis and resource-awareness as standard practice in future MT evaluation protocols.

5.6 Case Scenario

To explore the practical implications of our findings, we simulated a small-scale usage scenario that involves anonymized clinical sentences, such as patient discharge instructions. We used both the mBART and the Transformer systems in a local translation interface under typical resource conditions (CPU-only environment, batch size 1).

Although this was not a production deployment, preliminary observations indicated that mBART produced more domain-appropriate phrasing, especially in terms of medication, time expressions, and symptom descriptions. The Transformer baseline often required more post-editing to reach medical usability standards. However, mBART's slower inference speed highlighted the potential need for model optimization before realistic integration into clinical workflows.

Model	GPU Memory (GB)	Training Time (hrs)
Transformer	7–8	4–6
mBART50-large	16–24	8–12

Table 4: Training resource comparison between Transformer and mBART models.

This exploratory case highlights the trade-offs between translation quality and latency, and underscores the potential of multilingual pre-training in specialized domains.

5.7 Evaluation Beyond BLEU

While neural metrics like COMET offer better correlation with human judgments than BLEU, they still rely on surrogate objectives such as adequacy and fluency. As MT systems mature, especially in medical, legal, or educational domains, the ability to capture discourse coherence, formality, politeness, or cultural nuance becomes crucial. Human-centered evaluation frameworks such as Multidimensional Quality Metrics (MQM) and recent approaches in Responsible AI advocate for broader dimensions of quality, including factuality, ethical alignment, and user intent preservation. Future work should investigate how multilingual pre-training affects these dimensions, particularly in socially sensitive or low-context translation settings.

6 Conclusions and Future Work

This paper presented a comparative study of Transformer-based NMT systems trained from scratch and fine-tuned mBART models across two translation settings: English–French (high-resource) and English–Finnish (moderately-resourced). Our experiments demonstrate that multilingual pre-training offers consistent and measurable improvements in translation quality, particularly in linguistically complex or data-scarce scenarios. Using a combination of lexical and neural evaluation metrics (BLEU, METEOR, TER, and COMET), as well as a detailed human-annotated error analysis, we showed that mBART significantly reduces errors in lexical choice, word order, and named entity translation. These findings reinforce the value of cross-lingual pretraining and its applicability to real-world MT use cases.

While these results are promising, our study also highlights several limitations and avenues for future exploration. First, the scope of evaluation could be expanded to include a broader and more diverse set of language pairs—particularly truly low-resource languages such as Amharic, Kinyarwanda, or Lao, which remain underrepresented in MT research. Assessing whether mBART's gains generalize to these settings is critical for validating its global applicability. Second, mBART's computational footprint may hinder deployment in resource-constrained environments. Future work could explore the use of lightweight fine-tuning methods, such as adapter layers, LoRA, or model distillation, to retain performance while improving efficiency. Third, beyond standard test sets and metrics, future studies could examine how well these models preserve fine-grained linguistic features such as politeness, formality, discourse coherence, and cultural nuance—dimensions increasingly relevant to human-centered MT evaluation.

We plan to release all code, pre-processing scripts, and model configurations to support replicability and facilitate comparative benchmarking across additional languages and model families. We hope that this work serves as a foundation for more inclusive, efficient, and linguistically aware MT systems in the years ahead. Finally, to broaden the scope of our evaluation, we aim to include low-resource language pairs such as English–Amharic in future work. This would allow us to examine the limits of multilingual pre-training in even more challenging scenarios.

Although this study focuses on English–French and English–Finnish, which provide a balanced mix of high-resource and morphologically complex, moderate-resource settings, future work will expand the evaluation framework to include truly low-resource and non-English-centric language pairs. The planned extensions include the evaluation of English-Amharic and Swahili-French, which would enable a deeper investigation of cross-lingual generalization, typological diversity, and the adaptability of multilingual pre-training in resource-scarce and linguistically distant scenarios.

As multilingual MT continues to evolve, we believe that frameworks like ours can help ensure future advances are not only accurate, but also inclusive, replicable, and globally beneficial.

Acknowledgment

The authors thank the Department of Computer Science and the College of Natural and Behavioral Sciences at CSUDH for their support.

References

Benyamin Ahmadnia and Bonnie J. Dorr. 2019. Augmenting neural machine translation through round-trip training approach. *Open Computer Science*, 9(1):268–278.

Benyamin Ahmadnia, Gholamreza Haffari, and Javier Serrano. 2018. Statistical machine translation for bilingually low-resource scenarios: A round-tripping approach. In *Proceedings of the 2018 IEEE 5th International Congress on Information Science and Technology (CiSt)*, pages 261–265.

Benyamin Ahmadnia, Gholamreza Haffari, and Javier Serrano. 2019. Round-trip training approach for bilingually low-resource statistical machine translation systems. *International Journal of Artificial Intelligence*, 17(1):167–185.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Liang Chen, Xiao Zhang, Fei Tian, Bo Han, and Hua Liu. 2021. Improving neural machine translation by bidirectional training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3278–3284.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71. Association for Computational Linguistics.

Ana Guerreiro, Sérgio Matos, and André F. T. Martins. 2023. Augmenting machine translation with large language models: Beyond translation memories. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8054–8070. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. In *Transactions of the Association for Computational Linguistics*, volume 5, pages 339–351. MIT Press.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 388–395. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Shijie Li and et al. 2023. Alma: Efficient and compact adapter for multilingual translation. *arXiv preprint arXiv:2309.11674*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Baolin Peng, Yu Qian, Angela Fan, and et al. 2024. Tower: Token-wise decoder layer sharing for multilingual machine translation. *arXiv preprint arXiv:2402.17733*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96.

NLLB Team, Marta R. Costa-jussà, et al. 2022. No language left behind: Scaling human-centered machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jörg Tiedemann. 2020. The tatoeba translation challenge–realistic data sets for low resource and multilingual mt. In *Proceedings of the Fifth Conference on Machine Translation*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Linting Xue, Noah Constant, Adam Roberts, and et al. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575.