Tao–Filipino Neural Machine Translation: Strategies for Ultra–Low-Resource Settings

Adrian Denzel Macayan[*], **Luis Andrew Madridijo**[*],
Ellexandrei Esponilla, Zachary Mitchell Francisco
De La Salle University-Manila
Manila, Philippines
{adrian_macayan, luis_madridijo}@dlsu.edu.ph
{ellexandrei_esponilla, zachary_francisco}@dlsu.edu.ph

Abstract

Neural Machine Translation (NMT) performance degrades significantly in ultra-low-resource settings, particularly for endangered languages like Tao (Yami) which lack extensive parallel corpora. This study investigates strategies to bootstrap a Tao-Tagalog translation system using the NLLB-200 (600 million parameter) model under extremely limited supervision. We propose a multi-faceted approach combining domain-specific fine-tuning, synthetic data augmentation, and cross-lingual transfer learning. Specifically, we leverage the phylogenetic proximity of Ivatan, a related Batanic language, to pre-train the model, and utilize dictionary-based generation to construct synthetic conversational data. Our results demonstrate that transfer learning from Ivatan improves translation quality on in-domain religious texts, achieving a BLEU score of 34.85. Conversely, incorporating synthetic data enhances the model's ability to generalize to conversational contexts, mitigating the domain bias often inherent in religious corpora. These findings highlight the effectiveness of exploiting linguistic typology and structured lexical resources to develop functional NMT systems for under-represented Austronesian languages.

1 Introduction

Neural Machine Translation (NMT) systems have achieved substantial success for high-resource language pairs, largely due to the availability of extensive parallel and monolingual corpora, which often contain millions of sentence-aligned examples (Bahdanau et al., 2015; Vaswani et al., 2017). The abundance of such resources enables NMT models to process large datasets, learn robust cross-lingual representations, and produce high-quality translations across diverse domains. However, the majority of the world's languages—beyond English, French, Chinese, and a few others—lack such extensive parallel data. In particular, endangered and low-resource languages often possess only a few thousand computer-accessible sentence pairs, severely limiting the effectiveness of standard NMT training pipelines (Haddow et al., 2022). Recent multilingual and massively pre-trained sequence-to-sequence models have broadened NMT capabilities to low-resource settings. By sharing parameters across hundreds of languages, these models induce partially language-agnostic representations that enable zero-shot generalization (Costa-jussà et al., 2024). Nevertheless, performance remains uneven and inconsistent for many low-resource languages. Ultra-low-resource languages—those with limited corpora, no presence in pre-training data, or distinctive morphological characteristics—still experience degraded translation quality. This gap highlights persistent inequities in access to translation tools and underscores the need for improved strategies that go beyond traditional data scaling.

At the same time, many low-resource or "data-poor" languages possess extensive linguistic documentation accumulated through long-term fieldwork, community initiatives, and academic research, which can be leveraged computationally. Such resources—grammars, dictionaries, religious translations, and transcribed oral traditions—contain structured linguistic knowledge not captured by conventional corpora. Tao (Yami), a Batanic language spoken on Orchid Island, Taiwan, exemplifies this scenario. Although Tao lacks the large-scale parallel corpora typical of mainstream NMT approaches, it maintains rich lexical resources, ranging from Biblical translations to community-authored texts. These circumstances raise an important research question: can structured linguistic resources meaningfully compensate for data scarcity when developing NMT systems for endangered languages? To address this question, the present work investigates ultra-low-resource strategies for bootstrapping a Tao–English transla-

[*] Equal contribution

Proceedings for the Ninth Workshop on Technologies for Machine Translation of Low Resource Languages (LoResMT 2026), pages 27–36
March 28, 2026 ©2026 Association for Computational Linguistics

tion system under extremely limited supervision (fewer than 5,000 parallel sentences). Instead of relying solely on data volume, we focus on three complementary approaches that leverage linguistic structure and other indicators to reduce degradation and improve translation quality. Ultimately, this study proposes a scalable framework for transforming static linguistic archives into functional translation models, offering a roadmap for other under-represented Austronesian language.

Primary Corpus Construction We compile, digitize, and standardize a parallel Tao–English corpus drawn from diverse domains. Our primary sources include religious texts (specifically the Tao translation of the New Testament), educational publications, and community-authored narratives.

Synthetic Data Augmentation To address the scarcity of authentic parallel data, we employ two complementary data augmentation strategies: **Dictionary-Assisted Generation:** We generate synthetic sentence pairs by employing the comprehensive Tao–Chinese–English dictionary and morphological rules provided by Rau and Dong (2006). We map high-frequency lexical items and grammatical constructions to their target equivalents, expanding the coverage of the training data to include morphological patterns commonly found in conversational Tao that is not fully represented in the Bible corpus. **Pivot-Based Augmentation:** We supplement the limited authentic corpora by utilizing Mandarin and Tagalog as pivot languages. We utilize commercial translation systems (e.g., Google Translate) to translate Mandarin resources into Tagalog and English, creating pseudo-parallel pairs that align with our Tao data. This approach increases data diversity and introduces semantic variations that help the model generalize beyond the specific domains of the primary corpus.

Transfer Learning Strategy We adopt a transfer learning approach to leverage the phylogenetic relationships within the Austronesian language family. We initialize our Neural Machine Translation (NMT) models using weights pre-trained on high-resource languages. We specifically fine-tune on typologically related Batanic languages (such as Ivatan) and regionally dominant languages (such as Tagalog) before adapting to Tao. This curriculum learning strategy allows the model to leverage shared morphological features and cognates, with the intent to significantly enhance translation performance in ultra-low-resource conditions.

2 Related Works

Multilingual and Transfer-Learning Approaches for Low-Resource NMT: A well-established strategy for improving NMT for low-resource languages is leveraging multilingual and transfer-learning techniques. Early foundational work demonstrated that NMT models trained on high-resource language pairs can provide robust parameter initializations that improve translation quality when fine-tuned on low-resource pairs (Zoph and Knight, 2016). Sharing sub-word vocabulary and morphology between source and target languages further boosts performance when the languages are related, based heavily on the principle of linguistic proximity (Nguyen and Chiang, 2017). Recent studies explore meta-learning approaches for adaptation to low-resource languages, demonstrating that models can achieve competitive BLEU scores with as few as 600 parallel sentences (Gu et al., 2018). However, survey work confirms that these methods struggle when the target language is unseen in pre-training or is typologically distant from high-resource languages (Haddow et al., 2022; Costa-jussà et al., 2024).

Domain Divergence and "Auxiliary Fine-Tuning: A significant, yet often overlooked, challenge in this transfer learning paradigm is Domain Divergence. In ultra-low-resource settings, researchers are often forced to rely on auxiliary domains—most commonly religious texts— that diverge significantly from the target application of daily conversation. Ranathunga et al. (2024) explicitly addressed this in Exploiting Domain-Specific Parallel Data, investigating the impact of domain mismatch on Multilingual Sequence-to-Sequence Language Models (msLMs). Their study confirmed that while msLMs provide strong initialization, they fail to generalize when fine-tuned solely on divergent auxiliary data due to substantial lexical distribution mismatches. Crucially, they propose "auxiliary fine-tuning"—adapting the model to a high-resource related language in the target domain (e.g., Tagalog conversational text) prior to the final low-resource adaptation—as a mechanism to bridge this semantic gap.

Ultra-Low-Resource NMT: In scenarios where both parallel and monolingual corpora are extremely limited, semi-supervised and unsupervised approaches such as back-translation or noise-augmented self-training are often applied (Sennrich et al., 2016). However, these methods are constrained by the scarcity of usable monolingual data, especially in endangered or under-documented languages. Studies indicate that ultra-low-resource settings (fewer than 5,000 sentence pairs) require additional techniques beyond standard semi-supervised NMT to achieve reliable translation quality (Aharoni et al., 2019).

The Shift to Large Language Models (LLMs): The period from 2024 to 2026 has witnessed a paradigm shift from specialized encoder-decoder architectures (like NLLB) to general-purpose Large Language Models (LLMs). The Lin et al. (2025) study established the first comprehensive benchmark for Formosan languages (Atayal, Seediq, Paiwan), which share significant phylogenetic proximity to Tao (Yami). Their findings reveal a complex landscape: while off-the-shelf LLMs initially struggle with the VSO word order and focus systems typical of Austronesian languages, they exhibit remarkable few-shot learning capabilities when prompted with high-quality dictionary definitions. This suggests that the future of ultra-low-resource NMT may lie not in training from scratch, but in Parameter-Efficient Fine-Tuning (PEFT) of massive pre-trained models. Techniques such as Low-Rank Adaptation (LoRA) allow for the adaptation of large models (7B+ parameters) on consumer hardware by updating less than 1% of the parameters, effectively mitigating the risk of "catastrophic forgetting" often observed when over-training on tiny datasets.

Dictionary-Based Data Augmentation: Dictionary-based augmentation has emerged as a viable strategy for extremely low-resource MT. By leveraging bilingual lexica, morphological patterns, or rich wordlists, synthetic parallel sentence pairs can be generated to improve coverage of rare vocabulary and grammatical constructions (García et al., 2019; Zhang et al., 2023). These approaches are particularly useful when monolingual corpora are insufficient for self-supervised methods, and they complement transfer learning from related languages. Furthermore, recent scholarship argues that quantity of data is secondary to cultural and semantic fidelity. Lovenia et al. (2024) introduced SEACrowd, a collaborative data hub specifically for Southeast Asian languages. Unlike global datasets which often suffer from "translationese," SEACrowd standardizes corpora across nearly 1,000 indigenous languages, explicitly addressing the "cultural misrepresentation" prevalent in Western-centric models. This aligns with the release of Oepen et al. (2025), which employs advanced language identification filters to recover usable monolingual data for languages previously discarded as noise. For Tao, this implies that augmenting training data with "culturally aware" synthetic sentences—derived from folklore or community narratives rather than generic web text—is essential for preserving semantic nuance.

MT for Endangered, Indigenous, and Austronesian Languages: Recent work highlights translation efforts for endangered and indigenous languages, including South American, North American, and Austronesian languages (Cardoso et al., 2022; Rodríguez et al., 2023). Such studies emphasize the importance of domain-specific corpora (e.g., religious or educational texts) and careful selection of primary data for transfer learning. While Austronesian languages such as Hawaiian, Māori, and Ivatan have been examined in the context of MT, few studies explicitly address the Yami (Tao) language. However, significant foundational work in Yami computational linguistics exists. Yang et al. (2010) proposed a model for constructing a Yami WordNet, creating a crucial lexical database aligned with English semantic concepts. Subsequent studies expanded this into ontological resources, including an integrated semantic network for *ka-* verbs (Yang et al., 2011) and a computational analysis of Yami emotion phrases (Yang et al., 2012). These computational lexical resources provide a structured basis that can be leveraged and expanded on to enhance machine translation performance for the Tao language.

Critical Synthesis and Limitations of Existing Methodologies: While the literature provides robust individual strategies for low-resource NMT, a critical synthesis reveals a distinct gap in their application to Batanic languages. Existing transfer learning approaches predominantly focus on high-resource pivots (like English or Spanish), often neglecting the potential of "phylogenetic transfer" from closely related, yet still low-resource, sister languages (such as Ivatan). Furthermore, while dictionary augmentation is widely proposed, there is a lack of empirical research on how this specifically interacts with "Liturgical Bias"—the tendency of models trained on Bible corpora to hallucinate archaic formality in casual settings. Current methodologies largely fail to address the "Domain-Register Incongruence" that occurs when a model pre-trained on modern web text (Tagalog) is fine-tuned on archaic religious text (Tao Bible), and then expected to translate daily conversation. This study aims to address this specific intersection: bridging the gap between phylogenetic transfer learning and domain-adaptive synthetic data generation to construct a functional NMT system for an ultra-low-resource, endangered Austronesian language.

3 Language, Data and General Setup

The Yami Language: Yami, autonymically known as Tao, is an Austronesian language spoken by the indigenous Tao people of Orchid Island (Lanyu), located off the southeastern coast of Taiwan. Despite its geographic proximity to the main island of Taiwan, Yami is phylogenetically classified within the Batanic branch of the Malayo-Polynesian family, rather than the Formosan languages (Smith, 2017). Based on lexical innovations and archaeological evidence, Smith (2017) groups Batanic languages with the Northern Luzon and Greater Central Philippine languages, distinguishing them from other Malayo-Polynesian branches.Typologically, Yami exhibits the structural characteristics of a Philippine-type language (Rau and Dong, 2006). It features a dominant Verb-Initial (VSO) word order and a rich agglutinative morphology. A defining grammatical feature is its complex focus system, where verbal affixes—including prefixes, infixes, and circumfixes—signal the semantic relationship (Actor, Patient, Locative, or Instrumental) between the verb and the focused noun phrase. Currently, the language is classified as endangered, with inter-generational transmission threatened by the widespread adoption of Mandarin Chinese.

Dictionary: The primary lexical resource for the language is the *Yami Texts with Reference Grammar and Dictionary*, compiled by Rau and Dong (2006). This comprehensive volume provides Yami lexical entries with definitions in both English and Chinese, capturing the nuances of the language as spoken on Orchid Island.

Parallel Corpus: As Yami is a ultra-low-resource language, large-scale parallel corpora are not readily available. The parallel data used in this study primarily consists of the texts collected by Rau and Dong (2006), which include cultural narratives and daily conversations aligned with Tagalog and Chinese translations. Furthermore, we utilize the Yami translation of the New Testament Bible, which provides a substantial number of aligned sentence pairs. These sources were manually cleaned and aligned to create a Yami-Tagalog parallel dataset suitable for training and validation.

Model: To address the data scarcity of the Yami language, we employ a transfer learning approach using NLLB-200 (No Language Left Behind), a multilingual machine translation model developed by NLLB Team et al. (2022). Unlike general-purpose Large Language Models (LLMs), NLLB-200 utilizes a Transformer-based encoder-decoder architecture explicitly optimized for translation across 200+ languages. We utilize the 600-million parameter variant as our foundational base. The model is particularly well-suited for this task as it was pre-trained on a massive dataset including several Austronesian and Philippine-type languages (e.g., Tagalog, Ilokano, Cebuano, and Pangasinan) that share typological and genealogical features with Yami. We fine-tune the model on the Yami-English parallel corpus, leveraging the model's pre-existing cross-lingual representations to improve alignment and translation generation quality in an ultra-low-resource setting.

Evaluation Set: We establish a hybrid evaluation benchmark to assess performance on the Yami-Tagalog language pair. The set consists of (1) a Cognate Challenge Set: 20 manually curated sentence pairs derived from Rau and Dong (2006), where Yami terms are mapped to Tagalog equivalents and verified via morphological analysis of shared Austronesian cognates; and (2) a Religion Domain Set: 200 verse-pairs randomly sampled from the parallel Bible corpus. We have made sure that the parallel corpus and the evaluation set do not overlap.

Evaluation Metrics We evaluate the quality of our translation models using a combination of lexical, morphological, and semantic metrics. First, we report BLEU (Papineni et al., 2002), the standard metric for n-gram overlap, to provide a baseline comparable to existing literature. Second, given the rich morphology of the Yami language, we employ chrF++ (Popović, 2015), a character n-gram F-score. Unlike word-level metrics, chrF++ is less sensitive to tokenization errors and captures morphological accuracy in languages with complex affixation. Finally, to assess semantic fidelity beyond lexical overlap, we utilize SBERT (Sentence-BERT) (Reimers and Gurevych, 2019). We compute the cosine similarity between the embeddings of the generated translation and the reference text, which credits translations that are semantically correct even if they diverge lexically from the standard.

4 Methodology

4.1 Extracting the Bible Parallel Corpus

To extract the text required for machine translation of Yami to other languages, we decide to source the text from the various Philippine language translations of the Bible. As a key religious text, it is readily available as computer-readable data and sufficient size for a ultra-low-resource NMT model. Table 1 contains select Philippine languages and the respective source Bible edition that serves as the basis of the primary corpus for each language.

Table 1: List of Languages and Bible Versions

Language	Bible Version / Source
Bikolano	Marahay na Bareta Biblia
Ilokano	Ti Baro a Naimbag a Damag Biblia
Ivatan	VTSP (Bible.com)
Pangasinan	Maung A Balita Biblia
Tagalog	Ang Biblia (2001)
Yami	Seysyo No Tao

4.2 Selecting a Target Language for Yami Machine Translation

We select Tagalog as the target language of the Yami translation to strengthen general Filipino-based translation models and provide a versatile real-world use case, given its dominance in the country's capital, major urban and suburban areas, namely Metro Manila and the CALABARZON region. Furthermore, the presence of multilingual translation model, with regards to the Philippine branch, is centered on Tagalog—one of the most common representations on NLP platforms such as HuggingFace (Hugging Face, 2025).

4.3 Data Preprocessing

The parallel corpus for improving improve and fine-tuning the selected model consisted of aligned Bible verse pairs for Yami-Tagalog. In accordance with standard best practices in machine translation and text normalization, all alphabetic characters are converted to lowercase to reduce vocabulary sparsity and simplify the model's learning space. Whitespace inconsistencies were also standardized to avoid unintended tokens that could negatively affect tokenization quality. After normalization, all verse pairs were examined for alignment completeness. Verses in one language that did not have a corresponding verse in the paired language were removed to maintain a strictly parallel corpora, ensuring that every training example consists of a clean and fully aligned source–target pair. After filtering, the textual data was passed through a tokenizer, with each verse constrained to a maximum length of 128 tokens. This constraint follows common practice in transformer-based MT systems, which typically limit input lengths for efficient batch processing (Ahmad et al., 2024). Any verse exceeding 128 tokens was truncated to preserve uniformity in sequence length and computational feasibility.

4.4 Creating a Synthetic Yami-Tagalog Parallel Corpus

By using only the Bible parallel corpus to train the base NLLB-200 Model, there is a risk of the model being biased towards the often archaic sentence structure and vocabulary of Bible verses. In order to prevent the model from being limited to performance in translating Bible verses, we elect to obtain parallel sentences in different context outside of faith-based circumstances. While direct Yami-to-Tagalog human translation of sentences are severely lacking in resource, Yami-to-Chinese parallel corpora are more readily available, such as in the work of (Indigenous Languages Research and Development Foundation, 2020). For our synthetic corpus, we chose the parallel sentences provided by the dictionary because it contained conversational sentences, such as those common in classroom instruction and family settings. For the purposes of creating a parallel corpus, having Chinese as one of the languages allows effective use of other MT tools, such as Google Translate to translate the Chinese translation of Yami sentences to Tagalog. The tool is comparatively an improvement in terms translation quality over the local model of NLLB-200. Examples of NLLB-200's poor Chinese-to-Tagalog translation quality are shown in Table 2.

4.5 Improving Yami-to-Tagalog translation using Synthetic Data

Initially, only the Bible parallel corpus trained our first Yami-to-Tagalog model (Model 1). To further improve the initial model, we supplement the training data from the Bible parallel corpus with the synthetic data Yami-Tagalog parallel corpus created from the Yami-Chinese parallel corpus sourced from Indigenous Languages Research and Development Foundation (2020). With this new training, NLLB-200 model trained trained using the same initial hyperparameters to set comparison with the previous model against Model 2.

Table 2: Comparison of Yami Translations

Yami	**Chinese**	**English**	**Tagalog (NLLB-200 600M)**	**Tagalog (Google Translate)**
kokey kamo papoen tana	大家好。我始了。	Hello everyone. Here we go.	Bigyang mabuti... tayo ay nagsisimula...	Hello sa lahat. Magsimula tayo.
tanek kamo	起立! 敬!	Please stand up! Salute!	起立! 敬礼!	Mangyaring tumayo! Salute!
akokey a sinsi namen	老好。	Hello.	Teacher mabuti.	Hello, guro.
lisna kamo	坐下。	Please sit down.	坐下。	Mangyaring umupo.

However, considering that the synthetic data, being only 3966 in sentence count, is outnumbered by the real data (10498). To prevent the Bible data from dominating the conversational and cultural data in the synthetic parallel corpus, we duplicated the synthetic data three times, i.e. 3966 original sentences × 3 = 11898 total synthetic sentences. We chose the × 3 multiplier because it roughly equalize the count between the real and synthetic data. Moreover, further shuffling of the synthetic data and the Bible verses was done (Xu et al., 2019) to prevent catastrophic interference caused by feeding one data source followed by a new and separate data source (van de Ven et al., 2024).

4.6 Improving Yami-to-Tagalog translation using a Pretrained Model

Additionally, we performed pretraining of the base NLLB-200 model on Ivatan-to-Tagalog because Ivatan is another Batanic language. Although the NLB-200 model has training on some Philippine languages, it does not have any training on a Batanic language. By training the initial model on Ivatan-To-Tagalog, the model will have some initial representation of a closely related Batanic language before being introduced to Yami, which may improve the translation quality of the final model (Model 3). After training the model on Ivatan, we produce another fine-tuning it towards Yami using the Yami-to-Tagalog Bible parallel corpus and the synthetic Yami-Tagalog corpus.

5 Analysis of Performance

Table 3 presents the quantitative results across our three experimental configurations. We assess performance using BLEU, chrF++, and SBERT metrics on two distinct test sets: the in-domain *Bible Test Set* (200 pairs) and the out-of-domain *Validation Set* (20 pairs), which consists of conversational and daily-life vocabulary Yami text.

The Effectiveness of Language Transfer (Model 3): Our results provide strong empirical evidence for the utility of phylogenetic transfer in ultra-low-resource settings. Model 3, which utilizes pre-training on the closely related Ivatan language before fine-tuning on Yami, achieved the highest performance on the in-domain Bible dataset, securing a **BLEU score of 34.85** and a **chrF score of 58.36**. This represents a marked improvement over the baseline Model 1, which achieved a BLEU score of 32.00. We attribute this gain to the specific genealogical relationship between Ivatan and Yami, both of which belong to the Batanic branch of the Malayo-Polynesian family. Unlike general multilingual pre-training, initializing weights with Ivatan allows the model to "warm start" with a high degree of relevant morphosyntactic alignment. The significantly higher chrF score (58.36 vs 56.26) is particularly telling; since chrF operates at the character n-gram level, it suggests that Model 3 is far more successful at generating the correct agglutinative affixes and reduplication patterns shared by Batanic languages, even when exact lexical matches are unavailable. This confirms that in the absence of massive parallel corpora, exploiting the structural priors of a sister language is a potent strategy for stabilizing the decoder.

Generalization vs. Domain Specialization (Model 2): The performance of Model 2, which augments the training set with synthetic dictionary-based data, illustrates a critical trade-off between domain specialization and robust generalization. On the specific Bible test set, Model 2 experienced a slight degradation in performance, with the BLEU score dropping to **30.97** compared to the baseline's 32.00. This dip is attributable to the dilution of the model's probability distribution; by introducing conversational data, the model is

Table 3: Performance Comparison: Initial vs. Fine-Tuned vs. Pretrained (Ivv) Models

	Model 1		**Model 2**		**Model 3**	
Metric	**Bible**	**Val.**	**Bible**	**Val.**	**Bible**	**Validation**
BLEU	32.00	3.12	30.97	6.67	34.85	3.64
chrF	56.26	28.28	55.16	34.01	58.36	32.29
Median SBERT	92.26	82.77	92.55	79.80	90.54	77.94

less over fitted to the specific idiolect and archaic sentence structures of the religious text. However, this minor trade-off in the source domain yielded disproportionate gains in generalization. On the Validation Dataset—which mimics real-world usage—Model 2 effectively doubled the translation quality, achieving a **BLEU score of 6.61** compared to the baseline's 3.12. Furthermore, it achieved the highest chrF score on the validation set (**34.01**) among all models. This indicates that the inclusion of synthetic conversational pairs successfully mitigates the bias, allowing the model to recover basic conversational structures (e.g., greetings, imperatives) that are statistically rare in the corpus but remain essential for a functional translation system.

Semantic Divergence and Formality Bias (SBERT Analysis): The SBERT (Sentence-BERT) semantic similarity scores reveal a nuanced limitation of pure transfer learning regarding register and formality. While Model 3 was the strongest structural translator (highest BLEU), it recorded the *lowest* semantic similarity score on the conversational Validation set (**SBERT 77.94**), significantly lower than Model 1 (82.77). We postulate that this is driven by a "Formality Bias." Both the Ivatan pre-training data and the Yami Bible data consist of high-register, formal, and often archaic language. Consequently, when presented with a casual input from the Validation set, Model 3 is predisposed to generate a formally rigid or archaic Tagalog equivalent. While these translations may be grammatically sound, they diverge semantically from the modern, casual Tagalog references used in the validation set, resulting in lower embedding similarity. In contrast, Model 2, which was exposed to synthetic conversational pairs sourced from dictionary examples, maintained a higher SBERT score, suggesting that dictionary-based augmentation acts as a necessary "bridge" to modern semantics, preventing the model from becoming locked into an exclusively liturgical register.

6 Conclusion

This study presented a systematic investigation into bootstrapping Neural Machine Translation for Tao (Yami), an endangered ultra-low-resource language, by leveraging the NLLB-200 architecture. Our findings demonstrate that in settings where parallel data is fewer than 5,000 sentences, relying solely on data quantity is insufficient; instead, exploiting linguistic structures and rich lexical resources is key in translating low-resource languages. We establish that phylogenetic transfer learning—specifically pre-training on the closely related Ivatan language—is the most effective strategy for maximizing translation fidelity. This approach yielded the highest in-domain performance (BLEU 34.85), confirming that shared morphological and syntactic features between Batanic languages can be effectively leveraged to stabilize the decoder. However, our analysis also revealed a critical "liturgical bias" in models trained exclusively on religious texts. To address this, we demonstrate the effectiveness of dictionary-assisted synthetic data augmentation. While this approach incurred a minor trade-off in in-domain precision, it significantly enhances the model's generalization capabilities, improving the BLEU score on conversational out-of-domain data. This suggests a complementary framework for future work: utilizing transfer learning to establish the grammatical blueprint of the language, while employing synthetic augmentation to expand the semantic system required for modern communication. Ultimately, this work provides a reproducible blueprint for other under-represented ultra-low-resource Austronesian languages. The study highlights that even in the absence of large-scale corpora, functional translation systems can be constructed by intelligently combining digitized linguistic heritage—such as grammars and dictionaries—with pre-trained SOTA multilingual transfer learning. Such efforts are essential for fostering an inclusive technological landscape the meets the needs of marginalized languagess.

7 Recommendations

Based on the on the analysis of model performance and the limitations observed during evaluation and configuration for the training, the following recommendations are proposed for future system improvements and further research. While basic text normalization was used, such as lowercasing and character filtering, was effectively used in this study. Implementing a more advanced normalization, like true-casing, punctuation utilization, and Unicode aware normalization, may be able to improve the model's performance. Utilizing better hardware will also enable future researchers to configure the training process to be more optimized by using larger batch sizes, training for more epochs, or experimenting with the latest and larger token models. In addition, future work can look into using a more sophisticated tokenization approaches, like sub-word modeling, to improve the handling of rare words and terms, and to also handle the rich morphological structure of the Tao language. Lastly, we recommended that future ultra-low-resource research involves collaborating with native speakers and language experts to verify the quality of translation of the model and to help promote research interest on the indigenous languages of the Philippines.

Limitations

Although Bible-based corpus is a widely available source in various languages in the Philippines, the different Bible translations, due to translators having varying source-versions, interpretations, faithfulness to the source text, and target use cases for their translations, may differ significantly in wording and structure from one language to another. Additionally, Bible verses are not as effective representations of the typical sentence structure or vocabulary of a language.

Moreover, our paper only focuses on one ultra-low-resource Austronesian language in the Yami language of Orchid Island. The machine translation findings presented in this paper are not representative of all Austronesian low-resource languages, nor are these results representative of low-resource languages globally. Likewise, these results and findings are not representative of the effectiveness of the various machine translation improvement techniques presented in our paper, given that we only focused on one source-to-target language pair, namely Yami to Tagalog. Future researchers will benefit from incorporating more low-resource languages in their works to better understand the effectiveness of pre-training using related languages, augmentation using synthetic data, and other techniques not implemented in this paper, to improve low-resource machine translation.

The Yami-Tagalog parallel corpus presented in this paper can not be used for validation of machine translation quality because it was created synthetically from a Yami-Chinese parallel corpus. For this task, human translations are still preferred because of the various biases and inaccuracies of current machine translation models. However, this parallel corpus can be used to augment real Yami-Tagalog parallel data, such as the Bible parallel corpus, for fine-tuning or training machine translation models, given the scarcity of quality parallel corpora incorporating the Yami language.

The neural machine translation model used in this paper, namely Meta's No Language Left Behind model (600 million parameters) was initially selected for VRAM, memory and speed considerations. In lieu, models with more parameters may likely produce higher quality translations than the specific model used in this paper. Additionally, we were limited to a limit of 30 hours of GPU accelerator usage on Kaggle, which further underscored our choice of model. Future researchers will benefit from using improved hardware, latest models, or more efficient fine-tuning techniques for their neural machine translation models. Furthermore, we also used the default tokenizer of NLBB-200 in consideration towards the limited time frame of our study. Lastly, our evaluation of machine translation quality did not incorporate human evaluation. Although the metrics we used to measure translation quality allow for the models to be compared to other machine translation models, these metrics do not capture translations of figure speech and ambiguous sentences. Additionally, these metrics also fail to detect biases in translations (e.g., unexpected gender bias when translating a sentence without gendered pronouns).

Acknowledgments

We thank Nathaniel Oco for introducing and educating us on NLP and NMT. We further acknowledge the use of Gemini to assist with the LaTeX formatting of the technical manuscript. All scientific claims, experimental results, and manuscript text were done and verified by the authors.

Author Contributions

Adrian Denzel Macayan designed the transfer learning methodology, conducted the formal analysis and review of related literature, and contributed to both the original draft and the final review and editing.
Luis Andrew Madridijo led the conceptualization and methodology design, developed the software for model training, and contributed to the writing of the original draft.
Ellexandrei Esponilla contributed to the the data curation and pre-processing of the Bible corpus, performed data validation, and contributed to the writing of the original draft.
Zachary Mitchell Francisco was responsible for the curation and development of synthetic data and contributed to the writing of the original draft.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Mahmoud Ahmad, Auwal Khalid, Lukman Aliyu, Babangida Sani, and Mariya Abdullahi. 2024. Arewa NLP's participation at WMT24. In *Proceedings of the Ninth Conference on Machine Translation*, pages 829–832, Miami, Florida, USA. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, San Diego, CA, USA.

Filipe Cardoso, Rui Silva, and Luis Gomes. 2022. Improving neural MT of indigenous languages with multilingual transfer learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 299–307, Dublin, Ireland. Association for Computational Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Víctor M. García, Jeroni Suárez, Marta R. Costa-jussà, and José A. R. Fonollosa. 2019. Context-aware monolingual repair for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 389–395, Florence, Italy. Association for Computational Linguistics.

Jiatao Gu, Yong Wang, Yun Chen, Kyunghyun Cho, and Victor O.K. Li. 2018. Meta-learning for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3622–3631, Brussels, Belgium. Association for Computational Linguistics.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Hugging Face. 2025. Hugging face model hub: Models (tagalog). Accessed: 2025-12-05.

Indigenous Languages Research and Development Foundation. 2020. e (indigenous language e-paradise). Accessed: 2025-12-05.

Kaiying Kevin Lin, Hsiyu Chen, and Haopeng Zhang. 2025. FormosanBench: Benchmarking low-resource Austronesian languages in the era of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16527–16539, Online. Association for Computational Linguistics.

Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James V. Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onno P. Kampman, Joel Ruben Antony Moniz, Muhammad Ravi Shulthan Habibi, Frederikus Hudi, Railey Montalan, Ryan Ignatius, and 1 others. 2024. SEACrowd: A multilingual multimodal data hub and benchmark suite for Southeast Asian languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5155–5203, Miami, Florida, USA. Association for Computational Linguistics.

Toan Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, Taipei, Taiwan. Asian Federation of Natural Language Processing.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Stephan Oepen, Nikolay Arefyev, Ona de Gibert, Andrey Kutuzov, Sampo Pyysalo, Jörg Tiedemann, and 1 others. 2025. HPLT 3.0: Very large-scale multilingual resources for LLM and MT. *Preprint*, arXiv:2511.01066. arXiv preprint arXiv:2511.01066.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Surangika Ranathunga, Shravan Nayak, Shih-Ting Cindy Huang, Yanke Mao, Tong Su, Yun-Hsiang Ray Chan, Songchen Yuan, Anthony Rinaldi, and Annie En-Shiun Lee. 2024. Exploiting domain-specific parallel data on multilingual language models for low-resource language translation. *Preprint*, arXiv:2412.19522. arXiv preprint arXiv:2412.19522.

D. Victoria Rau and Maa-Neu Dong. 2006. *Yami Texts with Reference Grammar and Dictionary*. Institute of Linguistics, Academia Sinica, Taipei.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Ana Rodríguez and 1 others. 2023. Train global, tailor local: Minimalist multilingual translation into endangered languages. In *Proceedings of the 6th Workshop on Technologies for MT of Low Resource Languages (LoResMT 2023)*. Association for Computational Linguistics.

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, Berlin, Germany. Association for Computational Linguistics.

Alexander D Smith. 2017. The western malayo-polynesian problem. *Oceanic Linguistics*, 56(2):435–490.

Gido M van de Ven, Nicholas Soures, and Dhireesha Kudithipudi. 2024. Continual learning and catastrophic forgetting. *arXiv preprint arXiv:2403.05175*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 5998–6008.

Nuo Xu, Yinqiao Li, Chen Xu, Yanyang Li, Bei Li, Tong Xiao, and Jingbo Zhu. 2019. Analysis of back-translation methods for low-resource neural machine translation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 466–475. Springer.

Meng-Chien Yang, Si-Wei Huang, and D. Victoria Rau. 2011. Two ontological approaches to building an integrated semantic network for yami *ka-* verbs. In *2011 International Conference on Asian Language Processing (IALP)*, pages 1–4. IEEE.

Meng-Chien Yang, D. Victoria Rau, and Ann Hui-Huan Chang. 2010. A proposed model for constructing a yami wordnet. In *2010 International Conference on Asian Language Processing (IALP)*, pages 1–4. IEEE.

Meng-Chien Yang, D. Victoria Rau, and Yi-Hsin Wu. 2012. Analyzing and classifying the yami emotion phrases using ontological structure and computation. In *2012 International Conference on Asian Language Processing (IALP)*, pages 45–48. IEEE.

Ling Zhang and 1 others. 2023. Bilex Rx: Lexical data augmentation for massively multilingual Machine Translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15745–15763, Toronto, Canada. Association for Computational Linguistics.

Barret Zoph 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, Austin, Texas. Association for Computational Linguistics.