# Investigating Neural Machine Translation Strategies for Tagalog

**Caldera, Henry | Hollingshead, Victoria | Jakkinpali, Neha**
DATASCI 266 | Natural Language Processing
School of Information | University of California, Berkeley
{ecalderajr, v.hollingshead, neha.jakk}@berkeley.edu

## Abstract

In natural language processing (NLP) research, high-resource languages benefit from extensive digitized data, enabling more advanced and accurate computational models. Conversely, low-resource languages suffer from a scarcity of digitized material, posing significant challenges in developing robust NLP systems. In this study, three state-of-the-art pre-trained LLM models were investigated in combination with several augmentation techniques to assess model translation performance for Tagalog, a low resource language among over 150 dialects spoken in the Philippines. With an iterative approach, BLEU, BLEURT, and native speaker evaluation scores were used to measure model performance of GPT-3.5 turbo, mBART50 and M2M100, with follow up fine-tuning, hyperparameter tuning, and back translation experiments. For GPT-3.5 turbo, prompt engineering with baseline performed the best with a BLEU score of 21.84 and and BLEURT score of 54.81. In the mBART50 model, the 100 real to 25 synthetic performed most optimally with a BLEU score of 25.01 and BLEURT of 38.81. For M2M100, the back-translation setup of 100 real to 75 synthetic was identified as the optimal model exhibiting a BLEU score of 28.61 and a BLEURT score of 46.94.

## 1 Introduction

A pervasive issue in natural language processing is addressing the performance gap between high and low-resource languages, due largely to the asymmetric availability of digitized corpora. In this study, the first research objective was to understand the baseline translation performance of several state-of-the-art pre-trained LLM models relative to MNMT literature results, to include GPT-3.5 turbo (Brown et al., 2020), mBART50 (Tan et al., 2019) and M2M100 (Fan et al., 2020). Secondly, this study investigates several augmentation techniques to improve model translation performance, including fine-tuning, hyperparameter tuning, and back translation. Using an iterative approach, BLEU, BLEURT, and native speaker evaluation scores were used to measure model performance and make incremental experimental changes for each model iteration.

## 2 Background

There have been several advancements in the effort to improve multilingual neural machine translation (MNMT) for low resource languages. For such models, the similarity of the languages used for training is highly important, as transferring information between dissimilar languages may lead to negative transfer and degraded performance (Saleh et al., 2019). Approaches to mitigate this error include language clustering, where families of languages are identified using a language embedding attained from a universal MNMT and clustered in the embedding space (Tan et al., 2019) or determined using a combination of language embedding vectors and syntax features (Saleh et al., 2019). The Philippines language group, distinct but related in their Austronesian linguistic phylogeny, has been studied using MNMT with zero-shot and pivot-based techniques with BLEU scores that range from 28.15 to 51.51 for various English to target language pairs (Baliber et al., 2020).

Goyle et al. (2019) explored the combination of the mBART-CC25 encoder-decoder model with three augmentation strategies to evaluate the translation of four low resource languages to English: Sinhala, Nepali, Khmer, and Pashto. The three augmentation strategies included back translation, transfer learning, and changing the loss function to focal loss. Figure 1 outlines the back translation workflow, where a Translator was trained on the original parallel dataset to translate from target to source language and is applied to a monolingual dataset in the target language to create a synthetic parallel corpus that is used to bolster the training dataset for the ultimate source-to-target translator.

For this experiment, mBART-50 (Tang et al., 2020) is used instead of mBART-CC25 to leverage the bi-direction fine-tuning done on 25 additional languages.

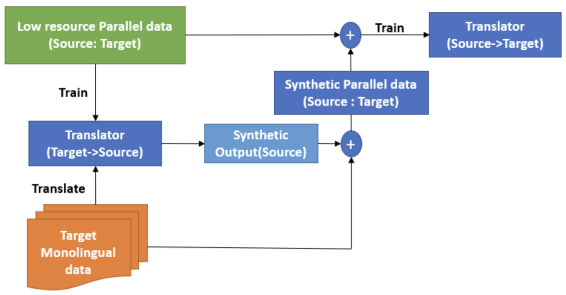


Figure 1: Back-translation Flow Diagram (Goyle et al., 2019)

This study also includes the Many-to-Many-100 model, or M2M100, which is particularly notable for substituting English-centric training data for a large dataset covering thousands of non-English translation pairs (Fan et al., 2020). M2M100 is designed to handle a myriad of tasks, showcasing its versatility in processing and generating human-like text across various domains. This makes it a valuable asset for global applications and research endeavors. Most recently, OpenAI's GPT-3.5 turbo has emerged as a cutting-edge language model, building upon its predecessor of GPT-3. With an expansive architecture and advanced training methodologies, GPT-3.5 turbo demonstrates extraordinary capabilities in natural language understanding, generation, and manipulation. This model has been optimized for performance, offering faster response times and enhanced efficiency in handling a wide array of linguistic tasks. The inclusion of M2M100 and GPT-3.5 turbo in this study signifies a commitment to exploring and harnessing the potential of these advanced language models. Their incorporation provides a comprehensive examination of the latest developments in the field, allowing for a nuanced understanding of the strengths and capabilities of each model.

In addition to model advancements, there have been several developments in parallel and monolingual corpora for low-resource languages, including FLORES200 (NLLB Team, 2022) and CulturaX (Nguyen et al., 2023), respectively. FLORES200 stands out as a comprehensive parallel corpus that has been catered to languages with limited linguistic data. It encompasses 200 languages, providing a diverse and extensive dataset for training and evaluating language models. CulturaX, another influential advancement in monolingual corpora goes beyond mere linguistic alignments by incorporating contextual and cultural information to enhance the depth and richness of language models. For example, the corpus includes sentences on topics that range from law to pop culture to politics. CulturaX is particularly valuable in scenarios where understanding the cultural context is crucial for accurate and meaningful language processing applications. By using and analyzing FLORES200 and CulturaX, we hope to derive insights on specific low-resource languages in hopes for more inclusive and effective language technologies.

## 3 Methodology

### 3.1 Data Pre-processing

For data pre-processing, CulturaX was parsed using regular expression to filter out non-ASCII glyphs and URLs. Examples containing Latin ASCII characters (hex codes: x00-x7F) with greater than 100 characters were kept for training purposes. Interval-based sampling was used to ensure samples from the diverse societal and cultural aspects of the CulturaX dataset were represented in training. From the FLORES200 dataset used in training, there were a total of 2009 parallel sentences, split 60/40/40 for training, validation, and testing sets, respectively. The specific breakdown is as follows: Training set: 1205 sentences, Validation set: 402 sentences, Test set: 402 sentences.

### 3.2 Tokenization

Tokenization was applied with padding to a multiple of 32 and truncation to a maximum length of 128. This approach was based on the understanding that truncating doesn't significantly decrease BLEU scores and leads to more efficient training (Popel and Bojar, 2018). The decision to use multiples of 32 was made for faster training by leveraging the architecture of NVidia's Ampere GPUs and aligning matrix dimensions as a multiple of Tensor core count (NVidia, 2023). The dataloader was configured with a batch size of 48, and it's noted that a bigger batch size is considered better for training efficiency and highly BLEU scores (Popel and Bojar, 2018).

### 3.3 Optimization and Learning Rate Scheduler

In terms of optimization and learning rate scheduling, the chosen optimizer was RAdam, with

a learning rate (LR) set to $2 \times 10^{-4}$. RAdam was selected to address issues with Adam/AdamW, where the variance of the adaptive learning rate in the early stages of training is overly large, leading to non-optimal convergence (Liu et al., 2021). The learning rate scheduler employed was Constant with Warmup, and the hyperparameters included warmup steps set to 85% of the first epoch (Liu et al., 2021).

During the training process, the model utilized the Early Stopping Callback with a configuration of 10 epochs, incorporating a tolerance of $1 \times 10^{-3}$, a patience of 5, and evaluating the performance based on the evaluation loss metric. To enhance the model's inference capabilities, a custom configuration was implemented, employing Diverse Beam Search. The hyperparameters for this configuration included 5 beam groups and 5 beams per group, contributing to a diversified and effective inference process (Vijayakumar et al., 2018).

### 3.4 Evaluation

The BLEU metric was employed with specific hyperparameters, setting the n-gram value to 4 and utilizing smoothing function #7. Notably, the 4-gram is the default setting for BLEU scoring. A smoothing function was introduced as a computational adjustment aimed to rectify a nuance in BLEU scoring where, if there is no overlap of n-grams for any order, the metric returns a value of 0. This is due to the precision for the non-overlapping order of n-grams being 0, and the geometric mean in the final BLEU score computation multiplying this 0 with the precision of other n-grams. This behavior often results in lower scores than expected (Chen and Cherry, 2014). On the other hand, BLEURT was employed with model checkpoint-20 (Pu et al., 2021). This metric is chosen for its reliance on training with native speaker evaluations, intending to reflect adequacy, which measures how well the original meaning of the text is retained in translation. Notably, in practical applications, BLEURT tends to exhibit a high correlation with fluency (Sellam et al., 2020).

### 3.5 Experiments

In the remainder of the study, the first Translator in Figure 1 will be referred to as the reverse-trained model. The second Translator in Figure 1 is each experiment is denoted in Table 2. The baseline model was a stock model, while the reverse-trained model underwent training from target to source to generate back-translations for subsequent experiments. The purpose of reverse training was to elucidate the success of back-translation experiments, attributing their efficacy to the reverse-trained model's strong performance in target-to-source translation. The fine-tuned model was a stock model forward-trained from source to target. In the context of back-translation experiments, a stock model was forward-trained on a dataset with a varying ratios of real-to-synthetic data. As part of a bonus experiment, the reverse-trained model (rather than the stock model) was used to forward-train on a synthetic dataset. Despite a BLEU score that exhibited a slower increase similar to the stock model performance, native speaker evaluations revealed outstanding results. The experiment achieved the highest score for fluency and the second-highest score for adequacy. These findings provide valuable insights into the effectiveness of different training approaches and their impact on translation performance, as highlighted in the discussion and results.

### 3.6 GPT-3.5 Turbo, fine-tuning and Hyperparameter Tuning

In the first iteration of experiments for GPT-3.5 turbo, the baseline model was fine-tuned and its performance assessed against baseline metrics. To streamline the fine-tuning process within the time constraints of this study and the operational limits set by the OpenAI API, the fine-tuning process was constrained to the FLORES200 dataset exclusively. The RateLimitError, which flags API calls when the number of tokens and requests exceed a threshold for a given time window, was prevalent for exceedingly large training datasets. Therefore, the FLORES200 dataset was partitioned with a roughly 60/20/20 split, resulting in 1205 lines for training, and 402 lines for each of the validation and test sets. These partitions were used to fine-tune and test the model. In addition, number of epochs, batch size, and learning rate multiplier were varied to assess if model performance improved.

### 3.7 Prompt Engineering

In the second iteration of experiments for GPT-3.5 turbo, a systematic approach to prompt engineering was taken to further refine the baseline model and improve results. This involved the construction of eight distinct prompts, each augmented with a variety of prompting strategies, including role assignment, contextualization, one-shot, and

few-shot learning. Concurrently, stringent attention was paid to ensuring that the contextual input for each prompt adhered to the maximum token limit of 4097, as delineated by the architecture of the model. The OpenAI Token Calculator was instrumental in estimating the length of context inputs prior to the initiation of experimental trials. Subsequently, the default parameters in OpenAI's Playground tool allowed for quick experimentation. These parameters included: temperature (also known as softmax temperature or softmax scaling), which governs the model's confidence level in its responses; maximum length, which truncates the response to a predetermined number of characters; top p, which serves as a threshold for confidence in evaluating resultant tokens; frequency penalty, which reduces the recurrence of tokens in the model's outputs to mitigate verbatim responses; and presence penalty, which diminishes the likelihood of token selection based on their prior occurrence in the output. The specified settings for these parameters were 0, 254, 1, 0, and 0, respectively. After the prompts were run using the parameters outlined above, the native speaker evaluation was applied to assess the relative fluency, adequacy, and formality of the machine translations. From this evaluation, the highest performing prompts were selected for additional experimentation.

### 3.8 Native Speaker Evaluation

Fluency and adequacy are widely recognized metrics in the human evaluation of machine translation outputs, where fluency pertains to the grammatical correctness and overall comprehensibility of a sentence and adequacy refers to the extent to which the original meaning of the sentence is preserved in the translation (Snover et al., 2009). While studies use BLEU and BLEURT to evaluate performance, it is important to acknowledge the limitations in both metrics when capturing fluency and adequacy. For example, because the BLEU metric relies on n-gram comparisons, it is not highly reliable in measuring the adequacy of machine generated text. Likewise, while the BLEURT metric captures adequacy (Sellam and Parikh, 2020), the reliability of BLEURT-20, used throughout this study, is dependent on the language being assessed, having only undergone testing on a restricted set of thirteen languages, conspicuously excluding Tagalog (Pu et al., 2021). While BLEURT has the potential to assess Tagalog on some level due to the presence of Tagalog in the multilingual C4 dataset, used to train RemBERT, one of the foundational models of BLEURT, it remains empirically unverified to what extent BLEURT can be used to reliably evaluation translations in this low resource language. To address the gaps presented in both established metrics, this study introduced a modest native speaker evaluation. To assess the actual semantic and syntactic coherence of the machine translations, a 7-point Likert Scale, derived from previous studies (kir) was structured according to three criteria: fluency, adequacy, and formality, outlined in Appendix A.

In addition to these conventional evaluation criteria, this study introduces 'formality' as a third evaluative criterion. While this metric was not used to assess translation quality, it added an additional dimension of analysis by quantifying the level of colloquialism present in the translation output. It is measured on a scale from 1 to 3, with 3 denoting the highest degree of formality. Each evaluation in this study was completed by one native speaker. As an effect of this limitation, the native speaker evaluations are averaged across no more than five English-Tagalog translation pairs per model. For consistency, every model examined in this study underwent evaluation using an identical set of five English-Tagalog translation pairs. If translations were not in Tagalog, or contained erroneous vocabulary or vocabulary from other dialects, the translation resulted in a score of 0.

## 4 Results and discussion

### 4.1 GPT-3.5 Turbo

In the first iteration of experiments, the baseline model was fine-tuned using the FLORES200 dataset. Contrary to the initial hypothesis, fine-tuning the baseline model gravely decreased performance for BLEU and BLEURT, which corroborated with marginal decreases for average fluency and adequacy in the native speaker evaluation. Despite FLORES200 being a recent, state-of-the-art parallel corpus, it is possible that the presence of degraded Tagalog translations negatively impacted the performance of the baseline model. Smaller, but highly refined datasets could be used to fine-tune the baseline model in future experiments for comparison.

In the second iteration, a series of hyperparameter tuning experiments were applied to the fine-tuned model. Both the baseline and fine-tuned model used the GPT-3.5 turbo default values for the number of epochs, batch size, and learning

| Model | BLEU | BLEURT | Fluency | Adequacy | Formality |
|---|---|---|---|---|---|
| Baseline | 21.21 | 54.43 | 5.6 | 4.6 | 3 |
| Fine-tuning | 16.44 | 43.21 | 5 | 4.2 | 1.6 |
| Fine-tuning + HP Tuning | 17.13 | 43.87 | 6.6 | 5.3 | 2.8 |
| Prompt #1 | 20.80 | 53.79 | 4.4 | 4.4 | 1 |
| Prompt #2 | 21.84 | 54.81 | 5.6 | 5.6 | 1.4 |
| Prompt #3 | 21.47 | 54.18 | 5.6 | 5.4 | 1.6 |

Table 1: GPT-3.5 turbo performance with prompt engineering and fine-tuning

rate multiplier, with the fine-tuned model resulting in a training and validation loss of 0.6274 and 0.7699, respectively. Following hyperparameter tuning, an increase in the number of epochs from 3 to 5, the training and validation loss decreased to 0.3832 and 0.4952, respectively. The hyperparameter tuned model also surpassed both the baseline and fine-tuned model for average fluency and adequacy. However, despite the lower validation loss and improved average fluency and adequacy in the hyperparameter tuned model, due to the significantly lower BLEU and BLEURT scores relative to the baseline, prompt engineering was applied to the baseline model.

In the third iteration, Prompts 3, 4, and 8 from Table 6 in Appendix A were selected for additional prompt engineering experimentation based on their native speaker evaluation and renamed to Prompts 1, 2, 3 for the remainder of the study, as outlined in Appendix A.

Relative to the baseline model, prompt engineering resulted in higher BLEU and average adequacy scores for both Prompts 2 and 3, with Prompt 3 also resulting in a higher BLEURT score by 0.38. Prompt 1 underperformed in all four performance metrics likely due to the relatively low initial adequacy score in comparison to the other two prompts. These results suggest that context and few shot learning can add marginal improvements in capturing the meaning of the text, while offering few gains in fluency. In addition, role and instruction alone are not sufficient in improving performance over baseline.

In comparison to the results presented in (Baliber et al., 2020), where the English to Tagalog translation resulted in a BLEU score of 36.61, the GPT-3.5 turbo baseline model, fine-tuned model, and subsequent prompt-engineered models, all underperform by over 10 BLEUR units. Future experiments could be constructed using GPT-3.5 turbo such that additional low resource Philippines Languages are trained on the same model to potentially take advantage of morphological similarities between dialects. Altering the translation direction during training could also be explored.

### 4.2 mBART50

In the first iteration of the mBART50 experiment, the baseline model was reverse-trained with the FLORES200 training and validation datasets. Following the reverse-training, the BLEU score increased by 17 units and the BLEURT score increased by 35 units from baseline, showing promise regarding the quality of the back-translated, synthetic dataset to be used in subsequent experiments. Fine-tuning the baseline model on 100% real and 0% synthetic, there was a 12 and 7 unit increase in the BLEU and BLEURT scores, respectively. Using the baseline model, several iterations of real-to-synthetic data were tested including 100:0, 100:25, 100:50, 100:75, 100:100, and 100:200. From the back-translation experiments, we found that the 100% real to 25% synthetic performed most optimally with a BLEU score of 25.01 and BLEURT of 38.81, which is 20 units above the BLEU and BLEURT baseline scores. Increased ratios of real-to-synthetic training data gradually diminished improvements over baseline performance. This can be seen by the inspection of back-translated sentences from the reverse-trained model. Given a sentence in Tagalog, the model gives the following prediction: `Bote a rock climbing the river and shattered` , while the ground truth is: `A bottle fell onto the floor and shattered`. The back-translation is relatively fluent, but fails to retain the meaning of the original text. Importantly, the back-translation has retained one Tagalog word `bote`, which if occurring frequently in other back-translations, would affect the evaluations scores of models fine-tuned on higher ratios of augmented data.

Analysis based on native speaker evaluations reveals that, despite the back-translation baseline

| Model | Checkpoint | BLEU | BLEURT | Fluency | Adequacy | Formality |
|---|---|---|---|---|---|---|
| mBART50 | Baseline | 4.20 | 18.39 | 0 | 0 | 0 |
| mBART50 | Fine-tuned (100:0) | 16.38 | 25.18 | 2.6 | 1.27 | 0.98 |
| mBART50 | Back-translation (100:25) | 25.01 | 38.81 | 3.8 | 1.63 | 1.35 |
| mBART50 | Back-translation (100:50) | 23.58 | 37.44 | 3.6 | 1.6 | 1.31 |
| mBART50 | Back-translation (100:75) | 22.61 | 35.50 | 4.8 | 1.97 | 1.97 |
| mBART50 | Back-translation (100:100) | 23.18 | 37.95 | 4.8 | 2.63 | 1.92 |
| mBART50 | Back-translation (100:200) | 18.96 | 31.72 | 2.4 | 1.07 | 0.92 |
| M2M100 | Baseline | 18.28 | 36.89 | 5.80 | 3.80 | 2.37 |
| M2M100 | Fine-tuned (100:0) | 25.67 | 40.75 | 3.40 | 2.23 | 1.52 |
| M2M100 | Back-translation (100:25) | 27.37 | 43.41 | 6.60 | 4.10 | 2.39 |
| M2M100 | Back-translation (100:50) | 27.62 | 46.57 | 5.20 | 3.20 | 2.06 |
| M2M100 | Back-translation (100:75) | 28.61 | 46.94 | 5.20 | 3.37 | 1.94 |
| M2M100 | Back-translation (100:100) | 27.16 | 45.48 | 5.00 | 2.50 | 1.93 |
| M2M100 | Back-translation (100:200) | 26.33 | 45.53 | 5.20 | 3.03 | 2.03 |

Table 2: mBART50 and M2M100 Results on FLORES200 dataset

of 100% real to 25% synthetic being identified as superior by BLEU and BLEURT scores, the back-translation baseline of 100% real to 100% synthetic exhibited the highest fluency and adequacy scores. This underscores the significance of incorporating native speaker perspectives in the comprehensive evaluation of model performance, acknowledging nuances that BLEU and BLEURT metrics may not fully capture, especially in the context of languages with limited resources.

### 4.3 M2M100

In the initial phases of the M2M100 iterations, the baseline model underwent reverse training using the FLORES200 training and validation datasets. This reverse-training procedure yielded a notable enhancement in a BLEU score of 26.96 and a BLEURT score of 58.27. These scores reflected an appreciable improvement of approximately 8 BLEU points and 22 BLEURT points over the baseline metrics. Subsequent fine-tuning of the baseline model with a composition of 100% real data and 0% synthetic data resulted in a 7 and 4 unit increase of BLEU BLEURT scores, respectively. Then, a series of back-translation iterations were conducted, varying the ratios of real and synthetic data, including 100:0, 100:25, 100:50, 100:75, 100:100, and 100:200. Notably, the back-translation iteration with a composition of 100% real to 75% synthetic emerged as the most successful, exhibiting a BLEU score of 28.61 and a BLEURT score of 46.94. This represented a significant advancement from the baseline metrics. Overall, the back-translation setup of 100% real to 75% synthetic was identified as the optimal model for M2M100 across all iterations, showcasing its superior performance in comparison to other configurations within the experimentation framework. BLEU and BLEURT scores are not a definitive assessment of the overall model performance, and it is best to also consider the native speaker evaluation when performing a holistic analysis on the models. Although the back-translation setup of 100% real to 75% synthetic was identified as the optimal model, from the native speaker evaluation, we can see that the back-translation of 100% real to 25% synthetic was the best with a 6.6 fluency, 4.1 adequacy, and 2.39 formality. Subsequently, the next most effective model aligns with the baseline, whereas the remaining fine-tuned models exhibit lower performance compared to the baseline.

## 5 Conclusion

Of all the models investigated, M2M100 consistently achieved the highest BLEU scores, reaching mid-20 BLEU scores across all experiments while GPT-3.5 turbo generally produced the highest BLEURT and native speaker evaluation scores. While all experiments generally achieved incremental improvements with each iteration, models underperformed by approximately 10 BLEU units in comparison to literature results. Future research would benefit from recruiting multiple native speakers to ensure the statistical integrity of the native speaker evaluations. For GPT-3.5 turbo, future experiments could be constructed such that additional low resource Philippines Languages are trained on the same model to potentially take advantage of morphological similarities between dialects. Altering the translation direction during training could also be explored. In the GPT-3.5 model, prompt engineering with baseline performed the best with

a BLEU score of 21.84 and and BLEURT score of 54.81. In the mBART50 model, the 100% real to 25% synthetic performed most optimally with a BLEU score of 25.01 and BLEURT of 38.81. In the M2M100 model, the back-translation setup of 100% real to 75% synthetic was identified as the optimal model exhibiting a BLEU score of 28.61 and a BLEURT score of 46.94. In the future studies, for the decoder-encoder models, an intermediate model to validate and eliminate subpar back-translations could be explored to enhance the overall performance of the final translator model (Galiano-Jimenez et al., 2023).

## References


Renz Iver D. Baliber, Charibeth K. Cheng, Kristine Mae M. Adlaon, and Virgion H. Mamonong. 2020. Bridging philippine languages with multilingual neural machine translation.

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

Boxing Chen and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 362–367, Baltimore, Maryland, USA. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation.

Aaron Galiano-Jimenez, Felipe Sanchez-Martinez, Victor M. Sanchez-Cartagena, and Juan Antonio Perez-Ortiz. 2023. Exploiting large pre-trained models for low-resource neural machine translation. https://www.dlsi.ua.es/~fsanchez/pub/pdf/galiano23a.pdf.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2021. On the variance of the adaptive learning rate and beyond.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages.

James Cross Onur Çelebi Maha Elbayad Kenneth Heafield Kevin Heffernan Elahe Kalbassi Janice Lam Daniel Licht Jean Maillard Anna Sun Skyler Wang Guillaume Wenzek Al Youngblood Bapi Akula Loic Barrault Gabriel Mejia Gonzalez Prangthip Hansanti John Hoffman Semarley Jarrett Kaushik Ram Sadagopan Dirk Rowe Shannon Spruit Chau Tran Pierre Andrews Necip Fazil Ayan Shruti Bhosale Sergey Edunov Angela Fan Cynthia Gao Vedanuj Goswami Francisco Guzmán Philipp Koehn Alexandre Mourachko Christophe Ropers Safiyyah Saleem Holger Schwenk Jeff Wang NLLB Team, Marta R. Costa-jussà. 2022. No language left behind: Scaling human-centered machine translation.

Martin Popel and Ondřej Bojar. 2018. Training tips for the transformer model. *The Prague Bulletin of Mathematical Linguistics*, 110(1):43–70.

Amy Pu, Hyung Won Chung, Ankur P. Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for mt.

Fahimeh Saleh, Wray Buntine, Gholamreza Haffari, and Lan Du. 2019. Multilingual neural machine translation: Can linguistic hierarchies help?

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. Bleurt: Learning robust metrics for text generation.

Thibault Sellam and Ankur P. Parikh. 2020. Evaluating natural language generation with bleurt. Google Research Blog.

Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 259–268, Athens, Greece. Association for Computational Linguistics.

Xu Tan, Yi Ren, Di He, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2019. Multilingual neural machine translation with knowledge distillation.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R. Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. Diverse beam search: Decoding diverse solutions from neural sequence models.

# A Appendix

| Score | Fluency | Adequacy |
|---|---|---|
| 7 | Flawless Tagalog | All |
| 6 | Fluent | Most |
| 5 | Slightly Fluent | Much |
| 4 | Non-Native Tagalog | Adequate |
| 3 | Slightly Disfluent | Some |
| 2 | Disfluent | Little |
| 1 | Incomprehensible | None |

Table 3: Likert Scale of fluency and adequacy evaluation.

| Score | Formality |
|---|---|
| 3 | Official government report |
| 2 | Local news broadcast |
| 1 | Very informal, slang |

Table 4: Evaluation of Formality

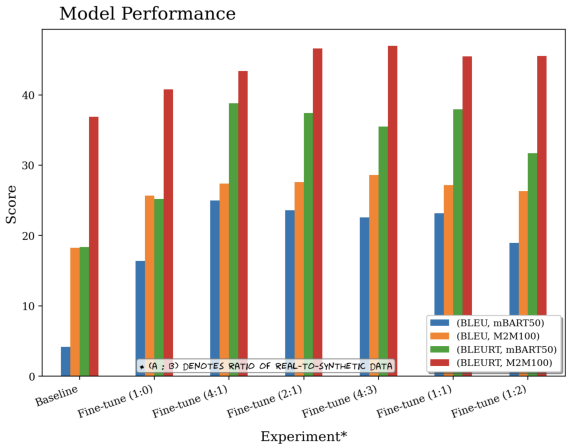


Figure 2: BLEU and BLEURT evaluation scores for mBART-50 and M2M100 fine-tuned on varying ratios of real-to-synthetic corpora

| No. | Technique | Native Speaker Evaluation | | |
|---|---|---|---|---|
| | | Fluency | Adequacy | Formality |
| 1 | Instruction | 7 | 6 | 2 |
| 2 | Role and instruction | 7 | 7 | 2 |
| 3 | Role and instruction | 7 | 6 | 3 |
| 4 | Context and instruction | 7 | 7 | 3 |
| 5 | Context, role, and instruction | 6 | 5 | 1 |
| 6 | Context, role, and instruction | 7 | 7 | 1 |
| 7 | Context, role, instruction, and one example. | 7 | 7 | 1 |
| 8 | Context, role, instruction, and two examples. | 7 | 7 | 1 |

Table 5: Native Speaker Evaluation of Different Techniques

| Model | BLEU | BLEURT | Fluency | Adequacy | Formality (Avg) |
|---|---|---|---|---|---|
| Baseline | 0.00 | 0.00 | 0 | 0 | 0 |
| Fine-tuning | -4.77 | -11.22 | -0.6 | -0.4 | -1.4 |
| Fine-tuning + HP Tuning | -4.09 | -10.56 | **1.00** | **0.67** | -0.16 |
| Prompt #1 | -0.42 | -0.64 | -1.2 | -0.2 | -2 |
| Prompt #2 | **0.63** | **0.38** | 0 | **1** | -1.6 |
| Prompt #3 | **0.26** | -0.25 | 0 | **0.8** | -1.4 |

Table 6: GPT-3.5 turbo performance relative to baseline.

| No. | Technique | Prompt |
|---|---|---|
| 1 | Role and instruction. | You are a fluent English-Tagalog speaker. Translate from English to Tagalog. |
| 2 | Context and instruction. | Tagalog is an Austronesian language spoken in Luzon and neighboring islands and forming the basis of the standardized national language of the Philippines (Filipino). Its vocabulary has been much influenced by Spanish and English, and to some extent by Chinese, Sanskrit, Tamil, and Malay. Translate from English to Tagalog. |
| 3 | Context, role, and instruction, and two examples. | Tagalog is an Austronesian language spoken in Luzon and neighboring islands and forming the basis of the standardized national language of the Philippines (Filipino). Its vocabulary has been much influenced by Spanish and English, and to some extent by Chinese, Sanskrit, Tamil, and Malay. You are a fluent English-Tagalog speaker. We would like to translate English sentences to Tagalog. Here are some examples of translations. Make sure to translate correctly.<br>English: The Cook Islands do not have any cities but are composed of 15 different islands.<br>Tagalog: Walang kahit among siyudad ang Cock Islands sublet kinabibilangan ng 15 iba-ibang pula.<br>English: The main ones are Rarotonga and Aitutaki.<br>Tagalog: Ang pinakamalaki sa mga ito ay ang Rarotonga at Aitutaki. |

Table 7: Selected prompts for pairing with baseline model.

| | Score | | Delta | | Native Speaker Evaluation | | |
|---|---|---|---|---|---|---|---|
| **Model** | **BLEU** | **BLEURT** | **BLEU** | **BLEURT** | **Fluency** | **Adequacy** | **Formality** |
| Baseline | 4.20 | 18.39 | - | - | 0.00 | 0.00 | 0.00 |
| Fine-tuned (100:0) | 16.38 | 25.18 | +12.18 | +6.80 | 2.60 | 1.27 | 0.98 |
| Backtranslation (100:25) | 25.01 | 38.81 | +20.82 | +20.42 | 3.80 | 1.63 | 1.35 |
| Backtranslation (100:50) | 23.58 | 37.44 | +19.38 | +19.06 | 3.60 | 1.60 | 1.31 |
| Backtranslation (100:75) | 22.61 | 35.50 | +18.41 | +17.12 | 4.80 | 1.97 | 1.97 |
| Backtranslation (100:100) | 23.18 | 37.95 | +18.99 | +19.56 | 4.80 | 2.63 | 1.92 |
| Backtranslation (100:200) | 18.96 | 31.72 | +14.76 | +13.34 | 2.40 | 1.07 | 0.92 |

Figure 3: mBART50 Results

| | Score | | Delta | | Native Speaker Evaluation | | |
|---|---|---|---|---|---|---|---|
| **Model** | **BLEU** | **BLEURT** | **BLEU** | **BLEURT** | **Fluency** | **Adequacy** | **Formality** |
| Baseline | 18.28 | 36.89 | - | - | 5.80 | 3.80 | 2.37 |
| Fine-tuned (100:0) | 25.67 | 40.75 | +7.39 | +3.86 | 3.40 | 2.23 | 1.52 |
| Backtranslation (100:25) | 27.37 | 43.41 | +9.09 | +6.52 | 6.60 | 4.10 | 2.39 |
| Backtranslation (100:50) | 27.62 | 46.57 | +9.34 | +9.67 | 5.20 | 3.20 | 2.06 |
| Backtranslation (100:75) | 28.61 | 46.94 | +10.33 | +10.05 | 5.20 | 3.37 | 1.94 |
| Backtranslation (100:100) | 27.16 | 45.48 | +8.88 | +8.59 | 5.00 | 2.50 | 1.93 |
| Backtranslation (100:200) | 26.33 | 45.53 | +8.04 | +8.64 | 5.20 | 3.03 | 2.03 |

Figure 4: M2M100 Results