Boosting Quality and Quantity of Back-Translations to Improve Performance for Low-Resource Language Translation

Jerry Qian
UC Berkeley MIDS
jerryq@berkeley.edu

Neta Tartakovsky
UC Berkeley MIDS
ntartakovsky@berkeley.edu

Abstract

Back-translation has been proven to be an effective method for improving performance of machine translation models for low-resource languages. In this study, we assess the impact of two variables on the final Swahili to English translation quality: the quality of the pre-trained model used to generate the back-translations for fine-tuning, and the amount of back-translated pairs used to fine-tune the baseline model. We created experimental back-translated datasets using 4 different pre-trained models - GNMT, M2M-100, GPT-3.5, and mBART-50 - and then used the datasets to fine-tune the mBART-50 baseline model. The final Swahili to English translations generated by the fine-tuned models were then evaluated by calculating BLEU scores using the FloRes-200 dataset. The findings support our original hypothesis - that applying higher quality pre-trained models to generate the back-translations, and using a larger amount of back-translated data for fine-tuning, will ultimately lead to higher translation quality in the final fine-tuned model.

1 Introduction

Neural machine translation models have achieved state-of-the-art performance in recent years for translation between high-resource languages (languages with vast amounts of available parallel data). However, the translation quality for low-resource languages has not been as impressive, due to an insufficient amount of available parallel language data (Sennrich and Zhang, 2019). This subpar performance has inspired research focused on identifying techniques for utilizing monolingual data to improve performance, and boosting fluency for low-resource language translations (Ranathunga et al. 2023).

Back-translation has been shown to improve performance of machine translation models on low-resource language translations (Przystupa and Abdul-Mageed, 2019). This technique involves using a pre-trained model to generate synthetic translations in the target language from real, human-generated monolingual text in the source language (Sennrich et al. 2016). The synthetic target translations are then fed back into the pre-trained model to translate the sentences back into the source language. These new synthetic sentence pairs are then used as additional data to train the machine translation model with, which creates more parallel data for low-resource language pairs, and improves the fluency of the resulting translations.

Our study extends the existing research around back-translation by investigating two further questions:

1. Is there a significant difference in the final model's quality of translations when different pre-trained models are used to generate the synthetic back-translated pairs?

2. Does varying the amount of back-translated pairs from each pre-trained model used to fine-tune the base model have a significant impact on the final translation quality?

2 Background

The work presented in Poncelas et al. (2019) investigated the effects of back-translated data originating from two different types of machine translation systems - statistical machine translation and neural machine translation. The researchers found that the difference in model architecture used to generate the synthetic back-translation pairs did have a significant impact on final translation quality. We extend this research in our study by investigating the difference in

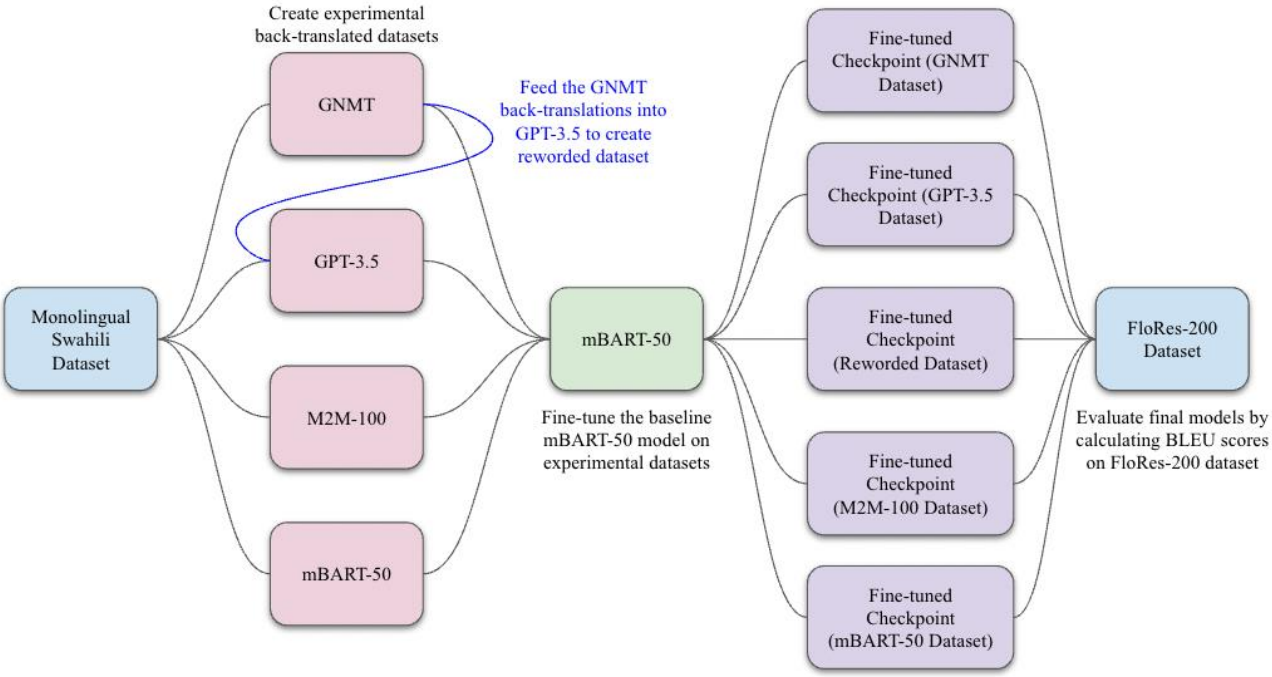

Figure 1: Summary of our full experimental flow: Original monolingual data is fed into 4 different pre-trained models to generate experimental back-translated datasets, which are then used to fine-tune the baseline mBART-50 model. The fine-tuned checkpoint models are then evaluated using the FloRes-200 dataset.

performance resulting from different pre-trained neural models.

Additionally, the impact of varying the amount of back-translated pairs used while fine-tuning the base machine translation model was also investigated by Poncelas et al. (2018). Their research established that while training the model on a large amount of synthetic back-translated pairs does lead to significant enhancements in the quality of machine translation for low-resource languages, performance actually starts to decrease when too many synthetic pairs are added. In our study, we explore this effect at the individual model level by varying the amount of data for each pre-trained model.

3 Methods

3.1 Experiment Design

In this study, we fine-tune a baseline mBART-50 model on multiple datasets containing synthetic back-translations to improve the quality of the final model's Swahili to English translations. We assess the impact of two variables on the final translation quality: the quality of the pre-trained model used to generate the back-translations, and the amount of back-translated pairs used to fine-tune the baseline model.

The study focuses specifically on Swahili to English translation. We are interested in boosting performance for a low-resource language, and while Swahili is spoken by approximately 100 million people, it is considered a low-resource language in the context of NLP due to limited availability of training data and resources. English, on the other hand, is the most widely spoken language in the world, with over 1.5 billion speakers. Improving Swahili to English translation can help facilitate communication and collaboration between Swahili speakers and the wider English-speaking world.

To assess the impact of back-translated texts on the performance of our fine-tuned models, we carried out two simultaneous experiments. In the first experiment, we utilized 4 pre-trained models - GNMT, M2M-100, GPT-3.5, and mBART-50 itself - to generate the synthetic back-translated texts used to fine-tune our baseline mBART-50 model. In the second experiment, we varied the amounts of back-translated texts used to fine-tune the mBART-50 model, ranging from 10,000 to 200,000 sentence pairs from each pre-trained model. We evaluated the impact of both variables on final Swahili to English translation quality by calculating the BLEU scores for the translations generated across all our fine-tuned models, using the FloRes-200 benchmark dataset (Costa-jussà et al. 2022).

3.2 Original Data

The CC100-Swahili dataset is one of the 100 corpora of monolingual data that was processed from CommonCrawl snapshots from the CC-Net repository by Conneau et al. (2020). This dataset provided us with a rich source of authentic Swahili text. We sampled 100,000 Swahili sentences from the CC100-Swahili dataset to create our primary monolingual dataset. These sentences were then used to generate synthetic translations in the target language (English) based on real, human-generated monolingual text in the source language (Swahili). The synthetic English translations were subsequently used to generate back-translations in Swahili.

3.3 Experimental Datasets

We created multiple back-translated datasets by feeding the original monolingual Swahili dataset into 4 different pre-trained models to generate synthetic Swahili to English translations. Specifically, we used GNMT, M2M-100, GPT-3.5, and mBART-50. Once we had the synthetic English translations, we fed them back into their respective models to generate back-translations into Swahili. This resulted in four different back-translated datasets, each generated by a different pre-trained model. The goal of this approach was to generate multiple versions of the original Swahili dataset, with the hope that the resulting datasets would contain different sentence structures, word choices, and other linguistic features.

3.3.1 GNMT

Google Neural Machine Translation (GNMT) follows the common sequence-to-sequence learning framework with attention, and has three components: an encoder network, a decoder network, and an attention network (Wu et al. 2016). It contains 8 LSTM encoder layers, and 8 decoder layers. GNMT was selected as it is widely regarded as the state-of-the-art translation model. It is often the go-to tool for many users when they require translations. Including GNMT in our experiment allowed us to use its performance as a benchmark, helping us assess the quality of the translations generated by the other models.

3.3.2 M2M-100

This Many-to-Many multilingual translation model (M2M-100) is based on the Transformer sequence-to-sequence architecture, and is composed of two modules, the encoder and the decoder, each with 24 layers (Fan et al. 2020). M2M-100 is pre-trained on a large number of language pairs, and has been shown to capture cross-lingual patterns and generate high-quality translations, making it a good candidate for generating synthetic translations.

3.3.3 GPT-3.5

GPT-3.5 is a fine-tuned checkpoint of the original GPT-3 (Generative Pre-Trained Transformer) model, an autoregressive language model with 175 billion parameters and 96 layers (Brown et al. 2020). GPT-3.5 is a state-of-the-art language model trained on a massive amount of diverse text, which can help it generate high-quality translations with a natural language style. Although it is primarily known for its performance on question-answering and reasoning tasks, we were interested in exploring its performance on translation tasks as well. Due to the cost associated with the API needed to utilize the GPT-3.5 model, we created a dataset with only 100k sentences, rather than the 200k sentence datasets created using the other pre-trained models.

3.3.4 mBART-50

The mBART-50 model is a fine-tuned checkpoint of the original mBART (Multilingual Bidirectional and Auto-Regressive Transformers) model, which is a sequence-to-sequence denoising auto-encoder pretrained on large-scale monolingual corpora in many languages, and is intended mainly for machine translation (Tang et al. 2020). Since mBART-50 is also our baseline model, we used this model to generate the translations to investigate the impact of fine-tuning a baseline model on its own outputs.

3.3.5 GPT-3.5 - Rewording

In addition to generating datasets with direct translations, we created an experimental dataset consisting of reworded translations generated by the GPT-3.5 model to investigate performance of a model fine-tuned on reworded synthetic data. To create this dataset, we selected the English translations and Swahili back-translations generated by GNMT, as its performance was expected to be the best among the models, and fed them into GPT-3.5, along with instructions to reword the sentences. We hoped this data

augmentation technique would enable us to create additional data in a low-resource scenario, and boost performance for low-resource language translations.

4 Training & Evaluation

We used the "mBART-large-50-many-to-many-mmt" fine-tuned checkpoint of mBART-50 as our baseline model. Throughout each round of training, we maintained a fixed set of hyperparameters to ensure a consistent evaluation environment (Table 1). This approach enabled us to isolate the effects of the back-translated dataset on model performance to obtain a more accurate understanding of the impact of the experimental datasets.

Hyperparameter	Value
Learning Rate	5e-4
Batch Size	64
Epochs	10
Max Sequence Length	32

Table 1: Hyperparameters used while training baseline mBART-50 model

We monitored the training and validation loss at each epoch to track the progress of our model and identify any trends or patterns that emerged as a result of the different experimental conditions.

We evaluated our final fine-tuned models using FLoRes-200, a benchmark dataset for machine translation between English and low-resource languages (Costa-jussà et al. 2022). It contains 200 language pairs, each with 2000 sentence pairs, and covers a wide range of languages from various language families. We used a subset of 2000 parallel sentence pairs in Swahili and English as our evaluation dataset.

5 Results & Discussion

5.1 Baseline Model

Prior to conducting the experiments, we first evaluated the baseline mBART-50 model on the FloRes-200 dataset to establish a point of comparison. The baseline model achieved a BLEU score of 0.404.

5.2 Experiment 1: Models used to Generate Back-Translations

When comparing the BLEU scores across the final fine-tuned models, it is clear that the quality of the pre-trained model chosen to generate the synthetic back-translations does have an impact on final translation quality.

Pre-Trained Model	Baseline BLEU Score
GNMT	34.630
M2M-100	13.529
GPT-3.5	25.287
mBART-50	0.404

Table 2: BLEU scores for each pre-trained baseline model, evaluated on FloRes-200 dataset

Prior to the experiment, we evaluated the baseline Swahili to English translation quality of all 4 pre-trained models by calculating the BLEU scores for the pre-trained models using the FloRes-200 dataset (Table 2). The final experimental BLEU scores for the models fine-tuned on back-translations generated by the 4 pre-trained models followed the pattern present in the original baseline BLEU scores for the models. These results can set the expectations for the performance of fine-tuning mBART on each pre-trained model's outputs.

The BLEU scores for the models fine-tuned on GNMT outputs consistently surpassed that of the other models at every dataset size, with the highest BLEU score of 14.035 for a dataset size of 200k sentences (Table 3). The BLEU scores for the models fine-tuned on GPT-3.5 outputs began to approach the performance of the GNMT scores at the 50k and 100k dataset sizes, indicating that GPT-3.5 models may catch up in performance with additional data. These findings are consistent with the baseline BLEU scores for the GNMT and GPT-3.5 models, which respectively received the highest and second-highest scores in both scenarios. This highlights the importance of using specialized models for specific tasks, as evidenced by GNMT outperforming GPT-3.5 in translation tasks.

The BLEU scores for the models fine-tuned on mBART-50 outputs are significantly lower than those of the other models. As the mBART-50

	10k	20k	50k	100k	200k
GNMT	0.381	1.709	6.769	11.457	14.035
M2M-100	2.156e-81	3.033-e05	4.435e-05	7.077	8.554
GPT-3.5	2.092e-81	7.412e-05	6.629	9.321	-
GPT-3.5 Reworded	1.979e-81	0.0357	0.797	0.999	-
mBART-50	0.173	8.730e-157	0.037	0.092	5.020e-05

Table 3: BLEU scores for fine-tuned models - split by pre-trained models & back-translated dataset sizes

baseline model is trained on its own outputs, it appears to be overfitting, leading to worse performance when exposed to more data. The BLEU scores support this observation, as they decrease with larger dataset sizes. This result also correlates with the baseline BLEU scores for the pre-trained models, as the mBART-50 model received the lowest baseline BLEU score (Table 2).

To further investigate the quality of the synthetic back-translations, we calculated the BLEU scores for the English and Swahili translations generated by each pre-trained model, using the GNMT-generated translations as the "ground truth" for the calculations. For the Swahili translations, our findings were consistent with the final performance of the fine-tuned models - with the GPT-3.5 translations achieving the highest BLEU score of 0.462, followed by the M2M-100 translations with a 0.273 BLEU score, and lastly trailed by the mBART-50 translations with a low BLEU score of 0.008. This is consistent with the rankings of the performance of the final BLEU scores for the fine-tuned models. However, the BLEU scores for the English translations did not show a clear pattern, with all the pre-trained models receiving relatively low BLEU scores ranging from 0.0002 to 0.003.

The rewording experiment demonstrated that while adding more synthetic data led to incremental improvements in the final model's performance, the BLEU scores achieved were significantly lower than those obtained using GNMT or GPT-3.5 translations. One possible explanation for this suboptimal performance is that the GNMT translations, although the best-performing in our study, were not perfect and introduced errors or inconsistencies when reworded. These imperfections in the reworded data may have caused confusion during the model's learning process, ultimately leading to a negative impact on its overall performance.

These findings are consistent with our original hypothesis - utilizing a pre-trained model with higher translation quality to generate the synthetic back-translations used for fine-tuning a baseline model will ultimately lead to higher quality of translations from the final fine-tuned model. This finding emphasizes the importance of creating high-quality synthetic back-translated data for fine-tuning models with the goal of improving performance for low-resource language translation tasks.

5.3 Experiment 2: Back-Translation Dataset Size

Prior research has shown that there is a threshold where fine-tuning with too much synthetic data can start to decrease the model's performance (Poncelas et al. 2018). We did not reach that threshold in our experiment. As the size of the dataset grew across all the pre-trained models, the BLEU scores increased along with it, indicating that the models were learning effectively with the new data, and that 200k sentences is not enough data to result in overfitting.

There was one exception to this finding - the mBART-50 BLEU scores did not consistently increase as the dataset size increased. This is likely due to the poor translation quality of the mBART-50 model in comparison to the other models, and a result of the model overfitting on its own low-quality outputs, resulting in a feedback loop of poor translations. Upon investigating the drop in performance when moving from the 100k

to 200k dataset, we fine-tuned another model using only the additional 100k rows of data present in the 200k dataset. The resulting BLEU scores were significantly lower than those of the model trained on the first 100k dataset. This finding provides evidence that the transition from the 100k to 200k dataset introduced poor translations that confused the learning process. Consequently, even with a larger dataset, the model exhibited worse performance, suggesting that the quality of additional data is crucial for improving translation performance.

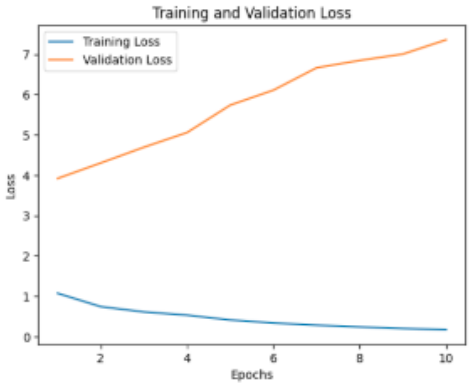

Figure 2: Training and validation loss for model fine-tuned on back-translations generated by mBART-50

Overfitting is further evidenced by the train and validation loss chart, which shows that while the training loss consistently decreases with each epoch, the validation loss increases at a faster rate (Figure 2). These results emphasize the risks associated with training a model on its own outputs, as doing so may lead to overfitting, reinforcement of errors, and ultimately, poorer performance.

6 Future Work

Future studies may extend this research by experimenting with other low-resource language pairs, and perhaps focusing on a low-resource to low-resource language translation, rather than the English-centric approach in our study. Another direction could focus on exploring the impact of the model architecture for the pre-trained models used to generate the back-translations. Prior research has investigated using statistical machine translation models vs. neural machine translation models for back-translation generation (Soto et al. 2020). Future studies might expand on this by also examining the effectiveness of using rule-based model architectures.

7 Conclusion

In conclusion, this study sought to investigate the impact of using back-translated synthetic data generated by various pre-trained models on the performance of a Swahili to English translation model. Our findings were consistent with our original hypotheses - that applying higher quality pre-trained models to generate the back-translations, and using a larger amount of back-translated data for fine-tuning, will ultimately lead to higher translation quality in the final fine-tuned model.

Our experiments demonstrated that the quality of back-translated texts indeed plays a significant role in model performance. We observed that model performance generally improved with larger datasets, as long as overfitting did not occur. Our rewording experiment indicated that although reworded data could provide incremental improvements, the gains were limited, potentially due to the quality of the initial translations used. The results from training the model using its own outputs revealed that such an approach could lead to overfitting and reinforcement of errors which resulted in poorer performance. This finding emphasizes the importance of using diverse and high-quality synthetic data when fine-tuning low-resource language translation models.

Overall, this study contributed insights into the role of synthetic data in the fine-tuning process for low-resource language translation models and brought attention to the factors that influence model performance and offered directions for future research in this area.

References

Brown, Tom B., et al. “Language Models Are Few-Shot Learners.” *ArXiv.org*, 22 July 2020, https://arxiv.org/abs/2005.14165.

Conneau, Alexis, et al. “Unsupervised Cross-Lingual Representation Learning at Scale.” *ArXiv.org*, 8 Apr. 2020, https://arxiv.org/abs/1911.02116.

Fan, Angela, et al. “Beyond English-Centric Multilingual Machine Translation.” *ArXiv.org*, 21 Oct. 2020, https://arxiv.org/abs/2010.11125.

Poncelas, Alberto, et al. “Combining SMT and NMT Back-Translated Data for Efficient NMT.” *ArXiv.org*, 9 Sept. 2019, https://arxiv.org/pdf/1909.03750.pdf.

Poncelas, Alberto, et al. “Investigating Backtranslation in Neural Machine Translation.” *ArXiv.org*, 17 Apr. 2018, https://arxiv.org/abs/1804.06189.

Przystupa, Michael, and Muhammad Abdul-Mageed. “Neural Machine Translation of Low-Resource and Similar Languages with Backtranslation.” *ACL Anthology*, 2019, https://aclanthology.org/W19-5431/.

Ranathunga, Surangika, et al. “Neural Machine Translation for Low-Resource Languages: A Survey.” *ArXiv.org*, 29 June 2021, https://arxiv.org/abs/2106.15115.

Sennrich, Rico, and Biao Zhang. “Revisiting Low-Resource Neural Machine Translation: A Case Study.” *ACL Anthology*, 2019, https://aclanthology.org/P19-1021/.

Sennrich, Rico, et al. “Improving Neural Machine Translation Models with Monolingual Data.” *ArXiv.org*, 3 June 2016, https://arxiv.org/abs/1511.06709.

Soto, Xabier, et al. “Selecting Backtranslated Data from Multiple Sources for Improved Neural Machine Translation.” *ArXiv.org*, 1 May 2020, https://arxiv.org/abs/2005.00308v1.

Tang, Yuqing, et al. “Multilingual Translation with Extensible Multilingual Pretraining and Finetuning.” *ArXiv.org*, 2 Aug. 2020, https://arxiv.org/abs/2008.00401.

Team, NLLB, et al. “No Language Left behind: Scaling Human-Centered Machine Translation.” *ArXiv.org*, 25 Aug. 2022, https://arxiv.org/abs/2207.04672.

Wu, Yonghui, et al. “Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation.” *ArXiv.org*, 8 Oct. 2016, https://arxiv.org/abs/1609.08144.