Federated Meta-Learning for Low-Resource Translation of Kirundi

Kyle Sang
University of Maryland
ksang@umd.edu

Tahseen Rabbani
Yale University
tahseen.rabbani@yale.edu

Tianyi Zhou
University of Maryland
tianyi@umd.edu

Abstract

In this work, we reframe multilingual neural machine translation (NMT) as a federated meta-learning problem and introduce a translation dataset for the low-resource Kirundi language. We aggregate machine translation models (→ en) locally trained on varying (but related) source languages to produce a global meta-model that encodes abstract representations of key semantic structures relevant to the parent languages. We then use PerFedAvg to fit the global model onto a specified target language in a few-shot manner. The target language may live outside the subset of parent languages (such as closely-related dialects or sibling languages), which is particularly useful for languages with limitedly available sentence pairs. We first develop a novel dataset of Kirundi-English sentence pairs curated from Biblication translation. We then demonstrate that a federated learning approach can produce a tiny 4.8M Kirundi translation model and a stronger NLLB-600M model which performs well on both our Biblical corpus and the FLORES-200 Kirundi corpus.

1 Introduction

The federated learning (FL) paradigm has drawn great interest for its inherent privacy, scalability, and performance across myriad vision and language tasks. Recent works have proposed federated learning as a solution for low-resource machine translation (Tupitsa et al., 2024; Moskvoretskii et al., 2024a). Centralized federated learning often focuses on optimizing a global model by aggregating weights over a cluster of clients trained on identical tasks (with varying local datasets). Current literature suggests a global model can also be used as a meta-model to increase model performance and convergence speed (Fallah et al., 2020; Chen et al., 2018). In the meta-learning setting, clients train on similar, but heterogeneous tasks, enabling few-shot adaptation to new tasks of the same flavor.

In this paper, we attempt to utilize federated learning, viewed through the meta-learning lens, to produce a seq2seq translation model for Kirundi, which despite having 11.2 million speakers, is rarely considered in literature and lacks translation resources. Here, the meta-task is → en machine translation, with varying source language. We aggregate a global model over a small cluster of parent seq2seq models. The parent models train higher-resource Bantu languages, specifically Luganda, Bemba, and Kinyarwanda.

To the best of our knowledge, the FLORES-200 dataset (Costa-jussà et al., 2022) is the only publicly-available parallel translation corpus of Kirundi, containing roughly 2000 sentences (aligned with 200 other languages). We produce a novel corpus of 29,506 English to Kirundi sentence pairs by scraping pairs from parallel corpuses of the New and Old Testament produced by The International Bible Society (available at `bible.com`). We demonstrate that the federated meta-learning strategy can boost performance on both the FLORES-200 Kirundi and our Bible corpus. We use our approach to construct a tiny, but performant 4 million parameter run → en model and to improve the performance of NLLB-600M, which has already been trained to translate Kirundi (among many other languages).

2 Algorithm and Preliminaries

The PerFedAvg algorithm combines FedAvg (McMahan et al., 2017), Reptile (Nichol et al., 2018), and personalization to increase convergence speed and stability. To rapidly adapt to a new language for a machine translation task (in our case,

Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL 2025), pages 190–194
March 2, 2025 ©2025 Association for Computational Linguistics

run → en), we split our approach into 3 steps, similar to PerFedAvg (Fallah et al., 2020).

1. **Global Model Training**. Using the FedAvg federated learning algorithm (McMahan et al., 2017), we aggregate gradients across multiple clients. Each client holds data for a language exclusive to them. Training is performed, and gradients are aggregated to update the global model, which is used by clients for the next epoch of training. It is well-known the hetereogeneous weighting of gradients during aggregation is required to achieve optimal performance (Moskvoretskii et al., 2024a; Fallah et al., 2020; Tupitsa et al., 2024; Kairouz et al., 2021). We use the Optuna library to fine-tune gradient weighting rather than using an even average (McMahan et al., 2017) or weighing by the the amount of client data (Fallah et al., 2020).

2. **Reptile Meta-Learning**. The fine-tuning is tested against a subset of Kirundi training data. We run the Reptile algorithm (Nichol et al., 2018) 10 times on our model after training the global model to improve model performance as outlined by PerFedAvg. Reptile enables our meta-model (i.e., federated global model) to quickly adapt to run → en translation by repeatedly sampling other parental translation tasks and performing SGD on each parent task, then updating the initialization parameters in the direction of the of the run → en loss minima. This will prepare our global model for rapid personalization towards our full Kirundi training sets.

3. **Kirundi Personalization**. We take our fine-tuned, Reptile-optimized global model and then perform full training over our Kirundi datasets.

3 Experiments

3.1 Kirundi Dataset

While machine translation work has been performed for other African languages(Vegi et al., 2022; Emezue and Dossou, 2022; Omwoma et al., 2024; Nyoni and Bassett, 2021), besides FLORES-200, there are no other widely-known parallel corpora for Kirundi. Despite having 11.2 million speakers, it is underrepresented in the machine translation community. One of the initiatives of this work was to curate a new dataset of sentence pairs to stimulate further work on this language.

Using the Kirundi Bible we were able to directly translate English sentences to their Kirundi counterparts. The Kirundi verse pairs were extracted and cleaned from the Kirundi Bible found at `https://www.bible.com/`. The dataset itself contains 29,506 sentence pairs. For training purposes, we truncated the full set down to sentence pairs with token lengths $<=11$ (for a total of 1317 pairs) during training with a train dev/test of $80\%/20\%$. We intend to release these sentence pairs on GitHub following the deanonymization of this submission.

3.2 Training

Small seq2seq model. For our tiny model scenario, we use 4.8M parameter Seq2Seq torch models with Bahdanau attention (Bahdanau et al., 2014), Adam optimizers, and NLL loss (Sutskever et al., 2014). Learning rate is set to $1e-5$, weight decay to $1e-4$. FedAvg for our global model is run for 50 communication rounds where every client participates in 1 local epoch per round. After 25 communication rounds, Optuna is used to finetune the gradient weights (i.e., model mixture) every 5 rounds. Reptile is run for 10 rounds. After the global meta-model is prepared for knowledge transfer, we run local Kirundi training (i.e., personalization) for 100 epochs. We source Luganda-English pairs from a published Zenodo set (Kimera et al., 2023) and Kinyarwanda-English pairs from a biblical translation.

NLLB. For federated training of NLLB-600M Kirundi, we adopt the same hyperparameters as the tiny model scenario, but we do not perform Optuna finetuning due to the sheer size of the model. That is, we use equal weighting of the parent Luganda, Bemba, and Kinyarwanda models, with all training and test data sourced from FLORES-200.

3.3 Translation Tasks

3.3.1 Kirundi Bible Corpus

In Table 1, we record the BLEU scores of various models on our Bible corpus. PerFedAvg refers to parental model weighting N_k/N where N_k is the number of training samples for client k and $N = \sum_k N_k$. Equal weighting sets federated weights equal to $1/k$ (in our case $k = 3$). Frozen weights applies an Optuna fine-tuned, Reptile-optimized global meta-model directly on the Kirundi bible test set. No global model trains the tiny seq2seq

Model	BLEU Score
Fine-Tuned FL + Personalization	**20.67**
PerFedAvg Weights	17.66
Equal Weights	17.89
Frozen Weights	17.01
No Global Model	17.70
NLLB-600M	23.85

Table 1: **Kirundi Bible Dataset.** Highest achieved BLEU scores of different algorithms averaged over 3 runs on our Kirundi Bible Corpus. NLLB is also included as a baseline.

model from scratch (no federated learning). Fine-Tuned FL + Personalization weights refers to Optuna+Reptile global model fine-tuning in addition to Kirundi bible train set personalization. We observe that a federated model with fine-tuned parent model mixtures can achieve the highest performance – lagging only the NLLB-600M model which is roughly 125x its size.

3.3.2 FLORES-200 Corpus

Model	BLEU Score
Fine-Tuned FL + Personalization FL	19.26
NLLB-600M (Unchanged Default Weights)	23.46
NLLB-600M (No FL + Personalization)	23.45
NLLB-600M (FL + Personalization)	**25.51**

Table 2: **FLORES-200 Dataset.** Highest achieved BLEU scores of different algorithms averaged over 3 runs on the FLORES-200 Kirundi dataset.

In Table 2, we study how our various models perform on the FLORES-200 Kirundi corpus of roughly 2000 sentence pairs (approximately 1000 pairs for train/test). Fine-tuned FL + Personalization performs respectably on FLORES-200 with no personalization the FLORES train set, indicating the Bible training corpus imbues our tiny model with general knowledge of modern Kirundi. We observe that federated learning is able to improve the performance of NLLB-600M, which is already pre-trained on massive web corpora of Bantu languages (Costa-jussà et al., 2022).

3.3.3 K-shot Learning

We can see across all of our ablation training curves, depicted in Figure 1, using a global model (Fine-Tuned FL) for pre-training leads to an increase in performance. It maintains this improvement in all k-shot tasks. We found that improvement was especially impressive in few-shot learning environments, with consistent increases despite a low amount of accessible training data.

In addition to this, we can also observe a much faster convergence for the pre-trained model in Figure 1. The pre-trained model can be seen converging 5 to 10 rounds before a model trained without a meta-model.

These improvements in training speed and accuracy can be explained by the pre-trained model having already seen similar examples during the training of the global model. With this in mind, using a global model as a meta-model presents an avenue for improving model performance when target language data is low, but data from related languages is available.

3.4 Weighting Algorithms

In Figure 2, we review different weighting strategies and their performance compared to our algorithm. Compared to the PerFedAvg strategy (weighting proportional to size of training data), we can see increased performance in our algorithm. PerFedAvg weights on sample count, but in our case, we have a low number of clients with differing amounts of data. As a result, PerFedAvg weighting results in overfitting to a specific language which is detrimental in obtaining optimal meta-model weights.

We also compare our algorithm to equally weighting gradients from all clients. If finding the truest average of our client languages during our global model training was the most effective for personalization, this strategy would yield the highest performance. However, during our weight tuning, we found that oftentimes certain languages would be weighted as more important to personalization. For example, during training, we found that weights from our Kinyarwanda client would be weighted slightly higher than other clients. Intuitively, this is because Kinyarwanda has a closer lexical similarity to our target language of Kirundi compared to Bemba or Luganda.

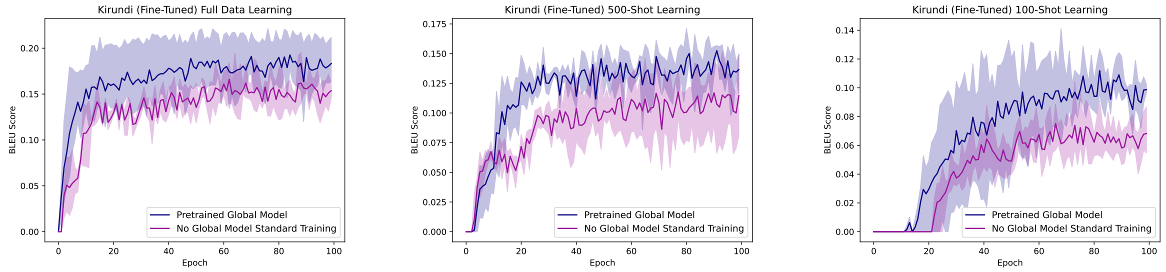

Figure 1: Comparing performances of fine-tuning from a pre-trained global model and training from scratch in different k-shot settings.

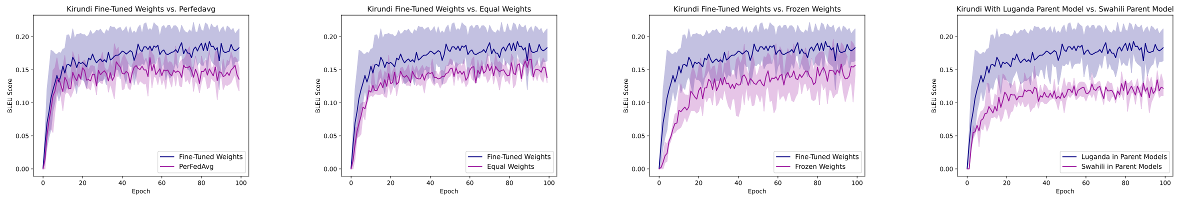

Figure 2: Comparing the performance of different weighting strategies applied during training of the global model.

We also analyze the performance of personalization with frozen intermediate weights. Again, our algorithm outperforms this setting. This demonstrates the task as more than a fine-tuning task, but a more complex meta-learning problem.

From these results, we can surmise that there exists a most optimal set of weights for each client that is based on the lexical similarity of the parent languages used in training the global model to our target languages.

3.5 Parental Model

We also explored the impact of other Bantu languages on our personalization step, replacing Luganda in our parent languages with Swahili. We previously discussed the correlation of lexical similarity to a target language and the importance of a parent language. Other studies have claimed unrelated parent models should not have an impact on the personalization step (Moskvoretskii et al., 2024b). However, from our experiment illustrated in Figure 2, we can see that an unrelated language has deleterious effects on performance. Despite being a Bantu language, Swahili is much less lexically related to Kirundi than Luganda. As a result, the drop in performance can be associated with our Swahili client effectively poisoning the weights of global model with an unrelated task.

4 Conclusion

In this work, we curate a dataset and develop an algorithm for English to Kirundi translation. Despite being a widely spoken Bantu language, there were no previously existing translation resources for Kirundi. Despite limited sentence pairs, our work shows a translation model can be developed with certain federated learning techniques to provide support for an underrepresented language.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Fei Chen, Mi Luo, Zhenhua Dong, Zhenguo Li, and Xiuqiang He. 2018. Federated meta-learning with fast convergence and efficient communication.

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

Chris C. Emezue and Bonaventure F. P. Dossou. 2022. Mmtafrica: Multilingual machine translation for african languages. Proceedings of the Sixth Conference on Machine Translation (2021) 398-411, Association for Computational Linguistics.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. 2020. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in neural information processing systems*, 33:3557–3568.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. 2021. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210.

Richard Kimera, Daniela N Rim, and Heeyoul Choi. 2023. Building a parallel corpus and training translation models between luganda and english. *arXiv preprint arXiv:2301.02773*.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.

Viktor Moskvoretskii, Nazarii Tupitsa, Chris Biemann, Samuel Horváth, Eduard Gorbunov, and Irina Nikishina. 2024a. Low-resource machine translation through the lens of personalized federated learning. *arXiv preprint arXiv:2406.12564*.

Viktor Moskvoretskii, Nazarii Tupitsa, Chris Biemann, Samuel Horváth, Eduard Gorbunov, and Irina Nikishina. 2024b. Low-resource machine translation through the lens of personalized federated learning.

Alex Nichol, Joshua Achiam, and John Schulman. 2018. On first-order meta-learning algorithms.

Evander Nyoni and Bruce A Bassett. 2021. Low-resource neural machine translation for southern african languages. *arXiv preprint arXiv:2104.00366*.

Vincent Omwoma, Lawrance Nderu, Kennedy Ogada, and Tobias Mwalili. 2024. Neural machine translation for low resource bantu languages in east and southern africa. *Research Square*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks.

Nazarii Tupitsa, Samuel Horváth, Martin Takáč, and Eduard Gorbunov. 2024. Federated learning can find friends that are advantageous.

Pavanpankaj Vegi, Sivabhavani J, Biswajit Paul, Prasanna K R, and Chitra Viswanathan. 2022. ANVITA-African: A multilingual neural machine translation system for African languages. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1090–1097, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.