PlayGround Low Resource Machine Translation System for the 2023 AmericasNLP Shared Task

Tianrui Gu, Kaie Chen, Siqi Ouyang, Lei Li
University of California, Santa Barbara
{tianruigu, kaiechen, siqiouyang, leili}@ucsb.edu

Abstract

This paper presents PlayGround's submission to the AmericasNLP 2023 shared task on machine translation (MT) into indigenous languages. We finetuned NLLB-600M, a multilingual MT model pre-trained on Flores-200, on 10 low-resource language directions and examined the effectiveness of weight averaging and back translation. Our experiments showed that weight averaging, on average, led to a 0.0169 improvement in the ChrF++ score. Additionally, we found that back translation resulted in a 0.008 improvement in the ChrF++ score.

1 Introduction

We participated in the AmericasNLP 2023 (Ebrahimi et al., 2023) shared task with the goal of advancing previous studies (Mager et al., 2021) on indigenous American languages. The task is to translate Spanish into 10 indigenous languages, including Ashaninka, Aymara, Bribri, Guarani, Hñähñu, Nahuatl, Quechua, Raramuri, Shipibo-Konibo, and Wixarika. Additionally, there was another language, Chatino[1], for which we did not participate in.

We started with the monolingual and bilingual data from Mager et al. (2021) and finetuned NLLB-600M, a multilingual pre-trained MT model from Meta's No Language Left Behind (NLLB) project (NLLBTeam et al., 2022) both bilingually and multilingually. On top of that, we employed weight averaging and back translation. For back translation, we additionally filtered the back translated sentence pairs to improve the data quality.

We demonstrate that training on model weights averaged from multiple checkpoints improves translation quality, as indicated by a 0.0169 increase in the ChrF++ score on average, without requiring additional computation resources. Additionally, we found that back translation can enhance translation quality for low-resource languages, although it is sensitive to the quality of synthetic data. To address this, we introduced a data filtering technique to improve the quality of synthetic data. With filtered back translation, our system achieved an average improvement of 0.008 in the ChrF++ score. Furthermore, our study reveals that multilingual fine-tuning achieves comparable translation quality to bilingual fine-tuning for low-resource languages.

We selected the bilingual model with weight averaging and back translation as our final submission. The implementation of this study is available in our Git repository[2].

2 Methods

2.1 Data

We adopted the data preparation method described by the University of Helsinki's submission to AmericasNLP 2021 (Vázquez et al., 2021) for our system. The details of the dataset can be found in Table 1. Our model training utilized the filtered parallel data (referred to as parallel data), which consisted of the training data provided by the organizers as well as additional data collected by the University of Helsinki (Vázquez et al., 2021). In order to generate synthetic parallel data (referred to as synthetic data), we employed monolingual data and applied back translation techniques (refer to Section 2.3). The development data was used for model selection purposes.

2.2 Pre-trained Model

Our models are based on the NLLB-600M Seq2Seq pre-training scheme introduced by the NLLB team (NLLBTeam et al., 2022). For tokenization, we utilize the SentencePiece tokenizer (Kudo and Richardson, 2018), following the NLLB configuration. The NLLB model was initially trained on

[1] https://scholarworks.iu.edu/dspace/handle/2022/21028

[2] https://github.com/KaieChen/ameircasnlp2023

Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 173–176

Lang	Filtered	Monoling	Dev
Ashaninka	3858	13195	883
Aymara	8352	16750	996
Bribri	7303	0	996
Guarani	14483	40516	995
Hñähñu	7049	537	599
Nahuatl	17431	9222	672
Quechua	228624	60399	996
Raramuri	16529	0	995
Shipibo-Konibo	28854	23595	996
Wixarika	11525	511	994

Table 1: Number of segments in dataset. Filtered data and monolingual data are collected and filtered by University of Helsinki team (Vázquez et al., 2021) from AmericasNLP 2021.

the Flores-200 dataset, which consists of Aymara, Guarani, Quechua, and Spanish.

2.3 Fine-tuned Models

We fine-tune NLLB-600M using the data mentioned in Table 1. For both X-to-Spanish and Spanish-to-X directions, we fine-tune NLLB-600M using filtered parallel data in both bilingual and multilingual way. This produces 20 bilingual models and 2 multilingual models.

We leverage the above X-to-Spanish models to generate back translated data to enrich the training corpus. Then we further fine-tune the Spanish-to-X models with parallel dataset extended with back translated sentence pairs.

The final models are obtained with weight averaging since the training can be unstable with insufficient data.

2.3.1 Back Translation

In order to make use of monolingual data in indigenous languages, we employed back translation. Specifically, we froze the decoder layers of NLLB model and performed fine-tuning of an X-to-Spanish model using parallel data. Then, we utilized this model to generate synthetic sentences.

Data filtering: Synthetic sentences may contain noise. To address this issue, we implement a data filter to select a subset of synthetic sentences that will expand the original parallel dataset (Ranathunga et al., 2023). In our task, we initially fine-tuned a Spanish-to-X model using the parallel data. Subsequently, we evaluated this model on the synthetic sentences and selected the top N samples with the lowest cross-entropy loss. The value of N is determined by the following:

$$N = \min(|Y_{par}|, |Y_{syn}|) \quad (1)$$

where $|Y_{par}|$ represents the number of segments in the parallel dataset, and $|Y_{syn}|$ represents the number of segments in the synthetic dataset.

Finally, we combined the selected synthetic data with the parallel data and proceeded to perform additional fine-tuning of the NLLB model.

2.3.2 Weight Averaging

Studies have shown that averaging the weights of multiple finetuned models can enhance accuracy (Wortsman et al., 2022). In our training approach, the weights of the next epoch are trained based on the average of the model weights from the previous K epochs. For inference, we compute the final model by averaging the model weights from the last K epochs. The model can be defined as follows:

$$\textbf{NLLB}(x; \Theta_t) = \textbf{NLLB}(x; \frac{1}{K}\sum_{k=1}^{K}\Theta_{t-k}) \quad (2)$$

where Θ_t represents the model parameters at epoch t.

This technique shares similarities with training different models using various hyperparameters (Wortsman et al., 2022; Xu et al., 2020). However, as we only need to train a single model, this technique can be particularly efficient for large language models. The effectiveness of this approach is further discussed in Section 3.

2.3.3 Hyperparameters

In the fine-tuning process, we froze the encoder layers of the NLLB model, considering its prior training on a vast amount of Spanish sentences. We optimized the model using AdamW (Loshchilov and Hutter, 2017) with hyperparameters $\beta = (0.9, 0.999)$, $\epsilon = 10^{-6}$. We employed a learning rate of 3×10^{-4} for a total of $10,000$ iterations. For regularization, we utilized the same dropout rate as the original NLLB model and a weight decay of 0.01. Furthermore, for weight averaging, we set the value of K to be 5.

2.4 Evaluation

We report the results using ChrF++ (Popović, 2017), following the evaluation script[3] provided by the AmericasNLP 2023 shared task. ChrF++

[3] https://github.com/AmericasNLP/americasnlp2023

Target language	Baseline (Test)	Multi	Multi+	Multi++	Bi	Bi++	Bi++ (Test)
Wixarika	**0.304**	0.277	0.294	0.294	0.266	0.279	0.288
Hñähñu	0.147	0.129	0.133	0.138	0.144	0.141	**0.148**
Aymara	0.283	0.291	0.328	0.326	0.336	0.326	**0.300**
Shipibo-Konibo	**0.329**	0.224	0.238	0.253	0.261	0.283	0.277
Nahuatl	**0.266**	0.241	0.252	0.275	0.282	0.283	0.237
Guarani	**0.336**	0.304	0.316	0.321	0.315	0.303	0.331
Asháninka	0.258	0.222	0.238	0.272	0.269	0.286	**0.280**
Quechua	0.343	0.324	0.341	-	0.337	-	**0.344**
Rarámuri	**0.184**	0.161	0.175	-	0.184	-	0.145
Bribri	**0.165**	0.210	0.237	-	0.231	-	0.148

Table 2: Result in ChrF++ on develop dataset, except for baseline and Bi++(test). Baseline model is the best submission for AmericasNLP 2021. The effectiveness of weight averaging (Multi+ and Bi+) and back translation is compared (Multi++ and Bi++). We also compared the performance of bilingual (Bi) and multilingual (Multi).

captures the character-level performance, making it particularly suitable for evaluating the polysynthetic properties observed in many indigenous languages (Zheng et al., 2021).

3 Results

The results are presented in Table 2 for both the development and test datasets. Our **Bi++** model demonstrates improvements in four languages: Hñähñu, Aymara, Asháninka, and Quechua, compared to the **Baseline** model provided by the organizer. In general, the trends in results for the development and training datasets are similar, except for Rarámuri and Bribri. This discrepancy may be attributed to the test dataset containing more unknown tokens, to which our model is sensitive.

Previous study (Mager et al., 2021) has primarily focused on fine-tuning bilingual machine translation models. However, the results from our **Multi++** and **Bi++** models demonstrate the promising potential of multilingual fine-tuning (Tang et al., 2020). On average, the ChrF++ score for **Multi++** is only 0.0012 lower than that of **Bi++**.

We also compared the effectiveness of weight averaging and back translation. Weight averaging improved translations for all target languages. On average, **Multi+** achieved a ChrF++ score that was 0.0169 higher than **Multi**. These results indicate that our simple technique can enhance low-resource machine translation without requiring additional computational resources.

However, the impact of back translation varied across languages, as observed in the results for **Multi+** and **Multi++**. On average, the implementation of back translation resulted in a 0.008 improvement in the ChrF++ metric. For Wixarika and Aymara, there was a slight drop in the ChrF++ scores after back translation. Despite performing data filtering, the quality of synthetic data largely depends on the performance of the X-to-Spanish model.

In summary, our fine-tuning technique has shown improvements in performance. However, with further refinements and design enhancements, there is potential for our model to achieve higher levels of performance.

4 Conclusion

In this paper, we presented our submission to the AmericasNLP 2023 shared task. Our system utilized the NLLB-600M pre-trained model to translate Spanish into 10 indigenous languages. We also investigated the potential of multilingual translation models, which showed promising results. Additionally, we found that averaging model weights from previous epochs proved to be an efficient and effective approach. While back translation demonstrated performance improvements, further methods are necessary to address noisy data. These findings highlight the positive outcomes of our study and provide valuable insights for future advancements in low-resource machine translation techniques.

References

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Enora Rice, Cynthia Montaño, John Ortega, Shruti Rijhwani, Alexis Palmer, Rolando Coto-Solano, Hilaria Cruz, and Katharina Kann. 2023. Findings of the AmericasNLP 2023 shared task on machine translation into indigenous languages. In *Proceedings of*

the Third Workshop on Natural Language Processing for Indigenous Languages of the Americas. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2021. Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217, Online. Association for Computational Linguistics.

NLLBTeam, Marta R. Costa-jussà, 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, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Comput. Surv.*, 55(11).

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. *CoRR*, abs/2008.00401.

Raúl Vázquez, Yves Scherrer, Sami Virpioja, and Jörg Tiedemann. 2021. The Helsinki submission to the AmericasNLP shared task. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 255–264, Online. Association for Computational Linguistics.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23965–23998. PMLR.

Yige Xu, Xipeng Qiu, Ligao Zhou, and Xuanjing Huang. 2020. Improving BERT fine-tuning via self-ensemble and self-distillation. *CoRR*, abs/2002.10345.

Francis Zheng, Machel Reid, Edison Marrese-Taylor, and Yutaka Matsuo. 2021. Low-resource machine translation using cross-lingual language model pretraining. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 234–240, Online. Association for Computational Linguistics.