ChatGPT is not a good indigenous translator

David Stap **Ali Araabi**
Language Technology Lab
University of Amsterdam
{d.stap, a.araabi}@uva.nl

Abstract

This report investigates the continuous challenges of Machine Translation (MT) systems on indigenous and extremely low-resource language pairs. Despite the notable achievements of Large Language Models (LLMs) that excel in various tasks, their applicability to low-resource languages remains questionable. In this study, we leveraged the AmericasNLP competition to evaluate the translation performance of different systems for Spanish to 11 indigenous languages from South America. Our team, LTLAmsterdam, submitted a total of four systems including GPT-4, a bilingual model, fine-tuned M2M100, and a combination of fine-tuned M2M100 with kNN-MT. We found that even large language models like GPT-4 are not well-suited for extremely low-resource languages. Our results suggest that fine-tuning M2M100 models can offer significantly better performance for extremely low-resource translation.

1 Introduction

This paper presents the participation of the Language Technology Lab (LTL) from the University of Amsterdam in the AmericasNLP 2023 Shared Task, which aims to develop Machine Translation (MT) systems for indigenous languages of the Americas. We submitted translation results for Spanish into all indigenous languages: Hñähñu (oto), Wixarika (hch), Nahuatl (nah), Guaraní (gn), Bribri (bzd), Rarámuri (tar), Quechua (quy), Aymara (aym), Shipibo-Konibo (shp), Asháninka (cni), and Chatino (czn). In the face of limited parallel and monolingual data, our approaches focus on maximizing the potential of available resources and models. Specifically, our objectives include: 1) evaluating the performance of GPT-4, a state-of-the-art language model, in extremely low-resource settings; 2) utilizing a carefully optimized transformer setting for low-resource NMT (Araabi and Monz, 2020; Zwennicker and Stap); 3) exploring the effectiveness of a fine-tuned version of the multilingual M2M100 (Fan et al., 2021) model; and 4) investigating the potential of augmenting a neural model with a k-nearest-neighbor machine translation (kNN-MT) (Khandelwal et al., 2021) component. Our experiments show that, while GPT-4 demonstrates comparable performance to bilingual models trained from scratch, it significantly lags behind the fine-tuned M2M100 models. Notably, our M2M100+kNN-MT system emerged as our top-performing approach.

2 Data and preprocessing

2.1 Data

We used the parallel data provided by the shared task organizers for training across all South American languages (Ebrahimi et al., 2023). For Chatino, we used the Chatino Speech Corpus Archive Dataset.[1]

2.2 Pre-processing

We apply punctuation normalization, tokenization, data cleaning, and true-casing using the Moses scripts (Koehn et al., 2007). The sentence length is limited to a maximum of 175 tokens during training. After replacing phrases with variables, we also apply BPE segmentation (Sennrich et al., 2016) with the BPE parameter effective for each data size (Araabi and Monz, 2020). Table 2 presents the number of training samples following preprocessing.

3 Models and experiments

3.1 GPT-4

Large Language Models (LLM) such as ChatGPT are strong translation models for high-resource languages (Jiao et al., 2023). For low-resource languages, ChatGPT produces relatively few halluci-

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

Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 163–167

The translations provided may not be entirely accurate but are the closest possible translations using available resources.
Here is the translation for `[TGT]`
(Note: The provided text does not contain correct Ashaninka punctuation.)
The 10th sentence cannot be translated since it goes against Bribri cultural values, as the language has certain rules that may prevent imperative structures or indications that may go against tribal communication.
These translations are based on the Mezquital Ñahñu dialect.
(Note: Nahuatl is an indigenous language spoken in Mexico and there exists variation in vocabulary and grammar across different dialects, regions and time periods, so these translations may differ from other possible versions)
(Seq items above included due to impossibility of translating Mary Todd, Abe Lincoln, Illinois, and A Christmas Carol.)
The translations provided are approximations based on the Bribri language structure and available lexical resources.
(Note: Unfortunately, there isn't much appropriate vocabulary available online for Hnahnu/Ñähñu Language, which is an indigenous language of Mexico. I tried my best to give some translations, but the result may not be accurate.)

Table 1: Some examples of unwanted output produced by ChatGPT during translation.

Language	code	#sentences	#subwords
Asháninka	cni	3869	5k
Aymara	aym	13000	10k
Bribri	bzd	7502	5k
Guaraní	gn	26011	20k
Nahuatl	nah	15898	20k
Hñähñu	oto	4838	5k
Quechua	quy	250709	20k
Rarámuri	tar	13754	10k
Shipibo	shp	29126	20k
Wixarika	hch	8963	10k
Chatino	czn	310	5k

Table 2: Number of training samples and vocabulary size after preprocessing.

nations under perturbation, and its hallucinations are qualitatively different from conventional translation models (Guerreiro et al., 2023). It remains unclear how well LLMs perform when translating into *extremely* low-resource languages.

We use the ChatGPT (`gpt-4`) API[2] to translate Spanish source languages into the indigenous target languages. Following (Jiao et al., 2023) we use the following translation prompt: "*Please provide the* `[TGT]` *translation for these sentences:*". We add the following role content: "*You are a machine translation system.*" (Peng et al., 2023). Initial experiments with Temperature set to 0 (Peng et al., 2023) produce results that are inferior to the default Temperature value, so we stick to the latter. During translation, ChatGPT frequently added boilerplate text to translations such as "*Feel free to make adjustments if you have a better understanding of the language.*". See Table 1 for additional examples of unwanted ChatGPT boilerplate outputs. While some of these outputs, such as the warnings about inaccurate translations, can be valuable to machine translation users, we remove this boilerplate text in a post-processing step before evaluating the translations.

3.2 Bilingual

To conduct our bilingual experiments, we employ Transformer models (Vaswani et al., 2017) with parameters proposed by Araabi and Monz (2020), specifically tailored to extremely low-resource data regime. We use the Fairseq library (Ott et al., 2019) for our experiments.

3.3 Finetuned M2M100

Following Adelani et al. (2022), we fine-tuned the multilingual M2M100 model (Fan et al., 2021) for translations from Spanish to Indigenous languages.

M2M100 necessitates specifying the target language tag during decoding. Given that the Indigenous languages of interest are not part of M2M100, we adopted the approach suggested by Adelani et al. (2022) and selected a language tag that *is* represented in the pre-trained model. Preliminary results indicated that the translation quality remained unaffected by the choice of the target language tag, so we chose Swahili as the target language.

We used the 418M parameter version of M2M100 and trained individual models for each of the 11 target languages. These models were fine-tuned using the HuggingFace toolkit (Wolf et al., 2020). We employed the default learning rate of

[2] `https://platform.openai.com/docs/api-reference/chat`

model	oto	hch	nah	gn	bzd	tar	quy	aym	shp	cni	czn	avg
GPT-4	0.119	0.169	0.161	0.160	0.106	0.141	0.264	0.203	0.180	0.194	–	0.170
bilingual	0.073	0.185	0.072	0.120	0.113	0.113	0.133	0.146	0.129	0.217	**0.293**	0.145
M2M100	0.131	0.287	0.299	0.301	0.198	0.141	0.360	0.295	0.203	0.262	0.146	0.238
+*k*NN	**0.178**	**0.458**	**0.527**	**0.402**	**0.401**	**0.292**	**0.475**	**0.459**	**0.470**	**0.425**	0.158	**0.386**
GPT-4	0.117	0.157	0.159	0.155	0.094	0.130	0.258	0.183	0.162	0.189	–	0.160
bilingual	0.078	0.210	0.070	0.119	0.123	0.114	0.150	0.140	0.124	0.216	**0.366**	0.155
M2M100	0.139	0.304	0.260	0.329	0.214	0.151	0.368	0.252	0.198	0.260	0.144	0.238
+ *k*NN	**0.145**	**0.319**	**0.273**	**0.341**	**0.261**	**0.180**	**0.370**	**0.276**	**0.279**	**0.300**	0.152	**0.263**

Table 3: Chrf++ scores for Spanish→X directions on the development set (top rows) and test set (bottom rows). Best results are depicted in **bold**, and best results that do not encode the development set are underlined.

5e−5, set the maximum source and target length to 200, and stop training after 3 epochs.

3.4 Finetuned M2M100 + *k*NN-MT

We made the decision to withdraw this model from the competition track due to its encoding of the development set. Although it does not technically violate the competition rule (which states: "*The only limitation is that we ask participants to not have the test input translated by hand or train on the development or test sets*"), our solution operates in a grey area and confers an unfair advantage over other submissions. That said, we describe the approach below.

We operated under the assumption that the provided development data is similar to the test data. Using the development data during training or an additional fine-tuning step is a clear strategy for leveraging this similarity when aiming for enhanced performance on the development and test set domains. However, we opted for an alternative approach that permits more fine-grained control over the degree to which the resulting model depends on the development data as opposed to the training data. Furthermore, we explicitly sought to prevent encoding information about the development set within the resulting model weights, as this could potentially lead to overfitting and reduced generalization capabilities. Such an outcome would undermine the primary objective of creating a robust and versatile MT system that can effectively handle a wide range of input data in the context of Indigenous languages.

k-nearest-neighbor machine translation (kNN-MT) is a semi-parametric model that combines a parametric component with a nearest neighbor retrieval mechanism that allows direct access to a datastore of cached examples (Khandelwal et al., 2021). The datastore consists of key-value pairs, where each key is a decoder output representation, and the value is the corresponding target token. At inference time, the model searches the datastore to retrieve the set of k nearest neighbors, and combines the resulting distribution with the NMT distribution through interpolation.

For our submissions, we encoded the development sets of all Spanish to X directions in separate datastores. We do a grid search over *k*NN hyperparameters $\lambda \in \{0.2, 0.3, ...0.7\}$, $k \in \{8, 16, 32\}$ and $T \in \{50, 100\}$ on oto and hch. Based on these results we fix λ to 0.3, k to 32, and T to 50 and report results for those. We use the *k*NN-transformers library (Alon et al., 2022) for our experiments.

4 Results

We report Chrf++ scores (Popović, 2017) in Table 3. In general, we observe similar patterns for the development and test sets. Comparing GPT-4 and our bilingual models, we conclude that GPT-4 is better for 7/10 directions on both the development and test set. Scores for both models are very low; neither ChatGPT nor bilingual NMT are good indigenous translators.

Our *k*NN approach yields best results for 10/11 language directions, and the fine-tuned M2M100 is the best model that does not encode the development set.

Compared to other submissions, our kNN model ranks first for Spanish-Bribri, Spanish-Asháninka, and Spanish-Nahuatl, but we decided to withdraw this model (see Section 3.4).

5 Conclusion

In this paper, we describe our submissions to the AmericasNLP 2023 Shared Task on Machine Translation into Indigenous Languages. We submitted translations for all 11 languages. Our best system is the result of finetuning M2M100 on an unseen indigenous language, and augmenting this model with a *k*-nearest-neighbor datastore based on the development set.

This model ranked first in the Spanish-Bribri, Spanish-Asháninka, and Spanish-Nahuatl language pairs in the competition. However, we have made the decision to withdraw this model due to its operation in a grey area with respect to the competition rules. The uncertainty surrounding its compliance raises concerns about fairness among all participants, prompting us to take this action after discussion with the organizers.

References

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022. A Few Thousand Translations Go a Long Way! Leveraging Pre-trained Models for African News Translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

Uri Alon, Frank Xu, Junxian He, Sudipta Sengupta, Dan Roth, and Graham Neubig. 2022. Neuro-symbolic language modeling with automaton-augmented retrieval. In *International conference on machine learning*, pages 468–485. PMLR.

Ali Araabi and Christof Monz. 2020. Optimizing Transformer for low-resource neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3429–3435, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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.

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. 2021. Beyond English-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Nuno M. Guerreiro, Duarte Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in Large Multilingual Translation Models. ArXiv:2303.16104 [cs].

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine. ArXiv:2301.08745 [cs].

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2021. Nearest Neighbor Machine Translation. ArXiv:2010.00710 [cs].

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*. The Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations*, pages 48–53. Association for Computational Linguistics.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards Making the Most of ChatGPT for Machine Translation. ArXiv:2303.13780 [cs].

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz

Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, Long Beach, California. Neural Information Processing Systems (NIPS). ArXiv: 1706.03762.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick Von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Just Zwennicker and David Stap. Towards a general purpose machine translation system for sranantongo. In *Proceedings of the 2022 EMNLP Workshop WiNLP*, Abu Dhabi, United Arab Emirates (Hybrid).