Advancing NMT for Indigenous Languages: A Case Study on Yucatec Mayan and Chol

Julio C. Rangel and **Norio Kobayashi**
RIKEN Information R&D and Strategy Headquarters,
2-1 Hirosawa, 351-0198 Wakoshi, Japan
{juliocesar.rangelreyes, norio.kobayashi}@riken.jp

Abstract

This study leverages Spanish-trained large language models (LLMs) to develop neural machine translation (NMT) systems for Mayan languages. For that, we first compile and process a low-resource dataset of 28,135 translation pairs of Chol and Yucatec Mayan extracted from documents of the CPLM Corpus (Martínez et al.). Then we implement a prompt-based approach to train one-to-many and many-to-many models. By comparing several training strategies for two LLMs, we found that, on average, training multilingual models is better, as shown by the ChrF++ reaching 50 on the test set in the best case. This study reinforces the viability of using LLMs to improve accessibility and preservation for languages with limited digital resources. We share our code, datasets, and models to promote collaboration and progress in this field [1].

1 Introduction

In recent times, there has been a push towards creating NLP tools for the native languages of the Americas (Mager et al., 2023). Within this context, Mayan languages have not received attention in machine translation (NMT) studies, despite their deep linguistic roots and large speaker populations. Our study aims to bridge this gap by specifically developing and refining NMT systems for Mayan languages. By leveraging advancements in large language models (LLMs) pre-trained in Spanish, we aim to overcome the scarcity of a comprehensive parallel corpus for Mayan languages. As a result, building NMT systems for languages could greatly benefit these language communities by enabling them to access services and information related to law, healthcare, and finance in their mother tongues.

2 Languages

The Mayan languages form a family spoken by the Maya peoples, is primarily spoken across various regions in Central America. This family stands as among the most thoroughly researched and documented in the Americas (Campbell, 2000). It is believed that the contemporary Mayan languages originated from the Proto-Mayan language, which was likely spoken over 5,000 years ago. This ancient language eventually branched out into at least six distinct lineages: Huastecan, Quichean, Yucatecan, Qanjobalan, Mamean, and Ch'olan–Tzeltalan [2]

2.1 Yucatec Mayan

Yucatec Mayan, commonly referred to as Maya is a language spoken in the Yucatán Peninsula and the northern regions of Belize. Being one of the Mayan languages Yucatec Mayan plays a vital role, in connecting us to the diverse cultural and historical legacy of the Mayan civilization. Unlike indigenous languages Yucatec Mayan boasts a substantial number of speakers with an estimated count of approximately 800,000 individuals [3].

2.2 Chol

The Chol people, a group, in Mexico mainly live in the mountains of Chiapas. Being part of the Maya community they speak Ch'ol or Chol which belongs to the Mayan language group. Ch'ol has three dialects (Sabanilla, Tilá and Tumbalá), these dialects,often considered a single language showcasing the language's vitality and regional diversity. Had approximately 140,806 speakers, in 2000 including individuals who speak only this language[4].

This paper outlines the design and implementation of a comparative study on two sophisticated

[1] https://github.com/RIKEN-DKO/iikim_translator

[2] https://en.wikipedia.org/wiki/Mayan_languages

[3] https://en.wikipedia.org/wiki/Yucatec_Maya_language

[4] https://en.wikipedia.org/wiki/Chol_people

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 138–142
June 21, 2024 ©2024 Association for Computational Linguistics

neural machine translation (NMT) models, T5S (T5 Spanish) and M2M100, specifically tailored for the translation of Yucatec Mayan and Chol languages. We focus mainly on comparing the models' accuracy in translating to the Mayan languages, aiming to determine the most effective approach for developing NMT systems that can serve as a starting point for future Mayan-based NMT systems in low and high resource instances.

3 Methodology

3.1 Dataset

In support of our research, we gathered a dataset comprising 28,135 translation pairs from Spanish to Chol languages using the CPLM (Parallel Corpus, for Mexican Languages) web tool[5]. The data extraction process involved downloading ZIP files, each potentially containing multiple files with parallel sentences in Spanish and one or more target languages. We utilized the *langdetect*[6] library to verify the presence of Spanish; ZIP files without Spanish were excluded. To identify relevant files for Yucatec Mayan and Chol, we looked for language codes 'yua' and 'MY' for Mayan and 'ctu' and 'CHL' for Chol. If codes were absent, we searched for language names such as 'maya' and 'chol'. Finally, we aligned the files to create Spanish-to-Chol and Spanish-to-Mayan parallel datasets. The number of parallel pairs per language is shown in Table 1.

Language	Original # Pairs	Cleaned	Train	Dev/Test
maya-spanish	16149	13528	12176	1352
chol-spanish	11986	10660	9594	1066

Table 1: Summary of the dataset used for training and testing the NMT models.

3.2 Data Preparation

Inspired by previous NMT systems for Indigenous Languages (De Gibert et al., 2023), in the post-processing phase, we applied a length ratio filter to improve the quality of our translation pairs, removing any with a character length ratio exceeding 4. This filtering step was critical for maintaining a high-quality dataset by excluding pairs that could adversely affect translation accuracy. We then randomly divided the sentences into training, development, and testing sets. The results of this data preparation phase, including the final counts of translation pairs, are detailed in Table 1.

4 Models

Our selection criteria focused on recent models with extensive pretraining in Spanish, as evidence suggests this significantly aids in translating to native languages (Vázquez et al., 2021). Accordingly, we selected the T5S (T5 Spanish) and M2M100 (480M version) models for our translation tasks. While both models adopt the encoder-decoder architecture foundational to Transformer models, they are distinguished by their underlying philosophies and optimizations.

This research aims to compare two approaches to language models (LLMs): T5S, which is versatile for various NLP tasks, and M2M100, which is specialized for translation purposes. This comparison intends to evaluate how well a general model like T5S can handle low resource translation scenarios and determine the performance of M2M100 in translating between less commonly spoken languages. Through this method, we aim to identify which model design and training approach are most effective in creating Mayan NMT systems.

4.1 T5S (T5 Spanish)

The T5 model, recognized for addressing a range of NLP tasks as text-to-text conversions—including translation, question answering, and classification—generates target text from input (Raffel et al., 2020). Its variant, IndT5, has been applied for translating Spanish into 10 Indigenous languages (Nagoudi et al., 2021). We utilize T5S (Araujo et al., 2023), an iteration adapted from T5.1.17, featuring an encoder-decoder structure with 12 layers, 12 attention heads, and 768 hidden dimensions. T5S was pretrained on Spanish data totaling approximately 674GB, comprising the OSCAR 21.09 corpus (160GB), mC4-es corpus (500GB), and SUC corpus (14GB).

4.2 M2M100

We use the M2M100 (480M) model, with 12 encoders, 16 decoder layers,a feed forward network (FFN) size of 4096, and embedding dimensions of 1024 that have been optimized for machine translation. It allows for translation among 100 languages, including Spanish, without requiring a language. Trained on a dataset of over 1.5 billion sequences (Fan et al., 2020) it aligns with our strategy of

[5] http://www.geco.unam.mx/

[6] https://github.com/Mimino666/langdetect

utilizing pre trained Spanish language models. Previous studies have showcased the effectiveness of M2M100 in translating to languages like Mixtec (Tonja et al., 2023) along with its performance in tasks such as the AmericasNLP 2024 Shared Task (Stap and Araabi, 2023).

5 Experiments and Results

This section outlines our training approach, experimental setup, and the results obtained from deploying various strategies on the T5S and M2M100 models.

5.1 Training Methodology

For our experiments, both models were trained until no improvement was observed for three consecutive epochs on the development set, with the best-performing checkpoint on this set being used for testing. To enable a single model to translate between multiple languages, we adopted a prompt format of "`{source_text} translate {source_lang} to {target_lang}: {source_text}`" before tokenization and training commenced. This approach facilitated the development of models capable of one-to-many (Spanish to Indigenous languages and vice versa) translations.

5.2 Experimental Setup and Results

Table 2 summarizes the experimental results, presenting both models' performances across different training configurations. "Mayan and Chol" refers to a one-to-many model trained on both languages. In contrast, "Mayan" and "Chol" indicate models trained exclusively on a single language. The "Zero shot" configuration evaluates model performance without fine-tuning. All models were trained for translation from Spanish to a native language, except those with prefixes 'bi', indicating bi-directional training.

In addition to the base dataset, we explored the impact of augmenting it with additional data from the Americas NLP2023 Shared Task[7] (AmeNLP), which introduces 11 more target languages. The inclusion of AmeNLP data initially led to a decrease in performance metrics for both models. However, implementing a uniform sampling strategy mitigated this degradation for combinations of Mayan languages with AmeNLP data but was less effective for Mayan and Chol alone. This observation suggests that the uniform sampling strategy is more advantageous when a model is trained across multiple datasets.

The M2M100 model outperformed T5S in translating Mayan and Chol languages, likely due to its specialization in translation tasks. For both models, the best average results were achieved when training with the Mayan and Chol datasets combined in a one-to-many approach. Interestingly, M2M100 showed a slight improvement on the Chol test set with the inclusion of "AmeNLP + uniform", suggesting that this strategy holds promise for enhancing multilingual model performance with additional data sources. The "biAmeNLP + uniform" strategy did not yield as positive results, possibly due to the requirement of specifying a target language tag for M2M100, which our Indigenous languages lack. Further investigation is needed to fully understand this aspect, despite indications that translation quality remains consistent irrespective of the chosen target language tag (Stap and Araabi, 2023).

For T5S, the "Maya and Chol" configuration was confirmed as the most effective strategy, with "biAmeNLP + uniform" emerging as the second-best approach. This suggests that for T5S, a bidirectional model is preferable, potentially because T5S does not necessitate explicit source or target language tags.

6 Conclusion and Future Work

This study has successfully demonstrated that the T5S and M2M100 models can be adapted for translation tasks between Spanish and Mayan languages, showcasing the potential of neural machine translation (NMT) in enhancing language preservation and accessibility. The M2M100 model, with its translation-focused architecture, excels in one-to-many translation scenarios. Conversely, the T5S model shows versatility in managing bidirectional translations, benefiting from its flexible design.

The incorporation of the "AmeNLP + uniform" strategy has emerged as a promising method to broaden the models' capabilities across multiple languages, though it introduces challenges that necessitate further exploration. Initial experiments have validated the potential of NMT for Yucatec Mayan and Chol, with both models performing effectively in low-resource settings. Despite the variation in translation quality, the results affirm the capacity of these models to acquire meaningful

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

Table 2: Comparative performance metrics. **Bold** denotes overall best; underscore for best ST5 results.

Dataset	Set	Maya		Chol		Average	
		ChrF++	BLEU	ChrF++	BLEU	ChrF++	BLEU
				T5S			
Mayan and Chol	dev	32.69	10.17	34.6	11.51	33.645	10.84
	test	33.13	10.39	35.1	11.42	34.115	10.905
biAmeNLP + uniform	dev	29.53	8.23	32.92	10.35	31.225	9.29
	test	29.91	8.11	33.03	10.17	31.47	9.14
AmeNLP + uniform	dev	21.27	4.38	23.24	5.1	22.255	4.74
	test	21.1	4.01	23.42	5.02	22.26	4.515
Mayan	dev	27.63	7.09				
	test	27.52	6.99				
Chol	dev			28.1	7.83		
	test			28.7	8.03		
Zero shot	dev	7.68	0.13	7.55	0.1	7.615	0.115
	test	7.65	0.09	7.44	0.08	7.545	0.085
				M2M100			
Mayan and Chol	dev	**50.56**	**27.5**	**48.22**	**23.92**	**49.39**	**25.71**
	test	51.48	28.85	48.88	25.11	**50.18**	**26.98**
AmeNLP + uniform	dev	49.11	25.48	47.85	24	48.48	24.74
	test	50.07	26.36	**48.89**	**25.17**	49.48	25.765
Mayan	dev	50.31	27.27				
	test	**51.55**	**29.13**				
Chol	dev			47.38	23.41		
	test			48.27	24.66		
biAmeNLP + uniform	dev	47.43	22.44	47.27	23.16	47.35	22.8
	test	47.99	22.98	48.21	24.22	48.1	23.6
Zero shot	dev	10.2	1.2	10.37	1.31	10.285	1.255
	test	9.89	0.62	10.26	1.25	10.075	0.935

translations from scant data.

This study marks the beginning of exploring NMT systems designed specifically for Mayan languages, highlighting both the possibilities and challenges of using NMT for languages, with resources. Moving forward, our future efforts will concentrate on expanding datasets and investigating active learning and few shot learning approaches. Furthermore, we plan to customize the M2M100 model for other native languages and delve into the nuances of tag selection to enhance translation accuracy. By progressing in these areas, we aim not only to improve the effectiveness of NMT systems for languages but also to contribute to the broader field of language technology and digital transformation (DX). Additionally, we plan to apply the techniques developed here to other low-resource scenarios, such as natural language to SPARQL translation.

References

Vladimir Araujo, Maria Mihaela Trusca, Rodrigo Tufiño, and Marie-Francine Moens. 2023. Sequence-to-sequence spanish pre-trained language models. (arXiv:2309.11259). ArXiv:2309.11259 [cs].

Lyle Campbell. 2000. *American Indian languages: the historical linguistics of Native America*, volume 4. Oxford University Press.

Ona De Gibert, Raúl Vázquez, Mikko Aulamo, Yves Scherrer, Sami Virpioja, and Jörg Tiedemann. 2023. Four approaches to low-resource multilingual nmt: The helsinki submission to the americasnlp 2023 shared task. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, page 177–191, Toronto, Canada. 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. (arXiv:2010.11125).

Manuel Mager, Rajat Bhatnagar, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2023. Neural machine translation for the indigenous languages of the americas: An introduction. (arXiv:2306.06804). ArXiv:2306.06804 [cs, stat].

Gerardo Sierra Martínez, Cynthia Montaño, Gemma Bel-Enguix, Diego Córdova, and Margarita Mota Montoya. Cplm, a parallel corpus for mexican languages: Development and interface.

El Moatez Billah Nagoudi, Wei-Rui Chen, Muhammad Abdul-Mageed, and Hasan Cavusogl. 2021. Indt5: A text-to-text transformer for 10 indigenous languages. (arXiv:2104.07483).

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67. ArXiv: 1910.10683 Citation Key: Raffel2020.

David Stap and Ali Araabi. 2023. Chatgpt is not a good indigenous translator. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, page 163–167, Toronto, Canada. Association for Computational Linguistics.

Atnafu Lambebo Tonja, Christian Maldonado-Sifuentes, David Alejandro Mendoza Castillo, Olga Kolesnikova, Noé Castro-Sánchez, Grigori Sidorov, and Alexander Gelbukh. 2023. Parallel corpus for indigenous language translation: Spanish-mazatec and spanish-mixtec. (arXiv:2305.17404). ArXiv:2305.17404 [cs].

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*, page 255–264, Online. Association for Computational Linguistics.