Parallel Corpus for Indigenous Language Translation: Spanish-Mazatec and Spanish-Mixtec

Atnafu Lambebo Tonja[1], **Christian Maldonado-Sifuentes**[2], **David Alejandro Mendoza Castillo**[2], **Olga Kolesnikova**[1], **Noé Castro-Sánchez**[3], **Grigori Sidorov**[1], **Alexander Gelbukh**[1]

[1]Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC), Mexico,

[2]Transdisciplinary Research for Augmented Innovation - Laboratory (TRAI-L), Mexico,

[3]Departamento de Ciencias Computacionales Tecnológico Nacional de México, Mexico

Abstract

In this paper, we present a parallel Spanish-Mazatec and Spanish-Mixtec corpus for machine translation (MT) tasks, where Mazatec and Mixtec are two indigenous Mexican languages. We evaluated the usability of the collected corpus using three different approaches: transformer, transfer learning, and fine-tuning pre-trained multilingual MT models. Fine-tuning the Facebook M2M100-48 model outperformed the other approaches, with BLEU scores of 12.09 and 22.25 for Mazatec-Spanish and Spanish-Mazatec translations, respectively, and 16.75 and 22.15 for Mixtec-Spanish and Spanish-Mixtec translations, respectively. The findings show that the dataset size (9,799 sentences in Mazatec and 13,235 sentences in Mixtec) affects translation performance and that indigenous languages work better when used as target languages. The findings emphasize the importance of creating parallel corpora for indigenous languages and fine-tuning models for low-resource translation tasks. Future research will investigate zero-shot and few-shot learning approaches to further improve translation performance in low-resource settings. The dataset and scripts are available at https://github.com/atnafuatx/Machine-Translation-Resources.

1 Introduction

Natural Language Processing (NLP), a sub-field of Artificial Intelligence (AI), has been attracting a lot of attention in terms of research and development as a result of the surge in the number of applications it has in a variety of different industries (Kalyanathaya et al., 2019). Machine Translation (MT), Sentiment or Opinion Analysis, POS Tagging, Question Classification (QC) and Answering (QA), Chunking, Named Entity Recognition (NER), Emotion Detection, and Semantic Role Labeling are currently highly researched areas in various high-resource languages (Tonja et al., 2023a).

The domain of machine translation (MT) is advancing at a rapid pace due to the growing prevalence of computational tasks and the expanding global reach of the Internet, which caters to diverse, multilingual communities (Kenny, 2018). MT systems have demonstrated remarkable translation outcomes for language pairs that possess abundant resources, such as English-Spanish, English-French, English-Russian, and English-Portuguese. However, in scenarios with limited or no resources, MT systems encounter difficulties due to the primary obstacle of inadequate training data for certain languages (Mager et al., 2018; Tonja et al., 2021, 2022, 2023b).

Low-resource languages have been suffering from a lack of new language technology designs. When the resources are limited and only a small amount of unlabeled data is available, it is very hard to reach a true breakthrough in creating powerful novel methods for language applications (Tonja et al., 2022), the problem becomes worse if there is no parallel dataset for certain languages.

Mexico is a multicultural and multilingual country with 68 officially recognized indigenous languages, 238 variants, and Spanish, a widely used language spoken by 90 percent of the population (Mager et al., 2021). Few language technologies have been developed for indigenous languages spoken in Northern and Southern America; moreover, many indigenous languages spoken in the Americas face a risk of extinction (Mager et al., 2018).

Indigenous language speakers often experience feelings of shame or reluctance to use their native languages, primarily due to limited opportunities for application in the presence of pervasive, dominant majority languages (Hornberger, 2008; Skutnabb-Kangas, 2000). This phenomenon can be attributed to social and cultural pressures that prioritize the use of majority languages over minority languages, thereby marginalizing indigenous linguistic communities and undermining the value

Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 94–102

of their linguistic heritage (Hinton, 2011).

In this paper, we introduce the first parallel corpus for machine translation tasks for two indigenous languages that are spoken in Mexico and benchmark experimental results. The contributions of our work are the following:

- We introduce the first parallel corpus for machine translation for Mazatec and Mixtec languages.
- We evaluate the performance of the collected corpus and present benchmark results by using transformers, transfer learning, and fine-tuning approaches.
- We open-source the parallel corpus and the scripts used in this paper.

The rest of the paper is organized as follows: Section 2 describes previous research related to this study, Section 3 describes the properties of Mazatec and Mixtec languages, Section 4 describes the statistics of the collected dataset, Section 5 describes models used for baseline experiments and their results, and Section 6 describes the conclusion of the paper.

2 Related works

Due to an increase in the enormous amount of data for different languages, machine translation is currently one of the most researched areas in NLP and has shown promising results in high-resource languages (Tonja et al., 2022). There are different MT approaches that have been used by different researchers, neural machine translation (NMT)is one of the current state-of-the-art approaches trained on huge datasets containing sentences in a source language and their equivalent target language translations (Belay et al., 2022). Basically, NMT takes advantage of huge translation memories with hundreds of thousands or even millions of translation units (Forcada, 2017). However, NMT for low-resource languages still under-performs due to the scarcity of parallel datasets (Tonja et al., 2022, 2023b).

Many researchers explored different approaches to solving low-resource machine translation problems. Zoph et al. (2016) proposed a transfer learning method to improve the MT performance of low-resource languages. The authors first train a high-resource language pair (the parent model), then transfer some of the learned parameters to the low-resource pair (the child model) to initialize and constrain training. The data augmentation approach proposed by Fadaee et al. (2017), targets low-frequency words by generating new sentence pairs containing rare words in new, synthetically created contexts. Pourdamghani and Knight (2019) proposed using high-resource language resources to improve MT performance for low-resource languages without requiring any parallel data. Copying monolingual data of the target language is proposed by Currey et al. (2017) to improve the performance of low-resource MT. Tonja et al. (2023b) proposed the use of source-side monolingual data as another way of improving low-resource MT performance. Transfer learning method, where one first trains a "parent" model for a high-resource language pair and then continues training on a low-resource pair only by replacing the training corpus was proposed by Kocmi and Bojar (2018). Mixing low-resource language resources during training, as proposed by Tonja et al. (2022) showed an improvement in MT performance for low-resource languages.

There have been promising research works done for indigenous languages; Feldman and Coto-Solano (2020) presented an NMT model and a dataset for the Bribri Chibchan language for Bribri-Spanish translation. Kann et al. (2022) compiled AmericasNLI, a natural language inference dataset covering 10 indigenous languages of the Americas. They conducted experiments with pre-trained models, exploring zero-shot learning in combination with model adaptation. Oncevay (2021) proposed the first multilingual translation models for four languages spoken in Peru: Aymara, Ashaninka, Quechua, and Shipibo-Konibo, providing both many-to-Spanish and Spanish-to-many models, outperformed pairwise baselines. Zheng et al. (2021) presented a low-resource MT system that improves translation accuracy using cross-lingual language model pre-training. The authors used an mBART implementation of fairseq to pre-train on a large set of monolingual data from a diverse set of high-resource languages before fine-tuning on 10 low-resource indigenous American languages: Aymara, Bribri, Asháninka, Guaraní, Wixarika, Náhuatl, Hñähñu, Quechua, Shipibo-Konibo, and Rarámuri. On average, their proposed system achieved BLEU scores that were 1.64 higher and chrF scores that were 0.0749 higher than the baseline. Nagoudi et al. (2021) introduced

IndT5, the first Transformer language model for 10 Indigenous American languages: Aymara, Bribri, Asháninka, Guaraní, Wixarika, Náhuatl, Hñähñu, Quechua, Shipibo-Konibo, and Rarámuri. To train IndT5, they built IndCorpus–a new dataset for ten indigenous languages and Spanish.

3 Languages

3.1 Mazatec

The Mazatec language comprises a collection of closely related indigenous languages spoken primarily in the Northern region of Oaxaca, with smaller populations in the adjacent states of Puebla and Veracruz in Mexico. Approximately 200,000 individuals speak Mazatec; however, this number may fluctuate depending on which particular dialects or linguistic variations are taken into account (Léonard et al., 2019).

Mazatec belongs to the Oto-Manguean language family, a large family of indigenous Mesoamerican languages which also includes Mixtec, Zapotec, Otomi, among others (Vielma Hernández, 2017). Linguistic characteristics of Mazatec include tonal distinctions (Garellek and Keating, 2011), complex consonant clusters, and a rich morphology (Léonard et al., 2012). The Mazatec languages are known for their agglutinative structure, where words are formed by combining multiple morphemes, each with a distinct meaning (Vielma Hernández, 2017).

3.1.1 Writing system

Vowels - Mixtec has five basic vowels, similar to those in Spanish:

- a (as in "car"),
- e (as in "bet"),
- i (as in "bit"),
- o (as in "bore"),
- u (as in "boot").

These vowels can also appear nasalized, indicated by a tilde ($\tilde{a}$, $\tilde{e}$, $\tilde{i}$, $\tilde{o}$, $\tilde{u}$), and long, indicated by a colon ($a:, e:, i:, o:, u:$). Tones can be associated with vowels, too.

Consonants - The Mazatec consonant inventory includes the following sounds:

- Stops: p, t, k, b, d, g,
- Affricates: ts, tʃ, dz, dʒ,
- Fricatives: s, ʃ, h, z, ʒ,
- Nasals: m, n, ng,
- Approximants: w, j (pronounced as "y" in "yes"),
- Lateral approximant: l,
- Rhotics: r.

Numerals/Numbers - Mazatec uses a vesimal numeral system (base-20). Here are the numbers 1 to 10 in Mazatec: (1) - kia̲, (2) - chjɨ, (3) - tsi, (4)- sti, (5) - nka̲, (6) - tsjɨ, (7) - kja, (8) - chjɨn, (9) - tsi̲, (10) - sti̲.

Word order - Typically, Mazatec exhibits a VSO (Verb-Subject-Object) word order; however, alternative structures such as SVO can also occur depending on the sentence, the focus of the statement, and the context.

Example sentence:

Kitsaara kji xi makjíñeni kua apana (I gave a pill for the headache to my father) - VSO order

3.2 Mixtec

The Mixtec language comprises a group of closely related indigenous languages predominantly spoken in the region known as La Mixteca, which spans the states of Oaxaca, Puebla, and Guerrero in Southern Mexico. Estimates indicate that there are approximately 500,000 speakers of Mixtec; however, this number may fluctuate depending on the specific dialects or language varieties considered (Josserand, 1983).

As Mazatec, Mixtec is a member of the Oto-Manguean language family (Rensch, 1977; Pike and Cowan, 1961; Hollenbach, 2000) possessing the characteristic mentioned in Section 3.1. It also shares the phonemic system with Mixtec (see vowels and consonants inventory in Section 3.1.1) as well as the word order features and the base-20 number system. Here are numbers from 1 to 10 in Mixtec: (1)- in, (2) - ña'a, (3) - ta'a , (4) - na'a, (5) - ma'a, (6) - chiko, (7) - chikue, (8) - chikuiin, (9) - chikunña'a, (10) - ndo'o.

And here are a couple of examples sentences:

- Ka'nu ña'a nuu ntaa (Sitting on the plain) - VSO order
- Ña'a nuu ntaa ka'nu (On the plain, sitting) - SVO order

Note that the Mixtec language has many dialects, so the phonetic inventory, numerals, word order, and example sentences provided here may vary across different Mixtec-speaking communities. The examples given here are intended to provide a general overview of the language's features

4 Parallel Dataset

Data is one of the crucial building blocks of any NLP application (Belay et al., 2022; Tonja et al., 2023a), and a parallel corpus is essential to the success of any machine translation task. For Mazatec and Mixtec, we were unable to find publicly available datasets for the MT task. We collected datasets for these two indigenous Mexican languages from two main domains: *religious* and *constitution*. We also collected additional resources for the Mixtec language from different *textbooks* which have a similar translation to Spanish. Table 1 shows the statistics of the collected parallel corpus for Mazatec and Mixtec.

Text Alignment - We took a base directory path where text files were stored as input. Then we read and merged the content of all text files in the directory, and obtained a list of lists containing the content of each file. We proceeded to iterate through each file in the directory and read their contents line by line. Each line was normalized using the Unicode Normalization Form KC (NFKC) before being appended to the resulting list. We added a function that takes a language code `lang` as input, which determines the filename of the text file to be read from a predefined folder. The function read the file line by line, normalized each line using NFKC, and concatenated the lines into a single string. The result was returned as an array.

With another function, we added the two lists as input: one containing the content of the files to be aligned, and the other containing the filenames for the output files. We then iterated through the content list and aligned the text by iterating through the chapters and paragraphs of each translation. The aligned text was written to the corresponding output file as tab-separated values (TSV). Then we defined the root path where the input files were located, initialized the name and content arrays, and called the function that populated the content array with the pre-processed text. Finally, the function that writes the file was called to align and write the output files.

Pre-processing - After aligning the texts of two indigenous languages with their equivalent translations in Spanish, we pre-processed the corpus before splitting it for our experiments. The pre-processing steps included removing the numbers and special character symbols such as ;,",?, etc. For the baseline experiment, we split the pre-processed corpus into training, development, and test sets in the ratio of 70:10:20, respectively. Table 2 shows the split of the dataset used for our experiments.

5 Baseline Experiment and Discussion

In this section, we discuss the models used for the baseline experiment, the hyper-parameter used, the benchmark results, and the discussion. We used three approaches to evaluate the usability of the collected corpus. These are :-

- **Transformer** - is a type of neural network architecture first introduced in the paper *Attention Is All You Need* (Vaswani et al., 2017). The key innovation of the Transformer architecture is the attention mechanism, which allows the network to selectively focus on different parts of the input sequence when making predictions. This is in contrast to traditional recurrent neural networks (RNNs), which process input sequentially and are prone to the vanishing gradient problem.

  In the transformer architecture, the input sequence is processed in parallel by multiple layers of self-attention and feed-forward neural networks. Each layer can be thought of as a "block" that takes the output of the previous layer as input and applies its own set of transformations to it. The self-attention mechanism allows the network to weigh the importance of each element in the input sequence when making predictions, while the feed-forward networks help to capture non-linear relationships between the elements.

  Currently, transformers are state-of-the-art approaches and are widely used in NLP tasks such as MT, text summarization, sentiment analysis, etc. We used the base transformer configuration as described in (Vaswani et al., 2017) work.

- **Transfer learning**- refers to the process of leveraging pre-trained language models to improve the performance of downstream NLP tasks. Specifically, transfer learning involves

Source	Mazatec (maq) - Spanish (spa)			Mixtec (xtn) - Spanish (spa)		
	#sentences	#tokens (maq)	#tokens (spa)	#sentences	#tokens (xtn)	#tokens (spa)
Religion	8,203	269,753	187,773	8,208	278,874	183,050
Constitution	1,596	138,504	68,392	1,185	104,497	68,393
Others	-	-	-	3,842	71,628	70,080
Total	9,799	408,257	256,165	13,235	454,999	321,523

Table 1: Parallel dataset distribution of Mazatec-Spanish and Mixtec-Spanish

Language pairs	**Number of Sentences**		
	Train	**Dev**	**Test**
Mazatec - Spanish	7,056	784	1,959
Mixtec - Spanish	9,529	1,059	2,647

Table 2: Dataset split used in baseline experiments

using a pre-trained model to initialize the parameters of an MT system and then fine-tuning the system on a smaller dataset specific to the target language pair or domain.

Transfer learning can be especially useful in MT because training a high-quality MT system from scratch requires a large amount of data and computational resources, which may not be available for all language pairs or domains. By leveraging pre-trained models, transfer learning allows MT systems to achieve high performance with fewer data and fewer resources. For our baseline experiments, we used English-Spanish as parent model with two (**opus-mt-es-en**[1] and **opus-mt-tc-big-en-es** [2]) pre-trained models available from Hugging Face[3] trained for English-Spanish on the OPUS dataset (Tiedemann and Thottingal, 2020) by *Helsinki-NLP group*.

- **Fine tuning** - is the process of taking a pre-trained MT model and adapting it to a specific translation task, such as translating between a particular language pair or in a specific domain. The process of fine-tuning involves taking the pre-trained model, which has already learned representations of words and phrases from a large corpus of text, and training it on a smaller dataset of specific task examples. This involves updating the parameters of the pre-trained model to better capture the patterns and structures present in the target translation task.

Fine-tuning can be useful in MT because it allows the pre-trained model to quickly adapt to a new task without having to train a new model from scratch. This is especially beneficial when working with limited data or when there is a need to quickly adapt to changing translation requirements. We used two commonly known pre-trained multilingual MT models:

  - **M2M100-48** - is a multilingual encoder-decoder (seq-to-seq) model trained for many-to-many multilingual translation (Fan et al., 2020). We used a model with 48M parameters due to computing resource limitations.
  - **mBART50** - is a multilingual sequence-to-sequence model pre-trained using the *Multilingual Denoising pre-training* objective (Tang et al., 2020).

Hyper-parameters - For the transformer approach we tokenized the source and target parallel sentences into subword tokens using Byte Pair Encoding (BPE) (Gage, 1994). The BPE representation was chosen in order to remove vocabulary overlap during dataset combinations. For other approaches we applied the tokenizer of each model, Table 3 shows hyper-parameters used in our baseline experiments.

5.1 Results

Table 4 and Figure 1 shows the benchmark experimental results for bi-directional neural machine translation for Mazatec(maq) - Spanish(spa) and Mixtec(xtn) - Spanish(spa). In our baseline experiments, we observed that employing a transformer model for low-resource languages shows sub-optimal results compared to transfer learning and fine-tuning methodologies. As demonstrated in Table 4 and Figure 1, the performance of the **transformer** was inferior to alternative approaches utilized in the study. This finding substantiates the hypothesis that the efficacy of transformer models is heavily reliant on the availability of exten-

[1]https://huggingface.co/Helsinki-NLP/opus-mt-es-en

[2]https://huggingface.co/Helsinki-NLP/opus-mt-tc-big-en-es

[3]https://huggingface.co/

Approaches	Models	Parameters
Transformer	transformer	- enc_layers: 6 - dec_layers: 6 - heads: 8 - hidden_size: 512 - optimizer: adam - warmup_steps: 4000 - training_steps: 30000 - learning _rate: 5e-2
Transfer learning	opus-mt-es-en	- max_seq_length: 128 - num_train_epochs: 3 - per_device_batch_size: 4 - num_beams: 5
	opus-mt-tc-big-en-es	
Fine-tuning	mBART50	
	M2M100-48	

Table 3: Hyper-parameters used for baseline experiments

Models	xx-spa BLEU score		spa-xx BLEU score	
	maq-spa	xtn-spa	spa-maq	spa-xtn
M1	5.89	6.23	11.41	12.62
M2	6.91	10.47	14.49	13.73
M3	8.45	12.44	19.61	17.27
M4	10.45	15.66	21.2	16.93
M5	**12.09**	**16.75**	**22.5**	**22.15**

Table 4: Benchmark experimental result for bi-directional Mazatec(maq)-Spanish(spa) and Mixtec(xtn)-Spanish(spa) neural machine translation, M1, M2, M3, M4, and M5 represents transformer, opus-mt-es-en, opus-mt-tc-big-en-es,mBART50, and M2M100-48 models respectively.

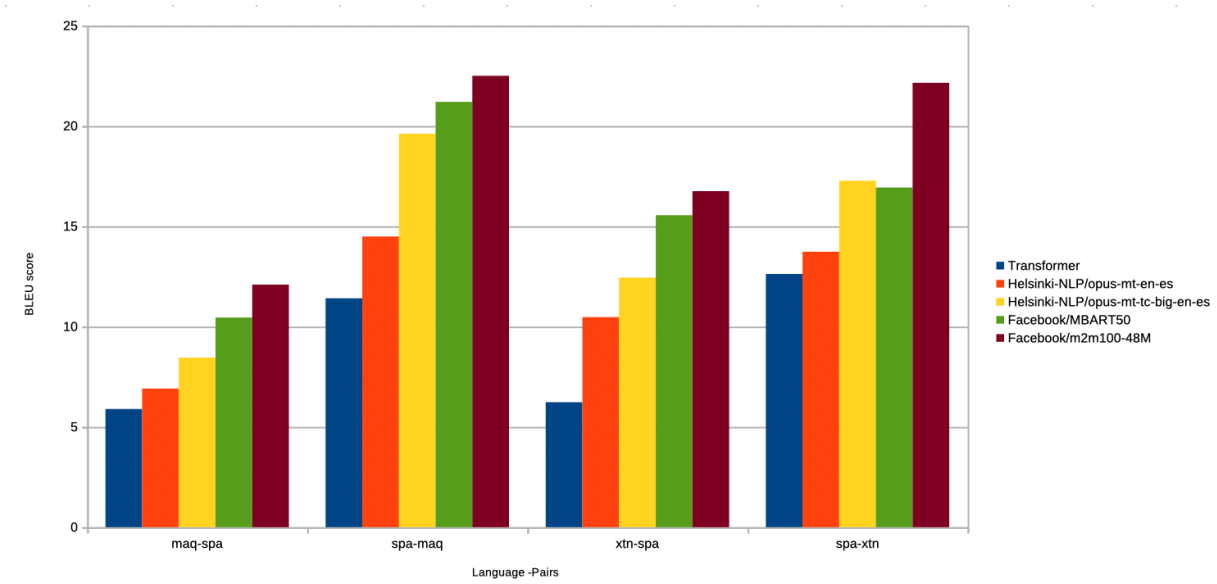

Figure 1: Benchmark results of selected approaches

sive parallel corpora for machine translation tasks. Upon further examination of language pair performance, we discovered that utilizing indigenous languages as the target language surpasses the performance achieved when using Spanish as the target language. This observation indicates that translating from Spanish to indigenous languages is a less complex task for the model as opposed to translating indigenous languages to Spanish.

Transfer learning approach showed more

promising results for the indigenous low-resource languages than the transformer approach. Out of the two models used in the transfer learning experiment, the model with *transformer-big* configuration outperformed the model with *transformer-base* configuration. This shows that the transfer learning approach depends on the size of the model parameter. Similarly, when using the transfer learning approach for indigenous low-resource languages by utilizing models trained on high-resource languages, better results were obtained when Spanish was used as the source language than when Spanish was used as the target language.

Fine-tuning approach outperformed the rest of the approaches used in our baseline experiment in both translation directions. This shows that using a multilingual pre-trained translation model for fine-tuning low-resource languages outperforms other models. From the two multilingual models used in the experiment, the **M2M100-48** model outperformed the **mBART50** multilingual model. The M2M100-48 model showed 4.7 and 5.5 BLEU scores on average for Mazatec (maq)-Spanish (spa) and Spanish (spa)-Mazatec (maq) translation. For Mixtec (xtn)-Spanish (spa) and Mixtec (xtn)-Spanish (spa), the M2M100-48 model showed a 10.2 and 7.5 BLEU score improvement on average when compared to the other models used in the experiments. When comparing the results of the two languages in all the approaches used, Mixtec (xtn)-Spanish (spa) translation showed better performance than Mazatec (maq)-Spanish (spa) translation when using Spanish as the target language, This shows that the availability of the parallel corpora for the language pairs has a high impact on the performance of the translation models. The overall results show that using multilingual MT models for fine-tuning in our selected indigenous low-resource languages gives promising results.

5.2 Discussion

In our analysis, we conducted an error analysis to identify the strengths and weaknesses of the three approaches: transformer, transfer learning, and fine-tuning. We found that the transformer approach, which relies on large parallel corpora, yielded sub-optimal results for low-resource languages. It struggled to capture the linguistic patterns and structures specific to indigenous languages. This limitation indicates that the transformer model's performance is highly dependent on the availability of extensive parallel corpora for effective machine translation.

On the other hand, the transfer learning approach showed more promising results for low-resource indigenous languages. We observed that models pre-trained on high-resource languages, such as Spanish, and fine-tuned on the indigenous languages improved translation quality. However, even with transfer learning, the performance was not satisfactory, and there were errors that persisted across all three approaches.

The general error that all three approaches failed to address adequately was the translation of domain-specific and culturally specific terms in Mazatec and Mixtec. These languages have unique vocabulary and cultural nuances that require a deeper understanding and context to ensure accurate translation. The limited availability of domain-specific parallel corpora for these languages hampered the models' ability to capture and translate such terms effectively.

6 Conclusion

In this paper, we presented a parallel corpus for two indigenous Mexican languages (Mazatec (maq) and Mixtec (xtn)) for machine translation tasks and evaluate the usability of the collected corpus using three different approaches. From the approaches, fine-tuning multilingual pre-trained MT models outperformed the rest of the experiments; Facebook's M2M100-48 outperformed all other models with BLEU scores of 12.09 and 22.25 for maq-spa and spa-maq, respectively, and 16.75 and 22.15 for xtn-spa and spa-xtn, respectively. We noticed from the experimental results that the dataset size has less impact when using indigenous languages as a target than the source. This observation highlights the potential benefits of focusing on developing and fine-tuning models specifically designed for translation tasks involving low-resource languages. Moreover, it underscores the value of creating and employing parallel corpora tailored to indigenous languages, as these resources can significantly improve machine translation performance, particularly when used in conjunction with advanced multilingual pre-trained models.

Our BLEU results for Mizatec and Miztec to Spanish translation were very low on the best configuration to have any usability in real-life applications, but the translation in the opposite direction demonstrated BLEU scores above 22 facilitating

uses, for example in government apps to present hints to Mixtec and Mazatec native speakers who have a low level of Spanish comprehension, in the government web pages. This could significantly improve the usefulness of the native language of the speakers, thus promoting communication of the language and its preservation.

In future research, we plan to investigate the efficacy of advanced techniques, including zero-shot and few-shot learning, for low-resource languages in the context of limited parallel datasets. These methodologies hold promise for effectively leveraging sparse data available in low-resource settings, as they capitalize on pre-existing knowledge from related tasks or languages without requiring extensive fine-tuning or additional annotated data. By exploring these approaches, we aim to uncover potential benefits and improvements in the machine translation performance of low-resource languages, thus contributing to developing more robust and accurate translation systems for underrepresented linguistic communities.

Acknowledgements

The work was done with partial support from the Mexican Government through the grant A1S-47854 of CONACYT, Mexico, grants 20220852, 20220859, and 20221627 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico, and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

References

Tadesse Destaw Belay, Atnafu Lambebo Tonja, Olga Kolesnikova, Seid Muhie Yimam, Abinew Ali Ayele, Silesh Bogale Haile, Grigori Sidorov, and Alexander Gelbukh. 2022. The effect of normalization for bi-directional amharic-english neural machine translation. In *2022 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pages 84–89. IEEE.

Anna Currey, Antonio Valerio Miceli-Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the second conference on machine translation*, pages 148–156.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. *arXiv preprint arXiv:1705.00440*.

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.

Isaac Feldman and Rolando Coto-Solano. 2020. Neural machine translation models with back-translation for the extremely low-resource indigenous language bribri. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3965–3976.

Mikel L Forcada. 2017. Making sense of neural machine translation. translation spaces, 6 (2), 291-309.

Philip Gage. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

Marc Garellek and Patricia Keating. 2011. The acoustic consequences of phonation and tone interactions in jalapa mazatec. *Journal of the International Phonetic Association*, 41(2):185–205.

Leanne Hinton. 2011. Language revitalization and language pedagogy: New teaching and learning strategies. *Language and Education*, 25(4):307–318.

Barbara E. Hollenbach. 2000. Mixtec - a new look at an old problem. *International Journal of American Linguistics*, 66(1):62–82.

Nancy H Hornberger. 2008. *Can schools save indigenous languages? Policy and practice on four continents*. Springer.

J. Kathryn Josserand. 1983. *Mixtec Dialectology: A Survey*. Tulane University.

Krishna Prakash Kalyanathaya, D Akila, and P Rajesh. 2019. Advances in natural language processing–a survey of current research trends, development tools and industry applications. *International Journal of Recent Technology and Engineering*, 7(5C):199–202.

Katharina Kann, Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, John E Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo A Giménez-Lugo, et al. 2022. Americasnli: Machine translation and natural language inference systems for indigenous languages of the americas. *Frontiers in Artificial Intelligence*, 5:266.

Dorothy Kenny. 2018. Machine translation. In *The Routledge handbook of translation and philosophy*, pages 428–445. Routledge.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1809.00357*.

Jean Léo Léonard, Vittorio Dell'Aquila, and Antonella Gaillard-Corvaglia. 2012. The almaz (atlas lingüístico mazateco): From geolinguistic data processing to typological traits. *STUF-Language Typology and Universals*, 65(1):78–94.

Jean Léo Léonard, Marco Patriarca, Els Heinsalu, Kiran Sharma, and Anirban Chakraborti. 2019. *Patterns of Linguistic Diffusion in Space and Time: The Case of Mazatec*, pages 139–170. Springer International Publishing, Cham.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza. 2018. Challenges of language technologies for the indigenous languages of the americas. *arXiv preprint arXiv:1806.04291*.

Manuel Mager, Alejandro Oncevay, Ali Ebrahimi, Juan Ortega, Alexander R Gonzales, Angela Fan, 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.

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 preprint arXiv:2104.07483*.

Arturo Oncevay. 2021. Peru is multilingual, its machine translation should be too? In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 194–201.

Kenneth L. Pike and Charles F. Cowan. 1961. *Language in Relation to a Unified Theory of the Structure of Human Behavior*. Mouton.

Nima Pourdamghani and Kevin Knight. 2019. Neighbors helping the poor: improving low-resource machine translation using related languages. *Machine Translation*, 33(3):239–258.

Calvin R. Rensch. 1977. *Oto-Manguean, Overview*. Summer Institute of Linguistics.

Tove Skutnabb-Kangas. 2000. *Linguistic genocide in education-or worldwide diversity and human rights?* Lawrence Erlbaum Associates Publishers.

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. *arXiv e-prints*, pages arXiv–2008.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Atnafu Lambebo Tonja, Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Moges Ahmed Mehamed, Olga Kolesnikova, and Seid Muhie Yimam. 2023a. Natural language processing in ethiopian languages: Current state, challenges, and opportunities. *arXiv preprint arXiv:2303.14406*.

Atnafu Lambebo Tonja, Olga Kolesnikova, Muhammad Arif, Alexander Gelbukh, and Grigori Sidorov. 2022. Improving neural machine translation for low resource languages using mixed training: The case of ethiopian languages. In *Advances in Computational Intelligence: 21st Mexican International Conference on Artificial Intelligence, MICAI 2022, Monterrey, Mexico, October 24–29, 2022, Proceedings, Part II*, pages 30–40. Springer.

Atnafu Lambebo Tonja, Olga Kolesnikova, Alexander Gelbukh, and Grigori Sidorov. 2023b. Low-resource neural machine translation improvement using source-side monolingual data. *Applied Sciences*, 13(2):1201.

Atnafu Lambebo Tonja, Michael Melese Woldeyohannis, and Mesay Gemeda Yigezu. 2021. A parallel corpora for bi-directional neural machine translation for low resourced ethiopian languages. In *2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pages 71–76. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jonathan Daniel Vielma Hernández. 2017. Panorama de los estudios lingüísticos sobre el mazateco. *Cuadernos de Lingüística de El Colegio de México*, 4(1):211–272.

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.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.