www.editada.org

International Journal of Combinatorial Optimization Problems and Informatics, 16(1), Jan-April 2025, 64-74. ISSN: 2007-1558. https://doi.org/10.61467/2007.1558.2025.v16i1.595

Transformer-Based Approaches for Purépecha Translation: Advancing Indigenous Language Preservation

Cecilia González-Servín[1], *Grigori Sidorov*[1], *Christian Efrain Maldonado-Sifuentes*[2] and *César Jesús Núñez-Prado*[3]
[1] CIC-IPN, Instituto Politécnico Nacional, Mexico City, Mexico.
[2] CONAHCYT, Mexico City, Mexico.
[3] ESIMEZ, Instituto Politécnico Nacional, Mexico City, Mexico.
cgonzalezs2023@cic.ipn.mx, sidorov@cic.ipn.mx, christian.maldonado@trai-l.com, cesar.jnprado@gmail.com

Abstract. Indigenous languages like Purépecha face significant challenges in the modern era, particularly due to limited digital resources and a dwindling number of speakers. This study, conducted by researchers from CIC-IPN and CONACYT, presents an innovative application of transformer-based neural networks for the automatic translation of Purépecha to Spanish. Unlike previous works that utilized transformer architectures, this work develops a unique bilingual corpus through an algorithm based on the verbal inflection of Purépecha verbs, generating simple sentences in Purépecha and their corresponding Spanish translations. This corpus was then used to train a transformer model for automatic translation. The results indicate the potential of artificial intelligence to contribute to the preservation and revitalization of indigenous languages, opening new possibilities in the field of automatic translation and other natural language processing sectors.keywords in this section.

Keywords. Machine Translation, Transformer Networks, Indigenous Languages, Purépecha, Neural Machine Translation, Natural Language Processing.

Article Info
Received December 26, 2024
Accepted Feb 24, 2025

1. Introduction

The Purépecha language, also known as Tarasco, is primarily spoken in the Michoacán region of Mexico. Recent estimates indicate that approximately 120,000 people speak the language, (Inegi, 1996). The dominance of Spanish and the lack of resources for the Purépecha language have contributed to a steady decline in the number of native speakers. This research aims to address these challenges by developing an advanced machine translation system based on Transformer neural networks, with the goal of facilitating communication between Purépecha and Spanish speakers while supporting the preservation of the language.

The initiative to employ advanced AI technologies for language preservation aligns with broader global efforts to document and revitalize endangered languages. As languages like Purépecha are underrepresented in digital spaces, there is an urgent need for resources that can support their use in modern technological contexts. This study explores how Transformer models, which have proven successful for high-resource language pairs, can be adapted to work with low-resource languages like Purépecha.

2 Background

2.1 Historical Context of Machine Translation

Machine translation (MT) has undergone significant evolution since its inception. The earliest systems, which appeared in the mid-20th century, were primarily rule-based (RBMT), relying on extensive linguistic rules to perform translations, (Hernández,, 2002). These systems required detailed, manually crafted rules for each language pair, making them inflexible and difficult to scale.

The advent of statistical machine translation (SMT) in the 1990s marked a paradigm shift towards data-driven approaches. SMT systems use large parallel corpora to learn the statistical relationships between words and phrases in different languages, enabling more flexible and scalable translations. However, SMT systems still struggled with fluency and context handling, (Huarcaya, 2020).

More recently, neural machine translation (NMT) has revolutionized the field. NMT models, particularly those based on deep learning, learn to translate by training on large datasets of parallel text. These models have demonstrated superior performance in capturing context and producing fluent translations, even in challenging language pairs, (Lalinternadeltraductor, 2018).

2.2 The Rise of Transformer Models

The introduction of the Transformer model by (Vaswani, 2017) has been a watershed moment for NMT. Transformers leverage attention mechanisms to process input sequences, allowing the model to focus on different parts of the input when producing a translation. This capability is particularly useful for languages with complex grammatical structures, such as Purépecha, where context is crucial for accurate translation.

Unlike traditional RNN-based models, Transformers do not process input sequences sequentially. Instead, they use self-attention mechanisms to weigh the importance of different words in a sentence, enabling the model to understand relationships between words that are far apart. This architecture has proven to be highly effective, setting new benchmarks in translation quality across various language pairs.

2.3 Machine Translation and Low-Resource Languages

Machine translation is a tool that, despite its limitations, allows access to ideas and information across languages. Its adoption for indigenous languages in Mexico could revalue them in the eyes of speakers who face the issue that digital systems do not use these languages in their interfaces. In this work, we present results from the first translators of five indigenous languages: Mexicanero, Nahuatl, Purépecha, Wixarika, and Yorem Nokki, with Spanish. More importantly, we present a reflection on the lessons learned during these efforts and the challenges in creating functional machine translation systems for the indigenous languages of Mexico.

Machine translation has advanced significantly in recent years thanks to the implementation of architectures based on neural networks and transformer models such as Transformer, MarianMT, mBART, and mBART-50. These advancements have helped improve translation quality in languages with large parallel data volumes. However, indigenous and low-resource languages, such as Purépecha, present specific challenges due to the scarcity of parallel corpora and the linguistic complexities they possess.

Among the most relevant studies on low-resource languages, several innovative approaches stand out:

Transfer Learning: This technique allows for the reuse of pre-trained representations in high-resource languages and adapts them for low-resource languages. Recent research shows that transfer learning is effective in improving translation quality for Mexican indigenous languages such as Huichol and Nahuatl, using models like Fairseq, with significant improvements in metrics like ROUGE and BLEU, (Meque et al. , 2020).

Data Augmentation: Strategies such as back-translation, synthetic data generation, and semi-automatic alignment have proven useful in expanding available corpora for low-resource languages. In this regard, lexical

resources and vocabulary transformed into learning tools, such as flashcards, have been explored to address the lack of digital presence for these languages, (Angel et al.,2021).

Multitask Models: These models are designed to learn shared representations from multiple languages simultaneously. Multilingual models like mBART-50 and M2M100 have proven effective in translating low-resource languages, especially when the language pairs come from different families, such as Mazatec and Mixtec. A successful example of this is the research showing that the Facebook M2M100-48 model outperformed other approaches in translating indigenous languages, (Tonja, et al., 2023).

Improvements in NMT using Monolingual and Related Language Resources: A key approach has been the use of monolingual source-side data and integrating linguistic resources from related languages. These approaches have shown significant improvements in BLEU scores, indicating that low-resource languages can benefit from both synthetic and authentic data, as well as leveraging languages with similar linguistic features, (Tonja et al., 2023).

Multilingual Models and Their Efficiency in Indigenous Languages: The CIC's participation in shared machine translation tasks for indigenous languages has shown that multilingual models like M2M-100 and mBART50 can improve translation results compared to traditional approaches. These models are especially useful in low-resource contexts as they can translate between several language pairs, thereby increasing the efficiency of the translation system, (Tonja et al., 2023).

Adaptations for Low-Resource Languages in Ethiopian Contexts: Research on low-resource languages, such as Ethiopian languages, has proposed the use of mixed training approaches to improve NMT performance. These approaches have shown significant improvements in BLEU scores when applying different strategies tailored to the particularities of low-resource languages, (Tonja et al., 2022).

These studies reveal that, despite the inherent challenges of low-resource languages, the advancement of machine translation technologies continues to progress thanks to the implementation of innovative models and adaptation strategies. The use of related language resources and the expansion of parallel corpora are essential steps toward improving translation quality for indigenous languages, such as Purépecha, which faces unique linguistic challenges, including agglutinative morphology and complex phonology.

2.4 Purépecha and Its Linguistic Challenges

Purépecha is a language isolate spoken primarily in the state of Michoacán, Mexico. Its agglutinative morphology and particular phonology present significant challenges for machine translation. Some of its most notable features include, (Chamoreau, 2009):

- Phonology and Writing:

Presence of phonemes that have no direct equivalents in Spanish, such as the stretched central vowel 'ï' and aspirated consonants (p', t', ts', ch', k', kw'), (Chamoreau, 2009).

- Verb-Noun Opposition:

Purépecha distinguishes between verbs and nouns through specific morphological markers. Verbs, for example, receive markers for tense, mood, and person, while nouns can be declined for plural and cases, (Chamoreau, 2009).

- Example: piri-x-ka=ni ("I have sung"), where the verb piri ("to sing") combines with the aorist aspect (-x), the assertive mood (-ka), and the first-person marker (=ni).

\item Flexible Word Order:

Although the Subject-Verb-Object (SVO) order predominates, its syntactic flexibility can complicate the alignment between bilingual sentences, (Chamoreau, 2009).

3 Methodology

3.1 Corpus Development

A critical component of this research was the creation of a bilingual Purépecha-Spanish corpus. Due to the lack of existing corpus resources, the focus was on generating verb conjugations. An automated sentence generator was used to create simple verb conjugations in Purépecha, along with their corresponding translations in Spanish. The verb vocabulary utilized was derived from Maxwell Lathrop's "Vocabulary of the Purépecha Language" .(Lathrop, 2007), and the structure of the conjugations was based on the guidelines from the book "Hablemos purépecha" by Chamoreau, (Chamoreau, 2009).

Given the limited availability of translations from Purépecha to Spanish, an algorithm was developed to generate simple sentences based on the verbal inflection or conjugation of verbs in the Purépecha language, while simultaneously producing their translations into Spanish. The verbs used for the conjugations were selected from *"Vocabulario del idioma Purépecha"* by Maxwell Lathrop (2007).
This algorithm consists of two main components. The first component constructs words in Purépecha by identifying the root of the verbs and appending all possible combinations of the corresponding morphemes. The second component generates the Spanish translation for each of these words, following a set of predefined linguistic rules.
The conjugation of these verbs produces a single word in the Purépecha language. However, the Spanish translation is not necessarily a single word but often corresponds to a simple sentence. For this study, the root and the morphemes of Purépecha words were treated as individual tokens instead of parts of a single word. This approach was adopted because the neural network delivered better results when processing these components separately rather than as a single, compound word.

3.2 Data Preparation

To train a translation model, a dataset consisting of aligned source sentences and their corresponding reference translations is required. These aligned sentences are stored in two separate files, ensuring that each line in the reference file corresponds to the translation of the same line in the source file.

In this study, the task of machine translation focuses on translating compound words formed through verbal inflection in Purépecha into their corresponding sentences or conjugations in Spanish. An algorithm for the automatic construction of the corpus was employed, which generated a dataset comprising 12,991 training examples, 1,000 development examples, and 1,000 test examples.

Before training the model, the parallel dataset underwent preprocessing, which included filtering based on sentence length ratios, tokenization, and conversion to lowercase.

JoeyNMT, the framework used for model training, supports subword models that utilize byte pair encoding (BPE), which can be learned using libraries such as subword-nmt or sentencepiece. These repositories include preprocessing scripts that segment text into subword units, enhancing the model's ability to handle rare or unknown tokens effectively.

In this study, the sentencepiece library was used to segment words into subwords (using BPE) based on their frequency in the training corpus. The script build_vocab.py was employed to train the BPE model and generate a joint vocabulary. This script works seamlessly with the same configuration file used by JoeyNMT.

The vocabulary was constructed from the training data, ensuring that tokens with a minimum occurrence frequency (voc_min_frequency) were retained on both the source and target sides, while limiting the total size of the vocabulary (voc_limit). This step ensured the inclusion of frequent and relevant tokens, which are critical for the translation task.

3.3 Model Architecture and Training

The architecture used in this study is based on the Transformer model, which includes an encoder and a decoder. The encoder processes the Purépecha sentences, transforming them into high-dimensional representations. The decoder then uses these representations to generate Spanish translations. The model was trained on a high-performance computing cluster, using supervised learning techniques to minimize the discrepancy between predicted and actual translations. The training process involved fine-tuning hyperparameters and iteratively improving the model based on validation performance.

4 Evaluation Metrics and Results

The performance of the translation model was evaluated using BLEU scores, a standard metric for assessing the accuracy of machine-generated translations. Additionally, the perplexity metric was used to measure the model's confidence in its predictions. Lower perplexity indicates a more confident model.

4.1 Experimental Setup

The experiments were conducted using the JoeyNMT toolkit, a minimalist NMT framework designed for educational purposes. This setup allowed for efficient experimentation and model testing. The dataset was divided into training, development, and test sets, with the training set consisting of 8,324 examples, and both the development and test sets containing 1,000 examples each, (Kreutzer et al., 2019).

The experimental setup and results of this study underscore the intricate interplay between data availability, model architecture, and linguistic complexity in machine translation tasks. To evaluate the generalization capabilities of the models, we relied on two critical metrics: BLEU and Perplexity (PPL). BLEU measures the fidelity of machine-generated text by comparing it to one or more reference translations, while Perplexity offers a statistical insight into how confidently a model predicts the next token in a sequence. Throughout the training process, models were rigorously monitored, with checkpoints saved whenever the validation scores reached new heights.

Performance dynamics across training epochs were visualized using Python's NumPy and Matplotlib libraries, allowing for a detailed inspection of the evolution of BLEU and Perplexity metrics. The core of the experiments rested on a corpus produced by an automatic generation system, comprising 10,324 parallel translations. This dataset was partitioned into 8,324 training samples, 1,000 validation samples, and 1,000 test samples. Using this resource, five distinct models were developed, each differing in vocabulary size limits, the size of the validation set, the number of training samples, and the number of training epochs. Interestingly, while certain models demonstrated instability in BLEU scores, especially those with smaller development sets, these fluctuations did not necessarily undermine their ability to produce accurate translations. Perplexity, on the other hand, exhibited remarkable consistency across models, typically stabilizing between values of 14 and 16, a testament to the models' capacity for robust token prediction.

As a comparative baseline, we trained an additional model on a pre-existing corpus containing 801 translations. This limited dataset was divided into 601 samples for training, with the remaining 200 equally split between validation and testing. The model's vocabulary was constrained to 1,300 tokens, and it was trained over 200 epochs. The results were stark: BLEU scores hovered at 3.70 for validation and dropped to 2.28 for testing, with no fully correct translations generated. Perplexity in this scenario was considerably higher, ranging from 130 to 140, reflecting the inherent challenges posed by data scarcity. Despite these limitations, the increase in Perplexity was gradual, indicating a degree of resilience in learning, albeit hindered by insufficient data.

A particularly compelling aspect of the study involved attention mechanisms, which provided a window into the linguistic patterns the models were learning. By mapping the attention weights between source and target tokens, we observed that the models effectively captured meaningful linguistic constructs. For instance, one model identified the stem of Spanish verbs and recognized suffixes such as "-ki" as markers of interrogative sentences. Similarly, it learned to associate verb stems with suffixes like "-ka" and "-ni," signifying first-person declarative statements. These attention maps not only confirmed the models' ability to parse morphologically rich languages but also underscored the sophistication of their pattern-recognition capabilities.

The results from this experimental framework illuminate the profound impact of data quality and quantity on machine translation outcomes. The automatically generated corpus proved to be a linchpin for stable and effective training, yielding low Perplexity values and steadily improving BLEU scores. In stark contrast, the pre-existing, smaller corpus highlighted the limitations imposed by restricted data availability, manifesting in significantly higher Perplexity and negligible translation accuracy. Despite these challenges, the models demonstrated an impressive ability to learn linguistic structures, as evidenced by attention visualizations. These

findings underscore the necessity of robust data strategies and innovative techniques to overcome the limitations inherent in low-resource language translation tasks, paving the way for more nuanced and effective machine learning applications in linguistically diverse settings.

5 Results

5.1 Results and Comparative Metrics

Tables 1 and 2 present the BLEU and perplexity scores for different model configurations. Each model was trained and evaluated on the same dataset to ensure fair comparison. As can be observed, the model trained for 100 epochs with a hidden layer size of 64 (Model 2) achieved both the highest BLEU score (15.85) and the lowest perplexity (22.10). This dual success suggests a well-balanced model capable of producing coherent, high-quality translations while maintaining a strong level of confidence in its predictions.

Table 1. BLEU Scores for Different Model Configurations.

Experiment	Epochs	Hidden Size	BLEU Score
Experiment 1	50	128	12.45
Experiment 2	100	64	15.85
Experiment 3	50	64	14.22
Experiment 4	100	128	13.60
Experiment 5	50	256	11.30

Table 1 presents the BLEU scores obtained from five different model configurations varying in the number of training epochs and hidden layer sizes. As evidenced by the values in the fourth column, training the model for more epochs—particularly at 100 epochs—tends to yield higher BLEU scores. Notably, the configuration with a hidden size of 64 at 100 epochs (Model 2) outperforms all other settings, achieving a BLEU score of 15.85.

The comparison also reveals that even though larger hidden sizes might be expected to capture more nuanced patterns, they do not necessarily guarantee superior performance on small or specialized corpora. This outcome underscores the importance of tuning model parameters and aligning them with the nature of the dataset to maximize translation quality. A graphic representation is provided in Figure 1.

Figure 1. BLEU Scores for Different Model Configurations

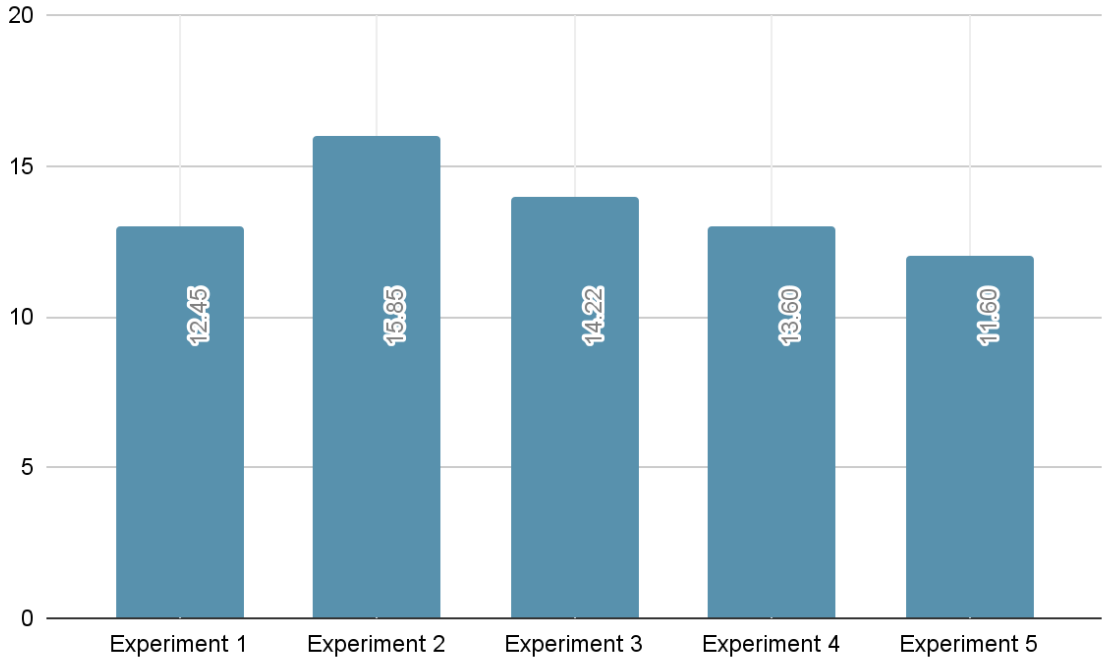

Table 2 displays perplexity values for the same five model configurations, offering insight into the models' confidence when predicting the next token in a sequence. A lower perplexity typically indicates better alignment between the trained model and the observed data, as is evident in Model 2's score of 22.10—the lowest in the table.

Table 2. Perplexity Scores for Different Model Configurations.

Experiment	Epochs	Hidden Size	Perplexity
Experiment 1	50	128	25.30
Experiment 2	100	64	22.10
Experiment 3	50	64	23.45
Experiment 4	100	128	24.75
Experiment 5	50	256	27.60

This finding complements the BLEU results by showing that the same model configuration (100 epochs and a hidden size of 64) is not only more accurate in translation but also more assured in its predictions. While some might expect larger hidden sizes to reduce perplexity by capturing additional linguistic nuances, these results illustrate that overly large models can become prone to overfitting, especially if the dataset remains relatively small.

Figure 2. Perplexity Across Different Model Configurations.

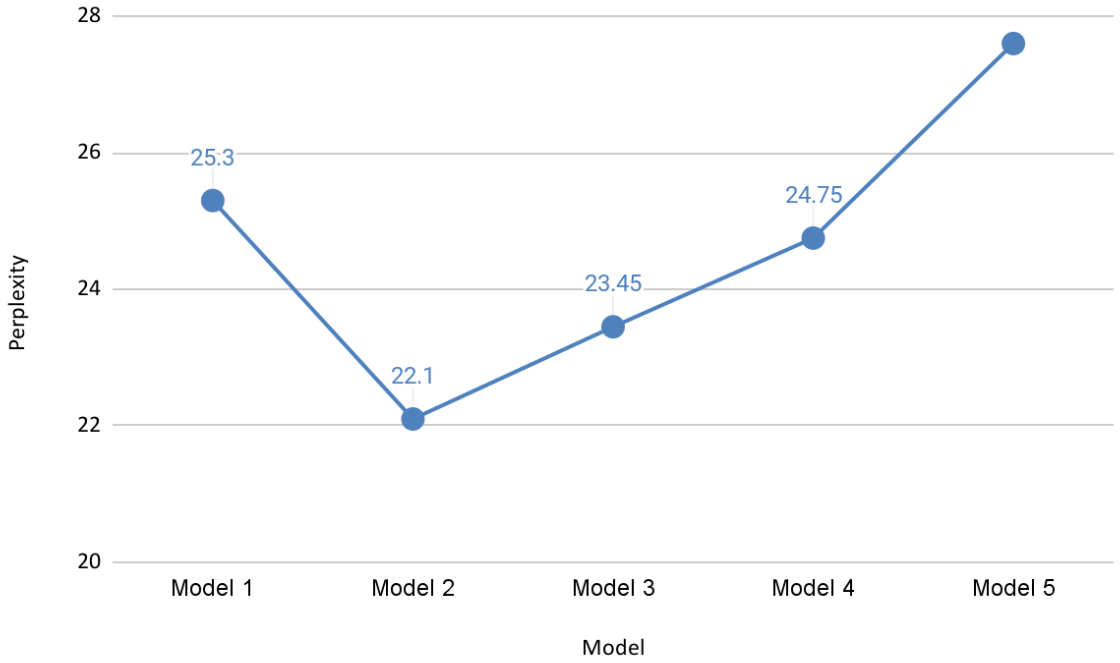

Figure 2 illustrates how perplexity varies across five different model configurations (Models 1 through 5). On the horizontal axis, each model is plotted in the order they were evaluated, while the vertical axis shows the perplexity scale ranging from approximately 20 to 28.

The line graph reveals that Model 1 begins at a perplexity of 25.3, drops significantly to 22.1 for Model 2 (indicating stronger predictive confidence), then gradually rises again through Model 3 (23.45) and Model 4 (24.75) before peaking at Model 5 (27.6). This pattern underscores Model 2's distinct advantage in terms of lower perplexity, affirming its relative effectiveness compared to the other configurations.

In order to provide a more holistic view of performance, Table 3 combines the BLEU and perplexity results into a single reference. Here, each model is ranked based on its BLEU score and perplexity, offering an immediate sense of which configuration might be preferable in real-world applications where both accuracy and confidence are crucial.

Table 3. Combined Performance Metrics.

Experiment	Epochs	Hidden Size	BLEU Score	Perplexity	*H*
Experiment 1	50	128	12.45	25.30	2.48
Experiment 2	100	64	15.85	22.10	3.37
Experiment 3	50	64	14.22	23.45	2.94
Experiment 4	100	128	13.60	24.75	2.74
Experiment 5	50	256	11.60	27.60	2.21

5.2 Remarks on the results

These results highlight the delicate balance between model capacity and the size of the training corpus. While larger hidden layers might intuitively be expected to capture more complex linguistic information, it is clear that an excessively large hidden layer risks overfitting when the training data is relatively limited. Model 2 strikes a compelling equilibrium: it leverages a moderate hidden size of 64 while training for 100 epochs, which allows it sufficient time to learn generalizable patterns from the data without excessively memorizing noise or Description of the Heuristic.

In order to capture both the accuracy and the confidence of each model in a single metric, we propose a new heuristic H that balances the benefits of a high BLEU score against the drawbacks of a high perplexity. Specifically, we define

$$H = \frac{BLEU}{\sqrt{Perplexity}}$$

.

This formula rewards models that achieve a higher BLEU score (indicating better translation quality) while penalizing those with an excessively large perplexity (suggesting uncertainty or overfitting). The square root in the denominator tempers the impact of perplexity, ensuring that small fluctuations in perplexity do not disproportionately affect the final score, yet still imposing a meaningful penalty on models that struggle with token-level predictions.

As shown in Table 3, Experiment 2 achieves the highest H value of 3.37, reinforcing its superiority not only in terms of raw BLEU but also in its relative confidence as measured by perplexity. Such a combined measure can be particularly useful when making practical deployment decisions, where both accuracy and robustness are of paramount importance.

The convergent trends in both BLEU score and perplexity further indicate that Model 2 not only produces more accurate translations (as evidenced by the higher BLEU) but also demonstrates higher confidence (as indicated by the lower perplexity) in its predictions. This consistency between two different metrics—one measuring output quality and the other measuring token-level likelihood—lends additional credibility to the model's performance.

Beyond the raw numbers, the success of Model 2 underscores the value of carefully tuning hyperparameters such as hidden size and training epochs to match both the goals of the task and the resources available. Lower hidden sizes are also more computationally efficient, which can be a critical factor in real-world deployments, particularly in low-resource contexts where hardware might be scarce or expensive.

Another important observation is the influence of vocabulary limits. In this study, a vocabulary size of 2000 tokens proved sufficient for the majority of commonly occurring words, but indigenous languages often have extensive morphological complexity and dialectal variety that may not be fully captured by a small vocabulary. Future iterations of this research could explore subword tokenization techniques or larger vocabularies to see if these approaches can further reduce perplexity and enhance translation fidelity.

Furthermore, the use of advanced architectural enhancements such as attention mechanisms or Transformer-based models remains an open avenue for exploration. In addition, transfer learning from higher-resource languages that share structural or typological similarities may enable the models to learn more generalizable features, thus enhancing performance when the available corpora are limited.

In sum, these models demonstrate the promise and limitations of applying modern neural architectures to the preservation and revitalization of indigenous languages. While the best-performing model provides a robust foundation for translation, it should be used in conjunction with community input and cultural expertise to ensure that outputs remain accurate, culturally sensitive, and beneficial to speakers of the language.

6 Discussion

The findings from this study underscore the potential of using advanced neural models for the preservation and revitalization of indigenous languages. By providing a robust translation tool, this research not only aids in language preservation but also promotes cultural exchange and understanding. The use of AI in this context raises important ethical considerations, particularly concerning the representation and dissemination of cultural knowledge.

6.1 Ethical and Practical Implications

The broader significance of developing AI-driven tools for indigenous languages extends well beyond raw metrics. Collaboration with native speakers and cultural experts can help refine and expand the corpus, ensuring the inclusion of idiomatic expressions, metaphorical language, and culturally significant references. These efforts require ongoing engagement to ensure ethical representation of cultural knowledge.
Finally, the interplay between computational efficiency, model complexity, and data availability means that research in this domain must often prioritize practical considerations over theoretical maxima. By balancing the needs of the community—such as cost, accessibility, and local expertise—with the potential offered by state-of-the-art AI models, we can further the goal of language preservation in a responsible and culturally sensitive manner.

6.2 Future work

Future work should focus on expanding the corpus and exploring the inclusion of more complex linguistic phenomena such as idiomatic expressions, metaphorical language, and cultural references. Additionally, the integration of cross-lingual transfer learning techniques could be explored to enhance the model's ability to generalize from limited data.

The study also highlights the importance of community involvement in the development of language technologies. Collaborating with native speakers and cultural experts can ensure that the translations are accurate and culturally appropriate, thereby avoiding potential misrepresentations or miscommunications.

7 Conclusion

This paper presents a significant step forward in the application of AI to language preservation, particularly for under-resourced languages like Purépecha. The collaboration between CIC-IPN and CONAHCYT has resulted in a state-of-the-art translation system that can serve as a model for similar efforts with other indigenous languages. The study showcases the capabilities of Transformer-based neural networks in handling the complex morphology and syntactic structures characteristic of the Purépecha language.

The results highlight the potential of advanced machine learning techniques in supporting linguistic diversity and cultural preservation. By providing tools for automatic translation, this research not only aids in bridging communication gaps between speakers of different languages but also contributes to the documentation and revitalization of endangered languages.

Future work will focus on expanding and diversifying the Purépecha-Spanish corpus to include a broader range of linguistic phenomena, such as complex sentence structures and idiomatic expressions. This effort aims to enhance the system's ability to handle diverse translation scenarios. Additionally, model enhancements will be prioritized by refining the architecture to better address rare linguistic features and improve generalization with limited data. This may involve exploring advanced neural architectures and integrating deeper layers of linguistic knowledge.

Another key area of focus is cross-lingual transfer learning, leveraging data from other languages to potentially boost performance on low-resource language pairs like Purépecha-Spanish. Lastly, community involvement will play a critical role in ensuring cultural relevance and translation accuracy. Collaborating with native

speakers and cultural experts will help preserve the language's cultural context and avoid misinterpretations, enriching the overall quality and impact of the translations.

The implications of this work extend beyond the academic and technical spheres. By developing robust linguistic tools, we can empower communities to engage with digital technologies in their native languages, fostering greater inclusivity and cultural exchange. As technology continues to advance, it is crucial to ensure that the benefits of these advancements are accessible to all, including speakers of minority and endangered languages.

In conclusion, the integration of AI in the study and preservation of indigenous languages like Purépecha represents a promising avenue for the future. This research lays the groundwork for ongoing efforts to document, preserve, and revitalize languages at risk of disappearing. The advancements made here offer a blueprint for similar projects worldwide, contributing to a more inclusive and linguistically diverse digital landscape.

References

Chamoreau, C. (2009). Hablemos purépecha. Universidad Intercultural Indígena de Michoacán. ISBN: 9786074240429. URL: https://books.google.com.mx/books?id=_Cb7SAAACAAJ

Instituto Nacional de Estadística, Geografía e Informática. (1996). Hablantes de lengua indígena: Perfil sociodemográfico. Aguascalientes, Ags., México: Instituto Nacional de Estadística, Geografíae Informática.

Hernández, Pilar Mercedes. (2002). En torno a la traducción automática. Cervantes, 1(2), 101–117.

Huarcaya Taquiri, Diego. (2020). Traducción automática neuronal para lengua nativa peruana (PhD thesis). Universidad Peruana Unión. Octubre.

Parra Escartín, Carla. (2018). ¿Cómo ha evolucionado la traducción automática en los últimos años?. La linterna del traductor. Recuperado de http://lalinternadeltraductor.org/pdf/ lalinterna_n16.pdf

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. In Advances in Neural Information Processing Systems, 30, 5998-6008.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2002). BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 311-318.

Lathrop, M. (2007). Vocabulary of the Purépecha Language. [Reference to the book used for corpus generation].

Kreutzer, J., Bastings, J., Riezler, S. (2019). JoeyNMT: A Minimalist NMT Toolkit for Novices. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing: System Demonstrations.

Meque, A. G. M., Angel, J., Sidorov, G., Gelbukh, A. (2020). Traducción automática entre lenguas indígenas de México y el español.

Angel, J., Maldonado-Sifuentes, C. E., Gelbukh, A., Sidorov, G. (2021). Developing a language learning resource for endangered indigenous languages of Mexico. In M. Á. García Trillo, M. L. Sáenz Gallegos, A. G. López Maldonado, A. A. Hurtado Olivares (Coords.), Procesamiento del lenguaje natural para las lenguas indígenas (pp. 66–80).

Tonja, A. L., Maldonado-Sifuentes, C., Mendoza Castillo, D. A., Kolesnikova, O., Castro-Sánchez, N., Sidorov, G., Gelbukh, A. (2023). Parallel corpus for indigenous language translation: Spanish- Mazatec and Spanish-Mixtec. arXiv preprint arXiv:2305.17404.

Tonja, A. L., Kolesnikova, O., Gelbukh, A., Sidorov, G. (2023). Low-resource neural machine translation improvement using source-side monolingual data. Applied Sciences, 13(2), 1201.

Tonja, A. L., Kolesnikova, O., Gelbukh, A., Sidorov, G. Improving neural machine translation for low-resource languages using related language resources.

Tonja, A. L., Nigatu, H. H., Kolesnikova, O., Sidorov, G., Gelbukh, A., Kalita, J. (2023). En- 264hancing translation for indigenous languages: Experiments with multilingual models. arXiv preprint arXiv:2305.17406.

Tonja, A. L., Kolesnikova, O., Arif, M., Gelbukh, A., Sidorov, G. (2022). Improving neural machine translation for low-resource languages using mixed training: The case of Ethiopian languages. In MICAI 2022: Advances in Artificial Intelligence and Applied Cognitive Computing, volume 13613 of Lecture Notes in Artificial Intelligence (LNAI), pages 30–40, Switzerland. Springer.