ISSN 1870-4069

Traducción automática entre lenguas indígenas de México y el español

Abdul Gafar Manuel Meque, Jason Angel,
Grigori Sidorov, Alexander Gelbukh

Instituto Politécnico Nacional,
Centro de Investigación en Computación,
México

{gafar_meque, sidorov}@cic.ipn.mx,
ajason08@gmail.com, gelbukh@gelbukh.com

Resumen. En los últimos años, hemos presenciado mejoras significativas en la precisión y velocidad de los sistemas de procesamiento de lenguaje natural. Particularmente, los métodos de traducción automática han abierto la posibilidad de conseguir mejores traducciones para lenguas de escasos recursos, tales como las lenguas indígenas de México. En este estudio usamos el modelo encoder-decoder Fairseq para evaluar la traducción automática desde el español de lenguas indígenas de México, incluyendo: el Huichol, el Mixteco, el Mazateco, el Mazahua, el Náhuatl de Guerrero y el Náhuatl de Puebla. Utilizando ROUGE y BLEU score como métricas de desempeño, nuestros resultados superan a trabajos anteriores para estas lenguas. Nuestras contribuciones incluyen la propuesta de un fuerte baseline para la evaluación de la traducción automática y la publicación de libre acceso del código y el dataset empleado.

Palabras clave: Traducción automática, lenguas indígenas, Náhuatl, Mazateco, Mixteco, Huichol, Mazahua.

Automatic Translation between Indigenous Languages of Mexico and Spanish

Abstract. In recent years, we have witnessed significant improvements in the accuracy and speed of natural language processing systems. In particular, automatic translation methods have opened the possibility of achieving better translations for low-resource languages, such as the indigenous languages of Mexico. In this study, we use the Fairseq encoder-decoder model to evaluate automatic translation from Spanish into Mexican indigenous languages, including: Huichol, Mixteco, Mazateco, Mazahua, Guerrero's Náhuatl, and Puebla's Náhuatl. Using ROUGE and BLEU score as performance metrics, our results outperform previous work for these languages. Our contributions include proposing a strong baseline for automatic translation evaluation and the open-source publication of the code and dataset used.

Keywords: Automatic translation, indigenous languages, Náhuatl, Mazateco, Mixteco, Huichol, Mazahua.

1. Introducción

La traducción automática de lenguas ha sido un tema de investigación y desarrollo en el área de Procesamiento del Lenguaje Natural durante varias décadas. Sin embargo, a pesar de las mejoras significativas en la calidad y velocidad de los sistemas de traducción automática, todavía es un desafío lograr traducciones precisas y fiables para lenguas de escasos recursos, es decir aquellas con cantidades muy limitadas de recursos digitales, como corpora, léxicos, pues al carecer de suficientes datos de entrenamiento los sistemas de traducción automática no pueden aprender con precisión los patrones linguisticos que deben utilizarse para leer o escribir en dichas lenguas.

Más aún, las lenguas de escasos recursos presentan retos significativos desde el punto de vista lingüístico para los modelos de procesamiento de lenguaje natural más efectivos de la actualidad, entre estos se evidencian la carencia de reglas ortográficas consistentes, amplias variaciones dialectales, mezcla de dialectos, neologismos en español y falta de consenso con respecto a los estándares ortográficos.

En este paper, revisamos los últimos avances en la traducción automática para lenguas de escasos recursos, y lo ejemplificamos usando 6 lenguas indígenas de México como lenguas objetivo para traducir textos de La Biblia desde el español. Nuestras contribuciones incluyen la propuesta de un fuerte baseline para la evaluación de la traducción automática y la publicación de libre acceso del código y el dataset empleado[1].

2. Antecedentes

La competencia de AmericasNLP 2021 [5] sobre traducción automática se centró en la traducción de lenguas indígenas habladas en el continente américano. El reto tuvo como objetivo promover la investigación en el área de traducción automática para idiomas de bajos recursos, particularmente aquellos con desafíos únicos como variaciones ortográficas, diferencias dialectales y falta de recursos escritos.

Durante la competencia se incluyeron diez lenguas indígenas alineadas con el español (ellas fueron: wixarika, Náhuatl, guaraní, bribri, rarámuri, aymara, shipibo-konibo, quechua, asháninka y Otomí) y se solicitó que participantes presentaran sistemas de traducción en ambas direcciones (esto es, español a lengua indígena y lengua indígena a español).

En AmericasNLP 2021 el modelo de referencia fue un modelo de secuencia a secuencia (sequence-to-sequence) implementado con Fairseq [6]. Los equipos participantes utilizaron varios enfoques, incluido el entrenamiento previo con datos monolingües, la incorporación de información fonética y el uso de modelos a nivel de caracteres.

Y aunque para la mayoría de los idiomas, muchos modelos pudieron mejorar considerablemente la línea de base, se hizo evidente la gran brecha que existe para traducir estas lenguas pues los sistemas con mejor desempeño lograron puntajes relativamente bajos en las métricas BLEU [7] y ChrF [8] respecto a los puntajes que obtienen las lenguas de con mayores recursos.

[1] huggingface.co/mekjr1

Tabla 1. Lenguas indígenas empleadas en esta investigación.

Lengua indígena	ISO 639-3	Hablantes	Estados de México
Mazahua	maz	120000	México y Michoacán
Huichol	hch	45000	Nayarit, Jalisco, Durango, y Zacatecas
Mazateco	maq	145000	Oaxaca
Mixteco	mim	490000	Guerrero
Náhuatl (de Puebla)	azz	170000	Puebla
Náhuatl (de Guerrero)	ngu	1500000	Guerrero, Puebla y Veracruz

De esta competencia se destaca que el ganador [11] logró los mejores resultados al combinar datos procedentes de La Biblia, Wikipedia y fuentes menores como constituciones politicas. Por otro lado, en cuanto a creación de corpus paralelos para lenguas indígenas, los autores en [3] describen el proyecto de creación de un corpus paralelo de español y Náhuatl junto con su interfaz de búsqueda.

El corpus se compiló a partir de libros no digitales, que presentaron varios desafíos durante el proceso de digitalización y alineación. El corpus paralelo comprende textos de diferentes fuentes que incluyen variaciones en dialecto, ortografía y cronología.

Su artículo enfatiza la escasez de recursos digitales para idiomas de bajos recursos como el Náhuatl (uno de los idiomas presentados en el trabajo actual) y cómo este corpus paralelo puede ser útil para los estudios lingüísticos y el desarrollo de tecnologías lingüísticas.

El documento proporciona ejemplos de cómo los corpus paralelos son valiosos para la traducción automática, la recuperación de textos multilingües y los estudios contrastivos y de traducción. Los autores discuten las diferencias entre las lenguas española y Náhuatl en términos de morfología, sintaxis y ortografía.

El documento está organizado en secciones que describen el proceso de compilación de los documentos paralelos, la interfaz de búsqueda, sus aplicaciones. Un trabajo similar al nuestro es [4] en el cual los autores crean un corpus paralelo entre el inglés y 5 lenguas africanas, utilizando distintos modelos neurales y el BLEU score como métrica de evaluación.

3. Fuentes de datos

Lo recursos lingüísticos utilizados en esta investigación son seis lenguas indígenas de México recuperadas de La Biblia [1], la cual es una fuente de datos bastante conocida al trabajar con lenguas de escasos recursos debido a que su contenido a sido traducido a una gran variedad de lenguas por motivos religiosos; ademas, resulta especialmente conveniente para la creación de modelos de traducción automática, pues la Biblia al estar estructurada en capítulos y versiculos permite una apropiada alineación entre los textos escritos en distintas lenguas.

A continuación se listan las lenguas indígenas empleadas en esta investigación, incluyendo el código ISO 639-3, el número de hablantes actuales y los estados de México donde principalmente se encuentran esos hablantes (Tabla 1).

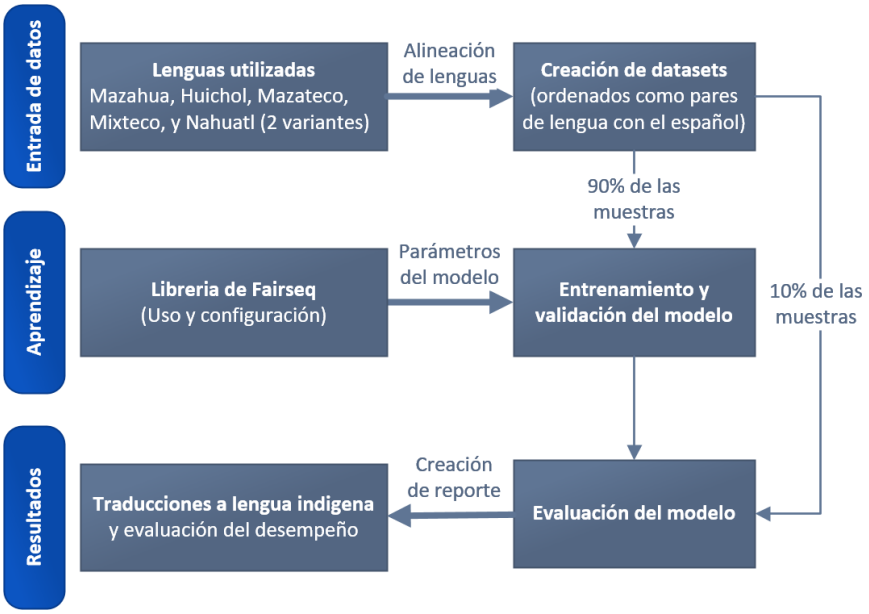

Fig. 1. Arquitectura del sistema de propuesto para implementar el baseline.

Con ello, el total de oraciones alineadas para cada lengua fue de 7930 aproximadamente, siendo el Mazateco (de Huautla de Jiménez) la lengua con menos oraciones alineadas al contar con un total de 7850 oraciones alineadas correctamente con el español.

Vale anotar además, que para esta investigación para la Biblia en español empleamos un "español simplificado" el cual utiliza un vocabulario mas sencillo que el que podría encontrarse en las biblias convenciales escritas en español, esto con el fin de facilitar las tareas de traducción desde el español hacia las lenguas indígenas (y viceversa si fuese necesario), pues como es sabido, las lenguas de escasos recursos cuentan con un vocabulario muy límitado que podria afectar el desempeño del modelo [2].

4. Método

El siguiente diagrama presenta la arquirectura del sistema implementado, el cual se divide en tres etapas: 'Entrada de datos' donde se preparan los datos para su uso en el modelo, 'Aprendizaje' donde se entrena el modelo, y 'Resultados' donde se evaluan las traducciones del modelo y se organizan los resultados generados para su posterior análisis y evaluación.

En cada una de estas etapas utilizamos Fairseq [6], una libreria escrita en PyTorch para entrenar modelos del tipo encoder-decoder que pueden ser personalizados para resolver una variedad de tareas de generación de texto, tales como traducir, resumir, parafrasear, entre otras. A continuación describimos en mayor detalle cada una de las etapas del sistema:

Tabla 2. Resultados del preprocesamiento de conjuntos de datos mediante fairseq-preprocess.

Lengua	Vocabulario	Valid UNK	Eval UNK
maz	25,480	2.41 %	2.43 %
hch	43,168	21.20 %	21.70 %
maq	22,000	1.97 %	1.81 %
mim	19,224	0.64 %	0.61 %
azz	37,632	7.31 %	7.20 %
ngu	33,296	8.80 %	8.11 %

1. **Entrada de datos:** El primer paso para crear nuestro conjunto de datos fue tomar cada una de las lenguas de interés y llevar a cabo un preprocesamiento riguroso mediante el cual se alinean los textos de La Biblia escrita en lengua indígena con La Biblia escrita en español. Este proceso se realiza utilizando los capítulos y versículos correspondientes de cada texto, lo que asegura una alineación precisa y confiable entre las dos versiones, de esta forma conseguimos representar los textos como pares de oraciones (i.e., ”texto-español, texto-lengua-indígena”). Una vez completada la alineación usamos la funcionalidad de Fairseq para preprocesar el texto antes de enviarlo al modelo de traducción automática. Este proceso implica tokenizar las palabras, segmentarlas y generar el vocabulario. El resultado final es un conjunto de datos apropiado para la tarea de traducción automática de cada lengua indígena.
2. **Aprendizaje:** Durante la etapa de aprendizaje se utilizó el 80 % de las muestras disponibles para entrenamiento y un 10 % para validación del modelo. Fairseq también permite una gran flexibilidad para personalizar los modelos generados usando distintos hiperparámetros para optimizar el desempeño del modelo. A continuación se presentan los parámetros empleados principalmente para configurar el modelo, pero puede ver la lista completa de ellos en el codigo fuente del sistema. encoder/decoder-embed-dim = 256, encoder/decoder layers = 2, dropout/attention-dropout = 0.2, learning-rate = 0.0005, optimizer adam. Finalmente, Fairseq también provee un conjunto de métricas para evaluar los resultados conseguidos, entre ellos el BLEU score que empleamos en esta investigación
3. **Resultados:** se lleva a cabo la evaluación del modelo utilizando el 10 % de las muestras disponibles. Para realizar la evaluación se emplea la métrica BLEU score, que es ampliamente utilizada en la evaluación de modelos de traducción automática. Además, se reporta el desempeño del modelo por n-gramas, desde 1-gram hasta 4-grams, lo que permite una evaluación detallada del rendimiento del modelo considerando diferentes niveles de precisión.

Cabe destacar que Fairseq también incluye una opción para manejar instancias de palabras (tokens) desconocidas, es decir, palabras que no fueron parte del vocabulario en la etapa de aprendizaje. Esta opción se llama –replace-unk y permite al usuario reemplazar tokens desconocidos con un token especial “unk” o con un token específico.

Al reemplazar tokens desconocidos con un token específico, el modelo aún puede aprender del contexto de las palabras circundantes y mejorar su capacidad para generar resultados precisos.

Tabla 3. Desempeño del modelo en los datos de validación usando las métricas ROUGE y BLEU.

Lengua	R-1	R-2	R-L	BLEU	1-gram	2-gram	3-gram	4-gram
maz	0.429	0.172	0.165	1.53	15.5	2.81	0.67	0.19
hch	0.465	0.203	0.189	2.42	17.18	3.75	1.19	0.45
maq	0.444	0.193	0.174	2.02	15.27	3.22	1.03	0.33
mim	0.538	0.251	0.196	2.43	17.61	3.97	1.27	0.4
azz	0.367	0.148	0.159	1.77	15.84	3.27	0.86	0.22
ngu	0.441	0.184	0.169	1.8	14.66	3.01	0.87	0.27

5. Análisis de resultados

En este artículo, presentamos un conjunto de datos para la traducción automática del español a seis lenguas indígenas. El conjunto de datos se preprocesó con fairseq-preprocess, siguiendo [10, 9] e informamos los resultados de referencia y las métricas de evaluación para cada par de idiomas.

5.1. Resultados del preprocesamiento del conjunto de datos

La tabla 2 muestra los resultados del preprocesamiento del conjunto de datos mediante el preprocesamiento de fairseq. La tabla informa el vocabulario de cada lengua, donde notamos que como varían entre cada una, siendo la mazahua la que tiene el vocabulario mas amplio (43,168) y el Mazateco el de menor vocabulario (22,000). También se informan los porcentajes de palabras desconocidas (UNK) para los conjuntos de de validación y evaluación con datos que van desde el 1,81 % hasta el 21,1 %.

5.2. Baseline results

La tabla 3 y la tabla 5.2 muestran los resultados de nuestro modelo de línea base en los conjuntos de validación y prueba, respectivamente, utilizando las métricas ROUGE y BLEU para medir el desempeño. Estas métricas son comúnmente utilizadas en la evaluación de modelos generativos de lenguahje, tales como resumen y traducción automática.

Especificamente, ROUGE mide la similitud entre el texto generado y el texto de referencia y para ello se consideran tres variantes: R1 que cuantifica la precisión de las palabras individuales que se superponen entre el resumen generado y el resumen de referencia, R2 mide la precisión de las secuencias de dos palabras superpuestas entre el resumen generado y el resumen de referencia, y RL mide la precisión de las secuencias de palabras superpuestas, teniendo en cuenta la longitud de la secuencia.

BLEU por otro lado mide la calidad de la traducción comparando la salida del sistema al contar el número de n-gramas en la traducción candidata que coinciden con los n-gramas en las traducciones de referencia. De manera complementaria ambas métricas proporcionan una medida cuantitativa de la calidad de la salida de los sistemas de resumen y traducción automática.

Tabla 4. Desempeño del modelo en los datos de evaluación usando las métricas ROUGE y BLEU.

Lengua	R-1	R-2	R-L	BLEU	1-gram	2-gram	3-gram	4-gram
maz	0.448	0.180	0.172	1.41	13.79	2.53	0.64	0.18
hch	0.466	0.206	0.192	2.33	17.12	3.74	1.15	0.4
maq	0.450	0.192	0.179	1.95	14.95	3.08	0.99	0.32
mim	0.525	0.247	0.192	2.62	18.74	4.25	1.32	0.45
azz	0.370	0.150	0.158	1.38	12.61	2.54	0.68	0.17
ngu	0.456	0.197	0.175	1.95	15.04	3.16	0.97	0.32

Cuanto mayor sea el valor de ROUGE o BLEU, mayor será la similitud entre el texto generado y el texto de referencia, o entre la traducción generada y la traducción de referencia, respectivamente. Como puede notarse los puntajes de las tablas 3 y 5.2 son relativamente bajos, siendo el Náhuatl de Puebla (azz) el que menor desempeño obtuvo en estos experimentos, mientras que el Mixteco y el Huichol obtuvieron los resultados mas altos.

Una posible razón podría ser el tamaño limitado del conjunto de datos. Como el número de oraciones y tokens en el conjunto de entrenamiento para cada par de idiomas es relativamente pequeño, es posible que el modelo no haya tenido suficientes datos para aprender los matices de los idiomas objetivo. Además, la complejidad y diversidad de los idiomas indígenas pueden representar un desafío significativo para los modelos de traducción automática.

Otra razón podría ser el preprocesamiento del conjunto de datos. Aunque utilizamos fairseq-preprocess, una herramienta ampliamente utilizada para el preprocesamiento de conjuntos de datos, es posible que una optimización adicional de los pasos de preprocesamiento pueda mejorar los resultados de la traducción.

Además, la calidad de las traducciones también puede verse afectada por la elección de la arquitectura del modelo, los hiperparámetros y el algoritmo de optimización. Por lo tanto, es necesario realizar más experimentos con diferentes modelos y técnicas de optimización para mejorar el rendimiento de la traducción.

Vale la pena señalar que, aunque los resultados de traducción obtenidos en nuestro estudio fueron relativamente bajos, son comparables a los informados en otros estudios con pares de lenguas de escasos recursos, como el benchmark de [5]. Por lo tanto, nuestros resultados brindan información valiosa sobre los desafíos para la traducción automática de estas lenguas y pueden informar futuras investigaciones en esta área.

6. Conclusión

La diversidad lingüística es un componente vital de la riqueza cultural de cualquier sociedad. Sin embargo, muchas lenguas indígenas están en peligro de extinción debido a factores como la globalización, la urbanización y la asimilación cultural. Con el progreso de las tecnologías para el procesamiento del lenguaje natural y su aplicación a las lenguas de escasos recursos creemos que es posible conseguir que estas lenguas no se pierdan, y con ello intentar preservar la historia y la identidad cultural de estas comunidades, lo que a su vez puede contribuir a la construcción de una sociedad más justa y equitativa.

En este artículo, creamos un conjunto de datos para la traducción automática del español a seis lenguas indígenas y evaluamos el rendimiento de la traducción utilizando las métricas de ROUGE y BLEU. Los resultados mostraron que las traducciones producidas por el modelo no fueron muy precisas, lo que indica la necesidad de una mayor mejora.

Para mejorar el rendimiento de la traducción, el trabajo futuro podría incluir el aumento del tamaño del conjunto de datos, la optimización de los pasos de preprocesamiento, la experimentación con diferentes arquitecturas de modelo, hiperparámetros y algoritmos de optimización, y la incorporación de conocimientos lingüísticos y culturales adicionales en el proceso de traducción.

En general, el desarrollo de sistemas de traducción automática efectivos para lenguas indígenas es crucial para preservar y promover la diversidad lingüística y el patrimonio cultural. Adicionalmente, como trabajo futuro planeamos agregar más lenguas indígenas considerando ademas sus respectivas variantes.

Específicamente, para el caso de México, existen 68 lenguas indígenas y según datos del Instituto Nacional de Lenguas Indígenas (INALI) se estima que existen alrededor de 364 variantes lingüísticas de estas 68 lenguas. Estas variantes reflejan la riqueza y la diversidad cultural de los pueblos indígenas de México y su patrimonio lingüístico.

Sin embargo, dada la escasez de recursos para estas lenguas, y mas aún, para sus variantes, es necesario explorar nuevas fuentes de datos que podamos incorporar para el entrenamiento de los modelos, algunas opciones incluyen: la constitución politica de México, y el uso de informacion multimodal como audio, el cual está publicamente disponible en algunos formatos de La Biblia.

Agradecimientos. Los autores agradecen al CONACYT los recursos de cómputo brindados a través de la plataforma de aprendizaje profundo para tecnologías del lenguaje del Laboratorio de Supercómputo del INAOE. Así como el uso de los recursos lingüísticos empleados en esta investigación.

Referencias

1. Bible (2023) Bible.com
2. Gu, J., Hassan, H., Devlin, J., Li, V.: Universal neural machine translation for extremely low resource languages. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, vol. 1 (2018) doi: 10.18653/v1/n18-1032
3. Gutierrez-Vasques, X., Sierra, G., Hernandez-Pompa, I.: Axolotl: A web accessible parallel corpus for Spanish-Nahuatl. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation, pp. 4210–4214 (2016)
4. Lakew, S. M., Negri, M., Turchi, M.: Low resource neural machine translation: A benchmark for five african languages. In: Proceedings of the 13th Conference on Language Resources and Evaluation, pp. 6654–6661 (2022)
5. Mager, M., Oncevay, A., Ebrahimi, A., Ortega, J., Rios, A., Fan, A., Gutierrez-Vasques, X., Chiruzzo, L., Giménez-Lugo, G., Ramos, R., Meza Ruiz, I. V., Coto-Solano, R., Palmer, A., Mager-Hois, E., Chaudhary, V., Neubig, G., Vu, N. T., Kann, K.: 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, Association for Computational Linguistics, pp. 202–217 (2021) doi: 10.18653/v1/2021.americasnlp-1.23
6. Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., Auli, M.: 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 (Demonstrations), Association for Computational Linguistics, pp. 48–53 (2019) doi: 10.18653/v1/N19-4009
7. Papineni, K., Roukos, S., Ward, T., Zhu, W. J.: BLEU: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, pp. 311–318 (2001) doi: 10.3115/1073083.1073135
8. Popović, M.: chrF: Character n-gram F-score for automatic MT evaluation. In: Proceedings of the Tenth Workshop on Statistical Machine Translation, Association for Computational Linguistics (2015) doi: 10.18653/v1/w15-3049
9. Ramesh-Harsha, R., Prasad-Sankaranarayanan, K.: Neural machine translation for low resource languages using bilingual lexicon induced from comparable corpora. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, vol. 1, pp. 1748–1759 (2018)
10. Sennrich, R., Zhang, B.: Revisiting low-resource neural machine translation: A case study. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, pp. 211–221 (2019) doi: 10.18653/v1/p19-1021
11. Vázquez, R., Scherrer, Y., Virpioja, S., Tiedemann, J.: The Helsinki submission to the AmericasNLP shared task. In: Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas, Association for Computational Linguistics (2021) doi: 10.18653/v1/2021.americasnlp-1.29