Low-resource Indigenous language translator

Emiliano Churruca[1], Jorge Santos[1], Santiago Durante[1], Marcia Nascimento[2], Juan Santos[1], and Luis Chiruzzo[3]

[1] Universidad Nacional de Hurlingham (UNAHUR), Buenos Aires, Argentina
[2] Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil
[3] Universidad de la República, Montevideo, Uruguay

Abstract. The line of research presented in this work aims to formulate and implement a methodology for the creation of translators for low-resource languages in South America. Language is part of the culture and identity of communities, especially among Indigenous peoples. Globalization and historical marginalization, among other factors, have caused many languages to recede until they disappear, while others are fading away.
The proposed methodology involves the search, collection, and cleaning of data for the creation of a bitext corpus that enables the development of language models sufficiently rich for the implementation of translation systems based on Transformers.
The languages under study, with which this work will initially be carried out, are Kaingang and Ayoreo. Regarding the former, the methodology has already begun to be applied by assembling a small corpus, and a translator based on the encoder-decoder scheme using Transformers has been implemented. The preliminary results presented here make it possible to draw some useful conclusions for the formulation of the methodology.

Keywords: Low-resource languages , Machine translation , Linguistic revitalization , Transformer

Traductor de lenguas originarias de bajos recursos

Resumen La línea de investigación presentada en este trabajo tiene como objetivo formular e implementar una metodología para la creación de traductores de lenguas con bajos recursos en América del Sur. El lenguaje forma parte de la cultura e identidad de las comunidades, especialmente en pueblos originarios. La globalización y la marginalización histórica, entre otros factores, provocaron que muchas lenguas se hayan retraído hasta desaparecer mientras que otras, se desvanecen. La metodología propuesta involucra la búsqueda, recolección y depuración de datos para la creación del corpus bitexto que permita la creación de

modelos de lenguaje suficientemente ricos como para la implementación de máquinas de traducción basadas en *Transformers*. Las lenguas objeto de estudio con las que inicialmente se trabajará son el Kaingang y el Ayoreo. En cuanto a la primera, se ha empezado a aplicar la metodología conformando un pequeño corpus y se ha implementado un traductor basado en el esquema codificador-decodificador usando *Transformers*. Los resultados preliminares expuestos aquí permiten obtener algunas conclusiones útiles para la formulación de la metodología.

Palabras clave: Lenguas de bajos recursos, Traducción automática, Revitalización lingüística, *Transformer*

1. Introducción

"Todas las personas deberían tener acceso a las Tecnologías del Lenguaje en sus lenguas maternas, incluidas las lenguas indígenas. El reto es doble: (i) preservar la cultura a través del lenguaje y (ii) facilitar la comunicación entre lenguas. Las lenguas que no tengan la oportunidad de adoptar las Tecnologías del Lenguaje serán cada vez menos utilizadas, mientras que las lenguas que se benefician de tecnologías interlingüísticas, como la Traducción Automática, serán cada vez más utilizadas" (Asociación Europea de Recursos Lingüísticos, 2019).

Según (Marmion et al., 2014), el objetivo de la revitalización para los miembros que así lo deseen no está relacionado con la expectativa que su lengua compita, sino con "fortalecer la conexión de las personas con su lengua y cultura, construir un sentido de identidad y bienestar, y aumentar la conciencia lingüística".

En este sentido, construir un traductor para lenguas originarias de bajos recursos en América Latina es una tarea muy importante y compleja debido a la baja disponibilidad de textos digitalizados. Sin embargo, existen ejemplos previos como (Mager et al., 2016), quienes presentaron un traductor wixarika-español. El wixarika es una lengua hablada por aproximadamente 60000 habitantes en México, para la tarea utilizaron un traductor estadístico. También, (Borges et al., 2021) introducen un método para identificar verbos en un texto en guaraní con el objetivo de mejorar el desempeño de un traductor basado en Transformers (Vaswani et al., 2017). Dicho método está basado en reglas morfológicas acerca de cómo los sufijos y prefijos caracterizan modo, tiempo, género y persona de un núcleo o raíz verbal presente en los tokens de un texto en Guaraní a ser traducido. En (Moreno et al., 2024) se describe la construcción de un conjunto bitexto para awajun-español y el ajuste fino (*fine tunnig*) de un modelo pre entrenado usando la técnica de transferencia de aprendizaje (*transfer learning*). Awajun es una lengua con cincuenta y cinco mil hablantes en el Perú la cual no había sido considerada, a la fecha del trabajo de estos autores, para el tratamiento de traducción automática con modelos *Transformers*. Aunque, si se habían usado

esa arquitectura para otras lenguas como el quechua ayacucho, quechua cuzco, aymara, shipibo-konibo y asháninka (Mager et al., 2021; Oncevay, 2021).

Asimismo, existe una competencia organizada por AmericasNLP en la que se entrenan traductores automáticos para lenguas originarias en América utilizado un corpus propuesto por la misma organización. En el certamen se fomenta la investigación en el procesamiento del lenguaje natural para lenguas originarias en América (Sur, Centro y Norte). Los trabajos de (Ebrahimi et al., 2023) y (De Gibert et al., 2025) muestran algunos resultados obtenidos en las competencias de 2023 y 2025 respectivamente.

En este trabajo mostramos una etapa inicial en la construcción de un traductor entre el kaingang y el español. El kaingang pertenece a la rama macro-jê y a la familia lingüística jê. Su orden constitutivo básico es el SOV (Sujeto, Objeto y Verbo). Es una lengua aislante, que presenta algunos procesos morfológicos en la categoría verbal (Nascimento, 2017). Se estima que el número de hablantes de esta lengua es del 50% de la comunidad kaingang, o sea alrededor de 25.000 personas. La mayor parte de la población se encuentra en el Sur de Brasil.

El proceso metodológico que utilizamos para desarrollar el traductor automático kaingang-español se muestra en la Sección 2 y, en la Sección 3, se pueden observar los resultados y discusión de las primeras traducciones.

Finalmente, las líneas de trabajo futuro están en la Sección 4.

2. Materiales y Métodos

El desarrollo de un traductor automático abarca dos tareas generales: el armado del corpus bitexto y el entrenamiento del traductor.

Para el armado del corpus bitexto en este trabajo preliminar utilizamos el Evangelio según San Mateo en (Anónimo judeocristiano, 90) y (Wycliffe Bible, 2011). Al estar ambos textos en formato pdf, se los sometió a un script de Python para normalizarlos. La tarea incluyó entre otras cosas: pasar todo el texto a minúsculas, eliminar caracteres especiales, encabezados, pie de página y numeración. Una vez normalizados los datos, se alinearon los documentos, esto es emparejar oraciones entre ambos textos. La alineación de realizó por versículo, quedando en total 1050 pares de oraciones.

Para el entrenamiento utilizamos una adaptación de un modelo *Transformer* propuesto por (Brownlee et al., 2022). La arquitectura adaptada tiene cuatro bloques de codificación y cuatro bloques de decodificación, dentro de cada bloque se utilizaron 8 cabezas de atención. El tamaño de la secuencia de entrada es 256, las matrices de consulta y claves tienen una dimensión de 128 mientras que la capa *Feed Forward* 512 unidades con función de activación ReLU en su capa oculta. Como optimizador utilizamos Adam con una tasa de aprendizaje dinámica bajo un esquema de *CustomSchedule*.

Con el corpus bitexto creado y el modelo *Transformer* preparado, se procedió al entrenamiento. Para ello, dividimos el corpus de 1050 pares de oraciones en conjuntos de entrenamiento, validación y testeo tomando el 70%, 15% y 15% respectivamente. Entrenamos el traductor con 2000 épocas y con 0.1 de *dropout*.

Los recursos utilizados para dicho entrenamiento es una computadora de escritorio con una placa de video GeForce NVIDIA RTX 3090 con 24 GB de memoria RAM. Las librerías utilizadas fueron Keras versión 3.12.1 y Tensorflow versión 2.20.0.

3. Resultados y discusión

En la Tabla 1 se detalla, para cada lengua, la cantidad total de *tokens* presentes en el *corpus* descripto en la Sección 2. También se informa el tamaño de los vocabularios obtenidos.

Table 1. Cantidad de *tokens*.

Lengua	Cantidad de *tokens*	Tamaño del vocabulario
kaingang	48629	1129
español	23226	3286

En la Figura 1 se observa la evolución de la pérdida tanto en el conjunto de entrenamiento como en el de evaluación.

Se puede observar un claro sobre ajuste (*over fitting*) a partir de la época 150 cuando la pérdida sobre el conjunto de evaluación empieza a incrementarse de manera significativa. Esto sucede aún habiendo usando un *dropout* de 0.1. La pérdida sobre el conjunto de entrenamiento cae rápidamente y se mantiene estable con un valor cercano a 0 a partir de la época 500. Esto se verifica cuando se le pide al traductor que traduzca oraciones que estaban en el conjunto de entrenamiento prediciendo la traducción exacta. En modo inverso, el traductor no tiene aciertos semánticos en traducir oraciones del conjunto de testeo aunque, desde el punto de vista gramatical la traducciones fueran correctas.

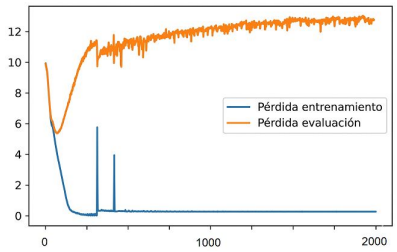

Fig. 1. Evolución de la pérdida durante el entrenamiento de 2000 épocas.

4. Conclusiones y trabajos futuros

Los resultados obtenido evidencian un sobre ajuste del modelo durante el entrenamiento lo cual es el resultado de usar un corpus muy pequeño. Esto conduce a una conclusión inmediata acerca de la necesidad de incrementar su tamaño, tarea que actualmente estamos llevando a cabo con el objetivo de llevar el conjunto de entrenamiento bitexto de 1050 a 25000 pares. Actualmente estamos procesando nuevos documentos en kaingang. Por otro lado, como parte de la comprensión del los recursos utilizados, buscaremos explorar diferentes híper parámetros de la arquitectura del *Transformer* utilizados a los efectos de mejorar el desempeño del mismo. Además, como parte de nuestro trabajo utilizaremos modelos multilingües pre entrenados aplicánoles un proceso de ajuste fino con nuestro corpus de datos a los efectos de poder comparar con la estrategia hasta ahora utilizada que es el entrenamiento desde cero incluyendo métricas estándar para la evaluación de las traducciones como BLEU y CHRF.

Referencias

Anónimo judeocristiano (ca. 80–90). Evangelio según san mateo. Archivo PDF. Autoría atribuida tradicionalmente al apóstol Mateo.

Borges, Y., Mercant, F., and Chiruzzo, L. (2021). Using guarani verbal morphology on guarani-spanish machine translation experiments. *Procesamiento del Lenguaje Natural*, 66:89–98.

Brownlee, J., Cristina, S., and Saeed, M. (2022). *Building Transformer Models with Attention: Implementing a Neural Machine Translator from Scratch in Keras*. Machine Learning Mastery.

De Gibert, O., Pugh, R., Marashian, A., Vázquez, R., Ebrahimi, A., Denisov, P., Rice, E., Gow-Smith, E., Prieto, J., Robles, M., et al. (2025). Findings of the americasnlp 2025 shared tasks on machine translation, creation of educational material, and translation metrics for indigenous languages of the americas. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 134–152.

Ebrahimi, A., Mager, M., Rijhwani, S., Rice, E., Oncevay, A., Baltazar, C., Cortés, M., Montaño, C., Ortega, J. E., Coto-Solano, R., et al. (2023). Findings of the americasnlp 2023 shared task on machine translation into indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 206–219.

Mager, J. M. M., Romero, C. B., and Ruiz, I. V. M. (2016). Traductor estadístico wixarika-español usando descomposición morfológica.

Mager, M., Oncevay, A., Ebrahimi, A., Ortega, J., Gonzales, A. R., Fan, A., Gutierrez-Vasques, X., Chiruzzo, L., Giménez-Lugo, G., Ramos, R., et al. (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.

Marmion, D., Obata, K., and Troy, J. (2014). *Community, identity, wellbeing: the report of the Second National Indigenous Languages Survey*. Australian Institute of Aboriginal and Torres Strait Islander Studies Canberra.

Moreno, O., Atamain, Y., and Oncevay, A. (2024). Awajun-op: Multi-domain dataset for spanish–awajun machine translation. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 112–120.

Nascimento, M. (2017). *Evidencialidade em Kaigang: descrição, processamento e aquisição*. PhD thesis, Tese de doutorado, UFRJ.

Oncevay, A. (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.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wycliffe Bible (2011). *Mateus Ty Rá: The New Testament in Kaingang*. Wycliffe Bible Translators, Inc. Archivo PDF.