Capítulo 6

EVALUACIÓN DE LA CALIDAD DE UN TRADUCTOR AUTOMÁTICO DE LA LENGUA INDÍGENA YOREM-NÓKKI

Reyna Elisa Montes Santiago
José Emilio Sánchez García
Yobani Martínez Ramírez
Carolina Tripp Barba

Universidad Autónoma de Sinaloa
Universidad Autónoma Indígena de México

https://doi.org/10.36825/SEICIT.2025.03.C06

Resumen

La traducción automática de lenguas indígenas de bajos recursos hoy en día sigue siendo un desafío latente, sobre todo por la escasez de recursos lingüísticos digitales. En este trabajo de investigación se propone la evaluación de la calidad de un traductor automático (que utiliza el modelo Transformer) del idioma Yorem-nókki al idioma Español y viceversa. La investigación se aborda desde un enfoque cuantitativo con alcance de tipo descriptivo-explicativo. Los resultados permitieron comprobar las hipótesis alternas CAL-BLUE-H1 y CAL-CHRF++-H1, en ambos casos, se comprobó que supera el 60% de calidad de traducción del idioma Yorem-nókki al idioma Español (y viceversa). Sin embargo, es importante mencionar que la evaluación de la calidad fue más estricta con la métrica BLUE y más realista con la métrica chrF++. Así también, aunque el resultado todavía está alejado de valor ideal (90-100%), esta es una primera versión de modelo de traducción con área de oportunidad de mejora. Para trabajos futuros se espera incrementar la base de conocimiento del idioma Yorem-nókki, utilizar modelos pre-entrenados de lenguas indígenas para mejorar la calidad de las traducciones y diseñar e implementar un software prototipo de traducción automática de libre acceso para la comunidad Yoreme.

Palabras clave: Evaluación de la calidad, traducción automática, métricas BLUE y chrF++, lengua indígena Yorem-Nókki.

Introducción

La Traducción Automática (TA) ha experimentado una evolución notable en la última década, principalmente gracias a la emergencia y perfeccionamiento de los modelos de TA Neuronal (TAN). Esta tecnología ha revolucionado la forma en que el contenido multilingüe es producido y consumido, permitiendo la traducción de grandes volúmenes de texto de manera rápida y eficiente. La demanda de comunicación transiligüe ha crecido exponencialmente en un mundo cada vez más interconectado, y la TA se ha posicionado como una herramienta indispensable para satisfacer esta necesidad.

En Sinaloa y Sonora, la lengua indígena Yoreme-Mayo (Yorem-nókki) se encuentra en un alto riesgo de desaparición por diversas razones como la falta de trasmisión intergeneracional de la lengua, la reducción de los ámbitos de uso, la cantidad y calidad de los materiales escritos en general. En este contexto, el uso de TA son una alternativa viable para su conservación.

No obstante, a pesar de los avances significativos en la TA, su aplicación a lenguas de bajos recursos, en específico el Yorem-nókki, es un desafío considerable, dada la severa escasez de grandes conjuntos de datos paralelos necesarios para entrenar un sistema de TA robusto. Por otra parte, esta limitación se agrava por la insuficiencia de métodos de evaluación efectivos y apropiados.

En este trabajo se evalúa la calidad de la traducción automática del idioma Yorem-nókki al idioma Español y viceversa con las métricas Bilingual Evaluation Understudy (BLEU) y Character-level F-score++ (ChrF++). Para ello se diseñó, entrenó y midió el rendimiento de un modelo de Red Neuronal Transformer (RNT) con una base de conocimiento extraída del diccionario Yorem-nókki del autor Aguilar Velázquez (2020). Esta es la primera etapa de la construcción de una aplicación inteligente que ayuda en la conservación del idioma Yorem-nókki.

El documento está estructurado en siete apartados: el primero consiste en la presente introducción; el segundo, expone los conceptos relacionados; el tercero, aborda los trabajos relacionados; el cuarto, presenta el proceso metodológico; el quinto muestra los resultados obtenidos; el sexto expone las conclusiones; y finalmente, en el séptimo apartado se enlistan las referencias en las que está sustentada la investigación.

Conceptos Relacionados

Traducción Automática

Como parte de la lingüística aplicada, la TA es relevante desde una perspectiva científica, ya que actúa como un campo experimental tanto para la lingüística como para la informática, especialmente en el

procesamiento y análisis automático del lenguaje natural. Esta disciplina también conecta con otras áreas de la lingüística aplicada, como la traductología, la terminología, la psicolingüística y la pragmática entre otras (Casacuberta Nolla & Peris Abril, 2017)automatic translation was dominated by systems based on linguistic information, but then later other approaches opened up the way, such as translation memories and statistical machine translation which draw on parallel language corpora. Recently the neuronal machine translation (NMT.

La TA se refiere al proceso automatizado de transformación de texto o voz de un idioma, conocido como idioma de origen, a otro, denominado idioma de destino. Implica el uso de modelos computacionales y algoritmos para analizar y comprender la información de entrada en el idioma de origen y generar una representación equivalente en el idioma de destino (Naveen & Trojovský, 2024).

Red Neuronal Transformer

Una red neuronal está formada por un conjunto de unidades de procesamiento simples, también conocidas como neuronas artificiales, que están fuertemente interconectadas. Su función consiste en calcular un producto escalar entre las entradas de la neurona y un vector de pesos asociado, seguido de una función de activación no lineal (Casacuberta Nolla & Peris Abril, 2017). Según (Vaswani et al., 2017), la red neuronal Transformer es un tipo de red recurrente que se ha convertido en la arquitectura más popular y robusta para el modelo codificador-decodificador en problemas de traducción automática neuronal. Este modelo utiliza mecanismos de autoatención que permiten al codificador y al decodificador considerar cada palabra en toda la secuencia de entrada.

Calidad de la TA

La calidad de TA se define como el grado en que un sistema de traducción automática produce textos en el idioma de destino que son comprensibles y precisos, manteniendo la fidelidad al contenido original. La calidad se evalúa mediante diferentes enfoques, incluyendo métricas automáticas

y evaluaciones humanas, para asegurar que la traducción conserve tanto el sentido como el estilo del texto fuente (Hiebl & Gromann, 2023). En el ámbito de la TA, la precisión y la fluidez son dos componentes fundamentales para evaluar la calidad de las traducciones generadas por máquinas.

- Precisión

La precisión en la TA se refiere a la exactitud con la que el sistema traduce el contenido del idioma fuente al idioma objetivo, manteniendo la fidelidad al significado original. Es decir, una traducción precisa transmite correctamente la información y el sentido del texto original sin distorsiones. Este concepto es esencial en la evaluación de la calidad de la TA, ya que una alta precisión indica una correspondencia cercana entre la TA y la traducción humana de referencia (Koponen, 2010)

- Fluidez

La fluidez se refiere a la naturalidad y coherencia de la traducción en el idioma objetivo. Una traducción fluida es aquella que, además de ser gramaticalmente correcta, resulta fácil de leer y entender para un hablante nativo del idioma objetivo. Este aspecto es esencial, ya que una traducción que carece de fluidez puede dificultar la comprensión del mensaje, aunque sea precisa en términos de contenido (Papineni et al., 2002).

Evaluación de la Calidad de la TA

La evaluación de la calidad de la TA se refiere a la medición y análisis de la efectividad y precisión de las traducciones generadas por sistemas automáticos (Briva-Iglesias, 2022), esta puede llevarse a cabo mediante métodos automáticos, que son eficientes y económicos, así como mediante evaluaciones humanas, que proporcionan un estándar de calidad más confiable. La calidad de la TA puede medirse mediante métodos automáticos como BLEU y chrF++ que se basan en la coincidencia léxica.

- BLEU (Bilingual Evaluation Understudy), es una métrica utilizada

para evaluar la calidad de traducciones generadas por sistemas de traducción automática. Fue propuesta por (Papineni et al., 2002) y se basa en el cálculo de la precisión de n-gramas al comparar la TA con una o más traducciones de referencia humanas.

- chrF++ (CHaRacter-level F-score), es una métrica que combina la evaluación basada en n-gramas de caracteres con n-gramas de palabras para aprovechar tanto la información morfológica como la semántica, resultando en una mejor correlación con evaluaciones humanas en diversas lenguas (Popović, 2017).

Lengua Indígena

Según la Real Academia Española (RAE, 2023) la definición de la palabra lengua se refiere al vocabulario y gramática propio de un grupo social. De acuerdo con la UNESCO (UNESCO, 2023) la lengua indígena no se refiere solamente a símbolos de identidad o pertenencia a un grupo, sino que también son vehículos de valores éticos. Representan una trama de sistemas de conocimientos mediante el cual un pueblo forma un todo con la tierra y es indispensable para su supervivencia.

Lengua Yoreme-Mayo

El Mayo es un grupo étnico originario del sur del estado de Sonora y del norte de Sinaloa, también llamado Yoreme (ArqueologiaMexicana, 2022). La agrupación lingüística mayo, o como sus hablantes lo denomina Yorem-nókki (que significa Yoreme-Mayo), pertenece a la familia Yuto-nahua. Se considera una sola lengua, puesto que no tiene variación interna (INALI, 2020).

Los Yoreme habitan en el noroeste de México, en parte de los estados de Sonora y Sinaloa, que comprende tres áreas naturales: la sierra, los valles y la zona costera, las cuales definen sus características productivas, así como sus problemáticas. En el estado de Sinaloa sus comunidades se localizan en la parte norte en los municipios de Ahome, El Fuerte y Guasave; mientras que, en Sonora, habitan en la porción sur, en los municipios de Etchojoa, Huatabampo y Navojoa, principalmente.

Trabajos relacionados

A continuación, se presentan diversas investigaciones relacionadas con la evaluación de la calidad de traductores automáticos de lenguas indígenas. Es importante decir, que una parte de estas investigaciones fueron detectadas en una revisión de literatura relacionada con el aprendizaje móvil de lenguas indígenas (Montes Santiago et al., 2024)

De Gilbert et al (2025) presentan una investigación que contribuye al desarrollo tecnológico, educativo y de evaluación para 14 lenguas indígenas de América a través de métodos de Procesamiento de Lenguaje Natural (PLN) adaptados, con participación de comunidades lingüísticas y hablantes nativos.

Aborda el problema de la escasez de recursos y herramientas para el PLN aplicado a las lenguas indígenas de América. Entre los desafíos que enfrenta en su estudio menciona la ausencia de métricas adecuadas y fiables para evaluar la calidad de la TA en lenguas indígenas. En la tarea compartida de desarrollo de métricas de TA utilizaron como métricas principales chrF++ y BLUE para medir el desempeño en la traducción de 14 lenguas indígenas de escasos recursos.

En el trabajo de investigación de Pinhanez et al. (2024), proponen el desarrollo de un marco tecnológico y metodológico para la revitalización de lenguas indígenas mediante herramientas de Inteligencia Artificial (IA) y PLN contribuyendo con esto al apoyo en documentación, preservación y revitalización de lenguas indígenas en peligro de desaparición. Construye asistentes de escritura con correctores ortográficos y predicción de palabras para lenguas indígenas como Guaraní, Mbyá y Nheengatu y prueba estos prototipos con jóvenes indígenas para fomentar el uso escrito de sus idiomas.

En el estudio realizan varios experimentos y evaluaciones para probar la efectividad de modelos de IA en la traducción y procesamiento de lenguas indígenas. Se comparó los modelos de IA en términos de

precisión de traducción. Las métricas utilizadas fueron BLEU Score que evalúa la similitud entre la traducción generada y la traducción esperada, así como BLEURT y BERTScore. Estos modelos están basados en redes neuronales para evaluar la calidad de la traducción.

Bautista Morales et al. (2024), realizan una investigación en donde presentan el desarrollo de una arquitectura de TA basada en Red Neuronal Transformer (RNT) para el idioma mixteco, una lengua indígena con escasos recursos. En su trabajo para calcular la calidad de traducción de español a mixteco, se utilizó la métrica BLEU demostrando con esto la viabilidad del modelo de traducción a pesar de la limitada disponibilidad de datos, evidenciando la necesidad de ampliar el corpus de entrenamiento. Proponen la integración de un sistema colaborativo con hablantes nativos y el desarrollo de un algoritmo basado en reglas gramaticales para optimizar la calidad de las traducciones. En este estudio se establecen las bases para futuras investigaciones en la preservación y revitalización de lenguas indígenas mediante IA.

En el trabajo de Le et al. (2023), presentan el primer estudio sobre el reconocimiento de entidades nombradas para la lengua indígena inuktitut de Canadá que carece de recursos lingüísticos y de grandes datos etiquetados. Hace una contribución importante al estudiar la transferencia de características lingüísticas del inglés al inuktitut, en función de reglas o incrustaciones de palabras bilingües. La investigación propone dos enfoques para el modelo de reconocimiento de entidades nombradas de la lengua indígena inuktitut. El enfoque basado en reglas y el basado en incrustaciones de palabras bilingües que buscan superar las limitaciones de datos anotados al aprovechar los recursos existentes.

La evaluación del modelo se realiza mediante varias métricas de desempeño en tareas de alineación de palabras, reconocimiento de entidades nombradas y TA. Los experimentos incluyen la utilización de las métricas como BLEU y chrF++ para analizar la calidad de la traducción; en términos de recuperación, precisión y puntuación obtiene resultados que muestran la eficacia de los métodos de reconocimiento de entidades nombradas mejorando el rendimiento de la TA neuronal

del inuktitut al inglés. En el estudio se implementan pruebas del modelo de reconocimiento de entidades nombradas, destacando dificultades debido a la escasez de datos y variabilidad lingüística. Los autores enfatizan la importancia de la colaboración con comunidades indígenas para la recolección y validación de datos y sugiere el uso de técnicas como aprendizaje transferido, así como datos sintéticos para mejorar la precisión de los modelos en lenguas con pocos recursos digitales.

Los autores Tonja et al. (2023), mencionan que existe una baja calidad en los sistemas de TA neuronal para lenguas con pocos recursos, como Wolaytta idioma de recursos limitados. En este sentido, debido a la escasez de datos paralelos existe una limitante en el acceso a tecnologías lingüísticas modernas para hablantes de estas lenguas. En su estudio proponen utilizar datos monolingües del lado fuente junto con datos sintéticos generados por un modelo inicial, aplicando un enfoque de autoaprendizaje (self-learning) y ajuste fino (fine-tuning) del modelo para mejorar la calidad de la TA neuronal en lenguas de bajos recursos.

En sus experimentos utilizan las métricas BLEU y chrF++ para evaluar objetivamente la calidad de las TA generadas. Estas métricas cuantificaron la efectividad de las técnicas propuestas (autoaprendizaje y ajuste fino con datos monolingües) para mejorar la TA neuronal en un escenario de bajos recursos (Wolaytta-inglés).

En el trabajo de Tonja et al. (2023), proponen el primer corpus paralelo español-mazateco y español-mixteco para modelos de TA. En su investigación realizan un gran avance en la digitalización de lenguas indígenas, proporcionando datos accesibles para investigadores y desarrolladores de PLN. En este estudio, la métrica BLEU se utilizó para evaluar el rendimiento de los modelos de TA desarrollados para las lenguas indígenas mazateco y mixteco, en relación con el español.

Este estudio llevó a cabo la evaluación de modelos de TA en lenguas de bajos recursos utilizando tres enfoques diferentes: transformación, aprendizaje por transferencia y ajuste fino de modelos de TA multilingües preentrenados, demostrando el impacto de los corpus paralelos en la

mejora de la traducción y promoción de la preservación lingüística a través del uso de tecnología para fomentar la enseñanza y documentación de estas lenguas mazateco y mixteco.

Los autores Billah Nagoudi et al. (2021), mencionan que la falta de corpus paralelos disponibles para lenguas indígenas dificulta la implementación de modelos de TA eficientes. Proponen el desarrollo de un Modelo de Traducción Transformer denominado IndT5 que permite mejorar la traducción. Para el entrenamiento de este modelo se elaboró un corpus con una colección de diez lenguas indígenas y el español, empleando técnicas de aprendizaje transferido que demuestran que los modelos preentrenados en español pueden mejorar la traducción en lenguas de bajos recursos.

En su trabajo evalúa el impacto de los datos en la TA, mostrando que la cantidad de corpus disponible es un factor clave en la calidad de las traducciones. Esta investigación, utiliza las métricas BLEU y chrF++ para evaluar la calidad de las TA generadas por el modelo IndT5. Los resultados muestran que los modelos de traducción basados en aprendizaje automático mejoran significativamente cuando se usa el corpus paralelo, aunque aún existen desafíos en la fluidez y precisión de las traducciones.

Metodología

En esta sección se explica la metodología aplicada para evaluar la calidad de la TA de una frase del idioma Yorem-nókki al idioma Español y viceversa.

La presente investigación se aborda desde un enfoque cuantitativo. Según (Hernández Sampieri et al., 2014), implica un conjunto de procesos de recolección y análisis de datos numéricos para responder a un planteamiento del problema de manera objetiva y medible.

El enfoque cuantitativo sigue un proceso sistemático, empírico y crítico que se centra en la recolección y el análisis de datos estadísticos, permitiendo realizar inferencias a partir de patrones y relaciones en los datos numéricos obtenidos. Tiene como objetivo medir y analizar

aspectos específicos del fenómeno de estudio, proporcionando resultados replicables y generalizables (Hernández & Mendoza, 2020).

El alcance de la investigación es de tipo descriptivo. Puede proporcionar datos descriptivos que permiten obtener una visión general de un fenómeno, lo cual es importante para entender su magnitud y características principales antes de pasar a etapas explicativas o experimentales.

A continuación, se presentan las hipótesis que guiaron este trabajo de investigación.

- Hipótesis Nula (CAL-BLUE-H0): La calidad de la TA del idioma Yorem-nókki no alcanza 60 puntos de la métrica BLEU.
- Hipótesis Alternativa (CAL-BLUE-H1): La calidad de la TA del idioma Yorem-nókki alcanza o supera 60 puntos de la métrica BLEU.
- Hipótesis Nula (CAL-CHRF-H0): La calidad de la TA del idioma Yorem-nókki no alcanza 60 puntos de la métrica chrF++.
- Hipótesis Alternativa (CAL-CHRF-H1): La calidad de la TA del idioma Yorem-nókki alcanza o supera 60 puntos de la métrica chrF++.

Enseguida, se explican las etapas del preprocesamiento de los datos para poder entrenar el modelo de red neuronal Transformer, y posteriormente, se describe el procedimiento que se siguió para evaluar el rendimiento del modelo y poder comprar las hipótesis planteadas.

Preprocesamiento de datos

El preprocesamiento de datos es de gran importancia en la preparación de la base de conocimiento de idioma cuando se realizan TA. En este sentido, se buscó limpiar y estandarizar cada palabra en el conjunto de datos para mejorar el rendimiento del modelo.

- Eliminación de la puntuación

Para garantizar la uniformidad de todo el conjunto de datos se eliminaron los signos de puntuación de idioma Yorem-nókki y del idioma Español:

coma, punto, punto y coma, dos puntos, puntos suspensivos, signos de admiración e interrogación, raya, guion, paréntesis, corchetes, comillas dobles y comillas simples.

- Limpieza de espacios y caracteres

Para garantizar un formato adecuado se eliminaron espacios adicionales entre palabras, espacios iniciales y espacios finales. También se identificaron y eliminaron caracteres no válidos.

- Minúsculas

Con la finalidad de tener coherencia en las frases se convirtieron a minúscula. Aunque en algunas ocasiones es importante utilizar mayúsculas y minúsculas para nombres propios, acrónimos, ubicaciones y términos significativos, en este caso no se consideraron. En la Tabla 1 se puede observar un primer parte del preprocesamiento.

Tabla 1
Preprocesamiento: eliminación de puntuación, limpieza de espacios y conversión a minúsculas

Frase	Idioma Yorem-nókki	Idioma Español
Normal	¡ Wá'am jákune'e wéyye !	¡ Vete por allá !
Después de Eliminar	waam jakunee weyye	vete por alla

Fuente: Elaboración propia.

- Manejo de la Tokenización

Debido a que el modelo de traducción no puede trabajar con las frases enteras, estás se dividieron en palabras. Entonces, se implementó el proceso de tokenización el cual consiste en dividir el texto en unidades básicas llamadas tokens, en este caso, la unidad básica es la palabra.

- Diccionario de datos

Debido a que el modelo solamente utiliza información numérica,

posteriormente, se procedió a construir un diccionario numérico con cada token. En este sentido, a cada token se le asignó un número único. En la Tabla 2 se puede observar una segunda parte del preprocesamiento.

Tabla 2
Preprocesamiento: manejo de tokenización y diccionario de datos

Frase	Idioma Yorem-nókki	Idioma Español
Preprocesada	waam jakunee weyye	vete por alla
Tokenización-palabras	[‘waam’, ‘jakunee’, ‘weyye’]	[‘vete’, ‘por’, ‘alla’]
Tokenización-números	[60, 28, 29]	[45, 46, 47]

Fuente: Elaboración propia.

- Longitud de las frases

Para garantizar que las frases de entrada del modelo tuvieran la misma longitud, se identificó la frase más larga dentro del conjunto de datos y se aplicó un rellenado (con valor numérico 0) para el resto de las frases. De esta manera, todas las frases alcanzaron la misma longitud de entrada en el modelo.

Entrenamiento

El modelo utilizado es una arquitectura de red neuronal de última generación llamada Transformer. Este modelo fue propuesto por (Vaswani et al., 2017) y se basa en mecanismos de autoatención.

Para el entrenamiento se utiliza como conjunto de datos el diccionario de palabras y frases del idioma Yorem-nókki recopilado por (Aguilar Velázquez Néstor, 2020). Aunque existen otros conjuntos de libros relacionados con cuentos, tradiciones y normas del idioma Yorem-nókki, el diccionario de Aguilar Velázquez (2020) reúne el mayor número de palabras relacionadas con el idioma.

El conjunto de datos del diccionario utiliza un alfabeto de 20 letras

(A, B, CH, E, H, I, J, K, L, M, N, O, P, R, S, T, U, W, X, Y) de los cuales hay 5 vocales y 15 consonantes. En el idioma Yorem-nókki utiliza el signo ortográfico llamado apóstrofo (´) para la correcta pronunciación de las palabras, pero en la limpieza de los datos fue necesario eliminarlo para mejorar la precisión de la traducción.

Debido a falta de información digital de idioma Yorem-nókki en este trabajo se utilizaron 39,510 palabras y frases en idioma Español y su correspondiente significado en idioma Yorem-nókki. Es importante mencionar que no se utilizó ninguna técnica especial para aumentar el conjunto de datos.

Para el entrenamiento se utilizó el entorno de desarrollo Google Colab (2025) y también, el servidor local Intel Xeon CPU con sistema operativo Windows 10. En Google Colab se generaron modelos de entrenamiento del idioma Yorem-nókki al idioma Español (y viceversa) con los siguientes hiperparámetros: 2500 épocas, 3 codificadores, 3 decodificadores, función de activación ReLU (Rectified Linear Unit), lotes de 64 ejemplos a la vez y optimizador Adam (Adaptive Moment Estimation). Del conjunto total solamente se utilizó un 75% (29362) para entrenamiento y un 25% para prueba (9878).

Para evaluar el rendimiento del modelo se utilizaron las métricas Bilingual Evaluation Understudy (BLEU) y Character-level F-score++ (ChrF++) en los conjuntos de datos de entrenamiento.

Procedimiento

El procedimiento para llevar a cabo la evaluación de la calidad de la traducción se describe a continuación:

1. El administrador ingresa al servidor donde está el conjunto de datos y el modelo de red neuronal Transformer.

2. El administrador define los parámetros de configuración del Modelo y se entrena la red neuronal Transformer en la plataforma Google Colab del idioma Yorem-nókki al idioma español y viceversa. Para

cada caso se obtiene un archivo con los modelos entrenados.

3. Luego, se configura el servidor local Intel Xeon CPU con los modelos entrenados del idioma Yorem-nókki al idioma español y viceversa, para posteriormente realizar pruebas de traducción.

4. Se prueban traducciones con 300 frases aleatorias (con más de una palabra) del idioma Yorem-nókki al idioma Español para evaluar la calidad de la traducción con la métrica BLUE y la métrica chrF++. El valor esperado de la métrica BLEU y chrF++ está en el rango de 0 a 1.

5. Posteriormente, se prueban 300 traducciones de una frase del idioma Español al idioma Yorem-nókki con el modelo neuronal Transformer ya entrenado y se obtiene el valor de la métrica BLEU BLUE y la métrica chrF++. El valor esperado está en mismo rango de 0 a 1.

6. Finalmente, con la información recabada se realiza un análisis estadístico de medias y desviación estándar con los valores de la métrica BLEU y chrF++.

Resultados y Discusiones

En esta sección se presentan los resultados obtenidos de la evaluación del modelo. Para las pruebas del modelo se consideró la evaluación cuantitativa, las variables a considerar fueron la métrica BLEU y la métrica chrF++.

Evaluación de la calidad

La evaluación de las traducciones se realizó considerando frases con más de una palabra. Para ello se seleccionaron de manera aleatoria 300 frases cortas del diccionario.

A continuación, se presentan los resultados de traducción:

- Resultados del idioma Yorem-nókki al idioma Español

En la Tabla 3 se puede apreciar los resultados de la evaluación de la calidad

de las algunas frases del idioma Yorem-nókki al idioma Español, con las métricas BLUE y chrF++.

Tabla 3
Resultados de la evaluación de la calidad del idioma Yorem-nókki al idioma Español

Palabra en Yorem-nókki	Palabra en Español	Traducción del Modelo Neuronal Transformer	Métrica BLEU	Métrica chrF++
kaa lawti aane	lentitud	lentitud	1	1
jaykimsu jiaayse	a que horas tu	a que tantas horas tu	0.1714	0.5972
ketche alheyya jachisu emow weyye tekkilpo	buenas tardes como te ha ido en el trabajo	buenas tardes como te levantaste hoy	0.3082	0.5606
kaa jiaaley	callaba	callaba	1	1
batchia bodeegam	silo	silo	1	1
achikola weena	circundar	conductor	0	0.1319
tebuxria buiyte	transcurrir	transcurrir	1	1
jammut sullame	galante	mujeriego	0	0.045
jitta nawnunuwuyeme	receptaculo	bandolero	0	0.0786
librom oorewaapo	librero	librero	1	1

Fuente: Elaboración propia.

Se puede apreciar en la Tabla 3 que la evaluación de la calidad de las métricas se encuentra en el rango de 0 a 1. En los valores más cercanos a 1 indica que la traducción realizada por la inteligencia se acerca a la traducción humana. En estos 10 casos se observa que la métrica chrF++ realiza una evaluación más realista de la calidad de la traducción. Por ejemplo, en la frase en Yorem-nókki “jaykimsu jiaayse” que significa en Español “a que horas tu”, la propuesta del traductor automático al idioma Español fue “a que tantas horas tu”. A pesar de ser muy similares la puntuación asignada por la métrica BLUE fue más estricta con un valor de 0.1714 (muy cercana a cero), mientras tanto, la métrica chrF++ le asignó

una puntuación más realista con un valor de 0.5972 (muy cercana a uno).

Para determinar la ubicación de los puntos donde se centran o se inclinan los resultados de las métricas de BLUE y chrF++, en la Tabla 4, se muestran los resultados de la media y la desviación estándar.

Tabla 4
Resultados estadísticos de la evaluación de la calidad del idioma Yorem-nókki al idioma Español

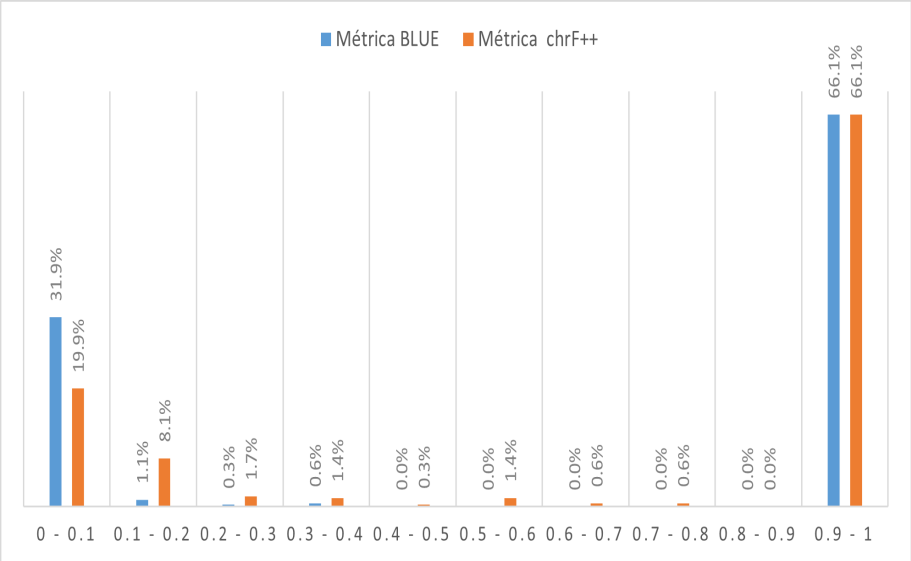
Métricas	Media	Desviación Estándar
Métrica BLUE	66.59%	0.4682
Métrica chrF++	70.87%	0.4179

Fuente: Elaboración propia.

En la Tabla 4 se aprecia que la métrica chrF++ presenta un mejor resultado de 70.87% y una desviación estándar baja de 0.4682 en cuanto a la calidad de la traducción. Por otra parte, aunque la métrica BLUE presenta un ligero bajo resultado de 66.59% y una desviación estándar de 0.4682, se pudo observar que muchas TA calificadas con BLUE fueron calificadas de manera más estricta (con valor cero a pasar existir relación entre la traducción original y la traducción automática).

En la Figura 1, se muestra la distribución de puntajes de la métrica BLUE y la métrica chrF++ de las traducciones del idioma Yorem-nókki al idioma Español. La Figura muestra los rangos de evaluación de la calidad (a través de las métricas BLUE y chrF++) de las frases del idioma Yorem-nókki al idioma Español. Los valores cercanos a 1 indican que las traducciones fueron casi idénticas a la traducción humana. En ese sentido, se puede apreciar que el mayor porcentaje (66.1% de TA) se encuentra en el rango de 0.9 – 1. También, se observa que la evaluación de la calidad de las TA en ambas métricas alcanzó el mismo valor.

Figura 1
Distribución de puntaje de la métrica BLUE y la métrica chrF++ de las traducciones del idioma Yoreme al idioma Español



Fuente: Elaboración propia.

Por otro lado, también se aprecia que la métrica BLUE fue más estricta en la evaluación y asignó una valoración de 0 – 0.1 a un 31.9% de las traducciones contra chrF++ que asignó una valoración de 19.9%. En la Figura las valores de chrF++ están más distribuidas, lo que significa que asignó valoraciones a traducciones candidatas que parcialmente coincidían con la traducción original.

- Resultados del idioma Español al idioma Yorem-nókki.

En la Tabla 5 se puede apreciar los resultados de la evaluación de la calidad de las algunas frases del idioma Español al idioma Yorem-nókki, con las métricas BLUE y chrF++.

Tabla 5
Resultados de la evaluación de la calidad del idioma Español al idioma Yorem-nókki

Frases en Español	Frases en Yorem-nókki	Traducción del Modelo Neuronal Transformer	Métrica BLEU	Métrica chrF++
de aquella manera	waneeli	aajimmak	0.0675	0.0463
le vendra a la memoria	waatina	waatina	1	1
dio por cierto	suaalek	suaalek	1	1
lo embrujaria	at moriatek-kipanua	bexjasu	0	0.0333
estando sentados	jookaka	jookari	0	0.469
se ha puesto escaso	poroxtila	poroxtila	1	1
le lanza	maabareka	maabaana	0	0.3649
le pasa rozando	jelixte	jelixte	1	1
pasandole encima	warakteka	warakteka	1	1
llegaron lejos	mekkayaixsuk	mekkayaixsuk	1	1

Fuente: Elaboración propia.

Se puede ver en la Tabla 5 que la métrica chrF++ también realiza una evaluación más a nivel de carácter de la calidad de la traducción. Por ejemplo, en la frase en Español "le lanza" significa en Yorem-nókki "maabareka", la propuesta del traductor automático fue "maabaana" que significa en idioma Español "le querra tirar", si se observa las palabras en Yorem-nókki de la frase original y candidata son muy similares, por lo que la puntuación asignada por la métrica chrF++ fue de 0.3649. Por otra parte, la métrica BLUE fue más estricta y le asignó una valoración de 0 (cero).

Para determinar la ubicación de los puntos donde se centran o se

inclinan los resultados de las métricas de BLUE y chrF++, en la Tabla 6, se muestran los resultados de la media y la desviación estándar.

Tabla 6
Resultados estadísticos de la evaluación de la calidad del idioma Español al idioma Yorem-nókki

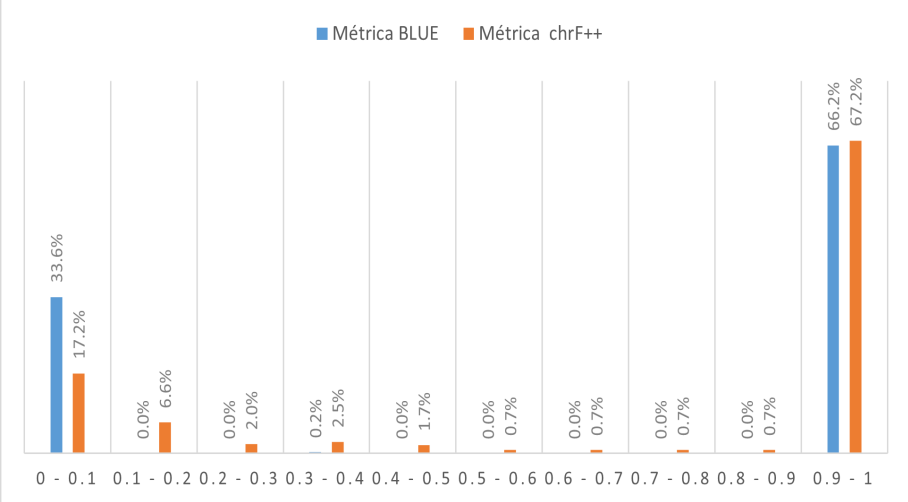
Métricas	Media	Desviación Estándar
Métrica BLUE	66.28%	0.4725
Métrica chrF++	73.22%	0.3993

Fuente: Elaboración propia.

También, en la Tabla 6, aunque se observa que la métrica chrF++ presenta un mejor resultado de 73.22% y una desviación estándar baja de 0.3993 en cuanto a la calidad de la traducción, esto se debe a que la valoración se hace a nivel de carácter por lo que su resultado es un poco más realista con la traducción humana. Por otro lado, la métrica BLUE presenta un resultado de 66.28% y una desviación estándar de 0.4725 en cuanto a la calidad de la traducción, esto se debe a que la métrica es más estricta en la valoración de las frases candidatas y de referencia.

En la Figura 2, se muestra la distribución de puntajes de las métricas BLUE y chrF++ de las traducciones del idioma Español al idioma Yorem-nókki. La Figura muestra los rangos de evaluación de la calidad (a través de las métricas BLUE y chrF++) de las frases del idioma Español al idioma Yorem-nókki. Los valores cercanos a 1 indican que las traducciones fueron casi idénticas a la traducción humana. Por lo anterior, se puede apreciar que el mayor porcentaje (66.2% de BLUE y 67.2% de chrF++, de las TA) se encuentra en el rango de 0.9 – 1. También, se observa que la evaluación de la calidad de las TA en ambas métricas alcanzó casi el mismo valor (con un 1% de diferencia porcentual).

Figura 2.
Distribución de puntaje de la métrica BLUE y la métrica chrF++ de las traducciones del idioma Español al idioma Yorem-nókki.



Fuente: Elaboración propia.

En la Figura 2, también se aprecia que la métrica BLUE fue más estricta en la evaluación y asignó una valoración de 0 – 0.1 a un 33.6% de las traducciones contra chrF++ que asignó una valoración de 17.2%. En esta Figura también se observa que las valores de chrF++ están más distribuidos, lo que significa que asignó valoraciones a traducciones candidatas que parcialmente coincidían con la traducción original, mientras que BLUE las penalizado.

Discusiones

Los resultados obtenidos de la evaluación de la calidad con el modelo de red neuronal Transformer entrenado con un corpus limitado de la lengua indígena Yorem-nokki alcanzó un promedio de 66.44% con la métrica BLEU y un 72.05% con la métrica chrF++. Estos datos guardan relación con los hallazgos encontrados en (De Gibert et al., 2025) que destaca que, aunque los modelos basados en transformers logran avances considerables en lenguas indígenas, los valores de métricas como BLEU

suelen estar por debajo del 70%, debido a la escasez de datos paralelos y a la morfología compleja de estas lenguas. Por otra parte, la presente investigación evidencia que chrF++ otorgó evaluaciones más realistas en comparación con BLEU, lo que permite mencionar que (De Gibert et al., 2025) sugiere que métricas a nivel de carácter, como chrF++, son más adecuadas para evaluar traducciones en lenguas con alta variabilidad morfológica y carencia de recursos lingüísticos estructurados. Esto valida la elección metodológica de emplear ambas métricas, permitiendo una visión más detallada y equitativa de la calidad de traducción.

Los hallazgos de esta investigación no solo aportan evidencia sobre la calidad alcanzable en la TA del Yorem-nókki, sino que también refuerzan el potencial de la IA como herramienta para la preservación y revitalización lingüística, tal como lo plantea (Pinhanez et al., 2024) en su estudio, los autores destacan que una tecnología lingüística efectiva debe estar al servicio de las comunidades hablantes, considerándose no únicamente desde su precisión mecánica, sino también desde su accesibilidad, utilidad cultural y capacidad de fomentar el uso cotidiano de la lengua.

En este sentido, la evaluación realizada en el presente trabajo representa una contribución inicial en esa dirección, al facilitar que estudiantes y docentes de comunidades Yorem-nókki puedan realizar traducciones inmediatas entre su lengua originaria y el español. Este uso práctico puede contribuir, en palabras de (Pinhanez et al., 2024), a "reinsertar las lenguas en la vida digital cotidiana", algo que se considera esencial para su supervivencia.

Conclusiones

De la evaluación de la calidad de las TA utilizando el modelo transformer con 300 frases aleatorias (con más de una palabra) se puede concluir lo siguiente:

- La métrica BLUE, del idioma Yorem-nókki al idioma Español, alcanzó una media de 66.59% con una desviación estándar de 0.4682, mientras

que del idioma Español al idioma Yorem-nókki, alcanzó una media de 66.28% con una desviación estándar de 0.4725. Esto confirma como valida la Hipótesis Alterna (CAL-BLUE-H1): La calidad de la TA del idioma Yorem-nókki alcanza o supera 60 puntos de la métrica BLEU.

- La métrica chrF++, del idioma Yorem-nókki al idioma Español, alcanzó una media de 70.87% con una desviación estándar de 0.4179, mientras que del idioma Español al idioma Yorem-nókki, alcanzó una media de 73.22% con una desviación estándar de 0.3993. Esto confirma como valida la Hipótesis Alterna (CAL-CHRF-H1): La calidad de la TA del idioma Yorem-nókki alcanza o supera 60 puntos de la métrica chrF++.

Se puede apreciar que la métrica chrF++ obtuvo una valoración de 4 puntos porcentuales por encima de la métrica BLUE, esto significa que su valoración de la calidad de la traducción es más granular, a nivel de carácter, por lo que se consideran incluso las aproximaciones parciales de una traducción. En este contexto, se considera una valoración más realista. En cambio, la métrica BLUE es más estricta ya que la valoración de la traducción es a nivel de palabra, por lo que penaliza las palabras que son diferentes incluso si es una letra.

Es importante mencionar que, aunque esto presenta un buen resultado en la medición cuantitativa del modelo de traducción del idioma Yorem-nókki al idioma Español (y viceversa), este resultado aún está alejado del valor ideal (90-100%) ya que esto en buena media depende de contar con una base de conocimiento más amplia, incluir diccionarios de sinónimos e inyectar más ciclos de entrenamiento al modelo propuesto.

Con esta idea en mente, se identificaron las siguientes áreas de oportunidad para trabajo futuro. Primero, incorporar más frases de libros publicados en Yorem-nókki e incorporar diccionarios de sinónimos y antónimos; segundo, probar con algunos modelos pre-entrenados de otros idiomas indígenas ya que según algunos autores se ha mejorado la calidad de las traducciones automáticas; tercero, realizar una evaluación cualitativa de la calidad del traductor automático con expertos del idioma; y cuarto, construir un prototipo de software de libre acceso para que

la comunidad Yorem-nókki del Norte de Sinaloa ayude a enriquecer el conjunto de datos.

Este proyecto de investigación está en desarrollo, pero a corto plazo estará en línea el software de traducción automática, el código fuente y el conjunto de datos del idioma Yorem-nókki de esta manera se pretende que esté al alcance y pueda ser utilizado como alternativa a futuras investigaciones en el campo de la traducción automática.

Referencias

Aguilar Velázquez Néstor. (2020). Diccionario (Yorem-Seewa) Yorem-Nókki-Español / Español-Yorem-Nókki (Gráficos de creativos 7 (Ed.); Primera ed).

ArqueologiaMexicana. (2022). Mayo. https://arqueologiamexicana.mx/

Bautista Morales, R., Martínez Ramírez, Y., Rocha Peña, L. E., & Montes Santiago, R. E. (2024). Arquitectura de un traductor automático para el idioma mixteco: un enfoque específico para lenguas indígenas con escasos recursos lingüísticos. Revista de Investigación En Tecnologías de La Información, 12(28), 71–81. https://doi.org/10.36825/RITI.12.28.007

Billah Nagoudi, E. M., Chen, W. R., Abdul-Mageed, M., & Cavusoglu, H. (2021). IndT5: A Text-to-Text Transformer for 10 Indigenous Languages. Proceedings of the 1st Workshop on Natural Language Processing for Indigenous Languages of the Americas, AmericasNLP 2021, 265–271. https://doi.org/10.18653/v1/2021.americasnlp-1.30

Briva-Iglesias, V. (2022). English-Catalan Neural Machine Translation: state-of-the-art technology, quality, and productivity. Revista Tradumatica, 20, 149–176. https://doi.org/10.5565/rev/tradumatica.303

Casacuberta Nolla, F., & Peris Abril, Á. (2017). Traducción automática

neuronal. Tradumàtica Tecnologies de La Traducció, 15, 66–74.

De Gibert, O., Pugh, R., Marashian, A., Vazquez, R., Ebrahimi, A., Denisov, P., Rice, E., Gow-Smith, E., Prieto, J., Robles, M., Manrique, R., Moreno, O., Lino, A., Coto-Solano, R., Alvarez, A., Agüero-Torales, M., Ortega, J. E., Chiruzzo, L., Oncevay, A. & Mager, M. (2025). Findings of the AmericasNLP 2025 Shared Tasks on Machine Translation, Creation of Educational Material, and Translation Metrics for Indigenous Languages of the Americas. Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), 134–152. https://doi.org/10.18653/v1/2025.americasnlp-1.16

Google. (2025, Agosto 30). Google Colaboratory [Software]. https://colab.research.google.com

Hernández, R., & Mendoza, C. (2020). Metodología de la investigación. Las rutas cuantitativa, cualitativa y mixta Las rutas Cuantitativa Cualitativa y Mixta. In McGRAW-HILL Interamericana Editores S.A. de C.V. http://repositorio.uasb.edu.bo:8080/bitstream/54000/1292/1/Hernández- Metodología de la investigación.pdf

Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, P. (2014). Metodología de la investigación (Sexta). Mc Graw Hill Education.

Hiebl, B., & Gromann, D. (2023). Quality in Human and Machine Translation: An Interdisciplinary Survey. Proceedings of the 24th Annual Conference of the European Association for Machine Translation, EAMT 2023, 375–384.

INALI. (2020, Agosto 30). Atlas de las lenguas indígenas nacionales de México. https://atlas.inali.gob.mx/inicio

Koponen, M. (2010). Assessing Machine Translation Quality with Error Analysis. In Mikael: Kääntämisen ja tulkkauksen tutkimuksen aikakauslehti (Vol. 4). https://doi.org/10.61200/mikael.129675

Le, N. T., Kasdi, I., & Sadat, F. (2023). Towards the First Named Entity Recognition of Inuktitut for an Improved Machine Translation. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 84–93. https://doi.org/10.18653/v1/2023.americasnlp-1.10

Montes Santiago, R. E., Sánchez García, J. E., Martínez-Ramírez, Y., & Bautista Morales, R. (2024). Aprendizaje móvil de lenguas indígenas: Revisión de literatura. In U. A. I. de México (Ed.), La educación y el impacto tecnológico actual con inteligencia artificial (1ra., pp. 169–188). Astra ediciones. https://doi.org/10.61728/AE24002929

Naveen, P., & Trojovský, P. (2024). Overview and challenges of machine translation for contextually appropriate translations. IScience, 27(10), 1–25. https://doi.org/10.1016/j.isci.2024.110878

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A method for automatic evaluation of machine translation. Proceedings of the Annual Meeting of the Association for Computational Linguistics.

Pinhanez, C., Cavalin, P., Storto, L., Finbow, T., Cobbinah, A., Nogima, J., Vasconcelos, M., Domingues, P., Mizukami, P. de S., Grell, N., Gongora, M., & Gonçalves, I. (2024). Harnessing the Power of Artificial Intelligence to Vitalize Endangered Indigenous Languages: Technologies and Experiences. 1–48. http://arxiv.org/abs/2407.12620

Popović, M. (2017). chrF++: words helping character n-grams. Proceedings of the Second Conference on Machine Translation, 2(1), 612–618. https://doi.org/10.18653/v1/W17-4770

RAE. (2023, Agosto 30). Definición de Aplicación Móvil. Diccionario de la Real Academia española. https://dle.rae.es/

Tonja, A. L., Kolesnikova, O., Gelbukh, A., & Sidorov, G. (2023). Low-Resource Neural Machine Translation Improvement Using

Source-Side Monolingual Data. Applied Sciences (Switzerland), 13(2). https://doi.org/10.3390/app13021201

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. Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), 94–102. https://doi.org/10.18653/v1/2023.americasnlp-1.11

UNESCO. (2023, Abril 28). Lenguas indígenas, conocimientos y esperanza. https://courier.unesco.org/es/articles/lenguas-indigenas-conocimientos-y-esperanza

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems 30 (NeurIPS 2017). http://arxiv.org/abs/1706.03762