Arquitectura de un traductor automático para el idioma mixteco: un enfoque específico para lenguas indígenas con escasos recursos lingüísticos

Architecture of a machine translator for mixteco: a specific approach for indigenous languages with scarce linguistic resources

Rolando Bautista Morales
Universidad Autónoma Indígena de México, Ahome, Sinaloa, México
rolandobautista@uaim.edu.mx
ORCID: 0009-0006-3889-1226

Yobani Martínez Ramírez
Universidad Autónoma de Sinaloa, Ahome, Sinaloa, México
yobanimartinez@uaim.edu.mx
ORCID: 0000-0002-4967-9187

Luis Enrique Rocha Peña
Universidad Autónoma de Sinaloa, Ahome, Sinaloa, México
enriquerocha@uas.edu.mx
ORCID: 0009-0009-3044-6656

Reyna Elisa Montes Santiago
Universidad Autónoma de Sinaloa, Ahome, Sinaloa, México
reynaelisa@ms.uas.edu.mx
ORCID: 0009-0003-3465-5785

doi: https://doi.org/10.36825/RITI.12.28.007

Recibido: Junio 14, 2024
Aceptado: Agosto 25, 2024

Resumen: La revitalización de las lenguas indígenas es de suma importancia para preservar la diversidad cultural y el conocimiento ancestral que representan. En la actualidad, las tecnologías como la inteligencia artificial (IA) ofrecen oportunidades sin precedentes para apoyar estos esfuerzos mediante la traducción automática (TA) y las redes neuronales (RN). Este trabajo de investigación se centra en detallar la arquitectura de un traductor automático para idiomas indígenas con escasos recursos lingüísticos. La propuesta considera el modelo de red neuronal *transformer*, el cual utiliza un corpus paralelo y técnicas de alineación automática. Para evaluar la calidad de las traducciones se implementa una arquitectura funcional mediante un sistema de traducción del idioma español al idioma mixteco, donde la métrica principal es BLEU. Para ello, se preparó un corpus de conocimiento con riqueza gramatical de 1.1K de palabras y frases del idioma mixteco. Los resultados de la calidad de la traducción son relativamente bajos, ya que se considera que este rendimiento es aún limitado debido al pequeño tamaño del corpus por lo que se concluye con propuestas interesantes para mejorar la calidad de la traducción automática.

Palabras clave: *Revitalización de Lenguas Indígenas, Traducción Automática, Transformer, Escasos Recursos Lingüísticos, Mixteco.*

Abstract: The revitalization of indigenous languages is of utmost importance to preserve the cultural diversity and ancestral knowledge they represent. Currently, technologies such as artificial intelligence (AI) offer unprecedented opportunities to support these efforts through machine translation (MT) and neural networks (NR). This research work focuses on detailing the architecture of a machine translator for indigenous languages with scarce linguistic resources. The proposal considers the transformer neural network model, which uses a parallel corpus and automatic alignment techniques. To evaluate the quality of the translations, a functional architecture is implemented by means of a translation system from Spanish to Mixtec, where the main metric is BLEU. For this purpose, a knowledge corpus with a grammatical richness of 1.1K words and phrases from the Mixtec language was prepared. The translation quality results indicate a low quality of translation from Spanish to Mixtec. It is considered that this performance is still limited due to the small size of the corpus so it is concluded with interesting proposals to improve the quality of machine translation.

Keywords: *Indigenous Language Revitalization, Machine Translation, Transformer, Scarce Linguistic Resources, Mixtec.*

1. Introducción

Los idiomas indígenas, son ricos en diversidad cultural y lingüística,[1] menciona que cuando se pierde una lengua se deja de nombrar eso que a cada cultura le resulta especialmente importante y significativo. Por eso, al perderse una lengua, se pierde también una parte importante de la cultura a la que nombra, o incluso, desaparece. Los autores[2] plantean que, "debido a la falta de recursos digitales, muchas de estas lenguas podrían extinguirse, y por ende se perdería la conexión con la cultura de los pueblos y las características de las lenguas".

En el panorama actual de la traducción automática (TA) para lenguas con escaso recursos lingüísticos (ERL), la brecha tecnológica se manifiesta de manera evidente. La inteligencia artificial (IA), con su capacidad para procesar y comprender patrones lingüísticos complejos, ha transformado la TA para los idiomas hegemónicos predominantes, dejando a un lado a las lenguas menos representadas. Al implementar estas tecnologías a menudo carecen de adaptabilidad y eficacia cuando se enfrentan con lenguas indígenas con ERL, contribuyendo así a la marginación digital de las comunidades. No obstante, la IA está redefiniendo su capacidad para abordar este desafío.

Hoy en día, la IA en el ámbito de la TA, particularmente las redes neuronales, ha conducido avances notables como la aplicación de algoritmos avanzados de aprendizaje automático. De acuerdo con [3] el modelo de red neuronal denominado *Transformer* cambia literalmente la concepción que se tiene de lo que era capaz de lograr la IA en el campo de la TA. Este algoritmo ha demostrado eficacia en la comprensión, generación y traducción de texto, por consecuente es la base de este trabajo de investigación para mejorar la TA en idiomas indígenas con ERL.

Investigaciones recientes han explorado el potencial del modelo *Transformer* en la TA, estos avances han demostrado mejoras significativas en la calidad de las traducciones, pero la mayoría se ha enfocado en idiomas de amplio uso. Los autores [4] mencionan que aun cuando en México existen 68 lenguas indígenas oficialmente reconocidas [5]el desarrollo de herramientas digitales para estas ha sido casi nulo. Esto revela una brecha en la aplicación de tecnologías de vanguardia para abordar las necesidades específicas de las comunidades indígenas.

Estudios como [6], [7] , [8] y [9], han destacado las dificultades en la alineación automática, la *tokenización* y el entrenamiento de modelos con idiomas indígenas de ERL. En este sentido, la aplicación y evaluación exhaustiva de herramientas en el contexto de lenguas minoritarias aún requiere mayor atención. Los autores antes mencionados evidencian la necesidad de implementar estrategias más adaptables, indicando un espacio propicio para esta investigación.

Por lo anterior, en este estudio se presenta una arquitectura de TA adaptada a lenguas indígenas con ERL, contextualizando la problemática global de la TA para lenguas minoritarias. En este contexto, se detalla una propuesta innovadora, se desglosan los componentes clave y se demuestra la funcionalidad de la arquitectura de TA. Para lograr este último punto se preparó una arquitectura funcional para evaluar la calidad de las traducciones

del idioma indígena mixteco.

Esta investigación se justifica por la urgente necesidad de preservar y revitalizar lenguas indígenas como el mixteco. Además, la falta de recursos tecnológicos en idiomas indígenas limita su uso en plataformas digitales, lo que motiva la creación de un traductor automático específico para la lengua mixteca. Nuestra propuesta no solo facilita la comunicación y preservación del idioma, sino que también abre nuevas vías para su estudio y enseñanza.

El Mixteco (*Tu'un Savi* o idioma de la lluvia), se habla en un espacio histórico de tres entidades federativas de México: Guerrero, Puebla y Oaxaca. De acuerdo con los datos del Instituto Nacional de Estadística y Geográfica (INEGI) es una de las lenguas indígenas que tiene mayor número de hablantes, ocupando el cuarto lugar a nivel nacional [10].

El presente artículo se encuentra estructurado en siete apartados: el primero, expone la actual introducción; el segundo, plantea los conceptos relacionados; el tercero, aborda los trabajos relacionados; el cuarto, detalla la propuesta de una arquitectura para traducción automática; el quinto, presenta la experimentación de una arquitectura funcional y discute los resultados; el sexto, expone las conclusiones; y finalmente, en el séptimo apartado se muestran las referencias en las que se apoya la presente investigación.

2. Conceptos relacionados

2.1. Arquitectura de software

La arquitectura de software es el diseño de más alto nivel de la estructura de un sistema, el cual consiste en un conjunto de patrones y abstracciones que proporcionan un marco claro para la implementación del sistema [11]. Los autores [12] definen a la arquitectura de software de un sistema como el conjunto de estructuras necesarias para razonar sobre el sistema. Comprende elementos de software, relaciones entre ellos, y propiedades de ambos.

2.2. Inteligencia Artificial

De acuerdo con [13] la definición de Inteligencia Artificial (IA) se relaciona con la capacidad de las computadoras para ejecutar tareas cognitivas similares a las realizadas por la mente humana, especialmente en el ámbito del aprendizaje y la solución de problemas. La IA es una rama de las ciencias computacionales encargada de estudiar modelos de cómputo capaces de realizar actividades propias de los seres humanos con base a dos de sus características primordiales: el razonamiento y la conducta [14].

2.3. Red Neuronal Transformer

Una red neuronal (RN) se compone de un conjunto de unidades de procesamiento simple (o neuronas artificiales) densamente conectadas entre sí y cuya función es realizar un producto escalar entre las entradas a la neurona y un vector de pesos (asociado a cada neurona) seguido de una función no lineal de activación[15]. De acuerdo con [3] la RN *transformer* es un modelo de RN recurrente usado para ser la más popular y más robusta arquitectura para la estructura del codificador – decodificador para la solución de problemas relacionados con la traducción automática neuronal (TAN). Según [16] este modelo emplea mecanismos de autoatención que permite tanto al codificador como el decodificador tener en cuenta cada palabra de toda la secuencia de entrada.

2.4 Traducción Automática

La Traducción Automática (TA), denominada en inglés *Machine Translation* (MT), es un área de la lingüística computacional que investiga el uso de software para traducir texto o habla de un lenguaje natural a otro tanto con o sin ayuda humana. La TA en sentido amplio abarca toda una variedad de sistemas que sólo comparten la utilización del ordenador como instrumento de traducción [17].

La TA, como rama de la lingüística aplicada, es importante desde el punto de vista científico, pues sirve como campo experimental de la lingüística y la informática, especialmente en el ámbito del procesamiento y análisis automático del lenguaje natural. Esta disciplina aplicada permite establecer vínculos con otras disciplinas de la lingüística aplicada como la traductología, la terminología, la sicolingüística y la pragmática, entre otras[15]. En adelante será indistinto hablar de TA o TA neuronal (TAN).

2.5. Evaluación de la Traducción Automática

El modelo BLEU (*Bilingual Evaluation Understudy*), una herramienta ampliamente utilizada para evaluar la calidad de las traducciones automáticas. BLEU compara las traducciones generadas por el modelo con traducciones de referencia humanas, evaluando la precisión y fluidez de la traducción [18]. Una puntuación BLEU es un número entre 0 y 100. Una puntuación de 0 indica una traducción de baja calidad en la que la traducción no coincide en absoluto con la referencia. Una puntuación de 100 indica una traducción perfecta, idéntica a la referencia. No es necesario alcanzar una puntuación de 100: una puntuación BLEU entre 40 y 60 indica una traducción de alta calidad[19].

3. Trabajos Relacionados

En los últimos años se ha realizado diferentes investigaciones para lenguas indígenas con ERL en diferentes partes del mundo donde diseñan e implementan arquitecturas con redes neuronales *Transformer*. A continuación, se citan los más actuales y relevantes.

En el trabajo de investigación de [20], se utiliza una técnica de aprendizaje profundo basada en redes neuronales para los idiomas Inglés y el idioma Urdu. La arquitectura utilizada en el estudio de traducción automática de inglés a urdu se basó en un modelo de codificador-decodificador con mecanismo de atención, específicamente empleando redes de memoria a corto y largo Plazo (Por sus siglas en inglés, LSTM). Se utilizan tamaños de corpus paralelos de alrededor de 30,923 oraciones. El corpus contiene oraciones del corpus paralelo inglés-urdu, noticias y oraciones que se utilizan con frecuencia en la vida cotidiana. En cuanto a los resultados obtenidos, se evaluaron utilizando métricas automáticas como el puntaje BLEU (*Bilingual Evaluation Understudy*). Se realizaron simulaciones múltiples del modelo y se obtuvo un puntaje BLEU promedio de 45.83, lo cual indica un nivel aceptable de calidad en las traducciones generadas por el sistema. Además, se compararon las salidas del modelo con el traductor de Google, mostrando similitudes en las traducciones producidas.

En la investigación de [9], aplica técnicas de TA para crear un modelo para la traducción entre el idioma español y el idioma quechua chanka. Con respecto a la arquitectura del modelo, se usó la arquitectura basada en mecanismos de atención denominada Transformers. Además, este trabajo proporciona nuevos recursos que comprenden un nuevo corpus con 119, 000 traducciones paralelas. Los resultados experimentales indicaron un puntaje de 39,5 en términos de BLEU (*Bi-Lingual Evaluation Understudy*), lo que representa una mejora notable en comparación con las técnicas convencionales de Traducción Automática Estadística y basadas en reglas.

En la investigación [12] los autores proponen un estudio sobre la lengua inuktitut mediante preprocesamiento y TAN, con el fin de revitalizar la lengua que pertenece a la familia inuit, un tipo de lengua polisintéticas que significa relación bi-unívoca entre una secuencia de morfemas y una secuencia de sílabas"[21]. Que se habla en el norte de Canadá. El enfoque se concentra en: (1) desarrollaron un modelo de segmentación de palabras bidireccional LSTM, y (2) crearon un sistema de TAN Inuktitut-Inglés basado en la arquitectura codificador-decodificador *Transformer*. Utilizaron modelos pre-entrenados tipo BERT y lograron mejoras significativas en comparación con la línea base en el conjunto de pruebas, ambas para lenguas indígenas de Canadá

En el artículo [22] los autores utilizaron una RN basada en la arquitectura *Transformer*. Plantean que los investigadores de la TA no pueden resolver solos el problema de la falta de recursos, por ello proponen la investigación participativa como un medio para involucrar a todos los agentes necesarios en el proceso de desarrollo de TA. Demostraron la viabilidad y escalabilidad de la investigación participativa con un estudio de caso sobre TA para lenguas africanas. Utilizaron métricas de evaluación como BLEU (*Bilingual Evaluation Understudy*), HTER (*Human-targeted Translation Error Rate*) y ChrF (*Character F-score*). Los resultados de las evaluaciones mostraron que hubo variabilidad en la calidad de las TA a los idiomas africanos de bajo recurso. Algunos puntajes de BLEU fueron moderados, como 34.85 para Igbo y 38.62 para yoruba, lo que indica una calidad aceptable, pero con margen de mejora. Sin embargo, los puntajes más altos, como 48.94 para Swahili, sugieren una mejor calidad en las traducciones. En general, los autores consideran que los resultados no fueron ni completamente buenos ni completamente malos, destacando la necesidad de seguir mejorando los modelos de TA para estos idiomas.

En el estudio [7] presentan el primer sistema de TAN para el idioma ayuuk. En sus experimentos tradujeron del idioma ayuuk al idioma español y viceversa. El ayuuk es un idioma hablado en el estado de Oaxaca en México por el pueblo Ayuukjä'äy (en español comúnmente conocido como Mixes). Usaron diferentes fuentes para crear

un corpus paralelo de bajos recursos, más de 6,000 frases. Para algunos de estos recursos confiaron en alineación automática. El sistema que propusieron se basa en la arquitectura neuronal *Transformer* y se basa en un entorno codificador-decodificador. Para el entrenamiento de su modelo utilizaron 2 configuraciones A y B. Los resultados de la métrica BLEU para la configuración B con 250 épocas de entrenamiento fueron de 5.83 en el conjunto de desarrollo de la traducción de ayuuk a español. Estos resultados muestran un rendimiento prometedor en la traducción, especialmente considerando las limitaciones de recursos en el contexto de lenguas indígenas.

A partir de los artículos analizados se presenta en la Tabla 1 una comparativa donde se identifican las herramientas que utilizaron los autores en el diseño de la arquitectura de los traductores automáticos.

Tabla 1. Herramientas utilizadas en el diseño de las Arquitecturas de los Traductores Automáticos.

Artículo	**Le y Sadat** [12]	**Zacarías y Meza** [7]	**Knowles *et al.*** [23]	**Moreno Veliz** [24]	**Feldman y Coto-Solano** [8]	**Huarcaya Taquiri** [9]	**Basit Andrabi y Wahid** [20]	**Xu** [6]
Modelo	Trans-former	Trans-former	Trans-former	Trans-former	Trans-former	Trans-former	No LSTM encoder	Trans-former
Métrica	+BLEU 8.40	BLEU 5.83	CHRF N/I	BLEU N/I	BLEU 16.9	BLEU 39.5	BLEU 4.5	BLEU 48.92
Toke-nizador	N/I	Subword-nmt library	13a	Sentence Piece	N/I	Tensor-flow	Tensor-flow	N/I
Len-guaje	N/I	Python	N/I	Python	Python Google Colab	Python Google Colab	Prolog Local	N/I Local
Biblio-teca	Marian-NMT Toolkit	Joey NMT	N/I	Toolkit Fairseq	PyTorch, OpenNMT	Tensor-flow	Tensor-flow	N/I
Corpus	1293.3K	6K	N/I	N/I	1.5K, 3K, 6K	145.2K	30.90K	100K

Fuente: Elaboración Propia.

En la Tabla 1, se muestra una comparativa de las características principales de las investigaciones anteriormente citadas, donde primariamente se identificó el modelo de red neuronal, las métricas de evaluación que utilizaron para evaluar la traducción, el lenguaje de programación, el entorno de desarrollo, el *tokenizador* utilizado en el análisis de texto, las bibliotecas de aprendizaje profundo y el tamaño del corpus. Esto proporciona una visión clara de las metodologías y tecnologías empleadas en cada estudio.

4. Propuesta de una Arquitectura de Traducción Automática

La arquitectura propuesta se divide en tres capas: 1) capa de presentación; 2) capa de aplicación; y 3) capa de datos. En la Figura 1, se puede apreciar la propuesta de una arquitectura de TA. En la Figura 1, se aprecia la arquitectura propuesta para realizar traducciones de lenguas indígenas con ERL. A continuación, se describen cada una de las capas de la arquitectura.

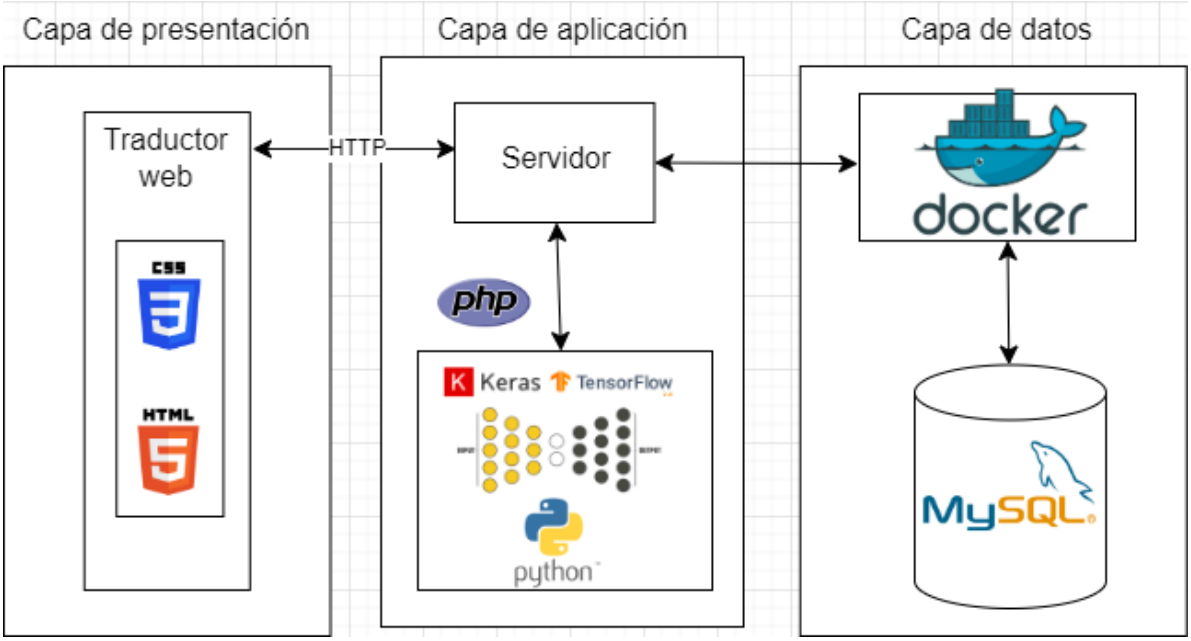

Figura 1. Propuesta de una Arquitectura para Traducción Automática.

4.1. Capa de Presentación

En la capa de presentación, se utiliza tecnología web para crear una interfaz de usuario que permita interactuar de manera intuitiva con el sistema de traducción automática. Esta capa se encarga de mostrar los resultados de la traducción y recibir las entradas del usuario (frases o palabras en español) que serán procesadas por el sistema. De acuerdo con [25] HTML es un lenguaje muy sencillo que permite describir hipertexto, es decir, texto presentado de forma estructurada y agradable, con vínculos o enlaces (*hyperlinks*) que conducen a otros documentos o fuentes de información relacionadas y con inserciones multimedia. Para [26] el CSS se emplea para estilizar y dar formato al contenido, asegurando una experiencia visual coherente y atractiva para los usuarios.

4.2. Capa de Aplicación

En la capa de aplicación utilizamos PHP (*Hypertext Preprocessor*) que actúa del lado del servidor y actúa como intermediario entre la capa de presentación y la capa de datos, en esta capa se complementa con el lenguaje de programación Python [27]. Según Python es un lenguaje de programación de propósito general creado por "Guido van Rossum", que es de alto nivel, fácil de aprender y dinámico [28] . En este sentido, Python procesa la lógica de negocio y las solicitudes del usuario, interactúa con la API de traducción y gestiona la comunicación con la base de datos. Por otra parte, para el desarrollo del *backend* se utiliza Django. Este es un *framework* de Python de alto nivel que fomenta un desarrollo rápido y un diseño limpio y pragmático. Para el desarrollo de la API de traducción se utiliza la librería de *Tensorflow* con *Keras*, por lo que se utiliza el modelo de red neuronal con arquitectura *Transformer* para realizar la TA [27].

4.3. Capa de Datos

En la capa de datos se almacena y gestiona la información necesaria para la TA. Se utiliza MySQL como sistema de gestión de bases de datos relacional, permitiendo almacenar y recuperar datos de manera eficiente [29]. Además, se emplea el sistema de archivos para almacenar modelos pre-entrenados, archivos de configuración y otros datos necesarios para el funcionamiento del sistema. Para ejecutar la base de datos dentro de un contenedor se utiliza Docker. Para [30] Docker es una tecnología permite crear aplicaciones en contenedores de software que son livianos, portátiles y autosuficientes. Los contenedores son un paquete de elementos que permiten crear un entorno en el que las aplicaciones se ejecutan independientemente del sistema operativo.

5. Experimentación

La arquitectura propuesta se utiliza para diseñar un sistema de TA del idioma español al idioma mixteco. Para ello, la red neural *Transformer* se entrena con un corpus extraído del libro titulado "Norma de Escritura del Tu'un Savi (Idioma mixteco)" [10].El corpus incluye oraciones seleccionadas por su riqueza gramatical, el cual la mayoría de los autores de investigaciones relacionadas recomiendan utilizar este tipo de oraciones. El modelo de red neuronal *Transformer* se entrena con el objetivo de potenciar su capacidad en la comprensión sintáctica y en la generación de textos en este idioma indígena. En la Figura 2, se presenta la arquitectura funcional del sistema de traducción automática.

En la Figura 2, se observa que se inicia con la entrada de una frase en idioma Español, que luego se *tokeniza* y se convierte en vectores de *embeddings* mediante un modelo como Word2Vec. El modelo Word2Vec es una familia de arquitecturas de modelos y optimizaciones que se pueden emplear para aprender incrustaciones de palabras a partir de grandes conjuntos de datos [31]. Por otra parte, los *embeddings* se procesan a través de un codificador *Transformer*, que utiliza múltiples capas de atención para capturar relaciones entre palabras. Luego, un decodificador *Transformer* genera la traducción en mixteco, también utilizando atención para alinear adecuadamente la traducción con la entrada. En la Figura 3, se presenta el flujo de trabajo y el diseño estructural del TA.

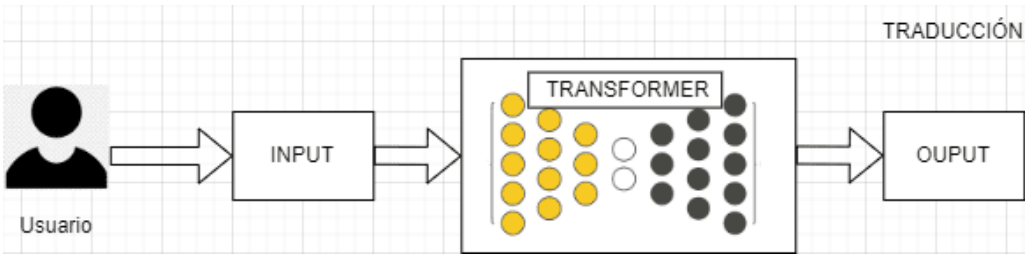

Figura 2. Arquitectura Funcional del Sistema de Traducción Automática.

En la Figura 3, se presenta una descripción general del proceso TA y los pasos involucrados en el desarrollo de un TA. La recopilación de datos de calidad, el preprocesamiento adecuado, la selección de técnicas de vectorización de palabras y algoritmos de aprendizaje automático, la validación rigurosa y la creación de una interfaz de usuario intuitiva son aspectos cruciales para lograr un traductor automático preciso y efectivo.

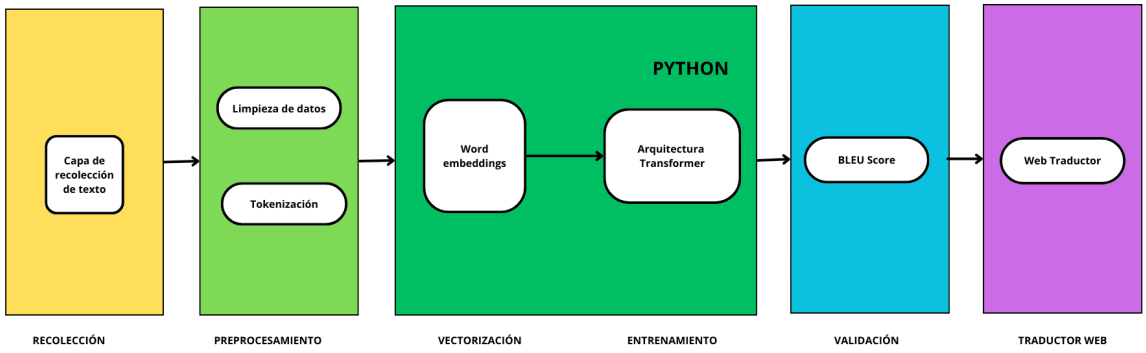

Figura 3. Flujo de Trabajo y el Diseño Estructural del Traductor Automático.

Con esta idea en mente, el flujo de trabajo del TA se divide en 5 etapas:

1. *Recolección de texto:* La primera etapa es recopilar el texto que se va a traducir. Los datos de entrenamiento para el TA pueden provenir de diversas fuentes, como libros, artículos, sitios web o incluso conversaciones habladas.

   Para este estudio, se preparó el corpus de conocimiento del idioma español al idioma mixteco con base a la información del libro [10]. Este libro proporciona una representación auténtica y unificada del idioma mixteco, incluyendo una variedad de textos que reflejan estructuras sintácticas complejas y reglas gramaticales específicas. La diversidad de géneros y estilos lingüísticos presentes permite al modelo entrenarse en diferentes contextos de uso del idioma, asegurando que pueda generar traducciones más

precisas y naturales en español y viceversa. Esto contribuye significativamente a mejorar la calidad de las traducciones y facilitar la comunicación efectiva entre hablantes de mixteco y otros idiomas.

2. *Preprocesamiento:* El preprocesamiento es la segunda etapa. En esta etapa, el texto se limpia y se prepara para su uso en el modelo de lenguaje neuronal. Esto puede implicar eliminar caracteres especiales, convertir todo el texto a minúsculas y *tokenizar* el texto en palabras individuales.
   Antes de entrenar el modelo, los datos de entrenamiento fueron debidamente limpiados y preprocesados. Esto implicó eliminar errores, corregir ortografía, estandarizar formatos y para este estudio el modelo fue entrenado con el formato de texto a nivel palabra. Es importante mencionar que el preprocesamiento de datos garantiza que el modelo reciba información consistente y de alta calidad, lo que mejora la precisión de la traducción
3. *Vectorización y Entrenamiento:* La etapa de vectorización significa convertir las palabras en representaciones numéricas que el modelo de lenguaje neuronal puede entender. Por otra parte, la etapa de entrenamiento implica que el modelo de lenguaje neuronal se entrena en un conjunto de datos de texto paralelo. Un conjunto de datos de texto paralelo es una colección de pares de oraciones, donde cada oración está en un idioma diferente. El modelo aprende a relacionar las palabras en un idioma con las palabras en el otro idioma.
4. *Validación:* La validación es la cuarta etapa en un sistema de traducción automática (TA). Este es un paso crucial para garantizar su precisión y confiabilidad. Entre los métodos de validación más utilizados se encuentra la evaluación de traducción automática bilateral (BiLingual Evaluation Understudy-BLEU, por sus siglas en inglés)
5. *Traducción:* La traducción es la etapa final. En esta etapa, el modelo de lenguaje neuronal se utiliza para traducir texto de un idioma a otro.

En relación con el diseño estructural del TA se puede decir que se compone de 3 capas:

- *Capa de entrada:* La capa de entrada recibe el texto en español que se va a traducir al mixteco.
- *Capa oculta:* La capa oculta es donde se realiza el trabajo real de traducción del español al mixteco. La capa oculta está compuesta por una red de neuronas artificiales que están interconectadas. Las neuronas procesan el texto de la capa de entrada y generan una representación del texto en el idioma de destino.
- *Capa de salida:* La capa de salida genera el texto traducido en mixteco.

6. Resultados

En este estudio, se diseñó y ajustó una red neuronal para la TA de idiomas indígenas con ERL, utilizando la red neuronal transformer. La red neuronal fue configurada con los siguientes parámetros - número de capas: 6, número de cabezas: 4, dimensionalidad de incrustación de entrada: 256, dimensionalidad de incrustación: 256, y tamaño de lote: 128. Estos parámetros fueron seleccionados basándose en investigaciones previas como [7], [10], donde se reportó que sus resultados fueron prometedoras al entrenar modelos con un corpus relativamente pequeño, comprendido entre 4,000 y 7,000 frases.

Es importante mencionar que, en este trabajo de investigación, el modelo fue entrenado con un corpus considerablemente reducido, compuesto por 1,055 unidades entre frases y palabras. Para evaluar el modelo entrenado, se seleccionaron 5 frases en español y su correspondiente traducción original proporcionada por un experto en el idioma mixteco. Luego, se utilizó el sistema de traducción automática (TA) para realizar la traducción del español al mixteco, y se calculó la calidad de la traducción utilizando la métrica BLEU.

A pesar de que la puntuación BLEU obtenida fue baja, es importante contextualizar estos resultados en comparación con otros sistemas de traducción automática que suelen entrenarse con millones de datos. Si bien la puntuación refleja las limitaciones del corpus, estos resultados son prometedores, ya que sugieren que, con un mayor volumen de datos, el rendimiento del modelo podría mejorar significativamente. Este hallazgo subraya la necesidad de ampliar nuestro corpus para futuros entrenamientos, lo que podría llevar a mejoras significativas en la calidad de las traducciones automáticas para idiomas indígenas como el mixteco.

Tabla 2 Evaluación de la traducción con la métrica BLEU.

Frase en idioma español	Traducción original de experto en idioma mixteco	Traducción automática neuronal en idioma mixteco	Puntuación BLEU
¡casa de la lluvia!	ve'e savi	ve'e	0
la flor blanca	ita kuiji	ita savi	0
hay muchas flores benditas	iyo kue'e ita ii	ita tina sí'ì válí	15.97
flor bonita del monte	ita vii yuku	ita	0
¡qué bonita está la flor que llevas!	¡anduu vii kaa ita ne'e un!	ita kuiji, ita ita yuku	6.87

Fuente: Elaboración propia.

En la Tabla 2, se observa que los resultados muestran que la mayoría de las traducciones tienen una puntuación BLEU muy baja, excepto una con una puntuación de alrededor de 16 y otra con alrededor de 7. Las puntuaciones BLEU bajas indican que las traducciones propuestas no coinciden bien con la traducción original del experto en idioma mixteco.

7. Discusión

La arquitectura propuesta para el traductor se compone de tres capas: presentación, aplicación y datos. Cada capa cumple una función específica y se integra con las demás para lograr un sistema modular, escalable y flexible.

La arquitectura propuesta presenta diversas ventajas:

- Modularidad y escalabilidad: Facilita el mantenimiento, la actualización y la expansión del sistema.
- Flexibilidad y adaptabilidad: Permite ajustar la solución a las necesidades específicas de cada lengua indígena.
- Aprovechamiento de tecnologías modernas: Utiliza herramientas confiables y ampliamente utilizadas.
- Abordaje de los desafíos de ERL: Facilita la implementación de estrategias para superar la escasez de recursos lingüísticos.

La selección de un corpus gramaticalmente rico fue una decisión estratégica basada en recomendaciones de investigaciones relacionadas, ya que se espera que la riqueza gramatical pueda compensar en cierta medida la escasez de datos. No obstante, los resultados con un puntaje muy bajo sugieren que, aunque la calidad del corpus es fundamental, la cantidad de datos sigue siendo un factor determinante. Además, dadas las limitaciones de recursos y la complejidad de traducir un idioma de bajo recurso como el español al mixteco, lograr puntajes de 16 y 7 puede ser un primer paso significativo en el desarrollo de un sistema de traducción automática para esta lengua.

8. Conclusiones

En este estudio, se presentó una arquitectura de TA para lenguas indígenas con ERL, el idioma que se utiliza para este caso de estudio es la lengua indígena mixteco con un corpus de conocimiento pequeño.

La arquitectura propuesta se basa en tecnologías modernas de TA para lenguas indígenas con ERL. Aunque el corpus seleccionado es gramaticalmente rico y contiene información precisa para la traducción del español al mixteco, el modelo muestra un gran potencial que puede ser optimizado. En este contexto, el modelo aun es limitado para realizar traducciones precisas, se prevé la ampliación del corpus disponible, lo cual permitirá entrenar de manera más efectiva la red neuronal *Transformer* y mejorar la precisión de las traducciones.

Estos hallazgos subrayan la importancia de disponer de un corpus más amplio para mejorar el rendimiento de los modelos de TA. Es importante mencionar que, aunque el corpus es gramaticalmente rico, la cantidad de datos

con las que fue entrenada afectó la capacidad del modelo para generar traducciones coherentes.

Se espera como trabajo futuro incorporar un sistema web colaborativo para que con el apoyo de expertos en el idioma mixteco se puede incrementar la base de conocimiento del idioma. De esta manera, se podrán integrar diferentes hablantes del idioma indígena, quienes podrán agregar nuevas frases y palabras al corpus. Así también, se pretende que este grupo de expertos corrija gramaticalmente las nuevas entradas, asegurando que cumplan con la gramática normalizada. Así también, como trabajo futuro está la propuesta de diseño de un algoritmo basado en reglas gramaticales del mixteco que mejore la calidad de las traducciones y extienda el corpus actual.

9. Referencias

[1] Schmelkes, S. (2022). El papel que ha jugado la escuela respecto a las culturas y las lenguas indígenas. En A. L. Gallardo, C. Rosa (Coord.) *Epistemologías e Interculturalidad en educación* (pp. 195-206). IISUE, UNAM. https://www.iisue.unam.mx/publicaciones/libros/epistemologias-e-interculturalidad-en-educacion

[2] Chakravarthi, B. R., Rani, P., Arcan, M., McCrae, J. P. (2021). A Survey of Orthographic Information in Machine Translation. *SN Computer Science, 2* (4), 1-19. https://doi.org/10.1007/s42979-021-00723-4

[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I. (2017). *Attention Is All You Need*. 31st Conference on Neural Information Processing Systems (NIPS). Long Beach California, USA. https://arxiv.org/pdf/1706.03762v5

[4] Mager, M., Meza, I. (2021). Retos en construcción de traductores automáticos para lenguas indígenas de México. *Digital Scholarship in the Humanities, 36*, (Supplement_1), i43–i48. https://doi.org/10.1093/llc/fqz093

[5] DOF. (2024). *Diario Oficial de la Federación*. https://www.dof.gob.mx/nota_detalle.php?codigo=5343116&fecha=30/04/2014#gsc.tab=0

[6] Xu, B. (2022). English-Chinese Machine Translation Based on Transfer Learning and Chinese-English Corpus. *Computational Intelligence and Neuroscience, 2022* (1), 1-10. https://doi.org/10.1155/2022/1563731

[7] Zacarías, Z., Meza, I. (2021). *Ayuuk-Spanish Neural Machine Translator.* https://github.com/anoidgit/yasa

[8] Feldman, I., Coto-Solano, R. (2020). *Neural Machine Translation Models with Back-Translation for the Extremely Low-Resource Indigenous Language Bribri*. 28th International Conference on Computational Linguistics. Stroudsburg, PA, USA. https://doi.org/10.18653/v1/2020.coling-main.351

[9] Huarcaya Taquiri, D. (2020). *Traducción automática neuronal para lengua nativa peruana* [Tesis de Grado]. Universidad Peruana Unión, Lima Perú. http://repositorio.upeu.edu.pe/handle/20.500.12840/4143

[10] INLI. (2022). *Norma de escritura del Tu'un Savi (idioma mixteco). Versión en español.* https://site.inali.gob.mx/INALIDhuchlab/assets/files/NormaTuunSavi-Espanol.pdf

[11] Blancarte Iturralde, O. J. (2020). *Introducción a la arquitectura de software. Un enfoque práctico.* https://pdfcoffee.com/introduccion-a-la-rquitectura-de-softwarepdf-2-pdf-free.html

[12] Le, N. T., Sadat, F. (2020). *Revitalization of Indigenous Languages through Pre-processing and Neural Machine Translation: The case of Inuktitut.* https://github.com/huggingface/transformers

[13] Baker, T., Smith, L., Anissa, N. (2019). *Educ-AI-tion rebooted? Exploring the future of artificialx intelligence in schools and colleges.* https://media.nesta.org.uk/documents/Future_of_AI_and_education_v5_WEB.pdf

[14] López Takeyas, B. (2007). *Introduccion a la Inteligencia*. https://nlaredo.tecnm.mx/takeyas/Articulos/Inteligencia%20Artificial/ARTICULO%20Introduccion%20a%20la%20Inteligencia%20Artificial.pdf

[15] Casacuberta Nolla, F., Peris Abril, Á. (2017). Traducción automática neuronal. *Revista Tradumàtica: tecnologies de la traducció*, (15), 66-74. https://doi.org/10.5565/rev/tradumatica.203

[16] Thinh Nguyen, T. (2019). *Machine Translation with Transformers* [Tesis de Maestría]. Universit¨at Stuttgart, Alemania. https://elib.uni-stuttgart.de/bitstream/11682/10621/1/thesis.pdf

[17] Barcena, M. E., Gamal Othman, M. (2017). *Traducción Automática y Asistida Por Ordenador*. https://www.researchgate.net/publication/323273904

[18] Microsoft. (2024*). ¿Qué es una puntuación BLEU?* https://learn.microsoft.com/es-es/azure/ai-services/translator/custom-translator/concepts/bleu-score

[19]Urban, E. (2024). *Desafío de aptitudes de IA.* https://learn.microsoft.com/es-es/azure/ai-services/translator/custom-translator/how-to/test-your-model

[20]Basit Andrabi, S. A., Wahid, A. (2022). Machine Translation System Using Deep Learning for English to Urdu. Computional Intelligence Neuroscience, 2022, 1-11. https://doi.org/10.1155/2022/7873012

[21]Moreno Cabrera, J. C. (2014). El español hablado como lengua aglutinante y polisintética. https://www.academia.edu/31414173/EL_ESPA%C3%91OL_HABLADO_COMO_LENGUA_AGLUTINANTE_Y_POLISINT%C3%89TICA

[22]Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohungbe, T., Oluwole Akinola, S., Muhammad, S., Kabongo Kabenamualu, S., Osei, S., Sackey, F., Niyongabo, R. A., Macharm, R., Ogayo, P., Ahia, O., Berhe, M. M., Adeyemi, M., Mokgesi-Selinga, M., Okegbemi, L., Martinus, L., et al. (2020). Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages. In T. Cohn, Y. He, Y. Liu (Eds.) *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 2144-2160). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.195

[23]Knowles, R., Stewart, D., Larkin, S., Littell, P. (2021). *NRC-CNRC Machine Translation Systems for the 2021 AmericasNLP Shared Task*. First Workshop on Natural Language Processing for Indigenous Languages of the Americas. Stroudsburg, PA, USA. https://doi.org/10.18653/v1/2021.americasnlp-1.25

[24]Moreno Veliz, O. (2021). *The REPU CS' Spanish–Quechua Submission to the AmericasNLP 2021 Shared Task on Open Machine Translation.* 1st Workshop on Natural Language Processing for Indigenous Languages of the Americas. Ciudad de México. https://doi.org/10.18653/v1/2021.americasnlp-1.27

[25]Casado-Vara, R. (2019). Introducción a HTML. En C. Pinzón Trejos (Ed.) *Knowledge extraction and representation* (pp. 279-506). Ediciones Universidad de Salamanca. http://hdl.handle.net/10366/139647

[26]Gather Consultores. (2023). *CSS: Diseñando el mundo web con estilo*. https://www.linkedin.com/pulse/css-dise%C3%B1ando-el-mundo-web-con-estilo-gather-consultores-pwhue/

[27]php NET. (2024). *¿Qués e PHP?* https://www.php.net/manual/es/intro-whatis.php

[28]Rawat, A. (2020). A Review on Python Programming. *International Journal of Research in Engineering, Science and Management, 3* (12), 8–11, https://journal.ijresm.com/index.php/ijresm/article/view/395

[29]Arias, A. (2015). *Bases de Datos con MySQL* (2da Ed.). Createspace Independent Publishing Platform.

[30]Ponsico Martin, P. (2017). *Tecnología de Contenedores Docker* [Tesis de Grado]. Escola Tècnica d'Enginyeria de Telecomunicació de Barcelona. Universitat Politècnica de Catalunya. Barcelona, España. https://upcommons.upc.edu/handle/2117/113040

[31]TensorFlow. (2022). *Incrustaciones de palabras*. https://www.tensorflow.org/text/guide/word_embeddings?hl=es-419

