Sistema intérprete automático neuronal mazahua del norte del Estado de México a español

Lizeth De La Cruz Piña Autor A, Juan Alberto Antonio Velázquez Autor B, Ivan Vladimir Meza Ruiz Autor C, Erika López González Autor D y Leopoldo Gil Antonio Autor E.

[a,b,d y e]Tecnológico de Estudios Superiores de Jocotitlán, Carretera Toluca-Atlacomulco km 44.8, Ejido de San Juan y San Agustin, Jocotitlán, 2017150480275@tesjo.edu.mx, juan.antonio@tesjo.edu.mx, erika.gonzalez@tesjo.edu.mx y leopoldo.gil@tesjo.edu.mx Estado de México.
[c]Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autonóma de México ivanvladimir@turing.iimas.unam.mx, Ciudad de México.

Resumen

En este artículo se presenta la creación del sistema de traducción automática neuronal para *lengua Mazahua* al español con variante del oriente, lengua hablada en el norte del Estado de México y en el Estado de Michoacán. Esta variante es conocida por los hablantes como *jñatrjo*. Las pruebas de traducción fueron realizadas del español al mazahua y en la creación del *corpus*, se utilizaron diferentes fuentes escritas, posteriormente digitalizadas e indispensables para la traducción automática. El *corpus* fue constituido por 8,000 mil frases estimadas de bajos recursos donde los datos fueron normalizados de manera automática. Posteriormente el sistema planteado se retroalimenta en la arquitectura neuronal *Transformer* y usa la *tokenización* a grado de subpalabras como acceso. Se expone el presente desempeño dado los recursos que fueron recolectados, dando como resultado de un *BLEU* de 0.60.

Palabras clave—*Lengua Mazahua, Traducción automática, Corpus, Transformer, Tokens, Subpalabras, BLEU.*

Abstract

In this work present to the creation of the neural automatic translation system for the Mazahua language into Spanish with an eastern variant, a language spoken in the north of the State of Mexico and Michoacán. This language variant is known to speakers as jñatrjo. The translation tests were carried out from Spanish to Mazahua language and in the creation of the corpus different written sources were used, later digitized and essential for automatic translation. The corpus was made up of 8,000 thousand keywords, estimated low-income sentences where the data was automatically normalized. Subsequently, the proposed system feeds back into the Transformer neural architecture and uses subword-grade tokenization as access. The present performance is exposed given the resources that were collected, thus resulting in a BLEU of up to .60

Keywords*—Mazahua language, Machine translation, Corpus, Transformers, Tokens,Subwords, BLEU.*

1. INTRODUCCIÓN

Este trabajo presenta un traductor automático básico entre las lenguas español (*Jñangicha*), a *Mazahua* (*Jñatrjo, Mazahua* de oriente) por medio del procesamiento de lenguaje natural (*PLN* o por sus siglas en inglés *NLP*).

La identificación del nombre *Mazahua*, fue dado históricamente por la población mexicana, que reconoce a un grupo indígena y también a su lengua. Dicho nombre es la forma castellanizada de *Mazahua*, ya que en el idioma *náhuatl* significa "*poseedores de venados*". Los hablantes de la *lengua Mazahua* suelen llamar a su lengua *Jñatjo*, lo cual tiene como significado lengua. El idioma *Mazahua* tiene dos variantes, el *Mazahua* del oriente y el de occidente. Al *Mazahua* de oriente, sus hablantes lo conocen como *Jñatrjo*, y al *Mazahua* de occidente le dicen *Jñatjo*. Ambos se hablan en los estados de Michoacán y el Estado de México, cabe mencionar que esta agrupación lingüística pertenece a la familia *oto-mangue*. Sus lenguas o idiomas hermanos más cercanos son el *Otomí*, el *Matlatzinca* y el *Tlahuica*. De todas las lenguas indígenas de la República Mexicana, el idioma *Mazahua* es un idioma hablado en una población de 13,600 y otro de 700,000 personas aproximadamente según el Sistema de Información cultural (SIC) [1].

Aunque en la actualidad el número de hablantes de la lengua indígena *Mazahua* se ha reducido en más de 12% en los últimos 15 años según cifras del Instituto Nacional de Estadística, Geografía e Informática (INEGI). Sin embargo, los datos del INEGI indican que un alto porcentaje de estas personas sólo son oyentes [2].

Dada la información anterior y pese al gran trabajo de las Universidades con enfoque cultural en la investigación y difusión del idioma; actualmente, solo existen diccionarios *español–mazahua* y libros de aprendizaje didáctico. Sin embargo, no existe un traductor automático que sea basado en el *PNL* de dicha lengua. Por ello se considera pertinente implementar una herramienta apoyada de la técnica *PNL* (*Programación Neurolingüística*), para el lenguaje *Mazahua*, para lo cual se investigó y trabajó con la variante *Mazahua* del oriente (*jñatrjo*) perteneciente al Norte del Estado de México.

Como parte de los antecedentes de investigación y de la revisión literaria, se revisaron los pasos del proyecto centrado en la traducción *Ayuuk-español* [3]. En la etapa de tokenización, se utilizó la biblioteca *subword-nmt6* [4] y para el entrenamiento de nuestros modelos se utilizó el *framework JoeyNMT7* [5], esta es una librería de *python* utilizada para los modelos de entrenamiento; al final con estas dos herramientas se desarrolló el código base que se puede consultar en línea excepto la parte del *corpus* debido a la restricción existente por derechos de autor.

2. TRABAJOS RELACIONADOS A LA TRADUCCIÓN AUTOMÁTICA CON UNA RED NEURONAL

En esta sección se describe la literatura analizada con base a la *traducción automática* relacionada el lenguaje, por lo que a continuación se muestran un conjunto de artículos y

trabajos analizados de varias investigaciones, y como resumen se menciona en el resumen que no se encontró ningún traductor automático de la *lengua Mazahua* de occidente *Jñatjo*, por lo cual a continuación se menciona el compendio de la literatura útil para el desarrollo de este proyecto.

En [3], se realiza la traducción automática neuronal *Ayuuk-español*, donde presentan el primer sistema de traducción automática neuronal para la lengua *Ayuuk*. En los experimentos tradujeron del *Ayuuk* al español y del español al *Ayuuk*. Esta es una lengua hablada en el Estado de Oaxaca (México), por el pueblo *Syuukjä'äy* (en español comúnmente conocido como *Mixes*). Utilizaron diferentes fuentes para crear un *corpus* paralelo de bajos recursos, más de 6000 frases. Para algunos de estos recursos se basaron en la alineación automática. El sistema propuesto se basa en la arquitectura neural *Transformer* y utiliza como entrada la tokenización a nivel de subpalabra. Muestran el rendimiento actual dado los recursos que hemos recogido para la variante de San Juan Güichicovi, son prometedores, hasta 5 *BLEU*. Cabe mencionar que su desarrollo se basó en el proyecto *Masakhane* para las lenguas africanas.

En [6], crearon un traductor automático híbrido *wixarika-español* con escasos recursos bilingües de las lenguas español y *wixarika*, también conocida como *Huichol*, por medio del Procesamiento de Lenguaje Natural. La lengua *wixarika* es importante como lengua indígena ya que es hablada en los Estados de Jalisco, Nayarit, Zacatecas y Durango, y tiene un aproximado de treinta y cincuenta mil hablantes. Para esto se usa el modelo de *Traducción Estadística por Frases* para resolver el problema, se creó un analizador y segmentador morfológico que permite la separación de las palabras aglutinadas *wixaritari* en morfemas, lo cual permite trabajar con la polisíntesis del idioma. También se escribieron herramientas básicas para el procesamiento de lenguaje natural, como es un normalizador, para lo que se estableció un alfabeto base del idioma y un *tokenizador*. Al final, los resultados obtenidos son buenos al compararlos con otros trabajos de traducción, tomando en cuenta la distancia entre lenguas traducidas y los escasos recursos con los que se cuenta.

Además, en [7], desarrollaron un traductor automático *español–Purépecha*, usando la herramienta *OpenNMSEl* traductor usa herramientas basadas en Procesamiento de Lenguaje Natural mediante redes neuronales utilizando el software *OpenNMT*. Para el desarrollo de tal proyecto se usa un *corpus* de 804 frases emparejadas entre los dos idiomas para el entrenamiento de las redes neuronales. Contemplando que el tamaño del *corpus* es demasiado limitado, este trabajo se basa en palabras simples, que aun que se pueda ingresar una frase en español muy compleja, la salida como traducción no será nada similar a su traducción verdadera. A su vez este proyecto contempla la elaboración de una interfaz para que cualquier persona pueda ingresar y hacer sus predicciones de traducciones.

En el proyecto [8], se realizó un traductor morfológico del Castellano al *Quechua* y viceversa, basado en la transferencia, que opera en tres fases (análisis, transferencia y generación), usando representaciones morfológicas para las palabras. Al traducir una palabra, el sistema no sólo devuelve la palabra convertida al otro idioma, sino también, muestra información lingüística de los componentes de la palabra. Para su desarrollo usaron herramientas de código abierto como Java, MySQL y Apache. El proyecto fue alojado en la Web, para que pueda ser fácilmente visitada y utilizada desde cualquier parte del mundo y por los usuarios que así lo deseen. Este sistema quedó también como prototipo para la optimización y el desarrollo de un nuevo sistema para la traducción de frases, oraciones y posteriormente textos.

En el trabajo [9], presentan el desarrollo de un *corpus* paralelo *Guaraní-Español*, con alineación a nivel de frase, las frases del *corpus* son en *Guaraní*, en donde el *corpus* es utilizado dentro del dialecto *Guaraní Jopara*; este dialecto del *Guaraní* es hablado en Paraguay, del cual se basa en la gramática *Guaraní*, pero puede incluir varias palabras del español. Salvo algunas excepciones, el idioma *Guaraní* permanece en gran medida inexplorado en los campos del Procesamiento del Lenguaje Natural y la Lingüística Computacional. Para la primera evaluación, se utilizó una pequeña muestra de 20 documentos (unas 150 frases). A partir de esta pequeña muestra el proceso de evaluación determinó que el 64,0% de los pares eran una coincidencia total, en el 25,5% de los casos había más información en el lado español, en el 4,6% de los casos había más información en el lado guaraní y en el 5,9% de los casos las frases no coincidían. Al final presentaron el desarrollo de un *corpus* paralelo de texto *Guaraní-Español* con alineación a nivel de frase. El *corpus* es de tamaño medio, conteniendo alrededor de 228.000 *tokens* en guaraní junto a los correspondientes 336,000 *tokens*.

1. Especificaciones del traductor Mazahua-Español

La lengua *Mazahua*, la cual está revisada y estandarizada por ISO 639-3 [10], es una lengua que se habla en el centro de México. Sus hablantes denominan a la lengua con el nombre de *jñatrjo*. El área históricamente ocupada por los *Mazahuas* se localiza en el Altiplano Central. Gran parte de los hablantes se ubican en la parte noroccidental y centro-occidental del Estado de México; en general se incluyen 15 municipios en los estados de Michoacán y el Estado de México. También al idioma *Mazahua* de oriente sus hablantes lo conocen como *jñatrjo*, y al idioma *Mazahua* de occidente le dicen *Jñatjo*. El idioma *Mazahua* pertenece al grupo lingüístico *oto-mangue*, de donde se deriva el tronco *otopame*, a la que pertenece la familia *otomí-mazahua*. Al igual que el *otomí* tiene artículo definido (*nu- 'el, la'*) e indefinido (*na- 'un, una'*), cuya forma en plural es común a ambos (*yo- 'los, las, unos, unas'*). Los nombres no distinguen normalmente singular de plural, aunque las formas poseídas de los mismos pueden distinguir si el poseedor es singular o plural, por ejemplo:

xin-ꞇumwɨ 'mi casa, mis casas'
xin-ꞇumwɨ-hi 'nuestra casa, nuestras casas'

En el verbo en cambio se distinguen tres números gramaticales: singular, plural y dual. El idioma *Mazahua* presenta un alineamiento nominativo-acusativo; el sujeto de

una oración transitiva y el sujeto de una oración intransitiva y se marcan de manera similar.
En la situación del idioma español, nuestro sistema crea traducciones al español mexicano, que forma parte de la variante del español latinoamericano, y con esto identificamos la lengua por el código ISO-639-1 [11].

3. IMPLEMENTACIÓN DEL TRADUCTOR

3.1 Desarrollo del *Corpus*

Para la creación del *corpus*, se realizó una investigación y trabajo de campo necesario para la recolección de documentos, especialmente bibliografía con traducciones disponibles en *Mazahua* con la variante del oriente a español, lo cual resultó complicado ya que existe un grave problema de escasez de datos que contenga texto con la variante del oriente, ver Tabla 1.

Tabla 1 Textos recolectados.

Titulo	*Autor*	*Tipo*	*Copyright*	*Tipo*	*Fuente*
El eterno retorno/Nupamapama nzhogú	Francisco Antonio León Cuervo	Libro	X	Digital	Mi universo mazahua
Revista electrónica/Ñu JñiñiJñatjo V.1, V.2, V3, V4, V5, V6.	Francisco Antonio León Cuervo	Revista	X	Digital	Mi universo mazahua
¿Sabías cuál fue el antecedente de la fundación del primer barrio de la cabecera Ixtlahuaca?	L.H. Sergio López Alcántara. Cronista Municipal de Ixtlahuaca	Articulo	X	Digital	Ayuntamiento Ixtlahuaca
Evolución arquitectónica de la casa de cultura "Químico José Donaciano Morales y Mier Altamirano"	L.H. Sergio López Alcántara. Cronista Municipal de Ixtlahuaca	Articulo	X	Digital	Ayuntamiento Ixtlahuaca
Santo Domingo v San Bartolomé: dos pueblos originarios del municipio de Ixtlahuaca vistos a través de su tradición indígena	L.H. Sergio López Alcántara. Cronista Municipal de Ixtlahuaca	Articulo	X	Digital	Ayuntamiento Ixtlahuaca
¿Sabías cómo fue la postura de los ixtlahuaquenses, durante los primeros años del movimiento armado de Independencia?	L.H. Sergio López Alcántara. Cronista Municipal de Ixtlahuaca	Articulo	X	Digital	Ayuntamiento Ixtlahuaca
La catedral mazahua de Ixtlahuaca: desde sus elementos sociales, religiosos y arquitectónicos. Siglo XVI-XVIII"	L.H. Sergio López Alcántara. Cronista Municipal de Ixtlahuaca	Articulo	X	Digital	Ayuntamiento Ixtlahuaca
El municipio de Ixtlahuaca, rumbo al bicentenario de vida institucional	L.H. Sergio López Alcántara. Cronista Municipal de Ixtlahuaca	Articulo	X	Digital	Ayuntamiento Ixtlahuaca
Ley de acceso de las mujeres a una vida libre de violencia del Estado de México	Ramón Gerardo Martínez Sánchez	Libro	X	físico	Ayuntamiento Ixtlahuaca
Diccionario mazahua-español y español-mazahua	Florencia Jiménez Berriozábal	Libro	X	físico	Ayuntamiento Ixtlahuaca
La cenicienta, cuento infantil traducido.	Charles Perrault/traducción: Carmen Sánchez	Cuento infantil	X	físico	División de lengua y cultura UIEM
Caperucita roja, cuento infantil traducido.	Hermanos Grimm/ traducción: Carmen Sánchez	Cuento infantil	X	físico	División de lengua y cultura UIEM
Aladino y la lámpara maravillosa, cuento infantil traducido.	Traducción: Carmen Sánchez	Cuento infantil	X	físico	División de lengua y cultura UIEM
Vocabulario mazahua-español y español-mazahua	Mildred Kiemele Muro	Libro	X	físico	Biblioteca UIEM

3.2 Preprocesamiento y alineación de los datos

En el desarrollo del *corpus*, se pudo recolectar documentos en dos formatos: digital y físico, sin embargo, la mayoría de los textos recolectados fueron físicos, por lo cual los textos tuvieron que digitalizarse. Por eso seleccionaron todos los libros y textos que cumplieran con la traducción de español a mazahua, por lo cual fue difícil conseguir textos con esas características ya que algunos solo se encuentran escritos en mazahua, pero no traducidos a español.
Para digitalizar todos los textos recopilados, se recurrieron a aplicaciones que utilizan el reconocimiento óptico de caracteres (OCR por sus siglas en inglés), donde fue necesario escanear cada página de los libros y posteriormente el texto se pueda almacenar en formato *pdf*, y así hacer el reconocimiento de caracteres. Pero en algunos casos el reconocimiento no fue posible, ya que, al utilizarse este mecanismo, influyó en un factor que derivó en un mal reconocimiento o la inefectividad del algoritmo OCR implementado; no debido al sistema sino a causa de una tipografía extraña, donde fue necesario normalizar la ortografía y algunas palabras, que en este caso se optó por sustituirlos a vocales sin algún tipo de símbolo especial (ver Tabla 2).

Tabla 2 Sustitución de caracteres correctos por caracteres con tipografía extraña.

Carácter	Sustitución
a̸	a
e̸	e
o̸	o
u̸	u
a̲	a
o̲	o
u̲	u

Las traducciones en su mayoría no estaban alineadas, ya que además tenían otro inconveniente, ya que, al ser demasiado extenso, fue necesario recurrir con separarlos a partir de puntos o comas y con esto determinar un inicio y un fin de cada frase. En otras traducciones recolectadas las frases ya permanecen alineados por defecto, y después deducir que la línea capturada corresponde a una frase. Los textos presentados ya normalizados, generaron dos archivos *archivo.es* (el cual contiene frases traducidas en español) y *archivo.maz* (que contiene frases traducidas en *Mazahua*). Cabe mencionar que este trabajo se basa de la propuesta realizada por [3]. Posterior a la normalización se descartan todas las alineaciones donde no hay texto o que estén vacías o dobles. Al final, se dividieron aleatoriamente las oraciones en conjuntos de entrenamiento, desarrollo y prueba.

Para configurar este conjunto de datos, es importante elegir los conjuntos en una distribución adecuada. El tamaño total del *corpus*, es decir el total de frases traducidas al español y mazahua fue de un total de 8,928 palabras. Se decidió colocar un 70% de frases en el *archivo.train*, obteniendo un total de 6428 frases, y dividir el restante en dos porciones equivalentes al 15% para los *archivos.dev* y *archivos.test* obteniendo así un total de 1250 frases para cada uno de los archivos (ver Tabla 3).

Tabla 3 Descripción general del Corpus.

	Frases	**Palabras**		**Caracteres**	
		Español	Mazahua	Español	mazahua
.dev	1250	9000	7617	58328	41575
.train	6428	41308	42060	249856	221589
.test	1250	7705	7126	48202	37742
Total	8,928	58,013	56,803	356,386	300,906

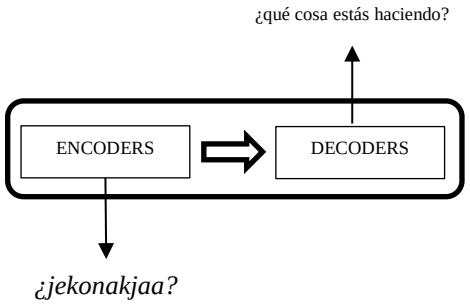
4 ARQUITECTURA NEURONAL

Nuestro modelo de traducción se basa en la arquitectura *Transformers* [12], el modelo es un tipo de arquitectura de redes neuronales que se usan en modelos basados en el lenguaje (*NLP*). Un transformador, en la situación de un sistema de traducción, transformará una sentencia redactada de un lenguaje a otro lenguaje. Partiendo de un ingreso y una salida (la deseada) y viendo al final a la arquitectura *transformer* como una caja negra.

Ilustración 1 EjemploTransformador

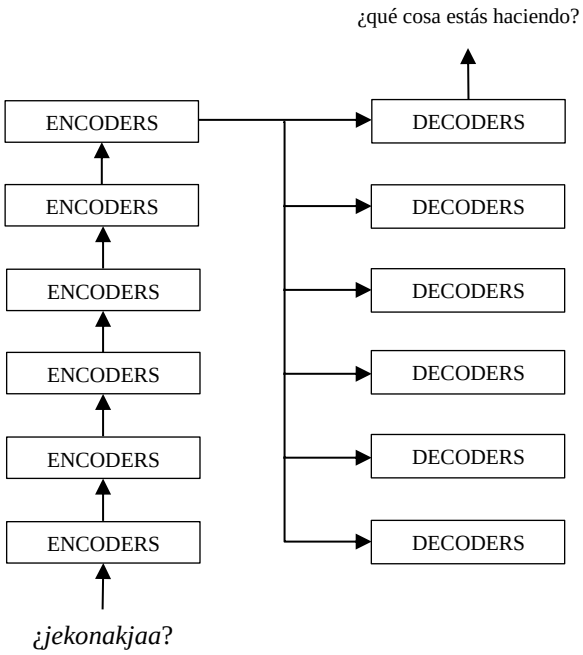
Si se colocan dentro de la caja negra lo que veríamos es un componente de encoders y otro de decoders:

Ilustración 2 Ejemplo codificador-decodificador



Tanto el componente de *encoders* como el de *decoders* es una pila de codificadores y decodificadores (el mismo número en ambos casos):

Ilustración 3 Pila codificadores-decodificadores



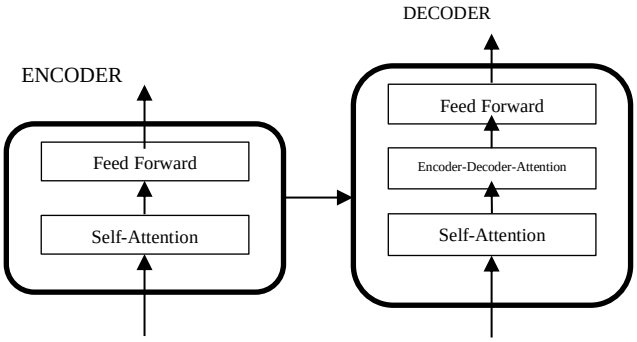
Los codificadores poseen una composición continuamente igual y los decodificadores realizan la transformación. Los primeros se parten en 2 capas que son:

- Capa de atención: ayuda al encoder a colocar la atención en las palabras que son importantes.
- *Feed forward neural network*: la salida de la autoatención se pasará a una red neuronal *feedforward* completamente conectada.

Los decodificadores se parten en 3 capas:

- Capa de atención: El número de parámetros de la red neuronal feedforward para cada codificador es el mismo, pero sus funciones son independientes.
- *Encoder/Decoder Atention*: ayuda al decodificador a centrarse en las piezas más importantes de la oración.
- *Feed forward*

Ilustración 4 Capas codificadores-decodificadores



A continuación, se presenta la Tabla 4, donde se utilizaron las cinco configuraciones diferentes basadas en el método *codificador-decodificador*.

Tabla 4 Configuración codificador-decodificador.

	Número capas	Número cabezales	Dimensión embedding entrada	Dimensión embedding	Tamaño lote
Configuración 1	3	4	64	64	128
Configuración 2	6	4	256	256	128
Configuración 3	6	4	256	256	32
Configuración 4	6	4	256	256	512
Configuración 5	2	2	64	64	128

5. EXPERIMENTOS Y RESULTADOS

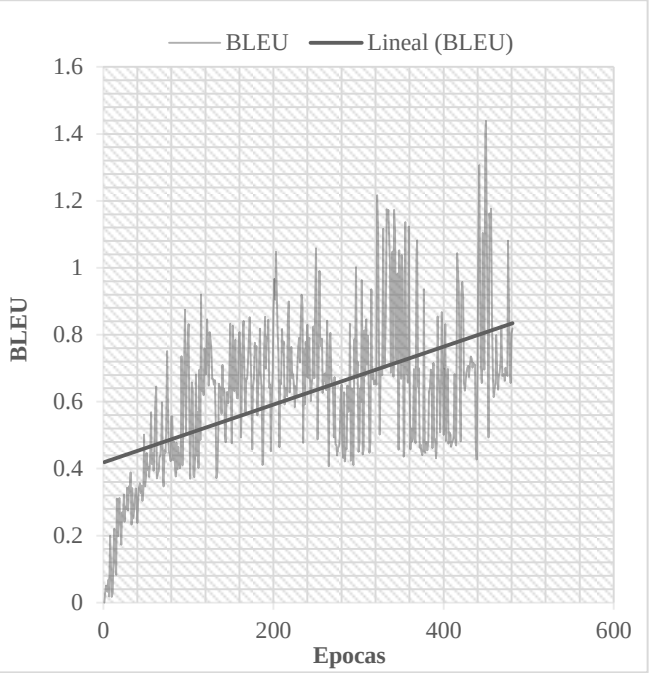
Según lo mencionado en la sección anterior también se modificó lo siguiente:

a) Épocas con un tamaño de 100 y 150.
b) El vocabulario del algoritmo de subpalabras *BPE* (2000 a 4000).
c) La longitud máxima frases (50 a80).

Estos modelos se entrenaron en una computadora de escritorio con las siguientes características: tarjeta gráfica *GeForce MX130*, *Driver Versión*: 460.73.01, *CUDA* Versión: 11.2

Posterior a la experimentación, se muestra la puntuación *BLEU* la cual fue de 0.69 dado el conjunto de desarrollo, prueba y entrenamiento hacia la traducción obtenida en el idioma español para la traducción y la curva de aprendizaje donde se presenta un rendimiento muy bajo, sin embargo, se tiene más ganancia en el modelo con más capas ocultas, ya que se cuenta con una pequeña cantidad de datos de entrenamiento. Por otro lado, la división como se esperaba muestra que es muy díficil de traducir, ya que las puntuaciones *BLEU* son mínimas. También se observó que con la configuración uno fue más fácil traducir al español con un total de cien épocas y un tiempo aproximado de doce horas. En las configuraciones de la dos a la cinco fueron menores las puntuaciones.

Ilustración 5 Puntuaciones BLEU transformer



Cabe mencionar que, para este modelo no encontramos ningún *corpus* o dato de entrenamiento alineado a frases. No obstante, nuestro trabajo muestra que un modelo con base en la arquitectura *Transformer* y con una configuración de recursos drásticamente baja debido a lo anteriormente mencionado; pero se espera que con un *corpus* más robusto se puedan esperar resultados prometedores. Todavía son bajos para los estándares, como se puede mostrar en la Tabla 6 con frases candidatas y los posibles objetivos generados por el traductor.

Ilustración 6 Traducciones generados por el traductor.

Candidato	**Objetivo**
¿mbekagokjogutsu?	¿qué te pasó?
ma mu jiasu	cuando amaneció
dyagipesinzhubu	no te preocupes
¿pjekogoxitsi?	¿qué te dijo?
bexkutjora mago	yo voy a ir aquí cerca
gonzhodukatsike	caminó un ratito
ka mi manojo	que era muy grande
ja mama gajanu	así lo cuentan

4. CONCLUSIONES Y RECOMENDACIONES

En este trabajo se presenta el sistema de traducción automática de textos para la *lengua Mazahua* variante del oriente al español. Donde la conjetura es que los sistemas de traducción automático, con base en redes neuronales son lo suficientemente adecuados para producir traducciones de las lenguas indígenas al español y viceversa. No obstante, codificar un sistema de *Traducción Automática* en un

escenario con escasos recursos (como desarrollar el *corpus*) y sin contar con los recursos, fueron una limitante computacional para llevar a cabo una traducción eficiente en esta lengua. Por esto se involucran diversas fases de desarrollo, a partir de la recolección de datos hasta llegar al entrenamiento del sistema.

La fase de trabajo de recolección de textos ha sido extenuante, debido a que hay escases de información de textos digitales y las traducciones resultan limitadas, por lo cual ha sido uno de los primeros inconvenientes que se encontraron en la investigación. Esta situación llevó a realizar una búsqueda de información en diversas instituciones, encontrando de esta forma un enorme desafío ya que ha sido difícil buscar textos alineados a las frases del idioma *Mazahua*. Con estos datos ya proporcionados por las distintas instituciones que se mencionan en la sección 3, se digitalizó la información, debido a que la mayor parte de datos recabados fueron especialmente en libros y esto ha sido un problema gracias a la gramática que se usa en la lengua *Mazahua* ya que contiene caracteres especiales en las vocales, ya que al utilizar herramientas de reconocimiento óptico de caracteres no se reconocen las frases y con esto el tiempo de captura de texto demoró el proceso.

Otro problema ha sido que la mayor parte del texto recolectado está protegida por derechos de autor, pese ello, con este trabajo se busca fomentar que la construcción de obras futuras sea abierta y con ello conseguir que personas hablantes y no hablantes interesados en la lengua *Mazahua* puedan obtener la información sin restricciones, con lo cual los investigadores en lenguaje natural y que investiguen la lengua *Mazahua*, no tengan problemas al crear el *corpus*.

Posteriormente el *corpus* fue normalizado para poder detectar ciertas frases y poder dividirlas con signos de puntuación y así al sistema de entrenamiento no le causara algún problema. De esta forma se trató de que la normalización quedara equitativa en los tres archivos de entrenamiento con un total de 8000 frases alineadas repartidas en los archivos de entrenamiento en un setenta por ciento para *train* y un treinta por ciento para *dev* y *test*; de las cuales se pudiesen obtener más frases ya que hay algunas frases que son extensas.

Por otro lado, este trabajo muestra que un modelo estándar con base en la arquitectura *Transformer* y con una configuración con recursos extremadamente bajos, obtiene resultados escasos, aún para los estándares normales en el campo de la traducción automática. A pesar de ello, los puntajes BLEU logrados son bajos pero eficientes, pero para trabajo futuro se espera que el *corpus* sea más amplio y así obtener resultados prometedores. En resumen, las principales contribuciones de esta investigación son la recolección, digitalización y alineación de textos de diferentes fuentes y con ello la creación de *corpus* para el entrenamiento pueda ser usado en futuros trabajos.

5. AGRADECIMIENTOS

Agradecemos a la Universidad Intercultural del Estado de México (IUEM), por el acceso a libros y textos proporcionados por su biblioteca y la división de Lengua y Cultura, así mismo agradecemos el proyecto “Traducción automática para lenguas indígenas de México” PAPIIT-IA104420, UNAM.

6. REFERENCIAS

[1] S. d. C. d. I. Cultural, «Mazahua,» 2020. [En línea]. Available: http://sic.gob.mx/ficha.php?table=inali_li&table_id=47. [Último acceso: 4 Octubre 2021].

[2] Inah.gob.mx, «Se pierde lengua Mazahua,» 4 Julio 2008. [En línea]. Available: https://www.inah.gob.mx/boletines/1519-se-pierde-lengua-mazahua . [Último acceso: 1 Octubre 2021].

[3] I. M. D. Zacarías, «Ayuuk-Spanish Neural Machine Translator,» 2021. [En línea]. Available: https://aclanthology.org/2021.americasnlp-1.19.pdf. [Último acceso: 1 Octubre 2021].

[4] B. H. A. B. Rico Sennrich, «Neural Machine Translation of Rare Words with Subword Units,» Association for Computational Linguistics, [En línea]. Available: https://aclanthology.org/P16-1162/. [Último acceso: 1 Octubre 2021].

[5] J. B. S. R. Julia Kreutzer, «Joey NMT: A Minimalist NMT Toolkit for Novices,» Association for Computational Linguistics, November 2019. [En línea]. Available: https://aclanthology.org/D19-3019. [Último acceso: 3 October 2021].

[6] I. V. M. R. Jesús Manuel Mager Hois, «Traductor híbrido wixarika-español con escasos recursos bilingües,» Febrero 2017. [En línea]. Available: http://code.kiutz.com/tesism/tesis.pdf. [Último acceso: 4 Octubre 2021].

[7] I. V. M. R. Miguel Salvador Soriano Garcia, «Traductor automático español–purépecha mediante openNMT,» Marzo 2018. [En línea]. Available: https://www.dropbox.com/s/0t1lxyasof5wqeb/soriano18.pdf?dl=0 . [Último acceso: 2021 Octubre 2].

[8] J. F. M. Indhira Castro Cavero, «Traductor morfológico del castellano y quechual,» [En línea]. Available: https://app.tecsup.edu.pe/file/sga/documentos/revistaIi/Ii_1/6.pdf. [Último acceso: 2021 Octubre 2].

[9] P. A. A. R. G. G. L. Luis Chiruzzo, «Development of a Guarani - Spanish Parallel Corpus,» European Language Resources Association, May 2020. [En línea]. Available: https://aclanthology.org/2020.lrec-1.320. [Último acceso: 2021 October 2].

[10] A. Mora-Bustos, «Construcciones escindidas en mazahua (otomangue),» December 2019. [En línea]. Available: https://periodicos.fclar.unesp.br/alfa/article/view/11433. [Último acceso: 2 October 2021].

[11] Glottolog, «Glottolog 4.5-Latin American Spanish,» Glottolog.org, 2013. [En línea]. Available: https://glottolog.org/resource/languoid/id/amer1254. [Último acceso: 2 October 2021].

[12] N. S. N. P. J. U. L. J. A. N. G. L. K. I. P. Ashish Vaswani, «Attention Is All You Need,» 6 December 2017. [En línea]. Available: https://arxiv.org/abs/1706.03762. [Último acceso: 2 October 2021].