Preserving Heritage: Developing a Translation Tool for Indigenous Dialects

Melissa Robles, Juan Camilo Prieto, Cristian Martínez, Sara Palacios y Rubén Manrique

Imagen tomada de https://www.onic.org.co/onic

Contenido

- Introducción
- Problema y propuesta
- Modelos anteriores y usados actualmente
- Traducción
- Comparación con estado del arte
- Recopilación de datos y construcción de bases
- Metodologías propuestas
- Resultados
- Avances y trabajo futuro

Introducción

- Según la ONIC, en Colombia hay 68 lenguas, de las cuales 65 son lenguas indígenas.
- Las lenguas que se encuentran los cuatro tipos morfológicos:

  - **Flexionales** como el griego o el latín (ej. Kogui).
  - **Aglutinantes** como el turco o el finlandés (ej. Wayuunaiki y Páez)
  - **Aislantes** como las lenguas malayo-polinesias (ej. embera del Chocó, criollo de San Andrés)
  - **Polisintéticas** como el esquimal (ej. kamsá).

Introducción

- Colombia es el tercer país en Latinoamérica con más diversidad lingüística, después de Brasil y México.
- Regiones con mayor diversidad:
  - Amazonía
  - Vaupés

- Se estima que cerca de **44 lenguas nativas han desaparecido en Colombia desde la época de la colonización.**
- De las 68 lenguas que sobreviven (tinigua, carijona, totoro y pisamira) son **habladas por menos de 30 personas.**

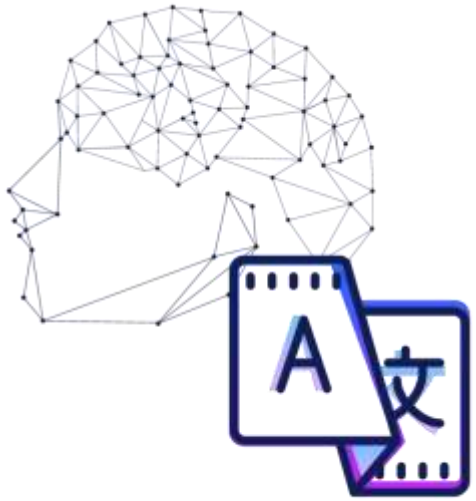
https://sinchi.org.co/lenguas-indigenas-de-la-amazonia-colombiana-en-alto-riesgo

https://eafit.edu.co/programas-academicos/pregrados/comunicacionsocial/noticias/noticias-2017/Paginas/apuesta-por-rescatar-las-lenguas-nativas-de-colombia.aspx

Problema

Muchas de estas lenguas nativas están en peligro de extinción debido a la reducción en el **número** de **hablantes nativos.**

Se prefiere mantener la **tradición oral** en lugar de la **escrita.**

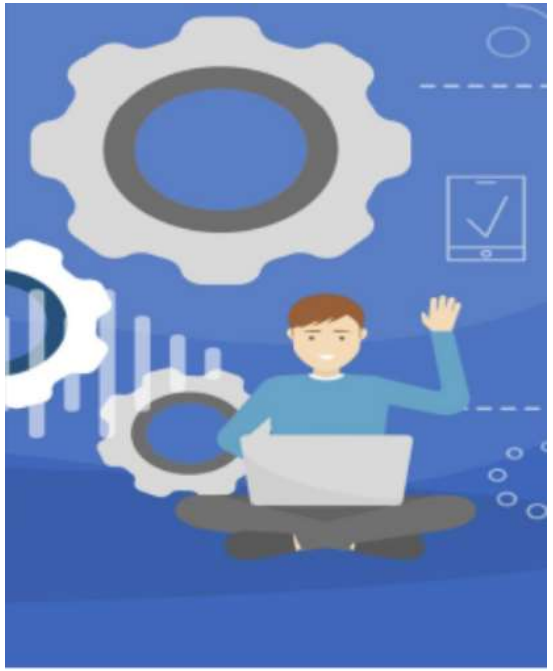


Propuesta

Se busca preservar las lenguas indígenas en Colombia a partir de la creación de un **traductor** español-lengua indígena para garantizar la inclusión y la igualdad lingüística.

Sistemas de Traducción Automática Basados en Reglas (RBMT)

- Se basan en reglas lingüísticas y gramaticales definidas manualmente por los lingüistas.
- Pueden tener dificultades para capturar todas las complejidades del lenguaje natural.
- Se necesita de profesionales en la lengua para tomar decisiones sobre estas reglas.
- Existen muchos casos y lógicas para lograr que funcione mínimamente bien.



Sistemas de Traducción Automática Basados en Estadísticas (SMT)

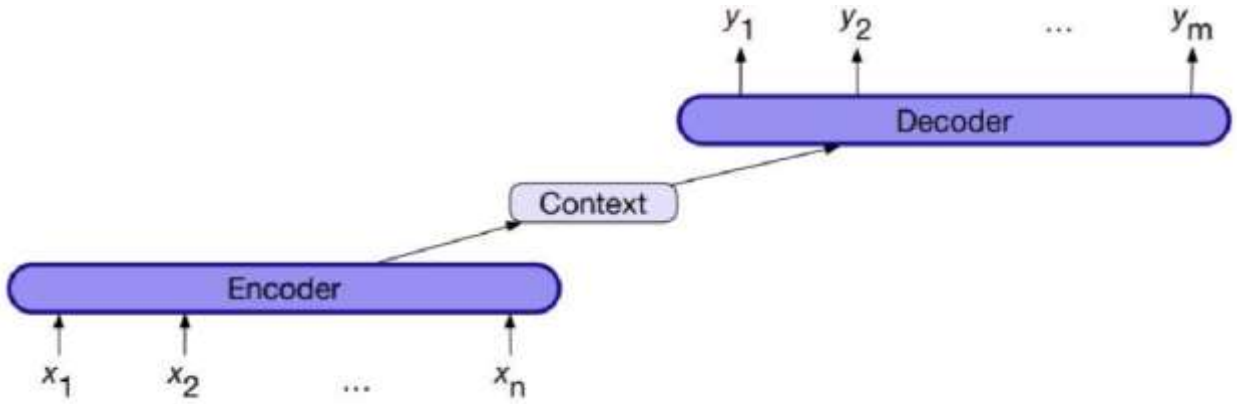
- Suelen estar basados en modelos de lenguaje n-gram y modelos de alineación de palabras, que calculan la probabilidad de que una secuencia de palabras en el idioma de origen se traduzca en una secuencia de palabras en el idioma de destino.
- La probabilidad condicional de una palabra dada su traducción.

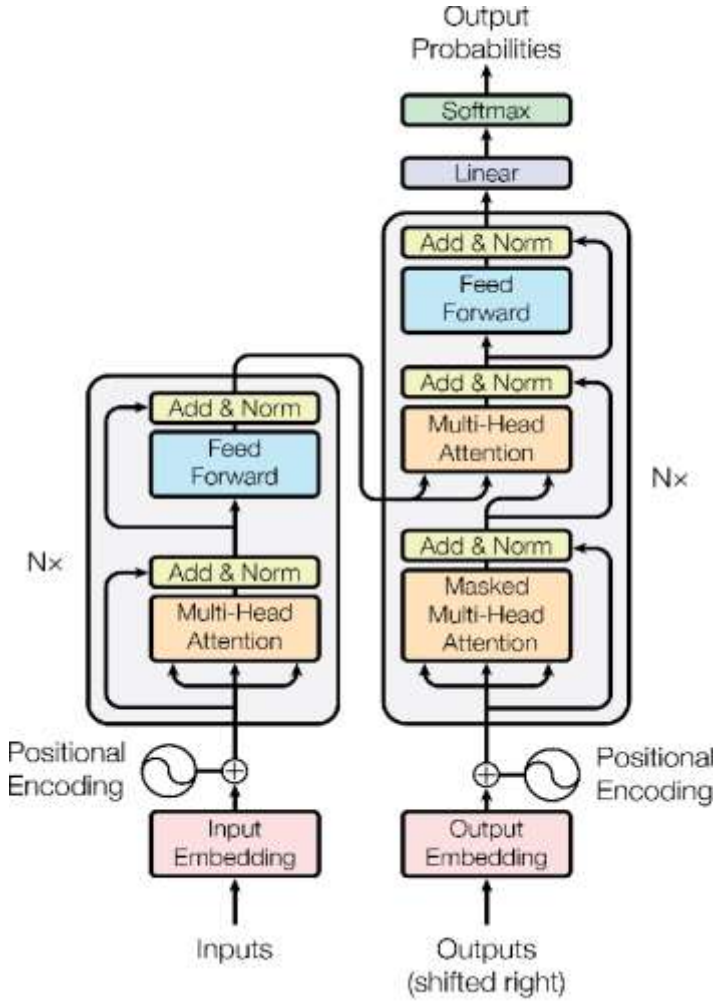
P(soy| yo) = 0.8

P(comer | yo voy a) = 0.3

Transformers

- Presentadas en 2017 por Vaswani et al. en "Attention is All You Need".
- Mecanismo de atención: Permite centrarse en partes específicas de la secuencia de entrada durante el entrenamiento e inferencia.
- Captura relaciones de largo alcance.
- Se puede paralelizar.

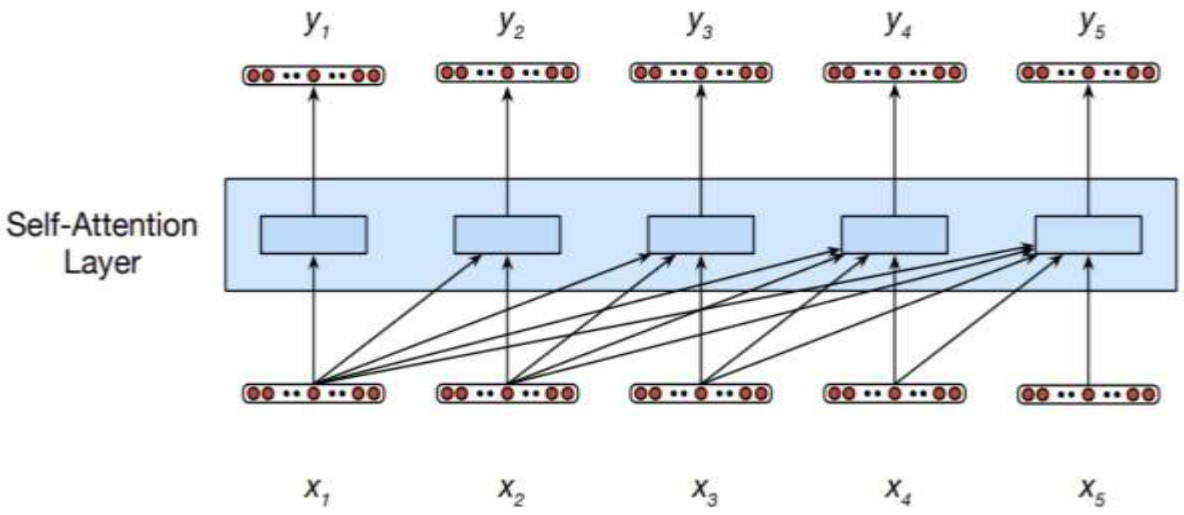


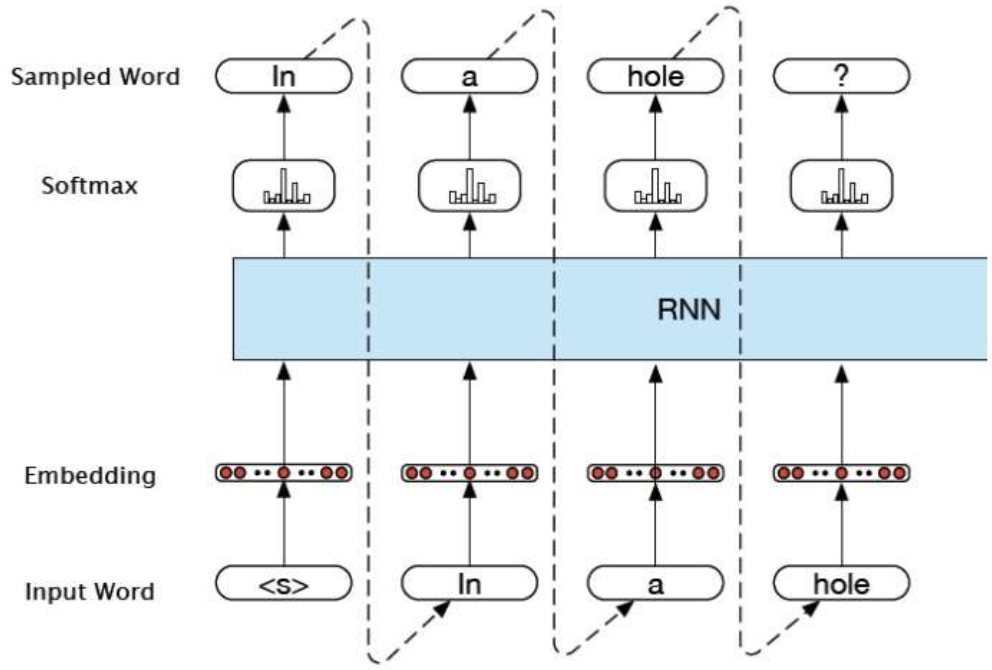


Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. Neural Information Processing Systems.

Mecanismo de atención

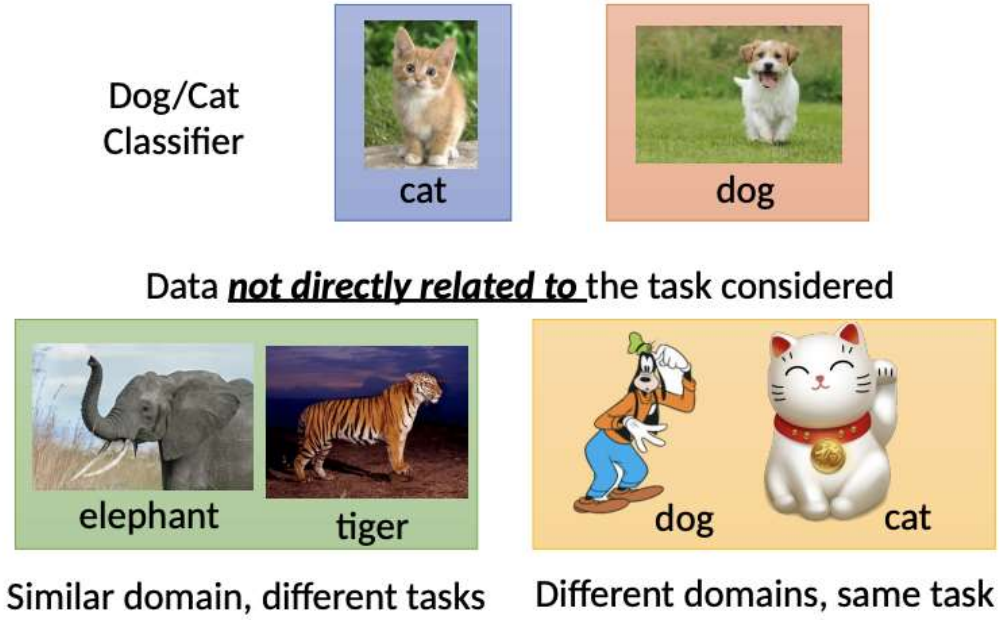
- El koala duerme felizmente abrazado de un árbol usando sus brazos.





Fine tuning

- Transferir el conocimiento de un modelo para realizar una nueva tarea.
- Aprovecha el conocimiento aprendido en grandes conjuntos de datos para mejorar el rendimiento en tareas específicas.
- Reducción del tiempo y recursos necesarios para entrenar modelos desde cero.
- Se puede entrenar un modelo con una base de datos más pequeña.



Fine tuning

Supongamos que no encuentra ningún modelo previamente entrenado para una tarea especifica

- Tarea: traducción (supervisada) de lenguas indígenas
- Datos de origen: Español-finlandés (2 millones de pares)
- Datos objetivo: Español-Wayuunaiki (83.600 pares)
- Idea: Entrenar un modelo según los datos de origen y luego ajustar el modelo según los datos de destino.
- Desafío: los datos objetivo son limitados (fácil generar overfitting)

- Beneficios del fine-tuning:
- Ya tendrá una comprensión básica de las estructuras lingüísticas y las relaciones entre los idiomas.
- Especialmente si comparten características lingüísticas similares.
- Más eficiente que entrenar un modelo desde cero.

Lenguas de bajos recursos

Existen más de 7000 lenguas en el mundo. Aún así, la mayoría de los modelos de lenguaje están entrenados únicamente con lenguas “de altos recursos”.

Una lengua de bajos recursos en el context de NLP se caracteriza por:

- Falta de texto anotado.
- Falta de datos conversacionales.
- Falta de otros recursos necesarios para entrenar modelos de lenguaje efectivos.

Resources for English (en) - Spanish (es) - (126 found)

Corpus	sentences	en tokens
CCMatrix *v1*	409,061,333	7,657,524,508
NLLB *v1*	409,061,333	7,657,593,794
ParaCrawl *v9*	269,400,355	4,374,060,920

Altos recursos

Inglés
Chino
Español
Francés
Japonés

Bajos recursos

Swahili → 140 millones de hablantes en el Este de África.

Chichewa → 9.3 millones de hablantes en Malaui.

Quechua → 7.8 millones de hablantes en Sudamérica.

https://opus.nlpl.eu/results/en&es/corpus-result-table

Lenguas de bajos recursos

Bajos recursos es diferente a pocos hablantes

Problemas de modelos en lenguas de bajos recursos:

- Overfitting
- Baja precisión
- Poco avance y recursos
- Perpetuación de sesgos

Soluciones

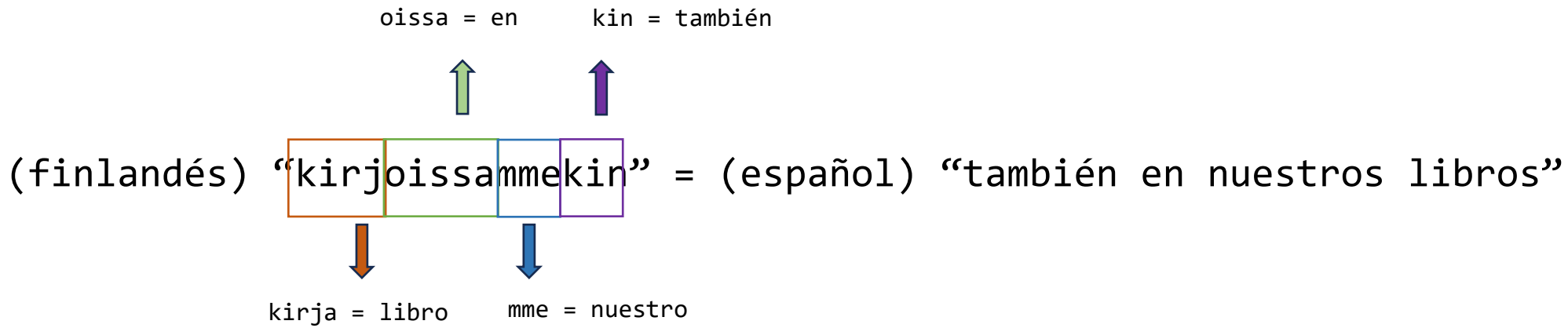
- Crear bases de datos, diccionarios o corpus etiquetados
- Cross-lingual transfer learning
- Aumentación de datos

140 millones de hablantes en el Este de África.

Diversidad en las lenguas

Lenguas aglutinantes: Finlandés, Japonés, Turco, Quechua, Guaraní, Maya, Nahuatl

Característica: palabras están conformadas por morfemas independientes.



Un **morfema** es la unidad mínima de significado en la estructura de una palabra.

Diversidad en las lenguas

Lenguas flexionales: Griego, Latín, Lenguas romance, Ruso, Kongui

Característica: Las palabras cambian su forma para expresar diferentes significados gramaticales, como el género, el número, el caso, el tiempo, el modo, la persona, etc.

(español) “yo comí”

Primera persona singular y en pretérito perfecto simple:

- Indica que la acción fue en el pasado
- Re afirma la primera persona

Diversidad en las lenguas

Lenguas polisintéticas: Groenlandés, Quechua, Nahuatl

Característica: morpheme-to-word ratio > 1

(groenlandés) “Aliikkusersuillammassuaanerartassagaluarpaalli”

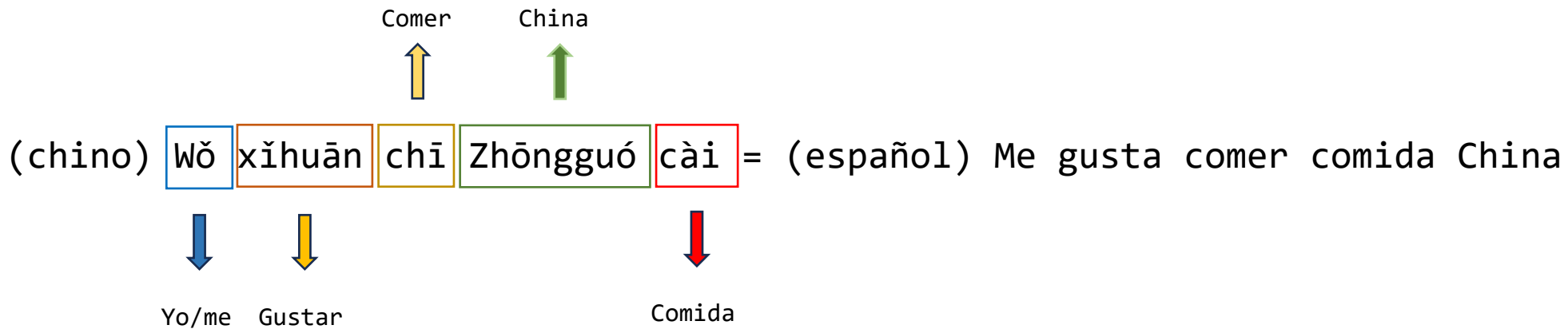
=

(inglés) “However, they will say that he is a great entertainer, but…”

Diversidad en las lenguas

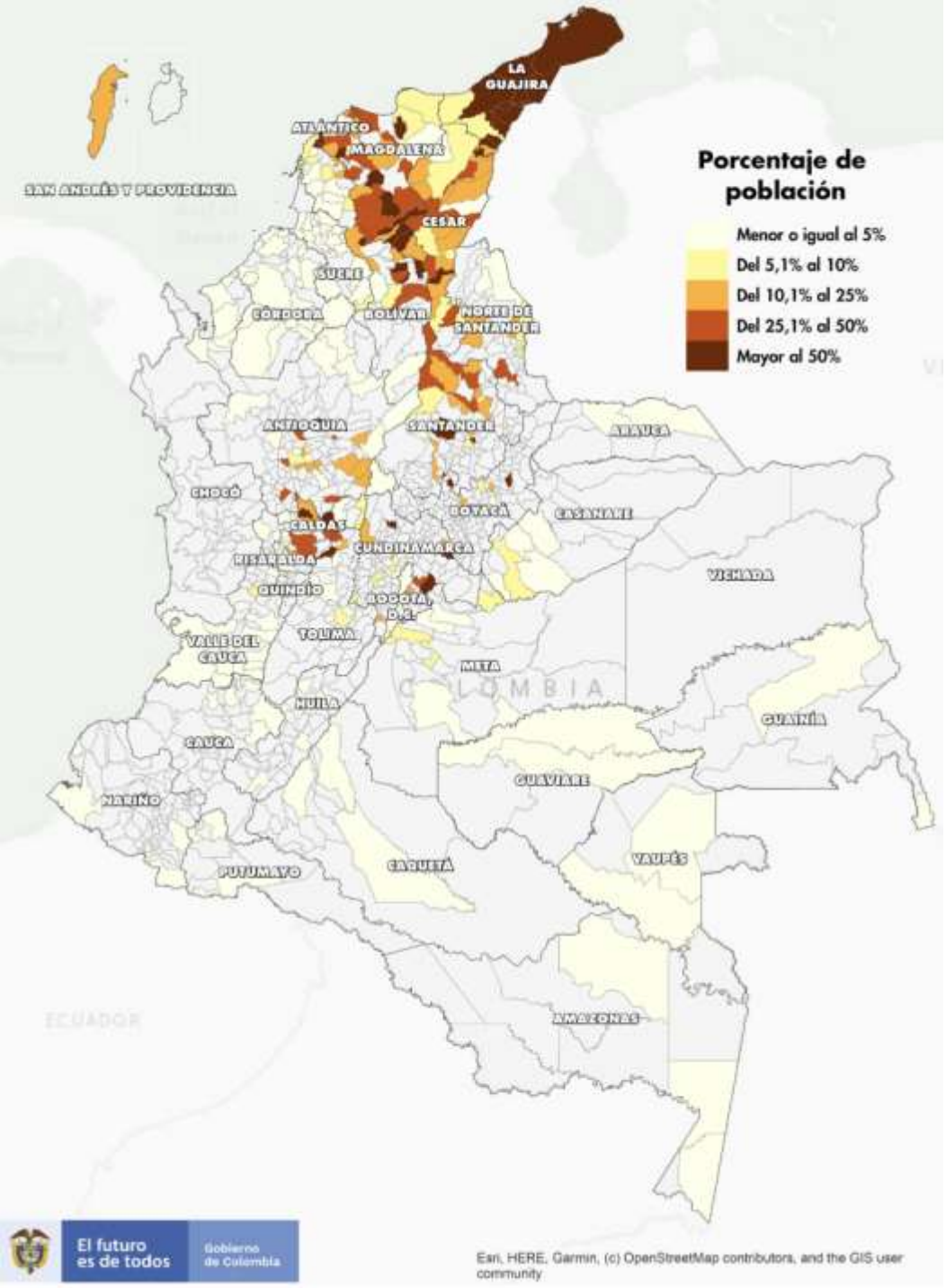
Lenguas Aislantes: Chino, Vietnamita, Yuroba, Embera

Característica: morpheme-to-word ratio = 1



Estado del arte

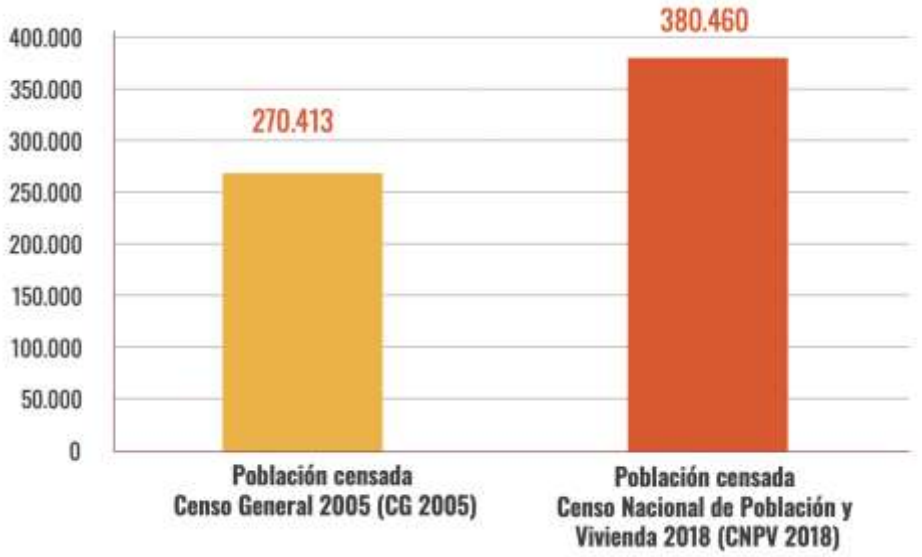
Título	Lengua indígena	Sistema de traducción	Cantidad de datos	Métricas y Resultados
Traductor estadístico wixarika - español usando descomposición morfológica. Hois, Mager et. al, 2016	Wixarika	Statistical Machine Translation	100 oraciones	WER: 38 TER: 0.84
Recopilación de corpus paralelo español-guaraní y experimentos iniciales con traductor automático estadístico. Álvarez López, Aldo Andrés. 2022	Guaraní	Statistical Machine Translation	33.367 Oraciones	BLEU: 47.1
Enriching Wayúunaiki-Spanish Neural Machine Translation with Linguistic Information. Graichen, Nora et. al. 2023	Wayuunaiki	Neural Machine Translation	43.501 Oraciones	BLEU: 4.5 ChrF2: 22.0 BLEURT: 0.22
Machine Translation Strategies for Low-Resource Colombian Indigenous Languages. Salazar, David. 2022	Nasa Yuwe Wayuunaiki	Neural Machine Translation	Nasa Yuwe: 7923 Wayúunaiki: 9081	BLEU: 2.82 ChrF: 29.52
Nuestra propuesta	Wayúunaiki Arhuaco	Neural Machine Translation	Wayúunaiki:83.606 Arhuaco: 5.722	BLEU: 10.54 ChrF2: 38.04



Wayuunaiki

El Wayuunaiki es una lengua indígena hablada principalmente en la región de **La Guajira**, en el norte de **Colombia** y partes de **Venezuela**.

Es una lengua **aglutinante**, es decir, las palabras se forman uniendo monemas independientes.

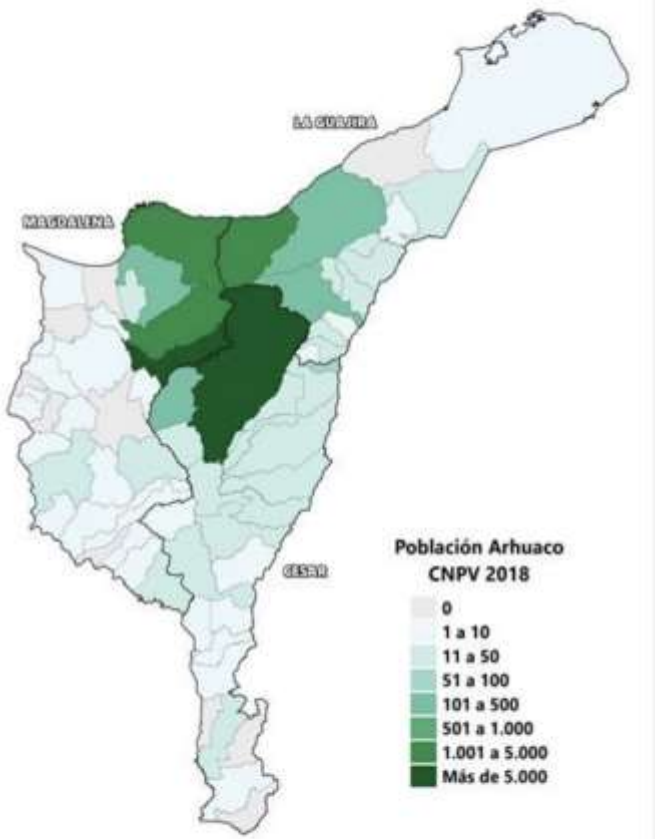


Recopilación de datos - Wayuunaiki

Para crear el corpus de traducción Wayuu-Español se utilizaron cuatro fuentes: la biblia, la constitución, un cuento traducido del Wayuu al Español y un diccionario.

Fuente	# Oraciones
Antiguo testamento (Génesis)	1473
Nuevo Testamento	7380
Constitución	131
Cuento	39
Diccionario	74583

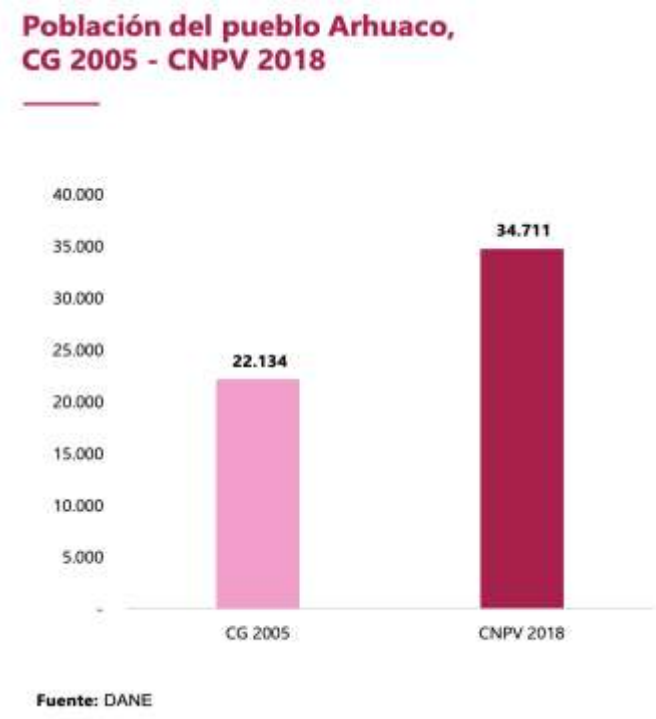
IKA (Arhuaco)



Mas del 70% de la población arhuaca vive en los departamentos de Magdalena, La Guajira y Cesar. Principalmente en la parte occidental y sur oriental de la Sierra Nevada de Santa Marta.

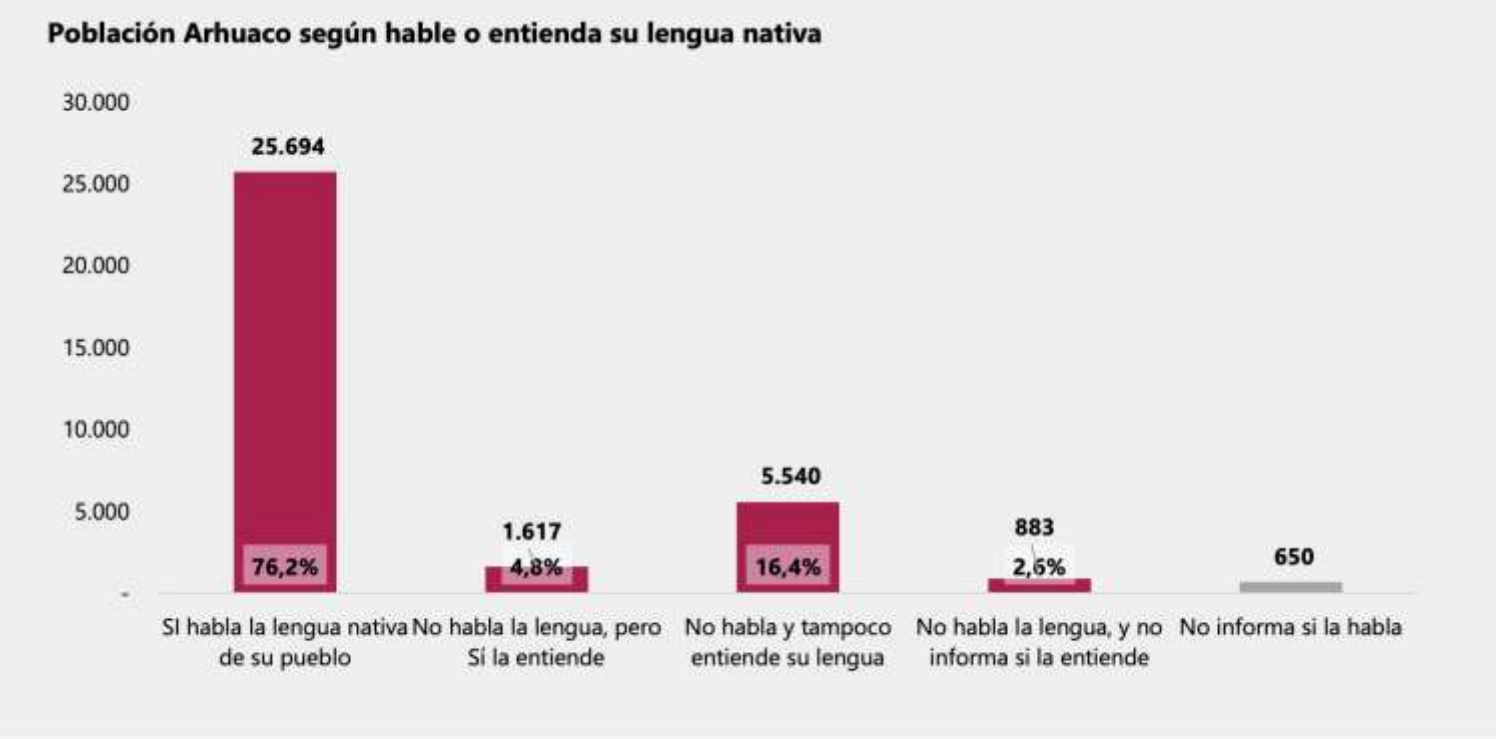
En esta lengua, las oraciones se construyen por medio de varios morfemas que se agregan a una raíz o lexema.

(idioma de la familia chibcha)



IKA (Arhuaco)

- El 76,2% de la población Arhuaca habla la lengua nativa de su pueblo.
- No hay muchos recursos escritos en internet.



Recopilación de datos
IKA(Arhuaco)

Fuente	# Oraciones
Antiguo testamento (Génesis)	1519
Nuevo Testamento	4044
Constitución	111
Cuento	29
Cuentos y poemas	19

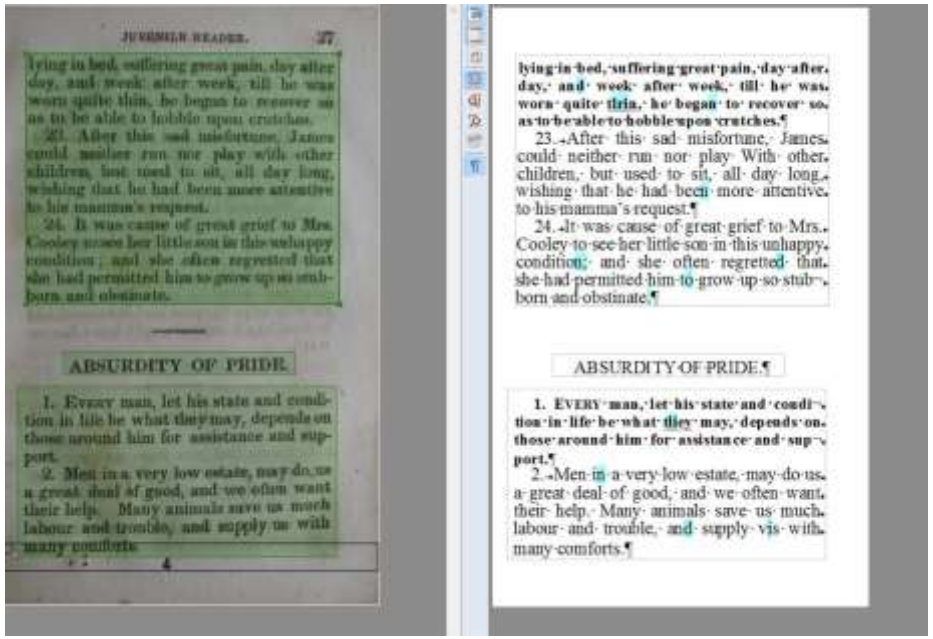
Para crear el corpus de traducción se utilizaron cuatro fuentes: la biblia, la constitución y un par de libros que contienen cuentos y poemas traducidos el arhuaco al español.

Procesamiento de los datos

- Realizamos scraping sobre páginas que tenían la biblia en español y nuestros lenguajes indígenas.
- Dividimos por cada versículo de la biblia

- Usamos Optical character recognition (OCR) para extraer las diferentes páginas de la constitución.
- Encontramos cartas, artículos y un pequeño diccionario para palabras concretas usadas en la constitución.

- Encontramos diferentes cuentos y poemas con su traducción en pdf, donde dividimos por oraciones.
- Realizamos un scrape de un diccionario en Wayuunaki que contenía traducciones de palabras y oraciones breves.



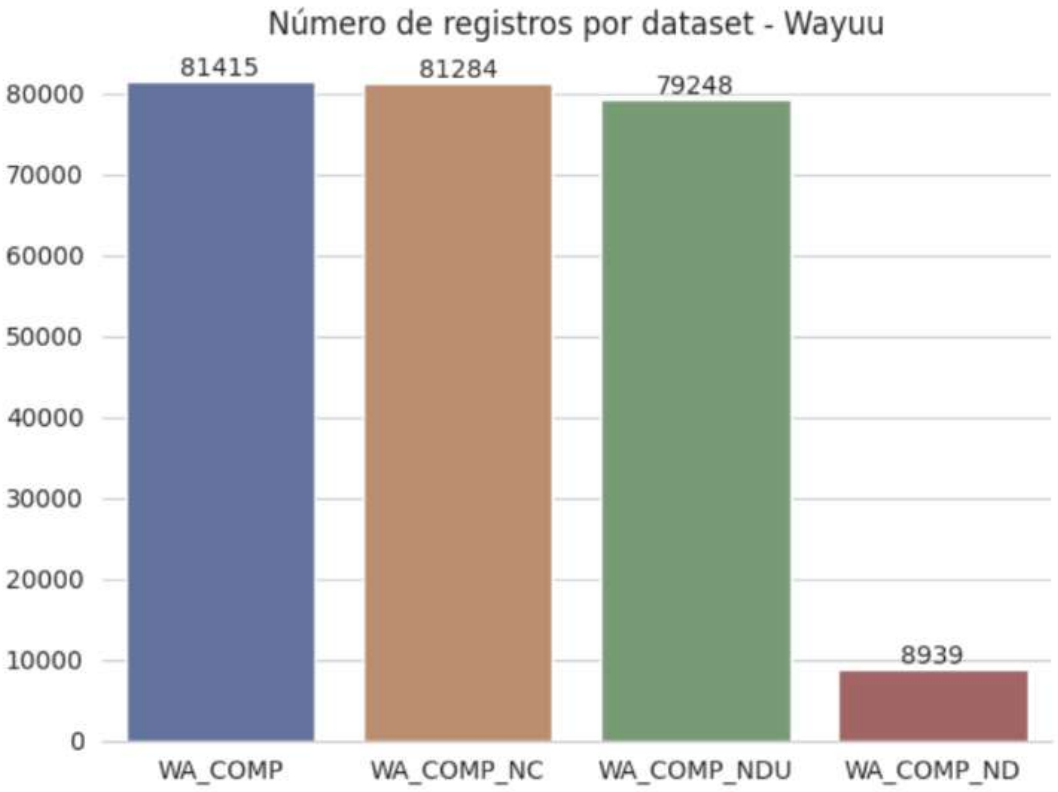
Consolidación base de traducción

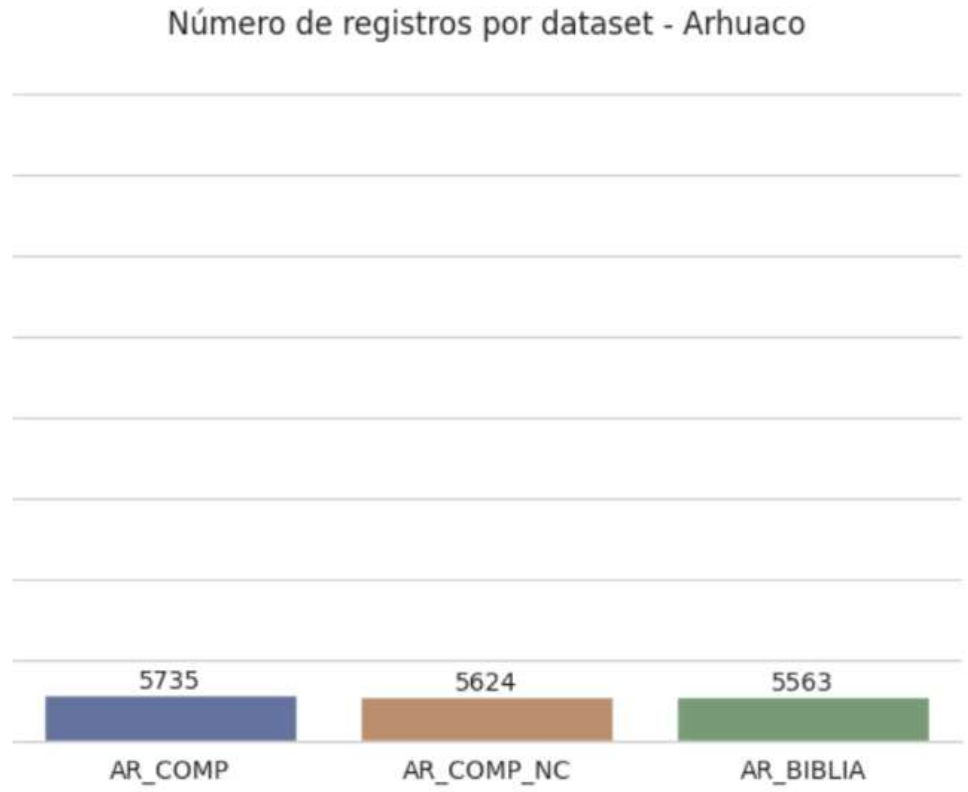
Los modelos se entrenaron con **distintas combinaciones** de las fuentes de datos con el fin de comparar el rendimiento final

Grupo	Lengua	Datasets Incluidos
WA_COMP	Wayuu	Dataset completo: Antiguo testamento, nuevo testamento, constitución, cuento y diccionario.
WA_COMP_NC	Wayuu	Dataset completo sin la constitucón
WA_COMP_NDU	Wayuu	Dataset completo sin considerar las palabras únicas del diccionario.
WA_COMP_ND	Wayuu	Dataset completo sin diccionario.
AR_COMP	Arhuaco	Dataset completo: Antiguo testamento, nuevo testamento, constitución, dos cuentos.
AR_COMP_NC	Arhuaco	Dataset completo sin la constitución
AR_BIBLIA	Arhuaco	Únicamente la biblia.

Consolidación base de traducción

Los modelos se entrenaron con **distintas combinaciones** de las fuentes de datos con el fin de comparar el rendimiento final

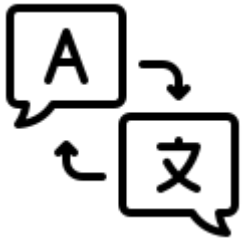


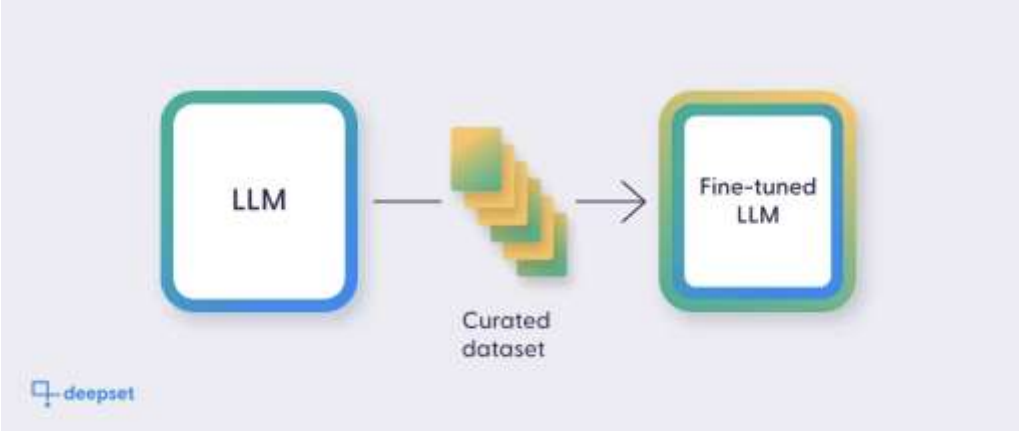


Metodología

Tres aproximaciones

1. Modelo Seq2Seq entrenado desde cero con arquitectura transformer.
2. Fine-tuning de un traductor finlandés.
3. Fine-tuning de un traductor multi-lengua.





Fine-tuning Large Language Models - https://www.deepset.ai/blog/llm-finetuning

Pasos en común:

1. División de datos en train y test.
2. Selección del tamaño máximo de la secuencia.
3. Selección y prueba del tokenizador.
4. Selección de hiperparámetros a explorar.
5. Entrenamiento de modelos.
6. Comparación de resultados.

Metodología

Modelo de traducción basado en transformers

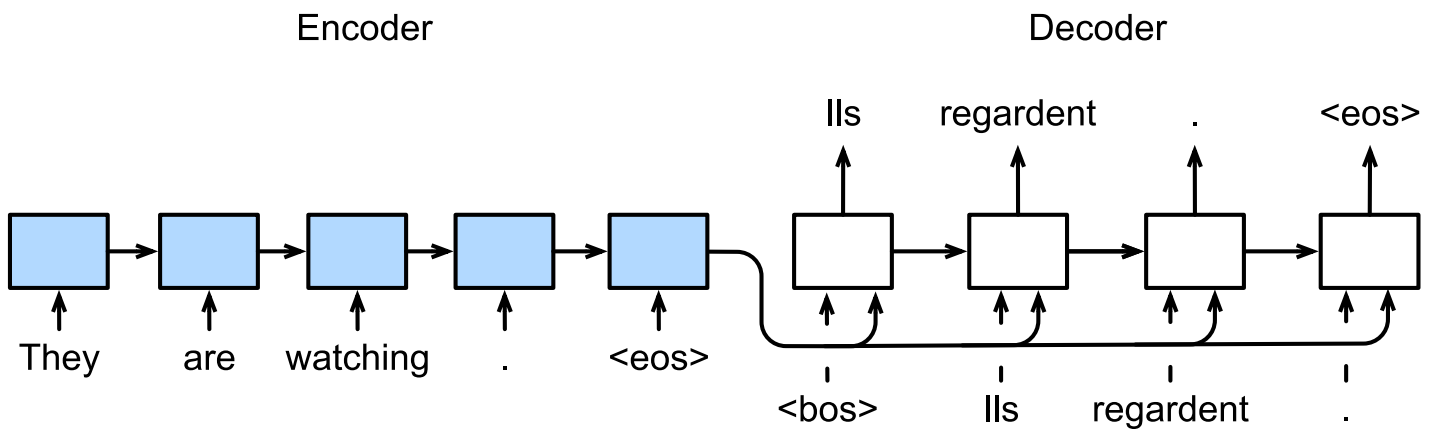
El modelo utiliza un encoder para analizar la frase de origen y comprender su significado, capturando las relaciones y características importantes.

Para el decoder recibe la representación codificada de la secuencia de entrada y comienza a generar la secuencia de salida en el idioma destino.

Se seleccionaron las 50**.000 palabras** más comunes en los tres idiomas como vocabulario y se definió una longitud de **sentencia máxima de 256 tokens.**

Hiperparámetros:

- Número de capas
- Número de cabezas
- Dropout
- Epsilon

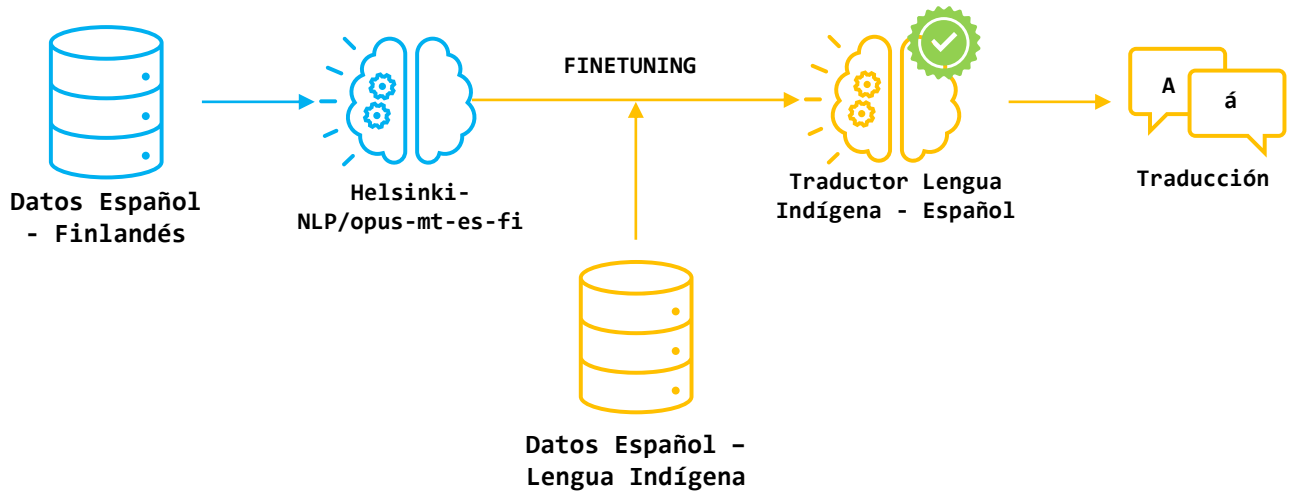


Metodología

Finetuning traductor español - finlandés

El finlandés es una lengua aglutinante: las palabras se forman mediante la unión de monemas independientes.

MarianNMT: Es una familia de más de 1000 modelos entrenados desarrollada por el equipo de Microsoft Translator. Se utilizó el modelo **Helsinki-NLP/opus-mt-es-fi** para posteriormente hacer fine-tuning con los datos consolidados.



Hipótesis: Suponemos que un modelo preentrenado en finlandés tiene un conocimiento que puede ser transferido y adaptado para mejorar la traducción del wayuu o arhuaco.

Metodología

Finetuning traductor español - finlandés

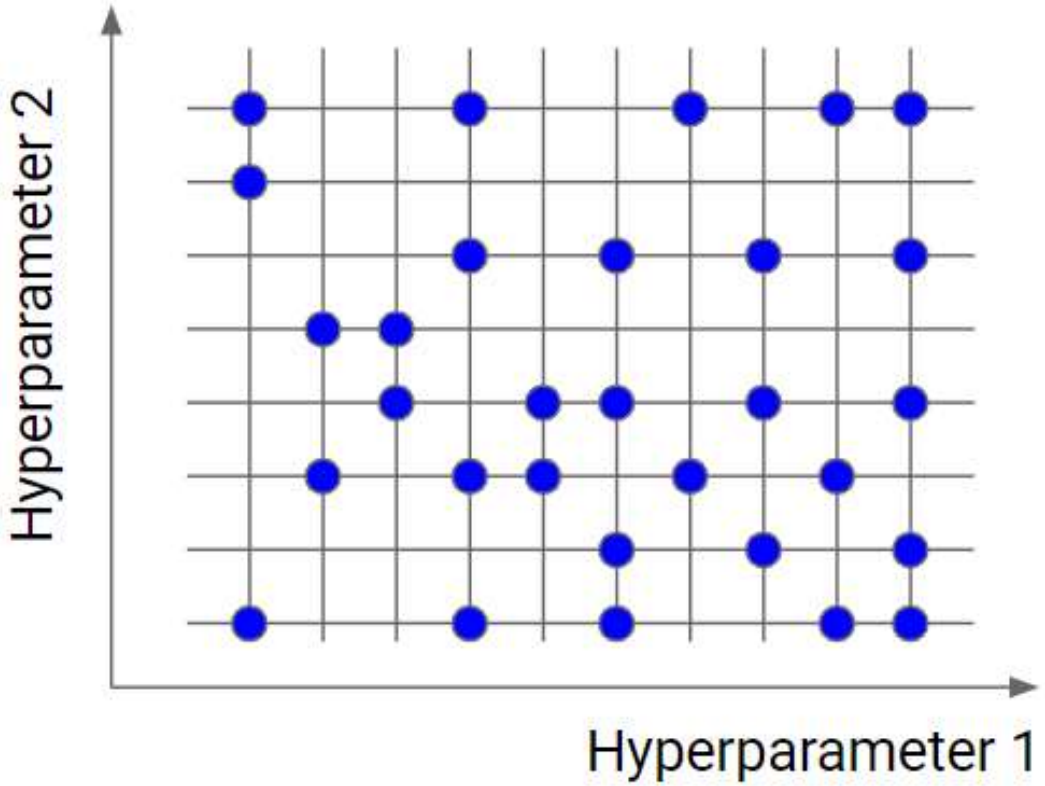
Hiperparámetros explorados:

- Número de épocas.
- Learning rate.
- Datasets.

Número total de modelos entrenados:

18 Español - Arhuaco

24 Español - Wayuu



https://www.yourdatateacher.com/2021/05/19/hyperparameter-tuning-grid-search-and-random-search/

Metodología

Finetuning traductor finlandés - español

Dado que los resultados de traducción de español a lengua indígena estaban teniendo buenos resultados, se decidió entrenar también modelos de lengua indígena a español para explorar si había una diferencia de dificultad entre las tareas.

MarainMT – Hugging Face - https://marian-nmt.github.io/quickstart/

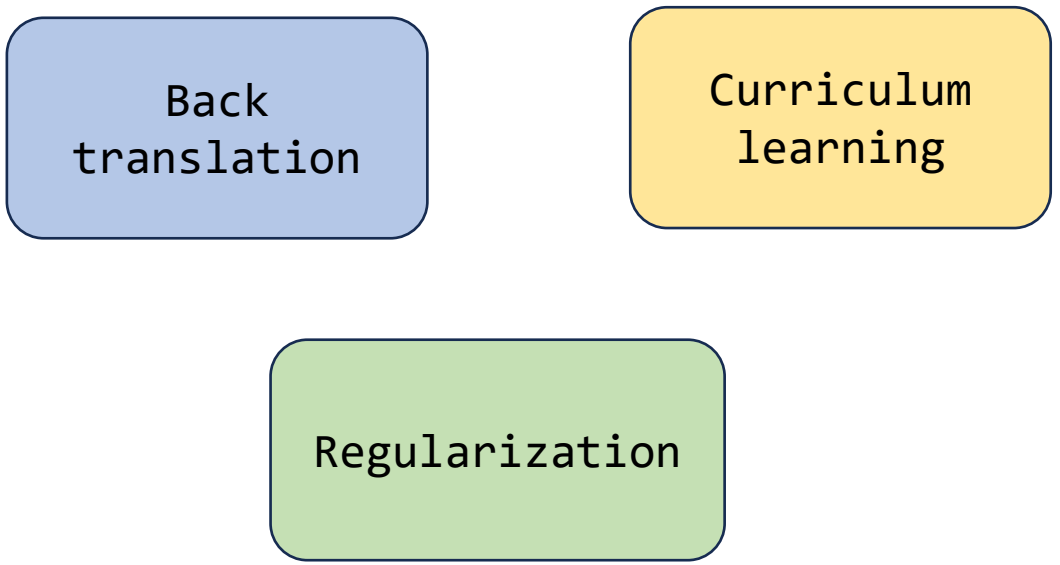
Se utilizó el modelo **Helsinki-NLP/opus-mt-fi-es** y se realizó la misma búsqueda de hiperparámetros que antes, obteniendo otros 42 nuevos modelos.

Metodología

Finetuning traductor multilengua

Se utilizó el modelo **No Language Left Behind (NLLB)** para buscar mejoras en el proceso de traducción de lenguas de bajos recursos ya que permiten que idiomas similares **compartan** datos durante el **entrenamiento**, mejorando significativamente la **calidad de la traducción** para idiomas con pocos recursos.

Problema: Los modelos se sobreajustan a medida que se entrena durante periodos largos.

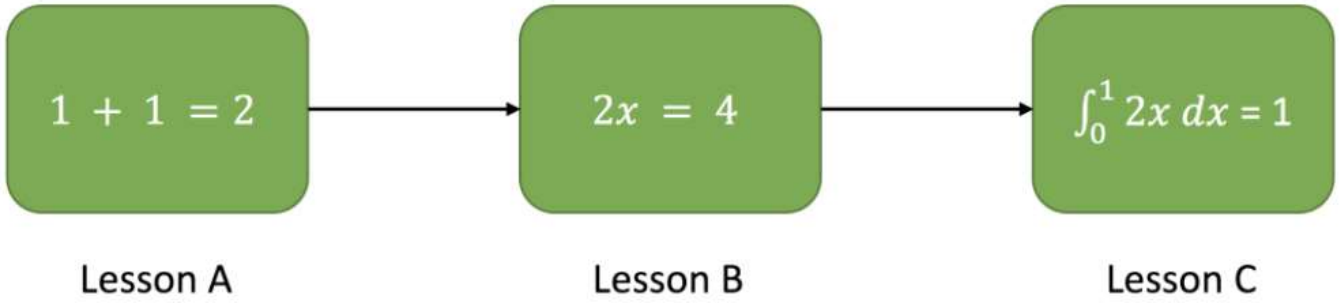


Soluciones de bajos recursos

Finetuning traductor multilengua

Curriculum learning

Los humanos y los animales aprenden mucho mejor cuando los ejemplos se organizan en un orden significativo que ilustra gradualmente más conceptos y, progresivamente, más complejos.



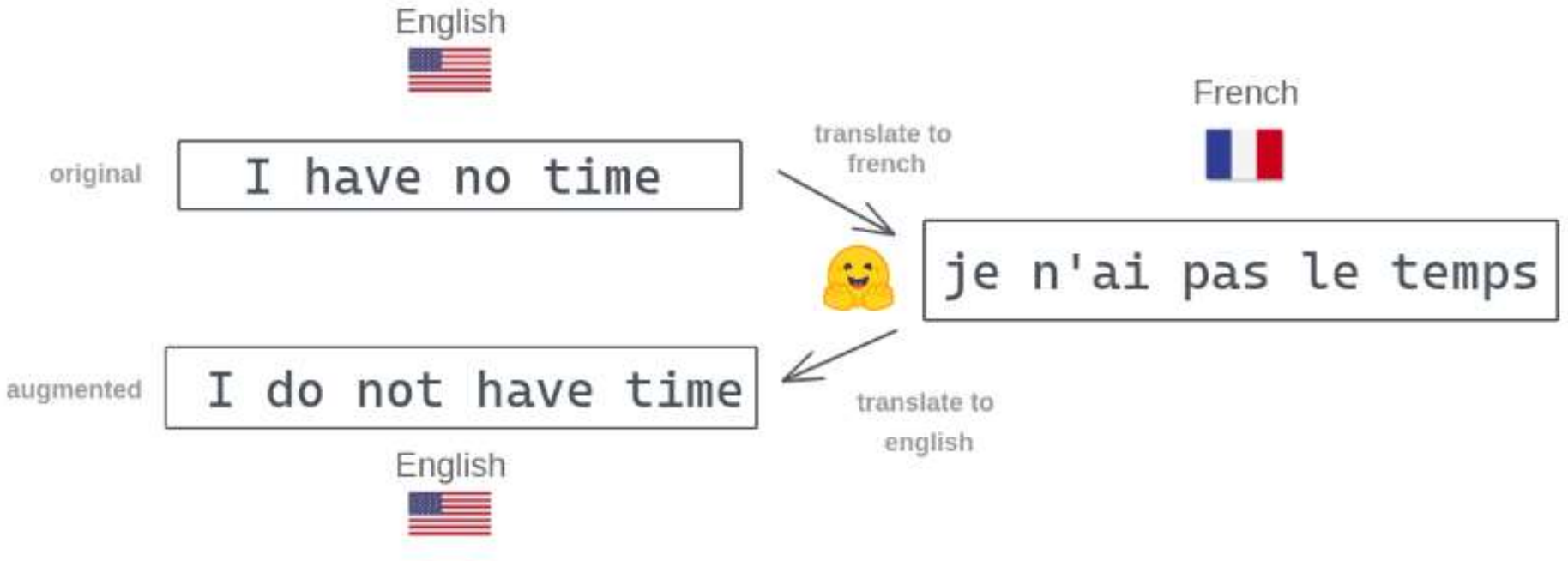
1. Entrenamiento con los idiomas de alto recurso durante unas cuantos épocas.
2. Luego se introdujeron los idiomas de bajos recursos.

https://medium.com/@pprocks/curriculum-learning-654aa6423abd

Soluciones de bajos recursos

Finetuning traductor multilengua

Back translation



https://amitness.com/back-translation/

Métricas

Comparar que los textos sean exactamente iguales no funciona para una métrica. “Tengo 40 años” != “Yo tengo cuarenta años”

BLEU

Comparar los N-gramas entre la traducción obtenida y la real.

Traducción

Real

$$Precisión\ unigramas = \frac{Num\ word\ matches}{Num\ words\ in\ generation} = \frac{4}{5}$$

Traducción

Real

$$Precisión\ unigramas = \frac{Num\ word\ matches}{Num\ words\ in\ generation} = \frac{5}{5} = 1$$

?

https://huggingface.co/learn/nlp-course/chapter1/1

Métricas

Comparar que los textos sean exactamente iguales no funciona para una métrica. “Tengo 40 años” != “Yo tengo cuarenta años”

BLEU

Comparar los N-gramas entre la traducción obtenida y la real.

Traducción

Real

Se restringe el número de veces que se cuenta cada palabra por el número de veces que aparece en la sentencia real

$$Precisión\ modificada\ unigramas = \frac{CLIP(Num\ word\ matches)}{Num\ words\ in\ generation} = \frac{1}{5}$$

Métricas

Comparar que los textos sean exactamente iguales no funciona para una métrica. “Tengo 40 años” != “Yo tengo cuarenta años”

BLEU

Comparar los N-gramas entre la traducción obtenida y la real.

$$Precisión\ modificada\ unigramas = \frac{4}{5}$$

?

$$Precisión\ modificada\ 4-gramas = \frac{0}{2} = 0$$

Métricas

ChrF2

Se centra en la concordancia de caracteres entre la traducción generada y la referencia.

A diferencia del BLEU, que se basan en coincidencias de palabras, ChrF2 evalúa la similitud a nivel de caracteres entre la traducción y la referencia.

Una limitacion es que no siempre captura la calidad de la traducción de manera exhaustiva, especialmente en contextos donde se requiere comprensión semántica o pragmática.

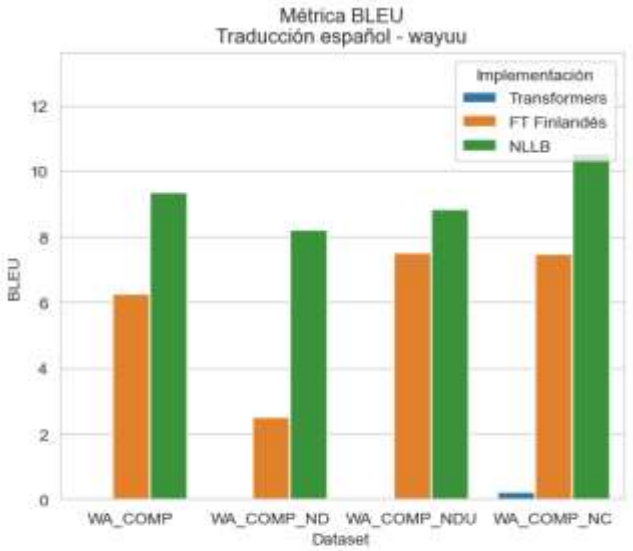
$$\mathrm{CHRF}\beta = (1+\beta^2)\frac{\mathrm{CHRP}\cdot\mathrm{CHRR}}{\beta^2\cdot\mathrm{CHRP}+\mathrm{CHRR}}$$

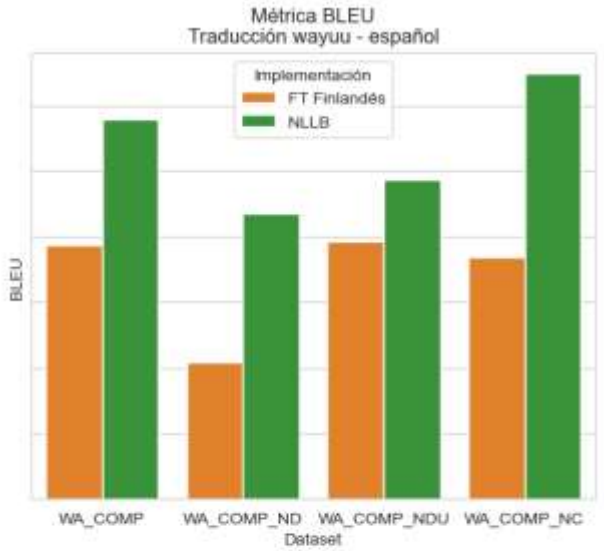
- CHRP: Porcentaje de n-gramas en la hipótesis que tienen una contraparte en la referencia
- CHRR: Porcentaje de n-gramas de caracteres en la referencia que también están presentes en la hipótesis
- B: Se da el valor de veces que le damos mas recall al que a la precision ; si β = 1, tienen la misma importancia.

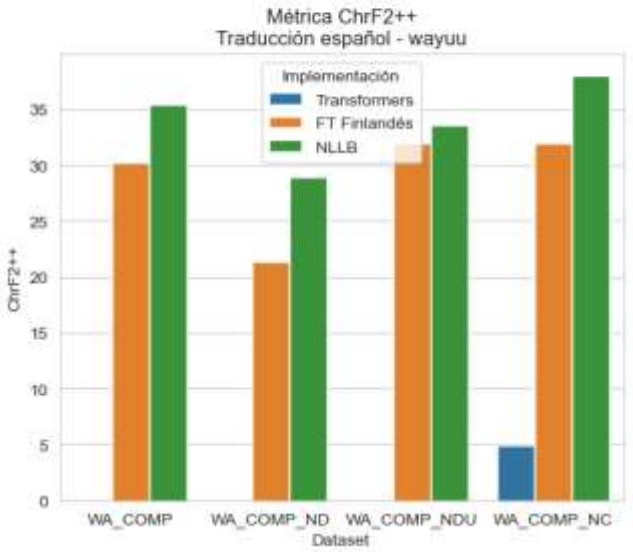
Resultados - Wayuu

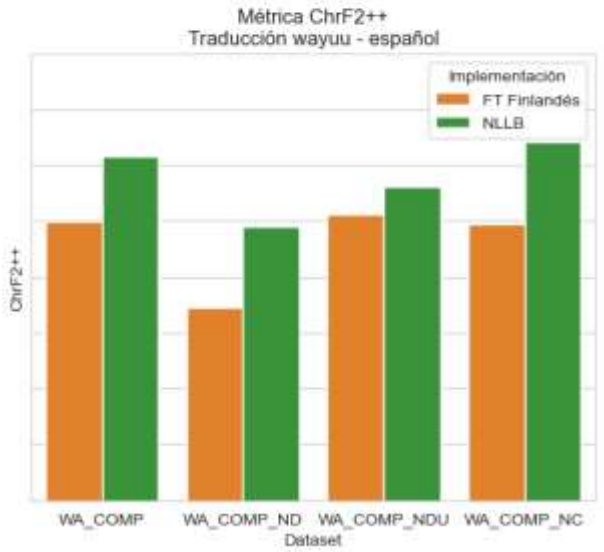
Métrcia BLEU

Métrcia ChrF2









Traducción español - wayuu

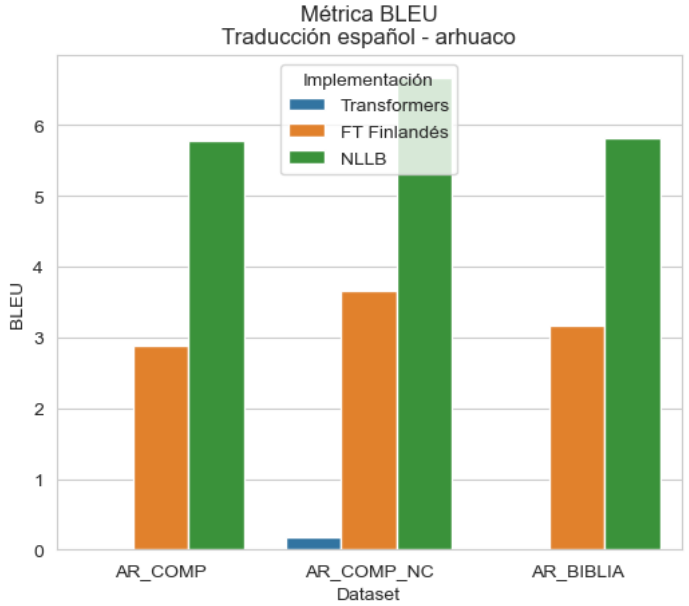
Traducción wayuu - español

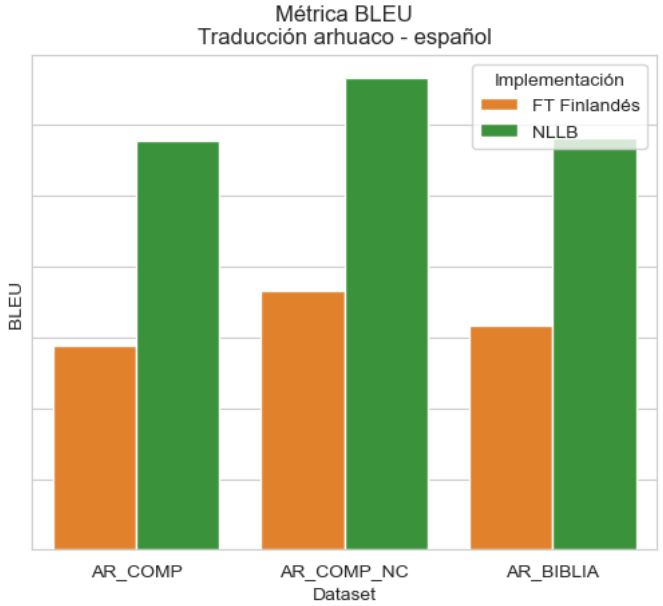
Se observa una eficiencia significativa en las métricas de **NLLB** a comparación de las otras dos aproximaciones.

El dataset **WA_COMP_NC** obtuvo mejores resultados. Además, El **diccionario** fue particularmente importante para la aproximación de finetuning con el traductor finlandes.

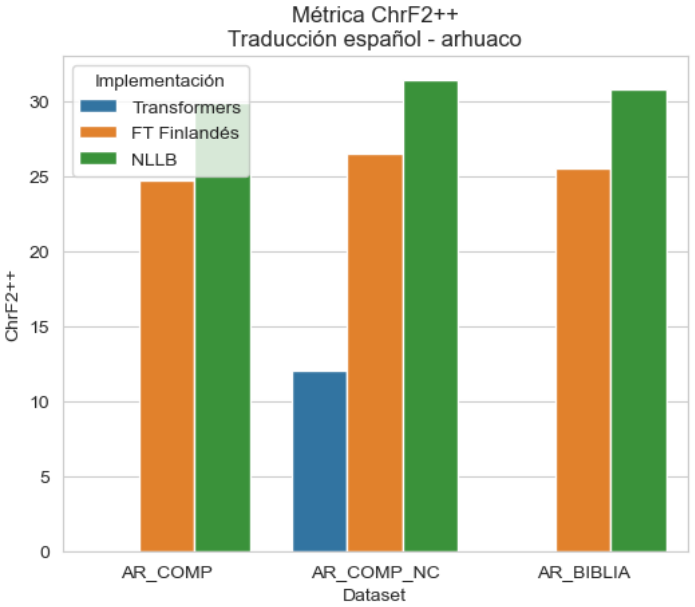
Resultados - Arhuaco

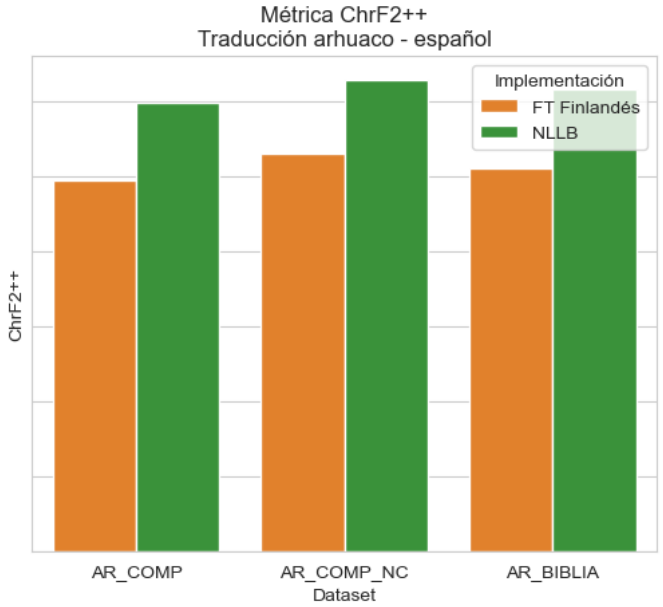
Métrcia BLEU





Métrcia ChrF2





Traducción español - arhuaco

Traducción arhuaco - español

La aproximación con mejores resultados fue el **fine-tuning de NLLB** para todos los casos, incluyendo ambas direcciones de traducción.

El dataset que mostró mejor rendimiento en todas las pruebas fue **AR_COMP_NC.**

Se evidencia una diferencia de **más del 50%** de la métrica **BLEU** entre la aproximación del fine-tuning de finlandés y el fine-tuning de NLLB. Esto puede deberse a que el modelo NLLB está entrenado con lenguas de bajos recursos.

Conclusiones

- La cantidad de datos es importante, al igual que la limpieza de los mismos. Por tal motivo se realizaron pruebas con diferentes conjuntos de datos.
- Realizar fine-tuning con modelos pre-entrenados de traducción mejora significativamente los resultados a comparación del entrenamiento de un modelo Seq2Seq desde cero.
- Para los modelos seq2seq los modelos con menos cabeza dieron mejores resultados dados la cantidad de datos.
- Utilizar una lengua aglutinante para el modelo base de traducción tuvo resultados favorables.
- La traducción de lengua indígena a español como la de español a lengua indígena son tareas de dificultad similar.
- Es posible que con más épocas el fine-tuning del traductor de finlandés mejore, dado que en la exploración de hiperparámetros siempre ganó el mayor número de épocas.
- Entre los modelos utilizados para fine-tuning el que presentó mejores resultados fue NLLB, posiblemente debido a la inclusión de lenguas de bajos recursos en el pre-entrenamiento del modelo.
- Las métricas obtenidas tanto en BLEU como en ChrF2 son competitivas con las obtenidas en el estado del arte para lenguas indígenas.

Avances y trabajo futuro

1. Creación de un dataset abierto para cuatro lenguas indígenas de Colombia

Language	Description	Sentences
Wayuunaiki	Dictionary (Amaya, 2021)	74583
Wayuunaiki	Bible (YouVersion, 2023)	6220
Wayuunaiki	Book (Álvarez, 2017)	534
Wayuunaiki	Book (Flórez et al., 2020)	467
Wayuunaiki	Book (Álvarez González, 2021)	229
Wayuunaiki	Book (Álvarez, 2016)	109
Wayuunaiki	Short story (Cue, 2012)	39
Wayuunaiki	Constitution (de Estudios de Lenguas Aborígenes, 1994)	37
Nasa Yuwe	Dictionary (originarios. Lenguas de América)	3729
Nasa Yuwe	Letters (Constitution) (de Estudios de Lenguas Aborígenes, 1994)	57
Nasa Yuwe	Common words (Constitution) (de Estudios de Lenguas Aborígenes, 1994)	53
Nasa Yuwe	Articles (Constitution) (de Estudios de Lenguas Aborígenes, 1994)	23
Arhuaco	Bible (para el Desarrollo de Pueblos Marginados)	5542
Arhuaco	Articles (Constitution) (de Estudios de Lenguas Aborígenes, 1994)	88
Arhuaco	Letters (Constitution) (de Estudios de Lenguas Aborígenes, 1994)	67
Arhuaco	Dictionary (Constitution) (de Estudios de Lenguas Aborígenes, 1994)	46
Arhuaco	Short stories (de la comunidad arhuaca de Jewrwa, 2014)	42
Inga	Dictionary (de Educación Inga de la Organización "Musu Runakuna", 1997)	3048
Inga	Constitution (de Estudios de Lenguas Aborígenes, 1994)	212

Table 1: Parallel data collected for each language

2. Participación en concurso de traducción de lenguas indígenas de américa.

3. Comunicación con hablantes para obtener fuentes actualizadas de datos.