PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

Escuela de Posgrado

Minería web de textos en lenguas indígenas para desarrollar tecnologías de lenguaje. Caso de estudio: quechua sureño.

Tesis para obtener el grado académico de Magíster en Informática con mención en Ciencias de la Computación que presenta:

Victoria Alejandra Ubaldo Gamarra

Asesor:

Felix Arturo Oncevay Marcos

Lima, 2022

RESUMEN

En la actualidad, para los más de 30 millones de peruanos, la información a la que accedemos se encuentra mayormente en el idioma español. Sin embargo Perú es un país multilingüe, posee una gran riqueza cultural y lingüística con alrededor de 47 lenguas originarias. Para esta población encontrar textos, noticias y contenido en internet en su lengua nativa es una tarea complicada. Existe un limitado acceso a información como lecturas, textos, noticias u otros contenidos que en modalidad digital es muy escaso. Esto se debe a que los pocos ciudadanos que se comunican en lenguas nativas son de manera oral y algunos hacen uso del español sobre sus lenguas nativas.

De ese modo, existen investigaciones en el campo de la inteligencia artificial donde a partir del poco material digital recolectado de lenguas nativas se construyeron corpus digitales para tareas de traducción automática y detección del lenguaje. Sin embargo, aún son corpus pequeños para elaborar traductores de calidad, presentan complicaciones en traducir textos completos, y además díficil el aprendizaje con algoritmos complejos, como redes neuronales profundas.

Por este motivo se propone realizar una minería web de textos en la lengua originaria quechua sureño para incrementar la cantidad de oraciones y diversidad de dominios, evaluar la calidad de los nuevos textos en un modelo de traducción automática de quechua a español, y desarrollar una web de libre acceso de consulta al corpus creado.

ABSTRACT

Currently, for more than 30 million Peruvians, the information we access is mostly in the Spanish language. However, Peru is a multilingual country, it has a great cultural and linguistic wealth with around 47 native languages. For this population, finding texts, news and internet content in their native language is a complicated task. There is limited access to information such as readings, texts, news, or other content that is very scarce in digital mode. This is because the few citizens who communicate in native languages do so orally and some use Spanish over their native languages.

In this way, there is research in the field of artificial intelligence where, from the little digital material collected from native languages, digital corpus was built for automatic translation and language detection tasks. However, they are still small corpus to develop quality translators, they present complications in translating complete texts, and difficult to learn with complex algorithms, such as deep neural networks.

For this reason, it is proposed to carry out a web mining of texts in the native language Quechua Sureño to increase the number of sentences and diversity of domains, evaluate the quality of the new texts in a model of automatic translation from Quechua to Spanish, and develop a web of free consultation access to the created corpus.

AGRADECIMIENTOS

A mi asesor Mg. Arturo Oncevay, por guiar y compartir sus conocimientos durante este proyecto y ser un buen mentor para lograr cumplir los objetivos.

A mis familiares y amigos por el gran apoyo y amistad durante lo largo del desarrollo del proyecto, gracias por la inspiración.

Índice General

RESUMEN 2
ABSTRACT 3
AGRADECIMIENTOS 4
1. GENERALIDADES 9
1.1. Problemática 9
1.2. Objetivos 11
1.2.1. Objetivo General 11
1.2.2. Objetivos Específicos 11
1.2.3. Resultados Esperados 11
1.3. Herramientas y Métodos 12
1.3.1. Herramientas 12
1.3.2. Metodología 13
1.4. Alcance y limitaciones 13
1.4.1. Hipótesis 13
1.4.2. Justificación 13
1.5. Limitaciones 14
2. MARCO CONCEPTUAL 14
2.1. Lenguas minoritarias 14
2.2. Corpus paralelo 14
2.3. Web Scraping 14
2.4. Codificación de byte pares(BPE) 15
2.5. Traducción automática 15
2.5.1. Modelo de secuencia a secuencia 15
2.5.2. Back-translation 15
2.6. Métricas de medición 16
2.6.1. Exactitud (Accuracy) 16
2.6.2. BLEU 16
2.6.3. ChrF 16
3. ESTADO DEL ARTE 17
3.1. Estrategia de búsqueda 17
3.1.1. Cadenas de búsqueda 17
3.1.2. Preguntas de revisión 17
3.2. Resultados de la revisión 17

4. EXPERIMENTACIÓN Y RESULTADOS 19
4.1. Modelo de traducción automática 19
4.1.3. Modelo Transformer Base 23
4.1.4. Modelo Transformer en corpus Helsinki y REPU 23
4.1.5. Modelo Transformer en corpus extra 24
4.1.6. Modelo Transformer y Fine-Tuning 24
4.1.7. Resultados obtenidos 25
4.2. Desarrollo de herramienta de consultas de corpus 28
5. CONCLUSIONES Y TRABAJOS FUTUROS 30
Conclusiones 30
5.1. Trabajos futuros 30
Bibliografía 31

Índice Figuras

Figura 4.1 Oraciones totales por corpus paralelos recolectados. 21
Figura 4.2 Ratio V/N en corpus quechua sureño. 21
Figura 4.3 Oraciones acumuladas por corpus. 22
Figura 4.4 Estadísticas resumen de corpus monolingües. 22
Figura 4.5 Indicador BLEU sobre modelos. 25
Figura 4.6 Indicador chrF sobre modelos. 26
Figura 4.7 Indicador BLEU sobre modelos con fine tuning 26
Figura 4.8 Indicador chrF sobre modelos con fine tuning 27
Figura 4.9 Pantalla inicial del buscador. 29
Figura 4.10 Búsqueda de palabras en corpus paralelo. 29

Índice Cuadros

Cuadro 1.1 Herramientas utilizadas en el proyecto. 12
Cuadro 4.1 Título y Fuente de lecturas adicionales. 20
Cuadro 4.2 Cantidad de oraciones en el conjuntos de entrenamiento modelo AmericasNLP. 23
Cuadro 4.3 Cantidad de oraciones en entrenamiento Helsinki y REPU. 23
Cuadro 4.4 Cantidad de oraciones en entrenamiento texto adicional. 24
Cuadro 4.5 Modelado con Fine-Tuning desde corpus Helsinki. 24
Cuadro 4.6 Modelado desde linea base AmericasNLP. 25
Cuadro 4.7 Modelado con Fine-Tuning desde corpus Helsinki 26
Cuadro 4.8 Steps y Accuracy por modelo inicial 27
Cuadro 4.9 Resultados Modelo Monolingüe Base 28
Cuadro 4.10 Steps y Accuracy por modelo en fine tuning 28

1. GENERALIDADES

1.1. Problemática

En los últimos años, previos al bicentenario del Perú, aproximadamente 4 millones de personas tienen como lengua materna a una de las 47 lenguas nativas o también llamadas minoritarias. (Sullon Acosta, 2013). De ese modo, son pocos los ciudadanos que se comunican oralmente con lenguas como el quechua (en todas sus variantes), aymara o shipibo-konibo, las cuales son transmitidas oralmente entre generaciones. Asimismo, debido a los cambios sociales y culturales para comunicarse ante los sucesos de esta última década, incluido la pandemia por el COVID-19, estos ciudadanos han tenido la necesidad de priorizar el español sobre sus lenguas nativas.

En este sentido, el quechua, en todas sus variantes, es una lengua minoritaria, el cual se describe como una lengua usada por un pequeño grupo de habitantes, llegando a incluso quedar olvidada en el tiempo, y junto a eso su historia y patrimonio cultural. De esta manera, el alcance para acceder a información en quechua sureño en textos, libros históricos, noticias y artículos es escaso, y se cuentan con pocos datos que permitan su enseñanza. La tarea de recopilar y digitalizar estos recursos bibliográficos y textos es costosa y requiere de muchos recursos, como horas de expertos. (Forcada, 2006)

En este escenario, el gobierno ha dado un avance desde el Ministerio de Educación el 13 de junio del 2015 cuando informaron que el país cuenta oficialmente con 24 alfabetos de lenguas originarias desde la Resolución Ministerial Nª 303-2015-MINEDU (Ministerio de Educación, 2015) y luego desde el Plan Nacional de Educación Intercultural Bilingüe (EIB) de acuerdo a la Resolución Ministerial N°629-2016-MINEDU (Ministerio de Educación, 2016). Por consiguiente, esto permitió que se generen textos bilingües para la educación siguiendo la política de Estado con el decreto Nº-005-2017-MC (Gobierno del Perú, 2017). En tanto, desde el artículo 2º en la ley Nº 29735 declara la preservación, desarrollo y fomento de lenguas originarias en el Perú (Poder Legislativo, 2011). Además, en el artículo Nº 9 oficializa el uso de lenguas originarias y soporta la preservación de estas lenguas nativas en sus comunidades de origen a partir de la elaboración de textos de educación bilingües.

Asimismo, desde el punto de vista de las ciencias de la computación, los estudios de traducción automática de lenguas minoritarias permiten revitalizarlas en la actualidad y generar su aprendizaje y estandarización a la población (Forcada, 2006). Estas tareas resultan desafiantes debido a que existen corpus pequeños para elaborar traductores de calidad y complicaciones en traducir textos completos.

Para lograr dicha mejora y por las razones descritas anteriormente, en el presente trabajo se propone una innovadora recolección de textos en quechua sureño a partir de nuevas fuentes de información como redes sociales en publicaciones de usuarios quechua hablantes o activistas, guías y PDFs en quechua con información COVID-19, además de nuevos escritos y sitios web en quechua.

Segundo, se propone evaluar la calidad de estos textos recolectados usando modelos de traducción automática, para comparar su rendimiento con una línea base de comparación a un corpus previamente estandarizado. La evaluación se realizará de forma automática usando las métricas de calidad estándar como BLEU y chrF.

Por último, se propone diseñar un sitio web de disponibilidad abierta u *open source*, que permita interactuar con los usuarios mediante la búsqueda de palabras en quechua o español, y encontrar su traducción con el corpus recolectado.

1.2. Objetivos

A continuación, se determinan los objetivos del siguiente proyecto, los resultados esperados, las herramientas a utilizar y la justificación de la investigación.

1.2.1. Objetivo General

Recolección de textos en lenguas originarias en la web para su aplicación en tareas automáticas de generación, como la traducción automática.

1.2.2. Objetivos Específicos

1. Identificar, extraer y limpiar textos escritos en quechua sureño desde fuentes no tradicionales.

2. Evaluar la calidad de los textos en una tarea de traducción automática de quechua sureño a español.

3. Diseñar e implementar una página web para consultar y visualizar los textos recolectados.

1.2.3. Resultados Esperados

1. Para el primer objetivo:

- Corpus de línea base con textos quechua sureño recolectado a partir de investigaciones anteriores.
- Scripts implementados para la extracción de textos en quechua sureño de PDFs, sitios web y redes sociales.
- Corpus procesado con scripts para su limpieza.

2. Para el segundo objetivo:

- Modelo de traducción automática, de quechua sureño a español, entrenado con el corpus de línea base.
- Modelo(s) de traducción automática, de quechua sureño a español, entrenado con los nuevos corpus recolectados y procesados.

- Comparación de la calidad de los corpus a partir de la evaluación de los modelos de traducción automática con las métricas BLEU y chrF.

3. Para el tercer objetivo:

- Página web implementada para consultar los textos recolectados, que permitirá analizar los datos paralelos entre español y quechua sureño.

1.3. Herramientas y Métodos

Las herramientas y métodos principales que se utilizan en el desarrollo del siguiente proyecto fueron las siguientes:

1.3.1. Herramientas

Para el siguiente proyecto se detalla las herramientas utilizadas para cada objetivo y resultado esperado.

Objetivo Específico	Herramientas
Identificar, extraer y limpiar textos escritos en quechua sureño desde fuentes no tradicionales.	• Jupyter Notebook[1] • Python[2] • Tweepy [3] • Facebook-scrapper[4] • PDFMiner2[5]
Evaluar la calidad de los textos en una tarea de traducción automática de quechua sureño a español.	• Jupyter Notebook • Pytorch[6] • Métrica BLEU[7]
Diseñar e implementar una página web para consultar y visualizar los textos recolectados.	• Django[8] • Elasticsearch[9]

Cuadro 1.1 Herramientas utilizadas en el proyecto.

[1] https://jupyter.org/
[2] https://www.python.org/
[3] https://www.tweepy.org/
[4] https://pypi.org/project/facebook-scraper/
[5] https://pypi.org/project/pdfminer2/
[6] https://pytorch.org/
[7] https://aclanthology.org/P02-1040.pdf
[8] https://www.djangoproject.com/
[9] https://www.elastic.co/

1.3.2. Metodología

En base a los objetivos planteados, se detalla la metodología a desarrollar:

O1) Construir un nuevo corpus a partir de la implementación de scripts que permitan recolectar datos de quechua sureño peruanas en nuevas fuentes, como redes sociales y lecturas en PDF. Seguidamente, se realiza una conversión a texto utilizando herramientas en Python como Tweepy, Facebook-Scrapper y PDFMiner.

O2) Implementar y evaluar al menos dos modelos de traducción automática. El primero con el corpus de línea base, y los siguientes con la corpora nueva de forma acumulada. Luego, representar las posibles mejoras en el modelo NMT, por ejemplo: el entendimiento de las palabras y frases, y el manejo de palabras raras que es un problema bastante conocido en este tipo de modelos. Para esto, se utilizará una red neuronal del tipo decoder-encoder con el fin de optimizar el modelo. Finalmente, se va a medir la calidad del texto traducido usando BLEU y chrF, las cuales son métricas usadas para evaluar la calidad de un texto traducido.

O3) Diseñar e implementar una página web para consultar y visualizar nuevos textos.
Se va a diseñar un sitio web que permita realizar búsquedas de palabras o frases de los corpus paralelos digitalizados Para esto, se utilizará Django para el back-end y Elasticsearch como motor de búsquedas y gestión de documentos.

1.4. Alcance y limitaciones

1.4.1. Hipótesis

Es posible desarrollar y mejorar la calidad de las traducciones hechas desde un modelo base de traducción automática usando nuevos corpus paralelos y monolingües desde nuevas lecturas y publicaciones de redes sociales en quechua sureño.

1.4.2. Justificación

La presente investigación se justifica por contribuir con nuevos corpus paralelos en quechua sureño que mejoren los algoritmos y modelos de traducción automática para futuros proyectos.

1.5. Limitaciones

Para la realización de este proyecto se ha identificado una limitación referente al tamaño del corpus base ya que este es pequeño y no cuenta con muchos datos. Para ello se ha planteado alimentar con nuevos datos de diferentes fuentes de información desde los corpus del concurso de desarrollo de sistemas de traducción automática para las lenguas indígenas de las Américas, AmericasNLP (Association for Computational Linguistics, 2021).

2. MARCO CONCEPTUAL

En la presente sección presentamos los principales conceptos teóricos para el entendimiento de la tesis.

2.1. Lenguas minoritarias

Las lenguas minoritarias son aquellas lenguas cuya cantidad de hablantes es menor a la cantidad total de ciudadanos del país donde se origina la lengua.
Recolectar esta gran cantidad de recursos lingüísticos de una lengua minoritaria es difícil y costoso, no se cuentan con material traducido o material digitalizado en la lengua.

2.2. Corpus paralelo

Conjunto de oraciones en pares de tal manera que exista una referencia entre las oraciones en el lenguaje original y aquellas en el lenguaje objetivo. Es indispensable para el proceso de traducción automática, permite extraer características y comportamiento de las unidades léxicas de los lenguajes a procesar.

2.3. Web Scraping

Es un procesamiento para extraer textos en la web y redes sociales. Esta tarea puede realizarse manualmente o con herramientas de software libre. Para incrementar el corpus paralelo en quechua del traductor automático, se usaron librerías de web scrapping Tweepy para publicaciones en quechua en Twitter, Facebook-scrapper para extraer textos de grupos públicos en Facebook y PDFMiner en lecturas en archivos PDF.

2.4. Codificación de byte pares(BPE)

Debido al uso de sufijos en las palabras en quechua, se necesita una herramienta que permita procesar las sub-palabras de los textos antes de entrenar el modelo.

La codificación de byte pares o BPE (Sennrich, Haddow, & Birch) nos permite comprimir los datos y tokenizar a nivel de caracteres, luego se genera un vocabulario en forma de diccionario, donde a cada palabra tokenizada se le asocia su cantidad de ocurrencias en el texto.

2.5. Traducción automática

Es una tarea informática que codifica un texto en un lenguaje origen y lo decodifica el texto en un lenguaje destino usando computadoras. Para cumplir este objetivo, este sistema usando algoritmos aprende el contenido semántico de los dos idiomas a partir de grandes cantidades de textos paralelo o corpus paralelo. (Al-Onaiza)

2.5.1. Modelo de secuencia a secuencia

Este modelo permite aprender desde una red neuronal recurrente (RNN), prediciendo una secuencia de salida a partir de una secuencia de entrada. El modelo se apoya desde esta red que codifica la secuencia de elementos de entrada en un vector, conteniendo información de los elementos pasados de esta secuencia, para luego decodificar y desplegar este vector en una nueva secuencia de elementos. (Vinyals, y otros, 2014)

2.5.2. Back-translation

Es una técnica para aprovechar los corpus monolingües, al generar una traducción sintética, a partir de un modelo de traducción automática entrenado previamente, para obtener un corpus paralelo nuevo.

2.6. Métricas de medición

2.6.1. Exactitud (Accuracy)

La exactitud mide el porcentaje de casos que el modelo ha acertado, es decir el rendimiento general del modelo (Liu & Udell, 2020).

2.6.2. BLEU

El *Bilingual Evaluation Understudy Score* o BLEU es una métrica aplicada en el modelo base a nivel de palabra y sub-palabra. Permite comparar n-gramas del texto evaluado o candidato con los n-gramas de una referencia y contar el número de coincidencias. (Papineni, Roukos, Ward, & Zhu, 2002)

Al existir mayores coincidencias, mayor será la corrección propuesta del candidato. El puntaje obtenido, que varía entre 0 a 100, es útil para comparar oraciones, sin embargo se propone una versión modificada que permite normalizar los n-gramas por su aparición para mejorar la puntuación de bloques de oraciones.

Como ejemplo :

referencia = [[[‘allinpuniqa’, ‘tukukuypis’], [[‘allinpuniqa’, ‘tukukuypis’, ‘kashan’]]
candidato = [‘allinpuniqa’, ‘tukukuypis’, ‘kashan’]

Bleu Score: 1.0

2.6.3. ChrF

Es una métrica a nivel de caracteres, y mide el número mínimo de ediciones de caracteres requeridas que coincida con una referencia. Varía entre 0 a 1. (Popovic, 2015)

Como ejemplo :

referencia = [‘Haykaqmi’, ‘ripunki’]
predicción = [‘Haykaqmi’, ‘hamunki’]

ChrF Score: 0.52

3. ESTADO DEL ARTE

En la presente sección se tiene como objetivo revisar y mostrar los corpus utilizados, algoritmos y técnicas de trabajos anteriores en el campo de traducción automática en lenguas originarias, a fin de contextualizar la investigación a realizar.

3.1. Estrategia de búsqueda

3.1.1. Cadenas de búsqueda

Para cumplir este objetivo se realizó la búsqueda en base de datos IEEE y ACL. Las cadenas de búsqueda fueron las siguientes:

TITLE-ABS-KEY (neural machine translation AND (less OR under OR low) resource language)

TITLE-ABS-KEY (corpus creation AND (less OR under OR low) resource language)

3.1.2. Preguntas de revisión

Seguidamente de recolectar los artículos que cumplen con la cadena de búsqueda, se tienen las siguientes preguntas por evaluar:

- Pregunta 1: ¿Qué métodos o técnicas se utilizan para el tratamiento de corpus paralelos o monolingües de lenguas minoritarias en traducción automática
- Pregunta 2: ¿Qué algoritmo y arquitectura utilizaron sobre el corpus recolectado en la tarea de traducción automática?
- Pregunta 3: ¿Cómo validan y evalúan la calidad de los resultados obtenidos?

3.2. Resultados de la revisión

Se obtuvieron 16 investigaciones donde se seleccionaron 5 más relevantes que cumplen con las respuestas de la revisión.

En el estudio de (Tshephisho, Nelisiwe, & Phillemon, 2021) para traducción de los lenguajes del Sur de África como Siswati y Setswana a inglés, se recolectaron textos de la Biblia desde páginas web. Para el tratamiento realizaron diversas actividades como el eliminado de espacios en blanco, palabras repetidas, versos con más de 20 palabras y solo contando las de menor cantidad, además utilizaron la técnica de BPE para segmentar sub-palabras. Para el modelado utilizaron un modelo Transformer usando JoeyNMT. Para evaluar la calidad se utilizó BLEU Score obteniendo una precisión en el modelo adecuado.

En la investigación de (Góngora, Giossa, & Chiruzzo, 2021) se buscaba traducir la lengua Jopara a español usando textos de redes sociales, específicamente publicaciones de Twitter. En la recolección y tratamiento se listaron las palabras frecuentes, que fueran mínimo 10 veces en el corpus. Seguidamente se contaron el número de tokens guaraníes presentes en cada tweet utilizando la lista larga y analizando manualmente los conjuntos extraídos. Se generaron varios clústeres de palabras con la misma semántica y se entrenaron varias variantes de colecciones de incrustaciones dimensionales utilizando la biblioteca Gensim (Rehurek, 2010) y Wikipedia en Guaraní.

Por otro lado, (Kchaou, Boujelbane, & Hadrich Belguith, 2020) realizó una investigación de corpus paralelos en textos en Árabe Tunecino, una variante minoritaria del Árabe, utilizando comentarios en Facebook en este lenguaje. Este corpus fue traducido por un hablante nativo. Utilizando el kit de herramientas IRSTLM (Federico, Bertoldi, & Cettolo, 2008) se calculan modelos probabilísticos de lenguaje basado en n-gramas. También utilizan la técnica de back-translation (traducción inversa). Como resultado se obtuvo un valor de BLEU de 15.03.

En la investigación de (Feldman & Coto-Solano, 2020) se generó un modelo NMT usando back-translation en el lenguaje indígena costarricense Bribri a español, a partir de textos coleccionados en libros de gramática, diccionarios, literatura oral y de educación. Los autores realizaron la estandarización del sistema de escritura y diacríticos con un sistema determinista basado en reglas, con un resultado que suaviza algunas variaciones específicas de estos textos y que se puede convertir a cualquier tipo de ortografía contemporánea. Se implementó Transformer en PyTorch [10]y el paquete OpenNMT[11]. En los resultados de la evaluación usando BLEU se obtuvo 16.9 de score.

[10] https://pytorch.org/

En (Zhang, Frey, & Bansal, 2020), el autor desarrolló un modelo de traducción automática con el lenguaje minoritario Cherokee a inglés de origen de Estado Unidos. Se recolectaron páginas web, revistas de noticias y libros. Para la recolección, se utilizó la herramienta para extraer documentos escaneados OCR (Smith, 2007) en textos bibliográficos y en el modelado se hizo uso de traducción automática con redes neuronales y BERT (Devlin, Chang, Lee, & Toutanova, 2019). En los resultados se obtuvo un BLEU de 15.8 dentro del dominio y 6.5 con datos fuera del dominio.

4. EXPERIMENTACIÓN Y RESULTADOS

4.1. Modelo de traducción automática

4.1.1. Introducción

Para el experimento se inicia con la recolección de datos en corpus paralelos desde repositorios académicos y textos en la web. El modelo inicial constará de un corpus inicial, donde incrementalmente se agregarán nuevos corpus para mejorar la diversidad del traductor. Para los experimentos se utilizó la librería OpenNMT[12], el cual permite configurar los hiperparámetros y traducir un corpus paralelo. Se utilizó la versión compatible con Pytorch.

4.1.2. Procesamiento

- **AmericasNLP**

  Conjunto de datos del workshop de la conferencia anual de NLP, estos datos en quechua y español contienen textos de dominio religioso, Ministerio de Cultura de Perú y un diccionario completo (Mager, y otros, 2021).

- **Helsinki**

  Se obtuvieron los datos de la universidad de Helsinki (Vázquez, Scherrer, Virpioja, & Tiedemann, 2021) al participar en el workshop AmericasNLP, la cual cuenta con la constitución en quechua sureño y español de Bolivia y Perú, Tatoeba y Wikipedia.

[11] https://opennmt.net/
[12] https://opennmt.net/

- **REPU**

Research Experience for Peruvian Undergraduates - Computer Science, en siglas: REPU (Moreno, 2021) recolecta información usando datos del vocabulario en Duran[13], textos de la web[14] y traducciones de un manual educativo[15] en quechua con corpus paralelo.

- **Lecturas adicionales**

Textos en quechua y español, disponibles gratuitamente en la web, las cuales pertenecen a UNICEF, el Ministerio de la Mujer y la Municipalidad de Lima.

Tìtulo	Fuente
#YoMeQuedoEnCasa y la pasamos bien[16]	Ministerio de la Mujer y Poblaciones Vulnerables - UNICEF
Guia Paciencia y amor para ganarle al Coronavirus[17]	Ministerio de Salud - UNICEF
Lírica Quechua[18]	Municipalidad de Lima
Manual de Escritura Quechua Sureño[19]	Ministerio de Educación
Vocabulario Quechua Sureño[20]	Ministerio de Educación

Cuadro 4.1 Título y Fuente de lecturas adicionales.

El Cuadro 1 resume la cantidad de oraciones que existen en cada conjunto de datos recolectado.

[13] http://quechua-ayacucho.org/es/index_es.php
[14] https:// lyricstranslate.com/
[15] Runasimita yachasun. Cesar Iter and Zenobio Ortiz Cárdenas. 2019.
[16] https://www.unicef.org/peru/informes/yomequedoencasa-y-la-pasamos-bien
[17] https://www.unicef.org/peru/informes/paciencia-y-amor-para-ganarle-al-coronavirus
[18] https://www.descubrelima.pe/coleccion-lima-lee/lirica-quechua-de-tradicion-oral-autoctona/
[19] https://repositorio.minedu.gob.pe/handle/20.500.12799/7190
[20] https://repositorio.minedu.gob.pe/handle/20.500.12799/7490

Figura 4.1 Oraciones totales por corpus paralelos recolectados.

La Figura 4.1 permite visualizar la cantidad de oraciones totales entre las bases, detectando que la diversidad de textos en AmericasNLP (medido en un ratio V/N o palabras únicas sobre el total de palabras), principalmente por los textos del corpus JW300 (Agić & Vulić, 2019) que son de dominio religioso, es el menor con 0.13 (ver Figura 4.2). Esto generaría un bajo resultado al implementar el modelo con esta base inicial.

Figura 4.2 Ratio V/N en corpus quechua sureño.

En el experimento, se usará los datos de AmericasNLP como línea base, sobre el cual se añadirán textos nuevos como los recolectados en REPU, textos adicionales en PDF y redes sociales. Como primera parte, AmericasNLP es un corpus paralelo por lo cual se consiguieron textos extra que cumplan lo mismo. En la Figura 4.3, se visualiza un resumen de los corpus paralelos recolectados y la cantidad de oraciones.

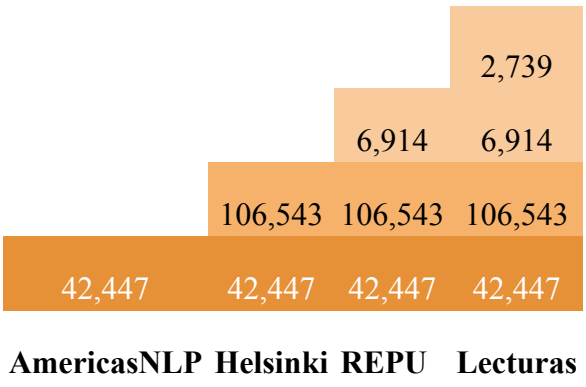

Figura 4.3 Oraciones acumuladas por corpus.

Las oraciones monolingües también forman parte importante al elaborar un modelo de traducción automática. Estas oraciones ingresan a un nuevo modelo usando una técnica de backtranslation, donde se generan oraciones sintéticas, obteniendo así un nuevo corpus paralelo que puede ser utilizado para tareas de traducción.

Para cumplir ello, se recolectaron textos y oraciones desde lecturas en archivos PDF y redes sociales, estas oraciones son monolingües y contienen diversas variaciones. Luego de un proceso de limpieza de caracteres y quedarnos solo con textos en quechua sureño, se obtuvo una única base para el modelamiento. En el caso de Twitter y Facebook, se realiza un listado inicial se usuarios y grupos que difundían quechua ayacuchano, seguidamente se realizó web scrapping seguro sobre estos enlaces, donde solo se obtuvo el texto de las publicaciones. La Figura 4.4 muestra la composición de este corpus.

En resumen, los datos recolectados cuentan con las siguientes características, donde visualmente las publicaciones de Facebook contienen gran cantidad de oraciones sobre las bases recolectadas, la hipótesis permite validar que la diversidad de textos obtenidos desde otras fuentes permita mejorar el resultado obtenido de esta base.

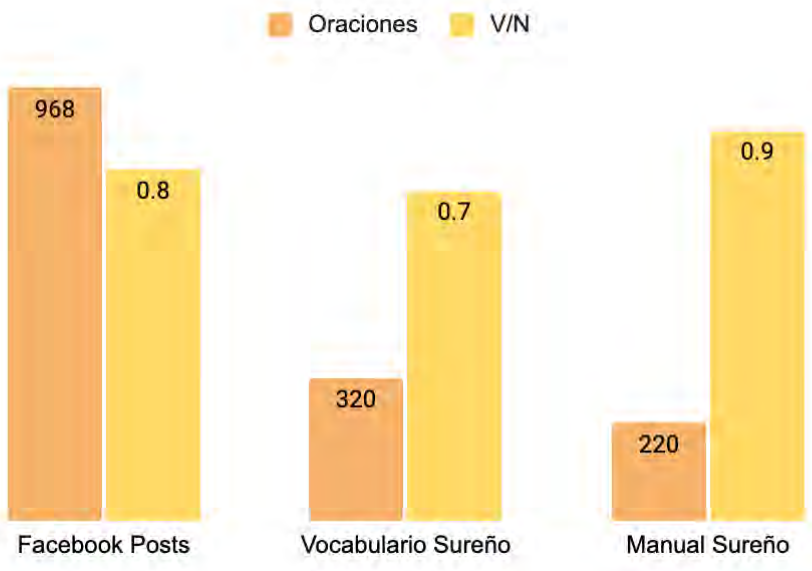

Figura 4.4 Estadísticas resumen de corpus monolingües.

4.1.3. Modelo Transformer Base

El modelo inicial consta de los textos del workshop AmericasNLP, en el siguiente cuadro visualizamos su distribución de entrada al modelo. Se escogen 1000 oraciones aleatorias de la base de entrenamiento y se agregan al conjunto de datos oficial de validación de AmericasNLP.

Fuente	Oraciones Iniciales	Train	Validación	Test
AmericasNLP	118,595	117,570	1,994	1,003

Cuadro 4.2 Cantidad de oraciones en el conjuntos de entrenamiento modelo AmericasNLP.

Como primer paso generamos la codificación BPE con los textos quechua sureño y español de entrenamiento de este repositorio. Seguidamente aplicamos esta codificación a nuestros archivos de validación y pruebas.

Luego de generar el vocabulario de nuestros textos, usando OpenNMT construiremos un modelo Transformer con 2 capas. Después de entrenar el modelo, validamos los textos con nuestro archivo de referencia.

4.1.4. Modelo Transformer en corpus Helsinki y REPU

Los siguientes dos modelos corresponden a los repositorios de Helsinki y REPU, ambos participantes del workshop AmericasNLP. Como se visualiza, existe mayor cantidad de oraciones de Helsinki sobre REPU para el entrenamiento.

Fuente	Oraciones Iniciales	Train	Validación	Test
AmericasNLP	118,595	117,570	1,994	1,003
Helsinki	153,653	271,223	1,994	1,003
REPU	9,443	280,666	1,994	1,003

Cuadro 4.3 Cantidad de oraciones en entrenamiento Helsinki y REPU.

De la misma forma que el modelo base, aplicamos la codificación BPE inicial a nuestros archivos de validación y pruebas de estos repositorios. Se desarrollan 2 modelos, ambos usando modelos Transformer en OpenNMT, entrenamos nuestros textos.

4.1.5. Modelo Transformer en corpus extra

En esta etapa, añadimos los textos adicionales recolectados, generando casi 281 mil oraciones que pasaran por el modelo Transformer de dos capas para generar un traductor más diverso. Se replican los pasos de preprocesamiento de textos y codificación.

Fuente	Oraciones Iniciales	Train	Validación	Test
AmericasNLP	118,595	117,570	1,994	1,003
Helsinki	153,653	271,223	1,994	1,003
REPU	9,443	280,666	1,994	1,003
Lecturas	1,093	281,759	1,994	1,003

Cuadro 4.4 Cantidad de oraciones en entrenamiento texto adicional.

4.1.6. Modelo Transformer y Fine-Tuning

Fine-Tuning es el proceso de tomar pesos de una red neuronal entrenada y usarla como inicialización para un nuevo modelo que se entrena con datos del mismo dominio. Esta actividad nos permitirá acelerar el entrenamiento y superar el tamaño pequeño del conjunto de datos.

Se escoge el modelo Helsinki como pre-entrenamiento y se añaden nuevos textos para mejorar el modelo. Para tokenizar las subpalabras se usa Byte Pair Encoding (BPE)[21] sobre el source y target de AmericasNLP y Helsinki. Se utilizan los mismos parámetros del modelo inicial.

Fuente	Oraciones Iniciales	Base Train
REPU	9,443	9,443
Lecturas	1,093	10,536
Monolingües con Back-Translation	1,414	11,950

Cuadro 4.5 Modelado con Fine-Tuning desde corpus Helsinki.

[21] https://aclanthology.org/P16-1162/

4.1.7. Resultados obtenidos

En el siguiente cuadro se visualizan los resultados obtenidos de los modelos usando las métricas BLEU y chrF. El valor de los modelos con las bases Helsinki y Lecturas obtienen mejor rendimiento en la traducción.

Modelos	Bases utilizadas	BLEU	chrF
Modelo A	AmericasNLP	6.76	0.18
Modelo B	AmericasNLP + Helsinki	**13.78**	**0.24**
Modelo C	AmericasNLP + Helsinki + REPU	11.46	0.21
Modelo D	AmericasNLP + Helsinki + REPU + Lecturas	**13.96**	**0.24**
Modelo E	AmericasNLP + Helsinki + REPU + Lecturas + Monolingües	13.57	0.23

Cuadro 4.6 Modelado desde linea base AmericasNLP.

En las Figuras 4.5 y 4.6 se visualizan la comparación entre los modelos implementados sobre el modelo base, comprobando que el Modelo D tiene mejor rendimiento en la traducción de textos de quechua sureño a español con un puntaje mayor en las dos métricas.

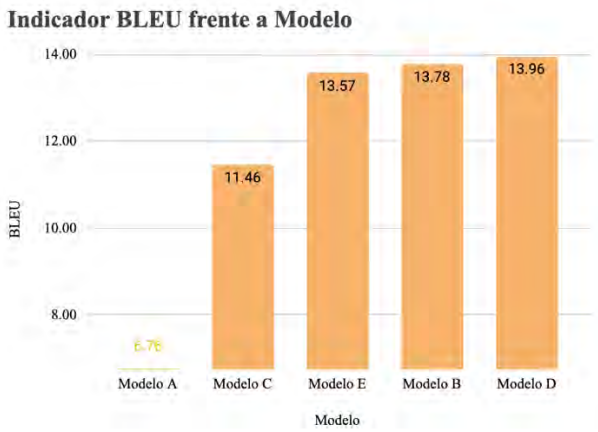

Figura 4.5 Indicador BLEU sobre modelos.

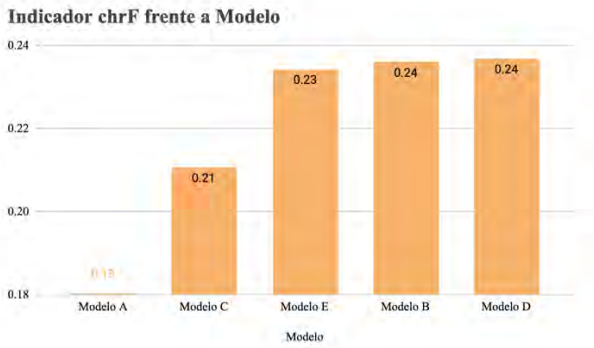

Figura 4.6 Indicador chrF sobre modelos.

Seguidamente se generaron modelos nuevos usando fine-tuning a partir de la base Helsinki. Al realizar el fine-tuning se visualiza que el rendimiento es mucho mejor en el uso de las bases Lecturas.

Modelos	Bases utilizadas	BLEU	chrF
Modelo C2	REPU	12.03	0.23
Modelo D2	REPU + Lecturas	**15.99**	**0.25**
Modelo E2	REPU + Lecturas + Monolingües	11.26	0.24

Cuadro 4.7 Modelado con Fine-Tuning desde corpus Helsinki

En las Figuras 4.7 y 4.8 se visualiza que modelo D2 obtienen mejores resultados en BLEU y chrF sobre el modelo base, esto debido al fine-tuning y a la diversidad textos y dominios nuevos de las lecturas recopiladas.

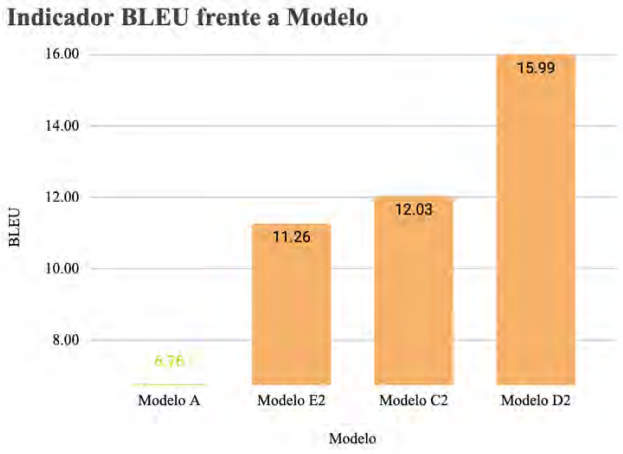

Figura 4.7 Indicador BLEU sobre modelos con fine tuning

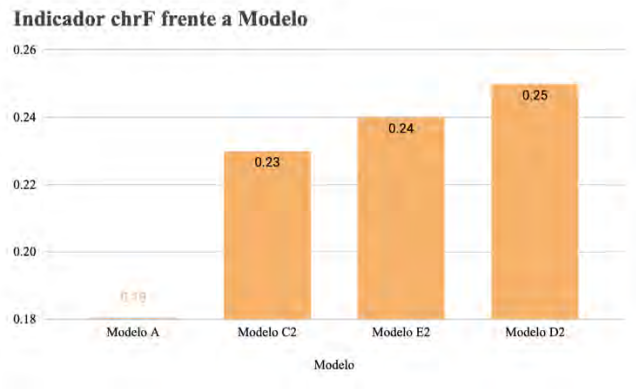

Figura 4.8 Indicador chrF sobre modelos con fine tuning

Para complementar el resultado, en el siguiente cuadro se añaden sus respectivos pasos de entrenamiento y el accuracy más alto alcanzado durante el entrenamiento.

Modelos	BLEU	chrF	Step	Accuracy
Modelo A	6.76	0.18	10,000	65.10
Modelo B	**13.78**	**0.24**	**29,500**	**43.09**
Modelo C	11.46	0.21	23,500	43.72
Modelo D	**13.96**	**0.24**	**23,000**	**43.10**
Modelo E	**13.57**	**0.23**	**26,500**	**44.65**

Cuadro 4.8 Steps y Accuracy por modelo inicial

Observamos que el Modelo B, que corresponde al generado usando la base AmericasNLP y Helsinki obtiene el máximo cantidad de pasos de entrenamiento, esto debido a la variedad y diversidad de textos.

El Modelo D , que utiliza una base de lecturas recolectadas incluidas las de UNICEF y el Ministerio de Salud, es el que tiene el BLEU más alto con 13.96, demostrando un mejor desempeño al traducir los textos de quechua sureño a español.

Seguidamente, al añadir las bases monolingües en el Modelo D, se visualiza una ligera baja en el indicador BLEU, alcanza mayores pasos de entrenamiento y accuracy. Este resultado no es originado por el corpus en sí, el modelo de traducción automático utilizado sobre esta base para obtener sus base paralelo en español tiene deficiencias y no tiene una alta calidad.

Para validar esto, se realizó un experimento adicional para entrenar únicamente el modelo monolingue sín acumular con las bases anteriormente mencionadas.

Modelos	BLEU	chrF	Step	Accuracy
Modelo Monolingüe base	2.88	0.17	3,000	9.42

Cuadro 4.9 Resultados Modelo Monolingüe Base

Como se observa en el cuadro 4.9, al utilizar unicamente el corpus con textos monolingues con su traducción sintética no genera un apropiado resultado, porque no es una traducción real.

Modelos	BLEU	chrF	Step	Accuracy
Modelo C2	3.66	0.24	32,500	81.14
Modelo D2	6.82	0.23	32,500	17.24
Modelo E2	**8.56**	**0.23**	**32,500**	**88.11**

Cuadro 4.10 Steps y Accuracy por modelo en fine tuning

En el cuadro 4.10 vemos que los 3 modelos utilizando fine-tuning desde la base Helsinki alcanzan los mismos pasos de entrenamiento. Notamos un BLEU y accuracy más alto al añadir los textos monolingües. Estos resultados se pueden deber a un mejor rendimiento al emplear fine-tuning desde un modelo estable y textos con contextos diversos. Sin embargo, estos resultados no son absolutamente mejores que los obtenidos por el Modelo D.

4.2. Desarrollo de herramienta de consultas de corpus

Se desarrolla un sitio web a partir de la plataforma *Elotl*(Esquite, 2022), que permita realizar una búsqueda de palabras o frases en corpus paralelos. A partir de todos los repositorios y textos recolectados en el proyecto, se consolidan a una base de datos maestra donde finalmente serán consultados desde esta aplicación. En la Figura 4.9 se muesta la pantalla inicial de búsqueda de textos.

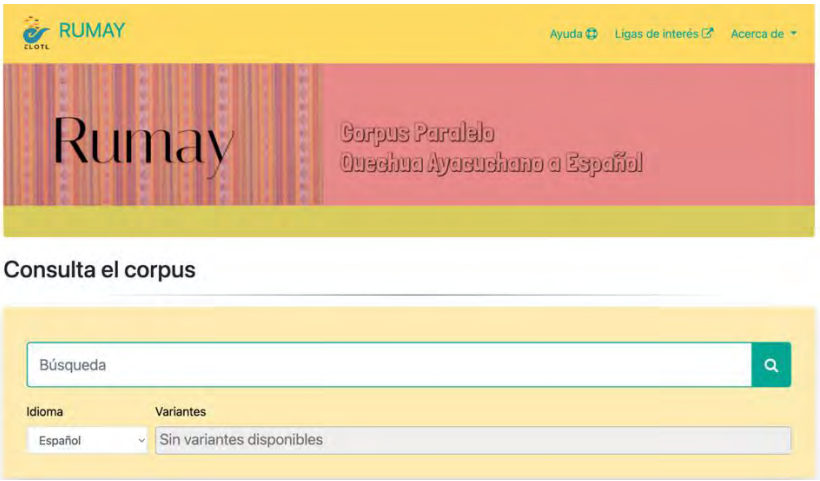

Figura 4.9 Pantalla inicial del buscador.

Este buscador permite consultar palabras en quechua sureño, y encontrar su traducción desde oraciones o párrafos en español que contengan dichas palabras. Además, se permite visualizar un documento de referencia en PDF con los textos utilizados.

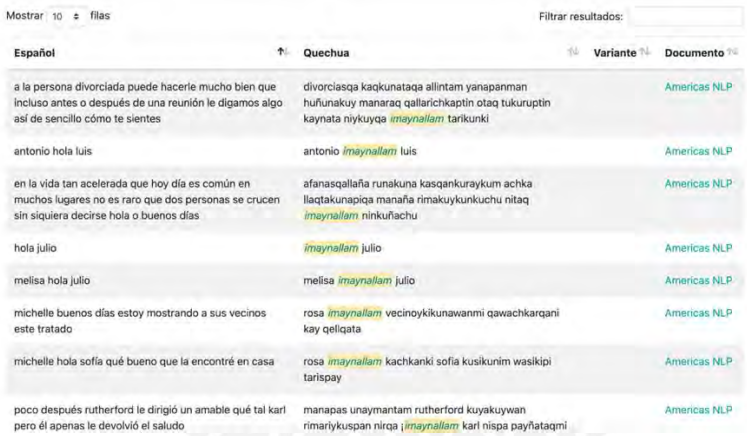

Figura 4.10 Búsqueda de palabras en corpus paralelo.

Para visualizar el funcionamiento del proyecto web, se puede consultar el siguiente enlace.

5. CONCLUSIONES Y TRABAJOS FUTUROS

Conclusiones

- Se recolectaron textos en quechua sureño en la web para su aplicación en traducción automática, usando como linea base los repositorios de AmericasNLP.
- Se extrayeron y limpiaron textos escritos en quechua sureño desde fuentes no tradicionales como publicaciones de Twitter y grupos de Facebook.
- Se propusieron modelos de traducción automatica de quechua sureño a español de calidad, usando el framework OpenNMT y un modelo Transformer, los cuales fueron evaluados con BLEU y chrF a partir de los textos recolectados. A partir de dichos modelos, se identificó que los corpus paralelos recolectados mejoraron la calidad de la traducción, pero no así los corpus monolingües.
- Se implementó una página web para consultar y visualizar palabras y frases en quechua sureño y sus textos en paralelo en español, para su difusión y aprendizaje.

5.1. Trabajos futuros

- Implementar modelos de traducción automática en la dirección de español a quechua sureño, y evaluar la salida del modelo con un hablante nativo de quechua sureño.
- Evaluar otro tipo de modelos diferentes a Transformer, así como continuar implementaciones a nivel de sub-palabras y palabras.
- Agregar más variantes de quechua, para así generar más diversidad en el modelo y en la plataforma de consulta de textos.

Bibliografía

Tshephisho, S., Nelisiwe, G., & Phillemon, S. N. (2021). *Transformer-based Machine Translation for Low resourced Languages embedded with Language Identification.* South Africa: IEEE.

Góngora, S., Giossa, N., & Chiruzzo, L. (2021). *Experiments on a Guarani Corpus of News and Social Media.* Montevideo: ACL.

Kchaou, S., Boujelbane, R., & Hadrich Belguith, L. (2020). *Parallel resources for Tunisian Arabic dialect translation.* Tunisia: ACL.

Feldman, I., & Coto-Solano, R. (2020). *Neural Machine Translation Models with Back-Translation for the Extremely Low-Resource Indigenous Language Bribri.* Barcelona: ACL.

Zhang, S., Frey, B., & Bansal, M. (2020). *ChrEn: Cherokee-English Machine Translation for Endangered Language Revitalization.* Chapel Hill: ACL.

Sullon Acosta, K. (2013). *Ministerio, ocumento nacional de lenguas originarias del Perú.* Ministerio de Educación (MINEDU).

Forcada, M. (2006). *Open source machine translation: an opportunity for minor languages.* LREC.

Ministerio de Educación. (2016). *Resolución Ministerial N°629-2016- MINEDU.* Lima.

Ministerio de Educación. (2015). *Resolución Ministerial Nª 303-2015-MINEDU* . Lima.

Gobierno del Perú. (2017). *Decreto Supremo N° 005-2017-MC.* Lima.

Poder Legislativo. (2011). *Ley Nº 29735.* Lima: El Peruano.

Association for Computational Linguistics. (2021). *AmericasNLP 2021 Shared Task on Open Machine Translation*. Retrieved from http://turing.iimas.unam.mx/americasnlp/st.html

Popovic, M. (2015). *CHRF: character n-gram F-score for automatic MT evaluation.* Germany.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). *BLEU: a Method for Automatic Evaluation of Machine Translation.*

Mager, M., Oncevay, A., Ebrahimi, A., Ortega, J., Rios, A., & Fan, A. (2021). *Findings of the AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas.* Association for Computational Linguistics.

Vázquez, R., Scherrer, Y., Virpioja, S., & Tiedemann, J. (2021). *The Helsinki submission to the AmericasNLP shared task.* Association for Computational Linguistics .

Moreno, O. (2021). *The REPU CS' Spanish–Quechua Submission to the AmericasNLP 2021 Shared Task on Open Machine Translation.* Association for Computational Linguistics .

Liu, B., & Udell, M. (2020). *Impact of Accuracy on Model Interpretations.*

Esquite, E. (2022). *Elotl*. Retrieved from https://elotl.mx/

Sennrich, R., Haddow, B., & Birch, A. (n.d.). *Neural Machine Translation of Rare Words with Subword Units.* Berlin: Association for Computational Linguistics.

Al-Onaiza, Y. (n.d.). *Statistical Machine Translation.* Final Report: JHU Summer Workshop.

Vinyals, O., Kaiser, L., Koo, T., Petrov, S., Sutskever, I., & Hinton, G. (2014). *Grammar as a Foreign Language.* CoRR abs/1412.7449.

Rehurek, R. (2010). *Software Framework for Topic Modelling with Large Corpora.* Valleta: ELRA.

Federico, M., Bertoldi, N., & Cettolo, M. (2008). *IRSTLM: an open source toolkit for handling large scale language models.* ISCA.

Smith, R. (2007). *An Overview of the Tesseract OCR Engine.* USA: ICDAR.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.*

Agić, Ž., & Vulić, I. (2019). *JW300: A Wide-Coverage Parallel Corpus for Low-Resource Languages.* Florence: ACL.