ISBN: 978-628-95207-8-1. ISSN: 2414-6390. Digital Object Identifier: 10.18687/LACCEI2024.1.1.1636

Development of a neural machine translation model optimized with BERT for translation from Quechua to Spanish

Beatrice Cueva Medina, Bachelor[1] 0009-0003-7144-0193, Gabriel Fabrizio Tuco Casquino, Bachelor[1] 0000-0003-3443-4504, and José Sulla-Torres, Doctor[1] 0000-0001-5129-430X
[1]Universidad Católica de Santa María, Arequipa-Perú, 75910710@ucsm.edu.pe, gabriel.tuco@ucsm.edu.pe
jsullato@ucsm.edu.pe

Abstract– ***Quechua, a Native American language spoken by over 3 million people in Peru, plays a significant cultural role but is at risk of decline due to limited resources and the dominance of Spanish. This paper proposes a Quechua-to-Spanish neural machine translation (NMT) model using a Transformer-based architecture and a semi-supervised approach known as LMfusion. The model is trained on parallel datasets, and PRPE morphological segmentation is employed during preprocessing. Initial results show promise, and integrating the QuBERT language model is expected to enhance translation quality. Additionally, a user-friendly web interface has been developed to facilitate Quechua-Spanish translation. This research aims to address the challenges of translating a low-resource language like Quechua and contribute to improved communication between Quechua and Spanish speakers, preserving cultural heritage and facilitating equitable access to information and services.***

Keywords-- ***Quechua, Neural Machine Translation, Low-resource Language, BERT, Transformer, PRPE.***

Digital Object Identifier: (only for full papers, inserted by LACCEI).
ISSN, ISBN: (to be inserted by LACCEI).
DO NOT REMOVE

Desarrollo de un modelo de traducción automática neuronal optimizado con BERT para la traducción del idioma quechua al español

Beatrice Cueva Medina, Bachelor[1] 0009-0003-7144-0193, Gabriel Fabrizio Tuco Casquino, Bachelor[1] 0000-0003-3443-4504, and José Sulla-Torres, Doctor[1] 0000-0001-5129-430X
[1]Universidad Católica de Santa María, Arequipa-Perú, 75910710@ucsm.edu.pe, gabriel.tuco@ucsm.edu.pe
jsullato@ucsm.edu.pe

Resumen– ***El quechua es una lengua nativa americana hablada por más de 3 millones de personas en Perú, desempeña un papel cultural importante, pero corre el riesgo de declinar debido a los recursos limitados y al predominio del lenguaje español. Este artículo propone un modelo de traducción automática neuronal (NMT) del quechua al español utilizando una arquitectura basada en Transformer y un enfoque semi-supervisado conocido como LMfusion. El modelo se entrena en conjuntos de datos paralelos y se emplea la segmentación morfológica PRPE durante el preprocesamiento. Los resultados iniciales son prometedores y se espera que la integración del modelo de lenguaje QuBERT mejore la calidad de la traducción. Además, se ha desarrollado una interfaz web fácil de usar para facilitar la traducción quechua-español. Esta investigación tiene como objetivo abordar los desafíos de traducir una lengua de bajos recursos como el quechua y contribuir a mejorar la comunicación entre los hablantes de quechua y español, preservando el patrimonio cultural y facilitando el acceso equitativo a la información y los servicios.***

Palabras clave-- Quechua, Traducción Automática Neuronal, Lenguaje de bajos recursos, BERT, Transformer, PRPE.

I. INTRODUCCIÓN

En el Perú, más de 4 millones de personas tienen como lengua materna una lengua nativa, entre ellos, alrededor de 3 millones 375 682 son quechua hablantes, esto equivale al 13,9% de la población peruana [1]. En la región Arequipa, según [2], se cuenta con 227 600 personas que aprendieron el quechua en la niñez, no obstante, la cantidad de personas hispanohablantes representa al 79,6% de toda esta región. El idioma quechua representa a una gran parte de historia y cultura peruana, sin embargo, está en peligro de desaparecer debido a diversos factores, entre ellos se presenta el poco interés o dificultad para el aprendizaje del idioma quechua, ya que en el Perú el lenguaje predominante es el español [3], lo que presiona a las poblaciones quechuas a tener que priorizar hablarlo y por ende se ven obligados a olvidar sus raíces con el fin de buscar un mejor futuro para ellos y sus familias en otras localidades del Perú [4] [5]. Por otra parte, según [6] se tienen registrados alrededor de 250 traductores e intérpretes del quechua en todo el Perú, donde solo 9 de ellos son pertenecientes a la región Arequipa, esta cifra no es suficiente para poder cubrir la demanda de intérpretes o traductores que conozcan tanto el idioma quechua como el español lo que condiciona que las personas quechua hablantes no puedan acceder a una comunicación efectiva [7].

La correcta utilización de tecnologías que involucran el lenguaje humano, como la traducción automática, ofrece numerosas posibilidades para revitalizar, difundir y aprender nuestros idiomas. Esto incluye fomentar la comunicación entre generaciones, como la interacción entre abuelos que hablan una lengua indígena y sus nietos que hablan español. Además, facilita el acceso a servicios e información de manera más equitativa y democrática. La lingüística computacional es la herramienta más potente para la revitalización de las lenguas nativas, sin embargo, la falta de recursos impide la aplicación de tecnologías del lenguaje humano. [8]

En el campo de la traducción automática neuronal (NMT) la falta de recursos de los lenguajes llega a tener un gran impacto en sus resultados. Para ello, diferentes autores proponen diferentes técnicas para superar este problema. Usualmente la NMT cae en los enfoques supervisados, semi supervisados y no supervisados. En los enfoques supervisados, se cuenta con una gran cantidad de datos paralelos, por lo que no puede ser tan ´útil para lenguajes con bajos recursos (LRL, de sus siglas en ingles), sin embargo, se puede utilizar técnicas como el aumento de datos, para la generación de datos paralelos sintéticos. En enfoques semi supervisados, se cuenta con un reducido conjunto de datos paralelos y grandes cantidades de datos monolingües, dentro de sus técnicas se puede encontrar el uso de datos monolingües para generar modelos de lenguaje (LM), los cuales pueden ser integrados en los modelos NMT.

Esta técnica se le denomina LM-fusion, generalmente se categoriza en Shallow fusion; donde se integra el LM en la salida o en el entrenamiento de los modelos NMT, y Deep fusión, donde se modifica la arquitectura del modelo NMT para integrar el LM. Por último, en los enfoques no supervisados, no se cuenta o se tiene muy pocos datos paralelos, para esto usualmente se usa Redes Generativas Adversariales (GAN, de sus siglas en ingles). [9]

El quechua es un lenguaje con pocos recursos puesto que no tiene una gran presencia internet, falta de apoyo por parte del gobierno peruano u otros motivos. además, los trabajos actuales [10] [11] también identifican esta situación. Como menciona [9] la lingüística computacional en bajos recursos, requiere de técnicas más allá del entrenamiento básico de modelos, ya que el procesamiento de una nueva lengua

Digital Object Identifier: (only for full papers, inserted by LACCEI).
ISSN, ISBN: (to be inserted by LACCEI).
DO NOT REMOVE

conlleva a nuevos desafíos como sistemas fonológicos especiales, problemas de segmentación de palabras, estructuras gramaticales borrosas, lenguaje no escrito, etc. Por otra parte, trabajos actuales brindan una base con diferentes técnicas, arquitecturas, modelos y la creación de nuevos conjuntos de datos (también llamados corpus) paralelos y monolingües, que pueden mejorar la aplicación de las tecnologías al quechua.

El quechua es una familia lingüística originada en América del Sur, mayormente concentrado en los países de Per´u, Bolivia, Ecuador y en menor concentración al sureste de Colombia, norte de Argentina y noreste de Chile [15]. El quechua es un lenguaje polisintético, el cual contiene una gran cantidad de morfemas (pequeñas palabras que pueden cambiar el significado), se dice que es polisintético ya que aglutina varios morfemas como sufijos que cambian el significado de la palabra raíz, ya sea un sustantivo o un verbo. [12]

Por lo anteriormente mencionado, este trabajo pretende brindar un traductor del quechua al español, utilizando un modelo NMT basado en transformer con un enfoque semi supervisado basado en la técnica LM-fusion el cual será implementado en una página web para uso.

II. Estado del Arte

En esta sección se cubrirá los trabajos relacionados a la traducción del quechua y la traducción de idiomas en lenguajes con pocos recursos. Actualmente se han aumentado el número de trabajos relacionados a la traducción y tratamiento del quechua, de los cuales se vio un incremento en la utilización de modelos traducción automática neuronal como se ve en el trabajo de [12] donde realizan varias técnicas e implementan modelos para la traducción del quechua al español. En este trabajo consideran la segmentación morfológica, donde muestran su algoritmo llamado algoritmo BPE-Guided para realizar la segmentación morfológica del quechua. Se presentó enfoques para NMT, para esto utilizan enfoques basados en reglas y en redes neuronales (incluyendo LSTM y Transformer). Para implementar los modelos NMT utilizaron openNMT; un kit de herramientas de código abierto para la traducción automática neuronal [13] . También se usó técnicas como la retro Traducción (BackTranslation) y la normalización, las cuales fueron integradas con el desarrollo de los modelos. Como resultados obtuvieron que los modelos basados en LSTM dieron un mejor resultado en la métrica BLEU, mencionan las técnicas de segmentación mejoraron los resultados de los modelos.

En [11] también se hace uso de los modelos NMT para la traducción del quechua al español, especialmente la arquitectura Transformer. Para la vectorización de los tokens los autores utilizaron las técnicas de análisis de componentes principales (PCA) y la distancia euclidiana. Para entrenar el su modelo utilizaron el corpus OPUS JW300 con aproximadamente 145 mil sentencias. Al cabo de 25 épocas consiguieron un puntaje de 30 puntos de BLEU.

Por otra parte, para ayudar a mitigar la falta de recursos en el quechua, los autores de [10] brindan un corpus considerablemente grande, curado y monolingüe de quechua sureño que consta de casi 450.000 segmentos. Además, se presenta técnicas de normalización y de tokenización, y un modelo preentrenado llamado QuBERT para la utilización de este corpus en tareas de part-of-speech (POS) tagging y Named Entity Recognition (NER). Como parte del modelo QuBERT se probó con diferentes tokenizadores como Byte-pair encoding (BPE), BPE-Guided y Prefix-Root-Postfix-Encoding (PRPE) y por último se hizo un ajuste fino en las tareas de NER Y POS. Los resultados obtenidos mostraron que en la tarea de NER se obtuvo un puntaje un poco más bajo que lenguajes de altos recursos, pero en el puntaje F1 se obtuvo una punta je en línea con otras investigaciones de bajos recursos y en la tarea de POS taggin se obtuvo buenos resultados obteniendo más de 80% en la puntuación de la precisión. Además, encontraron un mayor resultado en la segmentación con el algoritmo PRPE.

La falta de un conjunto de datos paralelo logra ser un gran obstáculo para la traducción automática. Comúnmente en diferentes investigaciones se trata de explotar corpus monolingües para aumentar el corpus paralelo, como en el trabajo de [14] donde presentan un enfoque que integra un modelo de lenguaje en el entrenamiento de un modelo de traducción automática. Para esto un LM se usa en el lado de la salida del modelo de traducción, sin modificar su arquitectura. Además, se agrega un término de regularización que orienta las distribuciones de salida del modelo de traducción para que sean coherentes con las distribuciones del modelo de lenguaje, esto permite al modelo de traducción omitir al ML cuando sea necesario. Resaltan que solo se utiliza durante el entrenamiento consiguiendo traducciones más rápidas que métodos de fusión. Para probar su nuevo enfoque, utilizaron distintos modelos de traducción para la traducción del lenguaje turco, dando como resultado un aumento de 1.8 puntos en BLEU.

El quechua es uno de los lenguajes suramericanos con una gran cantidad de hablantes nativos y estos suelen agregar alguno tonos o palabras diferentes según la localidad. [10] Según Alfredo Toreto [16], se reporta dos principales divisiones del quechua (Quechua I y Quechua II). El quechua II es mayormente hablado en el sur del Perú, el cual se divide en más variantes del quechua. Esta investigación se centrará generalmente en el quechua sureño.

III. Materiales y Metodología

A. Metodología.

La metodología planteada para este proyecto fue establecida con la investigación del estado del arte de diferentes artículos que hablan sobre procesamiento del lenguaje natural. En este artículo se utiliza la metodología CRIPS-DM el cual es un modelo de proceso para minería de datos, este consiste en 6 fases iterativas desde la comprensión

empresarial hasta la implementación [17], que pueden ser vistos en la figura 1. Esta metodología provee un conjunto de procedimientos para completar proyectos de machine learning. [18].

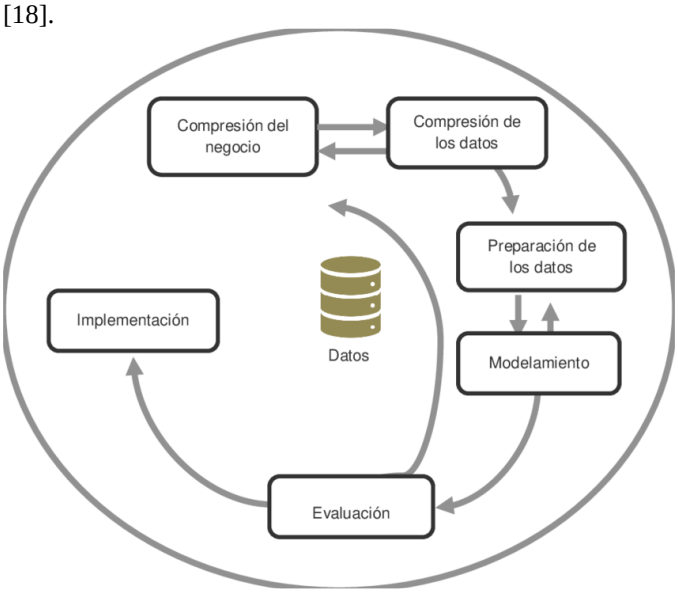

Fig. 1 Estructura de CRIPS-DM [19].

B. Materiales

1) Transformer: A partir de la creciente evolución del Procesamiento de Lenguaje Natural y debido a su continua mejora, surgieron diversos métodos y arquitecturas que permitirían generar resultados más precisos y acordes a la realidad en la comunicación humana, entre ellos se encuentra la arquitectura de los Transformers.

Los Transformers son modelos industrializados y homogenizados de aprendizaje pos-profundo diseñados para el funcionamiento paralelo en super computadoras. Esto permite a los Transformers a poder trabajar con una gran cantidad de datos sin etiquetar, es decir, datos en bruto sin necesidad de ajuste, en conjunto con una gran cantidad de parámetros para dichos datos mediante el aprendizaje auto-supervisado [20].

Por otra parte, según [21] indican que Transformer es una arquitectura de redes neuronales, la cual sirve para capturar patrones en largas secuencias de datos, así como para lidiar con grandes conjuntos de datos.

La arquitectura de un Transformer se basa principalmente en los módulos de codificación y decodificación como se puede observar en la Figura 2.

2) Bidirectional Encoder Representation from Transformer (BERT): La creación de los Transformers abrió paso al desarrollo de diversos modelos para el Procesamiento del Lenguaje Natural, entre ellos se encuentran el modelo de OpenAI GPT, T5 y, el que profundizaremos en esta investigación, BERT.

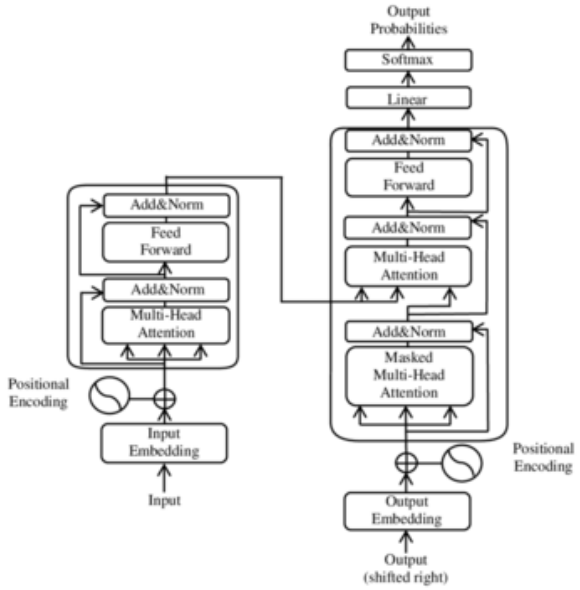

Fig. 2. Arquitectura de un Transformer

El Bidirectional Encoder Representation from Transformer es un modelo pre entrenado bidireccional, que permite crear modelos según el enfoque que se le brinde para una amplia gama de tareas entorno al NLP, sin necesidad de configurar o modificar la arquitectura de manera sustancial [22].

BERT es un modelo basado en el contexto, es decir, este modelo entiende el contexto y luego genera la incrustación para la palabra basada en el contexto. Este modelo fue creado con el objetivo de proveer mejores resultados en muchas tareas del NLP, como respuestas a preguntas, generación de textos, clasificación de oraciones. entre otras [23].

3) *Datos*: Para este trabajo se realizó una investigación de los diferentes trabajos con respecto a la traducción del quechua al español, de los cuales se recolecto diferentes conjuntos de datos paralelos (textos en quechua y su traducción al español) utilizados en estos trabajos. Se puede ver más detalle de los conjuntos de datos utilizados en este trabajo en la Tabla I.

TABLA I
TABLA DE RECOPILACIÓN DE DATOS

ID	Conjunto de datos	Entrenamiento	Prueba
1	AmericasNLP 2021 Shared Task on Open Machine Translation	103k	12.8k
2	OPUS the open parallel corpus	0.4k	

4) TensorFlow: TensorFlow es una plataforma de código abierto enfocada en el Machine Learning que permite a los principiantes o expertos crear sus propios modelos de Machine Learning [24]. Además, esta plataforma brinda modelos ya realizados a disposición de las personas, incluyendo modelos como Transformer o BERT. Por otro parte, se pueden encontrar diferentes conjuntos de datos publicados por la misma comunidad.

D. *Métodos*

1) Preprocesamiento: Se realiza la limpieza de los datos excluyendo los símbolos que no sean relevantes para el entrenamiento de la traducción. Además, se implementan distintos métodos, ya que al tratarse del quechua se deben tener en cuenta las normas que se aplican y su estructura semántica, como la segmentación morfológica, el cual es un proceso linguístico que implica dividir las palabras en sus componentes más pequeños, llamados morfemas, con el fin de comprender su estructura gramatical y semántica. Este proceso es especialmente relevante en idiomas altamente morfológicos o polisintéticos, como el quechua, donde pequeños cambios en la estructura de una palabra pueden alterar significativamente su significado [12]. Para realizar una segmentación se cuentan con dos métodos diferentes, sin embargo, en nuestro proyecto trabajamos con PRPE (Prefix-Root-Postfix-Encoding) es un algoritmo propuesto en este artículo que se utiliza como parte del preprocesamiento de texto en la traducción automática. Su objetivo principal es segmentar las palabras de manera cercana a su estructura morfológica en diferentes idiomas. Lo que hace especial a PRPE es su capacidad para adaptarse a diferentes idiomas con solo pequeñas modificaciones, lo que lo convierte en una herramienta potencialmente útil para idiomas que son ricos en morfología, pero carecen de analizadores morfológicos disponibles [25].

2) Modelo propuesto: El modelo propuesto se basa en un modelo NMT semi supervisado basado en la arquitectura Transformer, en esta se utilizará la técnica de LM-Fusion, específicamente la categoría Shallow fusion, donde el modelo de lenguaje utilizar será de la investigación de QuBERT [10], el cual servirá para el entrenamiento a priori del modelo NMT basado en Transformer, el modelo completo se puede observar en la figura 3, incluyendo los pasos de pre procesamiento.

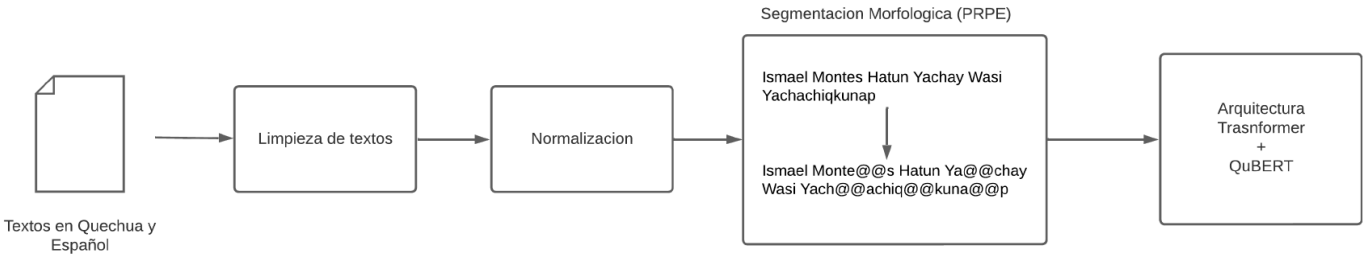

Fig. 3. Modelo Propuesto

Ofrecemos una interfaz web para probar el modelo de traducción del quechua al español, la cual se puede observar en la Figura 4. Para el desarrollo de esta interfaz se utilizará el framework React y se tomará una arquitectura cliente-servidor con peticiones HTTP para utilizar el modelo de traducción, ya que como el modelo es de un gran tamaño podría ralentizar el navegador. En la figura 4 se puede observar esta plataforma.

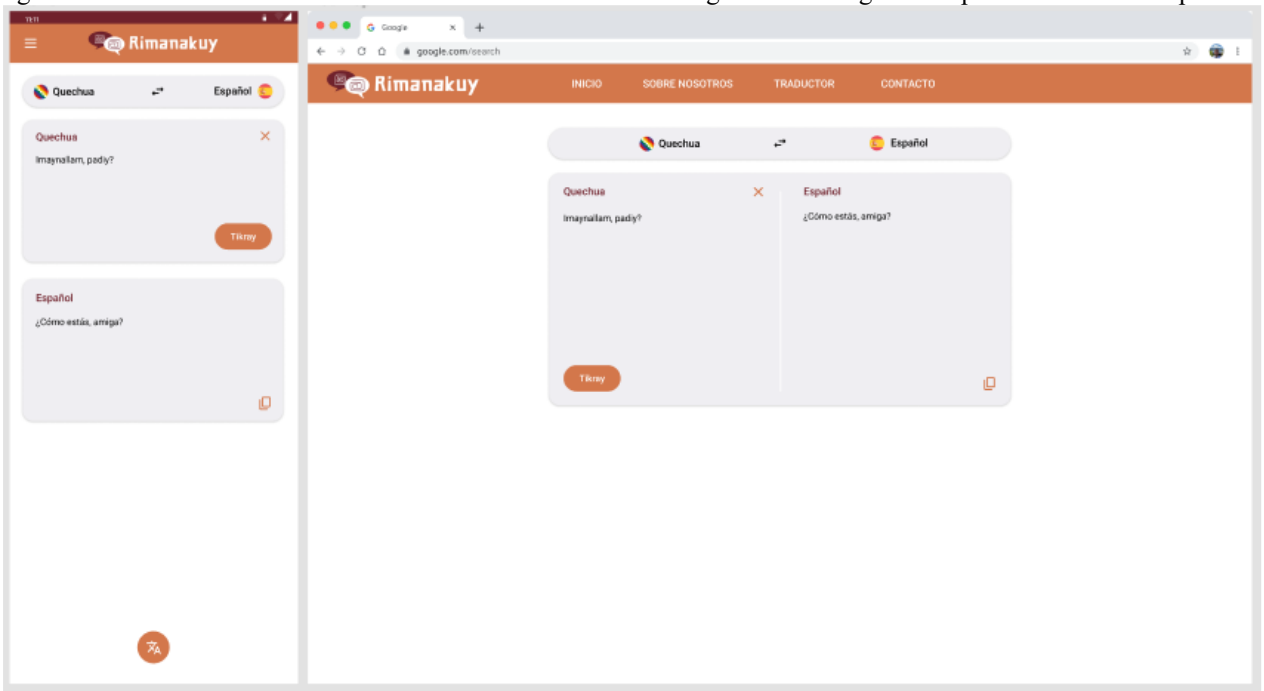

Fig. 4. Interfaz de la página web

3) Evaluación: Para la evaluación del modelo se está utilizando la métrica BLEU (Bilingual Evaluation Understudy, la cual es utilizada para medir la efectividad en una traducción automática comparado a una traducción humana profesional.

IV. Resultados

En la tabla II se muestran los resultados obtenidos por cada modelo hecho en esta investigación. En una primera instancia se probó el *corpus* recolectado con un modelo simple de Transformer, para ello se utilizó la herramienta OpenNMT para crear un modelo transformer, con el ajuste de hiperparámetros se pudo conseguir un puntaje BLEU de 4.5, lo cual es mucho menor que los resultados del estado del arte. Para comprobar la mejora que brinda la segmentación morfológica en la traducción del quechua, se utilizó el algoritmo PRPE que mejoró el modelo base, el cual llego hasta los 12.3 puntos en BLEU. Por último, con la mejora del modelo semisupervisado, se pudo lograr una mejora de 2 puntos, lo que se entiende que un modelo de lenguaje puede aportar en la traducción de lenguajes con pocos recursos.

TABLA II
Resultado de los modelos

Modelo	BLEU	Precisión	Recall	WER
Base con OpenNMT	10.1	34.23	13.40	17
Base (OpenNMT) + PRPE	12.3	43.56	16.30	14
Base (OpenNMT) + PRPE + Transformer BERT	14.6	55.31	19.56	8

En la Tabla III se puede observar algunas traducciones obtenidas por cada modelo, principalmente nuestro modelo no predice exactamente como el texto original pero el modelo llega a entender el contexto y la traducción llega a parecerse en cuanto al contexto que tradujo.

TABLA III
Traducciones hechas por los modelos

Modelo	Variable 1
Texto original	el deseo natural de todos los padres cristianos es educar a sus hijos en la disciplina y regulación mental de jehová.
Base	por ejemplo, los padres cristianos se burlan de quien es debido a que hasta la vejez y la disciplina de dios cuando tienen hijos se burlan de ellos
Base + PRPE	por ejemplo, como padre quiere que los hijos sean pequeños se les presenta como hijos de dios
Base + PRPE + BERT	los padres cristianos deben inculcar en sus hijos la ley de dios.

Además, en este estudio se realizó un cuestionario para realizar una evaluación humana del modelo, para ello, con la ayuda del aplicativo realizado, se probó con 27 personas quechua hablantes donde se obtuvo resultados positivos en el traductor, en las figuras 5, 6, 7 y 8 pueden verse estos resultados, sin embargo, se requiere de un traductor con mucha más precisión, como lo indica la pregunta 3. Además, la pregunta 2 indica que existen traductores precisos del quechua lo que puede conllevar a una mejora de dichos.

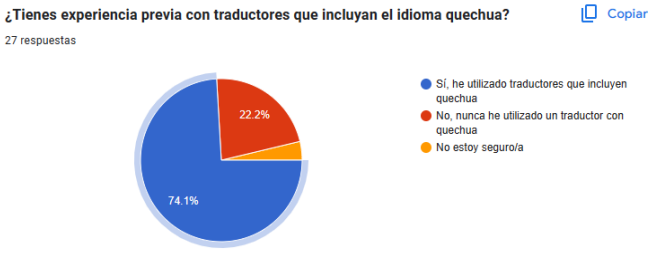

Fig. 5. Experiencia previa con traductores que incluyan el idioma quechua

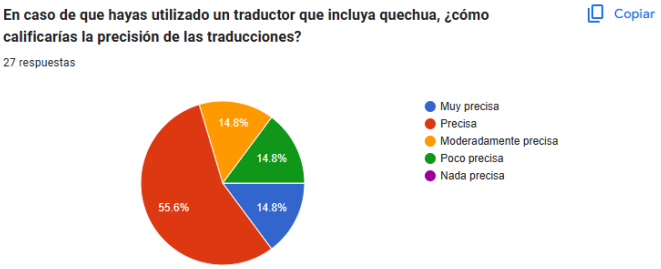

Fig. 6. Calificación de la precisión de las traducciones

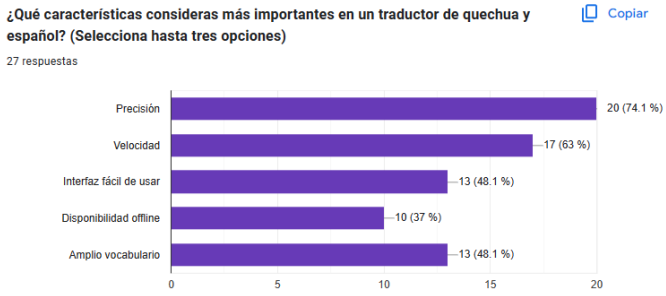

Fig. 7. Características más importantes en un traductor de quechua y español

Fig. 8. Recomendación del traductor de quechua y español

V. CONCLUSIONES

El presente trabajo presenta la creación de un modelo de traducción automática neuronal con un enfoque semi supervisado para la traducción del quechua sureño al español. Como algunas investigaciones y nuestro trabajo, afirmamos que la falta de recursos en los lenguajes sigue teniendo un gran impacto en el desarrollo de tecnologías como la traducción automática. Gracias a la ayuda de las técnicas semi supervisadas, se puede llegar a obtener mejores resultados en la traducción de idiomas con bajos recursos. El quechua es un

lenguaje complejo, lo cual complica su uso en tecnologías del lenguaje humano, por lo que requiere técnicas para poder tratarlo como lo es la segmentación morfológica. En nuestro estudio se evaluaron varios modelos para medir su efectividad en la traducción del quechua al español, utilizando métricas estándar como BLEU, precisión, recall y WER (Word Error Rate). En los resultados se observó un puntaje bajo al estado del arte, sin embargo, se espera mejorar este con un mejor ajuste de los hiperparametros y el aumento de los conjuntos de datos disponibles.

AGRADECIMIENTOS

Se agradece al Vicerrectorado de Investigación de la Universidad Católica de Santa María, Arequipa, por el apoyo en la publicación del artículo.

REFERENCIAS

[1] El Peruano, "El 13.9% de la población del perú tiene como lengua materna el quechua," El Peruano, 2021. [Online]. Disponible en: https://elperuano.pe/noticia/127783-el-139-de-la-poblacion-del-perutiene-como-lengua-materna-el-quechua

[2] El Instituto Nacional de Estadística e Informática Instituto Nacional de Estadística e Informática, "Perú: Perfil sociodemográfico. Censos nacionales 2017: Xii de población, vii de vivienda y iii de comunidades indígenas," Instituto Nacional de Estadística e Informática, 08 2018.

[3] F. K. Dueñas, "'llegando a secundaria les ha dado amnesia. . . ya no quieren hablar': Indigenous speakerhood socialization and the creation of language deniers in quechua education," Linguistics and Education, vol. 61, DOI 10.1016/j.linged.2020.100888, 12 2020.

[4] V. Zavala, "Youth, quechua and neoliberalism in contemporary per´u," International Journal of the Sociology of Language, vol. 2023, DOI 10.1515/ijsl-2022-0020, pp. 45–66, 04 2023.

[5] F. Kvietok, "Migrant bilingual youth, family, and school language policy: ethnographic insights for urban Quechua education," International Journal of the Sociology of Language, vol. 2023, no. 280, pp. 143–166, 2023.

[6] El Registro Nacional de Intérpretes y Traductores de Lenguas Indígenas Registro Nacional de Intérpretes y Traductores de Lenguas Indígenas, "Base de datos de pueblos indígenas u originarios." [Online]. Disponible en: https://traductoresdelenguas.cultura.pe/

[7] L. Melendez, "Documentación de la experiencia de los traductores e intérpretes de lenguas originarias en el perú," Sendebar, vol. 29, DOI 10.30827/sendebar.v29i0.6280, pp. 253–275, 06 2018.

[8] L. Camacho and R. Zevallos, "Siminchikkunarayku," Lingüística computacional para la revitalización y el poliglotismo. Hoja de Ruta. Fundación Siminchikkunarayku, Pontificia Universidad Católica del Perú, 2019.

[9] S. Ranathunga, E.-S. A. Lee, M. Prifti Skenduli, R. Shekhar, M. Alam, and R. Kaur, "Neural machine translation for low-resource languages: A survey," ACM Computing Surveys, vol. 55, no. 11, pp. 1–37, 2023.

[10] R. Zevallos, J. Ortega, W. Chen, R. Castro, N. Bel, C. Toshio, R. Venturas, H. Aradiel, and N. Melgarejo, "Introducing QuBERT: A large monolingual corpus and BERT model for Southern Quechua," in Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing, DOI 10.18653/v1/2022.deeplo-1.1, pp. 1–13. Hybrid: Association for Computational Linguistics, Jul. 2022. [Online]. Disponible en: https://aclanthology.org/2022.deeplo-1.1

[11] D. Huarcaya Taquiri, "Traducción automática neuronal para lengua nativa peruana," 2020.

[12] J. E. Ortega, R. Castro Mamani, and K. Cho, "Neural machine translation with a polysynthetic low resource language," Machine Translation, vol. 34, no. 4, pp. 325–346, 2020.

[13] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush, "Opennmt: Open-source toolkit for neural machine translation," arXiv preprint arXiv:1701.02810, 2017.

[14] C. Baziotis, B. Haddow, and A. Birch, "Language model prior for low resource neural machine translation," arXiv preprint arXiv:2004.14928, 2020.

[15] M. D´ıaz-Campos, The handbook of Hispanic sociolinguistics. John Wiley & Sons, 2011.

[16] A. Torero, Los dialectos quechuas. Univ. Agraria, 1964.

[17] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying crisp-dm process model," Procedia Computer Science, vol. 181, pp. 526–534, 2021.

[18] V. Babshetti and N. Ranjan, "Machine learning algorithm for work from home analysis during epidemic (2022)," Grenze International Journal of Engineering & Technology (GIJET), vol. 8, no. 2, 2022.

[19] G. Mancilla-Vela, P. Leal-Gatica, A. Sanchez Ortiz, and C. Vidal, "Factores asociados al éxito de los estudiantes en modalidad de aprendizaje en línea: un análisis en minería de datos," Formación universitaria, vol. 13, DOI 10.4067/S0718-50062020000600023, pp. 23–36, 12 2020.

[20] D. Rothman and A. Gulli, Transformers for Natural Language Processing: Build, train, and fine-tune deep neural network architectures for NLP with Python, PyTorch, TensorFlow, BERT, and GPT-3. Packt Publishing Ltd, 2022.

[21] L. Tunstall, L. Von Werra, and T. Wolf, Natural language processing with transformers." O'Reilly Media, Inc.", 2022.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. N. Toutanova, "Bert: Pretraining of deep bidirectional transformers for language understanding," 2018. [Online]. Disponible en: https://arxiv.org/abs/1810.04805

[23] S. Ravichandiran, Getting Started with Google BERT: Build and train state-of-the-art natural language processing models using BERT. Packt Publishing Ltd, 2021.

[24] "Tensorflow," https://www.tensorflow.org/overview, accessed: 2023-06-25.

[25] J. Zuters, G. Strazds, and K. Immers, "Semi-automatic quasimorphological word segmentation for neural machine translation," in International Baltic conference on databases and information systems, pp. 289–301. Springer, 2018.