PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

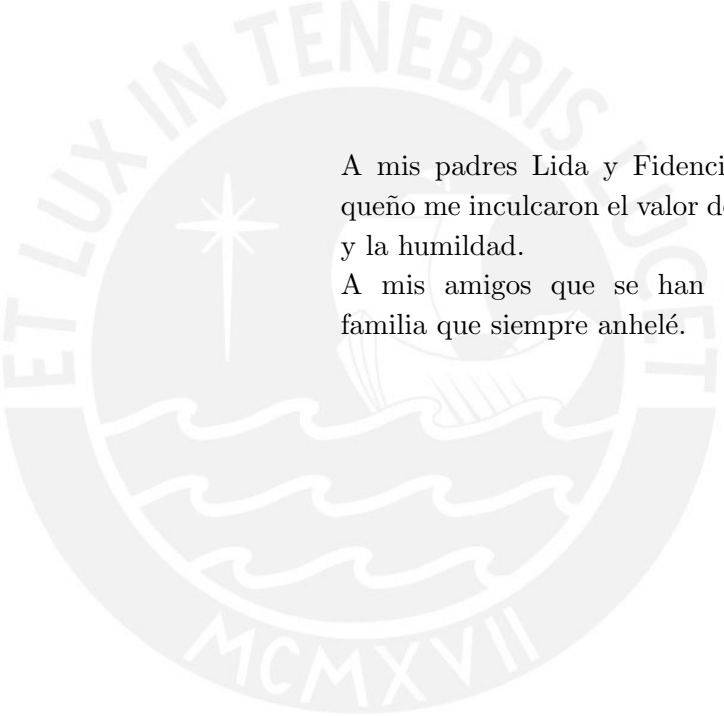
A crowd-powered conversational assistant for the improvement of a Neural Machine Translation system in native Peruvian language

Tesis para optar el grado académico de Magíster en Informática con mención en Ciencias de la Computación

Héctor Erasmo Gómez Montoya

Asesor: Mg. Felix Arturo Oncevay Marcos

Lima - 2019



A mis padres Lida y Fidencio, que desde pequeño me inculcaron el valor de la perseverancia y la humildad.
A mis amigos que se han convertido en la familia que siempre anhelé.

Agradecimientos

A Mg. Arturo Oncevay, por compartir el conocimiento obtenido mediante su trayectoria profesional, generando nuevas oportunidades de investigación y especialización a futuros investigadores.
A Fernando Alva y Marco Sobrevilla, un excelente grupo de investigadores que guiaron este proyecto de investigación para su mejoría.

Resumen

Para las comunidades más pequeñas y nativas en un país, es muy difícil encontrar información que se encuentre en su idioma original, esto debido a que su lengua no tiene el alcance ni la cantidad suficiente de hablantes, para poder seguir siendo transmitida. A este tipo de lengua se le denomina minoritaria o de pocos recursos.

Una de las principales formas en las que el gobierno incentiva el proceso de multilingüismo es proporcionando educación en el idioma nativo a su población, tal es el caso de los hablantes de Shipibo-Konibo que se encuentran dispersos a lo largo de la amazonía del Perú. Ellos cuentan con colegios donde se les imparten clases en su lengua nativa. para los niveles de primaria y secundaria. Sin embargo, una necesidad con la que cuentan los pobladores es que la cantidad de material educativo completamente traducido a shipibo-konibo es reducida. Esto debido a que el proceso de traducción es muy costoso y poco confiable.

El Grupo de investigación en Inteligencia Artificial de la PUCP (IA-PUCP, ex GRPIAA) ha desarrollado una plataforma que utiliza corpus paralelos la creación de un modelo estadístico de traducción automática para las lenguas Shipibo-Konibo y Español. Este modelo sufre de ciertas limitantes, entre las cuales tenemos: la cantidad de recursos bibliográficos y material completamente traducido, esto debido a que al ser una lengua minoritaria o de pocos recursos carecen de facilidades para la generación de nuevos corpus. Por otro lado, se desea mejorar el modelo actual en parámetros de eficiencia y obtener mejores resultados en las traducciones.

En este contexto nace la pregunta que motiva el presente trabajo: ¿de qué manera podemos incrementar el corpus paralelo de forma eficiente y confiable para la mejora del modelo actual de traducción automática?. Por consiguiente, en el presente trabajo se propone desarrollar un agente conversacional que permita la generación de nuevos corpus paralelos entre Shipibo-Konibo y Español que permitan mejorar un modelo de traducción automática neuronal en las lenguas ya mencionadas.

Publicaciones

Parte del material presentado en esta tesis está publicado en los anales de una conferencia académica internacional del área de Procesamiento de Lenguaje Natural.

1. Gómez-Montoya, H. E., Rivas, K., Oncevay, A. (2019). A continuous improvement framework of machine translation for Shipibo-Konibo. In Proceedings of the MT Summit XVII Workshop on Technologies for MT of Low Resource Languages (LoResMT 2019).

Índice General

Índice de Figuras

Índice de Tablas

1 Generalidades

1.1. Problemática

En el Perú, viven un conjunto de más de 30,000 ciudadanos que comparten un tesoro lingüístico. Ellos hablan y escriben en una lengua poco conocida por la mayoría de peruanos, el Shipibo-Konibo, que es la lengua más hablada de la familia lingüística Pano [1]. Si bien ellos utilizan su lengua nativa como primaria, esta población tiene la necesidad de utilizar el español para comunicarse con gente que no conoce su idioma, e incluso se ven forzados a reemplazar poco a poco su lengua materna debido a motivos socioculturales [24]. Esta es una situación comúnmente observable en todas las lenguas nativas habladas por minorías.

Se puede denominar como lengua minoritaria, a aquella lengua cuya cantidad de hablantes es sumamente menor a la cantidad total de ciudadanos del país donde se origina la lengua [2]. Actualmente, existe una gran variedad de proyectos gubernamentales que incentivan el uso de diferentes lenguas en el país. Para esto, se cuenta con colegios ubicados en la Amazonía de nuestro país, donde se imparten cursos en los niveles de primaria y secundaria en Shipibo-Konibo. Sin embargo, una de las principales razones por la cual este proceso de plurilingüismo es lento, es debido a la carencia de material o recursos bibliográficos que se encuentren escritos en la lengua minoritaria. Tal es el caso de leyes, noticias y planes de gobierno [4].
Una estrategia prometedora, desde el punto de vista informático, es la posibilidad de construir un traductor automático que nos proporcione de manera rápida y confiable textos traducidos de español a Shipibo-Konibo y viceversa [3]. Esta tarea consistiría en recopilar la mayor cantidad de recursos bibliográficos completamente traducidos (corpus en paralelo) y diseñar un modelo de aprendizaje de máquina que permita aprender el lenguaje para poder brindar traducciones eficientes.

El Grupo de investigación en Inteligencia Artificial de la PUCP (IA-PUCP, ex GRPIAA) realizó una investigación para desarrollar un traductor automático estadístico para la lengua nativa peruana Shipibo-Konibo[4], el cual obtuvo resultados alentadores sobre la viabilidad de modelos de traducción automática para lenguas con muy pocos recursos. No obstante, las tecnologías que brindan soporte a los mencionados modelos evolucionan cada vez más rápido. Esto debido a la necesidad de mejorar la tasa de eficiencia del modelo; es decir, la

[1] Disponible en: www.gob.pe/cultura
[2] Disponible en: www.unesco.org/new/es/lima

capacidad del sistema de no equivocarse y entregar resultados que sean lógicos y precisos [14].

El modelo actual permite la traducción de texto en español a Shipibo-Konibo; sin embargo, existen ciertas carencias que no se logran cubrir, las cuales se explican a continuación:
En particular, se sabe que los modelos estadísticos de traducción automática carecen de entendimiento del texto completo a traducir, debido a que dan una representación numérica a cada palabra, sin necesariamente tener una representación para palabras con significado parecido [20].

Por otro lado, debido a la falta de corpus y textos paralelos, el proceso de traducción se vuelve más desafiante, ya que se sabe que para poder entrenar un modelo estadístico de traducción automática se necesita contar con una gran cantidad de corpus traducido [9]. Cabe resaltar que recolectar esta gran cantidad de recursos lingüísticos para una lengua minoritaria, como es el Shipibo-Konibo, es bastante difícil y a la vez costoso, debido a que no se cuentan con material traducido o material digitalizado en la lengua.

En este contexto se describe la pregunta que motiva la presente investigación: ¿de qué manera podemos incrementar el corpus paralelo de forma eficiente y confiable para la mejora del modelo actual de traducción automática? De lo anterior, se propone desarrollar una herramienta de recolección de textos paralelos e implementar un nuevo modelo para la traducción automática que permita comprender lo que se está traduciendo de manera holística, usando los textos recolectados con la finalidad de mejorar el modelo constantemente.

Para lograr dicho propósito, se propone utilizar una serie de tecnologías que se describirán a continuación: Primero, desarrollar un modelo de NMT (*Neural Machine Translation*). Este es un enfoque relativamente nuevo para la traducción automática, y tiene como particularidad la capacidad de aprender de manera integral toda una oración y no frase por frase o palabra por palabra en contraste con modelos anteriores como son PBMT (Phrase based machine translation) o SMT (Statistical machine translation).

Segundo, debido a la poca cantidad de corpus inicial de documentos y la necesidad de incrementar estos de manera eficiente, se propone utilizar un esquema de AL (Active Learning), para que el sistema sea ayudado por el usuario, escogiendo aquellas oraciones cuya relevancia y necesidad de ser traducidas sea primordial para la mejora del modelo [6].

Por último, diseñar una herramienta basada en colaboración abierta, que permita interactuar con los hablantes de la lengua Shipibo-Konibo mediante un agente (bot) conversacional en una plataforma web. Esta herramienta permitirá realizar traducciones entre los lenguajes Shipibo-Konibo y español, y además, permitirá recopilar los nuevos corpus ya curados para poder mejorar el modelo de NMT.

Finalmente, es importante resaltar los beneficios que esta herramienta generaría en el dominio de traducciones automáticas en lenguajes minoritarios. Principalmente, se facilitará la creación de nuevos corpus paralelos por los hablantes de la lengua Shipibo-Konibo y español, se espera que estos documentos sean de ayuda para diferentes campos, ya sea para seguir incentivando el aprendizaje de la lengua o para servir como base a futuras investigaciones lingüísticas. Del mismo modo, utilizando una herramienta de carácter social (chatbot) y global, se logrará llegar a una mayor cantidad de hablantes e interesados, de manera que se logre que el lenguaje Shipibo-Konibo sea más reconocido.

1.2. Objetivos

En el presente capítulo, se determinan los objetivos del proyecto, propuesta de solución e hipótesis, así como la utilidad y justificación de la investigación.

1.2.1. Objetivo General

Implementar un marco de trabajo para la mejora continua de traductores automáticos neuronales (NMT) en beneficio de lenguas minoritarias.

1.2.2. Objetivos Específicos

1. Implementar y validar un modelo NMT que realice traducciones de Shipibo-Konibo a Español.
2. Implementar y validar un algoritmo de Active Learning para mejorar un modelo de NMT.
3. Diseñar e Implementar un prototipo de asistente de traducción automática conversacional basado en colaboración abierta distribuida para Shipibo-Konibo.

1.2.3. Resultados Esperados

1. Para el primer objetivo, un modelo de traducción automática basado en corpus que incorpore las técnicas más adecuadas para procesar textos de lenguajes con pocos recursos.
2. Para el segundo objetivo, un gráfico del desempeño obtenido aplicando la estrategia de aprendizaje activo.

3. Finalmente, para el tercer objetivo los resultados esperados son: un prototipo de un asistente conversacional que permita interactuar con el modelo propuesto e incrementar el repositorio de corpus paralelo entre Shipibo-Konibo y español.

1.3. Herramientas y Métodos

1.3.1. Herramientas

Las principales herramientas utilizadas en el desarrollo de este proyecto fueron:

- Para el desarrollo del modelo inicial se utilizó Keras y TensorFlow[3].
- Se utilizó las funciones de aprendizaje profundo que vienen con Wolfram Language 12.0, para el desarrollo de los modelos finales.
- Se utilizó las funciones de aprendizaje profundo que vienen con Wolfram Language[4], así mismo la interacción y desarrollo de los servicios para la aplicación de *crowdsourcing*, usando Wolfram Cloud[5].
- Se contó con una MacBook Pro core i7 (2.9GHz), 16Gb de RAM.

1.3.2. Metodología

En base a los objetivos planteados, se detallarán las actividades a desarrollar:

O1) Implementar un modelo NMT que realice traducciones entre Shipibo-Konibo y Español. Primero, se debe desarrollar un análisis del modelo estadístico anterior e identificar las limitaciones del mismo. Luego, representar las posibles mejoras en el modelo NMT, como por ejemplo: el entendimiento de las oraciones de manera general y holística o más conocida como secuencia inicial a secuencia final, y el manejo de palabras raras que es un problema bastante conocido en este tipo de modelos [7]. Para esto, se utilizará una red neuronal del tipo decoder-encoder [14], a fin de optimizar el modelo. Finalmente, se va a medir la calidad del texto traducido usando BLEU, el cual es una métrica usada para evaluar la calidad de un texto traducido [15].

O2) Implementar un algoritmo de Active Learning para re-alimentar un modelo de NMT. Primero, vamos a definir el método de selección de oraciones [2] y manejo de colas de prioridad para poder atender las diferentes interacciones entre usuarios simultáneos y el sistema, esto con la finalidad de escoger las consultas que puedan aportar al sistema en gran manera.

[3]https://www.tensorflow.org/guide/keras
[4]https://www.wolfram.com
[5]https://www.wolframcloud.com

Luego, se diseñará un sistema de votación automática con la finalidad de elegir aquellas traducciones que serán convertidas en nuevos corpus con la ayuda de los usuarios del sistema [10].

O3) Diseñar e Implementar un asistente de traducción automática conversacional (chatbot) basado en colaboración abierta distribuida en el dominio del lenguaje Shipibo-Konibo.

Se va a definir el conjunto de reglas a utilizar para la creación de historias de usuario que el bot será capaz de reconocer e interactuar. Las historias que se contemplan para este proyecto son las siguientes: Çomo usuario, yo puedo traducir un texto desde Español a Shipibo-Konibo y viceversa", Çomo usuario, puedo escoger y/o introducir una traducción válida a un texto traducido proporcionado por el sistema. Çomo usuario, yo puedo presentar al bot mis consultas en cualquiera de las lenguas ya sea Español o Shipibo-Konibo". Finalmente, realizar el despliegue de la herramienta en una plataforma web de mensajería instantánea, en este caso se utiliza Facebook Messenger, debido a que cuenta con una API-REST, bien documentada, la cual facilita el desarrollo de la aplicación.

1.4. Alcance y limitaciones

1.4.1. Hipótesis

Es posible aplicar Active Learning en un modelo de *Neural Machine Translation* para implementar una herramienta de colaboración masiva para mejorar las traducciones hechas por un modelo basado en corpus paralelos.

1.4.2. Justificación

La presente investigación se justifica por su implicación práctica debido a que, la aplicación de métodos de NMT y *Active Learning* mejoran las traducciones obtenidas que permitirá generar nuevos corpus paralelos para futuros proyectos de la línea de investigación.

2 Marco Conceptual

En el presente capítulo, se exponen los términos y conceptos requeridos para el entendimiento de este proyecto, el cual se divide en las siguientes 3 secciones:

2.1. Lenguas Minoritarias

2.1.1. Lengua de pocos recursos

Es un término acuñado por la UNESCO en su "Atlas interactivo UNESCO de las lenguas en peligro en el mundo"[1], empleado para identificar los lenguajes que son usados en una comunidad por una cantidad reducida de habitantes.

Para el interés de esta tesis y desde un punto de vista computacional, una lengua minoritaria es aquella lengua que no cuenta con *software*/datos digitalizados/recursos bibliográficos suficientes para ser procesados. [5]

2.1.2. Corpus paralelo

Para la tarea de traducción automática es necesario contar con un 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.
Este conjunto de datos es indispensable para el proceso de traducción automática, ya que permite la extracción de características y comportamiento de las unidades léxicas de los lenguajes a procesar [15].

2.2. Traducción automática neuronal

2.2.1. Representación Vectorial

Con la finalidad de encontrar una estructura estándar para la representación de características en palabras, frases u oraciones, se ve la necesidad de utilizar vectores que guarden esta

[1] http://www.unesco.org/languages-atlas/es/atlasmap.html

información de relación y similitud de manera compacta y que facilite su procesamiento computacional.
La creación de estos vectores sugiere el uso de redes neuronales y demás técnicas que permitan crear un modelo que represente el lenguaje[16].

2.2.2. Codificación de byte pares

La codificación de bytes pares (BPE, Byte pair encoding) es un enfoque utilizado en procesamiento de lenguaje natural para poder dividir palabras en sub-palabras y generar nuevos *tokens* para que sirvan de entrada al modelo a desarrollar[21]. En el caso particular de las lenguas nativas peruanas, se sabe que muchas tienen propiedades aglutinantes, es decir que utilizan sufijos para agregar significado. Es por ello que utilizar esta técnica es esencial ya que nos permitirá encontrar y procesar las sub-palabras encontradas.

2.2.3. Modelo secuencia a secuencia

La finalidad de este modelo es permitir transformar una secuencia de un dominio a otro.
Utilizando dos redes neuronales recurrentes (RNN) permite la transformación de una secuencia a otra. Una red de tipo codificadora condensa la secuencia de entrada en un vector y la red decodificadora procesa y despliega el vector en una nueva secuencia [22].

2.2.4. Aprendizaje transferido

En términos simples, aprendizaje transferido en procesamiento de lenguaje se puede definir como el proceso de entrenar un modelo en grandes conjuntos de datos para luego usar el modelo pre-entrenado para *transferir* conocimientos a tareas diferentes [17]. Actualmente este tipo de aprendizaje es utilizado en su mayoría por proyectos que involucren comparar conjuntos de datos que difieran en tamaño.

2.2.5. Aprendizaje activo

Es un tipo de aprendizaje que involucra al usuario para la verificación en casos en los que la anotación de corpus sea costosa.
Para la presente investigación se utiliza el concepto de selección de oraciones cuya relevancia en el modelo sea alta y poder priorizar y asegurar su traducción [2].

2.3. Colaboración abierta distribuida

Tarea que involucra un grupo de personas que a través de medios informáticos, como el internet, puedan obtener resultados eficientes para tareas determinadas.

Este proyecto al caracterizarse por trabajar con una lengua de pocos recursos, necesita una forma de crear un flujo entre el modelo y el usuario que permita interactuar con los datos con mucha más rapidez que el modelo planteado. Es por ello que se plantea usar *Crowdsourcing* para disminuir eventualmente el tiempo de procesamiento y la mejora de las traducciones [1].

3 Revisión de la literatura

En el presente capítulo se realizó una revisión de las técnicas que se vienen utilizando en el estado del arte con respecto a mejorar el rendimiento de máquinas de traducción automática. Así mismo se ha realizado una búsqueda de estudios previos donde se involucren lenguas de escasos recursos y las posibles estrategias que se han utilizado para tratarlas.

3.1. Objetivo de la revisión

El objetivo del estudio es identificar las estrategias y herramientas que se utilizan para intentar resolver la problemática planteada y obtener una línea base para esta investigación.

3.1.1. Cadenas de búsqueda

La búsqueda se realizó en las base de datos Scopus, ACL y ACM. Las cadenas de búsqueda fueron los siguientes:

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

TITLE-ABS-KEY (conversational assistant AND crowd-powered)

3.1.2. Preguntas de Revisión

En la revisión de los artículos recolectados, se buscó responder a las siguientes preguntas:

- Pregunta 1: ¿Qué métodos se utilizan para tratar corpus paralelo de lenguas de pocos recursos aplicados a traductores automáticos?
- Pregunta 2: ¿Cómo influyen las nuevas tecnologías en el desempeño de modelos de traducción automática ?

- Pregunta 3: ¿De qué forma se validan los resultados obtenidos?
- Pregunta 4: ¿De qué forma se validan los resultados obtenidos?

3.2. Resultados de la revisión

Se obtuvieron treinta y cinco resultados en total. Se seleccionaron ocho artículos que se consideraron los más relevantes para responder las preguntas de la revisión, incluyendo artículos recomendados y de cabecera para los temas a tratar.

En [20], se hace mención a la creación de modelos Long Short-Term Memory, debido a la necesidad de guardar información de oraciones para ser traducidas de manera holística, utilizando la propiedad recurrente de las redes para pasar recursos de una etapa a otra.

Este estudio, para el par de lenguas Inglés-Francés, logró una medición BLEU de 34.8 en comparación a un modelo estadístico que logra 33.3 en la escala. Se podría decir que este trabajo fue el pionero en utilizar redes neuronales para traducciones automáticas, optimizando los resultados de modelos anteriores.

En [14] evidenciaron un problema que tenían las redes neuronales aplicadas a procesos de traducción automática, ellos evidenciaban que las palabras "raras" en un corpus no eran fáciles de reconocer y por ende el método fallaba. Con palabra "rara" hacen referencia a palabras que estaban fuera del vocabulario generado por la red neuronal. Este problema fue resuelto utilizando una técnica basada en modelos anteriores, específicamente en modelos PBMT. Alinearon y anotaron el corpus inicial con las especificaciones de las palabras "raras" en la oración original. Esto ayudó a incrementar en 2.5 puntos su medida BLEU para el par de lenguas Inglés-Francés.

Google Translate es el servicio de traducción automática de la compañía Google, en [23] mencionan que la implementación de NMT para el proceso de traducción automática ha dado muy buenos resultados. Ellos proponen una GNMT, que se caracteriza por tener soporte eficiente para realizar traducciones automáticas reduciendo el número de problemas que se presentan actualmente.

Utilizan un modelo con unidades de memoria de corto plazo y redes neuronales profundas con 8 codificadores y 8 decodificadores, este diseño propone atender muchos de los problemas que actualmente tienen las NMT, uno de ellos es la simplificación de las palabras "raras" en sub-palabras para su mejor manejo, además incluir una penalización a la hora de escoger la traducción correcta de manera que se priorice una traducción holística. Lo más resaltante es

el uso de *reinforcement learning* para incrementar la puntuación BLEU que obtienen en los pares de lenguas Inglés-Francés e Inglés-Alemán.

En [8], asumen el problema de que existen muchas lenguas en las que no se cuenta con material o recursos traducidos, comúnmente llamados, recursos paralelos. Ellos proponen la idea de que estos lenguajes pueden llegar a ser traducidos también usando una NMT, pero utilizando lo que se sabe de alguna otra lengua con similar característica sintáctica y morfológica. Este enfoque, no se basa en la cantidad de documentos recolectados, si no que, estudia la relación entre lenguas de escasos recursos (LRL, low-resource language) y otras que sí tienen recursos suficientes.

Utilizando el Ruso como tercera lengua, se pudo realizar traducciones entre Bielorruso e Inglés. Este proyecto, obtiene también buenos resultados al usar una tercera lengua como el Italiano en traducciones Inglés-Castellano, debido a la similitud entre las lenguas de origen romano.

En [18], proponen una serie pasos a seguir para comparar y utilizar lenguas similares en el proceso de traducción automática utilizando redes neuronales profundas, el más resaltante es la comparación de métodos estadísticos y neuronales para la selección de estrategias, así como el uso de técnicas de aprendizaje transferido. Este conjunto de pasos ha sido probado usando Azerí o Azerbaijani (AZ) y Turco (TR) como par de lenguas relacionadas, siendo la primera aquella que no cuenta con recursos paralelos, con la finalidad de construir un traductor Azerí - Inglés.

En [11], se propone incrementar el número de lenguas relacionadas para realizar una traducción automática de manera que las palabras similares entre lenguas sean representadas en un espacio semántico par apoder ser utilizadas en el proceso de traducción de manera general. Esto permitirá entrenar un modelo con datos de entrada de diferentes lenguas, y obtener resultados para diferentes lenguas de destino que compartan una misma forma sintáctica o semántica. En experimentos iniciales, lograron crear una máquina de traducción automática usando tres lenguajes, Inglés, Italiano y Rumano.

Chorus [12], es un asistente creado con la finalidad de aprender a contestar preguntas sobre dominios específicos, utilizando la ayuda de usuarios que votan por las respuestas más adecuadas a las preguntas hechas. Chorus, logra mantenerse en el dominio de la pregunta en el 96 % de los casos, además el 93 % de veces responde correctamente a lo que el usuario le pide.

En [7], se propone realizar un asistente conversacional que reúna las características de los antiguos modelos y los nuevos en relación a chatbots. Ellos proponen generar un modelo que sea de buena calidad y bajo costo computacional, para esto la generación de un marco de

trabajo para construir un asistente conversacional es de suma importancia. Primero se enfocan en permitir el uso de diferentes ***chatbots*** para contribuir con el desarrollo de mejores respuestas. Segundo, la necesidad de aprender a utilizar respuestas ya sugeridas en preguntas similares que anteriormente se hayan hecho. Finalmente, aprender a orquestar correctamente las votaciones de los diferentes nodos de su red de ***crowdsourcing***. Ellos implementaron una aplicación móvil que pueda ser usada por los usuarios para que el modelo siga siendo realimentado.

En [10], se propuso estrategias para evaluar la aplicación de la colaboración de las masas para traducción automática. Su caso aplicado trabaja con las lenguas Ingles - Hindi, debido a la necesidad identificada en el dominio judicial de India. La corte suprema maneja sus juicios y trámites en inglés, mientras que los procedimientos de algunos tribunales se realizan en Hindi, la jurisdicción que ejerce la primera entidad sobre la segunda representa una necesidad de traducción latente. El proyecto se realizó con miras a posibles mejoras en el intercambio de información entre ambas cortes, aceleración de trámites y procesos judiciales. Para el análisis, se evalúa el compromiso de los usuarios con la tarea de traducción, las expectativas económicas que estos demuestran, la experiencia que tienen y si es suficiente para el dominio planteado, la difusión de las herramientas que hacen uso de la colaboración en masas y la experiencia de usuario que brindan las mismas como agente participativo.

En [2] se plantea una mezcla de dos técnicas bastante utilizadas por los autores para realizar traducciones automáticas en lenguas de pocos recursos. Utilizar aprendizaje activo (Active Learning) y colaboración distribuida (Crowd-Sourcing) o *Active Crowdsourcing Translation - ACT*. La primera apunta a reducir el coste del procesamiento utilizando selección de oraciones para priorizar y mejorar el modelo, la segunda proporciona un método de verificación y ayuda por parte de un grupo selecto de usuarios que son involucrados con la finalidad de actuar como expertos en masa. En resumen, logran construir un marco de trabajo para realizar la implementación de estas estrategias.

4 Experimentación y Resultados

En el presente capítulo se describen los resultados obtenidos luego de cumplir con los objetivos propuestos.

4.1. Modelo de traducción automática neuronal

4.1.1. Introducción

En esta sección se desarrollará el resultado esperado 1, correspondiente al modelo de traducción automática neuronal. En este se realiza el procesamiento de los corpus recolectados y la definición y prueba inicial del modelo.

4.1.2. Procesamiento de datos

En la etapa de evaluación se ha utilizado tres corpus diferentes, en la tabla **4-1** se puede observar la descripción del corpus utilizado. La justificación de utilizar diferentes corpus se debe a que cada modelo de traducción puede resultar más o menos efectivo en función del tipo de dominio en el que se encuentra.

El corpus etiquetado como "Religioso" consiste en frases shp-es extraídas de párrafos bíblicos. De este conjunto de datos se extrajeron oraciones que no formarán parte del entrenamiento

	$\mathcal{S}$	$\overline{r}_{\text{shp–es}}$	$\mathcal{T}$		$\lvert\mathcal{V}\rvert$		HLT	
			es	shp	es	shp	es	shp
Religioso	12,547	0.9476	195,887	185,638	13,620	19,091	6,426	11,115
Educativo	5,982	0.9148	53,710	49,135	4,351	6,568	1,649	4,044
Flashcards	7,740	1.0966	20,858	22,874	6,382	5,133	4,234	3,312
Total	26,269	0.9526	270,455	257,647	21,710	28,024	10,954	16,875

Tabla 4-1: Detalles del corpus paralelo para shp-es, por dominio y total: $\mathcal{S}$ = Número de oraciones; $\overline{r}_{\text{shp-es}}$ = Promedio del ratio$_{\text{shp–es}}$ por oración; $\mathcal{T}$ = número de token; $\lvert\mathcal{V}\rvert$ = tamaño del vocabulario; HLT = tokens con frecuencia igual a uno.

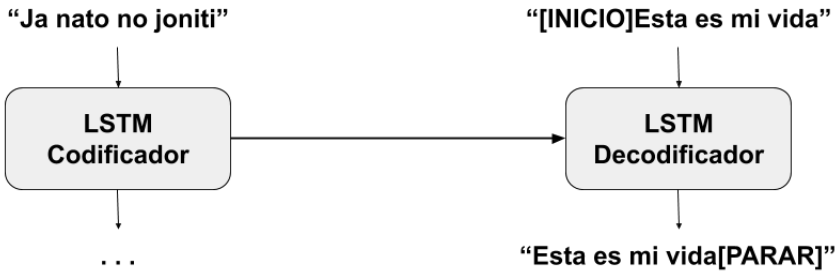


Figura 4-1: Modelo Base LSTM

y servirán para la validación del modelo, debido a que se debe contrastar los modelos entrenados con datos que no hayan sido usados por el modelo. De la misma manera se trató el conjunto etiquetado como “Educativo”, que fue extraído de textos escolares (cuentos, etc.), estos fueron validados por lingüistas y hablantes de la lengua. El último corpus empleado fue extraído y recopilado usando tarjetas de aprendizaje, que se han obtenido del proyecto CHANA[1]. Este juego de datos contiene aproximadamente 8000 oraciones traducidas en ambos idiomas.

Se probó, para cada dominio, un modelo con sus respectivos conjuntos de datos de validación, pero además se utilizó una estrategia que permitió no forzar el dominio para una traducción y probar con los conjuntos de validaciones opuestos. En esta investigación, se utilizó la métrica BLEU para la evaluación de las traducciones.

Luego de procesar las oraciones en paralelo para las dos lenguas (Shipibo y Castellano), se logró obtener una serie de patrones en más de 12 mil oraciones aproximadamente traducidas en paralelo.

4.1.3. Definición del modelo

Debido a que se trabaja con un conjunto de datos reducidos, se tuvo la necesidad de practicar un gran número de experimentos, con la finalidad de escoger las técnicas empleadas.
La linea base del proyecto, representada en la Figura **4-1**, consistió en utilizar una red neuronal recurrente *Sequence-to-Sequence*, y se definió el codificador y decodificador respectivo. Estos últimos fueron mejorados utilizando una compuerta GRU ya que tienen mejores resultados que una LSTM, debido a que contiene menos unidades de memoria, y esto hace que sea mucho más eficiente en cuestiones de procesamiento de la actualización de parámetros y generalización [19].

[1] chana.inf.pucp.edu.pe

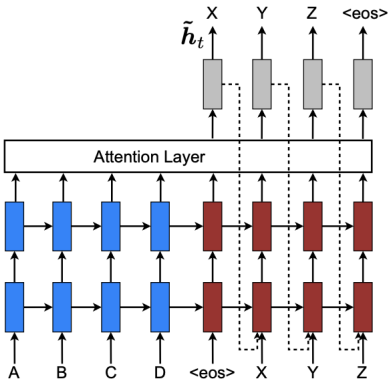


Figura 4-2: Mecanismo de atención global [13]

También se aplicó un método de aprendizaje por currículo en el codificador (ver Figura **4-2**). Esto con la finalidad de que en cada secuencia se decida ayudar o no al modelo pasando los resultados correctos; es decir, aquellos que se encuentran en los datos de entrenamiento y forzar el aprendizaje inicial. Esta técnica es utilizada generalmente en los modelos de traducción automática, y para este modelo se decidió lo siguiente:

- Solo aplicar al 10 % (en la literatura se utiliza de manera empírica este valor) de la data de entrenamiento
- En cada secuencia se elegirá si pasa o no a ser asistido, dependiendo del número de tokens que la secuencia contenga.
- El número máximo de tokens que una secuencia podrá tener para que se le pueda aplicar esta estrategia es de 10.

Luego se agregó un mecanismo de atención en el decodificador, con la finalidad de forzar al modelo en atender/enfocarse en ciertos tokens que son pasados como entrada, en vez de solo confiar en los vectores ocultos del decodificador. Los pesos pasados en esta instancia lograron crear relaciones entre las palabras o contexto de palabras.

Como resultado final, se obtuvieron tres modelos, estos se pueden observar en la tabla **4-2**. Como se aprecia al ser modelos iniciales o bases, se obtienen resultados o puntuaciones bastante bajas. Se aprecia que el grupo de datos del dominio religioso son los que peor resultados brindan, esto debido a la complejidad de las oraciones, al tamaño de las mismas y al numero de oraciones procesadas, el cual es bastante bajo, a diferencia del dominio *flashcards*, donde el tamaño y la simplicidad de las oraciones ayudan a obtener mejores resultados. En la figura **??**, podemos apreciar la perdida en el conjunto de validación y entrenamiento para el dominio *flashcards*.

	$BLEU_w$	$BLEU_{BPE}$	
		5k	15k
Religioso	1.29	**2.08**	1.33
Educativo	4.10	**4.91**	3.21
Flashcards	**11.95**	11.15	11.11
Total	3.76	**3.94**	2.46

Tabla 4-2: Puntuaciones BLEU obtenidas con el modelo base a nivel de palabra y subpalabras, usando BPE con 5,000 (5k) y 15,000 (15k) operaciones para este último.

4.1.4. Transferencia de aprendizaje

Para aplicar transferencia de aprendizaje (TL, Transfer Learning) en el contexto de traducción automática se han seguido los siguientes pasos:

- Seleccionar los recursos digitales de una tercera lengua con cuantiosos recursos digitales. Esta tercera lengua debe tener alguna relación gramatical con la lengua de pocos recursos.
- Los recursos seleccionados deberán seguir el pre-procesamiento adecuado explicado en secciones previas.
- Se entrena el modelo con los vectores obtenidos de la tercera lengua con la lengua española, a este modelo se le denominará **Padre**.
- Se entrenará el modelo con los vectores obtenidos de de la tercera lengua y la lengua shipibo-konibo, sin embargo, se utilizarán los pesos obtenidos en el paso anterior como iniciales para este modelo, el cual se llamará **Hijo**.

Este método se aplicó para las siguientes lenguas: Turco, Alemán, Inglés y Hebreo.

Los resultados se explican en la Tabla **4-4**. Estos nos muestran que la lengua Hebrea obtiene mejores resultados, esto debido a los similares accidentes gramaticales y la presencia de procesos aglutinantes en la construcción de palabras en la lengua, que hacen que se relacionen directamente con la lengua shipibo-konibo (ver Tabla **4-3**).

En el gráfico **4-4** podemos observar que algunas palabras fueron atendidas de manera directa, como por ejemplo joniti que se traduce a *vida* directamente. Pero las demás sufren alteraciones en cuanto a sentido, es por ello que aún falta mejorar este modelo. Cabe recalcar que debido a la propiedad aglutinante del lenguaje Shipibo-Konibo es inviable realizar un alineamiento palabra por palabra ya que difiere con la lengua española.

Es	Shp	He	Transliteración
sufrimiento	masati	לבס	sbl
yo sufro	teneti	סילבס	sblym

Tabla 4-3: Comparación lengua Hebrea y Shipibo

L (len.)	$\mathcal{S}_{\text{es–}L}$	$\text{BLEU}_{\text{es–shp}}$	$\text{Sim}_{(\text{shp},L)}$
Inglés	120,566	6.34	0.2822
Aleman	452,661	4.45	0.3382
Turco	7,177	9.22	**0.1764**
Hebreo	486,466	**12.34**	0.4264

Tabla 4-4: Experimentos de aprendizaje transferido usando es–L como par de lengua padre. $\mathcal{S}$ indica el tamaño del corpus, BLEU la puntuación obtenida de la traducción en la lengua par hija es–shp, y Sim es la puntuacioón de similitud entre L y shp

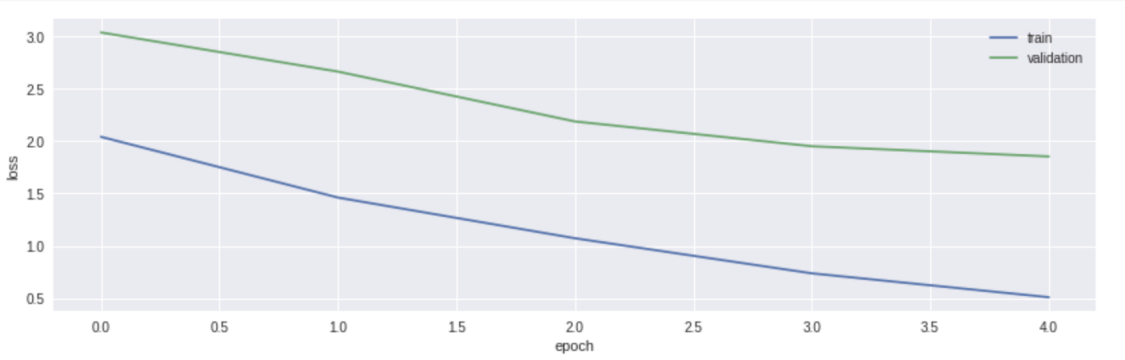


Figura 4-3: Entrenamiento vs. Validación.

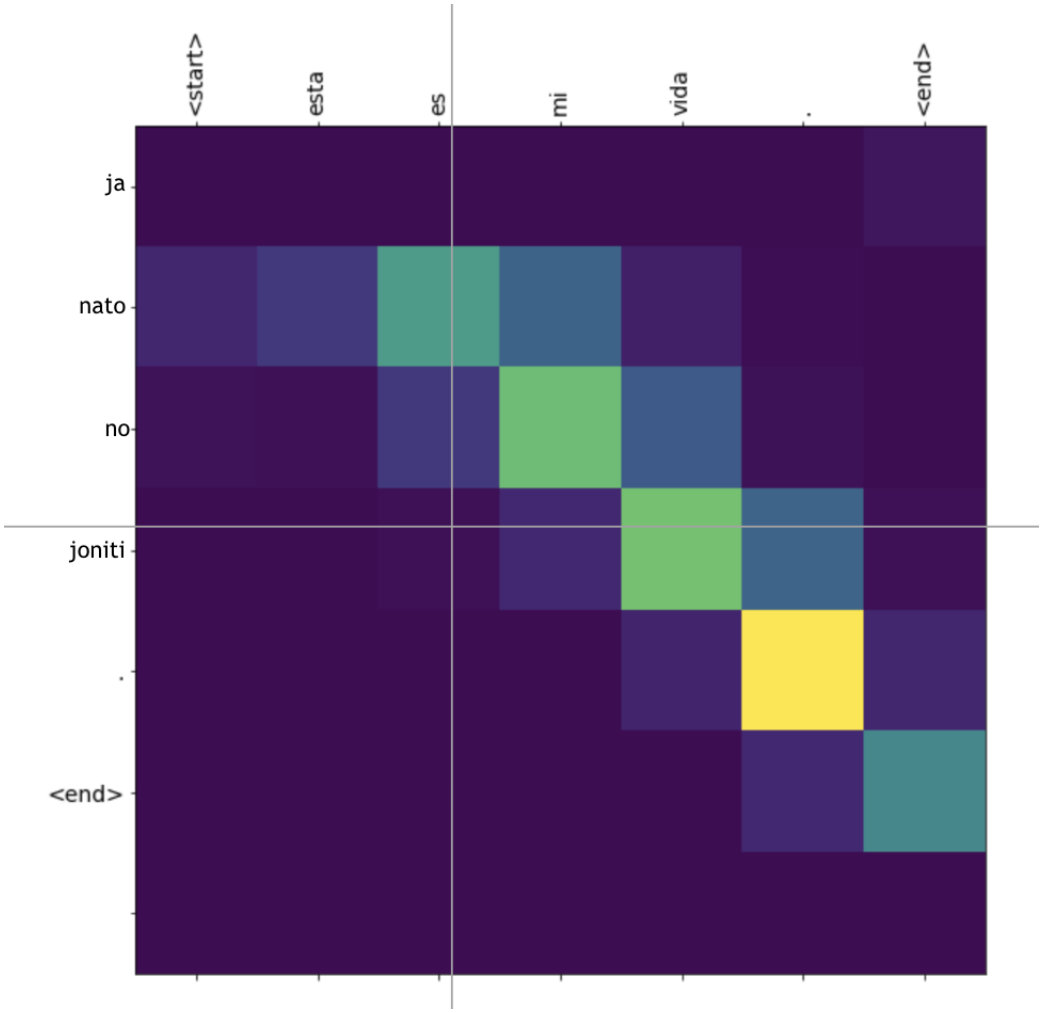


Figura 4-4: Diagrama de atención para una traducción de español a shipibo

4.2. Sistema de aprendizaje activo

4.2.1. Introducción

En esta sección se desarrollará el resultado esperado 2, correspondiente a la implementación de los algoritmos y experimentos de aprendizaje activo (AL, Active Learning) con la finalidad de mejorar el modelo base.

4.2.2. Selección de oraciones

Para la automatización del proceso de selección de oraciones a anotar se empleó la técnica de aprendizaje automático activo. Esta técnica propone la selección estratégica de los elementos a anotar cuando se tiene una gran cantidad de data y su anotación completa resulta costosa en tiempo y recursos humanos o económicos.

Esta estrategia se basa en la selección y clasificación de oraciones o palabras que de alguna manera pueda ayudar al modelo a realizar mejores traducciones, mediante un conjunto de pasos comprobados. Las siguientes características se consideraron para la selección de oraciones en el conjunto de datos en español y darles prioridad:

- Palabras fuera del vocabulario o **OOV** por sus siglas en inglés. Estas palabras representan una gran proporción y por lo tanto es prioridad poder tener una traducción completa de las mismas.
- Frecuencia de palabras, es importante tener una traducción exacta de las palabras que se usan con mayor frecuencia en el conjunto de oraciones paralelas.
- Tipos de palabra, los verbos representan la característica gramatical más importante en las oraciones que participan en el entrenamiento. Es por ello que se enfoca su traducción.
- Sinonimia, con la finalidad de obtener mejoras en un aspecto semántico.

Esta selección permite agrupar también el conjunto de datos para la realización de los experimentos, es por ello que en la siguiente sección se explicaran todos los grupos de datos que sirvieron como entrada para la selección de oraciones.

4.2.3. Inicialización del modelo y del experimento

Se siguen 2 esquemas diferentes para la realización de los experimentos de aprendizaje activo. El primero se utilizan los conjuntos de datos separándolos en los tres grupos principales según contexto. Esto quiere decir que para este esquema, lo llamaremos *Contextual*, tenemos los siguientes grupos:

	Inicial	+ Aleatorio	+ AL
Religioso	4.12	4.70	**5.78**
Educativo	5.65	5.89	**6.30**
Flashcards	10.20	12.30	**14.71**
Total	9.12	9.75	**10.43**

Tabla 4-5: Puntuación BLEU obtenida para incrementos de 40 % sobre 50 % de la configuración inicial del experimento de AL.

- Contexto Religioso (12,547 oraciones).
- Contexto Educacional (5982 oraciones).
- Flashcards (7740 oraciones).

Para cada uno de los grupos se siguió la siguiente estrategia para el experimento:
Se realizó el entrenamiento inicial al modelo escogido con el 50 % de la data total, con la finalidad de obtener un estado inicial. También se utilizó el mismo corpus para entrenar un modelo que no utilice la técnica de aprendizaje activo, con la finalidad de evaluar los resultados obtenidos por la estrategia planteada.

Luego del grupo sobrante de datos, se escogió el 40 % siguiente para utilizar la técnica de aprendizaje activo y comparar con una selección aleatoria. Para cada modelo se seleccionó un conjunto de oraciones utilizando el método de aprendizaje activo y aleatorio respectivamente. Y se procedió a seguir con el entrenamiento.

Con el 10 % de la data total restante, se tradujo utilizando ambos modelos los cuales presentaron los datos de la tabla **4-5** para el grupo separado por contextos.

El segundo grupo o *Total* comprende una agrupación aleatoria de los tres grupos iniciales de datos. Llegando a formar un solo grupo de datos de aproximadamente 12000 recursos paralelos.

El subgrupo *flashcards* obtuvo mejores resultados debido a la simplicidad de palabras utilizadas en ese conjunto de datos, a diferencia de los textos en el contexto religioso que son complejos y de gran tamaño.

Se comprobó que la selección estratégica de anotaciones muestra resultados apreciables en el entrenamiento de modelos de NMT. En general, todos los subgrupos se desempeñaron mejor utilizando la estrategia de aprendizaje activo.

4.3. Herramienta de colaboración masiva

4.3.1. Introducción

En esta sección se desarrollará el resultado esperado 3, correspondiente al prototipo de asistente de traducción automática basado en colaboración abierta. Se explicarán las estrategias de interacción con los expertos y el diseño e implementación del prototipo.

4.3.2. Estrategias de interacción

Utilizando los algoritmos explicados en la sección anterior podemos diseñar las estrategias para la interacción con el usuario utilizando un asistente conversacional.

Para aplicar las estrategias de aprendizaje en colaboración abierta, se diseñó un conjunto de modelos de persistencia que soportará la interacción.

Este modelo soporta las siguientes características:

- Almacenar posibles traducciones realizadas por los usuarios en un modelo de persistencia.
- Almacenamiento de oraciones no traducidas (en español) en espera de traducción.
- Selección de las oraciones para presentar al usuario y solicitar traducción.
- Incluir las nuevas traducciones como parte del flujo de entrenamiento.

4.3.3. Diseño e implementación del asistente conversacional

Se creó un marco de trabajo que permita la creación de un *Webhook* que soporte la interacción con la API de Facebook Messenger en su versión 3.2 [2] utilizando Wolfram Language en su versión 11.3[3]. Este webhook soporta dos tipos de interacción, para los fines de este proyecto los llamaremos *historias*.

La primera historia corresponde a la solicitud por parte del usuario a traducir una frase u oración escrita en castellano y recibir como respuesta la traducción en Shipibo-Konibo, un ejemplo de esta se observa en la figura **4-5**

La segunda historia corresponde a la interacción y solicitud del modelo para ser ayudado por el usuario a mejorar las traducciones. En la figura **4-6** se puede observar que el bot solicita

[2]https://developers.facebook.com/docs/graph-api
[3]https://reference.wolfram.com/language/

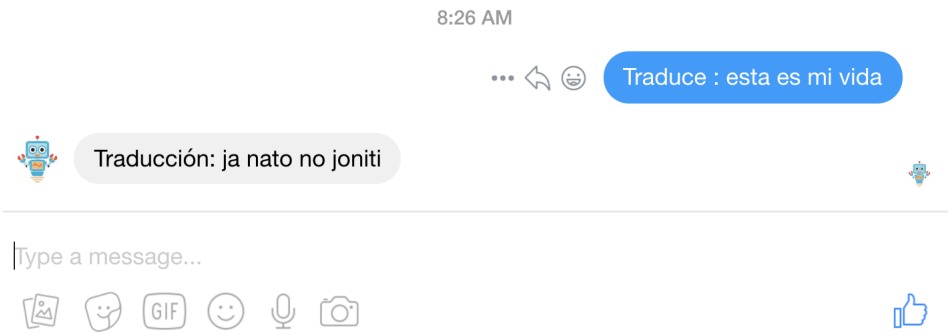


Figura 4-5: Prototipo de asistente de traducción, Historia 1

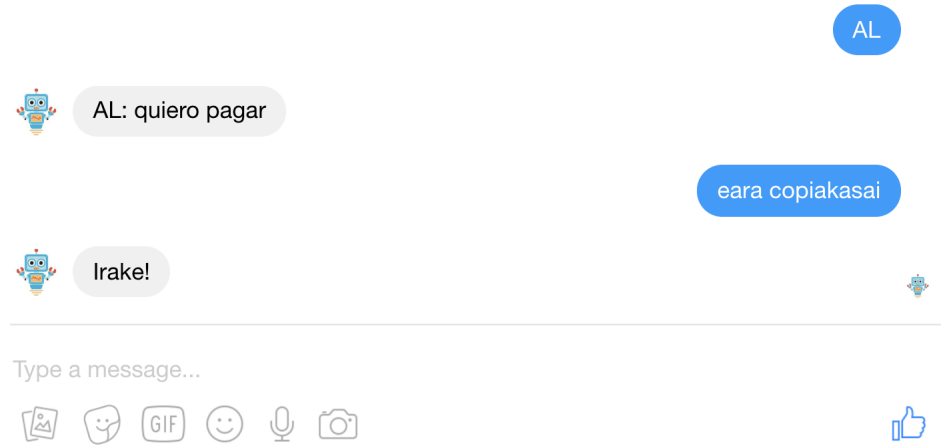


Figura 4-6: Prototipo de interacción con las frases seleccionadas, Historia 2

al usuario traducir una oración que ha sido extraída utilizando los algoritmos de aprendizaje activo explicados en la sección anterior.

5 Conclusiones y trabajos futuros

5.1. Conclusiones

- Se experimentó con varios lenguajes, entre ellos: Hebreo, Turco, Alemán e Inglés, para descubrir relaciones y elegir a cuales aplicar las técnicas de transferencia de aprendizaje y conseguir mejorar el modelo.
- Aplicamos las técnicas de transferencia de aprendizaje a dos pares de lenguas: Shipibo-Konibo-Español y Español-Hebreo y se muestra un incremento de 1.5 puntos aproximadamente en el indicador BLEU.
- Se utilizan técnicas de aprendizaje activo para la selección de oraciones para que los usuarios puedan ayudar en el proceso de traducción y se definieron estrategias para su uso, se muestra un incremento de 2 puntos aproximadamente en el indicador BLEU.
- Combinando las dos estrategias TL y AL, se logra incrementar la métrica BLEU en 3.5 puntos aproximadamente, lo cual demuestra que la propuesta puede ser mejorada.
- Se propuso un aplicativo de asistente conversacional en la plataforma *Facebook Messenger*, utilizando técnicas de ***Crowdsourcing***, con la finalidad que los hablantes de la lengua puedan interactuar y seguir ayudando en el proceso de recolección de nuevos conjuntos de datos paralelos.

5.2. Trabajos futuros

- Mejorar el modelo que se utilizó para el calculo de los vectores embebidos en Shipibo. Se debe tener en cuenta que el trabajo con este tipo de lenguas no puede ser tratado a nivel de palabras, por lo que la creación de vectores de sub-palabras es de carácter obligatorio, debido a la característica aglutinante de la lengua tratada.
- En el campo de aprendizaje activo, se puede utilizar un modelo más eficiente para la selección de oraciones por dominio y mantener esa relación en el transcurso del proceso, esto permitirá incrementar los recursos y realizar modelos específicos por cada dominio.

Bibliografía

[1] Ambati, Vamshi ; Vogel, Stephan ; Carbonell, Jaime: Collaborative workflow for crowdsourcing translation, 2012, p. 1191–1194

[2] Ambati, Vamshi ; Vogel, Stephan ; Carbonell, Jaime G.: Active Learning and Crowd-Sourcing for Machine Translation. En: *LREC* Vol. 1 Citeseer, 2010, p. 2

[3] Bahdanau, Dzmitry ; Cho, Kyunghyun ; Bengio, Yoshua: Neural Machine Translation by Jointly Learning to Align and Translate. En: *CoRR* abs/1409.0473 (2014)

[4] Galarreta, Ana P. ; H.AndrésMelgar, S. Andrés M. ; Oncevay-Marcos, Arturo: Corpus Creation and Initial SMT Experiments between Spanish and Shipibo-konibo. En: *RANLP*, 2017

[5] Gu, Jiatao ; Hassan, Hany ; Devlin, Jacob ; Li, Victor O.: Universal neural machine translation for extremely low resource languages. En: *arXiv preprint arXiv:1802.05368* (2018)

[6] Haffari, Gholamreza ; Roy, Maxim ; Sarkar, Anoop: Active Learning for Statistical Phrase-based Machine Translation. En: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Stroudsburg, PA, USA : Association for Computational Linguistics, 2009 (NAACL '09). – ISBN 978–1–932432–41–1, p. 415–423

[7] Huang, Ting-Hao (. ; Chang, Joseph C. ; Bigham, Jeffrey P.: Evorus: A Crowd-powered Conversational Assistant Built to Automate Itself Over Time. En: *CoRR* abs/1801.02668 (2018)

[8] Karakanta, Alina ; Dehdari, Jon ; Genabith, Josef: Neural Machine Translation for Low-resource Languages Without Parallel Corpora. En: *Machine Translation* 32 (2018), Juni, Nr. 1-2, p. 167–189. – ISSN 0922–6567

[9] Koehn, Philipp: *Statistical Machine Translation*. 1st. New York, NY, USA : Cambridge University Press, 2010. – ISBN 0521874157, 9780521874151

[10] Kunchukuttan, Anoop ; Roy, Shourya ; Patel, Pratik ; Ladha, Kushal ; Gupta, Somya ; Khapra, Mitesh M. ; Bhattacharyya, Pushpak: Experiences in Resource Generation for Machine Translation through Crowdsourcing. En: *LREC*, 2012

[11] Lakew, Surafel M. ; Di Gangi, Mattia ; Federico, Marcello: Multilingual Neural Machine Translation for Low-Resource Languages. (2018), 06

[12] Lasecki, Walter S. ; Wesley, Rachel ; Nichols, Jeffrey ; Kulkarni, Anand ; Allen, James F. ; Bigham, Jeffrey P.: Chorus: A Crowd-powered Conversational Assistant. En: *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology*. New York, NY, USA : ACM, 2013 (UIST '13). – ISBN 978–1–4503–2268–3, p. 151–162

[13] Luong, Minh-Thang ; Pham, Hieu ; Manning, Christopher D.: Effective Approaches to Attention-based Neural Machine Translation. En: *CoRR* abs/1508.04025 (2015)

[14] Luong, Thang ; Sutskever, Ilya ; Le, Quoc V. ; Vinyals, Oriol ; Zaremba, Wojciech: Addressing the Rare Word Problem in Neural Machine Translation. En: *CoRR* abs/1410.8206 (2014)

[15] Mager, Manuel ; Gutierrez-Vasques, Ximena ; Sierra, Gerardo ; Meza-Ruiz, Ivan: Challenges of language technologies for the indigenous languages of the Americas. En: *Proceedings of the 27th International Conference on Computational Linguistics*, Association for Computational Linguistics, 2018, p. 55–69

[16] Mikolov, Tomas ; Chen, Kai ; Corrado, Greg ; Dean, Jeffrey: Efficient Estimation of Word Representations in Vector Space. En: *CoRR* abs/1301.3781 (2013)

[17] Moon, Seungwhan ; Carbonell, Jaime G.: Completely Heterogeneous Transfer Learning with Attention-What And What Not To Transfer.

[18] Passban, Peyman ; Liu, Qun ; Way, Andy: Translating Low-Resource Languages by Vocabulary Adaptation from Close Counterparts. En: *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 16 (2017), September, Nr. 4, p. 29:1–29:14. – ISSN 2375–4699

[19] Su, J. ; Zeng, J. ; Xiong, D. ; Liu, Y. ; Wang, M. ; Xie, J.: A Hierarchy-to-Sequence Attentional Neural Machine Translation Model. En: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26 (2018), March, Nr. 3, p. 623–632. – ISSN 2329–9290

[20] Sutskever, Ilya ; Vinyals, Oriol ; Le, Quoc V.: Sequence to Sequence Learning with Neural Networks. En: *CoRR* abs/1409.3215 (2014)

[21] Tacorda, A. J. ; Ignacio, M. J. ; Oco, N. ; Roxas, R. E.: Controlling byte pair encoding for neural machine translation. En: *2017 International Conference on Asian Language Processing (IALP)*, 2017, p. 168–171

[22] Vinyals, Oriol ; Kaiser, Lukasz ; Koo, Terry ; Petrov, Slav ; Sutskever, Ilya ; Hinton, Geoffrey E.: Grammar as a Foreign Language. En: *CoRR* abs/1412.7449 (2014)

[23] Wu, Yonghui ; Schuster, Mike ; Chen, Zhifeng ; Le, Quoc V. ; Norouzi, Mohammad ; Macherey, Wolfgang ; Krikun, Maxim ; Cao, Yuan ; Gao, Qin ; Macherey, Klaus ; Klingner, Jeff ; Shah, Apurva ; Johnson, Melvin ; Liu, Xiaobing ; Kaiser, Lukasz ; Gouws, Stephan ; Kato, Yoshikiyo ; Kudo, Taku ; Kazawa, Hideto ; Stevens, Keith ; Kurian, George ; Patil, Nishant ; Wang, Wei ; Young, Cliff ; Smith, Jason ; Riesa, Jason ; Rudnick, Alex ; Vinyals, Oriol ; Corrado, Greg ; Hughes, Macduff ; Dean, Jeffrey: Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. En: *CoRR* abs/1609.08144 (2016)

[24] Zariquiey, Roberto: Reinterpretación fonológica de los préstamos léxicos de base hispana en la lengua shipibo-conibo. En: *Boletı´n de la Academia Peruana de la Lengua 41*, 2006