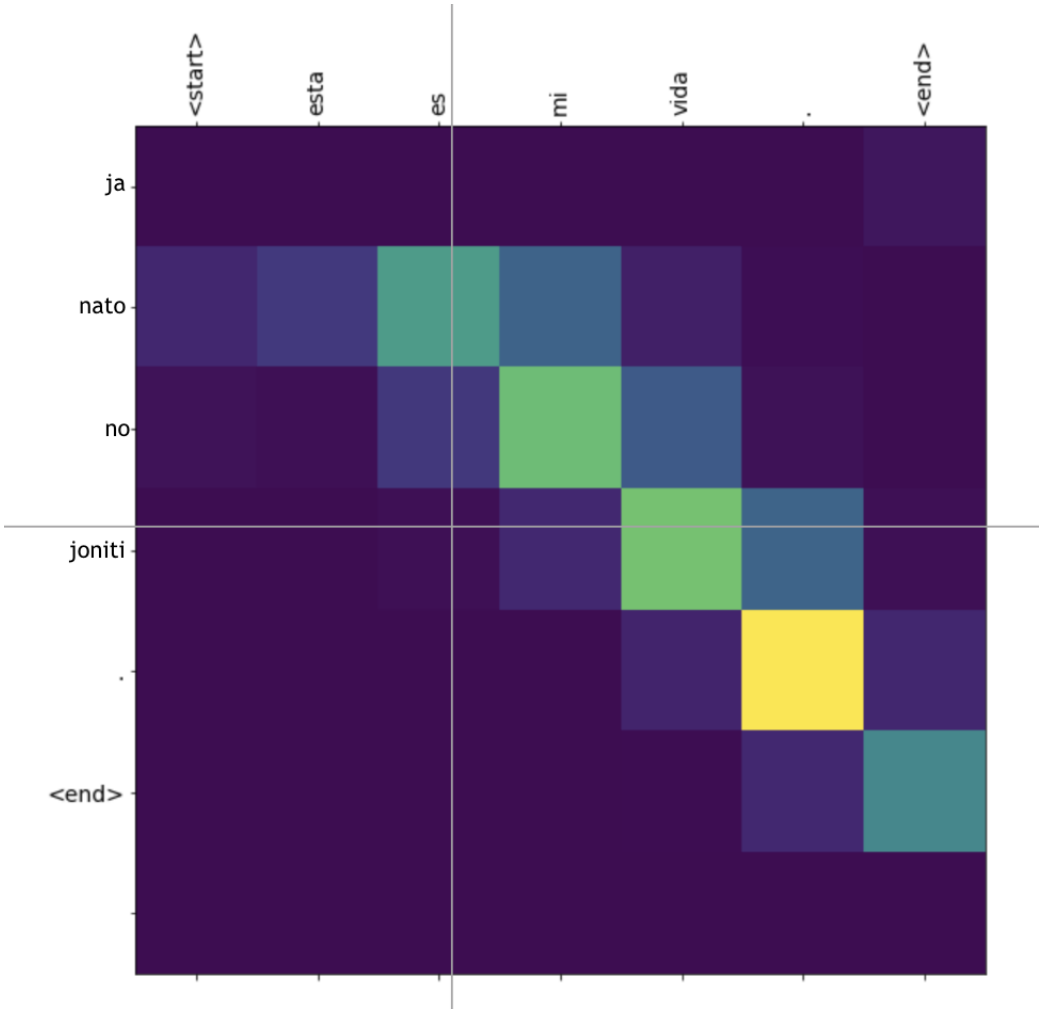

Figura 4-4: Diagrama de atención para una traducción de español a shipibo

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/

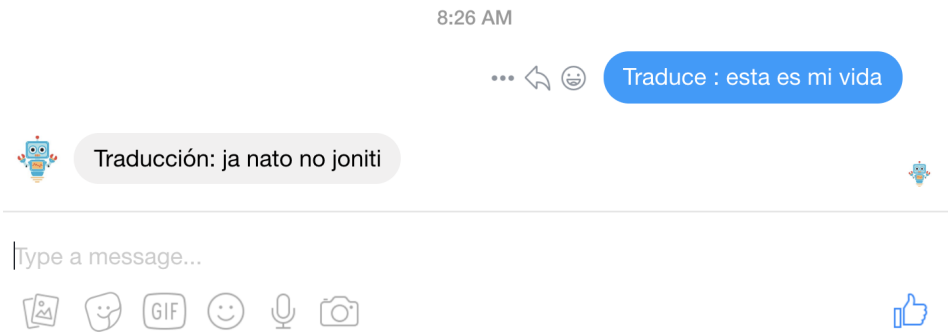

Figura 4-5: Prototipo de asistente de traducción, Historia 1

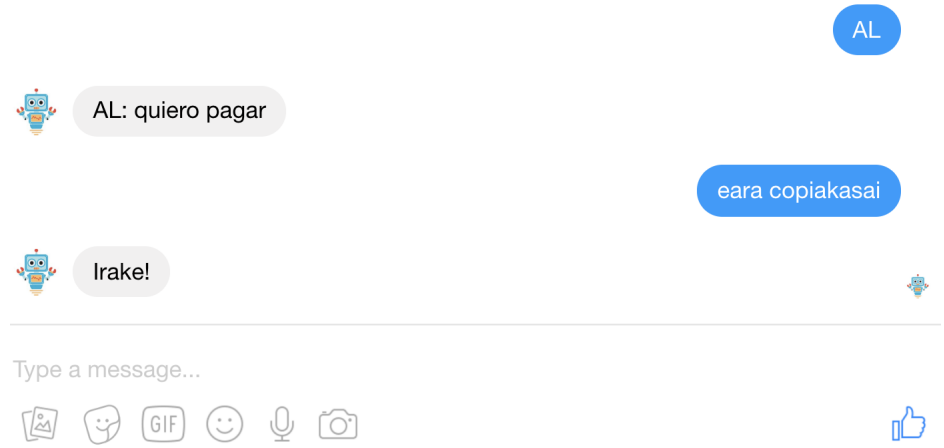

Figura 4-6: Prototipo de interacción con las frases seleccionadas, Historia 2

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