ISSN 1870-4069

Traductor automático neuronal ayuuk-español

Delfino Zacarías Márquez[1], Iván Vladimir Meza Ruiz[2]

[1] Universidad Nacional Autónoma de México,
Facultad de Estudios Superiores Acatlán,
México

[2] Universidad Nacional Autónoma de México,
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas,
México

delfino.zacarias@comunidad.unam.mx,
ivanvladimir@turing.iimas.unam.mx

Resumen. Este artículo presenta el primer sistema de traducción automática neuronal para la lengua ayuuk. En nuestros experimentos traducimos de ayuuk al español y de español a ayuuk. La lengua ayuuk es hablada en el estado de Oaxaca en México por los Ayuukjä'äy (en español comúnmente referidos como Mixes). Se usan diferentes fuentes escritas para crear en un corpus paralelo, esencial para hacer traducción automática, de más de $6,000$ frases que se considera como de bajos recursos. Para algunos de estas fuentes usamos la metodología de la alineación automática. El sistema propuesto se basa en la arquitectura neuronal Transformer y utiliza la tokenización a nivel de subpalabras como entrada. Mostramos el desempeño actual dado los recursos que hemos recolectado para la variante del municipio de San Juan Güichicovi, los resultados son prometedores, hasta 7 en BLEU. Cabe destacar que nuestro desarrollo parte del proyecto Masakhane para lenguas africanas.

Palabras clave: Lengua ayuuk, corpus, traductor automático, subpalabras, transformers, BLEU.

Ayuuk-Spanish Neural Automatic Translator

Abstract. This article presents the first neural machine translation system for the Ayuuk language. In our experiments we translate from Ayuuk to Spanish and from Spanish to Ayuuk. The Ayuuk language is spoken in the state of Oaxaca in Mexico by the Ayuukjä'äy (in Spanish commonly referred to as Mixes). Different written sources are used to create a parallel corpus, essential for automatic translation, of more than $6,000$ phrases that are considered low-resource. For some of these fonts we use the automatic alignment methodology. The proposed system is based on the Transformer neural architecture and uses tokenization at the subword level as input. We show the current performance given the resources we have collected for the variant of the municipality of San Juan Güichicovi, the results are promising, up to 7 in BLEU. It should be noted that our development is part of the Masakhane project for African languages.

Keywords: Ayuuk language, corpus, automatic translator, subwords, transformers, BLEU.

1. Introducción

En los últimos años se han incrementado los esfuerzos para preservar y promover la creación de herramientas de PLN para las lenguas indígenas de las Américas, en particular abordando los desafíos que este esfuerzo requiere [7]. La traducción automática (MT) se ha convertido en uno de los principales metas a perseguir, ya que a largo plazo puede ofrecer beneficios a las comunidades que hablan dichas lenguas.

Por ejemplo, MT podría brindar acceso al conocimiento en alguna lengua nativa y podría facilitar el acceso a servicios como asistencia legal, médica y financiera. En este trabajo trabajamos con la lengua *ayuuk* para la variante del municipio de San Juan Güichicovi, principalmente porque uno de los autores es un hablante nativo de esta variante.

Hasta donde sabemos, no ha habido una construcción de tal sistema para el ayuuk aunque existen recursos para otras variantes[3] por ejemplo en el corpus JW300 [1].Este trabajo se basó en múltiples esfuerzos previos. En el núcleo de nuestra propuesta, seguimos los pasos del proyecto Masakhane[4] que se centra en las lenguas africanas [9]. También contamos con las siguientes bibliotecas:

– Para la alineación automática de nuestros recursos utilizamos el alineador YASA[5] [5].

– Para la tokenización usamos la biblioteca subword-nmt[6] [12].

– Para el entrenamiento de nuestros modelos utilizamos JoeyNMT[7] [4].

Con estas herramientas desarrollamos nuestro código base que se puede consultar en línea junto con la parte del corpus que está disponible con licencia libre[8].

2. Ayuuk de San Juan Güichicovi

Ayuukjä'äy se puede traducir como gente de la lengua de las montañas, la mayoría de los herederos de esta cultura se concentra en 24 municipios del estado de Oaxaca. Ellos son hablantes nativos de la lengua ayuuk aproximadamente $139,760$ hablantes en México. La lengua ayuuk pertenece a la familia lingüística mixe-zoqueana.

Esta familia lingüística está compuesta por las subfamilias Mixe y Zoque[9]. En particular, la subfamilia Mixe incluye las lenguas Mixe de Oaxaca, Sayula Popoluca y Oluta Popoluca. Para el ayuuk hay seis variantes principales de la lengua, entre ellas el Mixe bajo a la que pertenece la variante de San Juan Güichicovi, con un código ISO 639-3 mir.

[3] Coatlán Mixe (ISO 639-3 mco), ayuuk de la región de Coatlán.

[4] https://www.masakhane.io/ (Última visita en marzo de 2021)

[5] https://github.com/anoidgit/yasa (Última visita en marzo de 2021)

[6] https://github.com/rsennrich/subword-nmt (Última visita en marzo de 2021).

[7] https://github.com/joeynmt/joeynmt (Última visita en marzo de 2021)

[8] https://github.com/DelfinoAyuuk/corpora_ayuuk-spanish_nmt

[9] Para obtener más información, visite acerca de la familia mixe-zoqueana https://glottolog.org/resourc/languoid/id/mixe1284

Tabla 1. Fuente de datos recopilados.

Recursos	**es**	**mir**
La Biblia	Abierto	No abierto
Cantos y poemas	No abierto	No abierto
Constitución Política de los Estados Unidos Mexicanos	Abierto	No abierto
Colección personal de Albino Pedro Juan	No abierto	No abierto
Fabulas de Esopo	Abierto	No abierto
Archivo Nacional de lenguas indígenas[10]	No abierto	Abierto
Redes sociales[11]	Abierto	Abierto
The dragon and the rabbit[11]	Abierto	Abierto[11]
Frases traducidos por el autor[11]	Abierto[12]	Abierto

En este municipio se puede estimar que hay aproximadamente $18,298$ hablantes ayuuk. Es importante notar que se estima que solo $3,205$ son monolinguës. La variante ayuuk de San Juan Güichicovi no tiene una ortografía normalizada, hay esfuerzos para acordar las convenciones ortográficas, sin embargo, hay posiciones encontradas referente al número de consonantes.

Una de estas posiciones, se conoce como "bodegeros" que propone 20 consonantes (ver 1b.a) [2] y otra se conoce como "petakeros" que propone una reducción a 14 (ver 1b.b) [10]. En términos de vocales, la variante de San Juan Güichicovi tiene seis (ver 2) que contrastan con las otras variantes del ayuuk que pueden tener hasta nueve vocales.

(1) a. b ch d ds g j k l m n ñ p r s t ts w x y ’
b. p t k x ts m n w y j l r s ’

(2) a e ë i o u .

Las siguientes frases son ejemplos de *ayuuk* de San Juan Güichicovi, estos fueron tomados de cuentos recogidos y escritos por Albino Pedro Juan, hablante nativo y promotor de la lengua.

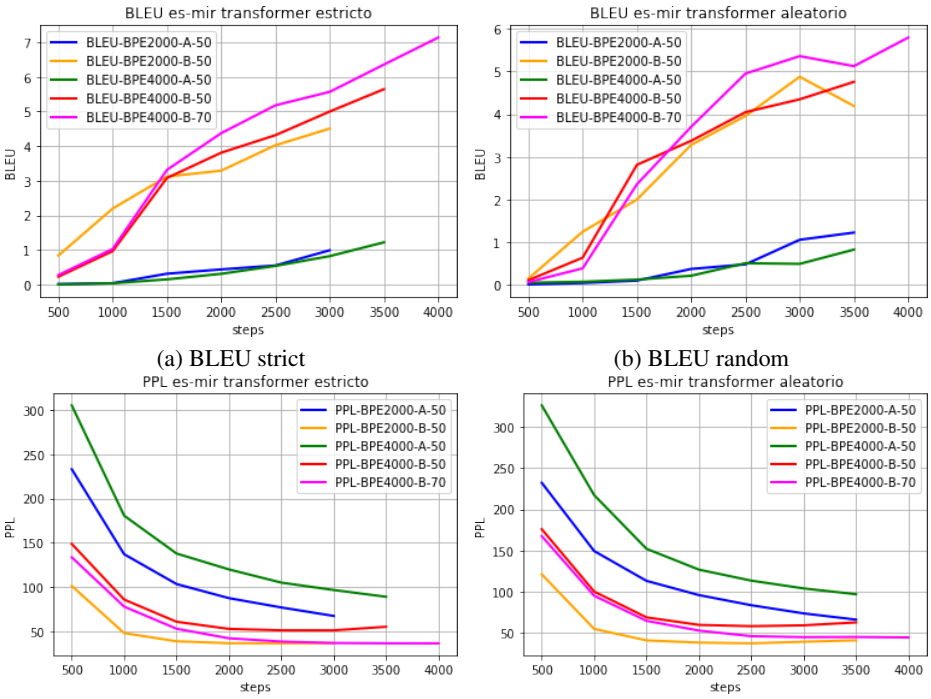
(1) Jantim xyondaak ja koy jadu’un.
El conejo se puso muy feliz.

(2) Kabëk je’e ti y’ok ëjy y’ok nójnë.
Cuando todo se quedó en silencio.

2.1. Español

En el caso del español, nuestro sistema produce traducciones al español mexicano que pertenece a la variante del español de América[13], identificamos la lengua por el código *es* del ISO-639-1.

[10] https://github.com/DelfinoAyuuk/corpora_ayuuk-spanish_nmt
[11] https://mexico.sil.org/es/resources/archives/55868
[12] https://www.manythings.org/anki/
[13] https://glottolog.org/resource/languoid/id/amer1254 (última visita en marzo de 2021)



(a) BLEU strict (b) BLEU random

(c) Perplexity strict (d) Perplexity random

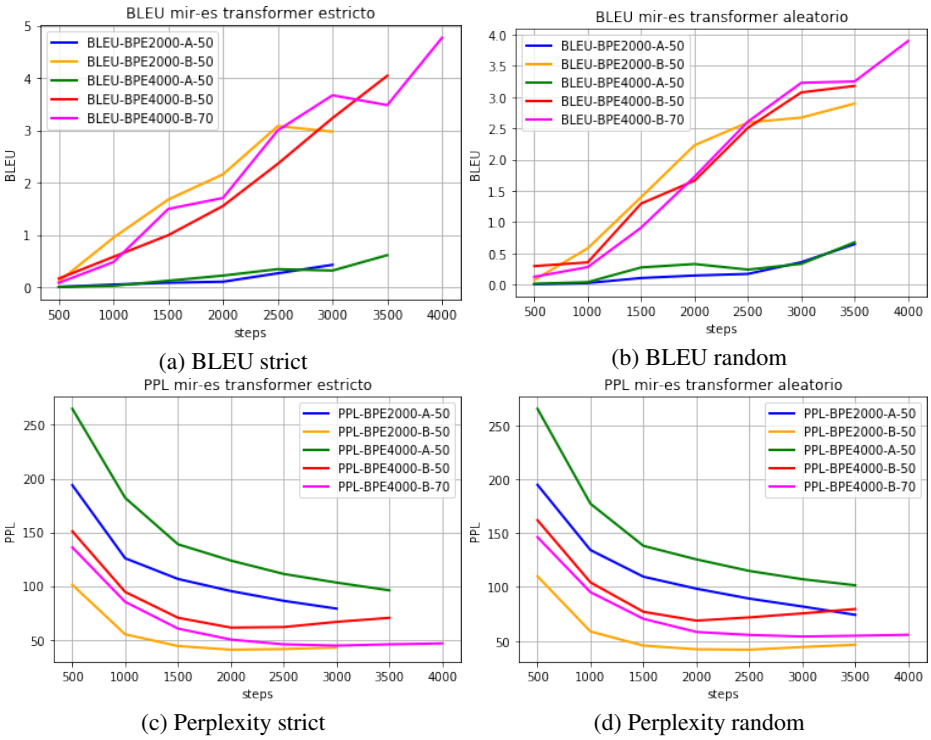
Fig. 1. Perplejidad y BLEU del entrenamiento con dirección es-mir en el conjunto de desarrollo.

3. Sobre el corpus paralelo

Para la creación del corpus paralelo, recolectamos textos de diferentes fuentes para las cuales había una traducción disponible entre ayuuk y español, ver Tabla 1. Dado que tenemos una diversidad de fuentes lingüísticas fue necesario normalizar la ortografía y algunas palabras. Para ello seguimos la propuesta derivada de la investigación de [11] quien ha seguido la unificación de la lengua ayuuk evitando tomar partido en la polémica sobre el número de consonantes. Principalmente hicimos dos reemplazos: ñ/ny y ch/tsy. Algunas de las obras ya estaban alineadas, otras no.

Para aquellos que no se estaban alineadas, creamos alineaciones automáticas usando la herramienta YASA [5]. Descartamos todas las alineaciones vacías y dobles. Finalmente, dividimos aleatoriamente las oraciones en conjuntos de entrenamiento, desarrollo y prueba. Para nuestro experimento, creamos dos versiones divididas, una estricta y otra aleatoria.

En la versión estricta usamos todas las frases del Archivo Nacional de lenguas indígenas [6] como test. Dado que estas oraciones están motivadas lingüísticamente y tienen como objetivo mostrar aspectos lingüísticos de la lengua, tienden a ser más difíciles de traducir. Esta división resultó en $5,847$ / 700 / 912 (train/dev/test). En la división aleatoria muestreamos oraciones al azar de nuestras fuentes, la división final resultó en $5,941$ / 700 / 912 (train/dev/test).



(a) BLEU strict (b) BLEU random

(c) Perplexity strict (d) Perplexity random

Fig. 2. Perplejidad y BLUE del entrenamiento con dirección mir-es en el conjunto de desarrollo.

Observe que la cantidad de frases entre las dos versiones cambia, esto se debe a que después de separar las frases de prueba (i.e., test) eliminamos frases repetidas o similares para los conjuntos train/dev.

Nuestra intuición era tener un entrenamiento/validación más uniforme, frases únicas, para la división aleatoria mientras que los ejemplos de test siguieran la distribución de las fuentes originales. Se siguió la misma metodología para la creación de la versión estricta.

4. Arquitectura neuronal

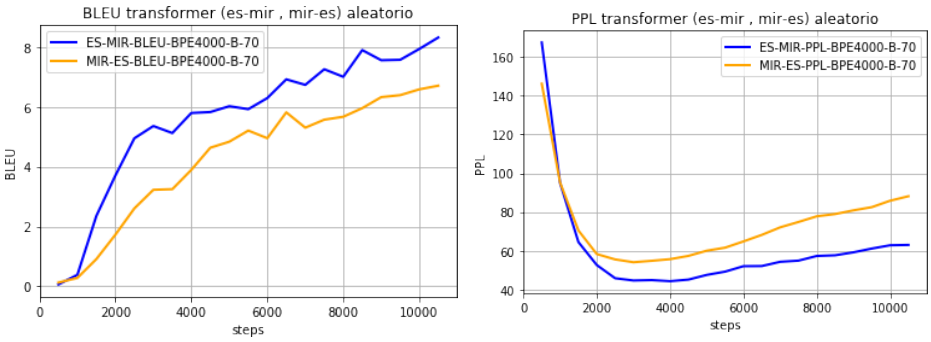
Nuestro modelo de traducción se basa en la arquitectura Transformer [13]. Usamos una configuración codificador-decodificador. Para nuestros experimentos tenemos dos configuraciones para el codificador y decodificador:

a) Número de capas: 3, número de cabezales: 4, dimensión de embedding de entrada: 64, dimensión embedding: 64, tamaño de lote: 128.

b) Número de capas: 6, número de cabezales: 4, dimensión de embedding de entrada: 256, dimensión de embedding: 256, tamaño de lote: 128.

Tabla 2. Puntuaciones BLEU de los entrenamientos con dirección es-mir y mir-es.

Configuracion A - 100 épocas	**Estricto es-mir**		**Aleatorio es-mir**		**Estricto mir-es**		**Aleatorio mir-es**	
BLEU	**dev**	**test**	**dev**	**test**	**dev**	**test**	**dev**	**test**
Longitud máxima 50 BPE 2000	1.72	0.05	1.66	1.71	0.64	0.10	0.91	0.66
Longitud máxima 50 BPE 4000	2.03	0.10	1.21	1.24	1.02	0.16	0.93	0.83
Configuración B - 100 épocas	**Estricto es-mir**		**Aleatorio es-mir**		**Estricto mir-es**		**Aleatorio mir-es**	
BLEU	**dev**	**test**	**dev**	**test**	**dev**	**test**	**dev**	**test**
Longitud máxima 50 BPE 2000	3.91	0.10	3.59	3.70	2.21	0.41	2.49	2.72
Longitud máxima 50 BPE 4000	5.02	0.13	4.17	4.20	2.33	0.28	2.13	2.23
Longitud máxima 70 BPE 4000	7.58	0.10	5.83	5.56	4.03	0.27	3.64	3.52
Configuración B - 250 épocas	**Aleatorio es-mir**				**Aleatorio mir-es**			
BLEU	**dev**		**test**		**dev**		**test**	
Longitud máxima 70 BPE 4000	5.83		5.56		3.64		3.52	



(a) BLEU strict

(b) BLEU random

Fig. 3. Perplejidad y BLEU del entrenamiento es-mir y mir-es con 250 epocas.

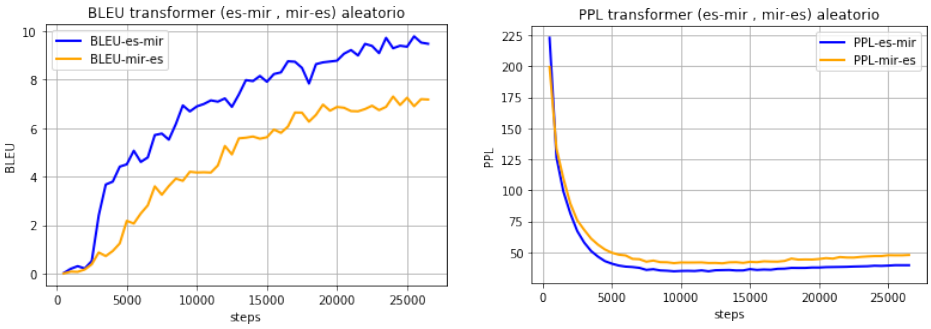
Estos modelos se entrenaron en un servidor con dos GPU Tesla V100. Para obtener el resultado de un modelo usualmente nos tomó alrededor de $2h$ por una cantidad de 100 épocas. También pudimos reproducir los experimentos en la plataforma Colaboratory.

5. Experimentos y resultados

Como se describió en la sección anterior, tenemos dos versiones diferentes de nuestras divisiones, estricto y aleatorio. Por cada división realizamos cinco experimentos, dos para la configuración con menos capas de la red Transformer (A) y tres para la configuración con más capas (B). También modificamos:

a) La longitud máxima de la frase (50 o 70).

b) El vocabulario del algoritmo de subpalabras BPE (probamos 2000 o 4000).



(a) BLEU strict (b) BLEU random

Fig. 4. Perplejidad y BLEU del entrenamiento es-mir y mir-es con 150 épocas.

Tabla 3. Puntajes BLEU de entrenamientos con la configuración con 150 épocas en dirección es-mir y mir-es.

Configuración C - 150 épocas	**Aleatorio es-mir**		**Aleatorio mir-es**	
BLEU	**dev**	**test**	**dev**	**test**
Longitud máxima 100 BPE 4000	7.34	7.29	6.18	5.82

La Figura 1 muestra la perplejidad y la puntuación BLEU en el conjunto de desarrollo durante el entrenamiento para la dirección del español (es) a ayuuk (mir). La primera parte de la Tabla 2, las columnas dos al cinco, presentan los resultados de los conjuntos de desarrollo y prueba. La Figura 2 muestra la curva de aprendizaje en la dirección de traducción ayuuk (mir) al español (*es*). La segunda parte de la Tabla 2, las columnas del seis al nueve, presentan los resultados sobre el desarrollo y la prueba para esta dirección de traducción. Como podemos apreciar, estos conjuntos de experimentos muestran que la traducción es posible.

Tenemos algunas ganancias en el modelo con más capas (B), esto no es trivial ya que tenemos una pequeña cantidad de datos de entrenamiento. Por otro lado, la división estricta como se esperaba muestra que es muy difícil de traducir, las puntuaciones BLEU son mínimas. Sin embargo, con las divisiones aleatorias, las puntuaciones BLEU son más prometedoras. También observamos que en la configuración actual es más "fácil" traducir del español al ayuuk que en la otra dirección. Dado los resultados prometedores de la configuración *B* en la división aleatoria, se realizó un experimento más grande con 250 épocas, siguiendo la intuición de que aún no se ha alcanzado el rendimiento correcto con 100 épocas.

En la Figura 3 se muestra la curva de aprendizaje en el conjunto de desarrollo del entrenamiento en ambas direcciones, la parte inferior de la Tabla 2 muestra los resultados finales. De acuerdo a los resultados, el entrenamiento con 250 épocas tiene el mismo puntaje BLEU que con 100 épocas, lo que destaca es una elevación de perplejidad en el step 4,000. Finalmente, realizamos un experimento con 150 épocas en ambas direcciones con la división aleatoria modificando la configuración *B*, donde reduce a 32 el número de lotes, se mantiene la BPE en $4,000$ y se aumenta a 100 la longitud máxima de frases.

Tabla 4. Traducciones candidatos generados por el traductor automático neuronal.

Traducción español - ayuuk variante de San Juan Güichicovi	
Origen	todos decían que soy malo
Objetivo	age nëm ajxy myana'any ko ëëtsy n'ëxëëgya'ayë
Candidato	a nëje'e ajxy ënajty myënaambë te'emjyëdu'un kyë'ëxë'ëky
Origen	ustedes me vieron ayer en el mercado
Objetivo	mijts axëëy xyijx jim ma too'ktaaktën
Candidato	xyijxëtsy mijts axëëy ma too'ktaaktën
Origen	le dijeron señor, ven y ve
Objetivo	nëm ajxy nyëmaay mëj windsën, jam ukte'emy'ix
Candidato	nëm ajxy y'adsooy mëj windsën, weenëtsy n'ijxë'ëky
Traduccion ayuuk variante de San Juan Güichicovi - español	
Origen	nëm ajxy nyëmaay mëj windsën, jam ukte'emy'ix
Objetivo	le dijeron señor, ven y ve
Candidato	y ellos le dijeron señor,
Origen	jim jaa koy y'aame'naay
Objetivo	el conejo estaba escondido
Candidato	estaba el conejo
Origen	nëm ja ya'ay ajxy y'adsooy ku ëdaa ya'eay ëxyëp kya'aku'uwandëyjëya'ayë, kap ëjts miitsy ëxyëp të nyajkë'ëdëgë'ëy
Objetivo	respondieron y le dijeron si éste no fuera malhechor, no te lo habríamos entregado
Candidato	ellos le respondieron si fuere necesario que no hagas

Dando como resultado final una disminución de perplejidad y un aumento en el puntaje BLEU, tanto para la dirección del español (es) al ayuuk (mir) como del sentido contrario, en la Figura 4 y en la Tabla 3 se muestran los resultados. Los puntajes BLEU y perplejidad dan una idea de cómo pueden ser las traducciones candidatas que proporciona el mejor modelo de traducción generado hasta el momento.

En la Tabla 4 se muestran algunos resultados de traducción que generan los modelos que han sido entrenados en este trabajo. Las frases de entrada se escogieron de manera aleatoria dentro del corpus test. Asimismo, en el tabla de resultados la fila origen corresponde a la lengua que se quiere traducir, la fila objetivo contiene la traducción correcta y la fila candidato es la traducción generada por el traductor automático neuronal.

6. Conclusiones y trabajos futuros

Las experiencias anteriores en MT basadas en la arquitectura de aprendizaje profundo, particularmente en la configuración de seq2seq, para las lenguas nativas de las Américas no habían sido prometedoras [8]. En particular, porque hay pocos o ningún dato de entrenamiento.

Sin embargo, nuestro trabajo muestra que un modelo estándar basado en la arquitectura Transformer y con una configuración de recursos extremadamente baja puede producir algunos resultados prometedores. Todavía son bajos para los estándares normales del campo de MT[14], sin embargo, son prometedores para un futuro donde hay más datos. Para mejorar el rendimiento del sistema, el trabajo futuro se centrará en:

1. Recopilar más datos, especialmente teniendo en cuenta las diferentes variantes del la lengua ayuuk. Hasta ahora en este trabajo abordamos una variante específica, pero existen múltiples variantes que también carecen de una ortografía estandarizada.
2. Aunque la configuración estricta penaliza fuertemente al sistema, creemos que las frases motivadas lingüísticamente podrían establecer una buena referencia para evaluar el progreso y el rendimiento de nuestro sistema de traducción automática. En esta dirección, seguiremos evaluando bajo esta configuración.
3. En este momento nos basamos en subpalabras, sin embargo, nuestro enfoque podría beneficiarse de un análisis morfológico más profundo [3].
4. Nuestra normalización seguirá respetando las posiciones de los "petakeros" y "bodegeros", y para otras variantes también incorporamos posiciones en cuanto al número de vocales.

Agradecimientos. Agradecemos a CONACYT por los recursos proporcionados a través de la Plataforma de Aprendizaje Profundo del Laboratorio de Supercomputación del INAOE para Tecnologías del Lenguaje. También agradecemos el proyecto "Traducción automática para lenguas indígenas de México" PAPIIT-IA104420, UNAM.

Referencias

1. Agić, Ž., Vulić, I.: JW300: A wide-coverage parallel corpus for low-resource languages. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, pp. 3204–3210 (2019) doi: 10.18653/v1/P19-1310
2. Graham, S., Hillman, V., Williams, J., Willett, T. L., Becerra-Bautista, M., Pérez-Luría, M., Eberle-Cruz, V., Araiza-Riquer, K., Dieterman, J., McCarty-Jr, J. M., Castañón-López, V., Castañón-Eugenio, M. D.: Breve diccionario del mixe del Istmo Mogoñé Viejo, Oaxaca. Instituto Linguistico de Verano (2018)
3. Kann, K., Mager Hois, J. M., Meza-Ruiz, I. V., Schütze, H.: Fortification of neural morphological segmentation models for polysynthetic minimal-resource languages. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, pp. 47–57 (2018) doi: 10.18653/v1/N18-1005
4. Kreutzer, J., Bastings, J., Riezler, S.: Joey NMT: A minimalist NMT toolkit for novices. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language

[14] Puntuaciones BLEU de entrenamientos de lenguas africanas https://github.com/masakhane-io/masakhane-mt/tree/master/benchmarks

Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations, Association for Computational Linguistics, pp. 109–114 (2019) doi: 10.18653/v1/D19-3019
5. Lamraoui, F., Langlais, P.: Yet another fast, robust and open source sentence aligner. Time to reconsider sentence alignment, XIV Machine Translation Summit, pp. 77–84 (2013)
6. Lyon, D. D.: Mixe de Tlahuitoltepec, Oaxaca, Archivo de Lenguas Indígenas de México. Colegio de México (1980)
7. Mager, M., Gutierrez-Vasques, X., Sierra, G., Meza-Ruiz, I.: Challenges of language technologies for the indigenous languages of the Americas. In: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, pp. 55–69 (2018)
8. Mager, M., Meza, I.: Hacia la traducción automática de las lenguas indıgenas de méxico. Proceedings of the DH, (2018)
9. Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohungbe, T., Akinola, S. O., Muhammad, S., Kabenamualu, S. K., Osei, S., Sackey, F., Niyongabo, R. A., Macharm, R., Ogayo, P., Ahia, O., Berhe, M. M., Adeyemi, M., Mokgesi-Selinga, M., Okegbemi, L., Martinus, L., Tajudeen, K., et al.: Participatory research for low-resourced machine translation: A case study in African languages. In: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, pp. 2144–2160 (2020) doi: 10.18653/v1/2020.findings-emnlp.195
10. Reyes-Gómez, J. C.: Aportes al proceso de enseñanza aprendizaje de la lectura y la escritura de la lengua ayuuk. Centro de Estudios Ayuuk–Universidad Indígena Intercultural Ayuuk (2005)
11. Sagi-Vela González, A.: El mixe escrit i el miratge del bon alfabet. Revista de Llengua i Dret, no. 71, pp. 146–157 (2019) doi: 10.2436/rld.i71.2019.3256
12. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, vol. 1, pp. 1715–1725 (2016) doi: 10.18653/v1/P16-1162
13. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, Curran Associates, Inc., vol. 30 (2017) doi: 10.48550/ARXIV.1706.03762