Ayuuk-Spanish Neural Machine Translator
Delfino Zacarías
Facultad de Estudios Superiores Acatlán,
Universidad Nacional Autónoma de México
delfino.zacarias@comunidad.unam.mx

Ivan Meza
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas,
Universidad Nacional Autónoma de México
ivanvladimir@turing.iimas.unam.mx

Abstract
This paper presents the first neural machine translator system for the Ayuuk language. In our experiments we translate from Ayuuk to Spanish, and from Spanish to Ayuuk. Ayuuk is a language spoken in the Oaxaca state of Mexico by the Ayuukjä’äy people (in Spanish commonly known as Mixes). We use different sources to create a low-resource parallel corpus, more than 6,000 phrases. For some of these resources we rely on automatic alignment. The proposed system is based on the Transformer neural architecture and it uses sub-word level tokenization as the input. We show the current performance given the resources we have collected for the San Juan Güichicovi variant, they are promising, up to 5 BLEU. We based our development on the Masakhane project which focuses on African Languages (Nekoto et al., 2020). We also rely on the following libraries:

- For the automatic alignment of our resources we use the YASA alignment (Lamraoui and Langlais)\(^3\)
- For the tokenization we use subword-nmt library\(^4\) (Sennrich et al., 2016)
- For the training of our models we use JoeyNMT\(^5\) (Kreutzer et al., 2019).

With these tools we developed our code base that can be consulted online together with the part of the corpus which is freely available. 6.

2 Ayuuk from San Juan Güichicovi

Ayuukjä’äy can be translated as people of the mountains, most them can be located in 24 municipalities of the Oaxaca state. They are the native speakers of the Ayuuk language with approximately 139,760 speakers in Mexico. The Ayuuk language, which has an ISO 639-3 code mir, belongs to the mixe-zoqueana linguistic family. This linguistic family is composed by the Mixe and Zoque subfamilies. 7 In particular, the Mixe subfamily also includes Mixe of Oaxaca, Sayula Popoluca and Oluta Popoluca languages. For Ayuuk there are six main variants of the language, among these the Mixe bajo to which the San Juan Güichicovi variant belongs to. At

\(^1\)Coatlán Mixe (ISO 639-3 mco), Ayuuk of the Coatlán region.

\(^2\)https://www.masakhane.io/ (last visited march 2021)
\(^3\)https://github.com/anoidgit/yasa (last visited march 2021)
\(^4\)https://github.com/rsennrich/subword-nmt (last visited march 2021).
\(^5\)https://github.com/joeynmt/joeynmt (last visited march 2021)
\(^6\)https://github.com/DelfinoAyuuk/corpora_ayuuk-spanish_nmt
\(^7\)For further information visit about the mixe-zoqueana family https://glottolog.org/resource/languoid/id/mixe1284
in this municipality it can be estimated there is approximately 18,298 speakers of the variant. It is important to notice that it is estimated that only 3,205 are monolinguis.

The San Juan Güichicoví Ayuuk variant does not have a normalized orthography, there are efforts to agree on orthographic conventions however there are strong positions related to number of consonants. One of these positions, it is known as the “bodegeros” position which proposes 20 consonants (see 1b.a) (Willett et al., 2018) vs “petakeros” which proposes a reduction to 13 (see 1b.b) (Reyes Gómez, 2005). In terms of vowels, this variant has six (see 2) which contrast with the other variants of Ayuuk which can have up to nine vowels.

(1) a. b c d e f g h i j k l m n ñ p r s t u w x y z
    b. p t k x ts m n wy j l r s

(2) a e ë i o u

The following are examples of San Juan Güichicoví Ayuuk these were taken from short stories recollected and written by Albino Pedro Juan a native speaker and preserver of the language.

(3) Jantim xyondaak ja koy jadu’un.
    The bunny become happy.
    El conejo se puso feliz.

(4) Kabëk je’e ti y’ok ëjy y’ok nójnë.
    When everything become silence.
    Cuando todo se silencia.

2.1 Spanish

In the case of Spanish, our system produces translations in Mexican Spanish which belongs to the American Spanish variant 8, we identify the language by the es ISO-639-1 code.

3 The parallel corpus

For the creation of the parallel corpus we collected samples from different sources for which there was a available translation between Ayuuk and Spanish, see Table 1.

Since we have a diverse source of linguistic sources it was necessary to normalize the orthography. For this we follow the proposal from Sagi-Vela González (2019) who has followed the unification of the Ayuuk language avoiding taking sides on the controversy about the number of consonants.

Mainly we made two replacements: ñ/ny and ch/tsy Some of the works were already aligned, others not. For those not aligned we created automatic alignments using the YASA tool (Lamraoui and Langlais). We discarded all empty and double alignments. Normalization and automatic alignments were manually verified by one of the authors. The corpus keep differences among both normalization variants: petakeros and bodegeros.

Finally, we randomly split the sentences into training, development and testing sets. For our experimentation we created two split versions, one strict and one random. In the strict version we use all the phrases from the National archive of indigenous languages (Lyon, 1980) as a test. Since these sentences are linguistically motivated and aim to show linguistic aspects of the language they tend to be harder to translate; This split resulted in 5,847/700/912 (train/dev/test). In the random split we randomly sample sentences from our sources, the final split resulted in 5,941/700/912 (train/dev/test). Notice that amount of phrases among splits changes, this is because after separating the test phrases, we remove repeated or similar phrases for the train/dev sets. Our intuition was to have a more uniform training/validation for the random split while the test follows the distribution of the original sources. We mimic this procedure for the strict sample.

4 Neural Architecture

Our translation model is based on the Transformer architecture (Vaswani et al., 2017). We use an encoder-decoder setting. For our experiments we

| Resource                                      | es | mir |
|-----------------------------------------------|----|-----|
| The bible                                     | Open | No open |
| Songs and poems                               | No open | No open |
| The Mexican constitution                     | Open | No open |
| Personal colection of Albino Pedro Juan      | No open | No open |
| Esopo Fables                                  | Open | No open |
| National archive of indigenous languages      | No open | Open |
| Social network                               | Open | Open |
| The dragon and the rabbit                     | Open | Open |
| Phrases translated by author                  | Open | Open |

Table 1: Source of data collected

8https://glottolog.org/resource/languoid/id/amer1254 (visited, last visited march 2021)
5 Experiments and results

As described in the previous section we have two different versions of our splits, strict and random. Per split we performed five experiments, two for configuration with fewer layers (A), and three for the configuration with more layers (B). We also modified: a) the maximum length of the phrase (50 or 70) b) the vocabulary of the BPE sub-word algorithm (we tested 2000 or 4000). Figure 1 shows the perplexity and the BLEU score in the development set during training for the direction Spanish (es) to Ayuuk (mir). The first part of the Table 2, columns two to five, presents the results on the development and test sets.

Figure 2 shows the learning curve on the direction of translation Ayuuk (mir) to Spanish (es). The second part of the table 2, columns six to nine, presents the results on the development and test for this translation direction.

These models were trained in a server with two Tesla V100 GPUs. To obtain a model it usually take us around 2h for a 100 epochs. We also were able to reproduce the experiments in the Colaboratory platform.

As we can appreciate these sets of experiments show that the translation is possible. We have some gains on the model with more layers (B), this is not trivial since we have a small amount of training data. On the other hand, the strict split as expected shows to be very difficult to translate, the BLEU scores are minimal. However with the random splits the BLEU scores are more promising. We also observe there that in the current setting it is more “easy” to translate from Spanish to Ayuuk than the other direction. Finally, we perform a larger experimentation with 250 epochs using the B configuration, following the intuition we haven reach the right performance with 100. Figure 3 shows the learning curve on the development set, the bottom part of Table 2 shows our final results using the random split.

6 Conclusions and Further work

Previous experiences on MT based on deep learning architecture, particularly on seq2seq settings, for native languages of the Americas have not been promising (Mager and Meza, 2018). In particular, because there is little to none training data. However, our work shows that a standard model based on the Transformer architecture and under
Table 2: BLEU scores of es-mir and mir-es.

| Configuration A 100 epochs | Strict es-mir | Random es-mir | Strict mir-es | Random mir-es |
|---------------------------|---------------|---------------|---------------|---------------|
| BLEU                      | dev | test | dev | test | dev | test | dev | test |
| Max length 50             | 1.72 | 0.05 | 1.66 | 1.71 | 0.64 | 0.10 | 0.91 | 0.66 |
| BPE 2000                  |     |      |     |      |     |      |     |      |
| Max length 50             | 2.03 | 0.10 | 1.21 | 1.24 | 1.02 | 0.16 | 0.93 | 0.83 |
| BPE 4000                  |     |      |     |      |     |      |     |      |

| Configuration B 100 epochs | Strict es-mir | Random es-mir | Strict mir-es | Random mir-es |
|---------------------------|---------------|---------------|---------------|---------------|
| BLEU                      | dev | test | dev | test | dev | test | dev | test |
| Max length 50             | 3.91 | 0.10 | 3.59 | 3.70 | 2.21 | 0.41 | 2.49 | 2.72 |
| BPE 2000                  |     |      |     |      |     |      |     |      |
| Max length 50             | 5.02 | 0.13 | 4.17 | 4.20 | 2.33 | 0.28 | 2.13 | 2.23 |
| BPE 4000                  |     |      |     |      |     |      |     |      |
| Max length 70             | 7.58 | 0.10 | 5.83 | 5.56 | 4.03 | 0.27 | 3.64 | 3.52 |
| BPE 4000                  |     |      |     |      |     |      |     |      |

| Configuration B 250 epochs | Random es-mir | Random mir-es |
|---------------------------|---------------|---------------|
| BLEU                      | dev | test | dev | test |
| Max length 70             | 5.83 | 5.56 | 3.64 | 3.52 |
| BPE 4000                  |     |      |     |      |

extremely low resource setting can produce some results. They are still low for normal standards of the MT field however they are promising for the future.

In order to improve the performance of the system future work will focus on:

1. Collecting more data, paying attention to other variants of the Ayuuk language.

2. Although the strict setting strongly penalizes the evaluation, we will continue using linguistic motivated phrases as a good bar to evaluate our progress.

3. At this moment we rely on sub-word of the phrases, however our approach could benefit from a deeper morphology analysis (Kann et al., 2018).

4. Our normalization will continue respecting the petakeros and bodegeros positions, and for other variants we also incorporate positions regarding the number of vowels.

Acknowledgements

The authors thank CONACYT for the computer resources provided through the INAOE Supercomputing Laboratory’s Deep Learning Platform for Language Technologies. We also thank the project “Traducción automática para lenguas indígenas de México” PAPIIT-IA104420, UNAM.

References

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Katharina Kann, Jesus Manuel Mager Hois, Ivan Vladimir Meza-Ruiz, and Hinrich Schütze. 2018. 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, Volume 1 (Long Papers), pages 47–57, New Orleans, Louisiana. Association for Computational Linguistics.

Julia Kreutzer, Jasmijn Bastings, and Stefan Riezler. 2019. Joey NMT: A minimalist NMT toolkit for novices. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations, pages 109–114, Hong Kong, China. Association for Computational Linguistics.

Fethi Lamraoui and Philippe Langlais. Yet another fast, robust and open source sentence aligner. time to reconsider sentence alignment. XIV Machine Translation Summit.
Don D. Lyon. 1980. *Mixe de Tlahuitoltepec, Oaxaca, Archivo de Lenguas Indígenas de México*. Colegio de México, México.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018. *Challenges of language technologies for the indigenous languages of the Americas*. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Manuel Mager and Ivan Meza. 2018. Hacia la traducción automática de las lenguas indígenas de México. *Proceedings of the DH*.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsu Umu, Salomon Kabongo Kabenamulu, Salomey Osei, Freshia Sackey, Rubungo Andre Nyongabo, Ricky Macharm, Perez Ogayo, Orevaoghen E. Ahia, Musir Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogweji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Gholah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefulu, Chris Chinemy Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabi, Arshath Ramkilowan, Alp Oktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

J. Carlos Reyes Gómez. 2005. *Aportes al proceso de enseñanza aprendizaje de la lectura y la escritura de la lengua ayuuk*. Centro de Estudios Ayuuk–Universidad Indígena Intercultural Ayuuk, Oaxaca, México.

Ana Sagi-Vela González. 2019. El mixe escrito y el espejismo del buen alfabeto. *Revista de Llengua i Dret*, (71).

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. *Neural machine translation of rare words with subword units*. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. *Attention is all you need*. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.