Neural Machine Translation in Low-Resource Contexts: Indigenous Languages of the Americas and Spanish

Jesus A. Hermosillo
Department of Linguistics
jadolfoh@stanford.edu

Abstract

Many indigenous languages of the Americas are polysynthetic languages, that is, languages whose words are composed of many morphemes and often the boundary between words and sentences tends to blur. These languages are also considered low-resource languages, or languages with very few written or digitized records. The limited existent data for indigenous languages is also prone to orthographic, dialectal, and diachronic variation due to a lack of standardization and strong oral traditions. These factors combined make neural machine translation a particularly difficult task. In this paper, we investigate how neural machine translation models can handle the tasks of translating from and into Nahuatl and Wixarika, two indigenous languages that fit both descriptions. In order to do so, we use an LSTM-based sequence-to-sequence model with attention and two mid-size Transformer-based models with different feed-forward (FFN) sub-layer dimensions. Our models achieve promising results for both language pairs in either translation direction. We find that NMT performance drops considerably in the task of translating into polysynthetic languages but adjusting the dimensionality of the FNN in the Transformer-based model can improve BLEU scores by at least 0.1 regardless of the typology of the target language.

1 Introduction

Machine translation for indigenous languages of the Americas is particularly tough. The limited amount of monolingual and parallel data available for these language poses a challenge on this and many other data-intensive tasks. Existing parallel corpora for these languages are composed of sentences with their translation equivalent in a language spoken by the majority of the population in the country in which they are spoken. In the Americas these languages are Spanish, Portuguese, English, or French. As a result, these parallel corpora contain language pairs with very distant typologies. This is particularly true for Nahuatl, a polysynthetic indigenous language from the Uto-Aztecan language family, and Spanish, a fusional language from the Indo-European language family. A significant difference between polysynthetic languages like Nahuatl and fusional languages like Spanish, is that words in the former are composed of multiple morphemes with different meanings, and words in the latter tend to squeeze in multiple units of meaning into one single morpheme. Further, in polysynthetic languages like Nahuatl, words can be composed of so many morphemes that their equivalent translation in a fusional language like Spanish is realized as a sentence with multiple words (see Figure 1). Data scarcity and distant typologies make machine translation a very difficult task, especially in the task of translating from a fusional to a polysynthetic language, as many morphemes in a polysynthetic languages do not have Spanish counterpart. In this paper, we investigate how neural models perform under low resource conditions on both translation directions: from a polysynthetic language to a fusional language and vice versa . We train our models on parallel

,

Nahuatl:	Nimitztētlamaquiltīz					
	ni-mits-teː-tɬa-maki-ltiː-s'					
	I-you.OBJ-someone-something-give-CAUSATIVE-FUTURE					
Spanish:	Yo	haré	que alguien	te	de	algo
	Yo	har-é	que alguien	te	d-e	algo
	I	make-1st.SG.FUT.	that someone	you.DAT	give-3rd.SG.SBJV	something
English:	"I shall make somebody give something to you"					

(Suarez, 1983)

Figure 1: Morpheme differences in Nahuatl, a polysynthetic language and Spanish, a fusional langauge, with the corresponding free translation in English.

corpora from Nahuatl and Wixarika, two genetically related polysynthetic languages, and Spanish, a fusional language.

2 Related Work

Machine Translation for low resources languages of the Americas has relied on a variety of methods, including rule-based, statistical, and neural approaches (Mager et al., 2018). Machine translation systems for Nahuatl and Wixarika use parallel corpora that have Spanish as the other parallel language. Word based SMT approaches have been used to explore morpheme to word (or token) alignment in Wixarika-Spanish data (Mager Hois et al., 2016; Mager et al., 2018) and Nahuatl-Spanish data (Mager et al., 2018). Morpheme segmentation has a positive impact on performance in the task of translating from Wixarika or Nahuatl into Spanish, however, the reverse suffers from information loss as many morphemes in these indigenous languages do not have a Spanish counterpart. Neural machine translation has also been explored for these language pairs on the task of translating into Spanish using an RNN-based sequence-to-sequence model (Mager et al). In their study, they use a parallel corpus of only 985 sentences. Under these extreme circumstances, a traditional SMT model (bleu = 10.14) outperforms the neural model (bleu = 3.04) in the task of translating from Nahuatl into Spanish. In the task of translating from Wixarika to Spanish both models performed poorly (bleu = 0). Work on other languages has found that hyper-parameter tuning of RNN-based models coupled with reducing vocabulary size can outperform previous statistical approaches (Sennrich & Zhang, 2019). More recent work has also explored the use of the Transformer (Vaswani et al), a sequence-to-sequence model that relies solely on attention mechanism rather than traditional recurrence (Araabi & Monz, 2020; van Biljon et al., 2020). In line with Sennrich & Zhang, these studies find that hyper-parameter tuning and reducing the vocabulary size was found improve model performance. More specifically, moderately reducing the size of original architecture and the dimension of sub layers leads to better performance.

3 Data

We use two parallel corpora provided by the organizers of the first Workshop on NLP for Indigenous Languages of the Americas. The first corpus is a portion of the Axolotl corpus (Gutierrez-Vasques et al., 2016), a digital parallel corpus of Spanish and Nahuatl. The Axolotl corpus is a collection of mostly printed documents from various domains including historical data, short stories, literature, didactic content, recipes, magazines, music, and legal data. We must note that this corpus contains diachronic, orthographic and dialectal variation as Nahuatl is a stand-in for the set of unspecified varieties present in the Axolotl corpus. The portion of the Axolotl corpus we will be working with has 16062 parallel sentences. The second corpus is the Wixarika-Spanish Parallel Corpus (Mager et al.). This corpus is a collection of Hans Christian Andersen's and the Brothers Grimm's classic fairy tales. Unlike the Axolotl corpus, this corpus is more uniform as it is written in one dialect, the Zoquipan Wixarika variety. The Wixarika-Spanish Parallel Corpus has 8966 parallel sentences. We

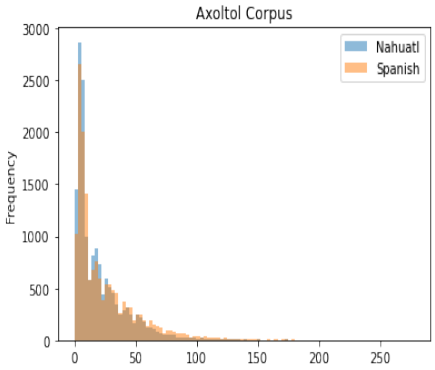


Figure 2: Distributions of sentence lengths for Nahuatl and Spanish in the Axolotl Corpus.

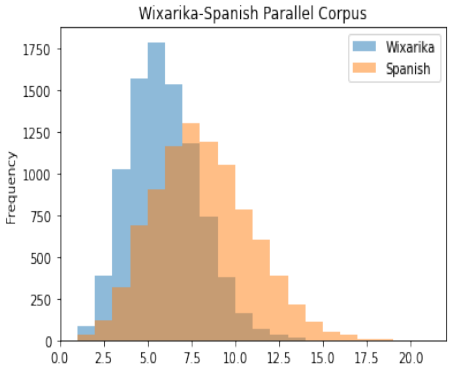


Figure 3: Distributions of sentence lengths for Wixarika and Spanish in the Wixarika-Spanish Parallel Corpus.

decided to use the only the training sets provided by Americas NLP to avoid further variation in the development sets.

3.1 Preprocessing

Subword tokenization has shown to improve performance in NMT (Sennrich et al., 2016). We implement subword tokenization using SentencePiece, a language independent subword tokenizer (Kudo & Richardson, 2018). Before we trained our sub-word tokenization models, we made sure our data sets were tokenized by words and that no empty lines were present in the corpus. We prepared our Wixarika-Spanish corpus by applying a word tokenizer to Wixarika sentences and Spanish word tokenizer from NLTK for Spanish sentences. The Axolotl corpus was already word tokenized, so no further action was necessary for either language. Given the amount of our data, we decided to train SentencePiece models with vocabulary sizes of 2000 as reducing the number of unique tokens improves performance in neural machine translation under low-resource conditions (Sennrich & Zhang, 2019). The models were trained on the whole data set for each language. After SentencePiece is applied, we truncate sentences so that they have at most 400 tokens. We then split our filtered parallel corpus into training, validation and test sets with 90%, 5% and 5% of data respectively.

4 Method

The main objective of this paper is to translate from a polysynthetic to a fusional language and vice versa. To do this we use two sequence-to-sequence architectures. A sequence-to-sequence architecture has two main components: an encoder and a decoder. The encoder takes a sequential input and produces a representation. This representation is then fed to the decoder which uses it to generate an output sequence. In this experiment, we use an LSTM-based model with attention as our baseline and compare it to a modified transformer-based model(see figures 3 and 4).

4.1 Baseline

Our baseline model is an LSTM-based Seq2Seq model with attention (Bahdanau et al., 2014). RNN-based models are very popular in this type of task. These models can be improved using attention mechanisms. This mechanism assigns weights to the inputs of the encoder based on their importance in the sequence, then these weights are additionally fed to the decoder to produce better translations. Our architecture consists of the 2 Bi-LSTMs in the encoder, 2 LSTMs in the encoder with 500 hidden units, and an attention-mechanism.

4.2 Transformers

We compare our baseline model to a mid-size Transformer (Vaswani et al, 2017). In this architecture, words in a sequence are converted into word embeddings of fixed dimension, which is in turn

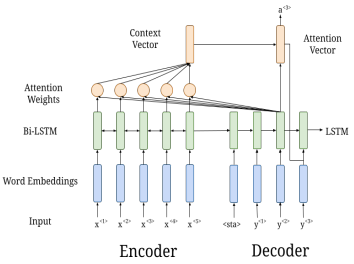


Figure 4: LSTM-based sequence-to-sequence model with attention.

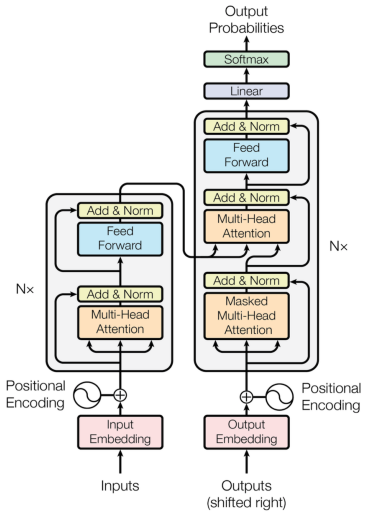


Figure 5: Transformer Model

the dimension of the output of every sub-layer of the network. Since the model does not rely on recurrence, information about the position of words is added to the word vectors using a positional encoding function. The encoder consists of a multi-head self-attention sub-layer and a position-wise feed forward sub-layer (FFN). There is also a residual connection around the sub-layers that is followed by normalization function. This will make sure that information in the input (e.g., positional encoding) is not lost along the way. The encoder has a similar structure to the encoder, but it has an additional self-attention sub-layer that allows it to learn information from the output of the encoder. Further, the self-attention sub-layer in the decoder only attends to subsequent positions so that the output only depends on previously known outputs for a given position. The output of the decoder goes through a linear transformation function and soft-max function that produce next-token probabilities. We follow the optimization of van Biljon et al.(2020) for a mid-size Transformer. In doing so, we modify the original transformer and reduce the number of layers in the encoder and decoder from 6 to 3, the number of attention heads from 8 to 4. Diffrent from Biljon et al., we chose a smaller input dimension, from 512 in the original architecture and 256 in van Biljon et al., to 128. We reduced the dimensions of the FFN, from 2048 to 1024 as in van Biljon et al in one model, to 512 in another. We chose a batch size of 8. We validated every 1,000 iterations for a total of 35,000 iterations per model. We evaluate model performance using bleu (Papineni et al 2002)and CHRF++ (Popovic, 2017). We use the PyTorch implementation of OpenNMT for our model (Klein et al., 2017).

5 Results Discussion

5.1 Experiment 1

In the first experiment, we translated from a polysynthetic language to a fusional language. In general, models trained on the Wixarika-Spanish parallel corpus (Wixarika-to-Spanish) achieved higher BLEU and CHRF++ scores than the models trained on the Axolotl corpus (Nahuatl-to-Spanish). The only exception was the Transformer-1024 which had the same CHRF++ score for both corpora (see Table 1, highest scores in bold). Bleu scores show that the Transformer-based models outperformed the LSTM-based model by 5.6 in Nahuatl-to-Spanish and by 3.3 in Wixarika-to-Spanish. CHRF++ scores also show that Transformer-based models outperformed the LSTM model by at least 0.08 in Nahuatl-to-Spanish and by 0.04 in Wixarika-to-Spanish. Increasing the dimension of the FFN increased bleu score by 0.2 in Nahuatl-to-Spanish and decreased it by 0.01 in Wixarika-to-Spanish. The CHRF++ scores increased by 0.01 in Nahuatl-to-Spanish but remained the same for Wixarika-to-Spanish.

	bleu		chrf++	
Model	nah-es	hch-es	nah-es	hch-es
LSTM	14.1	18.8	0.29	0.34
T-512	19.7	**22.1**	0.37	**0.38**
T-1024	**19.9**	22.0	**0.38**	**0.38**

Table 1: Bleu and Chrf++ scores for experiment 1

	bleu		chrf++	
Model	es-nah	es-hch	es-nah	es-hch
LSTM	6.1	6.9	0.16	**0.13**
T-512	6.3	**7.2**	**0.18**	**0.13**
T-1024	**6.5**	7.1	**0.18**	**0.13**

Table 2: Bleu and Chrf++ scores for experiment 2

5.2 Experiment 2

In the second experiment, we translated from a fusional language to a polysynthetic language. Similar to experiment 1, models trained on the Wixarika-Spanish parallel corpus achieved higher BLEU scores than the models trained on the Axolotl corpus, but, the models trained on the Axolotl corpus achieved higher CHRF++ scores(see Table 2, highest scores in bold). BLEU scores show that the Transformer-based models outperformed the LSTM-based model by at least 0.2 in both Spanish-to-Nahuatl and Spanish-to-Wixarika. CHRF++ scores show that Transformer-based models outperformed the LSTM model by 0.02 in Spanish-to-Nahuatl but did no effect was found in Spanish-to-Wixarika. Similar to experiment 1, increasing the dimension of the FFN increased BLEU scores by 0.2 in Spanish-to-Nahuatl and decreased them by 0.01 in Spanish-to-Wixarika. No effect was found in CHRF++ scores.

The best model based on both BLEU and CHRF++ scores was the transformer-512 trained on the task of translating from Wixarika to Spanish. The worst model based on bleu was the LSTM model trained on the task of translating from Spanish to Nahuatl. The worst model based on CHRF++ was also the LSTM model trained on task of translating from Spanish to Wixarika. The best architecture for the Axoltol corpus in both translation tasks was the Transformer-1024. Because the number of iterations was the same for both datasets, the models were trained for a little bit more than 2 epochs for the Axoltol corpus and around a little bit more than 3 for the Axoltol coprus. Thus, this difference in performance could change if we trained the models for the same number of epochs (see Appendix for performance improvement over iterations). Even then, our models achieve promising results for the tasks of translating from Nahuatl and Wixarika into Spanish and vice versa. As seen in statistical machine translation, translating from a polysynthetic language to a fusional language is very difficult. It is evident that information loss found in statistical machine translation is also present in neural machine translation. Nonetheless, even in such circumstances, adjusting the size of the FFN appears helpful in improving performance.

6 Conclusion

In this paper, we used an LSTM and two Transformer based sequence-to-sequence architectures to perform the tasks of translating from a polysynthetic to a fusional language and vice versa. Our models achieve promising results for both tasks for Nahuatl-Spanish and Wixarika-Spanish language pairs. We find that increasing the size of the FFN in a mid-size transformer counterproductive for smaller datasets regardless. In future work we would like to increase the number of epochs and reduced the dimensions of FFN sub-layers in the Transformer for languages with less data.

7 Acknowledgments

The author of this project would like to thank Professor Christopher Potts and Johnathan Li for their invaluable feedback and comments throughout the development of this project. The initial design of this project was completed in part for the guided research projects in Linguistics course.

8 References

Aarabi, A., & Monz, C. Optimizing Transformer for Low-Resource Neural Machine Translation. arXiv preprint *arXiv:2011.02266v1*

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473.*

Gutierrez-Vasques, X., Sierra, G., & Pompa, I. H. (2016). Axolotl: a Web Accessible Parallel Corpus for Spanish-Nahuatl. In LREC.

Klein, G., Kim, Y., Deng, Y., Senellart, J., & Rush, A. M. (2017). Opennmt: Open-source toolkit for neural machine translation. arXiv preprint *arXiv:1701.02810.*

Kudo, T., & Richardson, J. (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. arXiv preprint *arXiv:1808.06226.*

Mager Hois, J. M., Barrón Romero, C., & Meza Ruiz, I. V. (2016). Traductor estadístico wixarika-español usando descomposición morfológica, *COMTEL (6).*

Mager, M., Dionico, C., & Meza, I. The Wixarika-Spanish Parallel Corpus.

Mager, M., Gutierrez-Vasques, X., Sierra, G., & Meza, I. (2018). Challenges of language technologies for the indigenous languages of the Americas. arXiv preprint *arXiv:1806.04291.*

Mager, M., & Meza, I. (2018. Hacia la traducción automática de las lenguas indígenas de méxico. In *Proceedings of the 2018 Digital Humanities Conference*. The Association of Digital Humanities Organizations

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311-318).

Popović, M. (2017). chrF++: words helping character n-grams. In *Proceedings of the second conference on machine translation* (pp. 612-618).

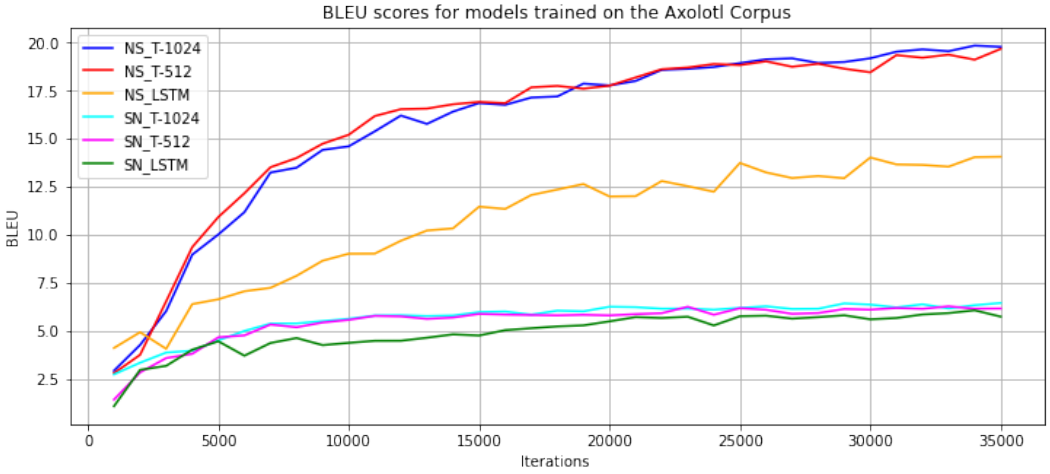
Sennrich, R. & Zhang, B. (2019). Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, ACL 2019, pages 211– 221.

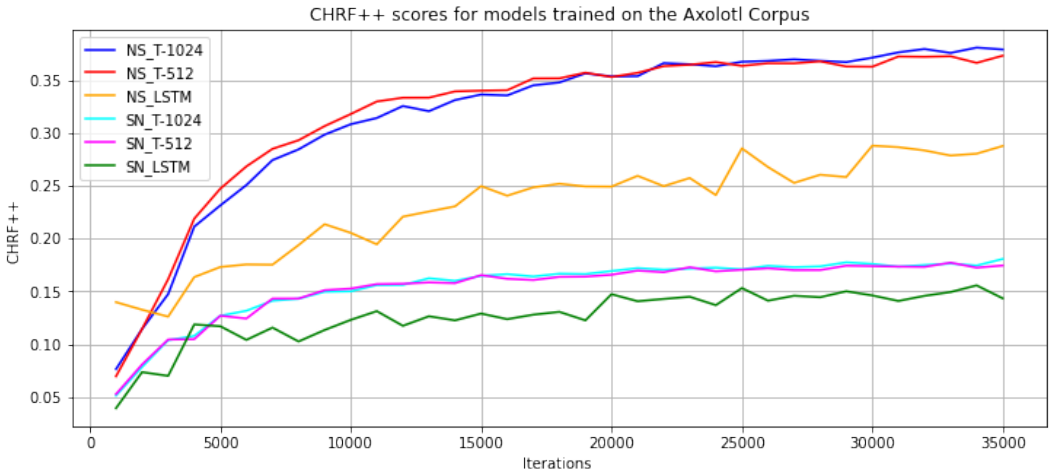
Sennrich, R, Haddow, B., & Birch, A. (2015) "Neural machine translation of rare words with subword units." arXiv preprint *arXiv:1508.07909* .

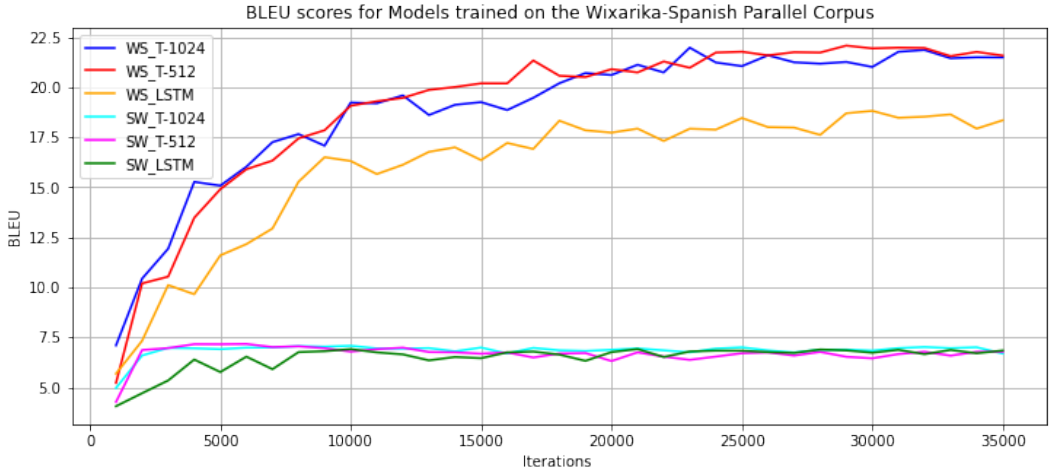
van Biljon, E., Pretorius, A., & Kreutzer, J. (2020). On optimal transformer depth for low-resource language translation. *CoRR, abs/2004.04418.*

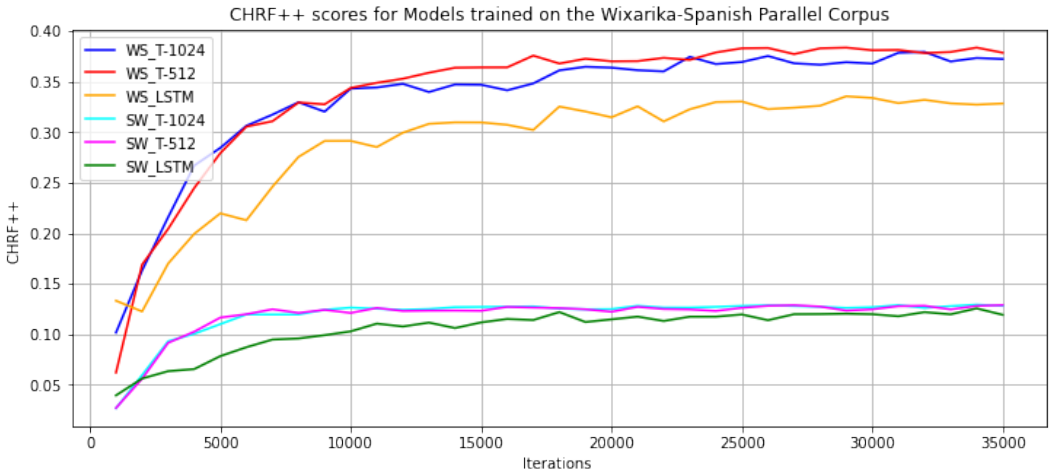
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*, pages 5998–6008.

9 Appendix




CHRF++ scores for models trained on the Axolotl Corpus
NS_T-1024
NS_T-512
NS_LSTM
SN_T-1024
SN_T-512
SN_LSTM
CHRF++
Iterations


BLEU scores for Models trained on the Wixarika-Spanish Parallel Corpus
WS_T-1024
WS_T-512
WS_LSTM
SW_T-1024
SW_T-512
SW_LSTM
BLEU
Iterations


CHRF++ scores for Models trained on the Wixarika-Spanish Parallel Corpus
WS_T-1024
WS_T-512
WS_LSTM
SW_T-1024
SW_T-512
SW_LSTM
CHRF++
Iterations