Open Machine Translation for Low Resource South American Languages (AmericasNLP 2021 Shared Task Contribution)

Shantipriya Parida[1], Subhadarshi Panda[2], Amulya Ratna Dash[3], Esaú Villatoro-Tello[1,4], A. Seza Doğruöz[5], Rosa M. Ortega-Mendoza[6], Amadeo Hernández[6], Yashvardhan Sharma[3], Petr Motlicek[1]

[1]Idiap Research Institute, Martigny, Switzerland
{firstname.lastname}@idiap.ch
[2]Graduate Center, City University of New York, USA
spanda@gradcenter.cuny.edu
[3]BITS, Pilani, India
{p20200105,yash}@pilani.bits-pilani.ac.in
[4]Universidad Autónoma Metropolitana Cuajimalpa, Mexico City, Mexico
evillatoro@cua.uam.mx
[5]Ghent University, Belgium
as.dogruoz@ugent.be
[6]Universidad Politécnica de Tulancingo, Hidalgo, Mexico
{rosa.ortega,amadeo.hernandez1911001}@upt.edu.mx

Abstract

This paper describes the team ("Tamalli")'s submission to AmericasNLP2021 shared task on Open Machine Translation for low resource South American languages. Our goal was to evaluate different Machine Translation (MT) techniques, statistical and neural-based, under several configuration settings. We obtained the *second-best* results for the language pairs "Spanish-Bribri", "Spanish-Asháninka", and "Spanish-Rarámuri" in the category "Development set not used for training". Our performed experiments will serve as a point of reference for researchers working on MT with low-resource languages.

1 Introduction

The main challenges in automatic Machine Translation (MT) are the acquisition and curation of parallel data and the allocation of hardware resources for training and inference purposes. This situation has become more evident for Neural Machine Translation (NMT) techniques, where their translation quality depends strongly on the amount of available training data when offering translation for a language pair. However, there is only a handful of languages that have available large-scale parallel corpora, or collections of sentences in both the source language and corresponding translations. Thus, applying recent NMT approaches to low-resource languages represent a challenging scenario.

In this paper, we describe the participation of our team (aka, Tamalli) in the Shared Task on Open Machine Translation held in the First Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP) (Mager et al., 2021).[1] The main goal of the shared task was to encourage the development of machine translation systems for indigenous languages of the Americas, categorized as low-resources languages. This year *8* different teams participated with *214* submissions.

Accordingly, our main goal was to evaluate the performance of traditional statistical MT techniques, as well as some recent NMT techniques under different configuration settings. Overall, our results outperformed the baseline proposed by the shared task organizers, and reach promising results for many of the considered pair languages.

The paper is organized as follows: Section 2 briefly describes some related work; Section 3 depicts the methodology we followed for performing our experiments. Section 4 provides the dataset descriptions. Section 5 provides the details from our different settings, and finally Section 6 depict our main conclusions and future work directions.

2 Related work

Machine Translation (Garg and Agarwal, 2018) is a field in NLP that aims to translate natural lan-

[1] http://turing.iimas.unam.mx/americasnlp/st.html

Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas, pages 218–223

guages. Particularly, the development of (MT) systems for indigenous languages in both South and North America, faces different challenges such as a high morphological richness, agglutination, polysynthesis, and orthographic variation (Mager et al., 2018b; Llitjós et al., 2005). In general, MT systems for these languages in the state-of-the-art have been addressed by the sub-fields of machine translation: rule-based (Monson et al., 2006), statistical (Mager Hois et al., 2016) and neural-based approaches (Ortega et al., 2020; Le and Sadat, 2020). Recently, NMT approaches (Stahlberg, 2020) have gained prominence; they commonly are based on sequence-to-sequence models using encoder-decoder architectures and attention mechanisms (Yang et al., 2020). From this perspective, different morphological segmentation techniques have been explored (Kann et al., 2018; Ortega et al., 2020) for Indigenous American languages.

It is known that the NMT approaches are based on big amounts of parallel corpora as source knowledge. To date, important efforts toward creating parallel corpora have been carried out for specific indigenous languages of America. For example, for Spanish-Nahuatl (Gutierrez-Vasques et al., 2016), Wixarika-Spanish (Mager et al., 2020) and Quechua-Spanish (Llitjós et al., 2005) which includes morphological information. Also, the JHU Bible Corpus, a parallel text, has been extended by adding translations in more than 20 Indigenous North American languages (Nicolai et al., 2021). The usability of the corpus was demonstrated by using multilingual NMT systems.

3 Methodology

Since the data sizes are small in most language pairs as shown in Table 1, we used a statistical machine translation model. We also used NMT models. In the following sections, we describe the details of each of these approaches.

3.1 Statistical MT

For statistical MT, we relied on an IBM model 2 (Brown et al., 1993) which comprises a lexical translation model and an alignment model. In addition to the word-level translation probability, it models the absolute distortion in the word positioning between source and the target languages by introducing an alignment probability, which enables to handle word reordering.

3.2 Neural MT

For NMT, we first tokenized the text using sentence piece BPE tokenization (Kudo and Richardson, 2018).[2] The translation model architecture we used for NMT is the transformer model (Vaswani et al., 2017). We trained the model in two different setups as outlined below.

One-to-one: In this setup, we trained the model using the data from one source language and one target language only. In the AmericasNLP2021[3] shared task, the source language is always Spanish (es). We trained the transformer model using Spanish as the source language and one of the indigenous languages as the target language.

One-to-many: Since the source language (Spanish) is constant for all the language pairs, we considered sharing the NMT parameters across language pairs to obtain gains in translation performance as shown in previous work (Dabre et al., 2020). For this, we trained a one-to-many model by sharing the decoder parameters across all the indigenous languages. Since the model needs to generate the translation in the intended target language, we provided that information as a target language tag in the input (Lample and Conneau, 2019). The token level representation is obtained by the sum of token embedding, positional embedding, and language embedding.

4 Dataset

For training and evaluating our different configurations, we used the official datasets provided by the organizers of the shared task. It is worth mentioning that we did not use additional datasets or resources for our experiments.

A brief description of the dataset composition is shown in Table 1. For all the language pairs, the task was to translate from Spanish to some of the following indigenous languages: Hñähñu (oto), Wixarika (wix), Nahuatl (nah), Guaraní (gn), Bribri (bzd), Rarámuri (tar), Quechua (quy), Aymara (aym), Shipibo-Konibo (shp), Asháninka (cni). For the sake of brevity, we do not provide all the characteristics of every pair of languages. The interested reader is referred to (Gutierrez-Vasques et al.,

[2] We also compared the BPE subword tokenization to word-level tokenization using Moses tokenizer and character level tokenization. We found that the best results were obtained using the BPE subword tokenization.

[3] http://turing.iimas.unam.mx/americasnlp/

Language-pair	Train(#Tokens)			Dev(#Tokens)			Test(#Tokens)	
	#Sentences	Source	Target	#Sentences	Source	Target	#Sentences	Source
es-aym	6531	128154	97276	996	11129	7080	1003	10044
es-bzd	7508	46820	41141	996	11129	12974	1003	10044
es-cni	3883	48752	26096	883	9605	6070	1003	10044
es-gn	26032	604841	405984	995	11129	7191	1003	10044
es-hch	8966	68683	48919	994	11129	10296	1003	10044
es-nah	16145	470003	351580	672	6329	4300	1003	10044
es-oto	4889	68226	72280	599	5115	5069	1003	10044
es-quy	125008	1898377	1169644	996	11129	7406	1003	10044
es-shp	14592	88447	62850	996	11129	9138	1003	10044
es-tar	14720	141526	103745	995	11129	10377	1003	10044

Table 1: Statistics of the official dataset. The statistics include the number of sentences and tokens (train/dev/test) for each language pair.

Task	Baseline		Tamalli			Best Competitor	
	BLEU	CharF	Submission#	BLEU	CharF	BLEU	CharF
es-aym	0.01	0.157	4	0.03	0.202	2.29	0.283
es-bzd	0.01	0.068	3	1.09	0.132	2.39	0.165
es-cni	0.01	0.102	1	0.01	0.253	3.05	0.258
es-gn	0.12	0.193	5	1.9	0.207	6.13	0.336
es-hch	2.2	0.126	1	0.01	0.214	9.63	0.304
es-nah	0.01	0.157	1	0.03	0.218	2.38	0.266
es-oto	0	0.054	1	0.01	0.118	1.69	0.147
es-quy	0.05	0.304	5	0.96	0.273	2.91	0.346
es-shp	0.01	0.121	1	0.06	0.204	5.43	0.329
es-tar	0	0.039	1	0.04	0.155	1.07	0.184

Table 2: Evaluation Results. All results are from the “Track2: Development Set Not Used for Training". For all the tasks, the source language is Spanish. The table contains the best results of our team against the best score by the competitor in its track.

2016; Mager et al., 2018a; Chiruzzo et al., 2020; Feldman and Coto-Solano, 2020; Agić and Vulić, 2019; Prokopidis et al., 2016; Galarreta et al., 2017; Ebrahimi et al., 2021) for knowing these details.

5 Experimental results

We used 5 settings for all the 10 pair translations. The output of each set is named as version [1-5] and submitted for evaluation (shown under column Submission# in Table 2). Among the 5 versions, version [1] is based on statistical MT, and version [2-5] is based on NMT with different model configurations. For model evaluation, organizers provided a script that uses the metrics *BLEU* and *ChrF* for machine translation evaluation. The versions and their configuration details are explained below. We included the best results only from all the versions [1-5] in Table 2.

Version 1: Version 1 uses the statistical MT. The source and target language text were first tokenized using Moses tokenizer setting the language to Spanish. Then we trained the IBM translation model 2 (Brown et al., 1993) implemented in `nltk.translate` api. After obtaining the translation target tokens, the detokenization was carried out using the Moses Spanish detokenizer.

Version 2: This version uses the one-to-one NMT model. First, we learned sentence piece BPE tokenization (Kudo and Richardson, 2018) by combining the source and target language text. We set the maximum vocabulary size to {8k, 16k, 32k} in different runs and we considered the run that produced the best BLEU score on the dev set. The

transformer model (Vaswani et al., 2017) was implemented using PyTorch (Paszke et al., 2019). The number of encoder and decoder layers was set to 3 each and the number of heads in those layers was set to 8. The hidden dimension of the self-attention layer was set to 128 and the position-wise feed-forward layer's dimension was set to 256. We used a dropout of 0.1 in both the encoder and the decoder. The encoder and decoder embedding layers were not tied. We trained the model using early stopping with a patience of 5 epochs, that is, we stop training if the validation loss does not improve for 5 consecutive epochs. We used greedy decoding for generating the translations during inference. The training and translation were done using one GPU.

Version 3: This version uses the one-to-many NMT model. For tokenization, we learned sentence piece BPE tokenization (Kudo and Richardson, 2018) by combining the source and target language text from all the languages (11 languages in total). We set the maximum shared vocabulary size to {8k, 16k, 32k} in different runs and we considered the run that produced the best BLEU score on the dev set. The transformer model's hyperparameters were the same as in version 2. The language embedding dimension in the decoder was set to 128. The encoder and decoder embedding layers were not tied. We first trained the one-to-many model till convergence using early stopping with the patience of 5 epochs, considering the concatenation of the dev data from all the language pairs. Then we fine-tuned the best checkpoint using each language pair's data separately. The fine-tuning process was also done using early stopping with patience of 5 epochs. Finally, we used greedy decoding for generating the translations during inference. The training and translation were done using one GPU.

Version 4: This version is based on one-to-one NMT. We have used the *Transformer* model as implemented in OpenNMT-py (PyTorch version) (Klein et al., 2017).[4]. To train the model, we used a single GPU and followed the standard "Noam" learning rate decay,[5] see (Vaswani et al., 2017; Popel and Bojar, 2018) for more details. Our starting learning rate was 0.2 and we used 8000 warm-up steps. The model *es-nah* trained up to 100K iterations and the model checkpoint at 35K was selected based on the evaluation score (*BLEU*) on the development set.

Version 5: This version is based on One-to-One NMT. We have used the *Transformer* model as implemented in OpenNMT-tf (Tensorflow version) (Klein et al., 2017). To train the model, we used a single GPU and followed the standard "Noam" learning rate decay,[6] see (Vaswani et al., 2017; Popel and Bojar, 2018) for more details. We used 8K shared vocab size for the models and the model checkpoints were saved at an interval of 2500 steps. The starting learning rate was 0.2 and 8000 warm-up steps were used for model training. The early-stopping criterion was 'less than 0.01 improvement in BLEU score' for 5 consecutive saved model checkpoints. The model *es-gn* was trained up to 37.5K iterations and the model checkpoint at 35K was selected based on evaluation scores on the development set. The model *es-quy* was trained up to 40K iterations and the model checkpoint at 32.5K was selected based on evaluation scores on the development set.

We report the official automatic evaluation results in Table 2. The machine translation evaluation matrices BLEU (Papineni et al., 2002) and ChrF (Popović, 2017) used by the organizers to evaluate the submissions. Based on our observation, the statistical approach performed well as compared to NMT for many language pairs as shown in the Table 2 (Parida et al., 2019). Also, among NMT model settings one-to-one and one-to-many perform well based on the language pairs.

6 Conclusions

Our participation aimed at analyzing the performance of recent NMT techniques on translating indigenous languages of the Americas, low-resource languages. Our future work directions include: *i)* investigating corpus filtering and iterative augmentation for performance improvement (Dandapat and Federmann, 2018), *ii)* review already existing extensive analyses of these low-resource languages from a linguistic point of view and adapt our methods for each language accordingly, *iii)* exploring transfer learning approach by training the model on a high resource language and later transfer it to a low resource language (Kocmi et al., 2018).

[4] http://opennmt.net/

[5] https://nvidia.github.io/OpenSeq2Seq/html/api-docs/optimizers.html

[6] https://nvidia.github.io/OpenSeq2Seq/html/api-docs/optimizers.html

Acknowledgements

The authors Shantipriya Parida and Petr Motlicek were supported by the European Union's Horizon 2020 research and innovation program under grant agreement No. 833635 (project ROXANNE: Real-time network, text, and speaker analytics for combating organized crime, 2019-2022). Author Esaú Villatoro-Tello, was supported partially by Idiap Research Institute, SNI-CONACyT, and UAM-Cuajimalpa Mexico during the elaboration of this work. Author Rosa M. Ortega-Mendoza was supported partially by SNI-CONACyT.

The authors do not see any significant ethical or privacy concerns that would prevent the processing of the data used in the study. The datasets do contain personal data, and these are processed in compliance with the GDPR and national law.

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.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Luis Chiruzzo, Pedro Amarilla, Adolfo Ríos, and Gustavo Giménez Lugo. 2020. Development of a Guarani - Spanish parallel corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2629–2633, Marseille, France. European Language Resources Association.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Comput. Surv.*, 53(5).

Sandipan Dandapat and Christian Federmann. 2018. Iterative data augmentation for neural machine translation: a low resource case study for english-telugu. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation: 28-30 May 2018, Universitat d'Alacant, Alacant, Spain*, pages 287–292. European Association for Machine Translation.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir, Gustavo A. Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando A. Coto Solano, Ngoc Thang Vu, and Katharina Kann. 2021. Americasnli: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages.

Isaac Feldman and Rolando Coto-Solano. 2020. Neural machine translation models with back-translation for the extremely low-resource indigenous language Bribri. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3965–3976, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ana-Paula Galarreta, Andrés Melgar, and Arturo Oncevay. 2017. Corpus creation and initial SMT experiments between Spanish and Shipibo-konibo. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 238–244, Varna, Bulgaria. INCOMA Ltd.

Ankush Garg and Mayank Agarwal. 2018. Machine translation: A literature review. *arXiv*.

Ximena Gutierrez-Vasques, Gerardo Sierra, and Isaac Hernandez. 2016. Axolotl: A web accessible parallel corpus for Spanish-Nahuatl. *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, pages 4210–4214.

Katharina Kann, Manuel Mager, Ivan Meza-Ruiz, and Hinrich Schütze. 2018. Fortification of neural morphological segmentation models for polysynthetic minimal-resource languages. In *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, volume 1, pages 47–57.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*.

Tom Kocmi, Shantipriya Parida, and Ondřej Bojar. 2018. Cuni nmt system for wat 2018 translation tasks. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 5th Workshop on Asian Translation: 5th Workshop on Asian Translation*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Tan Ngoc Le and Fatiha Sadat. 2020. Low-Resource NMT: an Empirical Study on the Effect of Rich Morphological Word Segmentation on Inuktitut. *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, 1(2012):165–172.

Ariadna Font Llitjós, Lori Levin, and Roberto Aranovich. 2005. Building Machine translation systems for indigenous languages. *Communities*.

Manuel Mager, Diónico Carrillo, and Ivan Meza. 2018a. Probabilistic finite-state morphological segmenter for wixarika (huichol) language. *Journal of Intelligent & Fuzzy Systems*, 34(5):3081–3087.

Manuel Mager, Carrillo Dionico, and Ivan Meza. 2020. The Wixarika-Spanish Parallel Corpus The Wixarika-Spanish Parallel Corpus. (August 2018).

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018b. 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, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Anna Currey, Vishrav Chaudhary, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager, Ngoc Thang Vu, Graham Neubig, and Katharina Kann. 2021. Findings of the AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas. In *Proceedings of theThe First Workshop on NLP for Indigenous Languages of the Americas*, Online. Association for Computational Linguistics.

Jesus Manuel Mager Hois, Carlos Barrón Romero, and Ivan Vladimir Meza Ruiz. 2016. Traductor estadístico wixarika-español usando descomposición morfológica. *Comtel*, pages 63–68.

C. Monson, Ariadna Font Llitjós, Roberto Aranovich, Lori S. Levin, R. Brown, E. Peterson, Jaime G. Carbonell, and A. Lavie. 2006. Building nlp systems for two resource-scarce indigenous languages : Mapudungun and quechua.

Garrett Nicolai, Edith Coates, Ming Zhang, and Miikka Silfverberg. 2021. Expanding the JHU Bible Corpus for Machine Translation of the Indigenous Languages of North America. 1:1–5.

John E Ortega, Richard Castro Mamani, and Kyunghyun Cho. 2020. Neural machine translation with a polysynthetic low resource language. *Machine Translation*, 34(4):325–346.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Shantipriya Parida, Ondřej Bojar, and Satya Ranjan Dash. 2019. Odiencorp: Odia–english and odia-only corpus for machine translation. In *Smart Intelligent Computing and Applications: Proceedings of the Third International Conference on Smart Computing and Informatics, Volume 1*, volume 159, page 495. Springer Nature.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Martin Popel and Ondřej Bojar. 2018. Training tips for the transformer model. *The Prague Bulletin of Mathematical Linguistics*, 110(1):43–70.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Prokopis Prokopidis, Vassilis Papavassiliou, and Stelios Piperidis. 2016. Parallel Global Voices: a collection of multilingual corpora with citizen media stories. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 900–905, Portorož, Slovenia. European Language Resources Association (ELRA).

Felix Stahlberg. 2020. Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69:343–418.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Shuoheng Yang, Yuxin Wang, and Xiaowen Chu. 2020. A Survey of Deep Learning Techniques for Neural Machine Translation. *arXiv e-prints*, page arXiv:2002.07526.