Few-shot Spanish-Aymara Machine Translation Using English-Aymara Lexicon

Nat Gillin
nat.gillin@gmail.com

Brian Gummibaerhausen
brian.gbh@gmail.com

Abstract

This paper presents the experiments to train a Spanish-Aymara machine translation model for the AmericasNLP 2023 Machine Translation shared task. We included the English-Aymara GlobalVoices corpus and an English-Aymara lexicon to train the model and limit our training resources to train the model in a *few-shot* manner.

1 Introduction

Aymara is a language spoken in Bolivia, Peru and Chile. It is one of the larger languages in the Americas, and has more than 2 million speakers[1], yet it has received worryingly little attention from NLP researchers. The development of language technologies encourage potential work in the documentation, promotion, preservation and revitalization of the languages (Galla, 2016; Mager et al., 2018). Recent initiatives to promote research on languages of the Americas brings NLP researchers closer to the Americas languages communities and activists (Fernández et al., 2013; Coler and Homola, 2014; Hois and Ruiz, 2018; Kann et al., 2018; Zhang et al., 2020; Ortega et al., 2020). Particularly, machine translation is a useful tool that encourages more research in the languages as it bridges the communication gaps in NLP researchers' understanding of the models' capabilities and limitations.

The AmericasNLP 2021 workshop hosted the Open Machine Translation (OMT) shared task focusing on indigenous and endangered Americas languages (Mager et al., 2021). The organizers provided a seed collection of publicly available corpora and highlighted the various nuances and variability of the translations due to the geographical and linguistic diversity between the language varieties. The Spanish data for development and test sets created in the AmericasNLP 2021 shared task are translated into the Aymara La Paz jilata variant, which is the same variant used in the Global Voices corpus (Tiedemann, 2012; Prokopidis et al., 2016). While Aymara is mutually intelligible across different dialects, they might differ in specific terminologies and minor grammatical preferences.

This paper presents our submission to the AmericasNLP 2023 machine translation shared task (Ebrahimi et al., 2023). We submitted our system that focuses only on translating from Spanish into Aymara. We fine-tuned a multilingual T5 model (Xue et al., 2021) by adding an Aymara-English lexicon[2] to the existing Spanish-Aymara and English-Aymara Global Voices corpus and the Spanish-Aymara shared task training data (Conneau et al., 2018; Ebrahimi et al., 2022).

Other than presenting the results of our AmericasNLP shared task submission, parts of this paper will also serve as a demonstration of how the model was modified from typical model training using HuggingFace suite of libraries (Wolf et al., 2020; Lhoest et al., 2021; McMillan-Major et al., 2021), this is especially useful for low-resource sequence-to-sequence tasks.

2 Pre-trained Tokenizer and New Languages

While the current state of vogue in using massively multilingual pre-trained models on low-resource languages allows researchers to extend the models' sub-word tokenizers, the models implicitly re-use the tokens from how it was previously pre-trained and simply ignore the new tokens by labelling them as `[UNKNOWN]`. In cases where the character set of the low-resource languages' orthography matches the languages that the models were pre-trained on,

[1] Statistics retrieved from Catalogue of Endangered Languages (2023)

[2] The lexicon is created from the notes of a student learning Aymara as a foreign-language, it is hosted on HuggingFace dataset hub. The original sources of the lexicon attributes to Parker (2008) *Webster Aymara-English thesaurus* and Peace Corps (1967) *Beginning Aymara* book.

Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 168–172
July 14, 2023 ©2023 Association for Computational Linguistics

it is possible that the models repurpose the sub-words to learn new parameter behaviors given sufficient computes and hyperparameter tuning experiments.

```
from transformers import AutoTokenizer
from datasets import load_dataset

lexicon_dataset = load_dataset(
    "alvations/aymara-english", on_bad_lines='skip')

tokenizer = AutoTokenizer.from_pretrained('google/mt5-base')

# Train a new tokenizer using the new dataset
# and the old tokenizer object.
new_tokenizer = tokenizer.train_new_from_iterator(
    lexicon_dataset, vocab_size=50_000)
new_tokens = set(new_tokenizer.vocab).difference(tokenizer.vocab)

# Before: 250100
print('Before:', len(tokenizer))
tokenizer.add_tokens(list(new_tokens))

# After (adding vocab): 251152
tokenizer.add_tokens(
    lexicon_dataset['train']['Aymara'] +
    lexicon_dataset['train']['English'])
print('After (adding vocab):', len(tokenizer))
```

To preserve the learned model parameters, a researcher using the multilingual model can extend its tokenizer's sub-word vocabulary by relearning the sub-word tokenizer from scratch, then apply it to dataset with the new language and finally extending the new sub-words to the pre-trained vocabulary. To assign new parameters in the model for these new sub-words tokens, the embedding layer of the model needs to be extended. The code snippet above demonstrates the function to extend the new language's vocabulary to existing pre-trained mT5 model.

The following snippet below presents the differences of the input token indices depending on how the tokenizer was extended for a new language.

```
from transformers import AutoTokenizer

tokenizer_old = AutoTokenizer.from_pretrained('google/mt5-base')
tokenizer_new = AutoTokenizer.from_pretrained('alvations/mt5-aym-lex')

sent = "1899n ahuicha yuriwayi"

tokenized_old_ids = tokenizer_old(sent)['input_ids']
tokenized_new_ids = tokenizer_new(sent)['input_ids']

tokens_old = [tokenizer.decode([s]) for s in tokenized_old_ids]
tokens_new = [tokenizer.decode([s]) for s in tokenized_new_ids]

print(tokens_old)
# Ouputs: ['1899', 'n', '', 'ahu', 'icha', 'yuri', 'way', 'i', '</s>']

print(tokens_new)
# Outputs: ['1899', 'n', '', 'ahuicha', 'yuri', 'way', 'i', '</s>']
```

Instead of using the subword tokenizer, users can pre-tokenize the new language data using a linguistic motivated rule-based tokenizer and add the tokens without further splitting these tokens into subwords to the models' vocabulary. However the tokenizer does not automatically recognize/determine spelling variants, e.g. "*ahuicha*" (i.e. "*grandma*" in English and "*abuela*" in Spanish) can also be spelled as "*awichajax*" in Aymara.

3 Experimental Setup

All models fine-tuned in this paper uses the mT5 architecture using A100 GPUs with 40GB RAM. We use the all default hyperparameters of the HuggingFace's `Seq2SeqTrainingArguments` except:

- `warmup_steps`[3] was set to 500, instead of the default 0
- `auto_find_batch_size` is enabled with the default algorithm to determine batch size automatically
- `max_steps` is set at 200,000. We cap the maximum number of model updates to 200K to limit the computing resources used for our experiments to approximately 24 hours per model, vis-a-vis '*few-shot*' training.

We fine-tuned a zero-shot lexicon-enriched system mT5 model with Aymara-English lexicon, the Spanish-Aymara and English-Aymara Global Voices corpus and Spanish-Aymara XNLI training data split for the training data. And we use the Spanish-Aymara XNLI development data split provided by the shared task organizers to select the best performing model.

Training Data	`mt5-base`	`mt5-zero`	`mt5-lex`
XNLI Train (spa-aym)	✓	✓	✓
Global Voices (spa-aym)	✓	✓	✓
Global Voices (eng-aym)		✓	✓
Lexicon (eng-aym)			✓

Table 1: Training Datasets used by the mT5 Variants

Our official submission to the shared task is selected from the best-performing system that scored the lowest perplexity loss and highest BLEU score. Other than the best performing zero-shot lexicon-enriched system (`mT5-lex`), we experimented and a baseline model that only fine-tuned Spanish-Aymara Global Voice and XNLI

[3]This hyperparameter is used to gradually increased the learning rate to make training more stable (Huang et al., 2020). The original transformer (Vaswani et al., 2017) set the warmup to 4,000.

dataset (mT5-base) and a second baseline that adds on the English-Aymara Global Voices data to Spanish-Aymara Global Voices and XNLI dataset (mT5-zero). Table 1 summarizes the datasets used to train the corresponding mT5 models.

4 Results

Our official submission to the shared-task scored a measly 0.12 BLEU (Papineni et al., 2002) and 9.22 ChrF score (Popović, 2015) on the AmericasNLP 2023 shared task test set. The best performing team in the shared task achieved 4.45 BLEU and 36.24 ChrF. The target Aymara text from the test set was not released publicly, hence we present the results of our model variants on the development set.

System	ChrF	BLEU
mt5-base	**30.59**	2.78
mt5-zero	23.98	**2.99**
mt5-lex	22.01	1.38

Table 2: Results on AmericasNLP 2023 Spanish-Aymara Development Set

We note the oracle effect of selecting the best model during training based on the development set, thus the results from Table 2 might be inflated.

As a sanity check, we translated the lexicon used to train mt5-lex from English into Spanish using the NLLB machine translation model (Costa-jussà et al., 2022) and count the tokens from the lexicon that matches the development texts. We found that the lexicon has little matches to the tokens in the development sets, see Appendix A for more details.

5 Conclusion

In this paper we present our participation in the AmericasNLP 2023 Spanish-Aymara machine translation shared task. We experimented with adding an English-Aymara lexicon and training

We share the follow resources created in our participation for future researchers to improve English/Spanish-Aymara translations.

- English-Aymara Lexicon
- mt-base model
- mt-zero model
- mt-lex model
- Model training script

References

Catalogue of Endangered Languages. 2023. University of Hawaii at Manoa.

Matthew Coler and Petr Homola. 2014. Rule-based machine translation for aymara. *Endangered Languages and New Technologies*, pages 67–80.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Enora Rice, Cynthia Montaño, John Ortega, Shruti Rijhwani, Alexis Palmer, Rolando Coto-Solano, Hilaria Cruz, and Katharina Kann. 2023. Findings of the AmericasNLP 2023 shared task on machine translation into indigenous languages. In *Proceedings of the Third Workshop on Natural Language Processing for Indigenous Languages of the Americas*. Association for Computational Linguistics.

Dayana Iguarán Fernández, Ornela Quintero Gamboa, Jose Molina Atencia, and Oscar Elías Bedoya. 2013. Design and implementation of an "web api" for the automatic translation colombia's language pairs: Spanish-wayuunaiki case. In *2013 IEEE Colombian Conference on Communications and Computing (COLCOM)*, pages 1–9. IEEE.

Candace Kaleimamoowahinekapu Galla. 2016. Indigenous language revitalization, promotion, and education: Function of digital technology. *Computer Assisted Language Learning*, 29(7):1137–1151.

Jesús Manuel Mager Hois and Ivan Vladimir Meza Ruiz. 2018. Hacia la traducción automática de las lenguas indígenas de méxico. *Digital Humanities 2018: Book of Abstracts/Libro de resúmenes.*

Xiao Shi Huang, Felipe Perez, Jimmy Ba, and Maksims Volkovs. 2020. Improving transformer optimization through better initialization. In *International Conference on Machine Learning*, pages 4475–4483. PMLR.

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.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2021. Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217, Online. Association for Computational Linguistics.

Angelina McMillan-Major, Salomey Osei, Juan Diego Rodriguez, Pawan Sasanka Ammanamanchi, Sebastian Gehrmann, and Yacine Jernite. 2021. Reusable templates and guides for documenting datasets and models for natural language processing and generation: A case study of the HuggingFace and GEM data and model cards. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 121–135, Online. Association for Computational Linguistics.

John E Ortega, Richard Alexander Castro-Mamani, and Jaime Rafael Montoya Samame. 2020. Overcoming resistance: The normalization of an Amazonian tribal language. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 1–13, Suzhou, China. Association for Computational Linguistics.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Philip M. Parker. 2008. *Webster's Aymara - English Thesaurus Dictionary*. ICON Group International.

Peace Corps. 1967. *Beginning Aymara: A course for English speakers*. Peace Corps.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

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).

Liling Tan, Josef van Genabith, and Francis Bond. 2015. Passive and pervasive use of bilingual dictionary in statistical machine translation. In *Proceedings of the Fourth Workshop on Hybrid Approaches to Translation (HyTra)*, pages 30–34, Beijing. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame,

Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

François Yvon and Sadaf Abdul Rauf. 2020. *Using lexical and terminological resources in neural machine translation*. Ph.D. thesis, LIMSI-CNRS.

Shiyue Zhang, Benjamin Frey, and Mohit Bansal. 2020. ChrEn: Cherokee-English machine translation for endangered language revitalization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 577–595, Online. Association for Computational Linguistics.

A Lexicon Matches in Development Set

There are 81 unique words that matches the Spanish translated lexicon to the tokens in the development set. The matches sum up to a frequency of 373 out of a total number of 53,135 in the development set on the Spanish source. However when we match the target Aymara text with the lexicon and we find only 4 unique words matches that occurred 9 times in the development set. Looking at the sentences that contains the Aymara word matches to the lexicon, the Aymara sentences from the development set contains loan words either from Spanish or English,

The 4 unique Spanish - Aymara lexicon matches are:

- *el vuelo -> fly*
- *mayo -> may*
- *firme -> firm*
- *hijo -> son*

The sentences that contains the target side matches are:

- *The firm Uk* ullartatï.
- Tamax may maya temanakanw yatiñ munapxchixa.
- Aka jan walt'awix may may lup'iypachatamxa, ukampis samart'awim suyt'am.
- Jichhurux awkixan nayra jakawipat arst'awaya ukatx kunawsatix Estados Unidos markar sarawayjix may may kast sarawinak utjirinakaw uñicht'ayätani
- *I'll fly away* uk ajlliristxa.
- Aruskipt'aw Hilbert, *Las mariposas son libres, El mago de Oz, Tierra de juguetes* y Vuelos ukanakatx purt'anirinakax uñjtawayapxaniwa.
- Ukampirus, niyapunix may uñjiristwa, uh, V6 inas.

We note that the underlined loan phrases matches contributes to the matching counts in the lexicon. And when it comes to the Aymara lexicon entry '*may*', it is a false-friend match, in both development setences that contains '*may may*', it phrase seems to be a grammatical/syntactic construct.

With the above anecdote, we find that lexicon effects in machine translation might not be evident in metrics scores if the lexicon matches in the test set is low, unlike previous studies of using lexicon in high resource languages (Tan et al., 2015; Yvon and Rauf, 2020).