The REPUcs' Spanish–Quechua Submission to the AmericasNLP 2021 Shared Task on Open Machine Translation

Oscar Moreno Veliz
Pontificia Universidad Católica del Perú
omoreno@pucp.edu.pe

Abstract

We present the submission of REPUcs[1] to the AmericasNLP machine translation shared task for the low resource language pair Spanish–Quechua. Our neural machine translation system ranked first in Track two (development set not used for training) and third in Track one (training includes development data). Our contribution is focused on: (i) the collection of new parallel data from different web sources (poems, lyrics, lexicons, handbooks), and (ii) using large Spanish–English data for pre-training and then fine-tuning the Spanish–Quechua system. This paper describes the new parallel corpora and our approach in detail.

1 Introduction

REPUcs participated in the AmericasNLP 2021 machine translation shared task (Mager et al., 2021) for the Spanish–Quechua language pair. Quechua is one of the most spoken languages in South America (Simons and Fenning, 2019), with several variants, and for this competition, the target language is Southern Quechua. A disadvantage of working with indigenous languages is that there are few documents per language from which to extract parallel or even monolingual corpora. Additionally, most of these languages are traditionally oral, which is the case of Quechua. In order to compensate the lack of data we first obtain a collection of new parallel corpora to augment the available data for the shared task. In addition, we propose to use transfer learning (Zoph et al., 2016) using large Spanish–English data in a neural machine translation (NMT) model. To boost the performance of our transfer learning approach, we follow the work of Kocmi and Bojar (2018), which demonstrated that sharing the source language and a vocabulary of subword units can improve the performance of low resource languages.

2 Spanish→Quechua

Quechua is the most widespread language family in South America, with more than 6 millions speakers and several variants. For the AmericasNLP Shared Task, the development and test sets were prepared using the Standard Southern Quechua writing system, which is based on the Quechua Ayacucho (quy) variant (for simplification, we will refer to it as Quechua for the rest of the paper). This is an official language in Peru, and according to Zariquiey et al. (2019) it is labelled as endangered. Quechua is essentially a spoken language so there is a lack of written materials. Moreover, it is a polysynthetic language, meaning that it usually express large amount of information using several morphemes in a single word. Hence, subword segmentation methods will have to minimise the problem of addressing "rare words" for an NMT system.

To the best of our knowledge, Ortega et al. (2020b) is one of the few studies that employed a sequence-to-sequence NMT model for Southern Quechua, and they focused on transfer learning with Finnish, an agglutinative language similar to Quechua. Likewise, Huarcaya Taquiri (2020) used the Jehovah Witnesses dataset (Agić and Vulić, 2019), together with additional lexicon data, to train an NMT model that reached up to 39 BLEU points on Quechua. However, the results in both cases were high because the development and test set are split from the same distribution (domain) as the training set. On the other hand, Ortega and Pillaipakkamnatt (2018) improved alignments for Quechua by using Finnish(an agglutinative language) as the pivot language. The corpus source is the parallel treebank of Rios et al. (Rios et al., 2012)., so we deduce that they worked with Quechua Cuzco (quz). (Ortega et al., 2020a)

In the AmericasNLP shared task, new out-of-

[1] "Research Experience for Peruvian Undergraduates - Computer Science" is a program that connects Peruvian students with researchers worldwide. The author was part of the 2021 cohort: https://www.repuprogram.org/repu-cs.

Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas, pages 241–247
June 11, 2021. ©2021 Association for Computational Linguistics

domain evaluation sets were released, and there were two tracks: using or not the validation set for training the final submission. We addressed both tracks by collecting more data and pre-training the NMT model with large Spanish-English data.

3 Data and pre-processing

In this competition we are going to use the AmericasNLP Shared Task datasets and new corpora extracted from documents and websites in Quechua.

3.1 AmericasNLP datasets

For training, the available parallel data comes from dictionaries and Jehovah Witnesses dataset (JW300; Agić and Vulić, 2019). AmericasNLP also released parallel corpus aligned with English (en) and the close variant of Quechua Cusco (quz) to enhance multilingual learning. For validation, there is a development set made with 994 sentences from Spanish and Quechua (quy) (Ebrahimi et al., 2021).

Detailed information from all the available datasets with their corresponding languages is as follows:

- JW300 (quy, quz, en): texts from the religious domain available in OPUS (Tiedemann, 2012). JW300 has 121k sentences. The problems with this dataset are misaligned sentences, misspelled words and blank translations.
- MINEDU (quy): Sentences extracted from the official dictionary of the Ministry of Education in Peru (MINEDU). This dataset contains open-domain short sentences. A considerable number of sentences are related to the countryside. It only has 650 sentences.
- Dict_misc (quy): Dictionary entries and samples collected and reviewed by Huarcaya Taquiri (2020). This dataset is made from 9k sentences, phrases and word translations.

Furthermore, to examine the domain resemblance, it is important to analyse the similarity between the training and development. Table 1 shows the percentage of the development set tokens that overlap with the tokens in the training datasets on Spanish (es) and Quechua (quy) after deleting all types of symbols.

Dataset	% Dev overlapping	
	es	quy
JW300	85%	45%
MINEDU	15%	5%
Dict_misc	40%	18%

Table 1: Word overlapping ratio between the development and the available training sets in AmericasNLP

We observe from Table 1 that the domain of the training and development set are different as the overlapping in Quechua does not even go above 50%. There are two approaches to address this problem: to add part of the development set into the training or to obtain additional data from the same or a more similar domain. In this paper, we focus on the second approach.

3.2 New parallel corpora

Sources of Quechua documents Even though Quechua is an official language in Peru, official government websites are not translated to Quechua or any other indigenous language, so it is not possible to perform web scrapping (Bustamante et al., 2020). However, the Peruvian Government has published handbooks and lexicons for Quechua Ayacucho and Quechua Cusco, plus other educational resources to support language learning in indigenous communities. In addition, there are official documents such as the Political Constitution of Peru and the Regulation of the Amazon Parliament that are translated to the Quechua Cusco variant.

We have found three unofficial sources to extract parallel corpora from Quechua Ayacucho (quy). The first one is a website, made by Maximiliano Duran (Duran, 2010), that encourages the learning of Quechua Ayacucho. The site contains poems, stories, riddles, songs, phrases and a vocabulary for Quechua. The second one is a website for different lyrics of poems and songs which have available translations for both variants of Quechua (Lyrics translate, 2008). The third source is a Quechua handbook for the Quechua Ayacucho variant elaborated by Iter and Cárdenas (2019).

Sources that were extracted but not used due to time constrains were the Political Constitution of Peru and the Regulation of the Amazon Parliament. Other non-extracted source is a dictionary for Quechua Ayacucho from a website called InkaTour [2]. This source was not used because we already had a dictionary.

Methodology for corpus creation The available vocabulary in Duran (2010) was extracted manually and transformed into parallel corpora using the first

[2] https://www.inkatour.com/dico/

pair of parenthesis as separators. We will call this dataset "Lexicon".

All the additional sentences in Duran (2010) and a few poems from (Lyrics translate, 2008) were manually aligned to obtain the Web Miscellaneous (WebMisc) corpus. Likewise, translations from the Quechua educational handbook (Iter and Cárdenas, 2019) were manually aligned to obtain a parallel corpus (Handbook).[3]

In the case of the official documents for Quechua Cusco, there was a specific format were the Spanish text was followed by the Quechua translation. After manually arranging the line breaks to separate each translation pair, we automatically constructed a parallel corpus for both documents. Paragraphs with more than 2 sentences that had the same number of sentences as their translation were split into small sentences and the unmatched paragraphs were deleted.

Corpora description We perform a large number or rare events (LNRE) modelling to analyse the WebMisc, Lexicon and Handbook datasets[4]. The values are shown in Table 2. The LNRE modelling for the Quechua Cusco datasets are shown in appendix as they are not used for the final submission.

	WebMisc		Lexicon		Handbook	
	es	quy	es	quy	es	quy
S	985		6161		2297	
N	5002	2996	7050	6288	15537	8522
V	1929	2089	3962	3361	4137	5604
V1	1358	1673	2460	1838	2576	4645
V/N	0.38	0.69	0.56	0.53	0.26	0.65
V1/N	0.27	0.55	0.34	0.29	0.16	0.54
mean	2.59	1.43	1.77	1.87	3.75	1.52

Table 2: Corpora description: S = #sentences in corpus; N = number of tokens; V = vocabulary size; V1 = number of tokens occurring once (hapax); V/N = vocabulary growth rate; V1/N = hapax growth rate; mean = word frequency mean

We notice that the vocabulary and hapax growth rate is similar for Quechua (quy) in WebMisc and Handbook even though the latter has more than twice the number of sentences. In addition, it was expected that the word frequency mean and the vocabulary size were lower for Quechua, as this demonstrates its agglutinative property. However, this does not happens in the Lexicon dataset, since is understandable as it is a dictionary that has one or two words for the translation.

Moreover, there is a high presence of tokens occurring only once in both languages. In other words, there is a possibility that our datasets have spelling errors or presence of foreign words (Nagata et al., 2018). However, in this case this could be more related to the vast vocabulary, as the datasets are made of sentences from different domains (poems, songs, teaching, among others).

Furthermore, it is important to examine the similarities between the new datasets and the development set. The percentage of the development set words that overlap with the words of the new datasets on Spanish (es) and Quechua (quy) after eliminating all symbols is shown in Table 3.

Dataset	% Dev overlapping	
	es	quy
WebMisc	18.6%	4%
Lexicon	20%	3.4%
Handbook	28%	10.6%

Table 3: Percentage of word overlapping between the development and the new extracted datasets

Although at first glance the analysis may show that there is not a significant similarity with the development set, we have to take into account that in Table 1, JW300 has 121k sentences and Dict_misc is a dictionary, so it is easy to overlap some of the development set words at least once.However , in the case of WebMisc and Handbook datasets, the quantity of sentences are less than 3k per dataset and even so the percentage of overlapping in Spanish is quite good. This result goes according to the contents of the datasets, as they contain common phrases and open domain sentences, which are the type of sentences that the development set has.

3.3 English-Spanish dataset

For pre-training, we used the EuroParl dataset for Spanish–English (1.9M sentences) (Koehn, 2005) and its development corpora for evaluation.

4 Approach used

4.1 Evaluation

From the Europarl dataset, we extracted 3,000 sentences for validation. For testing we used the devel-

[3] All documents are published in: https://github.com/Ceviche98/REPUcs-AmericasNLP2021

[4] We used the LNRE calculator created by Kyle Gorman: https://gist.github.com/kylebgorman/

opment set from the WMT2006 campaign (Koehn and Monz, 2006).

In the case of Quechua, as the official development set contains only 1,000 sentences there was no split for the testing. Hence, validation results will be taken into account as testing ones.

The main metric in this competition is chrF (Popović, 2017) which evaluates character n-grams and is a useful metric for agglutinative languages such as Quechua. We also reported the BLEU scores (Papineni et al., 2002). We used the implementations of sacreBLEU (Post, 2018).

4.2 Subword segmentation

Subword segmentation is a crucial process for the translation of polysinthetic languages such as Quechua. We used the Byte-Pair-Encoding (BPE; Sennrich et al., 2016) implementation in SentencePiece (Kudo and Richardson, 2018) with a vocabulary size of 32,000. To generate a richer vocabulary, we trained a segmentation model with all three languages (Spanish, English and Quechua), where we upsampled the Quechua data to reach a uniform distribution.

4.3 Procedure

For all experiments, we used a Transformer-based model (Vaswani et al., 2017) with default parameters from the Fairseq toolkit (Ott et al., 2019). The criteria for early stopping was cross-entropy loss for 15 steps.

We first pre-trained a Spanish–English model on the Europarl dataset in order to obtain a good encoding capability on the Spanish side. Using this pre-trained model, we implemented two different versions for fine-tunning. First, with the JW300 dataset, which was the largest Spanish–Quechua corpus, and the second one with all the available datasets (including the ones that we obtained) for Quechua.

5 Results and discussion

The results from the transfer learning models and the baseline are shown in Table 4. We observe that the best result on BLEU and chrF was obtained using the provided datasets together with the extracted datasets. This shows that the new corpora were helpful to improve translation performance.

From Table 4, we observe that using transfer learning showed a considerable improvement in comparison with the baseline (+0.56 in BLEU and +0.007 in chrF). Moreover, using transfer learning with all the available datasets obtained the best BLEU and chrF score. Specially, it had a 0.012 increase in chrF which is quite important as chrF is the metric that best evaluates translation in this case. Overall, the results do not seem to be good in terms of BLEU. However, a manual analysis of the sentences shows that the model is learning to translate a considerable amount of affixes.

Dataset	Size	Direction	BLEU	chrF
Europarl	1.9M	es→en	34.2	0.606
JW300 (baseline)	121k	es→quy	1.49	0.317
JW300 (fine-tuning)	121k	es→quy	2.05	0.324
All datasets (fine-tuning)	133k	es→quy	**2.18**	**0.336**

Table 4: Results of transfer learning experiments

Input (ES)	*El control de armas probablemente no es popular en Texas.*
Input (EN)	*Weapon control is probably not popular in Texas.*
Reference (QUY)	***Texas****piqa sutillapas* ***arma*** *controlayqa* ***mana****chusmi hin****achu*** *apa****kun***
Output	***Texas*** *llaqtapi* ***arma****kuna controlayqa manam runa****kun****apa run****achu***

Table 5: Subword analysis on translated and reference sentence

For instance, the subwords "arma", "mana", among others, have been correctly translated but are not grouped in the same words as in the reference. In addition, only the word "controlayqa" is translated correctly, which would explain the low results in BLEU. Decoding an agglutinative language is a very difficult task, and the low BLEU scores cannot suggest a translation with proper adequacy and/or fluency (as we can also observe this from the example). Nevertheless, BLEU works at word-level so other character-level metrics should be considered to inspect agglutinative languages. This would be the case of chrF (Popović, 2017) were there is an increase of around 3% when using the AmericasNLP altogether with the new extracted corpora.

Translations using the transfer learning model trained with all available Quechua datasets were submitted for track 2 (Development set not used for Training). For the submission of track 1 (Development set used for Training) we retrained the best transfer learning model adding the validation to the training for 40 epochs. The official results of the competition are shown in Table 6.

	Rank	Team	BLEU	chrF
Track 1	1	Helsinki	5.38	0.394
	3	**REPUcs**	3.1	**0.358**
Track 2	1	**REPUcs**	2.91	**0.346**
	2	Helsinki	3.63	0.343

Table 6: Official results from AmericasNLP 2021 Shared Task competition on the two tracks.Track 1: Development set used for Training, Track 2: Development set not used for Training

6 Conclusion

In this paper, we focused on extracting new datasets for Spanish–Quechua, which helped to improve the performance of our model. Moreover, we found that using transfer learning was beneficial to the results even without the additional data. By combining the new corpora in the fine-tuning step, we managed to obtain the first place on Track 2 and the third place on Track 1 of the AmericasNLP Shared Task. Due to time constrains, the Quechua Cusco data was not used, but it can be beneficial for further work.

In general, we found that the translating Quechua is a challenging task for two reasons. Firstly, there is a lack of data for all the variants of Quechua, and the available documents are hard to extract. In this research, all the new datasets were extracted and aligned mostly manually. Secondly, the agglutinative nature of Quechua motivates more research about effective subword segmentation methods.

Acknowledgements

This work could not be possible without the support of REPU Computer Science (Research Experience for Peruvian Undergraduates), a program that connects Peruvian students with researchers across the world. The author is thankful to the REPU's directors and members, and in particular, to Fernando Alva-Manchego and David Freidenson, who were part of the early discussions for the participation in the Shared Task. Furthermore, the author is grateful to the insightful feedback of Arturo Oncevay, Barry Haddow and Alexandra Birch, from the University of Edinburgh, where the author worked as an intern as part of the REPU's 2021 cohort.

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.

Gina Bustamante, Arturo Oncevay, and Roberto Zariquiey. 2020. No data to crawl? monolingual corpus creation from PDF files of truly low-resource languages in Peru. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2914–2923, Marseille, France. European Language Resources Association.

Maximiliano Duran. 2010. Lengua general de los Incas. http://quechua-ayacucho.org/es/index_es.php. Accessed: 2021-03-15.

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.

Diego Huarcaya Taquiri. 2020. Traducción automática neuronal para lengua nativa peruana. Bachelor's thesis, Universidad Peruana Unión.

Cesar Iter and Zenobio Ortiz Cárdenas. 2019. *Runasimita yachasun*.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252, Belgium, Brussels. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Philipp Koehn and Christof Monz. 2006. Manual and automatic evaluation of machine translation between european languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Lyrics translate. 2008. Lyrics translate. https://lyricstranslate.com/. Accessed: 2021-03-15.

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.

Ryo Nagata, Taisei Sato, and Hiroya Takamura. 2018. Exploring the Influence of Spelling Errors on Lexical Variation Measures. *Proceedings of the 27th International Conference on Computational Linguistics*, (2012):2391–2398.

John Ortega, Richard Alexander Castro-Mamani, and Jaime Rafael Montoya Samame. 2020a. 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.

John Ortega and Krishnan Pillaipakkamnatt. 2018. Using morphemes from agglutinative languages like Quechua and Finnish to aid in low-resource translation. In *Proceedings of the AMTA 2018 Workshop on Technologies for MT of Low Resource Languages (LoResMT 2018)*, pages 1–11.

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

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. 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.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Annette Rios, Anne Göhring, and Martin Volk. 2012. Parallel Treebanking Spanish-Quechua: how and how well do they align? *Linguistic Issues in Language Technology*, 7(1).

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.

Gary F. Simons and Charles D. Fenning, editors. 2019. *Ethnologue: Languages of the World. Twenty-second edition*. Dallas Texas: SIL international. Online version: http://www.ethnologue.com.

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

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 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Roberto Zariquiey, Harald Hammarström, Mónica Arakaki, Arturo Oncevay, John Miller, Aracelli García, and Adriano Ingunza. 2019. Obsolescencia lingüística, descripción gramatical y documentación de lenguas en el Perú: hacia un estado de la cuestión. *Lexis*, 43(2):271–337.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 1568–1575.

A Appendix

	Constitution		Regulation	
	es	quz	es	quz
S	999		287	
N	14295	9837	14295	3227
V	3404	4030	3404	1591
V1	2145	3037	2145	1248
V/N	0.2381	0.4097	0.2381	0.493
V1/N	0.1501	0.3087	0.1501	0.3867
mean	4.1995	2.4409	4.1995	2.083

Table 7: Description of the corpora extracted, but not used, for Quechua Cusco (quz). S = #sentences in corpus; N = number of tokens; V = vocabulary size; V1 = number of tokens occurring once (hapax); V/N = vocabulary growth rate; V1/N = hapax growth rate; mean = word frequency mean