Enhancing Spanish-Quechua Machine Translation with Pre-Trained Models and Diverse Data Sources: LCT-EHU at AmericasNLP Shared Task

Nouman Ahmed, Natalia Flechas Manrique, and **Antonije Petrović**
University of the Basque Country (UPV/EHU)
{anouman001, nflechas001, apetrovic001}@ikasle.ehu.eus

Abstract

We present the LCT-EHU submission to the AmericasNLP 2023 low-resource machine translation shared task. We focus on the Spanish-Quechua language pair and explore the usage of different approaches: (1) Obtain new parallel corpora from the literature and legal domains, (2) Compare a high-resource Spanish-English pre-trained MT model with a Spanish-Finnish pre-trained model (with Finnish being chosen as a target language due to its morphological similarity to Quechua), and (3) Explore additional techniques such as copied corpus and back-translation. Overall, we show that the Spanish-Finnish pre-trained model outperforms other setups, while low-quality synthetic data reduces the performance.

1 Introduction

The LCT-EHU team participated in the AmericasNLP 2023 low-resource machine translation shared task. The task involved machine translation from Spanish to 11 different indigenous languages. The languages in question are very much low-resource, with the number of speakers spanning from a few tens of thousands to a few million and with limited availability of parallel data. Monolingual data is not easily obtained either - Wikipedia is available only in a few of these languages, with the number of articles not being very high. Our team focused on Spanish-Quechua language pair with the approach consisting in:

- Finding and aligning new parallel data. We obtained bilingual legal documents of the Government of Ecuador (the constitution and some laws); the novel "The Little Prince", and the UN Declaration of Human Rights.
- Using pre-trained machine translation models trained on other language pairs. We experimented with Spanish-English, as a high-resource language pair, and Spanish-Finnish, with the linguistic intuition that using an agglutinative language on the target side would provide a closer set-up to the problem we were working on, as previously explored by Ortega and Pillaipakkamnatt (2018) and Ortega et al. (2020).
- Synthetic and monolingual data. We experimented with a copied corpus approach and synthetic parallel corpus creation from monolingual Spanish data.

The official metric used in the shared task is chrF++ (Popović, 2017). In the previous edition of the AmericasNLP shared task, the chrF score of 34.6 was obtained by the REPUcs team (Moreno, 2021) for the Spanish-Quechua language pair. However, this year's shared task takes the second-best result of 34.3 as a baseline.

All of the source code and newly collected data are available in the Github repository [1].

2 Related Work

Some previous work and approaches that were important for our experiments are explained in the following sub-sections.

2.1 AmericasNLP 2021 Shared Task

In the first edition of the AmericasNLP low-resource MT shared task, various contributions to the field of machine translation of American indigenous languages were published. The organizers provided training data collected from various sources, alongside manually translated development and test data. Two tracks were available: (1) development set used for training, and (2) development set not used for training.

Helsinki team (Vázquez et al., 2021) won the task in the majority of language pairs in both tracks,

[1] https://github.com/nouman-10/MT-SharedTask

Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 156–162
July 14, 2023 ©2023 Association for Computational Linguistics

using a two-phase transformer training. They also obtained additional parallel and monolingual data for Spanish-Quechua. Their Model A was a multilingual model with 11 languages, trained for 200 000 steps, which was then trained independently for each of the target indigenous languages for additional 2 500 steps. Model B was a multilingual model with Spanish as the only source language, and with 11 target languages (10 indigenous languages + English). The two-phase training was performed again. In the first phase, they trained the model with 90% of Spanish-English data, while the remaining 10% was divided between 10 indigenous languages, each taking 1% . In the second phase, the proportion of Spanish-English data is reduced to 50%, while including backtranslated data as well. Different versions of both Model A and Model B were trained, depending on whether the development data was used during training or not.

2.2 Synthetic translations and copied corpus

The use of synthetic translation approaches is born out of a common concern in machine translation: the lack of high-quality parallel data for many language pairs. To solve this, various solutions have been proposed. One of the most common ones is known as back-translation (Sennrich et al., 2016), which involves creating a synthetic parallel corpus by translating monolingual data from the target language into the source language (or source to target, in other approaches) and using this to augment the existing parallel data for training models. Another approach (Currey et al., 2017) involves using monolingual data from the target and aligning it with itself, to mimic parallel data (this is known as *copied corpus*). The authors try to explain the success of this approach by stating that there might be an improved accuracy on named entities and words that are identical in both source and target texts.

3 Data

In this section, we will describe the data used in the experiments.

3.1 Original parallel data

The following corpora were provided by the organizers of the competition (Agić and Vulić (2019) and Tiedemann (2012):

- **JW300 (quz & quy)** A collection of Jehovah's Witnesses Texts, both in Cuzco and Ayacucho Quechua.
- **MINEDU (quy)**: Sentences extracted from the official dictionary of the Ministry of Education (MINEDU) in Peru for Quechua Ayacucho.
- **Dict_misc (quy)**: Dictionary entries and samples collected by Diego Huarcaya.

The counts of sentences and domain information are presented in Table 1. The column *Count* refers to the number of sentences in this table and all subsequent ones.

Name	Domain	Count
JW300	Religious	121064
MINEDU	Dictionary	643
Dict misc	Dictionary	8998

Table 1: Original data of the AmericasNLP 2023 competition.

3.2 Additional resources

We also used resources that were introduced by some of the teams that participated in the 2021 competition. Details of the data introduced by the Helsinki-NLP team (Vázquez et al., 2021) are presented in Table 2.

Name	Domain	Count
Peruvian Constitution	Legal	1276
Bolivian Constitution	Legal	2193
Tatoeba (OPUS)	Misc.	163
Bible	Religious	31102

Table 2: Data introduced by the Helsinki-NLP 2021 team.

In Table 3 the details of the corpora used by the REPUcs-AmericasNLP2021 (Moreno, 2021) team are shown.

Name	Domain	Count
Web Misc	Misc.	985
Lexicon	Dictionary	6161
Handbook	Educational	2296
Peruvian Constitution	Legal	999
Regulations of the Amazon Parliament	Legal	287

Table 3: Data introduced by the REPUcs-AmericasNLP2021 team.

In addition to the data collected in the previous AmericasNLP task, we found some parallel data that was used to build *A Basic Language Technology Toolkit for Quechua* (Rios, 2016) [2]. The parallel data was used to create a multilingual treebank in the three languages of the machine translation systems, Spanish-German and Spanish-Cuzco Quechua. The majority of the corpus was Spanish-German, with the Quechua counterpart being translated by several native speakers in Peru. There were multiple aligned documents available here but most of them needed further cleaning and alignment. The three documents that were selected are:

- Strategy paper of the Swiss Agency for Development and Cooperation on the cooperation with Peru [3]
- 2009 Annual report of the Deutsche Welle Academy about Development and the Media [4]
- 2008 Annual report of a private foundation dedicated to education [5]

The sentence count of the documents is also shown in Table 4.

Name	Count
Cosude	529
DW	856
Fundeducation	440

Table 4: Additional resources of Cuzco Quechua

3.3 New resources

Apart from using the already existing resources, we have gathered, processed, and aligned publicly available documents found around the web. The summary of these resources is shown in Table 5. It is important to emphasize that, theoretically, Quechua should be regarded as a linguistic family rather than a single language, given that its various varieties exhibit limited mutual intelligibility when they are geographically distant. Within the specialized literature, the term "Quechua" is employed to refer to the varieties spoken in Bolivia and Peru, while the term "Quichua" is preferred for those spoken in Ecuador and Argentina, as indicated by Avellana (Avellana). For the sake of simplicity, when uncertainty arises regarding the specific Quechua variety being discussed, we adopt the `que` code as a macrolanguage identifier.

The documents were found in pdf format and were transformed into plain text using the `pdftotext` [6] tool, trying to keep the layout of the original pdf as intact as possible. Since most of the documents contained word wrapping to keep the fixed width of the document, we performed the unwrapping in such cases by joining the words at the ends of the lines which ended with the - sign. In this step, we made an effort to preserve the original document structure whenever feasible. For instance, with "The Little Prince," we maintained the chapter arrangement of the novel. Similarly, when dealing with the Ecuadorian constitution and laws [7], we retained the individual article divisions.

In the subsequent stage, we performed sentence segmentation at the chapter level while preserving the chapter boundaries. Our team experimented with several sentence segmenters such as `NLTK`, `spaCy`, and `stanza`. Following careful consideration, we ultimately chose `stanza` based on a higher alignment score, as explained in the next paragraph. For `stanza`, we opted for the Spanish sentence segmentation model for both Spanish and Quechua texts.

The `HunAlign` (Varga et al., 2007) tool was utilized to align the sentences. Additionally, we used a dictionary provided by AmericasNLP organizers as an input to the tool to improve the alignments. Overall, the legal document alignments were quite accurate, whereas the alignments of "The Little Prince" were slightly less precise. This could be attributed to the greater freedom often allowed in translations of literary works compared to the strict and rigid translations necessary in legal contexts. Even though `HunAlign` gives a confidence score for each alignment, we did not perform any filtering of the aligned sentences and decided to use all obtained alignments.

3.4 Synthetic translations

We collected three history books in Spanish. Specifically, old Chronicles of the Indies about the Incan empire and the subsequent colonial period. We hypothesized that because these books have plenty

[2] https://github.com/a-rios/squoia

[3] https://www.cooperacionsuiza.pe/cosude/

[4] http://www.dw.de/

[5] http://www.fundeducation.org/

[6] https://www.xpdfreader.com/

[7] https://www.asambleanacional.gob.ec/es/contenido/publicaciones

Name	Domain	Quechua variety	Count
The Little Prince	Literature	que	1312
UN Human Rights Declaration	Legal	qus	91
The Constitution of Ecuador	Legal	que	2243
Ley Soberania Alimentaria	Legal	que	174
Ley Consumo Drogas	Legal	que	69
Ley Organica Alimentacion	Legal	que	186

Table 5: Description of the gathered new parallel data

of words in Quechua language, they would be from a suitable domain. The three books were turned into plain text files and their sentences were segmented in the way described in the previous section. After that, the texts were translated into Quechua with the Spanish-Finnish model we fine-tuned on the original datasets and the additional resources introduced by participating teams in the 2021 competition (`train + extra`). Table 6 shows the final sentence counts of these books after being processed.

Name	Domain	Count
Comentarios Reales	History	1032
Nueva Cronica y Buen Gobierno	History	3578
Cronica del Peru	History	1798

Table 6: Chronicles of the Indies description.

3.5 Monolingual (Copied Corpus)

Following the approach in (Currey et al., 2017), we decided to add some monolingual Quechua data and copy it as is to create a parallel corpus. We used publicly available datasets on Huggingface and segmented the sentences based on line breaks, without any post-processing. The datasets included data cc100 (Conneau et al. (2020) and Wenzek et al. (2020) which was an attempt to recreate the dataset used for training XLM-R, and data from (Zevallos et al., 2022), which is a monolingual corpus of Southern Quechua and includes the Wiki and OSCAR corpora. Table 7 shows the sentence counts of these datasets

Name	Count
cc100	113931
Llamacha	182669

Table 7: Description of monolingual data used for Copied Corpus Approach

4 Models & Results

We experimented with 2 major model setups and 5 different kinds of dataset combinations. The two setups were based on fine-tuned machine translation models of Spanish-English and Spanish-Finnish (Tiedemann and Thottingal, 2020). On the one hand, the reason behind using a fine-tuned Spanish-English model was that both of them are high-resource languages, and thus the model has been trained on large amounts of data. This probably means that the model has learned a good Spanish encoder, and thus could be useful for further fine-tuning. On the other hand, the reasoning behind choosing a Spanish-Finnish model and fine-tuning on Spanish-Quechua was the similarity between Finnish and Quechua (specifically the agglutinative morphology of both languages), and Finnish having comparably more data than Quechua. All models were trained for 20 epochs, with evaluation being done after every 1000 steps. The best model was selected based on the chrF score on the development set. Here, we will define the different combinations of datasets used for our experiments:

- `train` : The original parallel-data provided in the AmericasNLP-2023 Shared Task (as mentioned in Table 1).
- `train + extra`: This includes the combination of original parallel data and extra Ayacucho Quechua (Quy) data gathered from different sources.
- `train + extra + aligned`: This includes the data above plus our newly gathered parallel data (as mentioned in Table 5).
- `train + extra + aligned + copied`: In addition to the above data, it also includes the monolingual copied corpus, (Table 7).

Model name	Pre-trained model	Data	Dev		Test		Sub
			chrF	BLEU	chrF	BLEU	
baseline			33.80	3.47	34.3	3.63	-
es_en_orig	es-en	train	36.70	3.11	-	-	-
es_en_extra		train+extra	36.57	2.42	-	-	-
es_en_aligned		train+extra+aligned	36.96	2.81	37.71	**3.47**	4
es_en_copied		train+extra+aligned+copied	31.48	1.22	-	-	-
es_en_quz		train+extra+quz	36.86	2.72	-	-	-
es_fi_orig	es-fi	train	36.93	2.86	-	-	-
es_fi_extra		train+extra	37.51	3.04	38.21	3.11	2
es_fi_aligned		train+extra+aligned	37.34	2.90	**38.59**	3.45	3
es_fi_copied		train+extra+aligned+copied	32.01	1.66	-	-	-
es_fi_quz		train+extra+quz	**37.70**	**3.36**	38.40	3.08	1
es_fi_all		all	36.40	2.54	37.26	3.06	5

Table 8: Results of the experiments on the development data, and official results on test data of Spanish-Quechua language pair. Column "Sub" describes the submission number to the official shared task evaluation.

- train + extra + aligned + quz: It includes all the data above excluding copied corpus, but also includes the additional data gathered from different sources pertaining to Cuzco Quechua (Quz). The reason for removing copied corpus was that it resulted in a decrease of the chrF score in all the experiments.
- all: It includes all the data above excluding the copied corpus, but includes the synthetic translations, as mentioned in Section 3.4.

4.1 Fine-tuned Spanish to English

Following (Vázquez et al., 2021), where including a majority of Spanish-English parallel data while building an MT system for low-resource languages improved the performance across all the languages, we decided to use an already fine-tuned Spanish-English MT model and fine-tune it again on our Spanish-Quechua parallel corpus. Concretely, we used the opus-mt-es-en model available at Huggingface [8]. As expected, we can see that these models perform quite close to the baseline system. Including more data seems to help as well, with the exception of copied corpus. The reason for this, we suspect, is due to the quantity of the data being higher than our total Spanish-Quechua parallel corpora (no analysis was done on the quality of the data). The best model in this case was fine-tuned on train + extra + aligned achieving a chrF score of 36.96 and 37.71 on the development and test set respectively with the train + extra + quz performing quite similarly as well.

4.2 Fine-tuned Spanish to Finnish

Lastly, we tried using a fine-tuned version of the Spanish-Finnish MT model. The model we used was opus-mt-es-fi, available at Huggingface [9]. The reason for choosing this specific model was firstly because of the similarity between Finnish and Quechua, i.e, both being agglutinative languages, and secondly, Finnish being a relatively high-resource language as compared to Quechua. This proved to be the best model among our experiments, which we believe is due to the reasons mentioned above. We can see in Table 8 that adding aligned data from Ayacucho Quechua seems to help more than adding Cuzco Quechua parallel sources. The best model among the experiments was trained on train + extra + aligned and achieved a chrF score of 37.34 and 38.59 on the development and test set respectively.

One final experiment was conducted on all of the collected data meaning train + extra + aligned + quz + bcktr. The model was able to achieve a chrF score of 36.40 and 37.26 on the Spanish-Quechua development and test set respectively. All the models are available on Huggingface [10]

[8] https://huggingface.co/Helsinki-NLP/opus-mt-es-en

[9] https://huggingface.co/Helsinki-NLP/opus-mt-es-fi

[10] https://huggingface.co/americasnlp-lct-ehu

5 Conclusion

To summarize our findings, in our submission to the AmericasNLP 2023 low-resource machine translation shared task for the Spanish-Quechua language pair, we have explored fine-tuning existing models in different language pairs, combining them with different data setups. We have collected and aligned new parallel data, created synthetic translations, and made use of copied corpus approach. The highest-performing model on the development data achieved 37.70 chrF. This model was obtained by fine-tuning OPUS MT's Spanish-Finnish model on the original training data, augmented with additional data presented by previous year's teams, both for Ayacucho and Cuzco Quechua. In the test set, however, the highest performing model was different, obtaining a chrF score of 38.59. This model was the same as the previous one, but the data consisted of the original training data, data from previous year's submissions (excluding Cuzco Quechua) and the novel alignments introduced in this work.

6 Acknowledgements

The authors would like to thank Erasmus Mundus European Masters Program in Language and Communication Technologies for its support. We would also like to thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Hábrók high performance computing cluster.

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.

Alicia Avellana. *Las Categorías Funcionales en el Español en Contacto con Lenguas Indígenas de la Argentina: Tiempo, Aspecto y Modo*. Ph.D. thesis, Universidad de Buenos Aires, year =.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156, Copenhagen, Denmark. Association for Computational Linguistics.

Oscar Moreno. 2021. The REPU CS' Spanish–Quechua submission to the AmericasNLP 2021 shared task on open machine translation. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 241–247, Online. 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, Boston, MA. Association for Machine Translation in the Americas.

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

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.

Annette Rios. 2016. *A Basic Language Technology Toolkit for Quechua*. Ph.D. thesis.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. 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).

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Daniel Varga, Peter Halacsy, Andras Kornai, Viktor Nagy, Laszlo Nemeth, and Viktor Tron. 2007. Parallel corpora for medium density languages. In *Recent Advances in Natural Language Processing IV. Selected papers from RANLP-05*, pages 247–258. Benjamins, Amsterdam.

Raúl Vázquez, Yves Scherrer, Sami Virpioja, and Jörg Tiedemann. 2021. The Helsinki submission to the AmericasNLP shared task. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 255–264, Online. Association for Computational Linguistics.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Rodolfo Zevallos, John Ortega, William Chen, Richard Castro, Nuria Bel, Cesar Toshio, Renzo Venturas, Hilario Aradiel, and Nelsi Melgarejo. 2022. Introducing qubert: A large monolingual corpus and bert model for southern quechua. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 1–13.