BSC Submission to the AmericasNLP 2024 Shared Task

Javier García Gilabert **Aleix Sant Savall** **Carlos Escolano**
Francesca De Luca Fornaciari **Audrey Mash** **Maite Melero**
Barcelona Super Computing Center (BSC),
{ javier.garcia1, aleix.santsavall, carlos.escolano,
francesca.delucafornaciari, audrey.mash, maite.melero}@bsc.es

Abstract

This paper describes the BSC's submission to the AmericasNLP 2024 Shared Task. We participated in the Spanish to Quechua and Spanish to Guarani tasks. In this paper we show that by using LoRA adapters we can achieve similar performance as a full parameter fine-tuning by only training 14.2% of the total number of parameters. Our systems achieved the highest ChrF++ scores and ranked first for both directions in the final results outperforming strong baseline systems in the provided development and test datasets.

1 Introduction

In this paper we present the submission of the Barcelona Super Computing Center to the Workshop on Natural Language Processing (NLP) for Indigenous Languages of the Americas (AmericasNLP) 2024 Shared Task on machine translation systems for Indigenous languages. We participated in two directions: Spanish to Guarani and Spanish to Quechua.

The Quechuan language family, comprising a diverse range of dialects, is spoken by over 6 million individuals across South America. It has a wide geographic spread, extending from Colombia through Ecuador, Peru, Bolivia, and Argentina. Quechua is an oral language, which results in a scarcity of textual resources. In this work we use an amalgamated data set for Quechua from Cuzco (quz) and Ayacucho (quy), which are the most widely spoken.

Guarani belongs to the Tupi–Guarani branch of the Tupian language family. It is one of the official languages of Paraguay and has more than 9 million speakers. The data used in this work contains a mix of dialects, from pure Guarani to more mixed Jopara.

In this work we fine-tuned various versions of NLLB-200 (Costa-jussà et al., 2022) as in previous year winning submission (Gow-Smith and Sánchez Villegas, 2023). Originally, NLLB-200 had been trained using parallel data in Quechua, Guarani, and Spanish. Our experiments aimed to answer the following questions: *(1)* How beneficial is fine-tuning with multilingual data in low resource scenarios? *(2)* Does translation quality improve with a greater model size for low resource languages? *(3)* Can we use LoRA for fine-tuning multilingual neural machine translation (MNMT) systems?

The remainder of this paper is organized as follows: Section 2 describes the data collection and preprocessing, Section 3 introduces our training methodology, Section 4 presents the results of our experiments and comments the results on the test sets, and finally, Section 5 summarizes the main findings.

2 Data collection and preparation

In this section we detail our data sources and the data cleaning pipeline used for cleaning training data. We utilize the parallel data provided by the organizers and additional sources for training our systems.

2.1 Data collection

AmericasNLP 2024 For Quechua, AmericasNLP training data included the Jw300 corpus (Agić and Vulić, 2019), MINEDU, and dict_misc. Additionally, English to Ayacucho Quechua (quy) and English to Cuzco Quechua (quz) translations of the Jw300 were also given. As for Guarani, the parallel data provided by the organizers was collected from web sources in a semi-automatic way and later supervised by a translator (Chiruzzo et al., 2020).

Helsinki The organizers made available parallel data from the Helsinki University (De Gibert et al., 2023). This data included resources from OPUS (Tiedemann, 2012). In addition, for Guarani they

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 143–149

also provided PYconst, News and Jojajovai. Furthermore, they generated synthetic data using two approaches: first, Spanish synthetic data was generated from Quechua and Guarani using Wikipedia monolingual corpus (Tiedemann, 2020), and second, by pivoting with English. They translated the English source sentences from the en-quy and en-quz Jw300 corpus into Spanish, resulting in synthetic es-quy and es-quz data, and translated the English source sentences from en-gn NLLB-seed corpus into Spanish, resulting in es-gn.

OPUS From the OPUS corpus collection, we get GNOME, Mozilla-I10n, Tatoeba, Ubuntu, and Wikimedia, all of which are available in Quechua and Guarani.

FLoRes-200 Developed by Costa-jussà et al., it is a widely-used multiparallel evaluation dataset which includes Guarani and Quechua. We extracted the corresponding Spanish to Guarani and Spanish to Quechua pairs from test and devtest sets, concatenated them, and utilized them as training data.

Hackathon-pln-es This dataset contains translations from Spanish to Quechua Ayacucho (quy). It is available at HuggingFace[1] and is a collection of webs and others datasets, namely: "Mundo Quechua", "Kuyakuykim (Te quiero): Apps con las que podrías aprender quechua", "Piropos y frases de amor en quechua", "Corazón en quechua", and parallel data from AmericasNLP 2021 Shared Task. We concatenated test set, validation set, and train sets.

Monolingual-Quechua The Monolingual-Quechua-IIC dataset (Zevallos et al., 2022) is a monolingual corpus of Quechua extracted from 50 different monolingual sources on the web which is available at HuggingFace[2]. Google Translate was employed to generate Spanish translations from this monolingual dataset, thereby obtaining parallel synthetic source generated data.

NLLB The NLLB corpus[3] is a dataset that was created by OPUS (Tiedemann, 2012) based on metadata for mined bitext released by the NLLB project (Costa-jussà et al., 2022). We extracted and incorporated pt-qu, pt-gn, en-qu and en-gn parallel data in our training dataset.

After collecting all the data, it was concatenated per language pair and then each set underwent the cleaning pipeline explained in the following subsection.

2.2 Data preprocessing

Given that training data was obtained from various sources and a significant part is synthetically-created, we employed a comprehensive data preprocessing pipeline to obtain good quality parallel data. In particular, we remove duplicates and near duplicates, we filter parallel sentences using heuristic-based and embedding-based methods and we remove off-target translations.

Cleaning step 1 We employed Moses (Koehn et al., 2007) standard preprocessing script to training data: `clean-corpus-n.perl` which removes sentences that have more than 150 tokens and removes sentence pairs that have a length ratio greater than 3. Then, Bifixer (Ramírez-Sánchez et al., 2020) was used to fix possible text issues.

Cleaning step 2 To remove duplicates and near-duplicates we used NLPDedup[4]. Then, we removed off-target translations using googletrans python library[5]. We found that there were a considerable number of near duplicates and off-target translations in the NLLB dataset collected from OPUS. In the English to Quechua direction, for example, 47% of the data were near duplicates, and 21.8% of the deduplicated data were off-target translations.

Cleaning step 3 We computed embedding-based similarity between a sentence pair using a sentence embedding model. We use SONAR (Duquenne et al., 2023) to embed both source and target sentences then compute a cosine similarity score between both. SONAR is a novel sentence embedding model that covers 200 languages and outperforms existing sentence embeddings such as LASER3 (Heffernan et al., 2022) and LabSE (Feng et al., 2022). It consists of an encoder-decoder architecture initialized from NLLB 1.3B dense model and trained by adding a pooling operation at the encoder's outputs to extract sentence representations.

[1] https://huggingface.co/datasets/hackathon-pln-es/spanish-to-quechua

[2] https://huggingface.co/datasets/Llamacha/monolingual-quechua-iic

[3] https://opus.nlpl.eu/NLLB/corpus/version/NLLB

[4] https://github.com/saattrupdan/NLPDedup

[5] https://github.com/ssut/py-googletrans

3 Methodology

3.1 Baseline Fine-tuning

We fine-tune separate NLLB models of 1.3B parameters for the Spanish to Quechua and Spanish to Guarani training data respectively. For each direction we experiment with different thresholds for choosing the training set according to the cosine similarity values computed using SONAR.

3.2 Multilingual Fine-tuning

Following similar lines as Gow-Smith and Sánchez Villegas, we fine tune NLLB-1.3B with multilingual data. We experiment with different variants of the training data to study whether a richer linguistic mixture of Quechua and Guarani is beneficial for the translation quality. Specifically, we include Portuguese and English directions to the Spanish ones. In table 1 we show the amount of parallel sentences collected for each translation pair.

Source	Quechua	Guarani
es `Original`	613,006	91,025
pt `Original`	2,071,571	684,883
en `Original`	2,874,090	2,959,122
es `Filtered`	132,884	42,504
pt `Filtered`	124,942	69,046
en `Filtered`	167,947	96,487
Total	427,773	208,037
Deduplicated	425,773	208,037

Table 1: Corpus statistics. `Filtered` refers to the number of pairs resulting from the data preprocessing pipeline.

Synthetic target generated data We study the impact of translation quality when adding synthetic generated data on the target side when fine tuning the model. Synthetic data was generated using Google Translate, translating Spanish monolingual data from TED2020 (Reimers and Gurevych, 2020) collected from OPUS website[6] which consists in 416,846 sentences in Spanish for about 4,000 TED talks covering a wide range of domains. We find that this data improves performance for the Spanish to Quechua direction (see Section 4).

Model size To investigate the effect of increasing the model capacity, we fine-tune the NLLB pre-trained model of 3.3B parameters with the best configuration found for the previous experiments with the model of 1.3B parameters.

3.3 LoRA Fine-tuning

Parameter-efficient fine-tuning (PEFT) techniques adapt pre-trained models by fine-tuning only a small subset of the model's parameters. The Low-Rank Adaptation (LoRA) technique (Hu et al., 2021) has been popularized for LLM training for its efficiency and often comparable results with full-parameter fine-tuning (Sun et al., 2023).

LoRA uses low-rank parametrized update matrices to reduce the number of trainable parameters. More specifically, given a pre-trained weight matrix $\mathbf{W}_0 \in \mathbb{R}^{d \times k}$, matrix $\mathbf{W}_0$ is updated with a low-rank decomposition as follows:

$$\mathbf{W}_0 + \frac{\alpha}{r} \Delta\mathbf{W} = \mathbf{W}_0 + \frac{\alpha}{r} \mathbf{BA}, \qquad (1)$$

where $\mathbf{B} \in \mathbb{R}^{d \times r}$, $\mathbf{A} \in \mathbb{R}^{r \times k}$, r is the rank of the trainable matrices and α is a scaling parameter that scales the learned weights. During the fine-tuning, the original matrix $\mathbf{W}_0$ remains unchanged and does not receive gradient updates. However, matrices $\mathbf{A}$ and $\mathbf{B}$ are updated during training.

We apply LoRA to the feed-forward and attention layers. The rank of the trainable matrices was set to 256 and α was set to 512.

3.4 Setup

For fine-tuning we use AdamW (Loshchilov and Hutter, 2019) optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-6}$, $\lambda = 0.01$. We use the inverse square root scheduler with an initial learning rate of 2e-4 and 15,000 warmup steps. We set a batch size of 4 with 2 gradient accumulation steps, and train on 10 epochs. We use 1,000 sentences from the training dataset as the validation dataset in training. All models are trained using the Transformers[7] library on H100 GPUs. For LoRA we use the Peft[8] library. We saved best checkpoints every 4,000 steps according to the best ChrF++ metric in the validation set.

3.5 Inference

We limit the translation length to 512 tokens. For the case of Quechua we do not allow the model to generate an apostrophe as the Quechua Ayacucho variety was the one used in development and test

[6] https://opus.nlpl.eu/

[7] https://huggingface.co/

[8] https://huggingface.co/docs/peft/index

sets which does not handle pentavocalism. For inference, we use beam search and we experiment with different values of the beam size and the repetition penalty term (β). We find that adjusting the repetition penalty term yields major improvements in the case of Guarani.

3.6 Evaluation

We evaluate our fine-tuned models with the evaluation script provided by the organizers. Specifically, models were evaluated with two metrics: BLEU (Papineni et al., 2002) and ChrF++ (Popović, 2017) using SacreBLEU[9] implementation (Post, 2018).

We compare our model's performance against the baseline models released by the University of Helsinki (De Gibert et al., 2023) and the University of Sheffield (Gow-Smith and Sánchez Villegas, 2023). In addition, we provide the performance of Google Translate and the pre-trained NLLB 1.3B and 3.3B without any fine-tuning.

4 Results

4.1 Dev Set Results

Baseline Fine-tuning We first fine-tune the NLLB 1.3B models using bilingual data. Figure 1 shows the results of a full parameter fine-tuning of the model using different thresholds for selecting training data given the cosine similarity scores computed using SONAR. Based on the results we decide to use a threshold of 0.2 and 0.3 for Guarani and Quechua respectively for our next experiments. Note that we do not experiment with thresholds greater than 0.4 for Guarani as the number of sentence pairs between Spanish and Guarani given that threshold is relatively low (less than 19K pairs).

Multilingual Fine-tuning Table 3 show the BLEU and ChrF++ metrics evaluated on the development dataset for the Spanish to Quechua and Spanish to Guarani directions. As shown in the table, using synthetic target generated data combined with multilingual training improves translation quality for the Spanish to Quechua direction. In comparison to baseline models, we observe that our systems exhibit strong performance. In the development set we outperform Google Translate by +1.41 ChrF++ points and Sheffield's baseline by +6.05 points.

Regarding the Spanish to Guarani direction, we find that training with multilingual parallel data

[9]Word n-gram order was set to 2 for SacreBLEU implementation of ChrF++.

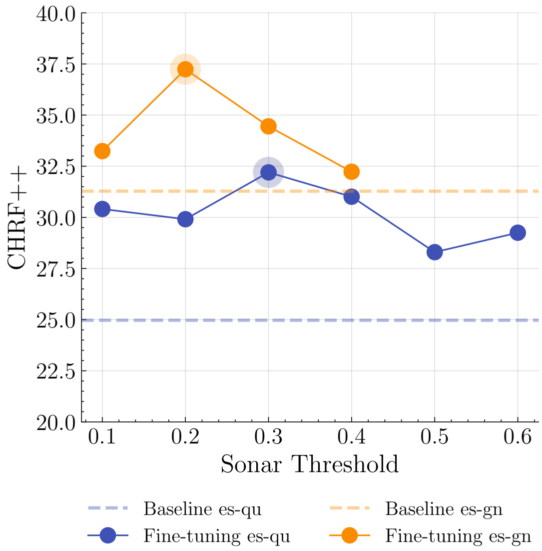

Figure 1: Performance evaluating on dev set. ChrF++ is in the vertical axis, and value for the threshold is in in the horizontal axis.

(en→gn and pt→gn) does not improve translation quality. However, our system trained only with Spanish to Guarani parallel data outperforms baseline models. We outperform Helsinki's baseline model by +2.74 ChrF++ points and Sheffield's baseline by +5.31 points.

Model size Table 2 compares the performance when increasing the model size. We compare the NLLB 1.3 Billion parameter model with the 3.3 Billion parameter one. When fine-tuning using the 3.3B model, we find that we do slightly worse (-1.35 ChrF++ points) than with the 1.3B model for the Spanish to Guarani direction, yet we gain +0.12 ChrF++ points for the Spanish to Quechua direction. These results indicate that increasing the model size does not yield superior performance. Mainly due to the small amount of data used for fine-tuning.

	Spanish → Quechua		Spanish → Guarani	
	ChrF++	**BLEU**	**ChrF++**	**BLEU**
NLLB-1.3B	36.27	3.77	**37.48**	**11.15**
NLLB-3.3B	**36.39**	**4.07**	36.13	10.20

Table 2: Experiments with model size.

LoRA Table 4 shows the results when fine-tuning using LoRA adapters. Notably, for the Spanish to Quechua direction, our system performs as well as the full fine-tuning with the NLLB 3.3B model

		Data			Spanish → Quechua		Spanish → Guarani	
		es	en	pt	ChrF++	BLEU	ChrF++	BLEU
Baseline	Helsinki				28.78	-	34.74	-
	Sheffield				30.22	-	32.17	-
	Google Translate				34.86	3.23	30.33	4.71
Inference	NLLB-1.3B				24.97	1.95	31.28	6.27
	NLLB-3.3B				26.84	1.64	32.03	6.31
Our	NLLB-1.3B	✓	✗	✗	32.20	3.28	**37.48**	**11.15**
	NLLB-1.3B	✓	✓	✗	32.71	3.18	33.94	7.89
	NLLB-1.3B	✓	✗	✓	32.40	3.35	33.92	7.88
	NLLB-1.3B	✓	✓	✓	31.59	2.94	34.01	8.26
Our	NLLB-1.3B + Synthetic	✓	✗	✗	35.12	3.55	32.50	6.08
	NLLB-1.3B + Synthetic	✓	✓	✓	**36.27**	**3.77**	32.48	6.00

Table 3: Evaluations computed on the development dataset. The symbol ✓ indicates that the parallel data in the corresponding direction was utilized for the fine-tuning whereas ✗ indicates that it was not used.

when fine-tuning with LoRA, even though it has 85.8% fewer trainable parameters. We perform just slightly worse when compared to the full fine-tuning in the Spanish to Guarani direction when using LoRA.

	Trainable Parameters	es→qu	es→gn
NLLB-3.3B	3,898,511,360	36.39	36.13
NLLB-1.3B	1,748,125,696	36.27	37.48
NLLB-3.3B + LoRA	553,648,128	36.40	35.24
NLLB-1.3B + LoRA	377,487,360	36.11	35.26

Table 4: Experiments with LoRA fine-tuning. ChrF++ scores are shown for Spanish to Quechua (es→qu) and Spanish to Guarani (es→gn) directions.

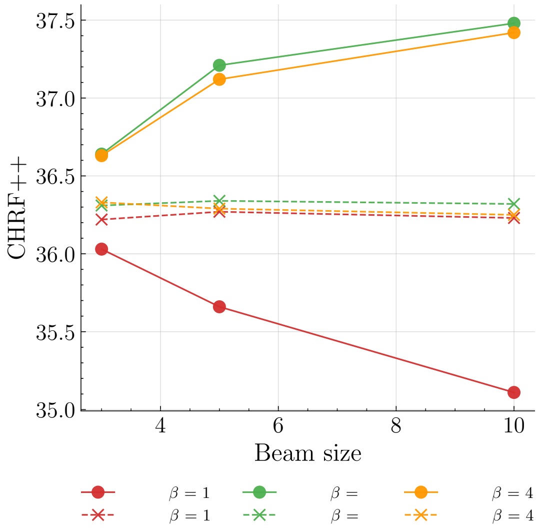

Figure 2: Performance evaluating on dev set grouped by repetition penalty. ChrF++ is in the vertical axis, and beam size is in the horizontal axis.

Inference experiments A grid search is performed to understand how the beam size and the repetition penalty term (β) parameters affect the translation quality. We plot the obtained results in Figure 2.

We observe that for the Spanish to Guarani direction, the penalty term has a significant impact on the translation quality as measured by the ChrF++ score. Specifically, we find that best ChrF++ results are obtained with a beam size of 10 and $\beta = 4$.

However, for the Spanish to Quechua direction, neither the repetition penalty term nor the beam size seem to have significant impact on the translation quality.

4.2 Test Set Results

In table 5 we show the results on the test sets. Our submitted systems ranked first in both directions. On Spanish to Quechua, we achieve 38.21 ChrF++ using LoRA adapters with the 1.3B model. We observe that our submitted systems significantly outperform baseline models by +4.2 points on Sheffield's baseline and by +9.4 points on Helsinki's baseline for this direction. On Spanish to Guarani direction we achieve 38.93 ChrF++ and our submitted systems outperform Helsinki's baseline model by up to +1.91 points and by up to

+3.17 points for Sheffield's baseline model.

Model	r_{qu}	es→qu	r_{gn}	es→gn
Baselines				
Helsinki		28.81		37.02
Sheffield		34.01		35.76
Full fine-tuning				
NLLB-3.3B	2	38.19	2	37.64
NLLB-1.3B	4	38.02	**1**	**38.93**
LoRA				
NLLB-3.3B + LoRA	3	38.10	3	37.63
NLLB-1.3B + LoRA	**1**	**38.21**	4	37.42

Table 5: Final test results. ChrF++ scores are shown for Spanish to Quechua (es→qu) and Spanish to Guarani (es→gn) directions. r_{qu} and r_{gn} denote the position in the final classification for Spanish to Quechua and Spanish to Guarani respectively.

5 Conclusions

In this paper we describe the Barcelona Supercomputing Center submission to the AmericasNLP 2024 shared task. We took part in the Spanish to Quechua and Spanish to Guarani tracks. We fine-tuned different versions of the NLLB-200 model. Our systems ranked in first place in both translation directions outperforming the provided baselines.

Our experiments show that increasing model size does not yield superior performance when data is scarce. Also, we show that training with multilingual data combined with synthetic data improves translation quality for the Spanish to Quechua direction. Finally, we show that by fine-tuning the models using LoRA, we can obtain a similar performance as a full parameter fine-tuning while training only 14.2% of the parameters.

Acknowledgments

This work has been supported by the Spanish project PID2021-123988OB-C33.

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.

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

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Ona De Gibert, Raúl Vázquez, Mikko Aulamo, Yves Scherrer, Sami Virpioja, and Jörg Tiedemann. 2023. Four approaches to low-resource multilingual NMT: The Helsinki submission to the AmericasNLP 2023 shared task. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 177–191, Toronto, Canada. Association for Computational Linguistics.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoit Sagot. 2023. SONAR: sentence-level multimodal and language-agnostic representations.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Edward Gow-Smith and Danae Sánchez Villegas. 2023. Sheffield's submission to the AmericasNLP shared task on machine translation into indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 192–199, Toronto, Canada. Association for Computational Linguistics.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2101–2112, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

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, Belgium, Brussels. Association for Computational Linguistics.

Gema Ramírez-Sánchez, Jaume Zaragoza-Bernabeu, Marta Bañón, and Sergio Ortiz-Rojas. 2020. Bifixer and bicleaner: two open-source tools to clean your parallel data. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 291–298, Lisboa, Portugal. European Association for Machine Translation.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Xianghui Sun, Yunjie Ji, Baochang Ma, and Xiangang Li. 2023. A comparative study between full-parameter and lora-based fine-tuning on chinese instruction data for instruction following large language model. *arXiv preprint arXiv:2304.08109*.

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. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Rodolfo Zevallos, John Ortega, William Chen, Richard Castro, Núria 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, Hybrid. Association for Computational Linguistics.