QUESPA Submission for the IWSLT 2023 Dialect and Low-resource Speech Translation Tasks

John E. Ortega[1], **Rodolfo Zevallos**[2], **and William Chen**[3]
[1]Northeastern University, USA, [2]Universitat de Pompeu Fabra, Spain
[3]Carnegie Mellon University, USA
contact email: j.ortega@northeastern.edu

Abstract

This article describes the QUESPA team speech translation (ST) submissions for the Quechua to Spanish (QUE–SPA) track featured in the Evaluation Campaign of IWSLT 2023: low-resource and dialect speech translation. Two main submission types were supported in the campaign: *constrained* and *unconstrained*. We submitted six total systems of which our best (primary) constrained system consisted of an ST model based on the Fairseq S2T framework where the audio representations were created using log mel-scale filter banks as features and the translations were performed using a transformer. The best (primary) unconstrained system used a pipeline approach which combined automatic speech recognition (ASR) with machine translation (MT). The ASR transcriptions for the best unconstrained system were computed using a pre-trained XLS-R–based model along with a fine-tuned language model. Transcriptions were translated using a MT system based on a fine-tuned, pre-trained language model (PLM). The four other submissions are presented in this article (2 constrained and 2 unconstrained) for comparison because they consist of various architectures. Our results show that direct ST (ASR and MT combined together) can be more effective than a PLM in a low-resource (constrained) setting for Quechua to Spanish. On the other hand, we show that fine-tuning of any type on both the ASR and MT system is worthwhile, resulting in nearly 16 BLEU for the unconstrained task.

1 Introduction

Low-resource machine translation (LRMT) can be considered a difficult task due to the low amount of parallel data on hand. (Haddow et al., 2022) By adding the task of automatic speech recognition (ASR), complexity can be even more difficult. Findings from the previous year's IWSLT 2022 (Antonios et al., 2022) have shown that for low-resource language pairs like Tamasheq–French, it is difficult to achieve more than 5 BLEU (Papineni et al., 2002) score points for the combined task of speech translation (ST), even in a unconstrained setting.

This year, the IWSLT 2023 (Agarwal et al., 2023) evaluation campaign for low-resource and dialect speech translation has included Tamasheq–French along with several other language pairs. One of the newly introduced language pairs is Quechua–Spanish deemed **QUE–SPA** by the organizers. Quechua is an indigenous language spoken in the Andes mountainous region in South America. It is spoken by millions of native speakers mostly from Peru, Ecuador and Bolivia. In those regions, the high-resource language is Spanish. Quechua displays many unique morphological properties of which high inflection and poly-synthetic are the two most commonly known. It is worthwhile to note that previous work (Ortega and Pillaipakkamnatt, 2018; Ortega et al., 2020) has been somewhat successful in identifying the inflectional properties of Quechua such as agglutination where another high-resource language, namely Finnish, can aid for translation purposes achieving nearly 20 BLEU on religious-based (text-only) tasks.

Since this is the first year that QUE–SPA has been included in the IWSLT 2023 campaign, we feel that it is important to set a proper baseline. The aim of our submission was to increase the viability of the use of a Quechua–Spanish ST system and we thus attempted several approaches that included the use of pipelines (cascade) approaches along with joint ASR + MT. We report on the six system submissions as a final takeaway for this article; however, we also compare other approaches that performed worse (1 BLEU or less). Our team is called **QUESPA** and consists of a consortium that spans across three universities: Northeastern University (USA), Universitat de Pompeu Fabra (Spain), and Carnegie Mellon University (USA). Our objective is to help to solve the LRMT prob-

Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023), pages 261–268

lem for Quechua with the intention of at some point releasing an ST system to the Quechua community where we have strategic partners located in areas of Peru where Quechua is mostly spoken. The authors of this article have participated in several other events and written literature that includes native Quechua annotations for natural language processing (NLP) systems including MT and more.

This article reports the QUESPA consortium submissions for the IWSLT 2023 dialect and low-resource tasks. We focus only on the low-resource task despite the mention of two dialects *Quechua I and II*. Our focus is on creating the optimum models we can for the *constrained* task and leveraging pre-trained models for the *unconstrained* task further described in Section 3.

The rest of this article is organized as follows. Section 2 presents the related work. The experiments for QUE–SPA low-resource track are presented in Section 3. Section 4 provides results from the six submitted systems and concludes this work.

2 Related work

In this section, we first cover work directly related to the ASR and MT tasks of QUE–SPA done in the past. Then, we introduce related work on ST models in general to provide an idea of what work is current in the field.

Quechua to Spanish MT approaches have become more abundant in the past few years. When it comes to ASR–>MT, or ST approaches, there are few attempts officially recorded. In this section, we list previous work in chronological order to better explain the MT approaches attempted. First, Rios (2015) provided an advanced linguistic Quechua toolkit that used finite state transducers (FSTs) to translate from Spanish to Quechua. Her work laid the foundation for future work and helped to promote the digitization of the Quechua language. After that, Ortega and Pillaipakkamnatt (2018) and Cardenas et al. (2018) introduced several new findings that included the ASR corpus used in the IWSLT 2023 task for both unconstrained and constrained purposes. Not long after, Ortega et al. (2020) introduced the first known attempt of a neural MT system that included several annotators along with the state-of-the-art techniques in subsegmentation such as byte-pair encoding (BPE) (Sennrich et al., 2015). Their work was then extended by others (Chen and Fazio, 2021) more recently to achieve 23 BLEU on religious-based text, the highest performing QUE–SPA for its time.

None of the approaches before Chen and Fazio (2021) work included the use of pre-trained language models (PLMs) for low-resource languages. However, the introduction of zero-shot models occurred at the low-resource machine translation workshop in 2020 (Ojha et al., 2020) and not long after in 2021 at the Americas NLP workshop (Mager et al., 2021). The Americas NLP 2021 workshop included the use of QUE–SPA, albeit for MT only achieving scores of 5.39 BLEU through the use of a multi-lingual model trained on 10 other indigenous languages. Their work did not include zero-shot task approaches as introduced by Ebrahimi et al. (2022) where fine-tuning was performed on a pre-trained XLM-R (Conneau et al., 2020) model that achieved impressive results (40–55 BLEU). More recent work (Weller-Di Marco and Fraser; Costa-jussà et al., 2022) did not surpass those results for MT of QUE–SPA.

To our knowledge only one competition/shared task has attempted to process QUE–SPA for speech translation purposes – Americas NLP 2022[1]. However, the findings for the task have not beeen published as of the writing of this article. Their competition used corpora similar to IWSLT 2023 but lacks MT data as a separate (constrained) resource. They also do not introduce the concept of constrained or unconstrained tasks as was done at IWSLT 2023.

Apart from those tasks that directly use the QUE–SPA language pair, several mainstream techniques are currently being used as alternatives to supervised (from scratch) training. For example, one of the most common approaches for both ST and MT approaches tend to use a transformer in some capacity along with a PLM. One such model that uses a multi-lingual low-resource corpus called Flores (Guzmán et al., 2019) is Facebook's NLLB (no language left behind) approach (NLLB Team et al., 2022). Their approach uses self-supervised learning (SSL) from previous innovation (Pino et al., 2020) for multi-lingual approaches that combines ASR with MT in a ST task alone and is made available through Fairseq (Wang et al., 2020). In our work, our primary systems use Fairseq and Facebook's PLMs with sentence embeddings based on previous work (Artetxe and Schwenk, 2019) and the M2M (multi-to-multi) model (Fan et al., 2021) consisting of 1.2 Billion parameters. This

[1] https://github.com/AmericasNLP/americasnlp2022

enables zero-shot cross-lingual transfer for many low-resource languages, including Quechua.

We provide reference to previous work that includes either a *direct* or *end-to-end* ST models (Berard et al., 2016; Weiss et al., 2017). More traditional approaches typically use a *cascade* approach which first transcribes using an ASR model and then translates using a MT model. While recent work (Bentivogli et al., 2021; Anastasopoulos et al., 2021; Antonios et al., 2022) has shown that the direct ST approaches are worthy, traditional approaches work well for low-resource situations too. In our system submissions, all of our systems with exception of the primary constrained used the cascade approach.

3 Quechua-Spanish

In this section we present our experiments for the QUE–SPA dataset provided in the low-resource ST track at IWSLT 2023. This is the first time that this dataset has been officially introduced in its current state which contains 1 hour and 40 minutes of *constrained* speech audio along with its corresponding translations and nearly 60 hours of ASR data (with transcriptions) from the Siminichik (Cardenas et al., 2018) corpus. AmericasNLP 2022's task used a smaller part of the dataset but the data was not presented or compiled with the same offering and, as of this writing, have not published their results. This dataset aggregates the QUE–SPA MT corpus from previous neural MT work (Ortega et al., 2020). The audio and corresponding transcriptions along with their translations are mostly made of of radio broadcasting, similar to the work from Boito et al. (2022) which contains 17 hours of speech in the Tamasheq language.

We present the six submissions for both the *constrained* and *unconstrained* as follows:

1. a primary constrained system that uses a direct ST approach;

2. a contrastive 1 constrained system consisting of a wav2letter (Pratap et al., 2019) ASR system and a neural MT system created from scratch;

3. a contrastive 2 constrained system consisting of a conformer-based (Gulati et al., 2020) ASR system and a neural MT system created from scratch;

4. a primary unconstrained system consisting of a multi-lingual PLM ASR model, a Quechua recurrent neural-network language model, and a fine-tuned neural MT system based on a PLM;

5. a contrastive 1 unconstrained system consisting of a multi-lingual PLM ASR model and a fine-tuned neural MT system based on a PLM;

6. a contrastive 2 unconstrained system consisting of a wav2letter ASR system and a fine-tuned neural MT system based on a PLM.

We present the experimental settings and results for all systems starting off with constrained systems in Section 3.1 and continuing with the unconstrained systems in Section 3.2. We then describe the other less successful approaches in Section 3.3. Finally, we offer results and discussion in Section 4.

3.1 Constrained Setting

The IWSLT 2023 constrained setting for QUE–SPA consists of two main datasets. First, the speech translation dataset consists of 1 hour and 40 minutes divided into 573 training files, 125 validation files, and 125 test files where each file is a .wav file with a corresponding transcription and human-validated translation from Simanchik (Cardenas et al., 2018). Secondly, there is a MT data set combined by previous work (Ortega et al., 2020) which consists of 100 daily magazine article sentences and 51140 sentences which are of religious context in nature.

3.1.1 Primary System

The Primary System consists of a direct ST approach. Since the constrained setting does not allow for external data, we used only the data provided. We use the Fairseq (Ott et al., 2019) toolkit to perform direct ST using the 573 training files, a total of 1.6 hours of audio. The system extracts log mel-filter bank (MFB) features and is based on the S2T approach by (Wang et al., 2020). We generate a 1k unigram vocabulary for the Spanish text using SentencePiece (Kudo and Richardson, 2018), with no pre-tokenization. Our model consists of a convolutional feature extractor and transformer encoder-decoder (Vaswani et al., 2017) with 6 encoder layers and 3 decoder layers. Error is measured using cross entropy and optimization is done using Adam. Our model was run for 500 epochs with a learning rate of .0002.

3.1.2 Contrastive 1 System

The Contrastive 1 System is a cascade system where first ASR is performed to produce transcriptions that are translated using a separate MT system. For the ASR system, we used the wav2letter++ (Pratap et al., 2019) model. The wav2letter++ model consists of a RNN with 30M parameters (2 spatial convolution layers, 5 bidirectional LSTM layers, and 2 linear layers) and a CNN with 100M parameters (18 temporal convolution layers and 1 linear layer). We use the convolutional gated linear unit (GLU) (Dauphin et al., 2017) architecture proposed in the recipe wav2letter (WSJ) (Collobert et al., 2016). Our experiments using wav2letter++ took 134 epochs to train, using Stochastic Gradient Descent (SGD) with Nesterov momentum and a minibatch of 8 utterances. The initial learning rate was set to 0.006 for faster convergence, and it was annealed with a constant factor of 3.6 after each epoch, with momentum set to 0. The model was optimized using the Auto Segmentation Criterion (ASG) (Collobert et al., 2016). During development, the ASR system WER was 72.15 on the validation set. The MT system was created from scratch using the OpenNMT framework (Klein et al., 2020) with the MT data provided for the constrained task along with the ASR training data. More specifically, the MT system's encoder and decoder are based on a transformer (Vaswani et al., 2017) (encode/decode) architecture of 6 layers. Hidden layer and vectors sizes were 512. Dropout was set to 0.1. Optimization was done using the Adam optimizer. Tokenization was done using SentencePiece (Kudo and Richardson, 2018). Both source and target vocabularies were 50k. Initial BLEU score on the validation set was 21.13.

3.1.3 Contrastive 2 System

Similar to the Contrastive 1 System, the Contrastive 2 system is a cascade approach. The ASR system, however, is distinct. It is derived using MFB features similar to previous work Berrebbi et al. (2022). It uses a conformer instead of the transformer encoder like Gulati et al. (2020). Training was performed using a hybrid CTC/attention loss (Watanabe et al., 2017). The model was optimized using Adam (Kingma and Ba, 2015) and a Noam learning rate scheduler (Vaswani et al., 2017) with 4000 warmup steps. The MT system is identical to the OpenNMT MT system mentioned for the Contrastive 1 submisison covered in Section 3.1.2.

3.2 Unconstrained Setting

For the unconstrained setting in IWSLT 2023, an additional 60 hours of speech data with their corresponding transcriptions was made available by the organizers. This allowed for greater mono-lingual fine-tuning of the ASR data. Additionally, for both the ASR and MT components of all three of our submitted unconstrained systems, PLMs were used along with fine-tuning. The three submissions were cascade systems.

3.2.1 Primary System

The Primary System for the unconstrained setting consists of two systems, the ASR and the MT system. Both systems are fine-tuned. First, the ASR system is multi-lingual model pre-trained on the 102-language FLEURS (Conneau et al., 2023) dataset. The model consists of a conformer (Gulati et al., 2020) encoder and transformer decoder and is trained using hybrid CTC/attention loss (Watanabe et al., 2017) and hierarchical language identification conditioning (Chen et al., 2023). The model inputs are encoded representations extracted from a pre-trained XLS-R 128 model (Babu et al., 2021) with its weights frozen, augmented with SpecAug (Park et al., 2019) and speech perturbation (Ko et al., 2015). In order to jointly decode, we also trained an RNN language model. The RNN consists of 2 layers with a hidden size of 650, trained using SGD with a flat learning rate of 0.1. The word-error rate on the validation set was 15. For the MT system, we use the Fairseq (Ott et al., 2019) tool kit for translation. The Flores 101 model was used (Guzmán et al., 2019) as the PLM and is based on a transformer (Vaswani et al., 2017) architecture used at WMT 2021[2] by Facebook. Fine-tuning was performed using the training ASR+MT data from the *constrained* task as was used for training in the Constrained Contrastive 1 task in Section 3.1.2.

3.2.2 Contrastive 1 System

The Constrastive 1 system is nearly identical to the Primary System for the unconstrained setting. The MT system is identical to that of the Primary System submission for the unconstrained setting. For the ASR system, a FLEURS approach is used identical to the unconstrained Primary System in Section 3.2.1. The only difference is that this Unconstrained Contrastive 1 system does not use a language model.

[2] https://www.statmt.org/wmt21/large-scale-multilingual-translation-task.html

3.2.3 Contrastive 2 System

The Contrastive 2 System is also a cascade (ASR+MT) system. The MT system is identical to that of the Primary System submission for the unconstrained setting. The ASR system architecture is identical to the Constrained Contrastive 1 System in Section 3.1.2, but with other hyperparameters. In this experiment took 243 epochs to train, using Stochastic Gradient Descent (SGD) with Nesterov momentum and a minibatch of 16 utterances. The initial learning rate was set to 0.002 for faster convergence, and it was annealed with a constant factor of 1.2 after each epoch, with momentum set to 0. In this system, we add the additional 60 hours of monolingual transcribed speech data from the *unconstrained* setting mentioned in the IWSLT 2023 low-resource task in addition to the 1.6 hours provided for the *constrained* setting.

3.3 Other Approaches

As noted in Section 2, there have been other successful approaches worth visiting. While we could not exhaustively attempt to use all of those approaches, we did focus on several that are worth noting.

For ASR approaches, we focused on experimenting with different model architectures. This included using different encoders (transformer, conformer) and decoders (auto-regressive Transformer, CTC-only). Regardless, all of the ASR systems achieved at best 100 WER in the constrained setting, limiting the effectiveness of any cascaded approach. In the unconstrained setting, we also looked at different ways to incorporate pre-training. For example, we tried directly fine-tuning a pre-trained XLS-R model (Babu et al., 2021; Baevski et al., 2020) instead of using extracted layer-wise features from a frozen model. These approaches were somewhat more successful by achieving up to 20.4 WER on the validation set; however, the top three systems reported performed better with ASR.

For MT approaches, several attempts were made to experiment with other systems. For example, the OpenNMT (Klein et al., 2020) toolkit now offers PLMs that include the Flores 101 (Guzmán et al., 2019) dataset. However, since Quechua was not included in the language list, the performance was extremely low on the validation set (0.06 BLEU). The Hugging Face version of the Flores 200 dataset was also tested and resulted in 23.5 on its own data. However, when testing on the validation set, the score was of 6.27 BLEU. The Flores 200 model is made available as the NLLB task on Fairseq, however, we experienced several conflicts with the machine infrastructure causing complexity with the Stopes tokenization that prevented us from moving forward.

For direct ST approaches, we also were unsuccessful using w2v feature encoding without major modification. Overall, the cascade approaches seemed to work better for this task and, thus, we made a decision to use those instead. The results for the *constrained* task, nonetheless, show that the direct s2t approach worked well using MFB features.

4 Results and Discussion

Team **QUESPA** BLEU and CHRF Scores			
Constrained			
System	**Description**	**BLEU**	**CHRF**
primary	mfb+s2t	1.25	25.35
contrastive 1	w2vl+onmt	0.13	10.53
contrastive 2	conformer+onmt	0.11	10.63
Unconstrained			
System	**Description**	**BLEU**	**CHRF**
primary	fleurs+lm+floresmt	15.36	47.89
contrastive 1	fleurs+floresmt	15.27	47.74
contrastive 2	w2vl+floresmt	10.75	42.89

Table 1: Team QUESPA results for the Quechua to Spanish low-resource task at IWSLT 2023.

Results are presented in Table 1. For the constrained task, we were unable to create a system that would be viable for deployment. Notwithstanding, we believe that the primary submission which used MFB features along with the default Fairseq S2T recipe could be used to further research in the field. Other systems, based on w2vletter (Pratap et al., 2019) and a conformer (Gulati et al., 2020) resulted in a near zero BLEU score and are probably only valid as proof of the non-functional status of the two systems when performing ASR on the QUE–SPA language pair. It is clear that with 1.6 hours of data for training, few constrained systems will perform better than 5 BLEU, as seen in previous IWSLT tasks.

For the unconstrained setting, our findings have shown that for both the ASR and MT models, the use of a PLM with fine-tuning is necessary. We were unable to create a system from scratch that would perform as well as those presented in previ-

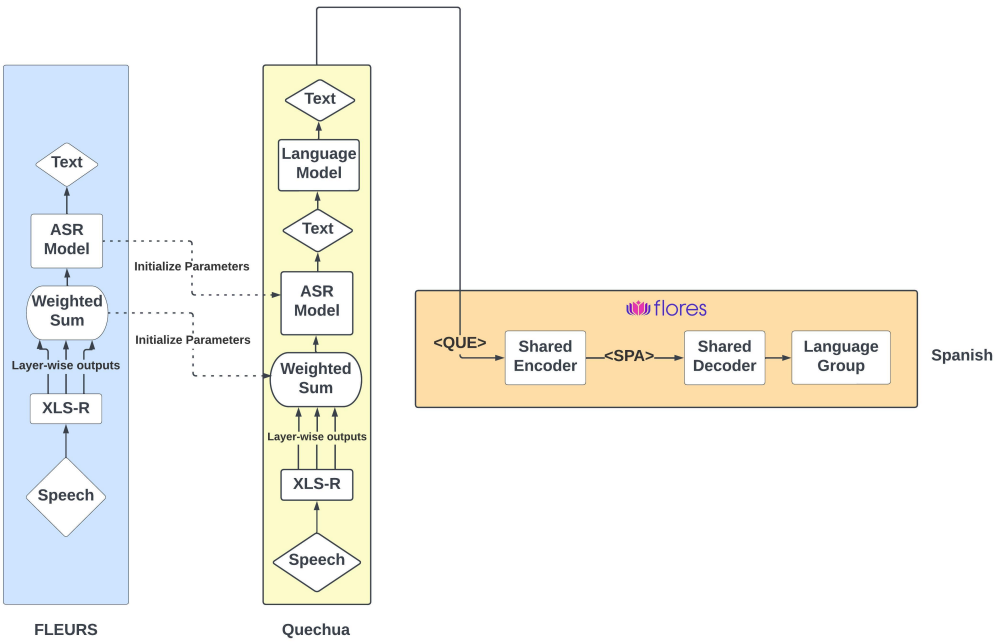

Figure 1: The best-performing *unconstrained* speech translation pipeline.

ous tasks. The combination of a language model and the FLEURS PLM for ASR along with the FLORES 101 PLM for MT constitutes our best performing system overall as shown in Figure 1. The language model slightly helped for the Primary system by a gain of nearly 0.10 points in BLEU. The other unconstrained system based on w2vletter (Pratap et al., 2019) performed much better than the constrained version making it worthwhile to explore for future iterations since it doesn't require other languages.

5 Conclusion

Concluding, we have experimented with several options for both the constrained and unconstrained settings. This constitutes the first time that experiments have been put together along with the other team submissions for the Quechua to Spanish task. We believe that the performance achieved here can serve as baselines for more sophisticated approaches. Additionally, it came to our attention that data splits provided by the organizers can be adjusted to better fit the data. There are multiple speakers in several of the audio files, we did not take advantage of this and hope to address it in the future. Also for the future, we believe that more work could be done using direct ST systems with fine-tuning. We did not follow that path in this work but feel it would be advantageous.

References

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Yannick Estève, Kevin Duh, Marcello Federico, Souhir Gahbiche, Barry Haddow, Benjamin Hsu, Phu Mon Htut, Hirofumi Inaguma, Dávid Javorský, John Judge, Yasumasa Kano, Tom Ko, Rishu Kumar, Pengwei Li, Xutail Ma, Prashant Mathur, Evgeny Matusov, Paul McNamee, John P. McCrae, Kenton Murray, Maria Nadejde, Satoshi Nakamura, Matteo Negri, Ha Nguyen, Jan Niehues, Xing Niu, Atul Ojha Kr., John E. Ortega, Proyag Pal, Juan Pino, Lonneke van der Plas, Peter Polák, Elijah Rippeth, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Yun Tang, Brian Thompson, Kevin Tran, Marco Turchi, Alex Waibel, Mingxuan Wang, Shinji Watanabe, and Rodolfo Zevallos. 2023. Findings of the IWSLT 2023 Evaluation Campaign. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*. Association for Computational Linguistics.

Antonios Anastasopoulos, Ondřej Bojar, Jacob Bremerman, Roldano Cattoni, Maha Elbayad, Marcello Federico, Xutai Ma, Satoshi Nakamura, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Alexander Waibel, Changhan Wang, and Matthew Wiesner. 2021. FINDINGS OF THE IWSLT 2021 EVALUATION CAMPAIGN. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 1–29, Bangkok, Thailand (online). Association for Computational Linguistics.

Anastasopoulos Antonios, Barrault Loc, Luisa Bentivogli, Marcely Zanon Boito, Bojar Ondřej, Roldano Cattoni, Currey Anna, Dinu Georgiana, Duh Kevin, Elbayad Maha, et al. 2022. Findings of the iwslt 2022 evaluation campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98–157. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, Matteo Negri, and Marco Turchi. 2021. Cascade versus direct speech translation: Do the differences still make a difference? *CoRR*, abs/2106.01045.

Alexandre Berard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. *CoRR*, abs/1612.01744.

Dan Berrebbi, Jiatong Shi, Brian Yan, Osbel López-Francisco, Jonathan Amith, and Shinji Watanabe. 2022. Combining Spectral and Self-Supervised Features for Low Resource Speech Recognition and Translation. In *Proc. Interspeech 2022*, pages 3533–3537.

Marcely Zanon Boito, Fethi Bougares, Florentin Barbier, Souhir Gahbiche, Loïc Barrault, Mickael Rouvier, and Yannick Estéve. 2022. Speech resources in the tamasheq language. *Language Resources and Evaluation Conference (LREC)*.

Ronald Cardenas, Rodolfo Zevallos, Reynaldo Baquerizo, and Luis Camacho. 2018. Siminchik: A speech corpus for preservation of southern quechua. *ISI-NLP 2*, page 21.

William Chen and Brett Fazio. 2021. Morphologically-guided segmentation for translation of agglutinative low-resource languages. *Proceedings of Machine Translation Summit XVIII*.

William Chen, Brian Yan, Jiatong Shi, Yifan Peng, Soumi Maiti, and Shinji Watanabe. 2023. Improving massively multilingual asr with auxiliary CTC objectives. *arXiv preprint arXiv:2302.12829*.

Ronan Collobert, Christian Puhrsch, and Gabriel Synnaeve. 2016. Wav2letter: an end-to-end convnet-based speech recognition system. *arXiv preprint arXiv:1609.03193*.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805.

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*.

Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. Language modeling with gated convolutional networks. In *International conference on machine learning*, pages 933–941. PMLR.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios Gonzales, Ivan Meza-Ruiz, et al. 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.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *The Journal of Machine Learning Research*, 22(1):4839–4886.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Proc. Interspeech 2020*, pages 5036–5040.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The flores evaluation datasets for low-resource machine translation: Nepali–english and sinhala–english. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR 2015, Conference Track Proceedings*.

Guillaume Klein, François Hernandez, Vincent Nguyen, and Jean Senellart. 2020. The opennmt neural machine translation toolkit: 2020 edition. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 102–109.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio augmentation for speech recognition. In *Proc. Interspeech 2015*, pages 3586–3589.

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

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega■, Annette Riosψ, Angela Fan, Ximena Gutierrez-Vasquesψ, Luis Chiruzzo, Gustavo A Giménez-Lugo, Ricardo Ramosη, et al. 2021. Findings of the americasnlp 2021 shared task on open machine translation for indigenous languages of the americas. *NAACL-HLT 2021*, page 202.

NLLB Team, 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.

Atul Kr Ojha, Valentin Malykh, Alina Karakanta, and Chao-Hong Liu. 2020. Findings of the loresmt 2020 shared task on zero-shot for low-resource languages. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 33–37.

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

John E Ortega and Krishnan Pillaipakkamnatt. 2018. Using morphemes from agglutinative languages like quechua and finnish to aid in low-resource translation. *Technologies for MT of Low Resource Languages (LoResMT 2018)*, page 1.

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 *NAACL (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.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

Juan Pino, Qiantong Xu, Xutai Ma, Mohammad Javad Dousti, and Yun Tang. 2020. Self-training for end-to-end speech translation.

Vineel Pratap, Awni Y. Hannun, Qiantong Xu, Jeff Cai, Jacob Kahn, Gabriel Synnaeve, Vitaliy Liptchinsky, and Ronan Collobert. 2019. Wav2letter++: A fast open-source speech recognition system. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6460–6464.

Annette Rios. 2015. *A basic language technology toolkit for Quechua*. Ph.D. thesis, University of Zurich.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

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.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020. fairseq s2t: Fast speech-to-text modeling with fairseq. *arXiv preprint arXiv:2010.05171*.

Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R. Hershey, and Tomoki Hayashi. 2017. Hybrid CTC/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253.

Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-Sequence Models Can Directly Translate Foreign Speech. In *Proc. Interspeech 2017*, pages 2625–2629.

Marion Weller-Di Marco and Alexander Fraser. Findings of the wmt 2022 shared tasks in unsupervised mt and very low resource supervised mt.