Findings of the Second AmericasNLP Competition on Speech-to-Text Translation

Abteen Ebrahimi[*] abteen.ebrahimi@colorado.edu
Manuel Mager[*] manuel.mager@ims.uni-stuttgart.de
Adam Wiemerslage[*] adam.wiemerslage@colorado.edu
Pavel Denisov[*] pavel.denisov@ims.uni-stuttgart.de
Arturo Oncevay a.oncevay@ed.ac.uk

Danni Liu[†], **Sai Koneru, Enes Yavuz Ugan, Zhaolin Li, Jan Niehues**
KIT danni.liu@kit.edu

Monica Romero[†], **Ivan G Torre**
MR&IG monica.romero@alumnos.upm.es

Tanel Alumäe[†], **Jiaming Kong**
TalTech tanel.alumae@taltech.ee

Sergey Polezhaev[†], **Yury Belousov**
Team-name sergey.polezhaev@outlook.com

Wei-Rui Chen[†], **Peter Sullivan, Ife Adebara, Bashar Talafha, Alcides Alcoba Inciarte, Muhammad Abdul-Mageed**
UBC-DL NLP weirui.chen@ubc.ca

Luis Chiruzzo luischir@fing.edu.uy
Rolando Coto-Solano rolando.a.coto.solano@dartmouth.edu
Hilaria Cruz layacruz@gmail.com
Sofía Flores-Solórzano sofia.flores.s@gmail.com
Aldo Andrés Alvarez López aldo.alvarez@fiuni.edu.py
Ivan Meza-Ruiz ivanvladimir@gmail.com
John E. Ortega j.ortega@northeastern.edu
Alexis Palmer alexis.palmer@colorado.edu
Rodolfo Zevallos rodolfojoel.zevallos@upf.edu
Kristine Stenzel kris.stenzel@colorado.edu
Thang Vu thang.vu@ims.uni-stuttgart.de
Katharina Kann katharina.kann@colorado.edu

Editors: Marco Ciccone, Gustavo Stolovitzky, Jacob Albrecht

[*] Equal contribution
[†] Corresponding author for participating teams. Teams sorted alphabetically.

Abstract

Indigenous languages, including those from the Americas, have received very little attention from the machine learning (ML) and natural language processing (NLP) communities. To tackle the resulting lack of systems for these languages and the accompanying social inequalities affecting their speakers, we conduct the second AmericasNLP competition (and the first one in collaboration with NeurIPS), which is centered around speech-to-text translation systems for Indigenous languages of the Americas. The competition features three tasks – (1) automatic speech recognition, (2) text-based machine translation, and (3) speech-to-text translation – and two tracks: constrained and unconstrained. Five Indigenous languages are covered: Bribri, Guarani, Kotiria, Wa'ikhana, and Quechua. In this overview paper, we describe the tasks, tracks, and languages, introduce the baseline and participating systems, and end with a summary of ongoing and future challenges for the automatic translation of Indigenous languages.

Keywords: natural language processing, machine translation, speech-to-text translation, automatic speech recognition, Indigenous languages, low-resource machine translation, low-resource languages

1. Introduction

Over the last decade, the field of natural language processing (NLP) has seen incredible progress, in large part due to the advent of deep learning models and pretraining as a transfer learning technique. However, these recent advancements require huge amounts of data, which make them inapplicable to languages with limited amounts of resources. For example, multilingual transformer models such as mBERT (Devlin et al., 2019) or XLM-R (Conneau et al., 2020) cover only around 100 languages – a tiny fraction of Earth's roughly 7000 languages. Similarly, Google Translate only supports 133 languages as of March 2023. This undesirable situation reinforces existing social inequalities. For example, speakers of high-resource languages such as French or German can easily get access to large amounts of information using the internet in combination with existing translation systems between English and their languages. However, speakers of many other languages cannot make use of most of that information, unless they also speak a high-resource language.

The Indigenous languages of the Americas are part of the large set of languages that have traditionally received little attention from the NLP community. Hence, for many NLP tasks and Indigenous languages, no systems are available. Furthermore, since Indigenous languages often differ from high-resource languages not only with regards to the available amounts of data but also with respect to their typology, it is unclear how well existing approaches can be applied. For instance, many Indigenous languages are polysynthetic or tonal, properties rarely found in high-resource European languages. Additionally, Indigenous researchers are underrepresented in the machine learning (ML) and NLP communities, which hinders evaluation and, thus, development of models even further.

This situation highlights the need for language technologies with real-world applicability to languages that are currently underrepresented. Our goal in this competition is to encourage and benchmark the development of speech-to-text translation systems for several Indigenous languages of the Americas, and, in doing so, increase the visibility of these, and other, Indigenous languages in the ML community. To this end, we have collected datasets for five language pairs (Bribri–Spanish, Guarani–Spanish, Kotiria–Portuguese, Wa'ikhana–Portuguese, and Quechua–Spanish), which are described in §3, and present three separate subtasks: (i) automatic speech recognition (ASR) (ii) machine translation (MT) (iii) speech-to-text translation (S2TT). We set out to explore two main questions: (1) **Which models or techniques are appropriate for speech-to-text translation of our five Indigenous languages into high-resource languages?** and (2) **Which models or techniques work well for speech-to-text translation in the low-resource setting?** Our hope is that the availability of translation systems for Indigenous languages of the Americas will have a huge impact in terms of reducing social inequalities by providing everyone with equal access to information, increasing avenues of communication for monolingual speakers of Indigenous languages, and supporting documentation and/or revitalization of endangered Indigenous languages.

This competition is the second iteration of the AmericasNLP Shared Task (Mager et al., 2021). Of the 10 languages featured in the 2021 edition, we repeat three, while introducing two new ones.

2. Tasks and Tracks

The competition consists of one main task, speech-to-text translation, and its two subtasks, automatic speech recognition and machine translation. Translation tasks involve translating from an Indigenous language to a high-resource language, either Spanish or Portuguese. ChrF scores (Popović, 2015) are used to evaluate translation performance, in order to remain consistent with prior work (Mager et al., 2021). The ASR task involves automatic recognition of speech in an Indigenous language, and is evaluated using character error rate (CER), as word boundaries are often not standardized, and the languages have a rich morphology. Teams are asked to evaluate their models on all 5 languages, and an average score is used to determine the final order.

Each task is itself comprised of two tracks: an unconstrained track (Track 1) and a constrained track (Track 2). In **Track 1**, all external data that the participants can collect – with the exception of the datasets used for evaluation – is valid to use as training data. The aim for this track is for teams to achieve the highest possible performance for the given languages. In **Track 2**, teams are only allowed to use the provided supervised training sets and any additional monolingual Spanish or Portuguese data they collect. For this track, the motivation is to push for the development and evaluation of novel models or strategies, e.g., approaches for data augmentation, as opposed to simply scaling the amount of data available. Both tracks allow for the use of established pretrained and multilingual models, regardless of if they have been trained on the target languages or not. Teams are allowed to submit models which satisfy the rules for Track 2 to both tracks if they wish.

3. Languages and Data

In this section we describe the languages and audio data used during the competition. Bribri, Guarani, and Quechua are translated to Spanish, while Kotiria and Wa'ikhana are translated to Portuguese. The ethics statement for the data can be found in Appendix B.

3.1. Bribri

Bribri, part of the Chibchan family, is a language spoken in southern Costa Rica by around 7,000 people (INEC, 2011). It is a tonal language following SOV word order. The

ISO	Family	MINUTES				INSTANCES			
		TRAIN	DEV	TEST	TOTAL	TRAIN	DEV	TEST	TOTAL
bzd	Bribri	29.09	11.66	41.93	82.69	495	250	1001	1746
gn	Guarani	19.39	7.36	12.99	39.74	293	93	160	546
gvc	Kotiria	161.95	17.91	77.60	257.46	1984	254	1001	3239
pir	Wa'ikhana	93.32	12.74	58.27	164.33	1419	250	1001	2670
quy	Quechua	100.29	124.55	129.79	354.63	573	250	415	1238

Table 1: Description of languages and available data.

Bribri data we use for the competition is taken from the Pandialectical Corpus of the Bribri Language,[1] which was collected through documentary fieldwork between 2013 and 2017. The corpus includes recordings of spontaneous speech, including stories and narration in the three major dialects of Bribri: Amubri, Coroma, and Salitre. We use slightly more than an hour of audio data and transcriptions.

3.2. Guarani

Guarani belongs to the Tupi-Guarani family and is spoken by between 6 and 10 million speakers, mainly in Paraguay but also in Bolivia, Argentina, and Brazil. It is a polysynthetic and agglutinative language with a very complex verbal and noun morphology. Guarani has a distinct set of oral (a, e, i, o, u, y) and nasal (ã, ẽ, ĩ, õ, ũ, ỹ) vowels that, when used within a word, affect the surrounding phonemes, as there must be harmonization in nasal or oral pronunciation (Academia de la Lengua Guaraní, 2018), which should be taken into account when doing speech recognition and transcription. Another challenge in Guarani transcription is the frequent use of the glottal stop, which in Guarani is considered a consonant called *puso* and written as ’.

We use the Guarani speech and transcription data from Mozilla Common Voice,[2] created by volunteers that record themselves speaking a series of sentences. The dataset contains 1883 spoken Guarani sentences together with their transcriptions, totalling about 2.3 hours of audio.[3] While some of these sentences were verified by volunteers in the platform, the rest were manually validated by competition organizers. For the competition, the dataset is translated to Spanish mainly by students in the Guarani-Spanish Bilingualism Program at Facultad de Humanidades in Universidad Nacional de Itapúa.

3.3. Kotiria and Wa'ikhana

Kotiria and Wa'ikhana are two closely related languages which together form a branch within the East Tukano language family, spoken in northwest Amazonia, near the border between Brazil and Colombia. There are around 1,500 to 2,000 speakers of each language. The languages are synthetic and agglutinative and rely heavily on suffixing morphology.

[1] http://bribri.net

[2] https://commonvoice.mozilla.org/

[3] Due to organizer miscommunication, only a subset of the data was released for the competition, however the full dataset is now available.

While subject position can vary, the languages are generally head-final. The data used is the result of documentary fieldwork, and consists of short stories. For the competition we use a combined 7 hours of existing audio data and translations.

3.4. Quechua

Quechua is an Indigenous language with several million speakers, mainly concentrated in Peru. However, there are also millions of speakers found in other countries such as Bolivia, Ecuador, Chile, and even Argentina. Its morphology can be considered both polymorphic and agglutinative and has been studied previously in the context of machine translation (Ortega and Pillaipakkamnatt, 2018; Ortega et al., 2020, 2021). Quechua is broken down into two main divisions (Quechua I and II) which are centered in Peru's mountainous region. The corpus that we present contains two main regional divisions spoken in Ayacucho, Peru (Quechua Chanka ISO: `quy`) and Cusco, Peru (Quechua Collao ISO: `quz`), both part of Quechua II. We take data from the Siminchik dataset (Cardenas et al., 2018), which consists of recorded radio conversations transcribed by volunteers, and translate them to Spanish for this competition.

4. Baselines

4.1. Speech Recognition

The competition's speech recognition baseline is implemented using the ESPnet2 toolkit (Watanabe et al., 2021) and relies on the XLS-R-300M pretrained model (Baevski et al., 2020a; Babu et al., 2021). All audios are encoded in 16 kHz mono WAVE format. Training data is augmented with 3-factor speech perturbation (Ko et al., 2015). Our model uses a weighted sum of the XLS-R layer outputs (Yang et al., 2021; Chang et al., 2021), followed by one self-attention encoder layer (Vaswani et al., 2017) with 8 heads and a dimension of 256. The XLS-R model parameters are kept frozen during training, except for the last two layers. The total number of model parameters is ~ 322M, out of which ~ 40M parameters are trainable. We train a separate model for each language. The output vocabulary consists of 100 subwords created using the Unigram (Kudo, 2018) version of SentencePiece (Kudo and Richardson, 2018), trained on the train set transcriptions.

Training is performed with Connectionist Temporal Classification (CTC; Graves et al., 2006) loss for 15 epochs using the AdamW optimizer (Loshchilov and Hutter, 2019) and a WarmupLR scheduler with a maximum learning rate of 10^{-4} and 300 warmup steps. The batch size is set to about 12 seconds of audio, or between 2 and 4 instances on average. It takes between 30 and 160 minutes to train each model on one NVIDIA GeForce GTX 1080 Ti GPU. The checkpoint with the lowest validation CER is used for decoding using beam search with a beam of size 5.

4.2. Translation

Machine Translation We follow the **random babbling** system from Bollmann et al. (2021), which they found to perform surprisingly well on character-level evaluation metrics. The system considers only the trigram distribution of a given target language and the lengths of the source sentences.

First, we compute the distribution of character trigrams in the target language training data $p(t)$. Then, we compute the *length ratio*, ϕ, of target trigrams to source trigrams across training pairs. At test time, we compute the number of trigrams in the source sentence $|s|$, and sample the top-n target trigrams from $p(t)$, where $n = \phi|s|$. Finally, we follow Bollmann et al. (2021) in implementing four heuristics to adjust the trigram sequence: (i) trigram ordering is randomized, (ii) consecutive whitespaces are reduced to a single whitespace, (iii) sentence-initial characters are uppercased, and non-word-initial characters are lowercased, and (iv) sentence-final punctuation from the source sentence is copied to the target.

Speech-to-Text Translation We use the same system as the MT baseline for S2TT, except that the source lengths are the duration of the wav file since the source language data is strictly audio.

5. Winning Submissions

In this section we describe the best performing systems for the ASR and MT tasks, and all submissions for the S2TT Task.

5.1. KIT

5.1.1. Automatic Speech Recognition (ASR)

Data Augmentation The KIT system builds upon the baseline data augmentation, which uses 3-factor speech perturbation (Ko et al., 2015) – effectively tripling the training data by adding utterances with 0.9 and 1.1 times the original speed. While a higher factor typically does not further improve performance, considering the extreme low resource condition, KIT further adds utterances with speed factors of 0.8 and 1.2. On average, this brings a 5.3% relative reduction in CER over the 5 languages. Moreover, they replace the original byte-pair-encoding with character-based outputs, which can be seen as an implicit form of data augmentation. With character-based targets, the loss is calculated over more positions, thereby providing the model with more training signals. This more granular output representation reduces relative CER by 5.7% on average. SpecAugment (Park et al., 2019) is also used on the wav2vec outputs, which gives another 5.7% relative CER reduction.

Scaling The baseline has most parameters of wav2vec frozen, except for the last two of its 12 layers. The KIT system trains all wav2vec parameters except the convolutional front-end. Unfreezing the wav2vec parameters brings the largest relative CER reduction of 13.7% on top of the improvements described previously. Further scaling up model size does not yield favorable results. Likewise, finetuning a recent large multilingual ASR model (Pham et al., 2022) based on wav2vec and the mBART decoder (Liu et al., 2020) also does not outperform the previous configuration.

5.1.2. Machine Translation (MT)

Finetuning Pretrained Models KIT's MT model is finetuned from the BASE configuration of DeltaLM (Ma et al., 2021), a pretrained encoder-decoder model with 256k subwords, 12 encoder layers, and 6 decoder layers. For potential knowledge transfer and simplicity in implementation, the multilingual model is trained on the 5 language pairs jointly. Only for the highest-resourced pair, Quechua-Spanish, is the model able to generate meaningful translations. As such, other translation methods explained below are explored.

Nearest-Neighbor Search Inspired by early retrieval-based spoken word recognition systems (*inter alia* Sakoe and Chiba, 1978), a similar approach is applied to MT. For each source sentence in the test set, the system goes through the training and development source sentences to find its nearest neighbour, where the distance is measured by character-level edit distance. After retrieving the nearest source sentence, the team uses its target side as output translation. This approach substantially outperforms the neural approach.

Data Augmentation For the *unconstrained* setup, the NLLB-200 (NLLB Team et al., 2022) is used to back-translate monolingual data from Spanish to Guarani and Quechua, the only two directions supported. The monolingual data is taken from the Spanish training data of the MuST-C dataset (Cattoni et al., 2021), which contains speech translation of TED talks. This creates around 300K back-translated sentences for each of the two directions. For Bribri–Spanish, Kotiria–Portuguese, and Wa'ikhana–Portuguese, parallel data is scraped from religious texts to train back-translation models, in the hope of utilizing additional Spanish and Portuguese monolingual data. However, in this case the additional back-translated data degrades performance. Potential reasons for this degradation include a strong domain mismatch and poor quality of the scraped data.

Prompting As large language models have shown promising results with few-shot learning, KIT explores prompting language models for translation. First, the team extracts word alignments from the training data. Then they finetune a Portuguese GPT-2 model (Guillou, 2020) with the word alignments formed as "source | word-to-word translations | target". At inference time, the model is prompted to generate the target translation with a source sentence and word-to-word translations as prefix. This approach, however, does not yield meaningful translations, which may be due to the poor word alignment quality and poor model adaptation to the source language. No prompting systems are included in the final submission.

5.1.3. Speech-to-Text Translation

The speech-to-text translation system is a cascade of the ASR and MT models above. While the two comparatively higher-resourced languages (Quechua/Guarani–Spanish) benefit from the unconstrained data condition, ASR coupled with nearest-neighbor-based MT gives the strongest results for the other three pairs.

5.2. MR&IG

5.2.1. Automatic Speech Recognition

The MR&IG team finetunes wav2vec2.0, a pretrained semi-supervised model. Two configurations, wav2Vec2-XLS-R-300M and wav2Vec2-XLS-R-1B, are considered, differing in the number of parameters and training process, but using the same amount of data during pretraining. In addition to the original training data, MR&IG collect 0.9 hours of transcribed speech for Bribri (Carla Victoria Jara Murillo, 2018), 7 hours for Kotiria (Faith Comes By Hearing) and 2.8 hours for Quechua (Matthew Brown, 2020). No external data is collected for Guarani and Wa'ikhana. The team also applies SpecAugment (Park et al., 2019) during training and collects additional text for training n-gram language models. The text includes speech transcriptions, online texts, and books. In all cases, they collect less than 100k words. Two model checkpoints are considered during training: one with the lowest loss, and one with the lowest WER on the validation set. During testing, the checkpoint with lowest loss obtained the best performance for all languages. For Kotiria, the best result comes from the 1B parameter wav2vec2.0 model, while for all the other languages the 300M parameter model performs best.

For decoding, 3-gram and 4-gram language models are trained for all languages and beam search is performed on the validation set. Beam search hyperparameters are selected based on Bayesian optimization. However, due to the lack of a standard normalization of the transcriptions and the low amount of data, this optimization does not lead to significant improvement and degrades performance for some languages. Therefore, the final decoding strategy is based on greedy search and heuristic corrections applied to correct textual errors such as capitalization, punctuation and reducing multiple spaces or letters.

5.3. TalTech

5.3.1. Automatic Speech Recognition

For all languages, the TalTech team finetunes the XLS-R-2B wav2vec2.0 model (Baevski et al., 2020a; Babu et al., 2021), which is pretrained on 500 000 hours of multilingual data using a self-supervised objective. Finetuning is performed by adding an output layer to the wav2vec2.0 model that corresponds to the character vocabulary of the particular language. The model is then trained with the provided ASR training data using the CTC objective. An effective batch size of 30 minutes is used, and the model is trained for 5000 updates. During the first 50 updates, only the output layer is trained. Heavy feature-space spectral masking and stochastic layer dropping are used.

The TalTech team also explores several data augmentation strategies. In the first strategy, the training data is perturbed using reverberation and mixing with background noises and music. Background noises from the Freesound portion of the MUSAN corpus (Snyder et al., 2015) and simulated small, medium and large room impulse responses (Ko et al., 2017) are used for data augmentation. As an alternative augmentation method, aligned data augmentation (ADA; Lam et al., 2021) is explored, where transcribed words and the corresponding speech segments within an utterance are replaced with randomly sampled words from other utterances, based on word alignment information obtained from a model trained without any augmentation. Specifically, the training dataset is replicated ten times, where 50% of the utterances are perturbed with a word replacement probability of 20%. However, the performance of the ASR model does not exhibit significant improvements with the use of any of the investigated augmentation techniques, as compared to their baseline.

5.3.2. Speech-to-text Translation

The TalTech system for speech-to-text translation uses a cascaded approach, where the source language utterances are first transcribed using the ASR model introduced in the previous section and then translated to the target language using a text-based machine translation model. The machine translation approach is based on finetuning the large-scale multilingual translation model produced from the No Language Left Behind (NLLB) project

(NLLB Team et al., 2022). The NLLB model is trained on carefully selected datasets, partially professionally translated within this project, and additionally uses a novel parallel text mining method to create hundreds of millions of aligned training sentences for low-resource languages. NLLB can deliver translations directly between any pair of over 200 languages. Of the competition languages, Guarani and Quechua Ayacucho are supported by NLLB.

Specifically, the NLLB-3.3B model is finetuned using the provided MT training data, for each language separately. Finetuning is performed with 500 updates, using an effective batch size of 16 sentence pairs. NLLB uses dedicated prefix tokens in source and target sentences to indicate the current language. For languages not covered by the NLLB model, the prefix token corresponding to Guarani (i.e., the Bribri–Spanish translation model is effectively finetuned from the Guarani–Spanish NLLB model) is used. Decoding is done using a beam search with size 16. For Quechua–Spanish and Guarani–Spanish, NLLB-3.3B has a relatively good MT performance out-of-the-box, but finetuning the model improves results further (around 25% relative improvement).

In addition to the NLLB-based model, the team explores using an IBM-style statistical MT model, trained for the three languages that are not covered by the NLLB model. However, the results obtained with the statistical model are significantly inferior to those obtained with the NLLB-based model, with around 30% lower relative chrF scores.

5.4. team-name

5.4.1. Speech-to-text Translation

The team-name system for speech-to-text translation task is based on the Whisper model (Radford et al., 2022), known to be effective in handling a wide range of languages and audio qualities. The model is finetuned on the training data for 20 epochs. Hyperparameters are manually tuned, with a final learning rate of 0.0001, weight decay of 0.01, and Adam epsilon of 1e-8. The team also experiments with cascading various SOTA models for ASR, including wav2vec2.0 (Baevski et al., 2020b) and Conformer-T (Gulati et al., 2020) with MT models. However, this approach does not outperform Whisper.

5.5. UBC DL-NLP

5.5.1. Machine Translation

The UBC DL-NLP system for MT is based on finetuning mBART50 Tang et al. (2020), a multilingual MT model pretrained on 50 languages. For each of the five language pairs, an independent model is trained. Each model is trained on the respective training set with a batch size of 20 for $15,000$ updates without early stopping. A beam size of 6 is used for decoding. To pick the best checkpoint, model performance is evaluated every 50 updates against the development set. The checkpoint with the least development loss is used for inference on the test set. The team also experiments with a statistical machine translation model, using KenLM (Heafield, 2011) and Giza++ (Och and Ney, 2003), as well as a bidirectional LSTM. However, they find that it does not improve over the neural model.

Task	Track	Team	bzd	gn	gvc	pir	quy	Avg.
ASR	1	MR&IG	34.70	15.59	36.59	35.23	12.14	26.85
		TalTech	37.01	18.04	37.42	34.37	11.76	27.72
		Karya-MSRI	44.77	20.94	42.64	37.80	18.74	32.98
		Baseline	51.14	30.57	45.00	43.94	26.12	39.35
		team-name	93.44	96.25	90.74	56.40	79.26	83.22
	2	TalTech	37.01	18.04	37.42	34.37	11.76	27.72
		KIT	36.09	13.89	47.19	36.92	12.70	29.36
		Baseline	51.14	30.57	45.00	43.94	26.12	39.35
		Factored AI	51.25	33.51	52.39	43.61	23.70	40.89
		NSU	62.16	32.15	45.33	41.91	45.89	45.49
		team-name	53.33	41.78	57.48	56.40	37.56	49.31
MT	1	KIT	71.43	17.05	29.32	22.25	40.29	36.07
		Baseline	38.46	20.18	28.85	28.55	28.55	28.92
	2	UBC DL-NLP	59.38	20.71	32.18	56.45	42.85	42.31
		Baseline	38.46	20.18	28.85	28.55	28.55	28.92
S2TT	1	KIT	71.43	27.60	45.04	49.68	48.89	48.53
		team-name	50.00	27.07	37.62	68.14	33.92	43.35
		Baseline	38.46	20.18	28.85	28.55	28.55	28.92
	2	KIT	71.43	17.05	45.04	49.68	44.63	45.57
		TalTech	51.23	25.28	29.96	24.12	30.30	32.18
		Baseline	38.46	20.18	28.85	28.55	28.55	28.92

Table 2: Main results of the competition. The ASR scores are measured using CER while MT and S2TT are measured using chrF.

5.6. Summary of Findings

For ASR, simple data augmentation methods, such as further adding utterances with more speed factors, improve performance over the baseline, while more complicated approaches, such as adding additional background noise or replacing random words, do not seem to help. Allowing for more fine-grained outputs and training more parameters of the model also improve performance. However, using a larger model does not always lead to reliable improvements. Teams using SpecAugment find it to be helpful. The best performing systems all use a version of XLS-R wav2vec2.0.

MR&IG perform best in the unconstrained track, finding additional data to be helpful in languages for which it is available. They report a 12.5 reduction in CER over the baseline. Despite the extra data for Quechua, TalTech, using a larger model with twice the parameters, outperform MR&IG without any external data. Indeed the TalTech system performs best in the constrained track with an absolute difference in CER of less than one point with respect to the unconstrained MR&IG system.

For MT, all teams experiment with finetuning pretrained multilingual transformer models, which is found to work best for the highest-resourced languages, Guarani and Quechua, despite the fact that we provide fewer translation instances than for other languages. Teams explore joint finetuning on all languages, as well as creating individual models. We find that traditional SMT or vanilla encoder-decoder models underperform the pretrained multilingual models, however a model motivated by retrieval-based word recognition offers stronger performance for the more under-resourced languages. In Bribri, for example, KIT report an absolute increase in chrF of 33.03 over the baseline. UBC DL-NLP, the best performing MT system overall, increase over the baseline for every language by finetuning mBART. The largest increase in accuracy is for Wa'ikhana: 27.9 absolute chrF over the baseline, which in turn outperforms the KIT system.

For S2TT, the finetuned Whisper model by team-name comfortably outperforms the baseline in the unconstrained track, with particularly strong results for Wa'ikhana. The cascaded system of KIT performs best on average for both tracks. TalTech also cascade their ASR and MT systems for S2TT, and outperform the baseline by a small margin.

6. Ongoing Challenges and Limitations

Speech-to-text translation for Indigenous languages of the Americas is a difficult task with multiple challenges, the first being the initial data collection. Recording and collecting speech data is time consuming and requires a close working relationship with native speakers. Furthermore, after recordings are collected, the transcription process is arduous and can be a bottleneck in the pipeline. This costly data collection step, along with potentially noisy and otherwise problematic recording conditions, results in small usable datasets for these languages. Translating transcriptions to a high-resource language adds further costs to the collection process. The second challenge is that the languages we feature are linguistically distant from almost all medium or high resource languages, preventing us from easily leveraging multilingual models through cross-lingual transfer. While the models submitted to the competition show improvements over the baseline, the raw performance is still far from an acceptable quality to be applicable for real-world use, even though the systems represent an important step towards achieving this goal.

7. Conclusion

The Second AmericasNLP Competition on Speech-to-Text Translation was organized with the goal of encouraging research on the technology necessary for automatic translation of languages with a strong oral tradition and to move Indigenous languages from the Americas into the focus of the ML, NLP, and Speech communities. In this overview paper, we described the tasks, tracks, and languages, as well as baseline and submitted systems. Most systems outperformed their respective baselines, and in some cases we saw a relative error reduction of more than 50%. We ended this paper with a description of ongoing and future challenges related to the translation of Indigenous languages from the Americas and hope that this competition will increase the visibility of and participation in research in this area.

References

Academia de la Lengua Guaraní. *Gramática Guaraní*. 2018.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. XLS-R: Self-supervised cross-lingual speech representation learning at scale, 2021. URL https://arxiv.org/abs/2111.09296.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020a. URL https://arxiv.org/abs/2006.11477.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020b.

Marcel Bollmann, Rahul Aralikatte, Héctor Murrieta Bello, Daniel Hershcovich, Miryam de Lhoneux, and Anders Søgaard. Moses and the character-based random babbling baseline: CoAStaL at AmericasNLP 2021 shared task. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 248–254, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.americasnlp-1.28. URL https://aclanthology.org/2021.americasnlp-1.28.

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

Alí García Segura Carla Victoria Jara Murillo. Portal de la lengua bribri SE'IE. Centro virtual de recursos para el estudio y la promoción de la lengua bribri, 2018.

Roldano Cattoni, Mattia Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. Must-c: A multilingual corpus for end-to-end speech translation. *Comput. Speech Lang.*, 66:101155, 2021. doi: 10.1016/j.csl.2020.101155. URL https://doi.org/10.1016/j.csl.2020.101155.

Xuankai Chang, Takashi Maekaku, Pengcheng Guo, Jing Shi, Yen-Ju Lu, Aswin Shanmugam Subramanian, Tianzi Wang, Shu-wen Yang, Yu Tsao, Hung-yi Lee, et al. An Exploration of Self-Supervised Pretrained Representations for End-to-End Speech Recognition. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 228–235. IEEE, 2021.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Faith Comes By Hearing. bible.is.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376, 2006.

Pierre Guillou. Gportuguese-2 (portuguese gpt-2 small): a language model for portuguese text generation (and more nlp tasks...). 2020.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

Kenneth Heafield. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, July 2011. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W11-2123.

INEC. Población total en territorios indígenas por autoidentificación a la etnia indígena y habla de alguna lengua indígena, según pueblo y territorio indígena. In Instituto Nacional de Estadística y Censos, editor, *Censo 2011*. 2011. URL https://admin.inec.cr/sites/default/files/media/resocialcenso2011-06.xls_2.xls.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. Audio augmentation for speech recognition. In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, pages 3586–3589. ISCA, 2015. URL http://www.isca-speech.org/archive/interspeech_2015/i15_3586.html.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur. A study on data augmentation of reverberant speech for robust speech recognition. In *ICASSP*, 2017.

T. Kudo. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. In *ACL*, pages 66–75, 2018.

T. Kudo and J. Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *EMNLP*, pages 66–71, 2018.

Tsz Kin Lam, Mayumi Ohta, Shigehiko Schamoni, and Stefan Riezler. On-the-fly aligned data augmentation for sequence-to-sequence ASR. In *Interspeech*, 2021.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Trans. Assoc. Comput. Linguistics*, 8:726–742, 2020. doi: 10.1162/tacl_a_00343. URL https://doi.org/10.1162/tacl_a_00343.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, and Furu Wei. Deltalm: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders. *CoRR*, abs/2106.13736, 2021. URL https://arxiv.org/abs/2106.13736.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.americasnlp-1.23. URL https://aclanthology.org/2021.americasnlp-1.23.

Karen Tucker Matthew Brown. Data from Quipu Project (12-2018), 2020.

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. No language left behind: Scaling human-centered machine translation, 2022. URL https://arxiv.org/abs/2207.04672.

Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

John E Ortega and Krishnan Pillaipakkamnatt. 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, 2018.

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

John E Ortega, Richard Alexander Castro Mamani, and Jaime Rafael Montoya Samame. Love thy neighbor: Combining two neighboring low-resource languages for translation. *Proceedings of Machine Translation Summit XVIII*, 2021.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. Specaugment: A simple data augmentation method for automatic speech recognition. In Gernot Kubin and Zdravko Kacic, editors, *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 2613–2617. ISCA, 2019. doi: 10.21437/Interspeech.2019-2680. URL https://doi.org/10.21437/Interspeech.2019-2680.

Ngoc-Quan Pham, Alexander Waibel, and Jan Niehues. Adaptive multilingual speech recognition with pretrained models. In Hanseok Ko and John H. L. Hansen, editors, *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, pages 3879–3883. ISCA, 2022. doi: 10.21437/Interspeech.2022-872. URL https://doi.org/10.21437/Interspeech.2022-872.

Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3049. URL https://aclanthology.org/W15-3049.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2022.

H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978. doi: 10.1109/TASSP.1978.1163055.

David Snyder, Guoguo Chen, and Daniel Povey. MUSAN: A Music, Speech, and Noise Corpus, 2015. arXiv:1510.08484v1.

Y. Tang, C. Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual translation with extensible multilingual pretraining and finetuning. *ArXiv*, abs/2008.00401, 2020.

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

Shinji Watanabe, Florian Boyer, Xuankai Chang, Pengcheng Guo, Tomoki Hayashi, Yosuke Higuchi, Takaaki Hori, Wen-Chin Huang, Hirofumi Inaguma, Naoyuki Kamo, Shigeki Karita, Chenda Li, Jing Shi, Aswin Shanmugam Subramanian, and Wangyou Zhang. The 2020 ESPnet Update: New Features, Broadened Applications, Performance Improvements, and Future Plans. In *2021 IEEE Data Science and Learning Workshop (DSLW)*, pages 1–6. IEEE, 2021.

Shu-Wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. SUPERB: Speech processing Universal PERformance Benchmark. *Proc. Interspeech 2021*, pages 1194–1198, 2021.

Appendix A. Acknowledgments

We would like to thank Roque Helmer Luna-Montoya (Academia Mayor de la Lengua Quechua in Cusco, Perú), Abel Ancalle Wilcha (court-certified freelance translator, Cusco, Perú), and Roxana Quispe Collantes (Universidad Nacional de San Antonio Abad del Cusco, Perú) for their help in translating the Quechua data. We would also like to thank the students at Facultad de Humanidades in Universidad Nacional de Itapúa for their translations of Guarani.

Appendix B. Ethics Statement

The Pandialectal Corpus of the Bribri Language was collected with grants from UAM-Santader Interuniversity Cooperation and University of Costa Rica. Corpus collection was conducted according to IRB approval from UAM - Autonomous University of Madrid and is the result of collaborative work between the language community and school teachers. The interviewed speakers signed an informed consent form for the data to be used for revitalization, educational, and scientific purposes.

For Guarani, the Mozilla Common Voice volunteers were given the option of optionally indicating their demographic information, but everything in this dataset is anonymized. Some of these sentences were verified by volunteers in the platform, and the rest were manually validated by competition organizers. Nine students participated in the translation process: eight Bilingualism students and one Computer Engineering student who is a native Guarani speaker. Each was given between 200 and 300 sentences to translate, taking between two and six weeks to complete the translations. All students are intermediate-advanced level speakers of Guarani, ensuring a high quality of translation.

The Kotiria and Wa'ikhana collections are the result of more than twenty years of fieldwork through grants to Kristine Stenzel from the Endangered Languages Foundation, the Wenner-Gren Foundation for Anthropological Research, the National Science Foundation, the National Endowment for the Humanities, the Hans Rausing Endangered Languages Project (ELDP), and the Brazilian National Council for Scientific and Technological Development-CNPq. All research was undertaken following ethical protocols and with full IRB approvals from Dr. Stenzel's academic institutions (University of Colorado, Federal University of Rio de Janeiro) and the Brazilian authorities: CNPq and FUNAI, the Brazilian National Foundation for Indigenous peoples. The documentation effort is the result of collaborative work between the language communities, who have granted permission for its use for revitalization, educational, and scientific purposes.