INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

CoVoST 2 and Massively Multilingual Speech Translation

Changhan Wang[⋆]*, Anne Wu*[⋆]*, Jiatao Gu, Juan Pino*[⋆]

Facebook AI, USA

{changhan,jgu,juancarabina}@fb.com, annewu018@gmail.com

Abstract

Speech translation (ST) is an increasingly popular topic of research, partly due to the development of benchmark datasets. Nevertheless, current datasets cover a limited number of languages. With the aim to foster research into massive multilingual ST and ST for low resource languages, we release CoVoST 2, a large-scale multilingual ST corpus covering translations from 21 languages into English and from English into 15 languages. This represents the largest open dataset available to date for volume and language coverage. Data checks provide evidence about the data quality. We provide extensive speech recognition (ASR), machine translation (MT) and ST baselines. We demonstrate the value of CoVoST 2 for multilingual ST research by leveraging it in 4 investigations: simplify multilingual training by removing ASR pretraining, study multilingual model scaling properties and investigate zero-shot and transfer learning capabilities of models trained on CoVoST 2.

1. Introduction

The development of benchmark datasets, such as MuST-C [1], Europarl-ST [2] or CoVoST [3] has greatly contributed to the popularity of ST research. MuST-C provides TED talks translations from English into 14 languages, thereby encouraging research into end-to-end ST [4] and one-to-many multilingual ST [5]. Europarl-ST offers translations between 6 European languages, with a total of 30 language pairs, enabling research into many-to-many multilingual ST [6]. VoxPopuli [7], also derived from the European Parliament proceedings is a large-scale multilingual corpus providing unlabeled speech, transcribed speech and interpretation data. The corpora described so far either involve European languages, or languages that are high resource for MT and ASR, or out of English language pairs. CoVoST [3] is a multilingual and diversified ST corpus from 11 languages into English, based on the Common Voice project [8]. Unlike previous corpora, it involves low resource languages such as Mongolian and it also enables many-to-one ST research. Nevertheless, for all corpora described so far, the number of languages involved is limited.

In this paper, we describe CoVoST 2, an extension of CoVoST which provides translations from English (En) into 15 languages—Arabic (Ar), Catalan (Ca), Welsh (Cy), German (De), Estonian (Et), Persian (Fa), Indonesian (Id), Japanese (Ja), Latvian (Lv), Mongolian (Mn), Slovenian (Sl), Swedish (Sv), Tamil (Ta), Turkish (Tr), Chinese (Zh)—and from 21 languages into English, including the 15 target languages as well as Spanish (Es), French (Fr), Italian (It), Dutch (Nl), Portuguese (Pt), Russian (Ru). The overall speech duration is extended from 700h to 2880 h and the total number of speakers is increased from 11K to 78K. The data is available at `https://github.com/facebookresearch/covost` under a CC0 license.

[⋆] Equal contribution.

Massive multilingual models are more maintainable, require fewer training resources and improve performance on lower resource languages. This paradigm has so far been investigated in the context of MT [9] and ASR [10] but not for ST due to the lack of suitable benchmark. In addition to providing extensive monolingual, bilingual and multilingual ASR, MT and cascaded and end-to-end baselines and in order to demonstrate the value of CoVoST 2, we conduct 4 investigations. First, we simplify multilingual training by omitting the well-known encoder pre-training step [11]. This can be achieved simply by including the ASR tasks and treat them as additional language pairs in training. We then study scaling data and model by combining CoVoST 2 and MuST-C. As CoVoST 2 and MuST-C are English centric, we also evaluate the capability of the multilingual models obtained for zero-shot translation in the Europarl-ST dataset. Finally, we study the transfer learning capabilities of multilingual models on two unseen language pairs, Mboshi-French [12] on which we obtained state-of-the-art performance and the Europarl-ST Romanian-English pair. The results obtained in these investigations are competitive and in some instances, state-of-the-art, which provides additional evidence of the quality of the data released.

2. CoVoST 2

2.1. Corpus Creation

Data Collection & Quality Control. CoVoST 2 is derived from Common Voice (CV) [8], a crowdsourced, read speech, multilingual ASR corpus with an open CC0 license. Translations are collected by sending deduplicated transcripts corresponding to validated voice clips (but without the corresponding audio) to professional translators. We then conduct checks based on language model (LM) perplexity, LASER scores [13] (computed with VizSeq [14]) and a length ratio heuristic in order to ensure the quality of the translations. Samples with low scores are manually inspected and retranslated. For LM perplexity checks, 20M lines are sampled from the OSCAR corpus [15] for each language, except for English and Russian for which pre-trained language models [16] are utilized. 5K lines are reserved for validation and the rest for training. BPE vocabularies of size 20K are then built on the training data, with character coverage 0.9995 for Japanese and Chinese and 1.0 for other languages. A Transformer *base* model [17] is then trained for up to 800K updates. Professional translations are ranked by perplexity and the ones with the lowest perplexity are manually examined and retranslated as needed. In the data release, we mark out the sentences that cannot be translated properly, mostly due to lack of context.

Data Splits. The original CV dataset splits utilize only one audio sample per sentence, while there are potentially multiple speakers available in the validated dataset. To allow higher data utilization and speaker diversity, we add part of the discarded samples back while keeping the speaker set disjoint and

http://dx.doi.org/10.21437/Interspeech.2021-2027

the same sentence assignment across different splits. We refer to this extension as CoVoST splits. Data utilization is thus increased from 44.2% (1273h) to 78.8% (2270h). We use CoVoST train splits for model training and CV dev and test split for validation and evaluation. The CoVoST dev/test splits are useful in multi-speaker evaluation [3] to analyze model robustness, but large amount of repeated sentences may skew the BLEU (WER) scores.

2.2. Statistics

Basic statistics (duration, speakers) are listed in Table 1. CoVoST 2 is diversified with large sets of speakers even on some of the low-resource languages (e.g. Fa, Cy and Nl). Moreover, the speakers are distributed widely across 66 accent groups, 8 age groups and 3 gender groups. CoVoST 2 represents the largest dataset for multilingual speech translation for language coverage (22) and duration (2880 hours).

Table 1: *Basic statistics of CoVoST 2 using original CV splits and extended CoVoST splits.*

	Hours (CoVoST ext.)			Speakers (CoVoST ext.)		
	Train	Dev	Test	Train	Dev	Test
X→En						
Fr	180(264)	22(23)	23(24)	2K(2K)	2K(2K)	4K(4K)
De	119(184)	21(23)	22(120)	1K(1K)	1K(1K)	4K(5K)
Es	97(113)	22(22)	23(23)	1K(1K)	2K(2K)	4K(4K)
Ca	81(136)	19(21)	20(25)	557(557)	722(722)	2K(2K)
It	28(44)	14(15)	15(15)	236(236)	640(640)	2K(2K)
Ru	16(18)	10(15)	11(14)	8(8)	30(30)	417(417)
Zh	10(10)	8(8)	8(8)	22(22)	83(83)	784(784)
Pt	7(10)	4(5)	5(6)	2(2)	16(16)	301(301)
Fa	5(49)	5(11)	5(40)	532(545)	854(908)	1K(1K)
Et	3(3)	3(3)	3(3)	20(20)	74(74)	135(135)
Mn	3(3)	3(3)	3(3)	4(4)	24(24)	209(209)
Nl	2(7)	2(3)	2(3)	74(74)	144(144)	379(383)
Tr	2(4)	2(2)	2(2)	34(34)	76(76)	324(324)
Ar	2(2)	2(2)	2(2)	6(6)	13(13)	113(113)
Sv	2(2)	1(1)	2(2)	4(4)	7(7)	83(83)
Lv	2(2)	1(1)	2(2)	2(2)	3(3)	54(54)
Sl	2(2)	1(1)	1(1)	2(2)	1(1)	28(28)
Ta	2(2)	1(1)	1(1)	3(3)	2(2)	48(48)
Ja	1(1)	1(1)	1(1)	2(2)	3(3)	37(37)
Id	1(1)	1(1)	1(1)	2(2)	5(5)	44(44)
Cy	1(2)	1(12)	1(16)	135(135)	234(371)	275(597)
En→X	364(430)	26(27)	25(472)	10K(10K)	4K(4K)	9K(29K)

3. Experiments & Results

3.1. Models

ASR and ST models share the same Transformer architecture [17, 18]. A convolutional downsampler reduces the length of inputs to $\frac{1}{4}$ of the original length. In the multilingual setting, we force decoding into a given language by using a language ID token as the first token during decoding [6]. For MT, we use a Transformer *base* architecture with l_e encoder layers, l_d decoder layers, 0.3 dropout, and shared embeddings for encoder/decoder inputs and decoder outputs. For multilingual MT models, encoders and decoders are shared across language pairs.

3.2. Experimental Setup

We normalize punctuation and build vocabularies with SentencePiece [19] without pre-tokenization. For ASR and ST, character vocabularies with 100% coverage are used for all experiments on CoVoST 2 only. Experiments on CoVoST 2 combined with MuST-C use a 10k unigram model. For bilingual MT models, joint 5k BPE [20] vocabularies are learned. For multilingual MT models, we learn joint 40k BPE vocabularies. For MT and language pair s-t, we also contrast using only s-t training data and both s-t and t-s training data. The latter setting is referred to as +Rev. We extract 80-dimensional log mel-scale filter bank features (windows with 25ms size and 10ms shift) with per-utterance cepstral mean and variance normalization and SpecAugment [21] to alleviate overfitting. The LB policy without time warping is used for all experiments using CoVoST 2 only while the LD policy is used when combining CoVoST 2 with MuST-C. Training samples with more than 3,000 frames or more than 512 characters are discarded.

For ASR and ST, we set $d_{model} = 256$ for bilingual models and set $d_{model} = 512$ or 1024 (denoted by a suffix "-M"/"-L" in the tables) for multilingual models. We pre-train non-English ASR and bilingual ST models with an English ASR encoder, and pre-train multilingual ST models with a multilingual ASR encoder. For MT, we set $l_e = l_d = 3$ for bilingual models and $l_e = l_d = 6$ for multilingual models.

We use a beam size of 5 and length penalty 1. The best checkpoint is selected by validation loss for MT, and we average the last 5 checkpoints for ASR and ST. For MT and ST, we report case-sensitive detokenized BLEU using sacreBLEU [22] except for English-Chinese and English-Japanese where we report character-level BLEU. For ASR, we report character error rate (CER) on Japanese and Chinese (without word segmentation) and word error rate (WER) on the other languages using VizSeq [14]. Before calculating WER (CER), sentences are tokenized by the default SACREBLEU tokenizer, lowercased and with punctuation removed (except for apostrophes and hyphens). All models are implemented in FAIRSEQ [23, 24] and we will open-source training recipes.

3.3. CoVoST 2 Baselines

Monolingual/Bilingual Baselines. We reports monolingual baselines for ASR and bilingual MT, cascaded ST (C-ST), end-to-end ST trained from scratch (E-ST) and pre-trained on ASR (ST) in Table 2. As expected, the quality of transcriptions and translations is very dependent on the amount of training data per language pair. The poor results obtained on low resource pairs can be improved by leveraging training data from the opposite direction for MT and C-ST. These results serve as baseline for the research community.

Multilingual Baselines. Multilingual MT (rows 3-4), multilingual cascaded ST (rows 5-8) and multilingual A2E/E2E/A2A ST (rows 9-12) baselines are reported in Table 3. The A2E MT model (row 3) underperforms the strong "+Rev" bilingual MT baseline since it is not taking advantage of out-of-English training data. On the other hand, the E2A MT model (row 3) outperforms bilingual MT on En-X, possibly due to a larger capacity. The A2A MT (row 4) is relatively competitive with bilingual MT (-1.4 BLEU overall) but with uneven performance per language group. Rows 5-8 show that A2A MT provides consistent improvements over A2E/E2A and that performance benefits from a larger multilingual ASR model (except for En-X). Rows 9-12 contrast A2E, E2A and A2A models with medium and large architectures. Similar to the conclusions for the cascaded performance, A2A is able to leverage additional data and outperforms A2E/E2A except for En-X; in addition, end-to-end multilingual models benefit from larger architectures.

Table 2: *Test WER for monolingual ASR and test BLEU for bilingual MT/ST ("C-ST" for cascaded ST, "E-ST" for end-to-end ST trained from scratch and "ST" for end-to-end ST with encoder pre-trained on English ASR). All non-English ASR encoders are also pre-trained on the English one. ⋆ We report CER and character-level BLEU on Chinese and Japanese text (no word segmentation available). +Rev†: leveraging CoVoST 2 data from the reversed directions for MT.*

		X→En						En→X					
	ASR	MT	+Rev†	C-ST	+Rev†	E-ST	ST	MT	+Rev†	C-ST	+Rev†	E-ST	ST
En	25.6												
Fr	18.3	37.9	38.1	27.6	27.6	24.3	26.3						
De	21.4	28.2	31.2	21.0	22.6	8.4	17.1	29.0	29.1	18.3	18.1	13.6	16.3
Es	16.0	36.3	36.2	27.4	27.4	12.0	23.0						
Ca	12.6	24.9	31.1	21.3	25.1	14.4	18.8	38.8	38.6	24.1	24.1	20.2	21.8
It	27.4	19.2	19.0	13.5	13.5	0.2	11.3						
Ru	31.4	19.8	19.4	16.8	16.8	1.2	14.8						
Zh⋆	45.0	7.6	16.6	7.0	9.9	1.4	5.8	35.3	38.9	24.6	25.9	20.6	25.4
Pt	44.6	14.6	13.9	9.2	9.2	0.5	6.1						
Fa	62.4	2.4	15.1	2.1	7.2	1.9	3.7	20.1	20.0	13.8	13.8	11.5	13.1
Et	65.7	0.3	13.7	0.2	4.4	0.1	0.1	24.0	24.3	14.5	14.5	11.1	13.2
Mn	65.2	0.2	5.4	0.1	1.9	0.1	0.2	16.8	17.1	11.0	10.7	6.6	9.2
Nl	52.8	2.6	2.5	1.8	1.8	0.3	3.0						
Tr	51.2	1.1	25.9	0.8	12.0	0.7	3.6	20.0	19.7	11.8	11.5	8.9	10.0
Ar	63.3	0.1	34.7	0.1	12.3	0.3	4.3	21.6	21.6	14.0	13.9	8.7	12.1
Sv	65.5	0.2	37.7	0.1	8.4	0.2	2.7	39.4	39.2	24.6	24.4	20.1	21.8
Lv	51.8	0.2	19.6	0.2	9.1	0.1	2.5	22.5	22.9	14.4	14.4	11.5	13.0
Sl	59.1	0.1	29.2	0.0	10.3	0.3	3.0	29.1	29.4	18.2	18.0	11.5	16.0
Ta	80.8	0.0	4.0	0.0	0.7	0.3	0.3	22.7	22.2	13.0	12.7	9.9	10.9
Ja⋆	77.1	0.0	14.6	0.0	2.6	0.3	1.5	42.8	42.2	32.1	29.3	26.9	29.6
Id	63.2	0.1	36.7	0.1	8.9	0.4	2.5	39.0	38.8	22.9	22.7	18.9	20.4
Cy	72.8	0.1	49.2	0.1	6.0	0.3	2.7	41.6	41.6	25.3	25.2	22.2	23.9

Table 3: *Average test BLEU for bilingual and multilingual models. The average WER for ASR-M is 47.6 and 42.9 for ASR-L. HR refers to pairs with Fr, De, Es, Ca as the source language and LR to the remaining ones. We apply temperature-based (T=2) sampling [25] to improve low-resource directions.*

Tasks	Pr.	Avg BLEU				
		X-En			En-X	Overall
		HR	LR	21	15	36
Bi. ST	✓	21.3	4.0	7.3	17.1	11.4
Bi. MT + Rev		34.2	21.0	23.5	29.7	26.1
A2E / E2A MT		33.3	14.1	17.7	32.1	-
A2A MT		36.5	16.6	20.4	30.8	24.7
ASR-M + A2E/E2A MT		25.3	8.1	11.4	18.9	-
ASR-L + A2E/E2A MT		25.8	8.9	12.1	19.6	-
ASR-M + A2A MT		27.1	9.3	12.7	18.1	14.9
ASR-L + A2A MT		27.7	10.2	13.6	18.8	15.7
A2E / E2A-M	✓	23.6	3.3	7.1	17.2	-
A2E / E2A-L	✓	23.2	3.0	6.9	19.4	-
A2A-M	✓	20.4	3.4	6.6	15.5	10.3
A2A-L	✓	24.0	3.7	7.5	18.3	12.0

3.4. ASR + ST

Table 4: *Analysis of the inclusion of the ASR task in training and its interaction with ASR encoder pretraining in two data conditions. CV: CoVoST 2; M-C: MuST-C; HR: high-resource; LR: low-resource.*

Data	Tasks	Pret.	Avg BLEU				
			X-En			En-X	All
			HR	LR	All	All	All
CV	ST	✓	24.0	3.7	7.5	18.3	12.0
		✗	16.6	3.4	5.9	13.1	8.9
	ASR/ ST	✓	26.8	5.0	9.2	20.9	14.0
		✗	25.8	4.1	8.2	19.3	12.8
CV/ M-C	ST	✓	25.9	4.2	8.4	21.7	13.9
		✗	23.4	3.9	7.6	20.4	12.9
	ASR/ ST	✓	27.1	4.6	8.9	21.0	13.9
		✗	27.4	6.4	10.4	21.4	15.0

To simplify training and to understand whether we can combine ST and ASR in one model, we jointly train multilingual ASR + ST models. We compare models with or without ASR pre-training, and including or excluding the ASR task from training. We do so on two data conditions, CoVoST 2 and CoVoST 2 with MuST-C. Table 4 shows the results. On average, the ASR+ST multilingual models, whether pre-trained or not, are on par or outperform corresponding ST models. While encoder pre-training helps to increase the performance of ST models, joint ASR+ST models without pre-training can also reach similar or better performance compared to the ST models with pre-training, thus simplifying the training process.

3.5. Scaling Up Speech Translation Models

We study basic scaling properties of multilingual ST models. Following subsection 3.4, MuST-C is combined with CoVoST 2, the ASR tasks are included in training with no ASR pretraining.

Table 5: *Effect of scaling model capacity up to 950M parameters on speech translation quality.*

# Enc. Lay.	# Dec. Lay	# Param	Avg BLEU	En-X BLEU	X-En BLEU
12		268M	14.8	21.3	10.1
16		318M	14.4	20.2	10.2
20		369M	15.1	22	10.2
24	6	419M	15.5	21.9	10.9
28		469M	15.2	21.4	10.7
32		520M	14.9	21.9	9.8
48		721M	15.6	21.2	**11.6**
12	8	301M	14.1	20.8	9.2
	12	369M	14.3	22	8.9
	16	436M	14.2	22.1	8.6
	20	503M	12.2	19.1	7.3
48	20	956M	**15.9**	**23.7**	10.2

Test set BLEU scores are reported in Table 5, averaging over all language pairs, En-X and X-En pairs. Increasing the encoder capacity improves performance, up to 0.8 BLEU when averaging across all pairs. The trend is consistent for X-En and En-X pairs. However, increasing the decoder capacity up to 16 layers has little effect. Beyond that (20 layers), performance degrades (-1.9 BLEU). We note that a larger decoder capacity may benefit En-X pairs but is consistently detrimental to X-En pairs. The last row increases the model capacity to 956M parameters, bringing an improvement of 1.1 BLEU over the baseline.

Table 6: *Zero-shot performance (BLEU) on Europarl-ST of a multilingual model pretrained on CoVoST 2 and MuST-C, averaged over 20 unseen language pairs.*

System	Avg BLEU
[26]	14.6
Zero-shot	4 (97.0%)
+ self-training	6.7 (98.1%)
Transfer learning	21.1

3.6. Zero-Shot Translation

We evaluate the zero-shot translation performance of a large multilingual model (row 7 in Table 4) on EuroParl-ST. Model training covers X-En and En-X pairs for the EuroParl-ST languages: Fr, De, Es, It and Pt, while translations between those languages are unseen. Results are shown in Table 6. The first row corresponds to a multilingual supervised baseline [26] with a Transformer "M" architecture and encoder pretraining on LibriSpeech [27]. The zero-shot performance is shown in row 2 with the accuracy of the predicted output language. Note that the model does not suffer from the same issue observed in the analogous zero-shot MT setting [28, 29]. Row 3 shows an average 2.7 BLEU improvement by applying self-training on the unseen language pairs and the output language prediction accuracy is further improved to 98.1%. Finally, in row 4, we also report a stronger upper bound by fine-tuning on Europarl-ST.

3.7. Low-Resource Transfer Learning

We investigate transfer learning capabilities of multilingual models trained on CoVoST 2 on Mboshi-French, a very low-resource language pair with 4 hours of speech [12], where Mboshi is unseen during pretraining and French is only seen as target in the French ASR task, and on Romanian-English from the v1.1 release of Europarl-ST with 24 hours of data, where both the Romanian source language and the Romanian-English pair are unseen during pretraining.

Table 7: *Transfer learning on low-resource pairs with unseen source language, with encoder pre-training (ASR-M) and fine-tuning from the A2A-M ST model.* †*: Our end-to-end baseline;* ‡*: [11].*

	Enc. PT	Dec. PT	BLEU
Mb-Fr	-	-	3.5‡
	300h En ASR	20h Fr ASR	7.1‡
	En ASR†		8.8
	En ASR	En ASR	7.1
	ASR-M	ASR-M	9.7
	CV ASR+ST	CV ASR+ST	12.6
Ro-En	En ASR†		10.8
	En ASR	En ASR	14.2
	ASR-M	ASR-M	16.7
	CV ASR+ST	CV ASR+ST	25.1

Results are shown in Table 7. In the first row [11], the model is trained from scratch. The second row [11] brings 3.6 BLEU improvement by pretraining the encoder on 300h of English ASR data and pretraining the decoder on 20h of French ASR data. We obtain similar performance by initializing the model with an English pretrained ASR model (Table 2) Note that initializing both the encoder and decoder underperforms initializing the encoder only (7.1 vs 8.8 BLEU), due to the difference in target language. Performance is further improved with multilingual ASR pre-training and with multilingual ASR+ST pre-training (Table 3), bringing a gain of 5.5 BLEU over the state-of-the-art on this dataset. We observe a similar trend for Romanian-English. The first row (our baseline) initializes the encoder from an English ASR pre-trained model. This time, the decoder initialization brings 3.4 BLEU gains since the target languages match. Similar to Mboshi-French, multilingual ASR pre-training and multilingual ASR+ST pre-training bring further gains, with an improvement of 14.3 BLEU over the baseline.

4. Conclusion

We introduced the largest ST corpus for language coverage and volume, with 21 languages into English and English into 15 languages. We provided extensive monolingual, bilingual and multilingual baselines for ASR, MT and ST. The dataset is free to use under a CC0 license and enables the research community to develop methods including massive multilingual modeling and ST modeling for low resource languages We demonstrated the usefulness of CoVoST 2 for multilingual speech translation research and the quality of the dataset by conducting 4 investigations. We showed how to simplify multilingual training by including ASR as another language pair; scaling properties of multilingual models were investigated; finally, we showed that multilingual ST models trained on CoVoST 2 have good zero-shot and transfer learning capabilities.

5. References

[1] M. A. Di Gangi, R. Cattoni, L. Bentivogli, M. Negri, and M. Turchi, "MuST-C: a Multilingual Speech Translation Corpus," 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 2012–2017.

[2] J. Iranzo-Sánchez, J. A. Silvestre-Cerdà, J. Jorge, N. Roselló, A. Giménez, A. Sanchis, J. Civera, and A. Juan.

[3] C. Wang, J. Pino, A. Wu, and J. Gu, "CoVoST: A diverse multilingual speech-to-text translation corpus," in *Proceedings of The 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 4197–4203.

[4] A. Berard, O. Pietquin, C. Servan, and L. Besacier, "Listen and translate: A proof of concept for end-to-end speech-to-text translation," in *Proceedings of the 2016 NeurIPS Workshop on End-to-end Learning for Speech and Audio Processing*, 2016.

[5] M. A. Di Gangi, M. Negri, and M. Turchi.

[6] H. Inaguma, K. Duh, T. Kawahara, and S. Watanabe.

[7] C. Wang, M. Rivière, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," *arXiv preprint arXiv:2101.00390*, 2021.

[8] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of The 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 4218–4222.

[9] R. Aharoni, M. Johnson, and O. Firat, "Massively multilingual neural machine translation," 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 3874–3884.

[10] V. Pratap, A. Sriram, P. Tomasello, A. Hannun, V. Liptchinsky, G. Synnaeve, and R. Collobert, "Massively multilingual asr: 50 languages, 1 model, 1 billion parameters," 2020.

[11] S. Bansal, H. Kamper, K. Livescu, A. Lopez, and S. Goldwater, "Pre-training on high-resource speech recognition improves low-resource speech-to-text translation," *arXiv preprint arXiv:1809.01431*, 2018.

[12] P. Godard, G. Adda, M. Adda-Decker, J. Benjumea, L. Besacier, J. Cooper-Leavitt, G.-N. Kouarata, L. Lamel, H. Maynard, M. Mueller, A. Rialland, S. Stueker, F. Yvon, and M. Zanon-Boito, "A very low resource language speech corpus for computational language documentation experiments," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018.

[13] M. Artetxe and H. Schwenk, "Margin-based parallel corpus mining with multilingual sentence embeddings," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 3197–3203.

[14] C. Wang, A. Jain, D. Chen, and J. Gu, "Vizseq: A visual analysis toolkit for text generation tasks," *EMNLP-IJCNLP 2019*, p. 253, 2019.

[15] P. J. Ortiz Suárez, L. Romary, and B. Sagot, "A monolingual approach to contextualized word embeddings for mid-resource languages," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 1703–1714.

[16] N. Ng, K. Yee, A. Baevski, M. Ott, M. Auli, and S. Edunov, "Facebook FAIR's WMT19 news translation task submission," in *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 314–319.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[18] G. Synnaeve, Q. Xu, J. Kahn, T. Likhomanenko, E. Grave, V. Pratap, A. Sriram, V. Liptchinsky, and R. Collobert, "End-to-end asr: from supervised to semi-supervised learning with modern architectures," in *ICML 2020 Workshop on Self-supervision in Audio and Speech*, 2020.

[19] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 66–71.

[20] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1715–1725.

[21] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[22] M. Post, "A call for clarity in reporting BLEU scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 186–191.

[23] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

[24] C. Wang, Y. Tang, X. Ma, A. Wu, D. Okhonko, and J. Pino, "fairseq s2t: Fast speech-to-text modeling with fairseq," *arXiv preprint arXiv:2010.05171*, 2020.

[25] N. Arivazhagan, A. Bapna, O. Firat, D. Lepikhin, M. Johnson, M. Krikun, M. X. Chen, Y. Cao, G. Foster, C. Cherry, W. Macherey, Z. Chen, and Y. Wu, "Massively multilingual neural machine translation in the wild: Findings and challenges," 2019.

[26] X. Li, C. Wang, Y. Tang, C. Tran, Y. Tang, J. Pino, A. Baevski, A. Conneau, and M. Auli, "Multilingual speech translation with efficient finetuning of pretrained models," *arXiv e-prints*, pp. arXiv–2010, 2020.

[27] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[28] J. Gu, Y. Wang, K. Cho, and V. O. Li, "Improved zero-shot neural machine translation via ignoring spurious correlations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 1258–1268. [Online]. Available: https://www.aclweb.org/anthology/P19-1121

[29] B. Zhang, P. Williams, I. Titov, and R. Sennrich, "Improving massively multilingual neural machine translation and zero-shot translation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 1628–1639. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.148