INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Improving Cross-Lingual Transfer Learning for End-to-End Speech Recognition with Speech Translation

Changhan Wang, Juan Pino, Jiatao Gu

Facebook AI, USA

{changhan, juancarabina, jgu}@fb.com

Abstract

Transfer learning from high-resource languages is known to be an efficient way to improve end-to-end automatic speech recognition (ASR) for low-resource languages. Pre-trained or jointly trained encoder-decoder models, however, do not share the language modeling (decoder) for the same language, which is likely to be inefficient for distant target languages. We introduce speech-to-text translation (ST) as an auxiliary task to incorporate additional knowledge of the target language and enable transferring from that target language. Specifically, we first translate high-resource ASR transcripts into a target low-resource language, with which a ST model is trained. Both ST and target ASR share the same attention-based encoder-decoder architecture and vocabulary. The former task then provides a fully pre-trained model for the latter, bringing up to 24.6% word error rate (WER) reduction to the baseline (direct transfer from high-resource ASR). We show that training ST with human translations is not necessary. ST trained with machine translation (MT) pseudo-labels brings consistent gains. It can even outperform those using human labels when transferred to target ASR by leveraging only 500K MT examples. Even with pseudo-labels from low-resource MT (200K examples), ST-enhanced transfer brings up to 8.9% WER reduction to direct transfer.

Index Terms: end-to-end speech recognition, cross-lingual transfer learning, speech translation, machine translation

1. Introduction

The attention-based encoder-decoder model paradigm [1, 2] has recently witnessed rapidly increased applications in end-to-end automatic speech recognition (ASR). It provides a generic framework for speech-to-text generation tasks, and achieves state-of-the-art performance on ASR [3, 4, 5] as an alternative to CTC (Connectionist temporal classification) models [6]. The recent surge of end-to-end speech-to-text translation (ST) studies [7, 8, 9, 10, 11] is also due to the application of attention-based encoder-decoder models. And very recent works [12, 13, 14] have demonstrated the possibility of combining the two related tasks, ASR and ST, under the same encoder-decoder architecture to achieve better performance. When targeting at ST only, transfer learning from ASR [15, 13] is helpful to warm-starting acoustic modeling (encoder) and enabling ST model training to focus more on learning language modeling and alignment (decoder).

In this paper, we study how to utilize ST to improve cross-lingual transfer learning for ASR. Transfer learning from high-resource languages [16, 17, 18, 19] is known to be an efficient way to improve end-to-end ASR for low-resource languages. Pre-trained or jointly trained encoder-decoder models, however, do not share the language modeling (decoder) for the same language, which is likely to be inefficient for distant target languages. We introduce ST as an auxiliary task to incorporate additional knowledge of the target language and enable transferring from that target language. Unlike previous ideas for leveraging translation data [20, 21, 22], our approach does not require any modification to the ASR model architecture. It leverages ST data instead of text-to-text translation data for ST training, which avoids speech-to-text modality adaption in the encoder. Moreover, we train ST with machine translation (MT) pseudo-labels on high-resource ASR transcripts, which overcomes the shortage of real ST data and consistently brings gains to the transfer learning. MT pseudo-labeling also simplifies ST model training (knowledge distilled data) and allows beam-searching diverse labels to alleviate overfitting.

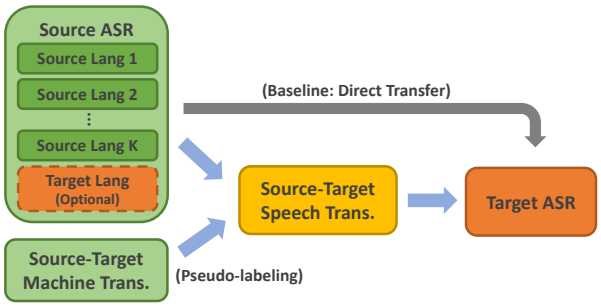

Figure 1: *An overview of proposed cross-lingual transfer learning pipeline. The color reflects data availability/quality.*

2. Methods

2.1. Attention-Based Encoder-Decoder Architecture

Our ASR and ST models share the same BLSTM-based encoder-decoder architecture [11] with attention mechanism, which is similar to the LAS architecture [23, 3, 4]. Specifically, on the encoder side, audio features $\mathbf{x} \in \mathbb{R}^{T \times d_0}$ are first fed into a two-layer DNN with $tanh$ activations and hidden sizes d_1 and d_2. Then two 2D convolutional layers with kernel size 3x3 and stride 2x2 are applied to reduce the sequence length to $\frac{T}{4}$. Both convolutional layers have 16 output channels and project the features to $4d_2$ dimensions after flattening. Finally, the features are passed to a stack of three bidirectional LSTM layers of hidden size d_3 to form encoder output states $\mathbf{h} \in \mathbb{R}^{T \times 2d_3}$. For the decoder side, a stack of two LSTM layers with hidden size $2d_3$ and additive attention [2] is applied, followed by a linear projection to size d_o.

For MT, we use one of Transformer *base* with 3 encoder/decoder layers, Transformer *base* and Transformer *big* models [24] (with original training hyper-parameters) depending on the MT dataset size.

http://dx.doi.org/10.21437/Interspeech.2020-2955

2.2. Speech Translation Trained with Pseudo-Labels

Word-level or sequence-level knowledge distillation (KD) is helpful to MT [25] and ST [26] model training, because it reduces noise and simplifies data distribution in the training set. Training end-to-end ST models is known to be difficult, given the fact that it needs to learn acoustic modeling, language modeling and alignment at the same time. When the training data distribution is complex, end-to-end ST models are likely to fit the data worse than cascading ASR and MT models. Moreover, ST labels are more expensive to obtain than ASR or MT ones. Existing ST corpora are strongly limited by size and language coverage, making ST model training even more difficult. To overcome the shortage of real data, we propose to pseudo-label ASR corpora with MT and train ST on the resulting datasets. This provides larger scale training data as well as more diversity (via different MT models and beam search) at little cost. Both are useful for alleviating overfitting. Moreover, training ST models with MT pseudo-labels can be viewed as a sequence-level KD process. Although potentially inaccurate pseudo-labels can hurt model training, pseudo-labels are easier to be fitted. This compensates its gap towards real labels, which are likely more difficult to learn. In our experiments, we show that ST models trained with pseudo-labels can even outperform those using real labels when transferred to the target ASR.

2.3. Pre-training ASR on Speech Translation

Instead of pretraining target (low-resource) ASR directly on (multilingual) source (high-resource) ASR, we pretrain target ASR on source-to-target ST. The latter is pretrained on source ASR and leverages MT pseudo-labels on source ASR data for training. Figure 1 provides an overview of our proposed transfer learning pipeline: $\text{ASR}_{\text{Source}} \rightarrow \text{ST}_{\text{Source-Target}} \rightarrow \text{ASR}_{\text{Target}}$. Our intuition is that this two-step approach helps to decouple transfer of language modeling (decoder) and acoustic modeling (encoder) to make transfer learning smoother and more effective. Moreover, the ST model leverages additional data (MT pseudo-labels) for the target language and hence is likely to model the target language better. We use the same model architecture for ASR and ST, so that they can be easily transferred between each other: $\text{ASR}_{\text{Source}} \rightarrow \text{ST}_{\text{Source-Target}}$[1] and $\text{ST}_{\text{Source-Target}} \rightarrow \text{ASR}_{\text{Target}}$. Pretraining ST with ASR warm-starts acoustic modeling so that ST training can be more focused on learning language modeling and alignment. We may simplify the transfer learning pipeline by training ST from scratch. This still outperforms direct $\text{ASR}_{\text{Source}} \rightarrow \text{ASR}_{\text{Target}}$ transfer in most of the cases as shown in our experiments. To another extreme, we may pre-train ST jointly on source+target ASR to warm-start from a better acoustic model: $\text{ASR}_{\text{Source+Target}} \rightarrow \text{ST}_{\text{Source-Target}}$.

3. Experiments

3.1. Data

For English and English+French ASR, we use Librispeech [27] and Common Voice [28] (v4, 2019-12-10 release).[2] We also use the ASR data in MuST-C [30] (En-Nl subset) for the analysis in section 3.3.2. For target ASR, we use Portuguese (Pt), Chinese (Zh-CN), Dutch (Nl) and Mongolian (Mn) from Common Voice v4[2] as well as Vietnamese (Vi) and Ht (Haitian) from IARPA Babel datasets (conversational telephone speech). Basic statistics of all used ASR corpora can be found in Table 1. For MT, we use a variety of datasets indexed by OPUS [31], which are listed in Table 2.

Table 1: *Source and target ASR data.*

	Dataset	Train	Speakers
	Source ASR		
CV	Common Voice: English	477h	15.2k
CV_{Fr}	Common Voice: French	264h	1.8k
LS	Librispeech	960h	2.3k
MC	MuST-C: En-Nl	422h	2.2k
	Target ASR		
Vi	IARPA Babel 107b-v0.7	96h	0.6k
Ht	IARPA Babel 201b-v0.2b	70h	0.3K
Pt	Common Voice v4	10h	2
Zh-CN	Common Voice v4	10h	22
Nl	Common Voice v4	7h	78
Mn	Common Voice v4	3h	4

Table 2: *MT data and Transformer models.*

	Dataset	En/Fr Sent.	Model
Vi	OpenSubtitles	4M/3M	Base
Ht	JW300	220K/220K	Base 3+3
Pt	OpenSubtitles	33M/23M	Big
Zh	MultiUN	10M/10M	Big
Nl	OpenSubtitles	37M/25M	Big
Mn	JW300+GNOME+QED	210K/203K	Base 3+3
Nl_W	WikiMatrix	511K/-	Base 3+3
Nl_S	OpenSubtitles	37M/-	Base 3+3
Nl_M	OpenSubtitles	37M/-	Base

3.2. Experimental Setup

For all texts, we normalize their punctuation and tokenize them with sacreMoses[3]. For ASR and ST, we lowercase the texts (except for Babel). For ASR, we remove all punctuation markers except for apostrophes. We use character vocabularies for ASR and ST, and use BPE vocabularies [32] for MT. We extract 80-channel log-mel filterbank features (windows with 25ms size and 10ms shift) using Kaldi [33], with per-utterance cepstral mean and variance normalization (CMVN) applied. We remove training samples having more than 3,000 frames or more than 512 characters for GPU memory efficiency.

The configuration of MT models can be found in Table 2. For ASR and ST models, we set $d_1 = 256$, $d_2 = 128$, $d_3 = 512$ and $d_o = 128$. We adopt SpecAugment [4] (LB policy without time warping) to alleviate overfitting. All models are implemented in Fairseq [34]. We use a beam size of 5 for decoding. We average the last 5 checkpoints for ASR and ST, and average the last 2 checkpoints for MT. For MT and ST, we report case-insensitive tokenized BLEU [35] using sacreBLEU [36]. For ASR, we report character error rate (CER) on Chinese (no word segmentation) and word error rate (WER) on the other languages using VizSeq [37].

[1] Full model transfer excluding the embedding and softmax layers.

[2] The original dataset splits contain only one sample per sentence. We instead use extended splits [29] to allow using all samples.

[3] https://github.com/alvations/sacremoses

Table 3: *Test WER (relative reduction in parentheses) for cross-lingual transfer from English and from English+French*

		Vi	Ht	Pt	Zh-CN	Nl	Mn
Baseline		57.2	66.1	62.3	90.3	96.5	109.7
From English							
CV	Src ASR	53.7	60.7	40.9	41.3	44.2	67.7
	+ ST	52.5 (-2.2%)	59.3 (-2.3%)	33.7 (-17.6%)	35.3 (-14.5%)	42.0 (-5.0%)	64.1 (-5.3%)
	Src+Tgt ASR	51.6	58.1	34.7	37.0	42.5	63.0
	+ ST	51.2 (-0.8%)	57.2 (-1.5%)	31.2 (-10.1%)	35.2 (-4.9%)	40.4 (-4.9%)	62.3 (-1.1%)
CV+LS	Src ASR	54.7	59.9	41.3	40.0	42.2	66.1
	+ ST	52.9 (-3.3%)	57.4 (-4.2%)	31.8 (-23.0%)	35.7 (-4.2%)	37.9 (-10.2%)	60.2 (-8.9%)
	Src+Tgt ASR	52.7	57.8	34.4	36.4	41.7	67.9
	+ ST	52.2 (-0.9%)	57.2 (-1.0%)	31.2 (-9.3%)	35.5 (-2.5%)	38.8 (-7.0%)	62.5 (-8.0%)
From English+French							
CV+CV$_{Fr}$	Src ASR	54.5	59.4	39.5	39.2	43.0	67.7
	+ ST	51.7 (-5.1%)	57.8 (-2.7%)	29.8 (-24.6%)	33.6 (-14.3%)	38.4 (-10.7%)	62.1 (-8.3%)
	Src+Tgt ASR	52.9	57.1	31.7	36.4	40.7	62.4
	+ ST	52.0 (-1.7%)	55.7 (-2.5%)	28.6 (-9.8%)	32.9 (-9.6%)	38.3 (-5.9%)	59.6 (-4.5%)

Table 4: *Test WER for source ASR*

	CV	+LS	+CV$_{Fr}$	MC	+CV
En/Fr	25.4/-	16.7/-	23.4/20.1	19.6/-	18.6/-

3.3. Results

3.3.1. Cross-Lingual Transfer via ST

We examine two settings for high-resource source ASR: monolingual (English) and multilingual (English and French). The test WER of source ASR models can be found in Table 4. Both settings use the same low-resource targets from different language families: Indo-European (Portugese and Dutch), Sino-Tibetan (Chinese), Austro-Asiatic (Vietnamese), French Creole (Haitian) and Mongolic (Mongolian). We experiment with different transfer learning strategies: with or without ST as an intermediate step, and with or without target ASR during ASR pre-training. The results (test WER) are presented in Table 3: $\text{ASR}_{\text{Source}} \rightarrow \text{ASR}_{\text{Target}}$ ("Src ASR"), $\text{ASR}_{\text{Source}} \rightarrow \text{ST}_{\text{Source-Target}} \rightarrow \text{ASR}_{\text{Target}}$ (the 2nd row "+ ST"); $\text{ASR}_{\text{Source+Target}} \rightarrow \text{ASR}_{\text{Target}}$ ("Src+Tgt ASR"), $\text{ASR}_{\text{Source+Target}} \rightarrow \text{ST}_{\text{Source-Target}} \rightarrow \text{ASR}_{\text{Target}}$ (the 4th row "+ ST"). We see that for both monolingual and multilingual settings, ST pre-training consistently brings gains to the direct transfer baseline. On Portuguese (Pt) and Dutch (Nl), there is over 9.3% and 4.9% WER reduction in all ST-enhanced transfers, respectively. There is also 1.0%-8.9% WER reduction on Haitian (Ht) and Mongolian (Mn), where MT is also low-resource with only around 200K training examples available. When the source ASR has larger scale data (from CV to CV+LS), the gains brought by ST may be enlarged, for example, from 5.0% reduction to 10.2% reduction for Dutch and from 5.3% reduction to 8.9% reduction for Mongolian.

3.3.2. MT Models for Pseudo-Labeling

In order to better understand how different MT pseudo-labels may affect the performance of ST as well as downstream target ASR, we experiment with Dutch pseudo-labels from different MT models for ST training: Nl_W and Nl_S both use Transformer *base* with 3 encoder/decoder layers but are trained on WikiMatrix (0.5M examples) and OpenSubtitles (37M examples), respectively; Nl_S, Nl_M and Nl are all trained on OpenSubtitles but use Transformer *base* with 3 encoder/decoder layers, Transformer *base* and Transformer *big*, respectively. We use MuST-C optionally with Common Voice ("+CV") as English source ASR for different data conditions. Results (MT and ST BLEU on MuST-C test set as well as Dutch ASR test WER) are available in Table 5. We notice that the ST model using Nl has almost the same ST BLEU as that using real labels, although Nl has only 35.6 MT BLEU. Real labels are more difficult to learned in ST (76% BLEU drop from MT to ST compared to pseudo-labels' 33%). Nl_W and Nl_S share the same architecture, while the latter is trained on a more noisy corpus. When using MC only, Nl_W outperforms Nl_S and smaller models on OpenSubtitles (Nl_S and Nl_M) are helpful to suppressing noise. When more data (MC+CV) is available, Nl_S performs as well as Nl_W on downstream ASR and larger models on OpenSubtitles are better for transfer.

Table 5: *Performance of different Dutch pseudo-labels*

	ST Label (NA for baseline transfer)					
	NA	Nl_W	Nl_S	Nl_M	Nl	Real
MT	-	24.8	34.0	34.1	35.6	**100.0**
ST	-	18.9	23.7	23.9	**24.0**	23.9
+CV	-	18.6	**23.3**	22.6	23.1	-
ASR	44.7	**42.4**	43.1	43.2	43.9	43.9
+CV	42.4	**38.7**	40.0	39.2	**38.7**	-

3.3.3. Pseudo-Label Sampling and Filtering

Pseudo-labels are from beam search decoding of MT models. There are up to k predictions per example given beam size k. Instead of using only the best ones for ST model training, we explore using the n-best ones ($2 \leq n \leq k$) to provide more diversity and alleviate overfitting. Specifically, in each epoch, training labels are uniformly sampled from the set of n-best candidates. Pseudo-labels from low-resource or out-of-domain

Table 6: *Test WER for transfers without ASR pre-training*

	Vi	Ht	Pt	Zh-CN	Nl	Mn
ASR	54.5	59.4	39.5	39.2	43.0	67.7
ASR→ST	51.7	57.8	29.8	33.6	38.4	62.1
ST	53.7	58.7	32.5	35.3	44.1	67.3

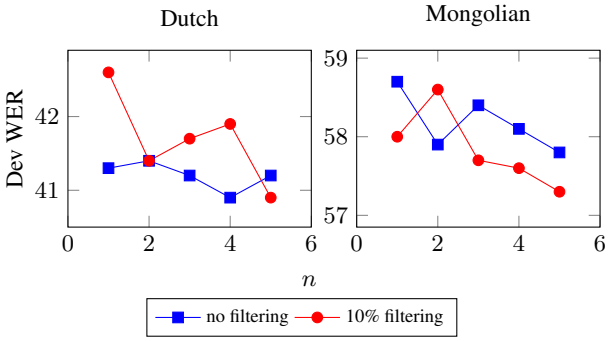

Figure 2: *Dev WER for Dutch (highest MT resource) and Mongolian (lowest MT resource) ASR pre-trained with ST using N-best MT pseudo-labels (optionally with filtering).*

MT models may have low quality on some of the examples. We optionally filter 10% examples by confidence scores (length-normalized log likelihood) to reduce noisy labels. We experiment with Dutch (highest MT resource) and Mongolian (lowest MT resource) for different values of n ($k = 5$). It can be seen from Figure 2 that n-best pseudo-labels lead to lower dev WER in most of the cases and filtering helps significantly when MT is low-resource (Mongolian).

3.3.4. *Effectiveness of ST Pre-training*

We introduce ST to the pipeline with the idea of bringing pre-trained models closer to the target ones. In other words, we expect the $\text{ST}_{\text{Source-Target}} \rightarrow \text{ASR}_{\text{Target}}$ transfer to be faster than the $\text{ASR}_{\text{Source}} \rightarrow \text{ASR}_{\text{Target}}$ transfer. We examine the training accuracy curves for Vietnamese (highest resource) and Mongolian (lowest resource) to verify our hypothesis (see Figure 3). We observe that the ST-enhanced transfer ("w/ ST") has substantially higher starting points (60 to 53 and 70 to 36) and keeps leading with a substantial gap throughout the training process.

3.3.5. *ST without ASR Pre-training*

Instead of using ST as an intermediate step during transfer, we can also train ST from scratch to simplify the transfer pipeline. We experiment with CV+CV_{Fr}, whose results can be found in Table 6. It is shown that the simplified ST-enhanced transfers can still outperform ASR-only ones in most of the cases, although the lack of ASR pre-training brings difficulties to ST model training.

4. Related Work

End-to-end models, such as CTC models and attention-based encoder-decoder models, work well for low-resource ASR [38]. It is known that multilingual training or pre-training with related languages improves low-resource end-to-end ASR significantly [16, 17, 18, 19]. Meta learning methods [39] have recently been introduced to improve the efficiency of multi-

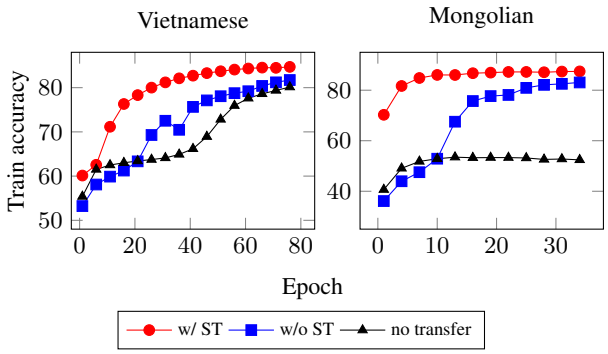

Figure 3: *Training accuracy curve for Vietnamese (highest resource) and Mongolian (lowest resource).*

lingual pre-training. Besides cross-lingual transfer learning, leveraging auxiliary data is another approach to improve low-resource ASR, for example, incorporating (synthetic) text translation data as additional inputs [21, 20, 22] or co-training with weakly supervised data [40] or text-to-speech (TTS) data [41].

5. Conclusions

We show that cross-lingual (high-resource to lower-resource) transfer learning for end-to-end ASR can be improved by adding ST as an intermediate step. It makes transfer learning smoother in the two-step process and incorporates additional knowledge of the target language to improve model performance. It leverages only MT pseudo-labels but no expensive human labels to train ST and does not require high-resource MT training data. Currently, our approach is based on attention-based encoder-decoder architecture. Our future work includes extending this transfer learning approach to other end-to-end architectures, such as CTC and RNN Transducer.

6. Acknowledgements

We thank Ann Lee, Yatharth Saraf, Chunxi Liu and Anne Wu for helpful discussions.

7. References

[1] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[2] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[3] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, N. Jaitly, B. Li, J. Chorowski, and M. Bacchiani, "State-of-the-art speech recognition with sequence-to-sequence models," 2017.

[4] 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.

[5] G. Synnaeve, Q. Xu, J. Kahn, E. Grave, T. Likhomanenko, V. Pratap, A. Sriram, V. Liptchinsky, and R. Collobert, "End-to-end asr: from supervised to semi-supervised learning with modern architectures," *arXiv preprint arXiv:1911.08460*, 2019.

[6] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[7] A. Bérard, O. Pietquin, C. Servan, and L. Besacier, "Listen and translate: A proof of concept for end-to-end speech-to-text translation," *arXiv preprint arXiv:1612.01744*, 2016.

[8] L. Duong, A. Anastasopoulos, D. Chiang, S. Bird, and T. Cohn, "An attentional model for speech translation without transcription," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 949–959.

[9] R. J. Weiss, J. Chorowski, N. Jaitly, Y. Wu, and Z. Chen, "Sequence-to-sequence models can directly translate foreign speech," *arXiv preprint arXiv:1703.08581*, 2017.

[10] L. C. Vila, C. Escolano, J. A. Fonollosa, and M. R. Costa-jussà, "End-to-end speech translation with the transformer." in *IberSPEECH*, 2018, pp. 60–63.

[11] A. Bérard, L. Besacier, A. C. Kocabiyikoglu, and O. Pietquin, "End-to-end automatic speech translation of audiobooks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6224–6228.

[12] A. Anastasopoulos and D. Chiang, "Tied multitask learning for neural speech translation," *arXiv preprint arXiv:1802.06655*, 2018.

[13] M. A. Di Gangi, M. Negri, and M. Turchi, "One-to-many multilingual end-to-end speech translation," *arXiv preprint arXiv:1910.03320*, 2019.

[14] Y. Liu, J. Zhang, H. Xiong, L. Zhou, Z. He, H. Wu, H. Wang, and C. Zong, "Synchronous speech recognition and speech-to-text translation with interactive decoding," *arXiv preprint arXiv:1912.07240*, 2019.

[15] 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.

[16] S. Dalmia, R. Sanabria, F. Metze, and A. W. Black, "Sequence-based multi-lingual low resource speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4909–4913.

[17] J. Yi, J. Tao, Z. Wen, and Y. Bai, "Adversarial multilingual training for low-resource speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4899–4903.

[18] J. Cho, M. K. Baskar, R. Li, M. Wiesner, S. H. Mallidi, N. Yalta, M. Karafiat, S. Watanabe, and T. Hori, "Multilingual sequence-to-sequence speech recognition: architecture, transfer learning, and language modeling," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 521–527.

[19] S. Zhou, S. Xu, and B. Xu, "Multilingual end-to-end speech recognition with a single transformer on low-resource languages," *arXiv preprint arXiv:1806.05059*, 2018.

[20] A. Anastasopoulos and D. Chiang, "Leveraging translations for speech transcription in low-resource settings," *arXiv preprint arXiv:1803.08991*, 2018.

[21] T. Hayashi, S. Watanabe, Y. Zhang, T. Toda, T. Hori, R. Astudillo, and K. Takeda, "Back-translation-style data augmentation for end-to-end asr," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 426–433.

[22] M. Wiesner, A. Renduchintala, S. Watanabe, C. Liu, N. Dehak, and S. Khudanpur, "Pretraining by Backtranslation for End-to-End ASR in Low-Resource Settings," in *Proc. Interspeech 2019*, 2019, pp. 4375–4379.

[23] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.

[24] 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.

[25] Y. Kim and A. M. Rush, "Sequence-level knowledge distillation," *arXiv preprint arXiv:1606.07947*, 2016.

[26] Y. Liu, H. Xiong, Z. He, J. Zhang, H. Wu, H. Wang, and C. Zong, "End-to-end speech translation with knowledge distillation," *arXiv preprint arXiv:1904.08075*, 2019.

[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] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," 2019.

[29] 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.

[30] M. A. Di Gangi, R. Cattoni, L. Bentivogli, M. Negri, and M. Turchi, "Must-c: a multilingual speech translation corpus," in *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2019, pp. 2012–2017.

[31] J. Tiedemann, "Parallel data, tools and interfaces in opus." in *Lrec*, vol. 2012, 2012, pp. 2214–2218.

[32] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," *arXiv preprint arXiv:1508.07909*, 2015.

[33] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

[34] 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.

[35] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.

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

[37] C. Wang, A. Jain, D. Chen, and J. Gu, "Vizseq: a visual analysis toolkit for text generation tasks," *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, 2019.

[38] A. Rosenberg, K. Audhkhasi, A. Sethy, B. Ramabhadran, and M. Picheny, "End-to-end speech recognition and keyword search on low-resource languages," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5280–5284.

[39] J.-Y. Hsu, Y.-J. Chen, and H.-y. Lee, "Meta learning for end-to-end low-resource speech recognition," *arXiv preprint arXiv:1910.12094*, 2019.

[40] K. Singh, D. Okhonko, J. Liu, Y. Wang, F. Zhang, R. Girshick, S. Edunov, F. Peng, Y. Saraf, G. Zweig *et al.*, "Training asr models by generation of contextual information," *arXiv preprint arXiv:1910.12367*, 2019.

[41] Y. Ren, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Almost unsupervised text to speech and automatic speech recognition," 2019.