INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

GigaST: A 10,000-hour Pseudo Speech Translation Corpus

Rong Ye, Chengqi Zhao* , Tom Ko, Chutong Meng† , Tao Wang, Mingxuan Wang, Jun Cao*

ByteDance

{yerong, zhaochengqi.d, tom.ko, wangtao.960826, wangmingxuan.89, caojun.sh}@bytedance.com, mengct00@gmail.com

Abstract

This paper introduces GigaST, a large-scale pseudo speech-to-text translation (ST) corpus. We create the corpus by translating the transcript in GigaSpeech, an English ASR corpus, into German and Chinese. The training set is translated by a strong machine translation system and the test set is translated by human. ST models trained with an addition of our corpus obtain new state-of-the-art results on the MuST-C English-German benchmark test set. We provide a detailed description of the translation process and verify its quality. We make the translated text data public and hope to facilitate research in speech translation. Additionally, we also release the training scripts on NeurST[1] to make it easy to replicate our systems. GigaST dataset is available at https://st-benchmark.github.io/resources/GigaST.

1. Introduction

End-to-end speech-to-text translation (ST) directly translates the speech in the source language into sentences in the target language, without outputting the source language transcription [1]. With the success of attention-based models for speech and text-related tasks, a typical and effective baseline model for ST is speech-transformer [2, 3, 4, 5], which has much of the same model structure as the commonly used MT model Transformer [6], except for the pre-processing down-sampling module for speech signals.

To train such an end-to-end ST model, a high-quality dataset is important. In general, the more data available, the better the model can be trained. For example, in the MT task, as the bilingual parallel data increases, so does the translation performance. As for the speech translation benchmark dataset MuST-C English-German (En-De), which is currently most widely compared by various models, for example, there are only 234k samples (408 hours), while there are more than 4M parallel text training samples for En-De text translation, and in comparison, they are not even in the same order of magnitude. If the ST training data is also upgraded to the same order of magnitude as the MT data, what will be the translation performance of various end-to-end models?

Therefore, in this paper, we try to build a large-scale speech-to-text translation dataset. Fortunately, GigaSpeech [7] provided 10,000 hours of English *speech-transcription* parallel data, which served as the speech source for our datasets. We extend the GigaSpeech ASR dataset to a massive ST corpus – GigaST, up to 25 times as large as the existing open-source dataset, such as MuST-C [8] and TEDx [9]. Specifically, the target-side translations of the training set are obtained by translating the transcription using high-quality MT models. The translations in the test sets are manually annotated and verified one by one, which avoided the alignment errors in the previous MuST-C dataset. The GigaST dataset contains both English-Chinese (En-Zh) and English-German (En-De) translation directions.

Using this dataset, we first evaluate the performance of the standard Speech-Transformer models. Then we evaluate the performance of SSL-Transformer models (SSL stands for self-supervised learning) by incorporating pre-trained speech encoders, Wav2vec2 [10] and HuBERT [11]. Results show that increasing the size of the training dataset with GigaST improves the BLEU scores in all test sets.

2. Dataset Creation

This section describes our method of creating a pseudo ST corpus from an existing ASR corpus. Pseudo-labeling has been proven effective in various machine learning tasks [12, 13, 14].

2.1. Training Set

We start from GigaSpeech [7], a multi-domain English speech recognition corpus with 10,000 hours of labeled audio, and create paired text-to-text translation with an MT model. To obtain high-quality pseudo labels, we train deep transformer-based machine translation models [15], with 24 layers of the encoder and 6 layers of the decoder. The training data for MT consists of WMT2021 [2] and CCMatrix, CCAlign and OpenSubtitles portions from OPUS [3]. We follow the data filtering and pre-processing methods described in [16, 17], and utilize iterative sequence-level knowledge distillation [18, 19] and back translation [20] techniques to improve the performance of MT.

The utterances of the original GigaSpeech corpus are segmented from long audio recordings. The discourse phenomena, such as pronominal anaphora and lexical consistency will be neglected by sentence-level MT systems, which makes it far worse than human translations [21, 22]. Therefore we apply *multi-resolution training* [23] to build a context-aware MT system. Specifically, we first concatenate each segment with its previous one and insert a special token [SEP] as the segment separator. Then the generated translation is split according to the [SEP] token.

Since our training dataset is artifacted from GigaSpeech ASR corpus based on the MT system described above, we want to verify how good the MT system is and whether it is close to

* Equal contribution.

† Work is done during internship at Bytedance.

[1] https://github.com/bytedance/neurst/

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

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

10.21437/Interspeech.2023-1233

the real translations of translators. To this end, we perform a comparison in terms of both automated metrics evaluation and human evaluation.

Table 1 illustrates the automated metrics evaluation (in BLEU) of our MT models for En-Zh and En-De directions on *newstest2021* set, where Online-B/W and Online-W/A are top-2 online systems individually for En-Zh and En-De reported by [24]. It shows that our MT model can obtain state-of-the-art performance and is comparable with online systems.

Table 1: *BLEU scores of our MT models versus online MT systems on* ***newstest2021*** *test set.*

Direction	System	BLEU
En-Zh	Online-B	48.5
	Online-W	44.8
	Our model	47.3
En-De	Online-W	51.0
	Online-A	47.6
	Our model	49.8

Furthermore, we conduct the human evaluation on the pseudo labels to verify the quality. First, we sample 30 audios from the training set and randomly select 1 to 20 continuous segments from each audio, with a total of 320 unique segments. Then 2 professional translators separately produce Chinese and German translations for this evaluation set, which we take as the ground truth. And another 6 evaluators (3 for En-Zh and 3 for En-De) are asked to rate the translations of both human and MT systems from 0 to 6. A rating of 6 indicates that the expression is fluent and the meaning of the translation is faithful to the source without any grammar errors, while a rating of 0 means the translation is incomprehensible and full of errors. Table 2 lists the averaged scores from the evaluators. The En-Zh MT system gets a rating of 4.14 and for En-De it is 4.82. The above 4 rating means our generated translations are semantically consistent with the source texts. The weakness mainly comes from fluency problems, such as unidiomatic word translations, informal expressions, and function word errors. Moreover, the rating of En-De MT (4.82) is close to that of human translations (5.06), which further verifies the quality of our produced training set.

Table 2: *Human evaluation results on the translations by human translators and machine translation models.*

Direction	Human	MT
En-Zh	4.92	4.14
En-De	5.06	4.82

2.2. Test Sets

We provide En-Zh and En-De test sets in GigaST. The En-Zh test set contains the translation of all GigaSpeech test utterances while the En-De test set contains a subset of it for this current release. The test sets are produced by human translators looking at the transcriptions. A small number of transcriptions are difficult to understand due to the lack of context and are ignored in our test sets.

The statistics of GigaST are listed in Table 3. Note that non-speech segments, such as music and noise, are not included in our statistics. The #tokens is counted in character and word for En-Zh and En-De respectively.

Table 3: *The statistics of GigaST*

Direction	Subset	#seg.	#hours	#tokens
En-Zh	S	210,012	243.1	4.1M
	M	835,846	974.3	16.7M
	L	2,084,274	2,337.7	41.8M
	XL	7,650,889	9,780.8	168.3M
	Test	19,888	35.3	0.6M
En-De	S	221,572	256.2	2.5M
	M	868,316	1,013.1	10.2M
	L	2,147,471	2,510.9	25.3M
	XL	7,815,436	9,997.9	101.6M
	Test	4,163	7.1	0.7M

3. Experiment

In this section, we conduct ST experiments with speech transformer models and SSL-Transformer models. All models are implemented using NeurST [5].

3.1. Setups

Preprocessing and Filtering For speech transformer models, we extract 80-channel log-Mel filterbank coefficients (fbank) of the audio, with windows of 25ms and steps of 10ms, and then apply CMVN (cepstral mean and variance normalization). The SSL-Transformer models use raw wave signals as the input. For the text side, words are encoded in subword-level. In detail, we lowercase the English transcriptions, remove all punctuations and use SentencePiece[4] with a vocabulary of 15,000. For Chinese text, we first segment sentences by Jieba[5], and then apply Byte-Pair Encoding (BPE)[6] [25] with 32,000 merge operations. German texts are first tokenized using Moses tokenizer, followed by BPE with 32,000 merge operations. During training, we truncate the audio to 30 seconds for GPU memory efficiency, that is, 480,000 for raw wave signals or 3,000 fbank frames. We remove training samples whose translation text longer than 120 tokens or the percentage of aligned words to the original English transcription is less than 40% (produced by `fast-align`[7] toolkit).
Training For En-Zh, we use GigaST as the training set and use TED `dev2010` and `tst2015` [26] as the validation set. For En-De, MuST-C is added to the training set and we use MuST-C `dev` set for validation.
Evaluation Apart from the GigaST test set described in Section 2.22.2, for both language directions, we add two additional test sets: an in-house test set containing a total of 8.5 hours of news and tech talks (3,917 sentences) for En-Zh and MuST-C `tst-COMMON` set for En-De. The metric we use is case-sensitive detokenized BLEU[8].

3.2. End-to-end Models

We compare various end-to-end ST models of different sizes.
Speech-Transformer [2] is our benchmark model. The model

[4] https://github.com/google/sentencepiece
[5] https://github.com/fxsjy/jieba
[6] https://github.com/rsennrich/subword-nmt
[7] https://github.com/clab/fast_align
[8] https://github.com/mjpost/sacrebleu

Table 4: ***En-Zh BLEU scores. S, M, L*** *and* ***XL*** *stand for training data of various scales. The test set includes GigaST test set as created in Section 2.2 and the in-house test set.*

Models		# params	GigaST Test				In-house Test			
			S	M	L	XL	S	M	L	XL
Speech-Transformer	S-Transf_S	37 M	24.2	28.8	29.8	29.9	20.1	24.6	25.5	25.9
	S-Transf_M	90 M	23.3	31.4	33	33.6	19.4	26.6	28.5	29.4
	S-Transf_L	322 M	21.2	31.4	35	36.3	17.8	26.7	30.2	31.5
SSL-Transformer	w2v2-base	359 M	27.2	32.1	34.2	37.4	22.1	27.0	28.8	32.1
	w2v2-large	581 M	27.0	31.6	33.9	36.9	19.8	24.5	26.9	30.4
	hubert-base	359 M	27.6	31.8	34.0	37.2	22.5	25.9	29.3	32.1
	hubert-large	581 M	**30.1**	**33.4**	**35.6**	**38.0**	**24.4**	**27.9**	**30.3**	**32.5**

Table 5: ***En-De BLEU scores.*** *The training sets include* ***S/M/L/XL*** *subsets plus MuST-C En-De training set. The test sets are GigaST En-De test set and MuST-C* `tst-COMMON` *set.*

Models		# params	GigaST Test				MuST-C tst-COM			
			S	M	L	XL	S	M	L	XL
Speech-Transformer	S-Transf_S	35 M	21.5	24.1	25.1	25.6	24.4	25.7	25.8	25.1
	S-Transf_M	87 M	22.7	27.3	28.8	29.6	24.9	27.6	28.2	28.4
	S-Transf_L	315 M	21.3	27.5	31.1	32.6	23.2	27.8	29.6	30.1
SSL-Transformer	w2v2-base	359 M	24.3	28.0	32.1	33.4	26.5	28.0	29.9	30.3
	w2v2-large	581 M	23.6	28.5	28.5	33.0	23.4	26.9	27.0	30.2
	hubert-base	359 M	24.1	28.3	30.2	33.7	24.9	27.7	29.2	30.5
	hubert-large	581 M	**27.1**	**30.6**	**31.8**	**33.5**	**27.7**	**29.6**	**30.1**	**30.6**

uses fbank features as the input, and stacks two 3×3 CNN layers with stride size 2 and a transformer encoder-decoder module. Specifically, we implement three different model sizes, namely `S-Transf_S`, `S-Transf_M`, `S-Transf_L`, with model hyper-parameters listed in Table 6. We follow the setup of the optimizer and the learning rate schedule as in [5], as well as using ASR pre-training and SpecAugment [27].

Table 6: *Hyper-parameters for Speech-Transformer models*

	S	M	L
Hidden Size	256	512	1024
Filter Size	2048	2048	4096
Attention Heads	4	8	16
Encoder Layers	12	12	12
Decoder Layers	6	6	6

SSL-Transformer As recent research on self-supervised learning (SSL) in speech has intensified, pre-trained speech encoders, such as Wav2vec2 [10] and HuBERT [11], have been applied to downstream tasks instead of spectral features [28]. We can adopt a similar idea for the ST task. We evaluate SSL-Transformer, which replace Fbank features in Speech-Transformer with representations extracted from pre-trained speech encoders. To reduce the sequence length, we add two layers of convolutional subsampler with stride=2 after the SSL module, and apply the Transformer encoder-decoder as the downstream module. For the SSL speech encoders, we try four of them which performed well on the SUPERB leaderboard[9], namely `w2v2-base`, `w2v2-large`, `hubert-base`, and `hubert-large`. For downstream Transformer, we use the same hyperparameter as `S-Trans_L`, with 6 layers of encoder and decoder, $d_{\text{model}} = 1024, d_{\text{ff}} = 4096, d_{\text{head}} = 8$. Combining the above modules, we get four models with model parameter sizes of 359M, 581M, 359M, and 581M, respectively. For the subsequent model notations, since the structures of CNN and Transformer modules are the same, for simplicity, we only use the name of the speech encoders as the name of the whole model. The raw waveform of the entire speech is fed into the model, and SSL modules are **NOT** frozen during training. We set warmup steps at $25,000$ and peak learning rate at $2e^{-4}$.

Results The BLEU scores of the En-Zh and En-De test sets are shown in Tables 4 and 5 for different models trained with varying training sets. It is obvious that, for every models, the performance improves as the training data size increases. And model capacity often determines how good the translation is. In general, with the same amount of training data, the larger the model size, the better the performance. It is interesting to note that `S-Transf_S`, with only 35M parameters, fails to improve substantially as the data size increases from 1k to 10k hours. On the other hand, large models tend to underfit when the training set is small. For example, when training with the GigaST_S subset, `S-transf_L` performs worse than `S-transf_S`.

3.3. Cascade Systems

By concatenating the ASR models and the MT models, we obtain cascade systems. Specifically, our ASR model has the same structure as `S-transf_L` and we get 11.8 WER on the GigaSpeech test set, which is on par with other ASR systems on the GigaSpeech leaderboard[10]. For the MT model, we provide two models, one is Transformer-large trained using GigaSpeech

[9] `https://superbbenchmark.org/leaderboard`

[10] `https://github.com/SpeechColab/GigaSpeech`

transcription-translation bilingual text, noted as *constrained*, and the other is the MT model introduced in Section 2.1, noted as *unconstrained*. The former is for a fair comparison between the cascade systems and the end-to-end models with the same training data, while the latter has a stronger translation performance with BLEU scores higher than 40.

Table 7 shows the performances of the cascade systems. When the models are trained with the same amount of data, the performance of different end-to-end models are better than those of the cascade systems. However, the unconstrained cascade models, boosted by the larger amount of text training data, have a higher quality of text translation, which leads to better speech translation than the end-to-end models. Now with a powerful MT model, the gap in BLEU between end-to-end and cascade models is around 2. We still have room to improve the end-to-end performance through pre-training of decoder, multi-task training and other techniques. We leave these for future work.

Table 7: *The BLEU scores of cascade models on GigaST test sets*

Direction	MT Condition	MT BLEU	ST BLEU
En-Zh	constrained	24.9	22.3
	unconstrained	44.3	39.8
En-De	constrained	37.6	33.4
	unconstrained	42.2	35.7

3.4. Analysis on Speech Representations

In addition to the results of the baseline Speech-Transformers, we also investigate the performance of pre-trained speech encoders under large-scale speech translation. Are these pre-trained encoders still effective and helpful under large amounts of training data?

Before answering this question, we need to figure out: how should these pre-trained speech encoders be incorporated into the training? Should they be frozen with parameters not updated, or should the whole model be fine-tuned based on ST supervised data? Taking two pre-trained encoders, `hubert-base` and `hubert-large`, as examples, Table 8 shows the results of En-Zh translation with and without the encoder frozen. It shows that with both base and large models, the frozen case perform much worse than the unfrozen case, with an average difference of as much as 2.6 BLEU. This differs from the practice of freezing speech encoders in the downstream ASR task, where freezing these encoders can still yield a word error rate of as low as 3.4 on LibriSpeech [28], but in ST tasks, we recommend fine-tuning the speech encoders together with other components, which can greatly improve ST performance. Therefore, we conduct the rest of our experiments without freezing the speech encoders. Detailed setups are introduced in Section 3.2.

The results of En-Zh and En-De experiments are summarized in Tables 4 and 5. It can be seen that SSL-Transformer models are generally better than the Speech-Transformer models. Even though `S-Transf_L` and `w2v2-base` have roughly the same order of magnitude of parameters (300m parameters), `w2v2-base` outperforms `S-Transf_L`. When the training data size is increased to 10,000 hours, according to the BLEU scores throughout the XL columns, the pre-trained speech module continues to play an essential role in improving ST performance. In addition, looking at the performances of the models

Table 8: *To freeze or not to freeze the SSL speech encoders?* `hubert-base` *and* `hubert-large` *were finetuned with GigaST XL subset and tested on the En-Zh test sets.*

SSL repr.	freeze?	GigaST	In-house	Avg.
`hubert-base`	✓	34.4	29.7	32.1
	✗	37.2	32.1	34.7
`hubert-large`	✓	35.3	30.0	32.7
	✗	**38.0**	**32.5**	**35.3**

trained with the S subset (only 250 hours), pre-trained speech encoders are particularly useful. Take En-Zh translation as an example (Table 4), using `hubert-Large` to train 250 hours of speech, the translation performance can reach or even exceed `S-Transf_S` training on 10,000 hours. Meanwhile, as the data size increases, we find that the gain from the pre-trained speech encoders for the downstream ST task becomes smaller. Despite the reduced benefit, we observe that pre-trained speech encoders are still useful in our setup.

On the other hand, there are performance differences between the four pre-trained speech encoders. Analyzing the performance of En-Zh and En-De translations comprehensively, the rank among the four is `hubert-large`, `hubert-base`, `w2v2-base`, and `w2v2-large`. It is surprising to see that `w2v2-large` (581M parameters) performs worse than the base models (359M parameters), where it violates the common belief that larger model capacity means more representation capability and with better performance. Overall, HuBERT models empirically perform better than Wav2vec2 models in our setup.

4. Conclusion

This paper presents GigaST, a new speech-to-text corpus suitable for training and evaluating ST systems. Our corpus is created by translating the transcript in GigaSpeech, which is one of the largest open-source English ASR corpora. We have demonstrated that models trained with an addition of our corpus can obtain new state-of-the-art results on the MuST-C English-German benchmark test set. We also establish new benchmark test sets for the two language directions. We release all the training data, human-translated test sets and our training scripts so that others can easily replicate our results. We believe our released dataset will open new avenues in speech translation research.

5. References

[1] A. Bérard, O. Pietquin, C. Servan, and L. Besacier, "Listen and translate: A proof of concept for end-to-end speech-to-text translation," in *NIPS workshop on End-to-end Learning for Speech and Audio Processing*, 2016.

[2] L. Dong, S. Xu, and B. Xu, "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition," in *Proc. of ICASSP*, 2018, pp. 5884–5888.

[3] H. Inaguma, S. Kiyono, K. Duh, S. Karita, N. Yalta, T. Hayashi, and S. Watanabe, "ESPnet-ST: All-in-one speech translation toolkit," in *Proc. of ACL*, 2020, pp. 302–311.

[4] C. Wang, Y. Tang, X. Ma, A. Wu, D. Okhonko, and J. Pino, "Fairseq s2t: Fast speech-to-text modeling with fairseq," in *Proc. of AACL*, 2020, pp. 33–39.

[5] C. Zhao, M. Wang, Q. Dong, R. Ye, and L. Li, "NeurST: Neural speech translation toolkit," in *Proc. of ACL - System Demonstrations*, Aug. 2021.

[6] L. Duong, A. Anastasopoulos, D. Chiang, S. Bird, and T. Cohn, "An attentional model for speech translation without transcription," in *Proc. of NAACL-HLT*, 2016, pp. 949–959.

[7] G. Chen, S. Chai, G.-B. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang, M. Jin, S. Khudanpur, S. Watanabe, S. Zhao, W. Zou, X. Li, X. Yao, Y. Wang, Z. You, and Z. Yan, "GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000 Hours of Transcribed Audio," in *Proc. Interspeech 2021*, 2021, pp. 3670–3674.

[8] M. A. Di Gangi, R. Cattoni, L. Bentivogli, M. Negri, and M. Turchi, "MuST-C: a Multilingual Speech Translation Corpus," in *Proc. of NAACL-HLT*, 2019, pp. 2012–2017.

[9] E. Salesky, M. Wiesner, J. Bremerman, R. Cattoni, M. Negri, M. Turchi, D. W. Oard, and M. Post, "The multilingual tedx corpus for speech recognition and translation," in *Proc. of Interspeech*, 2021.

[10] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. of NeurIPS*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020.

[11] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[12] F. Akhbardeh, A. Arkhangorodsky, M. Biesialska, O. Bojar, R. Chatterjee, V. Chaudhary, M. R. Costa-jussa, C. España-Bonet, A. Fan, C. Federmann, M. Freitag, Y. Graham, R. Grundkiewicz, B. Haddow, L. Harter, K. Heafield, C. Homan, M. Huck, K. Amponsah-Kaakyire, J. Kasai, D. Khashabi, K. Knight, T. Kocmi, P. Koehn, N. Lourie, C. Monz, M. Morishita, M. Nagata, A. Nagesh, T. Nakazawa, M. Negri, S. Pal, A. A. Tapo, M. Turchi, V. Vydrin, and M. Zampieri, "Findings of the 2021 conference on machine translation (WMT21)," in *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics, Nov. 2021, pp. 1–88. [Online]. Available: https://aclanthology.org/2021.wmt-1.1

[13] Y. Jia, M. Johnson, W. Macherey, R. J. Weiss, Y. Cao, C.-C. Chiu, N. Ari, S. Laurenzo, and Y. Wu, "Leveraging weakly supervised data to improve end-to-end speech-to-text translation," in *Proc. of ICASSP*. IEEE, 2019.

[14] K. Kuligowska and B. Kowalczuk, "Pseudo-labeling with transformers for improving question answering systems," *Procedia Computer Science*, vol. 192, pp. 1162–1169, 2021, knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 25th International Conference KES2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050921016082

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. of NeurIPS*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 5998–6008.

[16] B. Li, Y. Li, C. Xu, Y. Lin, J. Liu, H. Liu, Z. Wang, Y. Zhang, N. Xu, Z. Wang, K. Feng, H. Chen, T. Liu, Y. Li, Q. Wang, T. Xiao, and J. Zhu, "The NiuTrans machine translation systems for WMT19," in *Proceedings of the Fourth Conference on Machine Translation*, 2019.

[17] L. Wu, X. Pan, Z. Lin, Y. Zhu, M. Wang, and L. Li, "The volctrans machine translation system for wmt20," in *Proceedings of the Fifth Conference on Machine Translation (Volume 2: Shared Task Papers)*, Nov. 2020.

[18] Y. Kim and A. M. Rush, "Sequence-level knowledge distillation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016*, 2016.

[19] M. Freitag, Y. Al-Onaizan, and B. Sankaran, "Ensemble distillation for neural machine translation," *CoRR*, vol. abs/1702.01802, 2017.

[20] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*, 2016.

[21] R. Bawden, R. Sennrich, A. Birch, and B. Haddow, "Evaluating discourse phenomena in neural machine translation," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 1304–1313. [Online]. Available: https://aclanthology.org/N18-1118

[22] S. Läubli, R. Sennrich, and M. Volk, "Has machine translation achieved human parity? a case for document-level evaluation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 4791–4796. [Online]. Available: https://aclanthology.org/D18-1512

[23] Z. Sun, M. Wang, H. Zhou, C. Zhao, S. Huang, J. Chen, and L. Li, "Capturing longer context for document-level neural machine translation: A multi-resolutional approach," *CoRR*, vol. abs/2010.08961, 2020.

[24] F. Akhbardeh, A. Arkhangorodsky, M. Biesialska, O. Bojar, R. Chatterjee, V. Chaudhary, M. R. Costa-jussa, C. España-Bonet, A. Fan, C. Federmann, M. Freitag, Y. Graham, R. Grundkiewicz, B. Haddow, L. Harter, K. Heafield, C. Homan, M. Huck, K. Amponsah-Kaakyire, J. Kasai, D. Khashabi, K. Knight, T. Kocmi, P. Koehn, N. Lourie, C. Monz, M. Morishita, M. Nagata, A. Nagesh, T. Nakazawa, M. Negri, S. Pal, A. A. Tapo, M. Turchi, V. Vydrin, and M. Zampieri, "Findings of the 2021 conference on machine translation (WMT21)," in *Proceedings of the Sixth Conference on Machine Translation*. Online: Association for Computational Linguistics, Nov. 2021, pp. 1–88. [Online]. Available: https://aclanthology.org/2021.wmt-1.1

[25] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proc. of ACL*, 2016, pp. 1715–1725.

[26] Y. Liu, H. Xiong, J. Zhang, Z. He, H. Wu, H. Wang, and C. Zong, "End-to-end speech translation with knowledge distillation," in *Proc. of INTERSPEECH*, G. Kubin and Z. Kacic, Eds., 2019, pp. 1128–1132.

[27] 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," in *Proc. of INTERSPEECH*, 2019.

[28] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin *et al.*, "Superb: Speech processing universal performance benchmark," in *Proc. of INTERSPEECH*, 2021.