Interspeech 2022
18-22 September 2022, Incheon, Korea

Investigating Parameter Sharing in Multilingual Speech Translation

Qian Wang[1,2], *Chen Wang*[1,2], *Jiajun Zhang*[1,2]

[1]National Laboratory of Pattern Recognition, Institute of Automation, CAS
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
qian.wang@nlpr.ia.ac.cn, wangchen2020@ia.ac.cn, jjzhang@nlpr.ia.ac.cn

Abstract

End-to-end multilingual speech translation (ST) directly models the mapping from the speech in source languages to the text of multiple target languages. While multilingual neural machine translation has been proved effective in modeling the general knowledge with shared parameters and handling inter-task interference with language-specific parameters, it still lacks exploration of when and where parameter sharing matters in multilingual ST. This work offers such a study by proposing a comprehensive analysis on the influence of various heuristically designed sharing strategies. We further investigate the inter-task interference through gradient similarity between different tasks, and improve the parameter sharing strategy in multilingual ST under the guidance of inter-task gradient similarity. Experimental results on the one-to-many MuST-C dataset have shown that the gradient-guided sharing method can significantly improve the translation quality with a comparable or even lower cost in terms of parameter scale.

Index Terms: automatic speech translation, parameter sharing, multilingual translation

1. Introduction

End-to-end speech translation (ST) aims to translate the speech of one source language into the text of another target language [1]. Compared to traditional pipeline systems which are composed of automatic speech recognition and text machine translation, end-to-end ST model can reduce time delay and error propagation [2], so it has attracted intensive attention in recent years. In parallel to bilingual ST, multilingual ST, where a single model is used to translate source language speech into multiple target languages, becomes a new trend since it further reduces the cost of training and deploying compared to its bilingual counterparts.

Currently, the multilingual ST model often follows the *completely shared* paradigm, in which different languages share the same parameters. Despite the simplicity, the multilingual model also faces a capacity bottleneck on high-resource language pairs, which could hurt the translation accuracy [3]. Therefore, finding a proper trade-off between bilingual model and multilingual can potentially break the capacity bottleneck and improve model quality. In the field of multilingual neural machine translation (NMT), researchers have investigated various parameter sharing strategies, e.g., heuristically splitting the model parameters into shared parts and language-specific parts [4] or dynamically changing shared parameters into language-specific types [5]. The multilingual model using a mixture of shared and language-specific parameters can both preserve general knowledge and alleviate interference among different languages.

Corresponding Author: Jiajun Zhang.

While the parameter sharing methods have shown promising performances in multilingual NMT, we provide the first comprehensive study of different heuristically designed parameter sharing strategies in the multilingual ST model. To further improve the translation quality, we investigate the multi-task optimization trajectory via the gradients of different target languages on each parameter, and adopt the recently proposed gradient-guided method in multilingual text translation which improves the sharing strategy under the guidance of gradient similarities [5]. In the gradient-guided method, the parameters with opposite gradient directions are marked as language-specific, while other parameters with similar gradients are shared by different target languages. Experiments on the one-to-many MuST-C multilingual ST dataset show that the multi-task optimization trajectory in multilingual ST is quite different to multilingual NMT, and the traditional heuristically designed sharing strategies can only benefit from larger parameter scales compared to the complete shared model. On the other hand, the gradient-guided sharing strategy provides more fine-grained control of parameter sharing and significantly outperforms the strong baselines with comparable parameter scales.

2. Related Work

End-to-end Speech Translation is the task of converting speech utterances to their translations in other language without generating the intermediate transcriptions [1]. Although the end-to-end ST model has achieved great success in recent years [6, 7], it still suffers from data scarcity. Many techniques, such as pretraining [8, 9, 10], multi-task learning [11, 12], knowledge distillation [2], and generating synthesis data [13, 14] have been proposed to exploit the data from related tasks. Performing multilingual translation also alleviates data scarcity since it transfers knowledge across different languages, and has been proved effective in speech translation [15, 16]. [3] focuses on efficient fine-tuning with pretrained multilingual models, and the performance is comparable to supervised learning in the zero-shot translation directions. To the best of our knowledge, existing multilingual ST methods only consider a completely shared model but ignore the language specificity.

Multilingual Neural Machine Translation aims at achieving translation between multiple languages in a single model [17]. In the early stage, researchers share different modules to reduce the parameter scales in bilingual models. [18] uses a shared encoder and separate decoders for one-to-many translation. [19] shares the attention module to bridge separate encoders and decoders to enable many-to-many translation. More aggressively, [20] proposes a completely shared model, which significantly reduces the parameter scale and enables effective knowledge transfer among languages, but lacks the ability for retaining language-specific knowledge. To improve the model capacity while preserving effective knowledge transfer,

10.21437/Interspeech.2022-552

researchers resort to manually designed language-specific modules with parameter sharing strategies [4, 21]. To work around the limitation that needs to be manually designed for specific modules, [5] proposes parameter differentiation that dynamically transfers shared parameters into language-specific ones. Since using a mixture of shared and language-specific parameters has been proved effective in multilingual text translation, this work offers the first comprehensive study of various parameter sharing strategies in multilingual speech translation.

3. Parameter Sharing in Multilingual ST

In this section, we first briefly introduce the backbone multilingual ST model (Section 3.1), followed by heuristically designed parameter sharing strategy which pre-defines the shared components of multiple decoders (Section 3.2). Finally, we describe the gradient-based sharing strategy that provides a more fine-grained parameter sharing configuration without manual design (Section 3.3).

3.1. The Backbone Multilingual ST Model

As shown in Figure 1, we use the Transformer [22] as our backbone model as it has been proved successful in speech translation [7]. Different from the Transformer encoder used in machine translation, a small convolutional neural network is prepended to the encoder to downsample the speech feature sequence. The remaining part of the encoder includes multiple identical layers and each layer is composed of two sub-layers: the self-attention sub-layer and the feed-forward sub-layer. The decoder is also stacked with multiple layers and each layer has three sub-layers: the self-attention sub-layer, the encoder-decoder attention sub-layer, and the feed-forward sub-layer.

As shown in the right part of Figure 1, in the decoder layer, each self-attention sub-layer includes 4 parameter matrices: the query projection W_Q^1, the key projection W_K^1, the value projection W_V^1, and the final projection W_F^1. Similarly, the parameters in each encoder-decoder attention sub-layer are W_Q^2, W_K^2, W_V^2, W_F^2. The feed-forward sub-layer includes two linear projections: W_{L_1} and W_{L_2}, where the former projects the hidden representation into a wider dimension and the latter projects the output back to model dimension.

3.2. Heuristic Sharing Strategy

We first investigate the effectiveness of different heuristically designed sharing strategies in multilingual ST. Since this work mainly focuses on one-to-many translation, the most intuitive setting is using a shared encoder for source language speech and individual decoders for each target language. Next, we move a step forward and introduce more shared parameters among the different decoders to find a proper trade-off between this base setting and the *complete sharing* model in which all decoders share the same set of parameters.

Following the practice in multilingual text translation [4], the parameter sharing strategies explored in this work are described below:

- θenc: The base case includes a shared encoder and individual decoders for each target language.
- $\theta enc, W_Q^1, W_K^1, W_V^1, W_F^1$: Apart from the encoder, the self-attention sub-layers of the decoders are shared among different target languages.
- $\theta enc, W_Q^2, W_K^2, W_V^2, W_F^2$: The parameters in the encoder-decoder attention sub-layers are shared.

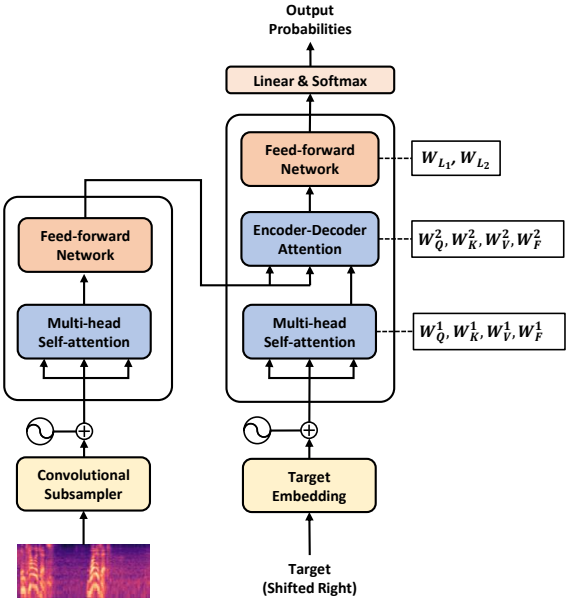

Figure 1: *The Transformer based speech translation model.*

- $\theta enc, W_K, W_V$: Sharing the key projection and the value projection in the self-attention sub-layer and the encoder-decoder attention.
- $\theta enc, W_Q, W_K$: Sharing the query projection and the key projection in the self-attention sub-layer and the encoder-decoder attention.
- $\theta enc, W_Q, W_K, W_V$: Sharing the query projection, the key projection, and the value projection in the self-attention sub-layer and the encoder-decoder attention.
- $\theta enc, W_Q, W_K, W_V, W_F$: The self-attention sub-layer and the encoder-decoder attention sub-layer are shared among different languages.
- $\theta enc, W_{L_1}, W_{L_2}$: Sharing only the feed-forward sub-layer while keeping the attention sub-layers language-specific.

3.3. Gradient-Guided Sharing Strategy

The heuristically designed method can help to find a better sharing strategy through extensive experiments, but it still lacks flexibility and fine-grained control of parameter sharing. As for the flexibility, the sharing configurations are identical in each decoder layer, which ignores the difference between different layers since previous studies have proved that morphological information evolves from shallower to deeper layers [23]. As for the granularity, each parameter matrix can only be in two states: shared by all target languages or completely language-specific. However, considering the complicated relationship between languages, the optimal sharing strategy may include parameters shared by only parts of the languages while keeping them language-specific for other languages.

To improve the flexibility and find the optimal parameter sharing configuration, we investigate the optimization trajectory of different target languages and adopt the recently proposed parameter differentiation method in multilingual NMT that dynamically changes the shared parameters into more specific types based on inter-task gradient similarities [5], which is referred to as gradient-guided sharing strategy.

	Method	#Parameter	DE	ES	FR	IT	NL	PT	RO	RU	Avg.
Adapter	Multilingual†	76.3M (1.0X)	24.18	28.28	34.98	24.62	28.80	31.13	23.22	15.88	26.39
	Fine-tuning†	610.4M (8.0X)	24.50	28.67	34.89	24.82	28.38	30.73	23.78	16.23	26.50
	Adapter†	153.1M (2.0X)	24.63	28.73	34.75	24.96	28.80	30.96	23.70	16.36	26.61
LNA	LNA-D†	76.3M (1.0X)	24.16	28.30	34.52	24.46	28.35	30.51	23.29	15.84	26.18
	LNA-E†	76.3M (1.0X)	24.34	28.25	34.42	24.24	28.46	30.53	23.32	15.89	26.18
	LNA-E,D†	76.3M (1.0X)	24.27	28.40	34.61	24.44	28.25	30.53	23.27	15.92	26.21
Baseline	Bilingual‡	610.4M (8.0X)	22.70	27.20	32.90	22.70	27.30	28.10	21.90	15.30	24.76
	Multilingual‡	76.3M (1.0X)	24.50	28.20	34.90	24.60	28.60	31.10	23.80	16.00	26.46
	Jointly Fine-tuning	76.3M (1.0X)	24.53	28.34	34.91	24.46	28.73	31.25	23.55	15.97	26.47
Heuristic	θ_{enc}	324.6M (4.3X)	25.17	28.85	35.33	<u>25.10</u>	<u>29.31</u>	<u>31.57</u>	23.97	<u>16.47</u>	<u>26.97</u>
	$\theta_{enc}, W_Q^1, W_K^1, W_V^1, W_F^1$	208.8M (2.7X)	25.12	28.84	<u>35.52</u>	24.77	29.14	31.40	24.09	16.25	26.89
	$\theta_{enc}, W_Q^2, W_K^2, W_V^2, W_F^2$	208.8M (2.7X)	25.09	28.58	35.06	24.90	29.01	31.33	23.92	16.12	26.75
	θ_{enc}, W_K, W_V	208.8M (2.7X)	25.20	28.50	35.35	24.92	29.11	31.39	23.92	16.34	26.84
	θ_{enc}, W_Q, W_K	208.8M (2.7X)	<u>25.25</u>	28.78	35.24	25.00	28.98	31.44	23.94	16.23	26.86
	$\theta_{enc}, W_Q, W_K, W_V$	186.7M (2.4X)	24.86	28.79	35.10	24.80	28.94	31.17	<u>24.16</u>	16.18	26.75
	$\theta_{enc}, W_Q, W_K, W_V, W_F$	164.7M (2.2X)	24.88	28.72	34.99	24.78	28.83	31.48	23.66	16.26	26.70
	$\theta_{enc}, W_{L_1}, W_{L2}$	164.7M (2.2X)	24.99	<u>28.88</u>	35.26	24.76	29.01	31.33	23.56	16.21	26.75
Random	1.5 X	114.0M	24.80	28.59	35.20	24.78	28.91	31.24	24.11	16.02	26.71
	2.0 X	156.3M	24.76	28.67	35.50	24.80	29.08	31.38	23.87	16.08	26.77
	2.5 X	190.2M	25.09	28.63	35.23	24.96	29.29	31.37	23.80	16.09	26.81
	3.0 X	228.5M	24.94	28.90	35.30	24.90	29.01	31.63	24.18	16.40	26.91
	3.5 X	252.9M	25.22	28.49	35.37	25.11	29.33	31.49	24.09	16.18	26.91
Grad-Guided	1.5 X	114.3M	24.99	29.26	35.60	25.36	29.44	31.93	23.95	16.41	27.12
	2.0 X	156.2M	25.26	29.16	35.43	25.52	29.64	31.54	23.85	16.90	27.16
	2.5 X	190.6M	25.37	29.81	35.88	25.61	29.99	**32.12**	24.68	**17.22**	27.59
	3.0 X	228.9M	25.57	**29.94**	35.97	**25.89**	**30.09**	31.82	24.50	17.10	**27.61**
	3.5 X	252.9M	**25.84**	29.58	**36.05**	25.58	30.04	31.95	**24.79**	16.82	27.58

Table 1: *BLEU scores on the MuST-C dataset.* † *indicates that the corresponding results are taken from [3], while the results with* ‡ *are taken from [7]. We use the pretrained model of Multilingual*‡ *to initialize the parameters of all the following models. The best results of heuristically designed sharing strategy are <u>underlined</u> and **bold** indicates the best result of all methods.*

Specifically, the gradient-guided approach involves the following steps. We first build the model as completely shared and initialize the parameters with a pretrained model[1]. Second, we calculate the gradients of each task on each parameters with a multi-way aligned data[2], and evaluate the pairwise gradient cosine similarities s between tasks t_j and t_k on each parameter matrices θ_i to obtain a quadruplet (θ_i, t_j, t_k, s). Third, we share the parameters by tasks with higher similarity and make others language-specific, e.g., the parameter θ_i is shared by tasks t_j and t_k if the corresponding similarity s if higher. Finally, we continue to train the multilingual ST model on the combined data until convergence.

4. Experiments

4.1. Dataset and Settings

We conduct experiments on the one-to-many MuST-C multilingual ST dataset [25], which contains up to around 504 hours of English speech to text of 8 European languages including German (DE), Spanish (ES), French (FR), Italian (IT), Dutch (NL), Portuguese (PT), Romanian (RO), and Russian (RU). We use the MuST-C *dev* data for validation and the *tst-COMMON* for evaluation. The multi-way aligned data includes $1,208$ speech utterances and corresponding translations in 8 target languages. We preprocess the audio signals following [7], in which 80-dimensional log-mel filterbank features with $25ms$ window size and $10ms$ step size are extracted from raw audio files. The training samples which are larger than $3,000$ frames are removed. We use a shared target side vocabulary with a size of $10K$ jointly learned with unigram encoding algorithm [26].

We conduct our experiments with the Transformer architecture and adopt the *speech_transformer* setting in fairseq[3], which includes 12 encoder layers and 6 decoder layers. The model dimension is set to 512 and the inner dimension of the feed-forward sub-layer is set to $2,048$. After pretraining, we fine-tune all models with gradient accumulation of 8 steps to simulate multi-GPU training and each step contains a batch of up to $40,000$ source frames and target tokens. During decoding, we use beam search decoding with a beam size of 5 and a length penalty of 0.6 to obtain the final translations.

4.2. Results

The results are shown in Table 1. We first investigate the baseline methods as well as other multilingual ST methods such as the adapter tuning [3] and the LNA tuning [27]. It is obvious that the *Multilingual* model outperforms the *Bilingual* model by

[1]We use the pretrained model in the public fairseq S2T repository (`https://github.com/pytorch/fairseq/blob/main/examples/speech_to_text/docs/mustc_example.md`).

[2]Similar to [5, 24], we build the multi-aligned data including translations in all target languages for each source language speech to minimize the gradient variance caused by inconsistent sentence semantics.

[3]`https://github.com/pytorch/fairseq`

up to 1.7 BLEU and significantly reduces the parameter scale (from 610.4M to 76.3M). On the other hand, introducing additional parameters (Adapter) or continuing training for more steps (individually Fine-tuning with bilingual data or Jointly Fine-tuning with multilingual data) can only bring negligible improvement on translation quality (by up to 0.22 BLEU).

We then compare heuristically designed parameter sharing strategies with the baselines. We find that using a shared encoder for the English speech and individual decoders for each target language text (θ_{enc}) performs best but maintains the largest model size. The results of different heuristic sharing methods also show a positive correlation between model size and translation quality, i.e., a larger model usually performs better than a smaller model.

To further investigate whether the performance gain in heuristic sharing methods comes from larger parameter scales, we design a *Random* sharing strategy, in which the model randomly selects parameters and share the parameters with randomly selected tasks until the model size reaches a threshold (e.g., 1.5X of the original model). By comparing the rows of *Random* and *Heuristic* in Table 1, we find that the *Random* sharing strategy can also bring comparable translation quality with the heuristically designed strategies. The results prove that sharing predefined specific modules does not apply to multilingual ST, though it works out in multilingual NMT [4].

Finally, we evaluate the gradient-guided sharing strategy with different parameter scales. Different from the *Random* sharing strategy, the model selects parameters and tasks for sharing following the order of pairwise gradient cosine similarities instead of random selection. From the rows of *Grad-Guided*, we can easily find that the gradient-guided sharing strategy significantly outperforms the other multilingual ST methods by up to 1.15 BLEU (27.61 v.s. 26.46). When it comes to similar model sizes, e.g., the *Adapter*, the $\theta_{enc}, W_{L_1}, W_{L_2}$ in *Heuristic*, the *Random* $2.0X$ and the *Grad-Guided* $2.0X$, the gradient-guided sharing strategy also performs better than other methods.

4.3. Analyses

4.3.1. The Effects of Model Size in Gradient-Guided Strategy

From the results of *Heuristic* and *Random* sharing strategies, we find a positive correlation between model size and translation quality. However, a larger parameter scale indicates the model includes more language-specific parameters which may hurt the positive knowledge transfer among languages. To investigate the optimal trade-off between language-specific parameters and shared parameters, we alter the model size in the *Grad-Guided* sharing strategy. As shown in the rows of *Grad-Guided* in Table 1, the translation quality increases with the model size grows from 114.3M to 228.9M, and then decreases with more language-specific parameters involved. The model with 228.9M parameters performs best and the maximum marginal utility comes with 2.5X model size which gains 0.43 BLEU compared to the 2.0X model.

4.3.2. Why Heuristic Sharing Fails in Multilingual ST

In the literature of multilingual text translation, researchers tend to explore layer-agnostic parameter sharing strategies and different layers share the same topological structure. However, in multilingual ST, the inter-task relationship between layers may be more distinct than that between sub-layers. We hereby investigate how gradient similarities evolve across layers.

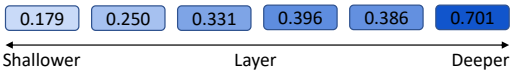

Figure 2: *Gradient cosine similarity averaged over all target languages across decoder layers.*

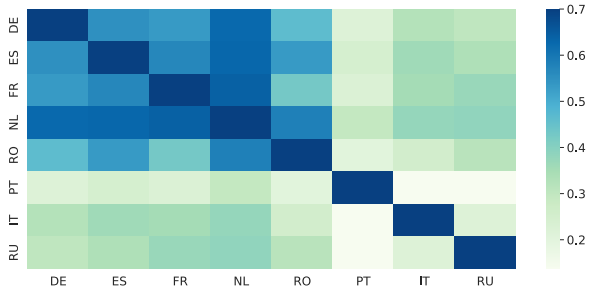

Figure 3: *Cosine similarity of gradients averaged across layers between different target languages.*

As shown in Figure 2, the gradient is more similar in deeper layers (0.701 of the last layer v.s. 0.179 of the first layer). This observation is different from the one-to-many text translation in which the middle layers share more inter-task gradient similarities, and proves the importance to distinguish different layers in multilingual ST parameter sharing. On the other hand, the gradient-guided sharing strategy provides a more fine-grained control of parameter sharing and brings significant improvement over the heuristically designed strategies.

4.3.3. Gradient Similarity and Linguistic Proximity

Intuitively, the gradients should be more similar between close languages (measured by linguistic feature like language family or language branch). We visualize the correlation between gradient cosine similarities and different languages in Figure 3. Specifically, we find that the languages from the same family (like German and Dutch of Germanic languages) share similar gradients while the ones from different family (like German of Germanic language and Russian of Slavic language) share distinct gradients. However, there also exist exceptions like Portuguese, which indicates a more sophisticated relationship among languages in the multilingual ST model.

5. Conclusion

In this work, we empirically explore various parameter sharing methods in multilingual ST and show that the heuristically designed sharing strategy in multilingual text translation cannot work well in multilingual speech translation. We further explain how and why the heuristic sharing fails by analyzing the difference between multilingual text translation and multilingual speech translation from the view of multi-task optimization trajectory. From the above observation, we adopt the gradient-guided method which use the inter-task gradient similarity to improve the sharing strategy. The results and analyses show that the gradient-guided method can significantly improve the translation quality over strong baselines.

6. Acknowledgement

This work is supported by the Natural Science Foundation of China under Grant No. 62122088, U1836221, and 62006224.

7. References

[1] A. Bérard, O. Pietquin, L. Besacier, and C. Servan, "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] Y. Liu, H. Xiong, J. Zhang, Z. He, H. Wu, H. Wang, and C. Zong, "End-to-end speech translation with knowledge distillation," in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria*, 2019.

[3] H. Le, J. M. Pino, C. Wang, J. Gu, D. Schwab, and L. Besacier, "Lightweight adapter tuning for multilingual speech translation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, (Volume 2: Short Papers)*, 2021.

[4] D. S. Sachan and G. Neubig, "Parameter sharing methods for multilingual self-attentional translation models," in *Proceedings of the Third Conference on Machine Translation: Research Papers, Belgium, Brussels*, 2018.

[5] Q. Wang and J. Zhang, "Parameter differentiation based multilingual neural machine translation," *CoRR*, vol. abs/2112.13619, 2021.

[6] H. Inaguma, S. Kiyono, K. Duh, S. Karita, N. Yalta, T. Hayashi, and S. Watanabe, "Espnet-st: All-in-one speech translation toolkit," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020.

[7] C. Wang, Y. Tang, X. Ma, A. Wu, D. Okhonko, and J. M. Pino, "Fairseq S2T: fast speech-to-text modeling with fairseq," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, 2020.

[8] 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," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, Volume 1 (Long and Short Papers)*, 2019.

[9] A. Alinejad and A. Sarkar, "Effectively pretraining a speech translation decoder with machine translation data," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.

[10] C. Wang, Y. Wu, S. Liu, Z. Yang, and M. Zhou, "Bridging the gap between pre-training and fine-tuning for end-to-end speech translation," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, New York, NY, USA*, 2020.

[11] A. Anastasopoulos and D. Chiang, "Tied multitask learning for neural speech translation," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, Louisiana, USA, Volume 1 (Long Papers)*, 2018.

[12] Y. Tang, J. M. Pino, C. Wang, X. Ma, and D. Genzel, "A general multi-task learning framework to leverage text data for speech to text tasks," in *IEEE International Conference on Acoustics, Speech and Signal Processing, Toronto, ON, Canada*, 2021.

[13] Y. Jia, M. Johnson, W. Macherey, R. J. Weiss, Y. Cao, C. Chiu, N. Ari, S. Laurenzo, and Y. Wu, "Leveraging weakly supervised data to improve end-to-end speech-to-text translation," in *IEEE International Conference on Acoustics, Speech and Signal Processing, Brighton, United Kingdom*, 2019.

[14] C. Wang, A. Wu, J. Pino, A. Baevski, M. Auli, and A. Conneau, "Large-scale self- and semi-supervised learning for speech translation," in *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia*, 2021.

[15] H. Inaguma, K. Duh, T. Kawahara, and S. Watanabe, "Multilingual end-to-end speech translation," in *IEEE Automatic Speech Recognition and Understanding Workshop, Singapore*, 2019.

[16] M. A. D. Gangi, M. Negri, and M. Turchi, "One-to-many multilingual end-to-end speech translation," in *IEEE Automatic Speech Recognition and Understanding Workshop, Singapore*, 2019.

[17] R. Dabre, C. Chu, and A. Kunchukuttan, "A survey of multilingual neural machine translation," *ACM Comput. Surv.*, vol. 53, no. 5, 2020.

[18] D. Dong, H. Wu, W. He, D. Yu, and H. Wang, "Multi-task learning for multiple language translation," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Beijing, China, Volume 1: Long Papers*, 2015.

[19] O. Firat, K. Cho, and Y. Bengio, "Multi-way, multilingual neural machine translation with a shared attention mechanism," in *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA*, 2016.

[20] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. B. Viégas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean, "Google's multilingual neural machine translation system: Enabling zero-shot translation," *Trans. Assoc. Comput. Linguistics*, vol. 5, 2017.

[21] Y. Wang, J. Zhang, L. Zhou, Y. Liu, and C. Zong, "Synchronously generating two languages with interactive decoding," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China*, 2019.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA*, 2017.

[23] Y. Belinkov, N. Durrani, F. Dalvi, H. Sajjad, and J. R. Glass, "What do neural machine translation models learn about morphology?" in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, Canada, Volume 1: Long Papers*, 2017.

[24] Z. Wang, Y. Tsvetkov, O. Firat, and Y. Cao, "Gradient vaccine: Investigating and improving multi-task optimization in massively multilingual models," in *9th International Conference on Learning Representations, Virtual Event, Austria*, 2021.

[25] M. A. D. 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, Minneapolis, MN, USA, Volume 1 (Long and Short Papers)*, 2019.

[26] 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, EMNLP 2018: System Demonstrations, Brussels, Belgium*, 2018.

[27] X. Li, C. Wang, Y. Tang, C. Tran, Y. Tang, J. M. Pino, A. Baevski, A. Conneau, and M. Auli, "Multilingual speech translation from efficient finetuning of pretrained models," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, (Volume 1: Long Papers), Virtual Event*. Association for Computational Linguistics, 2021.