Interspeech 2024
1-5 September 2024, Kos, Greece

Knowledge-Preserving Pluggable Modules for Multilingual Speech Translation Tasks

*Nan Chen, Yonghe Wang, Feilong Bao**

College of Computer Science, Inner Mongolia University, China
Inner Mongolia Key Laboratory of Mongolian Information Processing Technology
chennannlp@gmail.com, cswyh@imu.edu.cn, csfeilong@imu.edu.cn

Abstract

Multilingual speech translation tasks typically employ retraining, regularization, or resampling methods to add new languages. Retraining the model significantly increases training time and cost. Moreover, using existing regularization or resampling methods to balance performance between new and original languages might lead to catastrophic forgetting. This can degrade the translation performance of the existing languages. To mitigate the above issues, we store the knowledge of new languages in additional models. We then introduce them as pluggable modules into existing multilingual speech translation models. This approach does not significantly increase training costs and affect the translation performance of existing models. The experimental results demonstrate that our method improves the translation performance of new languages without affecting existing translation tasks. Our code is available at https://github.com/myaxxxxx/transfer-st.

Index Terms: speech translation, transfer learning, feed-forward module, adapters

1. Introduction

Multilingual speech-to-text translation tasks [1, 2, 3], which use a single model to translate in multiple directions, have been successful in recent years. However, multilingual speech translation models often need to be retrained when new languages are added [4, 5], significantly increasing training time and cost. In addition, directly training a multilingual translation model with new language data may result in catastrophic forgetting [6] due to differences in knowledge representation, resulting in a decline in the translation performance of the original languages. Therefore, exploring new training methods has become the focus of current multilingual speech translation research.

Multilingual speech translation models integrating new translation languages have been greatly developed for two main reasons. The first methods involve using adapters [7, 8, 9], which achieve good performance in low-resource tasks by introducing additional parameter-trainable modules into the existing model and freezing the original model parameters for modeling [10]. However, adapters typically consist of two linear layers and a non-linear activation function [11], and their modeling capabilities are limited when the amount of data for new language translation pairs is large. Another approach utilizes regularization [12, 13] and resampling [14, 15] techniques to assess the performance between new and existing translation directions, aiming to enhance the model's ability to model new language pairs. However, these methods require updating all model parameters and may adversely affect the translation performance of existing languages due to the issue of catastrophic forgetting problems [16, 6].

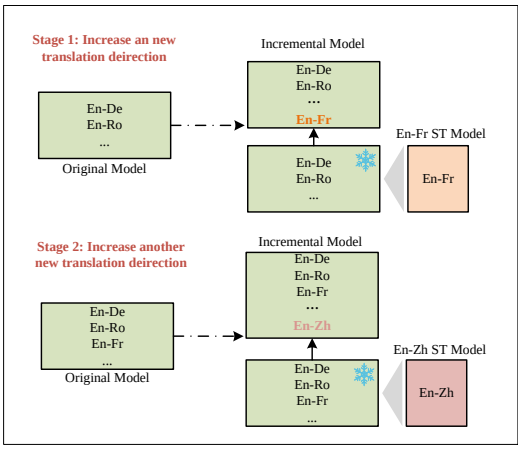

Figure 1: *Illustration of transfer learning for speech translation.*

How can we enhance the translation capability of a multilingual speech translation model for new language pairs while maintaining its existing translation performance? Previous studies [17, 18, 19] have identified the Feed Forward Network (FFN) module in the Transformer architecture as a promising solution for storing knowledge as key-value memories. In Transformer-based language models, the FFN module effectively converts discrete data into continuous feature representations and stores it as knowledge. In multilingual speech translation, we can leverage additional FFN modules to store the translation knowledge of new language pairs, ensuring that the original model remains unaffected.

Inspired by this, we propose a plug-and-play knowledge enhancement method for a multilingual speech translation model to accommodate new language pairs, as shown in Figure 1. Our method contains two novel mechanisms: 1) we train an additional speech translation model for the added languages using the Transformer architecture and save its FFN and embedding modules as pluggable modules. 2) we introduce pluggable modules into the multilingual speech translation model and freeze the original parameters for training. In this way, our proposed method stores the knowledge of the new language in pluggable modules and maintains the translation quality of the original language. The experimental results indicate that our proposed method does not influence the performance of existing speech translation languages. In contrast, compared to the adapters method, our method achieves an average performance improvement of approximately 1.5%, with the model's performance being more stable.

The main contributions of this paper are summarized as

10.21437/Interspeech.2024-2346

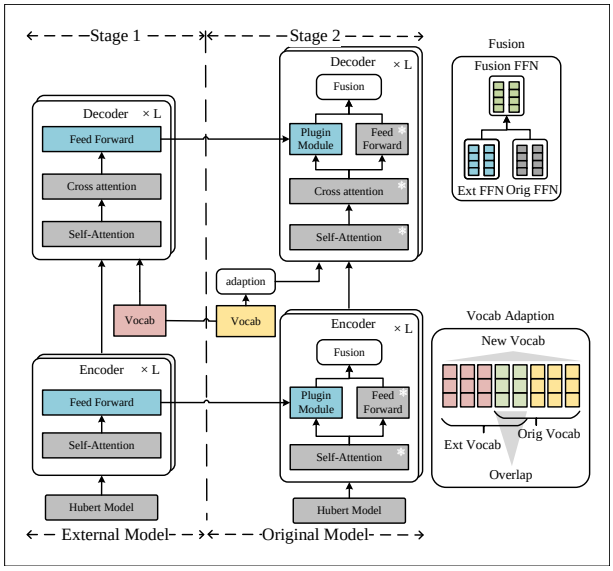

Figure 2: *Overview of our method. Ext represents the external model. Orig represents the original model.*

follows:

- We propose a multilingual speech translation architecture with pluggable modules to acquire knowledge of new languages, enabling competitive translation quality when adding new language pairs.
- Our architecture can maintain the performance of the original language pairs while effectively adapting to different language pairs in transfer learning.

2. Approach

As shown in Figure 2, our method consists of two steps. First, we separately train a translation model for the newly added language, aiming to convert the parallel corpus of the new language into continuous feature representations and store them as additional knowledge in the model. Second, the FFN and embedding modules of the added language pair translation model are migrated as plug-in additional components to the original multilingual speech translation module, followed by a second round of training.

2.1. Pluggable Modules

To adapt to new language pairs, in addition to adapting the FFN module, the multilingual speech-to-text translation task also needs to consider adapting the vocabulary. Therefore, we have added vocabulary adaptation functionality to the model.

Vocabulary Adaptation. For the speech-to-text translation task, the encoder takes speech signals as input and the decoder outputs text. Since the vocabulary of the new language cannot fully cover the vocabulary of the existing multilingual speech translation model, there will be some out-of-vocabulary (OOV) words. Therefore, we merge the vocabulary of the new language with the existing vocabulary to expand the size of the vocabulary and compensate for the differences between vocabularies. Assuming the vocabulary $\mathbf{V_1}$ of the existing multilingual speech translation model contains words $\mathbf{V_1} = \{\mathbf{w_1}, \mathbf{w_2}, \ldots, \mathbf{w_n}\}$, and the vocabulary $\mathbf{V_2}$ of the newly added language contains words $\mathbf{V_2} = \{\mathbf{w'_1}, \mathbf{w'_2}, \ldots, \mathbf{w'_m}\}$. The process of merging $\mathbf{V_2}$ into $\mathbf{V_1}$ can be represented as:

$$\mathbf{V_1} \cup \mathbf{V_2} = \left\{\mathbf{w_1}, \mathbf{w_2}, \ldots, \mathbf{w_n}, \mathbf{w'_1}, \mathbf{w'_2}, \ldots, \mathbf{w'_m}\right\}. \quad (1)$$

where $\cup$ denotes the union operation of sets, which merges all words from $\mathbf{V_1}$ and $\mathbf{V_2}$ into a new vocabulary.

FFN Adaptation. Each FFN layer consists of a non-linear activation function followed by two linear layers. The calculation process is as follows:

$$\mathrm{FFN}(\boldsymbol{x}) = \mathrm{ReLU}(\boldsymbol{x}\mathbf{W_1} + \mathbf{b_1})\mathbf{W_2} + \mathbf{b_2}. \quad (2)$$

where ReLU is the non-linear activation function, $\boldsymbol{x}$ is the inputs, $\mathbf{W_1}$ and $\mathbf{W_2}$ are the weight matrices, $\mathbf{b_1}$ and $\mathbf{b_2}$ are the bias vectors.

According to the research of the FFN module in transformer [17, 18, 19], the FFN module can be seen as key-value memories used for storing knowledge. Therefore, when adding new language pairs, we utilize the FFN layers of an external model to store additional knowledge and transfer this knowledge to the original multilingual speech translation model to achieve shared language knowledge. Additionally, the FFN layers of the external model can serve as a plug-in module, making it convenient to add new languages. We calculate the merged FFN output as follows:

$$\mathbf{H} = \mathrm{FFN}_{\text{original}}(\boldsymbol{x}) + \mathrm{FFN}_{\text{external}}(\boldsymbol{x}). \quad (3)$$

where $\mathrm{FFN}_{original}$ represents the FFN module of the multilingual machine translation model, and $\mathrm{FFN}_{external}$ represents the FFN module of the newly added language translation pair model.

2.2. Two Stage Training

In form, the Multilingual Neural Machine Translation (MNMT) model first selects a set of available parallel data $D = \{D_1, ..., D_i, ..., D_N\}$, covering N languages, where D_i represents the original parallel training corpus corresponding to the i-th language. As shown in Figure 2, we divide the training phase into two parts.

External Model Training. We first use the training data for the added language to train an additional speech translation model. When we use the new data $D' = \{D_{N+1}, ..., D_j, ..., D_M\}$, where D_j represents the incremental parallel training corpus corresponding to the j-th language. The training objective of the incremental model is to maximize the log-likelihood L:

$$\mathcal{L}_{\mathcal{D}'}(\hat{\theta}) = \sum_{\mathcal{D}_j \in \mathcal{D}'} \sum_{(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{D}_j} \log p(\boldsymbol{y} \mid \boldsymbol{x}; \hat{\theta}). \quad (4)$$

where $\hat{\theta}$ represents the trainable parameters of the additional speech translation model. After training the additional speech translation model, we only retain the FFN layer and embedding layer as pluggable modules to prepare for the next training phase.

Pluggable Module Training. In the second stage, we train the original multilingual speech translation model. We freeze the main parameters of the model and only train the pluggable modules, including the FFN module of the additional model and the word vector module. The training objective of the second stage is as follows:

$$\mathcal{L}_{\mathcal{D}'}\left(\hat{\theta}_e, \hat{\theta}_f\right) = \sum_{\mathcal{D}_j \in \mathcal{D}'} \sum_{(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{D}_j} \log p\left(\boldsymbol{y} \mid \boldsymbol{x}; \hat{\theta}_e, \hat{\theta}_f\right). \quad (5)$$

Table 1: *The BLEU scores of four additional languages (En-De, En-fr, En-Es, En-Ro). Parallel indicates that the pluggable module can be placed parallel to the corresponding module in the original model. Serial indicates that the pluggable module can be added after the corresponding module in the original model.*

Method	**Modules**	BLEU			
		En-De	En-Fr	En-Es	En-Ro
Baselines		25.3	35.7	30.5	23.8
Adapters [8]	Serial	25.24	35.21	28.65	22.41
	Parallel	23.35	33,53	26.22	20.59
Mono Adapters [10]	Serial	25.03	35.15	28.77	22.26
	Parallel	23.67	33.45	27.76	21.83
TPA adapter [20]	Serial	24.11	34.77	28.69	22.33
	Parallel	23.59	34.25	28.01	21.94
Ours	Serial	**26.67**	36.91	**30.73**	**24.25**
	Parallel	26.62	**36.97**	30.65	24.11

where $\hat{\theta}_e$ and $\hat{\theta}_f$ respectively represent the additional trainable parameters in the pluggable modules, which include the FFN layer and embedding layer.

3. Experiments

3.1. Datasets

Datasets: We conduct experiments on the MuST-C [21] dataset which comprises 8 translation directions, including English (En) to German (De), French (Fr), Spanish (Es), Romanian (Ro), Russian (Ru), Italian (It), Portuguese (Pt), and Dutch (Nl). These languages differ in terms of type and geographic location. The MuST-C dataset consists of at least 350 hours of TED talks, with corresponding transcripts and translation data available for each translation direction. To simulate the scenario of adding a new language to an existing multilingual speech translation model, we first constructed a multilingual translation model using four language pairs from the Must-C dataset, including Russian (Ru), Italian (It), Portuguese (Pt), and Dutch (Nl).

Model settings: We use Hubert [22] as the speech pre-trained model and use multi-task learning architecture [2, 3]. Next, we evaluated the performance of our approach on four other language pairs of speech translation, including German (De), French (Fr), Spanish (Es), Romanian (Ro). It is important to note that the original model does not support the addition of new language pairs. We use the Sentencepiece to build the vocabulary.

3.2. Baselines and Evaluation

The baselines can be listed as follows:

- **Language-pair adapters** [7] use a different adapter module per language pair in each encoder and decoder layer.
- **Monolingual language adapters** [10] use one type adapter per language. other settings are the same as Language-pair adapters.
- **Two Parallel Adapter (TPA)** [20]: is added to the residual connection of the feedforward module. It employs a decoder with randomly initialized parameters.

For serial and parallel, we introduce adapters in the serial or parallel connection manner, and our pluggable modules in the FFN layers can also be converted into a serial manner. For evaluation, we use the `tst-COMMON` dataset for testing and use beam search with size 10 for the decoding stage and evaluate with BLEU [23] scores.

Table 2: *BLEU score of the original speech translation direction when adding a new language. Regu-Based [12] denotes a training method based on regularization. Replay-Based [14] denotes a training method based on different sampling strategies.*

Method	En-Ru	En-It	En-Pt	En-Nl
Original	17.15	26.56	31.24	29.65
Replay-Based	15.73(-1.42)	25.31(-1.25)	30.04(-1.20)	28.33(-1.32)
Regu-Based	15.22(-1.93)	24.78(-1.78)	29.66(-1.58)	27.83(-1.82)
Ours	17.15(-0)	26.56(-0)	31.24(-0)	29.65(-0)

4. Experiments and Analysis

4.1. Main Results

Table 1 compares BLEU scores across four language pairs for different speech translation methods. Our method demonstrates a competitive advantage over others (excluding FP16 Baselines). Regarding serial module placement, our method achieves BLEU scores of 26.67, 36.91, 30.73, and 24.25 for the language pairs En-De, En-Fr, En-Es, and En-Ro, respectively. When applied in parallel, our method achieves 26.62, 36.97, 30.65, and 24.11, respectively. These results indicate robust performance, with minor differences between serial and parallel module placements, suggesting stability across configurations. In comparison to other models, our approach exhibits performance improvements. Specifically, when compared to specific alternative adapter methods, our approach has shown an average performance enhancement of approximately 1.5% to 2%, thus affirming the capability of the Feed-Forward Neural Network (FFN) module to retain additional knowledge efficiently. Notably, our method under serial and parallel configurations surpasses other adapter-based methods, illustrating our efficacy. The consistency in BLEU scores across different language pairs further underscores the generalizability and reliability of our method.

4.2. The Performance Stability of Pluggable Modules

We compare our proposed pluggable module method with existing methods based on resampling and regularization. After adding a new language, we evaluate the model's speech translation performance on existing language pairs. Experimental results are shown in Table 2.

Table 2 presents BLEU scores of speech translation methods when a new language is added. Notably, our method preserves original performance levels across language pairs such as English-to-Russian (En-Ru), English-to-Italian (En-It), English-

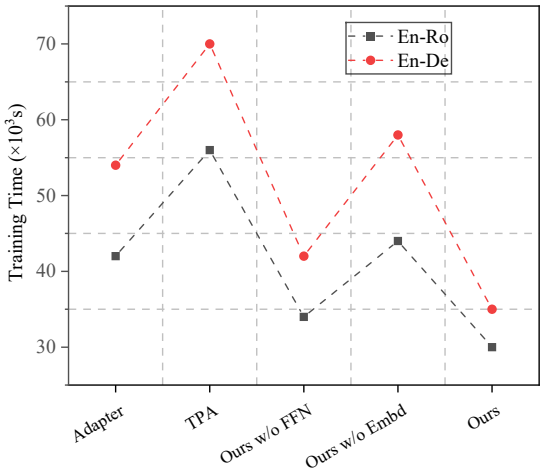

Figure 3: *The training time comparison of our proposed and adapter baselines.*

Table 3: *BLEU score of ablation study of pluggable module. Embedding refers to the embedding layer of the speech translation model's decoder. FFN represents the feed-forward layer of the speech translation model's decoder.*

No.	Transfer Scopes		Incremental Language			
	Embedding	FFN	En-De	En-fr	En-Es	En-Ro
1	✓	✘	25.24	35.57	29.11	22.94
2	✘	✓	26.35	36.51	30.24	23.92
3	✓	✓	26.67	36.91	30.73	24.25

to-Portuguese (En-Pt), and English-to-Dutch (En-Nl). This is in stark contrast to other approaches: Replay-based methods see some performance decrease, and Regularization-Based methods present a more distinct drop. Our proposed method demonstrates its effectiveness in adapting to new languages without compromising translation quality.

4.3. The Ablation Study of Pluggable Module

To further investigate the impact of pluggable modules on model performance, we conduct ablation experiments to validate the performance of models with and without added vocabulary. It is worth noting that in this experiment, we randomly initialize the language's vocabulary but do not update the parameters.

Table 3 presents BLEU scores from an ablation study of a speech translation model's pluggable module, exploring the importance of the embedding layer and the feed-forward network (FFN). The study shows that incorporating the FFN, whether alone or alongside the embedding layer, significantly improves performance across four language pairs: English-to-German (En-De), English-to-French (En-Fr), English-to-Spanish (En-Es), and English-to-Romanian (En-Ro). Using both the embedding and the FFN yields the best results, with BLEU scores peaking at 26.67, 36.91, 30.73, and 24.25, respectively. These findings underscore the FFN's pivotal role in enhancing the model's translation capabilities, confirming that it is crucial for retaining and effectively applying linguistic knowledge.

Table 4: *The ablation study of different modules in transformer and fusion strategy.*

Method	En-De	En-Fr
Ours	26.67	36.91
+=Self-Attention	26.17	36.17
+=Gate-Fusion	25.84	35.95
+=Dropout	25.93	36.03

4.4. The Importance of FFN Module

Our proposed method integrates the Feed-Forward Neural Network (FFN) module as additional knowledge into the original model. Furthermore, we investigate the impact of incorporating different modules and various fusion strategies on model performance, with relevant results detailed in Table 4.

Table 4 shows the BLEU scores for our speech translation model with various pluggable modules across two language pairs, English-to-German (En-De) and English-to-French (En-Fr). Our base model achieves 26.67 for En-De and 36.91 for En-Fr. Modifications with Self-Attention show a limited improvement, indicating a marginal contribution to knowledge retention. The Gate-Fusion module, a gating mechanism, leads to a drop in performance, suggesting the FFN module houses the bulk of translatable knowledge. The Dropout function also results in lower scores, further indicating the FFN module's role in utilizing linguistic information.

4.5. The Comparison of Training Cost

To further illustrate the efficiency of our method, we investigate the training time compared with other baselines. Figure 3 compares the training times for various speech translation models. It demonstrates the efficiency of our proposed method against baseline models like the Adapter and TPA. Specifically, the figure shows the training durations for translating between English and Romanian (En-Ro) and English and German (En-De). Our method consistently shows lower training times, reducing the computational resources required. This advantage is crucial for scalable systems where quick retraining is necessary, especially when incorporating new languages or updating existing ones. The data also suggests that removing the FFN or embedding layer increases training time, underscoring their importance in model optimization.

5. Conclusion

In this paper, we have proposed a novel approach for enhancing multilingual speech-to-text translation models by introducing pluggable modules for new language pairs. Our method effectively stores knowledge for new languages without affecting the performance of existing languages. Experimental results demonstrate that our approach outperforms traditional adapter methods, achieving an average performance improvement of approximately 1.5% while maintaining stability. This work contributes to the field of multilingual speech translation by providing a more efficient and stable method for integrating new languages into existing models.

This work is supported by the Natural Science Foundation of China (62366037, 62066033); Outstanding Youth Project of Inner Mongolia Natural Science Foundation (2022JQ05); Young science and technology talents cultivation project of Inner Mongolia University (21221505).

6. References

[1] Q. Fang and Y. Feng, "Understanding and bridging the modality gap for speech translation," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, A. Rogers, J. L. Boyd-Graber, and N. Okazaki, Eds. Association for Computational Linguistics, 2023, pp. 15 864–15 881. [Online]. Available: https://doi.org/10.18653/v1/2023.acl-long.884

[2] Y. Tang, J. M. Pino, X. Li, C. Wang, and D. Genzel, "Improving speech translation by understanding and learning from the auxiliary text translation task," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Association for Computational Linguistics, 2021, pp. 4252–4261. [Online]. Available: https://doi.org/10.18653/v1/2021.acl-long.328

[3] 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, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*. IEEE, 2021, pp. 6209–6213. [Online]. Available: https://doi.org/10.1109/ICASSP39728.2021.9415058

[4] L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, M. Duppenthaler, P.-A. Duquenne, B. Ellis, H. Elsahar, J. Haaheim *et al.*, "Seamless: Multilingual expressive and streaming speech translation," *arXiv preprint arXiv:2312.05187*, 2023.

[5] Q. Fang, R. Ye, L. Li, Y. Feng, and M. Wang, "STEMM: Self-learning with speech-text manifold mixup for speech translation," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 7050–7062. [Online]. Available: https://aclanthology.org/2022.acl-long.486

[6] R. M. French, "Catastrophic interference in connectionist networks: Can it be predicted, can it be prevented?" in *Advances in Neural Information Processing Systems 6, [7th NIPS Conference, Denver, Colorado, USA, 1993]*, J. D. Cowan, G. Tesauro, and J. Alspector, Eds. Morgan Kaufmann, 1993, pp. 1176–1177.

[7] A. Bapna and O. Firat, "Simple, scalable adaptation for neural machine translation," in *Proc. of EMNLP*, 2019.

[8] J. Philip, A. Berard, M. Gallé, and L. Besacier, "Monolingual adapters for zero-shot neural machine translation," in *Proc. of EMNLP*, 2020.

[9] C. Baziotis, M. Artetxe, J. Cross, and S. Bhosale, "Multilingual machine translation with hyper-adapters," in *Proc. of EMNLP*, 2022.

[10] J. Philip, A. Berard, M. Gallé, and L. Besacier, "Monolingual adapters for zero-shot neural machine translation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Association for Computational Linguistics, 2020, pp. 4465–4470. [Online]. Available: https://doi.org/10.18653/v1/2020.emnlp-main.361

[11] A. F. Agarap, "Deep learning using rectified linear units (relu)," *CoRR*, vol. abs/1803.08375, 2018. [Online]. Available: http://arxiv.org/abs/1803.08375

[12] Y. Huang, X. Feng, X. Geng, and B. Qin, "Omniknight: Multilingual neural machine translation with language-specific selfdistillation," *arXiv preprint arXiv:2205.01620*, 2022.

[13] Y. Zhao, J. Zhu, L. Xiang, J. Zhang, Y. Zhou, F. Zhai, and C. Zong, "Life-long learning for multilingual neural machine translation with knowledge distillation," *CoRR*, vol. abs/2212.02800, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2212.02800

[14] X. Garcia, N. Constant, A. P. Parikh, and O. Firat, "Towards continual learning for multilingual machine translation via vocabulary substitution," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds. Association for Computational Linguistics, 2021, pp. 1184–1192. [Online]. Available: https://doi.org/10.18653/v1/2021.naacl-main.93

[15] Z. Liu, G. I. Winata, and P. Fung, "Continual mixed-language pre-training for extremely low-resource neural machine translation," in *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, ser. Findings of ACL, C. Zong, F. Xia, W. Li, and R. Navigli, Eds., vol. ACL/IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 2706–2718. [Online]. Available: https://doi.org/10.18653/v1/2021.findings-acl.239

[16] I. J. Goodfellow, M. Mirza, X. Da, A. C. Courville, and Y. Bengio, "An empirical investigation of catastrophic forgeting in gradient-based neural networks," in *Proc. of ICLR*, 2014.

[17] M. Geva, R. Schuster, J. Berant, and O. Levy, "Transformer feed-forward layers are key-value memories," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, M. Moens, X. Huang, L. Specia, and S. W. Yih, Eds. Association for Computational Linguistics, 2021, pp. 5484–5495. [Online]. Available: https://doi.org/10.18653/v1/2021.emnlp-main.446

[18] D. Dai, W. Jiang, Q. Dong, Y. Lyu, and Z. Sui, "Neural knowledge bank for pretrained transformers," in *Natural Language Processing and Chinese Computing - 12th National CCF Conference, NLPCC 2023, Foshan, China, October 12-15, 2023, Proceedings, Part II*, ser. Lecture Notes in Computer Science, F. Liu, N. Duan, Q. Xu, and Y. Hong, Eds., vol. 14303. Springer, 2023, pp. 772–783. [Online]. Available: https://doi.org/10.1007/978-3-031-44696-2_60

[19] R. Vázquez, H. Çelikkanat, V. Ravishankar, M. Creutz, and J. Tiedemann, "A closer look at parameter contributions when training neural language and translation models," in *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, N. Calzolari, C. Huang, H. Kim, J. Pustejovsky, L. Wanner, K. Choi, P. Ryu, H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, and S. Na, Eds. International Committee on Computational Linguistics, 2022, pp. 4788–4800. [Online]. Available: https://aclanthology.org/2022.coling-1.424

[20] N. Chen, I. Shafran, Y. Zhang, C. Chiu, H. Soltau, J. Qin, and Y. Wu, "Efficient adapters for giant speech models," *CoRR*, vol. abs/2306.08131, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2306.08131

[21] 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 2012–2017. [Online]. Available: https://doi.org/10.18653/v1/n19-1202

[22] W. Hsu, B. Bolte, Y. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 3451–3460, 2021. [Online]. Available: https://doi.org/10.1109/TASLP.2021.3122291

[23] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proc. of ACL*, 2002.