NMT for low-resource languages: fighting catastrophic forgetting in hierarchical transfer learning using data interleaving

Lasse Becker-Czarnetzki Andrea Lombardo Matteo Tafuro
{lasse.becker-czarnetzki, andrea.lombardo, matteo.tafuro}@student.uva.nl

1 Introduction

Powered by deep learning, Neural Machine Translation (NMT) has become the dominant approach in the community: compared to traditional Statistical MT (SMT), neural networks can better capture the contextual information and therefore result in high-quality and fluent translations [1]. Yet, because of the complexity of the network and a large number of parameters, NMT models are highly dependent on the quality and the availability of parallel corpora [2]. Although this is not an issue for high-resource languages (HRLs), it is not a realistic assumption for the majority of the 7151 living languages [3, 4]. Therefore, recent years have seen a remarkable increase in NMT research that focuses specifically on low-resource languages (LRLs).

In this paper, we aim to combine several low-resource settings NMT techniques. As noted in the comprehensive survey of LRL-NMT by Ranathunga et al. [5], transfer learning using a pre-trained multilingual parent has been identified as a promising approach [6, 7]. Therefore, we aim to investigate whether combining a hierarchical fine-tuning approach with data interleaving can benefit a low-resource language translation, by fighting catastrophic forgetting in the chosen intermediate language. As depicted in Figure 1, we intend to improve the Azerbaijani *(AZ)*↔English *(EN)* translation performance of M2M100 [8] by adding an intermediate fine-tuning step on a related language Turkish *(TR)*↔EN, before finally fine-tuning on AZ↔EN with data interleaving. The hypothesis is that if we forget less about the related language, the low-resource translation can benefit more from the information learned during previous stages of training. In short, our **research question** is formalized as follows:

Does fighting catastrophic forgetting in a related language through simple data interleaving help the downstream translation performance in a hierarchical transfer learning approach?

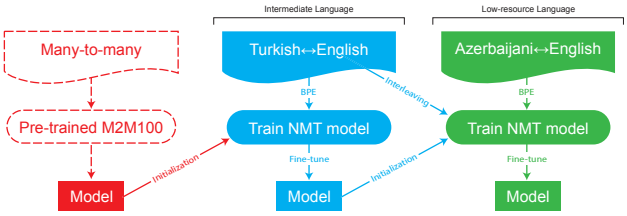

Figure 1: A diagram of our hierarchical transfer learning with data interleaving.

2 Related work

Many strategies have been proposed to overcome the data scarcity problem that is intrinsic to LRLs. A simple and intuitive approach is to exploit monolingual data to enrich parallel data, for instance using *data augmentation* [9], *back translation* [10], or *dual learning* [11]. Alternatively, one can also exploit data from auxiliary languages: languages in the same family may share linguistic properties and consequently, corpora of related languages can be utilized to assist the translation between LRL pairs [1]. One example is *multilingual training* [12], where the LRL pairs benefit from related rich-resource language pairs through joint training. Another effective method is *transfer learning* [13], wherein a model is firstly trained on HRLs and then fine-tuned on an LRL pair. This fine-tuning process can also be conducted hierarchically in a multi-stage fashion [4], to exploit the data volume advantages of HRLs and the syntactic similarity advantages of cognate languages. While this approach has proven to be beneficial for LRLs, it is prone to the *catastrophic forgetting* issue whereby a neural network forgets previously learned information after learning new ones [14].

Firstly defined by McCloskey and Cohen [14] in 1989, catastrophic forgetting is a phenomenon that haunts every learning system. Any model without lossless memory will, at some point, have to forget past information in order to learn new ones. The research has seen many efforts to fight catastrophic forgetting in NMT, for instance in the context of domain adaptation [15] and hierarchical transfer learning [4, 16]. Examples include adjusting the network weights during

training to control the forgetting of the model [17, 18] or expanding the model's capacity to support new tasks [19, 20]. However, Carrión-Ponz and Casacuberta [16] noted that these approaches tend to be quite hard to implement and are often very computationally intensive. For this reason, they proposed a simple method to mitigate the effects of catastrophic forgetting without a significant increase in computational costs. They interleave past data in the learning of new language pairs and introduce a loss function that actively re-weights past samples to minimize forgetting. We acknowledge the potential of this approach, and thus attempt to expand it with further experiments. More specifically, we investigate whether the method retains its effectiveness if the first training stage involves a multilingual NMT model and if it can benefit the downstream fine-tuned low-resource translation when using a related language in the intermediate step.

3 Methodology

This section first introduces the fundamental concepts. Then we formalize our approach and define the metrics we use to compare it to the baseline model. Finally, we describe the datasets used.

3.1 Fundamental concepts

Our work revolves around three main concepts: multilingual NMTs, transfer learning, and data interleaving.

Multilingual NMT. Multi-NMT systems are those handling translation between more than one language [21]. For our study, the most relevant feature of multi-NMT models is their ability to learn a shared semantic representation between languages [12].

Transfer Learning. Typical transfer learning methods train an NMT model on HRL pairs (*parent* model) and then fine-tune all or some of the model parameters on an LRL pair (*child* model). We embrace the recent advancements that have demonstrated the benefits of including intermediate stages (i.e. hierarchical approach), whereby the model is fine-tuned on a cognate language pair with wider data availability [4].

Data interleaving. Data interleaving fights the catastrophic forgetting phenomenon introduced by hierarchical transfer learning. As already specified in Sec. 2, we follow the lines of Carrión-Ponz and Casacuberta [16] who interleave past data in the learning of new language pairs[1]. The intuition behind this approach is that passing data from the past task during a later stage of training prevents large variations of the weights responsible for the previous tasks. In simple terms, it serves as a *memory refresher* for the model.

3.2 Models

Baseline. Our first baseline model is a pre-trained M2M100 [8] that is solely run in inference mode on the AZ↔EN translation task.

Baseline + fine-tuning. Before experimenting with our approach, we fine-tune the aforementioned M2M100 baseline on the AZ↔EN task *without* the multi-stage strategy nor the interleaving of the data.

Baseline + hierarchical fine-tuning. Our third and final baseline fine-tunes the pre-trained M2M100 baseline on the AZ↔EN task *with* the multi-stage strategy but *without* data interleaving.

Our approach. Finally, our approach combines the hierarchical transfer learning introduced by Luo et al. [4] with the data interleaving strategy proposed by Carrión-Ponz and Casacuberta [16]. Since the final task is AZ↔EN, the intermediate language must be chosen carefully such that the space mismatch is minimized [5]. We select a medium-resource language from the Turkic language family, i.e. Turkish, and use it for the intermediate fine-tuning step. To verify that the space mismatch of our chosen languages is indeed minimum, we investigate their vocabulary overlap. The reader can consult Appendix A.

Our approach is schematically depicted in Figure 1. It comprises the following steps: (i) a pre-trained M2M100 is utilized for the first stage of the hierarchical process; (ii) the weights of this model are used to initialize the training on Turkish↔English; (iii) the current weights are used to initialize the Azerbaijani↔English model and the training is conducted using data interleaving. Along the lines of the original authors, we vary the ratio of past data used per batch during the learning of the new language pair.

3.3 Evaluation metrics

BiLingual Evaluation Understudy (BLEU) [22] is the most popular metric used in the NLP field to quantify model performance. However, given its sensitivity to the chosen parameters and implementation, we used SacreBLEU [23] for our experiments.

3.4 Datasets

We obtain our parallel corpora (EN↔TR, EN↔AZ) from the TED dataset [24], a growing collection of

[1] Given the limited time and resources allocated to this project, we only experiment with data interleaving and disregard the proposed loss function.

Dataset	*Train*	*Dev*	*Test*
TR - EN	182,450	4,045	5,029
AZ - EN	5,946	671	903

Table 1: Number of sentences per language pair for each split.

parallel corpora extracted from TED talks. Table 1 shows the data sizes. We first considered using the WMT17 dataset [25] for the Turkish corpus during hierarchical training but noticed no improvements. For this reason, we stick to the same dataset to mitigate the need for domain adaptation the model has to learn.

4 Experiments

4.1 Experimental setup

To investigate our research question, we utilize HuggingFace's implementation of the M2M100 model. Due to resources and time scarcity, we resort to their 418M checkpoint and use this pre-trained model as our baseline. Then, we fine-tune it further according to our hierarchical approach (see Sec 3)[2]. Training such large-size models can be memory challenging. To this extent, we make use of block-wise dynamic quantization, since it enables stable 8-bit optimizers (in our case, AdamW[3] and maintains 32-bit performance at a fraction of the memory footprint [26]. Following the same reasoning, we also use sharded distributed training from FairScale [27], which is supported in HuggingFace, on two Nvidia 1080Ti.

For all datasets, the preprocessing of the translation data is performed using the M2M100Tokenizer of Huggingface. We fine-tune the M2M100 model for 10 epochs on the intermediate language (TR↔EN) and for 20 epochs on AZ↔EN. Note that at each stage we fine-tune *all* of the model parameters. We utilize a learning rate (LR) of $2e^{-5}$ and the linear HuggingFace scheduler [4] with warmup and weight decay.

We use early stopping on the best BLEU score on the corresponding validation set with a patience of 6. We twice per epoch. To fight overfitting we use a weight decay of 0.01. These values were set after a brief hyperparameter search. We also set dropout and attention dropout to 0.1. Higher values didn't result in improvments. Due to memory usage concerns, we use gradient accumulation and a batch size of 4 *per device*, effectively resulting in a batch size of 32. We also use mixed precision fp16 training supported by Hugginface to speed up training.

We report the BLEU score during inference on the validation and test sets using greedy decoding and beam search with a beam size of 5.

Finally, during downstream fine-tuning, we still record the performance on the intermediate language validation set to track the severity of catastrophic forgetting.

4.2 Data Interleaving Settings

Along the line of the original authors, we vary the ratio of past data used per batch (1%, 5%, 10%, 20%) during the learning of the new language pair. More specifically, in every epoch, we sample random sentences from the TR↔EN training set and add them to the full AZ↔EN training set. The amount of sampled sentences depends on the chosen interleaving percentage. By resampling past data at each epoch we prevent the model from memorizing specific sentences, and rather enforce it to generally perform well on Turkish. We provide our training scripts and modified code of the Transformers library to enable some of the training procedures on our github[5].

4.3 Results

(Due to time unavailability and other contingencies encountered along the way, we were only able to experiment with the AZ→EN translation direction. The code, however, is also available for the other direction (EN→AZ) and we encourage its use in possible future works.)

The results of the experiments are summarized and collected in Table 2. The BLEU scores highlight a noticeable improvement of the fine-tuned models over the baseline, which is especially prominent for the hierarchical model.

Fighting catastrophic forgetting of the intermediate language by interleaving data of the past task does not seem to help the downstream translation performance. However, further investigations reveal that catastrophic forgetting is *not* actually catastrophic and that the TR→EN translation performance only drops by $\sim$ 2 BLEU points after fine-tuning on AZ→EN (refer to Appendix B, Table 4). Accordingly, changing the percentage of interleaved data leaves the performance unchanged (see Appendix B, Table 5).

The improvements of the fine-tuned models over the baseline are also qualitatively reflected in the trans-

[2]The **baseline + fine-tuning** model is trained using the same AZ↔EN experimental setup. The only difference is that the model weights are initialized using the pre-trained M2M100, instead of using the intermediate fine-tuned model.

[3]We utilize the bitsandbytes AdamW optimizer as implemented here.

[4]Refer to the Hugging Face website.

[5]https://github.com/B-Czarnetzki/transformers

Model	Azerbaijani → English				English → Azerbaijani			
	Validation		**Test**		**Validation**		**Test**	
	Greedy	*Beam*	*Greedy*	*Beam*	*Greedy*	*Beam*	*Greedy*	*Beam*
Baseline	4.49	7.32	2.93	6.50	1.91	2.87	1.60	2.36
Baseline + fine-tuning	22.18	22.88	17.55	18.62	12.27	13.04	9.46	8.72
Baseline + hierarchical fine-tuning	23.27	24.40	18.81	19.72	x	x	x	x
Best interleaved model (1%)	23.21	24.35	18.81	19.75	x	x	x	x
Literature [28]	12.78				5.06			

Table 2: BLEU scores of the different models for AZ↔EN. To the best of our knowledge, Aharoni et al. [28] achieved the highest performance in the AZ↔EN task using the TED corpus. The best interleaved model for the AZ↔EN task uses a ratio of 1%, as seen in Table 5. Moreover, as already mentioned in Section 4.3, we were unable to complete the experiments for the EN→AZ translation direction (hence the missing *x* values).

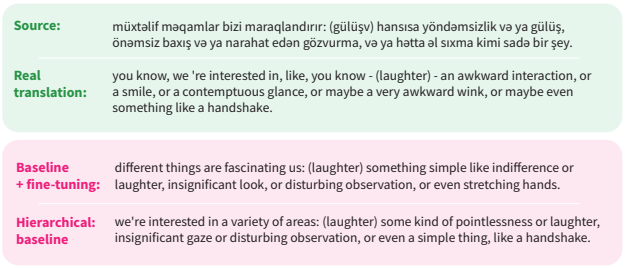

Figure 2: Qualitative analysis of the translation of a sentence containing OOV tokens. The hierarchical model exhibits better translation quality.

lations of the sentences. To prove this, we selected the out-of-vocabulary (OOV) tokens of the Azerbaijani test set and their frequencies in the training set to choose 10 sentences with the highest amount of OOV tokens. We then run inference with our models. Hierarchical fine-tuning exhibits clear translation improvements, as illustrated in Figure 3.

5 Discussion

As the results show, fine-tuning massive multilingual models is still very beneficial for a downstream translation of a low-resource language. We also observe the significant improvement that the hierarchical approach provides. This is likely due to the improved translation of OOV tokens that we transfer-learned from the intermediate language. However, in this context, the fighting against catastrophic forgetting was proven to be unnecessary, since fine-tuning on the downstream task led to minimal performance degradation on the intermediate task. Therefore, it is reasonable that interleaving data does not affect the final performance of the model.

We suspect that this is due to three main reasons. Firstly, catastrophic forgetting seems to be most prevalent when a model reaches a saturation point of training on more than a few tasks in a hierarchical setting [16], while we only have a two-stage fine-tuning. Secondly, due to the language tags of M2M100, the model has an internal memory functionality for past tasks that could be responsible for reducing catastrophic forgetting. Finally, the reasoning why the downstream translation can benefit from the hierarchical fine-tuning on a similar language can be reverted. It could be the case that the high vocabulary overlap and general language similarity prevents the model from radically changing the distribution of its weights.

Despite our unfortunate experimental setup, we still believe in the potential of the original research direction and we leave further investigations to future works. First, one could try to answer our research question in a deeper hierarchical setting, where the number of translation tasks is increased. Additionally, one could also investigate more fine-grained aspects of the model, e.g. whether fine-tuning on one language group increases catastrophic forgetting on a different language group. Alternatively, other research efforts could explore how stable the memorization of distant languages is through the sole use of language tags.

6 Conclusion

In this paper we investigated whether the downstream translation performance of a hierarchical transfer learning approach could be improved by fighting catastrophic forgetting in a related language through data interleaving. The experimental results demonstrated that, under our specific setup, the model retains almost its original performance on the intermediate language, hence it does not experience major forgetting. Thus, the use of interleaving also does not have any significant effect on mitigating forgetting or improving the downstream translation performance.

Appendix

A Vocabulary overlap between intermediate and target language

The hierarchical transfer learning approach requires the intermediate language to be affine to the target language. Only in that case, the approach retains the benefits of both large data volumes and linguistic similarities [4].

Given our target language (Azerbaijani), we prove the appropriateness of Turkish as an intermediate language by observing their vocabulary overlap. For the sake of analysis, we also include Portuguese (PT)[6] in the comparison in the hope to observe a major discrepancy between PT and AZ/TR.

We do the following: (i) take the vocabulary of the pre-trained M2M100 model; (ii) tokenize the training sets of all parallel corpora (AZ-EN, TR-EN, PT-EN); (iii) obtain the frequency of each vocabulary entry for AZ, TR and PT. Given the different dataset sizes, we (iv) normalize the frequency vectors so they sum to 1.

We compare the vocabularies both qualitatively and quantitatively. Regarding the qualitative analysis, we found a plot with three overlapping histograms to be extremely chaotic and not meaningful in terms of comparison. Therefore, we generate a Cumulative Distribution Function plot (Figure 3) that displays very clearly the vocabulary overlap.

To quantify the vocabulary overlap, we compute the *histogram overlap* of the normalized frequency vectors, which calculates the similarity of two discretized probability distributions. Given two histograms I and M with n bins, the histogram overlap is computed as:

$$\sum_{j=1}^{n} min(I_j, M_j) \quad (1)$$

The resulting values are collected in Table 3.

Both qualitative and quantitative analyses demonstrate the high overlap between Turkish and Azerbaijani, especially when these are compared to Portuguese.

B Results

Table 4 shows the BLEU scores of the different models for TR→EN. The hierarchical fine-tuned model achieves the best performance, as interleaving does not seem to make a difference. Table 5 shows the scores on the Azerbaijani→English translation task, using different interleaving ratios of Turkish→English data. The results seem to indicate that varying the ratio does not significantly improve the score.

[6]The parallel PT-EN corpus was also taken from the TED dataset, as described in Section 3.4.

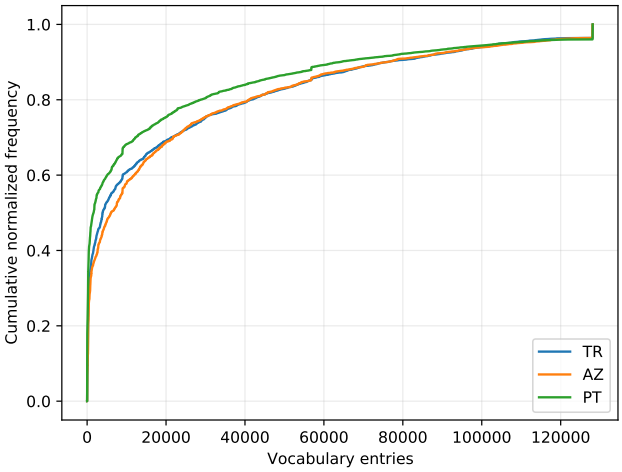

Figure 3: Cumulative Distribution Function plot of the normalized frequency vectors of Turkish (TR), Azerbaijani (AZ) and Portuguese (PT).

Language pair	Histogram Intersection
AZ-TR	0.486
AZ-PT	0.218
TR-PT	0.235

Table 3: Histogram intersection values calculated on the normalized frequency vectors for Turkish (TR), Azerbaijani (AZ) and Portuguese (PT).

	Turkish → English				English → Turkish			
Model	**Validation**		**Test**		**Validation**		**Test**	
	Greedy	*Beam*	*Greedy*	*Beam*	*Greedy*	*Beam*	*Greedy*	*Beam*
Baseline + hierarchical fine-tuning (at the intermediate stage)	33.21	34.03	31.13	31.70	20.85	21.69	18.76	19.42
Baseline + hierarchical fine-tuning (after the final stage)	31.84	32.78	29.77	30.88	x	x	x	x
Best interleaved model (5%)	31.76	32.72	29.71	31.01	x	x	x	x
Literature [28]	29.75				17.68			

Table 4: BLEU scores of the different models for TR↔EN. To the best of our knowledge, Aharoni et al. [28] achieved the highest performance in the AZ↔EN task using the TED corpus. The best interleaved model for the TR↔EN task uses a ratio of 5%, as seen in Table 5. Moreover, as already mentioned in Section 4.3, we were unable to complete the experiments for the EN→AZ translation direction (hence the missing x values).

	Turkish → English				Azerbaijani → English			
Interleaved ratio	**Validation**		**Test**		**Validation**		**Test**	
	Greedy	*Beam*	*Greedy*	*Beam*	*Greedy*	*Beam*	*Greedy*	*Beam*
0%	31.84	32.78	29.77	30.88	23.27	24.40	18.81	19.72
1%	31.90	32.63	29.73	30.97	23.21	24.35	18.81	19.75
5%	31.76	32.72	29.71	31.01	23.16	24.23	18.75	19.51
10%	31.73	32.65	29.81	30.96	23.10	24.38	18.85	19.68
20%	31.76	32.78	29.86	30.65	23.07	24.29	18.64	19.71

Table 5: BLEU scores of our approach on the Azerbaijani→English translation task, using different interleaving ratios of Turkish→English data.

References

[1] Rui Wang, Xu Tan, Renqian Luo, Tao Qin, and Tie-Yan Liu. A survey on low-resource neural machine translation, 2021.

[2] Shuoheng Yang, Yuxin Wang, and Xiaowen Chu. A survey of deep learning techniques for neural machine translation, 2020.

[3] D.M. Eberhard, G.F. Simons, and C.D. Fennig. *Ethnologue: Ethnologue: Languages of the World, Twenty-Fifth Edition*. Ethnologue Series. Sil International, Global Publishing, 2022. URL http://www.ethnologue.com.

[4] Gongxu Luo, Yating Yang, Yang Yuan, Zhanheng Chen, and Aizimaiti Ainiwaer. Hierarchical transfer learning architecture for low-resource neural machine translation. *IEEE Access*, 7:154157–154166, 2019. doi: 10.1109/ACCESS.2019.2936002.

[5] Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. Neural machine translation for low-resource languages: A survey, 2021. URL https://arxiv.org/abs/2106.15115.

[6] Raj Dabre, Atsushi Fujita, and Chenhui Chu. Exploiting multilingualism through multistage fine-tuning for low-resource neural machine translation. 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)*, pages 1410–1416, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1146. URL https://aclanthology.org/D19-1146.

[7] Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1032. URL https://aclanthology.org/N18-1032.

[8] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. Beyond english-centric multilingual machine translation, 2020. URL https://arxiv.org/abs/2010.11125.

[9] Marzieh Fadaee, Arianna Bisazza, and Christof Monz. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2090. URL https://aclanthology.org/P17-2090.

[10] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data, 2015. URL https://arxiv.org/abs/1511.06709.

[11] Yingce Xia, Di He, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. Dual learning for machine translation. 2016. doi: 10.48550/ARXIV.1611.00179. URL https://arxiv.org/abs/1611.00179.

[12] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation. Technical report, Google, 2016. URL https://arxiv.org/abs/1611.04558.

[13] Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1163. URL https://aclanthology.org/D16-1163.

[14] Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. volume 24 of *Psychology of Learning and Motivation*, pages 109–165. Academic Press, 1989. doi: https://doi.org/10.1016/S0079-7421(08)60536-8. URL https://www.sciencedirect.com/science/article/pii/S0079742108605368.

[15] Danielle Saunders. Domain adaptation and multi-domain adaptation for neural machine translation: A survey, 2021. URL https://arxiv.org/abs/2104.06951.

[16] Salvador Carrión-Ponz and Francisco Casacuberta. Few-shot regularization to tackle catastrophic forgetting in multilingual machine translation. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 188–199, Orlando, USA, September 2022. Association for Machine Translation in the Americas. URL https://aclanthology.org/2022.amta-research.14.

[17] Zhizhong Li and Derek Hoiem. Learning without forgetting. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 614–629, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46493-0.

[18] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, mar 2017. doi: 10.1073/pnas.1611835114. URL https://doi.org/10.1073%2Fpnas.1611835114.

[19] Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks, 2016. URL https://arxiv.org/abs/1606.04671.

[20] Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks, 2017. URL https://arxiv.org/abs/1708.01547.

[21] Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. A survey of multilingual neural machine translation. *ACM Comput. Surv.*, 53 (5), sep 2020. ISSN 0360-0300. doi: 10.1145/3406095. URL https://doi.org/10.1145/3406095.

[22] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

[23] Matt Post. A call for clarity in reporting bleu scores. 2018. doi: 10.48550/ARXIV.1804.08771. URL https://arxiv.org/abs/1804.08771.

[24] Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2084. URL https://aclanthology.org/N18-2084.

[25] Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 169–214, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W17-4717.

[26] Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 8-bit optimizers via block-wise quantization. *CoRR*, abs/2110.02861, 2021. URL https://arxiv.org/abs/2110.02861.

[27] FairScale authors. Fairscale: A general purpose modular pytorch library for high performance and large scale training, 2021. URL https://github.com/facebookresearch/fairscale.

[28] Roee Aharoni, Melvin Johnson, and Orhan Firat. Massively multilingual neural machine translation, 2019. URL https://arxiv.org/abs/1903.00089.