Domain Generalisation of NMT: Fusing Adapters with Leave-One-Domain-Out Training

Thuy-Trang Vu[◇*] **Shahram Khadivi**[†] **Dinh Phung**[◇] **Gholamreza Haffari**[◇]
[◇]Department of Data Science and AI, Monash University, Australia
[†] eBay Inc.
{trang.vuthithuy,first.last}@monash.edu
skhadivi@ebay.com

Abstract

Generalising to unseen domains is under-explored and remains a challenge in neural machine translation. Inspired by recent research in parameter-efficient transfer learning from pre-trained models, this paper proposes a fusion-based generalisation method that learns to combine domain-specific parameters. We propose a leave-one-domain-out training strategy to avoid information leaking to address the challenge of not knowing the test domain during training time. Empirical results on three language pairs show that our proposed fusion method outperforms other baselines up to +0.8 BLEU score on average.

1 Introduction

Building robust machine translation (MT) models that can perform well on a test set outside the domain of training examples is highly desired in real-world scenarios. Despite recent great progress in neural machine translation (NMT) research, NMT models have been found sensitive to distribution shift and adversarial examples (Koehn and Knowles, 2017; Belinkov and Bisk, 2018; Müller et al., 2020). While improving an NMT model to a new domain has been studied extensively in domain adaptation settings where in-domain parallel or monolingual data is given (Chu and Wang, 2018), generalising NMT models to unseen domains is under-explored (Specia et al., 2020).

Domain generalisation is a problem setting in machine learning that tackles the challenge of learning a robust model for unseen domains from multiple existing domains. This problem is closely related to several settings such as multi-task learning, transfer learning, and domain adaptation in terms of learning models from one or more given tasks/domains to enhance performance on some target tasks/domains. The main difference and challenge in domain generalisation is that the test domains are *unknown* in advance. Previous works on domain generalisation focused on learning invariant features by minimising the difference in the representations of the given domains for the classification tasks (Li et al., 2018; Wang et al., 2020b; Gulrajani and Lopez-Paz, 2020). Learning a domain-invariant representation is applicable to the classification problems where such invariances may be sufficient to predict the target classes. However, it may be inadequate for translation tasks. A good translation should not only preserve the invariant features such as syntax and grammar, but also be able to maintain the domain-specific features such as style of the source sentence.

In this paper, we propose a fusion-based approach to the domain generalisation problem for NMT. Our method comprises two training stages. The first stage is to learn domain-specific features through adapter modules added to the pre-trained encoder-decoder model. Previous works have shown that the task-specific adapter is an effective alternative method to fine-tuning. It allows fast adaptation of pretrained language models to downstream tasks (Houlsby et al., 2019), and multilingual NMT models to new language pairs (Philip et al., 2020; Berard, 2021). In the second stage, we propose to use an AdapterFusion module (Pfeiffer et al., 2021) and train it to effectively combine features of the existing domains in order to handle unseen domains.

Unlike (Pfeiffer et al., 2021) who trains the AdapterFusion module for transfer learning (from existing tasks to a *seen* target task), we do *not* have access to the test domain during training. To address this challenge, we propose a novel leave-one-domain-out (LODO) training strategy by creating homogeneous mini-batches consisting of training examples from a single domain and disabling the corresponding domain adapter when optimising the fusion layer. This training strategy is related to model selection for domain generalisation (Gulra-

[*]Work done while doing internship at eBay Inc.

Findings of the Association for Computational Linguistics: ACL 2022, pages 582 - 588
May 22-27, 2022 ©2022 Association for Computational Linguistics

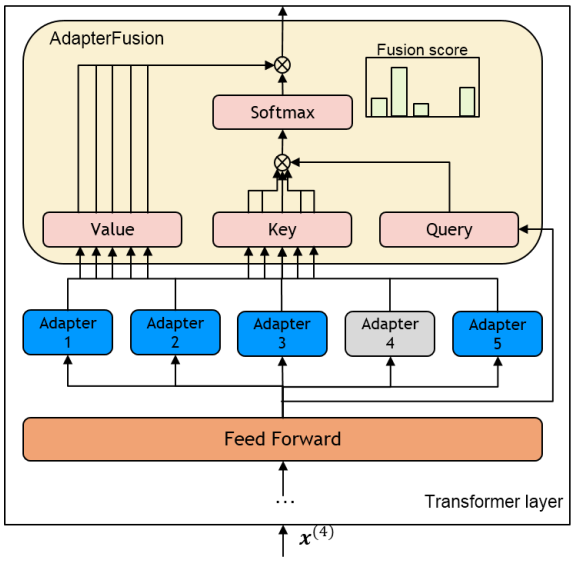

Figure 1: Leave-one-out adapter fusion training strategy

Algorithm 1 LODO Training of AdapterFusion

Function: trainFusion
Input: Training data $\mathcal{D} = \{D^1, ..., D^K\}$, NMT model θ, adapters $\{\omega_1, ..., \omega_K\}$
Output: Fusion layer ψ

while not converge **do**
  shuffle $\mathcal{D}$
  $\{b_j^{k_j}\}_{j=1}^{J} \leftarrow \{(\boldsymbol{x}, \boldsymbol{y}) \in D^{k_j} | D^{k_j} \in \mathcal{D}\}$ // create homogeneous minibatches
  for $j = 1$ **to** J **do**
    $d_j \leftarrow \{1, ..., K\} \setminus \{k_j\}$ // active domains
    for all transformer layer l **do**
      $\boldsymbol{h}^{(l)} \leftarrow \theta^{(l)}(b_j^{k_j}, \boldsymbol{h}^{(l-1)})$ // hidden state
      $\{\tilde{\boldsymbol{h}}_k^{(l)}\} \leftarrow \{\boldsymbol{h}^{(l)} + \omega_k(\boldsymbol{h}^{(l)})\}_{k \in d_j}$ // adapter outputs
      $\boldsymbol{h}^{(l)} \leftarrow \psi^{(l)}(\boldsymbol{h}^{(l)}, \{\tilde{\boldsymbol{h}}_k^{(l)}\}_{k \in d_j}, \{\tilde{\boldsymbol{h}}_k^{(l)}\}_{k \in d_j})$ // fusion output
    end for
    $\psi \leftarrow \psi - \gamma \nabla_\psi \mathcal{L}_{NMT}(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{h}^{(L)})$
  end for
end while
return ψ

jani and Lopez-Paz, 2020), where the aim is to maximise the expected performance w.r.t an unknown meta-distribution over domains. Unlike (Gulrajani and Lopez-Paz, 2020), we use LODO to train the AdapterFusion module instead of model selection.

Our contributions can be summarised as follows: (i) We extend AdapterFusion for domain generalisation, where the target domain is not available during training; (ii) We propose a novel leave-one-out training strategy to avoid over-fitting of the fusion layer to the given training domains; (iii) We demonstrate the efficacy of our proposed fusion method on three language pairs and four unseen domains. Empirical results show that our approach outperforms the learning invariant feature baseline on most of unseen test domains with an improvement up to +0.8 BLEU score on average[1].

2 Our Approach

Problem Formulation. We define domain generalisation for NMT as the problem of learning an NMT model on training datasets from multiple domains $\mathcal{D} = \{D^1, ..., D^K\}$ such that it performs well at some unseen test domain D^{K+1}. A dataset $D^k = \{(\boldsymbol{x}_i^k, \boldsymbol{y}_i^k)\}_{i=1}^{n_k}$ in domain k contains n_k examples from a distribution $\Pr_k(X, Y)$ where X ranges over sentences in the source language, and Y is its translation in the target language.

Domain-specific parameter learning. We insert a small adapter module in transformer layers of the pretrained NMT model to capture domain-specific features (Houlsby et al., 2019). Adapters are task-specific modules introduced to a pretrained network to enable fast adaptation to new tasks. Following (Pfeiffer et al., 2021), we add the adapter only after the last feed-forward layer. The adapter module includes a down-projection $\boldsymbol{W}_{down}^{(l)}$ followed by an up-projection $\boldsymbol{W}_{up}^{(l)}$ to project the hidden state $\boldsymbol{h}^{(l)}$ in layer l to a lower-dimension space then project back to high-dimension space

$$\boldsymbol{h}^{(l)} = \boldsymbol{h}^{(l)} + f(\boldsymbol{h}^{(l)}\boldsymbol{W}_{down}^{(l)})\boldsymbol{W}_{up}^{(l)} \quad (1)$$

where $f(.)$ is a nonlinear activation function. We denote $\omega = \{\boldsymbol{W}_{down}^{(l)}, \boldsymbol{W}_{up}^{(l)}\}_{l=1}^{L}$ as adapter parameters. We learn the adapter modules while freezing the encoder and decoder parameters. At training time, we only train the adapter with the training data from the corresponding domain.

Domain generalisation with AdapterFusion To help the NMT model generalise to unseen domains, we learn a fusion layer (Pfeiffer et al., 2021) to combine domain-specific adapters while freezing all other parameters. Since we do not have access to unseen domains at training time, we propose *leave-one-domain-out* (LODO) training strategy to train the fusion layer. We create *homogeneous* batches of training data from individual domains. For a given batch $b_j^{k_j}$ from domain k_j where $j \in \{1, ..., J\}$ ranges over all possible batches in a training epoch, the adapter corresponding to this domain is disabled, and the fusion layer learns to combine the

[1]Source code will be available at https://github.com/trangvu/lodo-nmt.

Domain	Train			Dev			Test		
	De-En	**Fr-En**	**Pl-En**	**De-En**	**Fr-En**	**Pl-En**	**De-En**	**Fr-En**	**Pl-En**
WMT	37.4M	35.1M	7.1M	2K	2K	5.3K	1.4K	3K	1K
LAW	454K	596K	1.3M	2K	2K	2K	2K	2K	2K
MED	705K	705K	666K	2K	2K	2K	2K	2K	2K
IT	158K	230K	97K	2K	2K	2K	2K	2K	2K
KORAN	17.8K	28K	30K	2K	2K	2K	2K	2K	2K
SUB	494K	492K	491K	2K	2K	2K	2K	2K	2K
BOOK	44K	114K	0.9K	2K	2K	459	2K	2K	516
TED	164K	190K	174K	4.1K	4.2K	4K	4.4K	4.8K	4.9K
ROBUST	-	-	-	-	-	-	1K/5.6K	-	-
TICO19	-	-	-	-	971	-	-	2.1K	-

Table 1: Number of sentences in train, dev and test set for each domain and language pairs. There are no training data released in ROBUST and TICO19 dataset. For En-De ROBUST dataset, there are two different test set for En⟶De (1K) and De⟶En (5.6K) directions.

output of other $K-1$ adapters as shown in Figure 1.

Algorithm 1 describes our proposed LODO training strategy. We denote $d_j \leftarrow \{1, ..., K\} \setminus \{k_j\}$ the index set of active adapters for a batch $b_j^{k_j}$ in domain k_j (line 5). The adaptive hidden state for domain $k \in d_j$ in transformer layer l is computed using eq. (1) (line 8). The fusion module in transformer layer l parametrised by $\psi^{(l)}$ combines the adapter outputs $\{\boldsymbol{h}_k^{(l)}\}$ using the self-attention mechanism with the adapter input as query, adapter outputs as key and value. We train the fusion module with cross-entropy loss while freezing other parameters (line 11).

3 Experiments

We evaluate our proposed approach to generalise a pretrained NMT model to unseen domains on three language pairs English-German (En-De), English-French (En-Fr), and English-Polish (En-Pl).

3.1 Experimental Setup

Dataset. The pretrained NMT models are trained on generic domain datasets from WMT2014 for En-Fr, WMT2020 for the other language pairs. Following the recipe in Koehn and Knowles (2017), we create five source domains: legal (LAW), IT (IT), Koran (KORAN), Medical (MED), and Subtitles (SUB) from OPUS (Tiedemann, 2012). We consider BOOK dataset from OPUS (Tiedemann, 2012), TED talk (TED) (Qi et al., 2018), TICO-19 (TICO19) (Anastasopoulos et al., 2020) and WMT20 Robustness task (ROBUST) (Specia et al., 2020) as unseen test domains. Data statistics are reported in Table 1.

Baselines. We consider two backbone pretrained NMT models: (i) generic-domain (*mBART*$_{WMT}$) - an mBART model (Liu et al., 2020) finetuned on WMT dataset; and (ii) multi-domain (*mBART*$_{MD}$) - an mBART model finetuned on the combination of training data from all available source domains. We evaluate our proposed domain generalisation approach against the following baselines:

- **Zeroshot** uses the pretrained backbone models *mBART*$_{WMT}$ and *mBART*$_{MD}$ to evaluate on the unseen domains.
- **Finetuning (FT)** which further trains the backbone on multi-domain datasets.
- **Adversarial domain discriminator (disc)** which adds a domain discriminator on top of the encoder to learn domain-invariant features by jointly training with MT and adversarial domain discrimination loss (Britz et al., 2017).

We also report the BLEU score of finetuning mBART on the test domain, which serves as a supervised oracle.

Architecture and hyperparameters. We finetune mBART-based models using a batch size of 4048 tokens with mixed-precision training up to 200K update steps and early stopping on 4 V100 GPUs. We apply Adam with an inverse square root schedule, a linear warmup of 5000 steps and a learning rate of 3e-5. We use dropout and label smoothing with a rate of 0.3 and 0.2. For multi-domain training, we use temperature-based sampling with $T = 1.5$ to balance training size between domains (Arivazhagan et al., 2019).

Backbone	Frozen	TrainAlg (Data)	En-De BOOK	En-De ROBUST	En-De TED	En-De avg	En-Fr BOOK	En-Fr TICO19	En-Fr TED	En-Fr avg	En-Pl BOOK	En-Pl TED	En-Pl avg	#params trained
Translate to English														
mBART	✗	sup. oracle	30.55	-	48.05	-	23.77	-	50.28	-	7.84	36.44	-	610M
mBART_{MD}	✓	zeroshot	18.87	29.10	29.87	25.95	15.04	27.50	**34.46**	25.67	5.51	22.76	14.14	-
mBART	✗	disc(MD)	18.92	30.17	29.82	26.30	14.52	28.86	34.02	25.80	5.66	**22.82**	14.24	610M
mBART_{MD}	✓	LODO(MD)	**19.15**†	**30.92**†	**30.38**†	**26.82**	**15.06**	**29.61**†	34.41	**26.36**	**6.06**†	22.81	**14.44**	37M
mBART_{WMT}	✓	zeroshot	25.99	30.93	37.30	31.41	15.26	33.52	33.43	27.40	8.23	22.69	15.46	-
mBART_{WMT}	✗	FT(MD)	16.76	30.14	30.28	25.73	14.72	27.82	33.61	25.38	5.97	21.82	13.90	610M
mBART	✗	FT(all)	18.16	30.40	28.80	25.79	13.85	27.08	33.12	24.68	5.99	20.03	13.01	610M
mBART_{WMT}	✓	LODO (MD)	**26.68**†	**31.28**†	**37.77**†	**31.91**	**15.77**†	**33.82**†	**34.00**†	**27.86**	**8.50**†	**22.85**	**15.68**	37M
Translate from English														
mBART	✗	sup. oracle	21.45	-	35.33	-	27.45	-	50.53	-	3.65	25.13	-	610M
mBART_{MD}	✓	zeroshot	11.55	28.01	25.08	21.55	20.57	25.92	32.86	26.45	4.03	19.21	11.62	-
mBART	✗	disc(MD)	11.07	28.15	**26.30**†	21.84	20.62	25.67	32.90	26.40	4.20	**20.14**	12.17	610M
mBART_{MD}	✓	LODO(MD)	**12.12**†	**28.67**†	25.89	**22.23**	**20.81**	**26.39**†	**33.03**	**26.74**	**4.38**	20.09	**12.24**	37M
mBART_{WMT}	✓	zeroshot	17.13	31.19	33.59	27.30	19.61	27.14	34.23	26.99	3.96	20.44	12.20	-
mBART_{WMT}	✗	FT(MD)	12.34	28.48	29.46	23.43	20.01	27.37	32.68	26.69	**4.34**†	20.14	12.24	610M
mBART	✗	FT(all)	11.44	28.28	24.21	21.31	19.48	27.02	30.97	25.82	3.81	20.02	11.92	610M
mBART_{WMT}	✓	LODO (MD)	**17.67**†	**32.34**†	**34.02**†	**28.01**	**20.29**†	**27.59**†	**34.57**†	**27.47**	3.89	**20.82**†	**12.36**	37M

Table 2: BLEU score on unseen test domains. The first two columns show the pretrained backbones and whether they are frozen during training: off-the-self mBART (*mBART*), finetuned mBART on WMT data ($mBART_{WMT}$), and finetuned mBART on multi-domain data ($mBART_{MD}$). The third column presents the training methods with the data used in brackets: zeroshot, finetuning (*FT*), domain discriminator (*disc*), and our proposed method (*LODO*) on the multi-domain data (MD) or all data including WMT and MD. Best and second best scores of each column are marked in **bold** and underline respectively. † indicates that the best score is statistically significant difference to the second best (p-value ≤ 0.05) using paired bootstrap resampling.

For adapter modules, we use the adapter architecture of Pfeiffer et al. (2021), which is added once only after the last feed-forward layer for each transformer layer of encoder and decoder. We set the bottleneck dimension to 256 in all experiments and use ReLU as the nonlinear activation function. We train the adapters for each domain separately with a learning rate of 2e-4 up to 120K steps with early stopping and 2000 warmup steps. Other hyperparameters are the same as in the mBART finetuning.

Following Pfeiffer et al. (2021), we initialise the value matrix V of the fusion layer with a diagonal of ones and the rest with random weights of a small norm 1e-6. The query matrix Q and key matrix K are initialised randomly. We train the fusion layers with a learning rate of 5e-5 up to 200K steps with early stopping and 10K warmup updates.

Evaluation. We report BLEU scores calculated by SacreBLEU (Post, 2018)[2].

3.2 Main Result and Ablation

We present the results on unseen domains for from-English and to-English translation of three language pairs in Table 2. There are big gaps between the supervised oracles and the domain generalisation methods, except for the En-Pl BOOK domain where the training data is relatively small. Overall, the $mBART_{WMT}$ backbones outperform the $mBART_{MD}$ backbones on the unseen domains. It is expected that WMT datasets can be considered as generic, and the $mBART_{MD}$ backbone may overfit to the seen domains.

Finetuning the $mBART_{WMT}$ backbones on multi-domain datasets (*FT(MD)*) degrades performance on unseen domain significantly. We observe a similar trend when finetuning on both WMT and multi-domain datasets (*FT(all)*). It may be due to dataset imbalance and negative interference between domains. Learning domain-invariant fea-

	En⟶De Book	En⟶De ROBUST	En⟶De TED	De⟶En Book	De⟶En ROBUST	De⟶En TED
LODO-homo	**17.67**	**32.34**	34.02	**26.68**	**31.28**	**37.77**
LODO-mixed	17.28	32.15	**34.62**	26.24	31.16	35.67
all-homo	16.82	31.76	33.88	25.79	30.79	32.33
all-mixed	16.12	31.82	33.52	26.33	30.22	32.75

Table 3: Ablation of fusion layer training strategies on (i) leave-one-domain-out training (*LODO*) vs. fusion all adapters (*all*), and (ii) whether to have homogeneous batches (*homo*) or mixed-domain batches (*mixed*). All models are trained with the $mBART_{WMT}$ backbone.

[2] `nrefs:1|case:mixed|eff:no|tok:none|smooth:exp|version:2.0.0`

tures (*disc(MD)*) are able to improve BLEU score on unseen domains over the $mBART_{MD}$ backbone. On average, our proposed fusion method outperforms other baselines in most translation directions without retraining the backbones.

Ablation on AdapterFusion training strategy. We do an ablation study of our LODO training strategy with homogeneous batches on En-De with the $mBART_{WMT}$ backbone in Table 3. Compared to LODO, we observe performance drop on all domains when activating all adapters. However, there is no significant difference between homogeneous and mixed batches.

4 Related works

Domain generalisation for NMT. Domain generalisation has been mostly studied in computer vision (Wang et al., 2021b). The main approaches include invariant feature learning (Li et al., 2018; Wang et al., 2020b), data augmentation (Wang et al., 2020a), and meta learning (Balaji et al., 2018; Wang et al., 2021a). Although domain mismatch is a known challenge in NMT (Müller et al., 2020), domain generalisation has just recently drawn attention with the introduction of zeroshot evaluation in WMT2020 Robustness shared task (Specia et al., 2020), but is still under-explored.

Adapters. Adapter-based methods have been shown effective in transferring to new languages in multilingual NMT (Üstün et al., 2021; Berard, 2021; Cooper Stickland et al., 2021; Zhu et al., 2021) and fast adaptation to new domains (Bapna and Firat, 2019). Combining task-specific adapters with attention mechanism (Pfeiffer et al., 2021) or ensemble (Wang et al., 2021c) allows efficient transfer to low-resource natural language understanding (NLU) and NMT tasks. When target domain examples are unavailable, adapters can be combined during inference to better generalise to unseen domains for NLU tasks (Gururangan et al., 2021).

5 Conclusion

In this paper, we propose a fusion-based approach to the domain generalisation problem for NMT. Our method first captures domain-specific features via adapters, then learns to combine them with leave-one-out strategy training. Experiments show the effectiveness of our methods without retraining the NMT backbone. Hence, it is a potential method to quickly incorporate newly arriving domains into the existing NMT systems.

Acknowledgments

This research is supported by an eBay Research Award and the ARC Future Fellowship FT190100039. This work is partly sponsored by the Air Force Research Laboratory and DARPA under agreement number FA8750-19-2-0501. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The authors are grateful to the anonymous reviewers for their helpful comments to improve the manuscript.

References

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federman, Dmitriy Genzel, Francisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the Translation initiative for COvid-19. arXiv:2007.01788.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. 2018. Metareg: Towards domain generalization using meta-regularization. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for 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 1538–1548, Hong Kong, China. Association for Computational Linguistics.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Alexandre Berard. 2021. Continual learning in multilingual NMT via language-specific embeddings. In *Proceedings of the Sixth Conference on Machine Translation*, pages 542–565, Online. Association for Computational Linguistics.

Denny Britz, Quoc Le, and Reid Pryzant. 2017. Effective domain mixing for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Asa Cooper Stickland, Alexandre Berard, and Vassilina Nikoulina. 2021. Multilingual domain adaptation for NMT: Decoupling language and domain information with adapters. In *Proceedings of the Sixth Conference on Machine Translation*, pages 578–598, Online. Association for Computational Linguistics.

Ishaan Gulrajani and David Lopez-Paz. 2020. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*.

Suchin Gururangan, Mike Lewis, Ari Holtzman, Noah A Smith, and Luke Zettlemoyer. 2021. Demix layers: Disentangling domains for modular language modeling. *arXiv preprint arXiv:2108.05036*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. 2018. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5400–5409.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Mathias Müller, Annette Rios, and Rico Sennrich. 2020. Domain robustness in neural machine translation. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (AMTA 2020)*, pages 151–164, Virtual. Association for Machine Translation in the Americas.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

Jerin Philip, Alexandre Berard, Matthias Gallé, and Laurent Besacier. 2020. Monolingual adapters for zero-shot neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4465–4470, Online. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. 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. Association for Computational Linguistics.

Lucia Specia, Zhenhao Li, Juan Pino, Vishrav Chaudhary, Francisco Guzmán, Graham Neubig, Nadir Durrani, Yonatan Belinkov, Philipp Koehn, Hassan Sajjad, Paul Michel, and Xian Li. 2020. Findings of the WMT 2020 shared task on machine translation robustness. In *Proceedings of the Fifth Conference on Machine Translation*, pages 76–91, Online. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of LREC*.

Ahmet Üstün, Alexandre Berard, Laurent Besacier, and Matthias Gallé. 2021. Multilingual unsupervised neural machine translation with denoising adapters. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6650–6662, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bailin Wang, Mirella Lapata, and Ivan Titov. 2021a. Meta-learning for domain generalization in semantic parsing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 366–379, Online. Association for Computational Linguistics.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, and Tao Qin. 2021b. Generalizing to unseen domains: A survey on domain generalization. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4627–4635. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Xinyi Wang, Yulia Tsvetkov, Sebastian Ruder, and Graham Neubig. 2021c. Efficient test time adapter ensembling for low-resource language varieties. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 730–737, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yufei Wang, Haoliang Li, and Alex C Kot. 2020a. Heterogeneous domain generalization via domain mixup. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3622–3626. IEEE.

Zhen Wang, Qiansheng Wang, Chengguo Lv, Xue Cao, and Guohong Fu. 2020b. Unseen target stance detection with adversarial domain generalization. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Yaoming Zhu, Jiangtao Feng, Chengqi Zhao, Mingxuan Wang, and Lei Li. 2021. Counter-interference adapter for multilingual machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2812–2823, Punta Cana, Dominican Republic. Association for Computational Linguistics.