On the Complementarity between *Pre-Training* and *Random-Initialization* for Resource-Rich Machine Translation

Changtong Zan[1*], **Liang Ding**[2†], **Li Shen**[2], **Yu Cao**[3], **Weifeng Liu**[1†], **Dacheng Tao**[2,3]

[1]College of Control Science and Engineering, China University of Petroleum (East China)
[2]JD Explore Academy [3]The University of Sydney
`b20050011@s.upc.edu.cn, dingliang1@jd.com`

Abstract

Pre-Training (PT) of text representations has been successfully applied to low-resource Neural Machine Translation (NMT). However, it usually fails to achieve notable gains (sometimes, even worse) on resource-rich NMT on par with their Random-Initialization (RI) counterpart. We take the first step to investigate the complementarity between PT and RI in resource-rich scenarios via two probing analyses, and find that: ❶ PT improves NOT the *accuracy*, but the *generalization* by achieving flatter loss landscapes than that of RI; ❷ PT improves NOT the confidence of lexical choice, but the *negative diversity*[1] by assigning smoother lexical probability distributions than that of RI. Based on these insights, we propose to combine their complementarities with a model fusion algorithm that utilizes optimal transport to align neurons between PT and RI. Experiments on two resource-rich translation benchmarks, WMT'17 English-Chinese (20M) and WMT'19 English-German (36M), show that PT and RI could be nicely complementary to each other, achieving substantial improvements considering both translation accuracy, generalization, and negative diversity. Probing tools and code are released at: https://github.com/zanchangtong/PTvsRI.

1 Introduction

Pre-training (**PT**; Devlin et al., 2019; Liu et al., 2019 has achieved tremendous success in natural language processing fields. Inspired by BERT (Devlin et al., 2019), recent works (Song et al., 2019; Lewis et al., 2020; Liu et al., 2020) attempt to leverage sequence-to-sequence PT for neural machine translation (NMT; Bahdanau et al., 2015; Gehring et al., 2017; Vaswani et al., 2017) by leveraging a large amount of unlabeled (i.e. monolingual) sentences.

While recent studies have empirically shown their benefit for the *low-resource* translation task where the labeled (i.e. parallel) sentences are limited (Lewis et al., 2020; Song et al., 2019; Liu et al., 2020), we are generally confronted with *resource-rich* scenarios, e.g. millions of parallel sentence pairs, in WMT evaluations (Akhbardeh et al., 2021) and industries. For these resource-rich tasks, PT becomes less effective (sometimes, even worse) than their Random-Initialization (**RI**) counterparts, for example, as Zhu et al. 2020; Liu et al. 2020 reported, the PT underperforms RI if improperly utilized or significant amount of bi-text data is given. However, there is limited understanding of: ① Why does PT fail compared to RI in terms of translation accuracy in high-resource settings? ② What is the difference between the optimized PT and RI models? ③ How can we harmonize PT with RI? Can we just leverage their advantages?

To this end, we introduce two probing analyses (i.e. loss landscape and lexical probability distribution) for the fully-optimized PT and RI models to investigate ① and ②, respectively. We find that: ❶ Supervised training with significant amount of parallel data, i.e. RI, is enough to optimize the model towards a better optimum point compared with PT, thus leading to better in-domain (same domain between test and training set) performance; ❷ PT mainly contributes to the generalization ability because of its flatter loss landscape and smoother lexical probability distribution, thus resulting in better out-domain performance and diversified generation. Motivated by this finding, we propose a simple combination method to reach ③ by aligning the neurons and weights

---

[1]Li et al. 2020 mentioned that the maximum likelihood estimation training treats all non-ground truth but semantic-relevant lexical predictions as being equally incorrect, potentially leading to the sharp lexical probability distribution. Here we use this term to refer to the higher probabilities assigned to the negative tokens.

(a) Loss Landscape of RI **(b) Loss Landscape of PT** **(c) Loss Landscape of Ours**
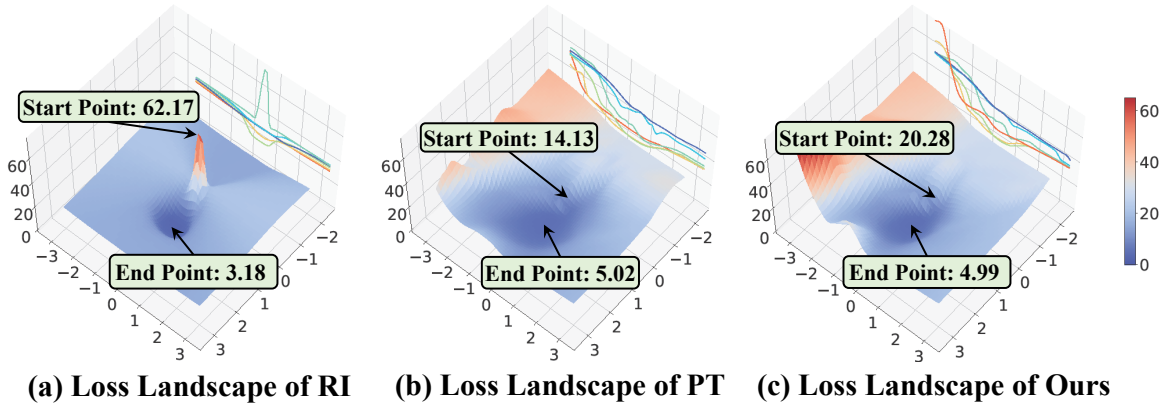
Figure 1: Loss landscapes with contour lines on WMT19 En-De dataset. "RI" means the transformer training from scratch, while "PT" denotes the pretrained model mBART with finetuning. Both models have the same architecture.
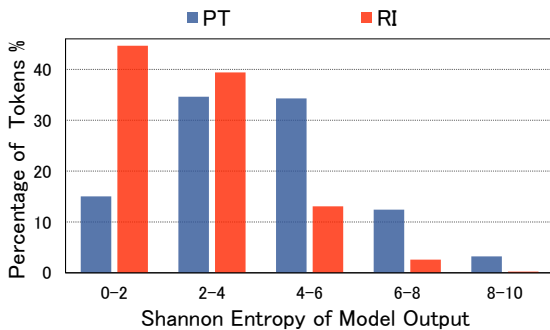

Figure 2: Token ratio distribution over difference entropy scales on WMT19 En-De test set. We compute the shanon entropy of the decoder output for each token.

of PT and RI and then fusing them into a single model based on optimal transport (Monge, 1781; Singh and Jaggi, 2020). Experiments conducted on the WMT17 English-Chinese and WMT19 English-German benchmarks show that PT can nicely complement RI, leading to better translation accuracy, generalization, and negative diversity.

2   Experimental Setup

Data   We conducted experiments on WMT17 English-Chinese (En-Zh) and WMT19 English-German (En-De) translation tasks, which are widely-used resource-rich benchmarks, and include 20M and 36M sentence pairs, respectively.

Setting   To make a fair comparison, all the model backbone architectures are the same as the pretrained `mBART.cc25`[2]. We tokenize with 250K SentencePiece model (Kudo and Richardson, 2018) of mBART and remove tokens that are not present

in the downstream task from the vocabulary. For PT, we also initialize the model with trimmed embedding layers. For evaluation, we use Sacre-BLEU (Post, 2018) to measure the translation quality with default tokenizer.

3   Understanding PT and RI

In this section, we aim to better understand the similarities and differences between PT and RI by introducing two probing analyses. The analyzed PT and RI models are trained on WMT19 En-De benchmark. We first present the loss landscape visualization of both models in Section 3.1, and then show their difference in lexical probability distribution in Section 3.2.

3.1   Impact on Loss Landscape

We follow Hao et al. (2019b) to visualize the two-dimensional (2D) loss surface, providing some insights on why PT fails compared to RI for in-domain accuracy while resulting better out-domain performance. Let θ_0 denotes the initialized parameters, i.e., mBART parameters or randomly initialized parameters. θ_1 and θ_2 are the parameters of models trained on target dataset and auxiliary dataset respectively. We plot the loss surface with function as follows:

$$f(\alpha, \beta) = \mathcal{L}(\boldsymbol{\theta}_0 + \alpha\boldsymbol{\delta}_1 + \beta\boldsymbol{\delta}_2), \qquad (1)$$

where $\mathcal{L}$ is the loss function, α and β are scalar values represent current coordinate, $\boldsymbol{\delta}_1 = \theta_1 - \theta_0$ is the the optimization direction on target dataset, and $\boldsymbol{\delta}_2 = \frac{\|\theta_1 - \theta_0\|}{\|\theta_2 - \theta_0\|} (\theta_2 - \theta_0)$ denotes the normalized optimization direction on auxiliary dataset, $\|\cdot\|$ denotes the Euclidean norm. In our setting, we select

mBART finetuned on WMT17 En-Zh data as the second axis to plot the loss landscape on WMT19 En-De benchmark.

Figure 1(a) and Figure 1(b) present the loss landscape with contour line of model training with PT and RI, where the left-side and right-side axis are corresponding to α and β in Equation 1 respectively. We also mark the start point and the end point of training path. The start point denotes the loss of model without training, while the end point indicates the converged model loss. Firstly, we can see that PT obtains flatter loss landscape than RI with contour line covers a wider area, which denotes that initialized with pre-trained parameters leads model to be more robust. This supports us to understand the generalization of PT from the perspective of the flatter optima. Secondly, the start point of PT is much smaller than RI (14.13 vs. 62.17), which demonstrates that the knowledge learned during pre-training stage is also suitable for the downstream task. We also observe that RI achieves the lowest loss at the end point (3.18). This shows that the RI model is well optimized on the in-domain data, as the large amount sentence pairs provide enough supervised signal.

3.2 Impact on Lexical Prob. Distribution

Previous work (Jiang et al., 2019) shows that the lexical probability distribution is correlated with the diversity of generation, i.e., over-confidence in certain tokens leads to lower diversity (Miao et al., 2021). This finding inspires us to measure the density of PT and RI outputs. We use Shannon entropy of decoder output according to test set to probe this attribute. The smaller entropy represents less assigned probability density on incorrect tokens.

As shown in Figure 2, we present how much percentage of character is involved in different intervals. We firstly see that most tokens prefer a relatively smaller entropy, which denotes that both models are well optimized, and predict tokens with higher confidence. Besides, PT achieves a smoother distribution than RI with most outputs achieving relative higher entropy. This suggests that PT could generate more diversiform translations, which is consistent with the report from Wang et al. (2022a).

4 Harmonizing PT and RI

Based on the above experiments, we prove that PT mainly contributes to the generalization ability and

RI leads to better in-domain performance. To combine the advantage of both models, we fuse both models via optimal transport based on activations following Singh and Jaggi (2020).

4.1 Model Fusion

Discrete Optimal Transport Given two discrete measures $\sum_{i=1}^{n} \alpha_i \delta_{\alpha_i}$ and $\sum_{i=1}^{n} \beta_i \delta_{\beta_i}$, where δ_{β_i} denotes the Dirac measures on β_i, and $\sum_{i=1}^{n} \alpha_i = 1$ and $\sum_{i=1}^{m} \beta_i = 1$. Then, with the cost matrix $\mathbf{C} \in \mathbb{R}^{n \times m}$, which computed with predefined cost function, the discrete optimal transport formulates:

$$\text{OT} := \min_{\mathbf{T} \in \Pi(\boldsymbol{\alpha}, \boldsymbol{\beta})} \langle \mathbf{T}, \mathbf{C} \rangle_F \qquad (2)$$

where $\langle \cdot, \cdot \rangle_F$ is the Frobenius product, and $\Pi(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \{\mathbf{T} \in \mathbb{R}^{n \times m} : \mathbf{T}\mathbb{1}_m = \boldsymbol{\alpha} \text{ and } \mathbf{T}^\top \mathbb{1}_n = \boldsymbol{\beta}\}$ is the set of all admissible transport matrix between the two measures ($\mathbb{1}_n$ is a vector of ones of size n).

Model Fusion For layers ℓ_{ri} and ℓ_{pt} of RI and PT models respectively, we define the activation based probability measures over neurons as $\mu^{\ell_{ri}} = (\alpha^{\ell_{ri}}, \delta^{\ell_{ri}})$ and $\nu^{\ell_{pt}} = (\beta^{\ell_{pt}}, \delta^{\ell_{pt}})$, where $\delta^{\ell_{ri}}$, $\delta^{\ell_{pt}}$ are activations for given samples. And, we set uniform distribution for α and β.

In term of the alignment procedure, we first align the weights of previous layer $\mathbf{W}_{ri}^{\ell, \ell-1}$ by constructing convex combination with the weights of corresponding transport matrix $\mathbf{T}^{\ell-1}$, normalized appropriately via the inverse of corresponding column marginals α. After getting the aligned $\widehat{\mathbf{W}}_{ri}^{\ell, \ell-1}$, we compute the optimal transport map $\mathbf{T}^\ell$ between $\mu^{(\ell_{ri})}$ and $\nu^{(\ell_{pt})}$, which is formulated as Equation 2. And now, we align ℓ_{ri} with respect to ℓ_{pt},

$$\widetilde{\mathbf{W}}_{ri}^{(\ell, \ell-1)} \leftarrow \text{diag}\left(\frac{1}{\vec{\beta}^{(\ell)}}\right) \mathbf{T}^{(\ell)\top} \widehat{\mathbf{W}}_{ri}^{(\ell, \ell-1)}. \quad (3)$$

We will refer $\widetilde{\mathbf{W}}_{ri}^{(\ell, \ell-1)}$ to the aligned weights of RI model, which we can directly add with $\mathbf{W}_{pt}^{(\ell, \ell-1)}$. We carry out this procedure over all layers sequentially. Then, we fine-tune the fused model on corresponding dataset to obtain the final model.

4.2 Results

As detailed in Section 2, we conduct the experiments on two widely used benchmarks WMT17 En-Zh and WMT19 En-De. To obtain the fused model, we set weights of 0.9 for PT and 0.1 for RI during model fusion and then fine-tune on the corresponding dataset.

Model	En-De	En-Zh
RI	40.1	32.9
PT	39.3	32.3
+Fusion	39.7$^{\Uparrow+0.4}$	32.4$^{\Uparrow+0.1}$

Table 1: In-Domain translation quality of model trained on the WMT19 En-De and WMT17 En-Zh benchmarks. We evaluate models on official test sets. PT denotes the fine-tuned mBART, while RI indicts model trained from scratch.

	PT	**RI**	**Fusion**	**Δ**
1) OD	34.9	35.0	35.2	**+0.2**
2) DuA	-12.1	-16.7	-12.3	**+4.4**
3) Multi-Ref	75.4	75.8	76.0	**+0.2**
4) TTR	19.1	19.1	19.2	**+0.1**

Table 2: Results of analysis experiments. "OD" (Out-Domain) evaluates the performance on the out-domain test set. "DuA" (Drop under Attack) reports the drop of BLEU when adding noise into source sentences. "Multi-Ref" is the BLEU score with multiple references. "TTR" denotes the Type-Token-Ratio. "Δ" indicts difference between RI and Fusion.

As present in Figure 1(c), we present the loss landscape of fused model. The start point is the fused model before fine-tuning. We can see that our model inherits the smooth loss landscape from PT and leads to smaller loss value.

The in-domain translation results are shown in Table 1. We present the results of PT and RI, which has the same network architecture, for comparison with our model. As seen, RI performs better than PT for two benchmarks (39.3 vs. 39.7 for En-De, 32.3 vs. 32.4 for En-De). This empirically verifies the claim we make in Section 3.1. Besides, our fused model achieves a better performance than PT (39.3 vs. 39.7 for En-De, 32.3 vs. 32.4 for En-De), which denotes that model fusion could promote the performance of PT. while still under-performs RI (40.1 vs. 39.7 for En-De, 32.9 vs. 32.4 for En-De). This may draw from that our approach can be seen as a trade-off of PT and RI.

4.3 Analysis

We conducted several analyses to understand whether it inherits the advantages of PT. We evaluate our approach on four tasks, compared with PT model and RI model, including model generalization, performance drop under attack, translation di-

versity and type-token-ratio. Note that we can measure more properties, e.g. copy errors (Liu et al., 2021b) and encoded linguistic knowledge (Hao et al., 2019a; Ding et al., 2020), but our aim is simply to show that PT and RI can be harmonized. All results are reported on the WMT19 En-De benchmark. For simplification, we list scores in one table.

1) Effects of Model Generalization To understand how our model performs on the out-domain data. We evaluate trained model on two different domains, including It and Medical, and report Out-Domain Score by averaging SacreBLEU scores. As it presents in the first line of Table 2, the fused model achieves +0.2 score compared with RI. This verifies the generalization ability of our model.

2) Performance Drop under Attack To further demonstrate that our fused model could be more robust than RI. We attack the model by adding noises into source sentence during the inference stage and report the drop in score. Specifically, we replace 10% words with a special token. The less drop mean model is more robust. As seen in the third line of Table 2, both models suffer a huge drop with the perturbed input. However, our model still performs better than RI (-29.6 vs. -31.2). This result shows that our model is less sensitive to noises, which leads to more robustness.

3) Effects of Translation Diversity Following Wang et al. (2022a), we evaluate the translation diversity of models with multi-reference SacreBLEU. We inference on the test set released by Ott et al. (2018), which consists of ten additional reference translations for 500 English sentences extracted from WMT14 En-De test set. As shown in the second line of Table 2, our model achieves a higher multi-reference SacreBLEU score than RI. This empirically proves that our fused model could generate a more diverse translation with different word orders.

4) Effects of TTR We also probe diversity of the lexical with **TTR** = $\frac{num.\ of\ types}{num.\ of\ tokens}$ following Wei et al. (2022). The higher score of **TTR** inflects that the model tends to generate translation with different word types. As it present in the last line of Table 2, our fusion model achieves +0.16 gains over RI. This shows that our model could not only generate translation with more word orders but also predict words of more types.

5 Conclusion and Future Work

This paper provides several insights for pre-training (PT) and random-initialization (RI) on resource-rich machine translation. We carefully investigate the impact of PT and RI on loss landscape and negative diversity and reveal that RI leads to a lower loss value of in-domain data, while PT is more beneficial to flatter loss and smoother lexical distribution. We also propose to combine them with model fusion via optimal transport. Experimental results on WMT17 En-Zh and WMT19 En-De benchmarks show that model fusion is a nice trade-off method to utilize the complementarity of PT and RI.

For future work, we would explore other effective fusion algorithms (Liu et al., 2022; Wang et al., 2022b) to combine the merits of PT and RI. Also, it will be interesting to investigate whether harmonizing PT and RI could complement existing data augmentation methods for NMT, e.g. back translation (Liu et al., 2021a), data rejuvenation (Jiao et al., 2020), bidirectional distillation (Ding et al., 2021a) and training (Ding et al., 2021b).

Acknowledgements

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina Espaa-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (wmt21). In *WMT*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Liang Ding, Longyue Wang, Xuebo Liu, Derek F. Wong, Dacheng Tao, and Zhaopeng Tu. 2021a. Rejuvenating low-frequency words: Making the most of parallel data in non-autoregressive translation. In *ACL*.

Liang Ding, Longyue Wang, Di Wu, Dacheng Tao, and Zhaopeng Tu. 2020. Context-aware cross-attention for non-autoregressive translation. In *COLING*.

Liang Ding, Di Wu, and Dacheng Tao. 2021b. Improving neural machine translation by bidirectional training. In *EMNLP*.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *ICML*.

Jie Hao, Xing Wang, Baosong Yang, Longyue Wang, Jinfeng Zhang, and Zhaopeng Tu. 2019a. Modeling recurrence for transformer. In *NAACL*.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2019b. Visualizing and understanding the effectiveness of BERT. In *EMNLP*.

Shaojie Jiang, Pengjie Ren, Christof Monz, and Maarten de Rijke. 2019. Improving neural response diversity with frequency-aware cross-entropy loss. In *The Web Conference*.

Wenxiang Jiao, Xing Wang, Shilin He, Irwin King, Michael Lyu, and Zhaopeng Tu. 2020. Data rejuvenation: Exploiting inactive training examples for neural machine translation. In *EMNLP*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *EMNLP*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*.

Zuchao Li, Rui Wang, Kehai Chen, Masso Utiyama, Eiichiro Sumita, Zhuosheng Zhang, and Hai Zhao. 2020. Data-dependent gaussian prior objective for language generation. In *ICLR*.

Chang Liu, Chenfei Lou, Runzhong Wang, Alan Yuhan Xi, Li Shen, and Junchi Yan. 2022. Deep neural network fusion via graph matching with applications to model ensemble and federated learning. In *ICML*.

Xuebo Liu, Longyue Wang, Derek F Wong, Liang Ding, Lidia S Chao, Shuming Shi, and Zhaopeng Tu. 2021a. On the complementarity between pre-training and back-translation for neural machine translation. In *Findings of EMNLP*.

Xuebo Liu, Longyue Wang, Derek F. Wong, Liang Ding, Lidia S. Chao, Shuming Shi, and Zhaopeng Tu. 2021b. On the copying behaviors of pre-training for neural machine translation. In *Findings of ACL*.

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. *TACL*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv*.

Mengqi Miao, Fandong Meng, Yijin Liu, Xiao Zhou, and Jie Zhou. 2021. Prevent the language model from being overconfident in neural machine translation. In *ACL*.

Gaspard Monge. 1781. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *ICML*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *WMT*.

Sidak Pal Singh and Martin Jaggi. 2020. Model fusion via optimal transport. In *NeurIPS*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: masked sequence to sequence pre-training for language generation. In *ICML*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.

Wenxuan Wang, Wenxiang Jiao, Yongchang Hao, Xing Wang, Shuming Shi, Zhaopeng Tu, and Michael Lyu. 2022a. Understanding and improving sequence-to-sequence pretraining for neural machine translation. In *ACL*.

Zhenyi Wang, Xiaoyang Wang, Li Shen, Qiuling Suo, Kaiqiang Song, Dong Yu, Yan Shen, and Mingchen Gao. 2022b. Meta-learning without data via wasserstein distributionally-robust model fusion. In *UAI*.

Xiangpeng Wei, Heng Yu, Yue Hu, Rongxiang Weng, Weihua Luo, and Rong Jin. 2022. Learning to generalize to more: Continuous semantic augmentation for neural machine translation. In *ACL*.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. 2020. Incorporating BERT into neural machine translation. In *ICLR*.