On the Complementarity between Pre-Training and Back-Translation for Neural Machine Translation

Xuebo Liu[1]*, Longyue Wang[2], Derek F. Wong[1], Liang Ding[3],
Lidia S. Chao[1], Shuming Shi[2] and Zhaopeng Tu[2]
[1]NLP^{2}CT Lab, Department of Computer and Information Science, University of Macau
[2]Tencent AI Lab [3]The University of Sydney
[1]nlp2ct.xuebo@gmail.com, {derekfw,lidiasc}@um.edu.com
[2]{vinnylywang,shumingshi,zptu}@tencent.com
[3]ldin3097@sydney.edu.au

Abstract

Pre-training (PT) and back-translation (BT) are two simple and powerful methods to utilize monolingual data for improving the model performance of neural machine translation (NMT). This paper takes the first step to investigate the complementarity between PT and BT. We introduce two probing tasks for PT and BT respectively and find that PT mainly contributes to the encoder module while BT brings more benefits to the decoder. Experimental results show that PT and BT are nicely complementary to each other, establishing state-of-the-art performances on the WMT16 English-Romanian and English-Russian benchmarks. Through extensive analyses on sentence originality and word frequency, we also demonstrate that combining Tagged BT with PT is more helpful to their complementarity, leading to better translation quality. Source code is freely available at https://github.com/SunbowLiu/PTvsBT.

1 Introduction

Neural machine translation (NMT; Bahdanau et al., 2015; Gehring et al., 2017; Vaswani et al., 2017) models are data-hungry and their performances are highly dependent upon the quantity and quality of labeled data, which are expensive and scarce resources (Leong et al., 2021). This motivates the research line of exploiting unlabeled monolingual data for boosting the model performance of NMT. Due to simplicity and effectiveness, pre-training (PT; Devlin et al., 2019; Song et al., 2019) and back-translation (BT; Sennrich et al., 2016b) are two widely-used techniques for NMT, by leveraging a large amount of monolingual data.

While empirically successful, the understandings of PT and BT are still limited at best. Several attempts have been made to better understand them at the data level, e.g. exploring different kinds of noises for the source data (Edunov et al., 2018; Lewis et al., 2020). However, there are few understandings at the model level that how PT and BT affect the internal module (e.g. encoder and decoder) of NMT models. As recent studies start to combine PT and BT for better model performance (Conneau and Lample, 2019; Liu et al., 2020b; Ding et al., 2021c), there is a pressing need to broaden the understandings of them.

To this end, we introduce two probing tasks to investigate the effects of PT and BT on the encoder and decoder modules, respectively. We find that PT mainly contributes to the encoder module while BT brings more benefits to the decoder module. This provides a good explanation for the performance improvement of simply combining PT and BT. Motivated by this finding, we explore a better combination method by leveraging *Tagged BT* (Caswell et al., 2019). Experiments conducted on the WMT16 English-Romanian and English-Russian benchmarks show that PT can nicely cowork with BT, leading to state-of-the-art model performances. Extensive analyses show that the tagging mechanism is helpful for enhancing the complementarity between PT and BT by improving the translation of source-original sentences and low-frequency words.

Our **main contributions** are as follows:

- We design two probing tasks to investigate the impact of PT and BT on NMT models.
- We empirically demonstrate the complementarity between PT and BT.
- We show that Tagged BT further improves the complementarity between PT and BT.

2 Preliminaries

2.1 Background

Pre-Training for NMT Self-supervised PT (Devlin et al., 2019; Song et al., 2019), which can ac-

*Work was done when Xuebo Liu and Liang Ding were interning at Tencent AI Lab.

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2900–2907
November 7–11, 2021. ©2021 Association for Computational Linguistics

quire knowledge from unlabeled monolingual data, has shown its effectiveness in improving the model performance of NMT, especially for those language pairs with smaller parallel corpora (Conneau and Lample, 2019).

The first research line treats pre-trained models as external knowledge to guidance NMT to learn better representations (Yang et al., 2020a; Zhu et al., 2020) and predictions (Chen et al., 2020). These methods are effective but costly since the NMT architecture needed to be elaborately designed. Another research line is directly taking the weights of pre-trained models to initialize NMT models, which is easy to use and advancing the state-of-the-art (Rothe et al., 2020; Lewis et al., 2020). In this paper, we treat pre-trained mBART (Liu et al., 2020b) as our testbed for parameter initialization, whose benefits have been sufficiently validated (Tran et al., 2020; Tang et al., 2020; Liu et al., 2021a) by multiple translation directions.

In general, previous studies focus on designing novel architectures (Song et al., 2019) and artificial noises for source sentences (Lin et al., 2020; Yang et al., 2020b) but are still unclear why pre-training can boost the model performance of NMT, which is this paper aims to investigate.

Back-Translation for NMT BT is an alternative to leverage monolingual data for NMT (Sennrich et al., 2016b). It first trains a reversed NMT model for translating target-side monolingual data into synthetic parallel data, and then complements them with the original parallel data to train the desired NMT model. To improve BT, previous works put attention to the importance of diversity and complexity in synthetic data, showing that adding symbols (e.g., noises and tags) to the back-translated source can help NMT distinguish the data from various sources and learn better representations (Fadaee and Monz, 2018; Wang et al., 2019; Edunov et al., 2018; Caswell et al., 2019; Marie et al., 2020). The claims and understandings from these works are chiefly at the data-level rather than the model-level.

There also exists some works that combine PT and BT to further boost the model performance (Conneau et al., 2020; Song et al., 2019; Liu et al., 2020b). However, the relation between BT and PT has not been fully studied. In this paper, we take the first step to understand BT and PT at the model-level and improve the complementarity between PT and BT.

2.2 Experimental Setup

Data We conducted experiments on the WMT16 English-Romanian (En-Ro) and English-Russia (En-Ru) translation tasks, which are widely-used benchmarks of data augmentation methods for NMT. The training/validation/test sets of the En-Ro include 612K/2K/2K sentence pairs, while those of En-Ru include 2M/3K/3K pairs. Towards better reproducibility, we directly used the BT data provided by Sennrich et al. (2016a)[1], consisting of 2.3M synthetic data for the En-Ro and 2.0M data for the En-Ru. All the data are tokenized and split into sub-words (Sennrich et al., 2016c) by the mBART tokenizer (Liu et al., 2020b).

Setting To make a fair comparison, all the model architectures and parameters are the same as the pre-trained `mBART.cc25`.[2] The NMT model augmented with PT directly uses the mBART weights for parameter initialization, while the other models randomly initialize their parameters. The training follows Liu et al. (2020b) except that we tuned the learning rate within [3e-5,1e-3] and the dropout within [0.3,0.5] for the vanilla model and BT models. We used the single model with the best validation perplexity for testing. The length penalty is 1.0 and the beam size is 5. We used sacreBLEU (Post, 2018) to calculate BLEU (Papineni et al., 2002) and TER (Snover et al., 2006) scores with the specific tokenization (Liu et al., 2020b) for Romanian and the default tokenization for Russian.

3 Understanding PT and BT

In this section, we aim to better understand the similarities and differences between PT and BT on improving model performance. We design two probing tasks to study the research question: *Which module of NMT do PT and BT respectively play a greater role in enhancing translation quality?*

3.1 Effects of PT on NMT

Given a pre-trained model, it is common to use its part or all parameters to initialize the downstream tasks. We design four NMT models, which differ from the NMT components (Encoder vs. Decoder) with parameter initialization manners (Random vs. Pre-trained). As shown in Table 1, the variants

[1] `http://data.statmt.org/rsennrich/wmt16_backtranslations`

[2] `https://github.com/pytorch/fairseq/tree/master/examples/mbart`

Enc.	Dec.	PT		BT	
		BLEU	Δ	BLEU	Δ
N	N	33.7	-	33.7	-
N	Y	33.5	-0.2	**37.8**	**+4.1**
Y	N	**36.9**	**+3.2**	35.8	+2.1
Y	Y	37.7	+4.0	38.3	+4.6

Table 1: The probing tasks of PT and BT. NMT models are trained and evaluated on the WMT16 En-Ro benchmark. "Y" denotes the corresponding parameters are activated when augmented with PT or BT, while "N" denotes the inactive operation. PT and BT respectively contribute more to the NMT encoder and decoder.

are: 1) **NN** is a vanilla NMT model, of which parameters are randomly initialized; 2) **NY** means that parameters of NMT encoder are randomly initialized while those of decoder are initialized with pre-training; 3) **YN** is contrary to **NY**; and 4) **YY** indicates that the whole NMT parameters are initialized with the pre-trained model. After that, the NMT models are fine-tuned on the parallel data with the same training strategy.

PT Mainly Contributes to Encoder As seen, the **YY** initialization strategy significantly improves the vanilla NMT model by +4.0 BLEU scores, which reconfirms the effectiveness of PT for translation tasks (Liu et al., 2020b). By comparing **NY** and **YN**, we find that the pre-trained encoder can still help the NMT to achieve +3.2 BLEU improvements while the pre-trained decoder can only perform on par with the vanilla model (i.e. -0.2 BLEU). This demonstrates that PT mainly contributes to the encoder part of NMT model, and this claim is consistent with the conclusion with other pre-trained models. For instance, Rothe et al. (2020) show that the NMT encoder initialization is superior to the decoder one when using pre-trained weights of BERT. We hypothesize that the performance boost with PT mainly comes from the better ability of source-side understanding, which is significant to NMT such as on disambiguating word senses (Tang et al., 2019).

3.2 Effects of BT on NMT

A vanilla NMT model is trained on the original bi-text and then fine-tuned on the mixture of the original and synthetic (i.e. back-translated) data. We also design four NMT models, which differ from which parts of parameters are updated at the fine-tuning stage. As shown in Table 1, the variants are: 1) **NN** is a vanilla NMT model only trained on the original data; 2) **NY** indicates that parameters of the NMT encoder are fixed while those of decoder are updated during model fine-tuning; 3) **YN** acts in an opposite way compared with **NY**; 4) **YY** means that the whole NMT parameters are updated at the fine-tuning stage.

BT Mainly Contributes to Decoder BT has been sufficiently validated to improve the performance of NMT models (Edunov et al., 2018, 2020). By exploiting additional target sentences, the NMT decoder can be enhanced to generate more fluent sentences in the target language. In contrast, the synthetic source sentences contain noises, which may be less useful for improving the ability of understanding. The results verify our hypothesis: BT mainly improves the decoder module of NMT. As seen, fine-tuning the whole NMT model (i.e. **YY**) with BT data can gain the best performance (+4.6 BLEU than the vanilla model), which shows the effectiveness of BT method. Surprisingly, only fine-tuning the decoder (i.e. **NY**) can perform close to **YY** model (37.8 vs. 38.3 BLEU), which confirms our claims. Compared with **NY**, the **YN** model obtains relatively fewer improvements (+4.1 vs. +2.1 BLEU), showing that BT brings more benefits to the decoder than the encoder.

4 Improving PT and BT

The answer of the research question in Section 3 is: *PT and BT respectively contribute more to the NMT encoder and decoder, demonstrating that they are orthogonal and complementary to each other.* This finding motivates us to better combine these two individual techniques together for further improving NMT models.

4.1 Experiments

As detailed in Section 2.2, we conducted experiments on two commonly-used benchmarks En-Ro and En-Ru. Besides, we train the BT models from scratch instead of fine-tuning in Section 3.2. As **YY** models (in Table 1) always achieve best performances when augmented PT or BT, we update all parameters of NMT models in next experiments.

The results are shown in Table 2. We use the vanilla model as our baselines, which are trained on original datasets with random initialization. Besides, we report results on existing PT models as our strong baselines, including XLM-R, mRASP,

Model	En-Ro		En-Ru	
	BLEU	**TER**	**BLEU**	**TER**
Existing Baselines				
XLM-R (Conneau et al., 2020)	35.6	-	-	-
mRASP (Lin et al., 2020)	37.6	-	-	-
mBART (Liu et al., 2020b)	37.7	-	-	-
Our Implemented Systems				
Vanilla NMT	33.7	48.6	28.8	61.6
+ PT	37.7	45.0	31.6	58.5
+ BT	38.4	45.0	31.1	59.2
+ BT + PT	41.2	42.6	33.2	57.1
+ Tagged BT	38.6	44.9	31.2	59.3
+ Tagged BT + PT	**41.6**	**42.1**	**33.6**	**56.5**

Table 2: Translation quality on the En-Ro and En-Ru benchmarks. "+" means incorporating PT and (Tagged) BT into NMT models.

mBART. As seen, PT can significantly improve the translation quality in all cases compared with vanilla baselines (averagely +2.5 BLEU), which is consistent with (or better than) existing PT models (37.7 vs. 35.6∼37.7 BLEU). Furthermore, two BT methods[3] (i.e. BT and Tagged BT) perform closely, which improves the standard NMT models by +3.5/+3.7 BLEU points on average. Simply combining them (+BT+PT) can further boost performances for NMT models across different sizes of datasets, showing the robustness and effectiveness of this approach. Encouragingly, the combination of Tagged BT and PT performs better than the simple one, leading to state-of-the-art performances on the two benchmarks. Similar tendencies are observed in terms of the TER scores. The above results illustrate the better complementarity between PT and Tagged BT on improving translation quality for NMT models.

4.2 Analysis

We conducted extensive analyses to better understand the improvement of our approach. All results are reported on the En-Ro benchmark.

Effects of Sentence Type Recent studies have shown that the evaluation of BT is sensitive to the sentences types, thus we report BLEU scores on the subsets of source-original (Src-Ori) and target-original (Tgt-Ori) datasets (Zhang and Toral, 2019; Liu et al., 2021a; Wang et al., 2021).[4] Generally speaking, the translation of Src-Ori is more important than that of Tgt-Ori for practical NMT systems (Graham et al., 2020), thus its performance should be taken seriously. As shown in Table 3, PT performs better on Src-Ori than BT (33.8 vs. 31.9 BLEU) while BT achieves higher scores on Tgt-Ori than PT (45.6 vs. 42.0 BLEU). Besides, simply combining PT and BT can improve the translation quality on both Src-Ori and Tgt-Ori sentences, but the improvement of Src-Ori is lower than only using PT. By introducing tagged BT, the model can achieve better performance than the simple one, especially on source-original sentences. **Takeaway:** *1) PT and BT complementary in terms of originality of sentences; 2) Tagged BT can alleviate the bias of translating Tgt-Ori sentences which is significant to practical NMT systems.*

Model	All	Src	Tgt
Vanilla	33.7	29.4	38.3
+ PT	37.7	33.8	42.0
+ BT	38.4	31.5	45.4
+ BT + PT	41.2	33.3	48.6
+ Tagged BT	38.6	31.9	45.6
+ Tagged BT + PT	**41.6**	**34.8**	**48.7**

Table 3: Translation quality of source-original and target-original sentences on the En-Ro benchmark. "Src" and "Tgt" respectively denote the sub-testsets of source-original and target-original while "All" means the whole testset.

Model	All	Low	High
Vanilla	62.8	48.5	64.6
+ PT	65.8	58.2	66.7
+ BT	65.9	57.5	67.1
+ BT + PT	67.8	60.8	68.8
+ Tagged BT	66.1	57.5	67.3
+ Tagged BT + PT	**68.3**	**61.8**	**69.1**

Table 4: F-measure of word translation according to frequency on the En-Ro benchmark. "Low" and "High" respectively denote the buckets of low- and high-frequency words while "All" means the whole words in the test set. Simply combining PT and BT improves the model performance, while adding tags to BT data further improves

[3] Tagged BT is to add a special token at the beginning of each back-translated source sentence.

[4] Src-Ori denotes the testing data originating in the source language, while Tgt-Ori denotes the data translating from the target language.

Effects of Word Frequency Data augmentation is an effective way to improve the translation quality of low-frequency words (Sennrich et al., 2016b). Thus, we compare the performance of the models on translating different frequencies of words. Specifically, we employed *compare-mt* (Neubig et al., 2019) to calculate the f-measure of translating low- and high-frequency words (<50 vs. ≥50). As shown in Table 4, PT improves more on translating low-frequency words (58.2 vs. 57.5 scores) while BT performs better on high-frequency words (67.3 vs. 66.7 scores). Furthermore, the combination of PT and tagged BT achieves the best performance on both low- and high-frequency words, leading to an overall improvement on the whole words. Similar phenomenons can be observed by combining self-training and BT (Ding et al., 2021b). **Takeaway:** *1) PT and BT complementary in terms of frequency of words; 2) Tagged BT are more complementary to PT on lexical translation.*

5 Conclusion and Future Works

This paper broadens the understandings of pre-training (PT) and back-translation (BT). We propose two probing tasks to investigate the impact of PT and BT on each NMT module and find that PT is more beneficial to the encoder while BT mainly improves the decoder. Experimental results on the WMT16 English-Romanian and English-Russian benchmarks show that PT is nicely complementary to BT. We also demonstrate that Tagged BT (i.e., adding tags to BT data) can further improve the complementarity of translating source-original sentences and low-frequency words.

In the future, we would like to apply curriculum learning (Liu et al., 2020a; Zhan et al., 2021; Ding et al., 2021a) to better organize the learning of PT and BT. It is also worthwhile to explore other kinds of methods utilizing monolingual data (e.g., self-training (Zhang and Zong, 2016; He et al., 2020; Jiao et al., 2021)) and validate the findings on practical NMT systems (Huang et al., 2021) and more generation tasks (Liu et al., 2021b).

Acknowledgement

This work was supported in part by the Science and Technology Development Fund, Macau SAR (Grant No. 0101/2019/A2), and the Multi-year Research Grant from the University of Macau (Grant No. MYRG2020-00054-FST). We thank the anonymous reviewers for their insightful comments.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy. Association for Computational Linguistics.

Yen-Chun Chen, Zhe Gan, Yu Cheng, Jingzhou Liu, and Jingjing Liu. 2020. Distilling knowledge learned in BERT for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7893–7905, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Liang Ding, Longyue Wang, Xuebo Liu, Derek F. Wong, Dacheng Tao, and Zhaopeng Tu. 2021a. Progressive multi-granularity training for non-autoregressive translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2797–2803, Online. Association for Computational Linguistics.

Liang Ding, Longyue Wang, Xuebo Liu, Derek F. Wong, Dacheng Tao, and Zhaopeng Tu. 2021b. Rejuvenating low-frequency words: Making the most of parallel data in non-autoregressive translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages

3431–3441, Online. Association for Computational Linguistics.

Liang Ding, Di Wu, and Dacheng Tao. 2021c. Improving neural machine translation by bidirectional training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 2020. On the evaluation of machine translation systems trained with back-translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2836–2846, Online. Association for Computational Linguistics.

Marzieh Fadaee and Christof Monz. 2018. Back-translation sampling by targeting difficult words in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 436–446, Brussels, Belgium. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252. PMLR.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81, Online. Association for Computational Linguistics.

Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. 2020. Revisiting self-training for neural sequence generation. In *International Conference on Learning Representations*.

Guoping Huang, Lemao Liu, Xing Wang, Longyue Wang, Huayang Li, Zhaopeng Tu, Chengyan Huang, and Shuming Shi. 2021. Transmart: A practical interactive machine translation system. *CoRR*, abs/2105.13072.

Wenxiang Jiao, Xing Wang, Zhaopeng Tu, Shuming Shi, Michael Lyu, and Irwin King. 2021. Self-training sampling with monolingual data uncertainty for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2840–2850, Online. Association for Computational Linguistics.

Chongman Leong, Xuebo Liu, Derek F. Wong, and Lidia S. Chao. 2021. Exploiting translation model for parallel corpus mining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2829–2839.

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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. Pre-training multilingual neural machine translation by leveraging alignment information. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2649–2663, Online. Association for Computational Linguistics.

Xuebo Liu, Houtim Lai, Derek F. Wong, and Lidia S. Chao. 2020a. Norm-based curriculum learning for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 427–436, Online. Association for Computational Linguistics.

Xuebo Liu, Longyue Wang, Derek F. Wong, Liang Ding, Lidia S. Chao, Shuming Shi, and Zhaopeng Tu. 2021a. On the copying behaviors of pre-training for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4265–4275, Online. Association for Computational Linguistics.

Xuebo Liu, Longyue Wang, Derek F. Wong, Liang Ding, Lidia S. Chao, and Zhaopeng Tu. 2021b. Understanding and improving encoder layer fusion in sequence-to-sequence learning. In *International Conference on Learning Representations*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020b. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Benjamin Marie, Raphael Rubino, and Atsushi Fujita. 2020. Tagged back-translation revisited: Why does it really work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5990–5997, Online. Association for Computational Linguistics.

Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, and Xinyi Wang. 2019. compare-mt:

A tool for holistic comparison of language generation systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 35–41, Minneapolis, Minnesota. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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. 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, Brussels, Belgium. Association for Computational Linguistics.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016c. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200. Citeseer.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5926–5936. PMLR.

Gongbo Tang, Rico Sennrich, and Joakim Nivre. 2019. Encoders help you disambiguate word senses in 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 1429–1435, Hong Kong, China. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Chau Tran, Yuqing Tang, Xian Li, and Jiatao Gu. 2020. Cross-lingual retrieval for iterative self-supervised training. In *Advances in Neural Information Processing Systems*, volume 33, pages 2207–2219. Curran Associates, Inc.

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 *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Shuo Wang, Yang Liu, Chao Wang, Huanbo Luan, and Maosong Sun. 2019. Improving back-translation with uncertainty-based confidence estimation. 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 791–802, Hong Kong, China. Association for Computational Linguistics.

Shuo Wang, Zhaopeng Tu, Zhixing Tan, Shuming Shi, Maosong Sun, and Yang Liu. 2021. On the language coverage bias for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4778–4790, Online. Association for Computational Linguistics.

Jiacheng Yang, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Weinan Zhang, Yong Yu, and Lei Li. 2020a. Towards making the most of bert in neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9378–9385.

Zhen Yang, Bojie Hu, Ambyera Han, Shen Huang, and Qi Ju. 2020b. CSP:code-switching pre-training for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2624–2636, Online. Association for Computational Linguistics.

Runzhe Zhan, Xuebo Liu, Derek F. Wong, and Lidia S. Chao. 2021. Meta-curriculum learning for domain adaptation in neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14310–14318.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Austin, Texas. Association for Computational Linguistics.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81, Florence, Italy. Association for Computational Linguistics.

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 *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.