Progressive Multi-Granularity Training for Non-Autoregressive Translation

Liang Ding[*]
The University of Sydney
ldin3097@sydnye.edu.au

Longyue Wang[*]
Tencent AI Lab
vinnylywang@tencent.com

Xuebo Liu
University of Macau
nlp2ct.xuebo@gmail.com

Derek F. Wong
University of Macau
derekfw@um.edu.com

Dacheng Tao
JD Explore Academy, JD.com
dacheng.tao@gmail.com

Zhaopeng Tu
Tencent AI Lab
zptu@tencent.com

Abstract

Non-autoregressive translation (NAT) significantly accelerates the inference process via predicting the entire target sequence. However, recent studies show that NAT is weak at learning high-mode of knowledge such as one-to-many translations. We argue that modes can be divided into various granularities which can be learned from easy to hard. In this study, we empirically show that NAT models are prone to learn fine-grained lower-mode knowledge, such as words and phrases, compared with sentences. Based on this observation, we propose progressive multi-granularity training for NAT. More specifically, to make the most of the training data, we break down the sentence-level examples into three types, i.e. words, phrases, sentences, and with the training goes, we progressively increase the granularities. Experiments on Romanian-English, English-German, Chinese-English and Japanese-English demonstrate that our approach improves the phrase translation accuracy and model reordering ability, therefore resulting in better translation quality against strong NAT baselines. Also, we show that more deterministic fine-grained knowledge can further enhance performance.

1 Introduction

Non-autoregressive translation (NAT, Gu et al., 2018) has been proposed to improve the decoding efficiency by predicting all tokens independently and simultaneously. Different from autoregressive translation (AT, Vaswani et al., 2017) models that generate each target word conditioned on previously generated ones, NAT models suffer from the multimodality problem (i.e. multiple translations for a single input), in which the conditional independence assumption prevents a model from properly capturing the highly multimodal distribution of target translations. To reduce the modes of training data, sequence-level knowledge distillation (KD) (Kim and Rush, 2016) is widely employed via replacing their original target samples with sentences generated from an AT teacher (Gu et al., 2018; Zhou et al., 2020; Ren et al., 2020).

Granular.	AT	NAT			
		Raw	△	KD	△
WORD	59.8	57.1	-2.7	59.0	-0.8
PHRASE	36.0	31.7	-4.3	34.2	-1.8
SENTENCE	29.2	24.5	-4.7	27.0	-2.2

Table 1: Translation performance at different granularity on the WMT14 English⇒German dataset. "△" indicates the performance gap between the NAT and AT.

Although KD reduces the learning difficulty for NAT, there are still complicated word orders and structures (Gell-Mann and Ruhlen, 2011) in the synthetic sentences, making the NAT performance sub-optimal. To answer this challenge, Saharia et al. (2020); Ran et al. (2021) propose to lowers the bilingual modeling difficulties under the *monotonicity assumption*, where bilingual sentences are in the same word order. However, they make extensive modifications to model structures or objectives, limiting the applicability of their methods to a boarder range of tasks and languages.

Accordingly, we turn to break down the sentence-level high modes into finer granularities, i.e. bilingual words and phrases, where we assume that finer granularities are easy to be learned by NAT. As shown in Table 1, we analyzed the translation accuracy at three linguistic levels (i.e. word, phrase and sentence) and found that although KD brings promising improvements at three granular-

[*] Liang Ding and Longyue Wang contributed equally to this work. Work was done when Liang Ding and Xuebo Liu were interning at Tencent AI Lab.

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 2797–2803

	Source	Targets
Word	bank	银行 岸 储库
Phrase	hollow structural	中空 结构 空心 的 结构 镂空 结构
Sentence	He is very good at English.	他 英文 很 好。 他 非常 擅长 英语 。 他的 英语 水平 很 高 。

Table 2: Examples of different translation granularities.

ities, there are still some gaps with AT teacher. Also, we showed that finer granularities are easier to be learned, that is, accuracy gap "Δ" of WORD is small than that of PHRASE, and SENTENCE (0.8<1.8<2.2). Thus, we propose a simple and effective training strategy to enhance the ability to handle the sentence-level high modes. More specifically, we generate bilingual lexicons from parallel data by leveraging word alignment and phrase extraction in statistical machine translation (SMT, Zens et al., 2002). Then we guide the NAT model to progressively learn the bilingual knowledge from low to high granularity. Experimental results on four commonly-cited translation benchmarks show that our proposed PROGRESSIVE MULTI-GRANULARITY (PMG) training strategy consistently improves the translation performance. The main contributions are:

- Our study reveals that NAT is better at learning fine-grained knowledge. Training with sentences merely may be sub-optimal.
- We propose PMG training to encourage NAT models to learn from easy to hard. The fine-grained knowledge distilled by SMT will be dynamically transferred during training.
- Experiments across language pairs and model structures show the effectiveness and universality of PMG training.

2 Methodology

2.1 Motivation

We investigated theories in second-language acquisition: one usually learns a foreign language from word-to-word translation to sentence-to-sentence translation, namely from local to global (Onnis et al., 2008). Bilingual knowledge is at the core of adequacy modeling (Tu et al., 2016), which is a major weakness of the NAT models due to the lacks of autoregressive factorization. Table 2 demonstrates the English⇒Chinese multimodality at different granularities (i.e. word, phrase, sentence levels). As seen, the sentence-level consists of various kinds of modes, including word alignment ("English" vs. "英语"/"英文"), phrase translation ("be good at" vs. "...非常 擅长..."/"...水平 很 高"), and even reordering ("英语" can be subject or object). However, phrase-level modes are less complex with similar structure and word-level modes are simple with token-to-token mapping. Generally, the lower level of bilingual knowledge, the easier for NAT to learn. This example explains why the sentence level performance gaps between NAT and AT are significant than that of word and phrase in Table 1. Based on the above evidence, it is natural to suspect that the existing sentence-level NAT training is sub-optimal.

2.2 Fine-grained Bilingual Knowledge

Phrase table is an essential component of SMT systems, which records the correspondence between bilingual lexicons (Koehn and Callison-Burch, 2009). For each training example in the original training set, we sample its all possible inter-sentence bi-lingual phrases from the phrase table that obtained with phrase-based statistical machine translation (PBSMT) model (Koehn et al., 2003). The GIZA++ (Och and Ney, 2003) was employed to build word alignments for the training datasets. We leave the exploitation of more advanced forms bilingual knowledge such as syntax rules (Liu et al., 2006) and discontinuous phrases (Galley and Manning, 2010) for future work. Take the sentence pair in Table 2 for example, we can obtain the bilingual En-Zh phrase pairs "*very good* ||| 很好", "*good at English* ||| 擅长英语" from original sentence pair, informing the NAT model the explicit phrase boundaries.

2.3 Progressive Multi-Granularity Training

We present an extremely simple progressive multi-granularity (PMG) training fashion. Concretely, we progressively schedule the PMG: learn from "low" to "high" granularity, i.e. word→phrase→sentence. And we empirically set the training steps for each training stage. Our work can be seen as a typical determinism-based curriculum learning (CL) (Bengio et al., 2009) method, where the finer granular-

Models	Speed	BLEU			
		Ro-En	En-De	Zh-En	Ja-En
AT Models					
Transformer-BASE (Ro-En Teacher)	1.0×	34.1	27.3	24.4	29.2
Transformer-BIG (En-De / Zh-En / Ja-En Teacher)	0.8×	n/a	29.2	25.3	29.8
Existing NAT Models					
NAT (Gu et al., 2018)	2.4×	31.4	19.2	n/a	n/a
Iterative NAT (Lee et al., 2018)	2.0×	30.2	21.6	n/a	n/a
DisCo (Kasai et al., 2020)	3.2×	33.3	26.8	n/a	n/a
Levenshtein (Gu et al., 2019)	3.5×	33.3	27.3	n/a	n/a
Mask-Predict (Ghazvininejad et al., 2019)	1.5×	33.3	27.0	23.2	n/a
Context-aware NAT (Ding et al., 2020b)	1.5×	33.2	27.5	24.6	29.4
Our NAT Models					
Levenshtein (Gu et al., 2019)	3.5×	33.2	27.4	24.4	29.1
+PMG Training		33.8†	27.8	25.0†	29.6
Mask-Predict (Ghazvininejad et al., 2019)	1.5×	33.3	27.0	24.0	28.9
+PMG Training		33.7	27.6†	24.5	29.5†

Table 3: Comparison with previous work on WMT16 Ro-En, WMT14 En-De, WMT17 Zh-En and WAT17 Ja-En datasets. "†" indicates that the proposed method was significantly better than baseline at significance level $p<0.05$.

ities are more deterministic than sentences. Thus we compare with typical CL works (Zhang et al., 2019; Platanios et al., 2019) in Section 3.2.

3 Experiment

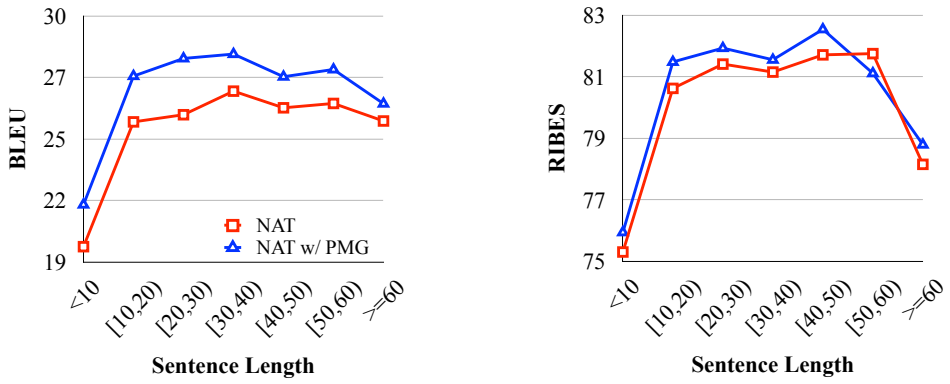
3.1 Setup

Data Experiments were conducted on four widely-used translation datasets: WMT14 English-German (En-De), WMT16 Romanian-English (Ro-En), WMT17 Chinese-English (Zh-En) and WAT17 Japanese-English (Ja-En), which consist of 4.5M, 0.6M, 20M and 2M sentence pairs, respectively. It is worthy noting that Ro-En, En-De and Zh-En are low-, medium- and high- resource language pairs, and Ja-En is word order divergent language direction. We use the same validation and test datasets with previous works for fair comparison. To avoid unknown works, we preprocessed data via byte-pair encoding (BPE) (Sennrich et al., 2016) with 32K merge operations. We evaluated the translation quality with BLEU (Papineni et al., 2002) with statistical significance test (Collins et al., 2005). For fine-grained bilingual knowledge, e.g. word alignment and phrase table, to ensure the source to target mapping more deterministic, we set 0.05 as the probability threshold. Taking WMT14 En-De for example, there are 3M words and 156M phrases in the original phrase table extracted by SMT methodology. We then filter the items whose translation probability is lower than 0.05 and obtain 0.3M words and 56.5M phrases as the final data.

Non-Autoregressive Models We validated our progressive multi-granularity training strategy on two state-of-the-art NAT model structures:

- *Mask-Predict* (MaskT, Ghazvininejad et al. 2019) that uses the conditional mask LM (Devlin et al., 2019) to iteratively generate the target sequence from the masked input;
- *Levenshtein Transformer* (LevT, Gu et al. 2019) that introduces three steps: deletion, placeholder prediction and token prediction.

For regularization, we empirically set the dropout rate as 0.2, and apply weight decay with 0.01 and label smoothing with $\epsilon = 0.1$. We train batches of approximately 128K tokens using Adam (Kingma and Ba, 2015). The learning rate warms up to 5×10^{-4} in the first 10K steps, and then decays with the inverse square-root schedule. We train 50k steps on word-level data and 50k steps on phrase-level data, respectively. And then update the remaining 200K steps for sentence-level training. Following the common practices (Ghazvininejad et al., 2019; Kasai et al., 2020), we evaluate the performance on an ensemble of 5 best checkpoints (ranked by validation BLEU) to avoid stochasticity.



(a) Translation Quality (b) Reordering Difference

Figure 1: Performances of our proposed approach on different length bins against the vanilla NAT model.

Autoregressive Teachers We closely followed previous works to apply sequence-level KD. More precisely, we trained two kinds of Transformer (Vaswani et al., 2017) models, including Transformer-BASE and Transformer-BIG. The main results employ BIG for all directions except Ro-En, which is distilled by BASE. The architectures of Transformer-BIG utilizes a large batch (458K tokens) training strategy.

3.2 Experimental Results

Main Results Table 3 lists the results of previous competitive NAT models (Gu et al., 2018; Kasai et al., 2020; Gu et al., 2019; Ghazvininejad et al., 2019). Clearly, our approach "+PMG Training" consistently improves translation performance (BLEU↑) over four language pairs. Specifically, our PMG training strategy achieves on average +0.53 BLEU scores improvements on four language pairs upon two NAT model structures. Note that our approaches introduce no extra parameters, thus does not increase any latency ("Speed").

Comparison to Curriculum Learning The existing CL methods can be divided into two categories, "Discretized CL (DCL)" (Zhang et al., 2019) and "Continuous CL (CCL)" (Platanios et al., 2019). Sentence length is the most significant variable in our multi-granularity data, therefore we implemented discretized and continuous CL with the sentence length (source side) criteria.

Our DCL setting explicitly predefined the number of data bins, while CCL method continuously samples the shorter examples with the training progresses. For DCL, we split the training samples into a predefined number of bins (5, in our case). As for CCL, we employ their length curriculum and square root competence function. We find that on WMT14 En-De dataset with MaskT model, DCL performs worse than KD baseline (-0.6 BLEU) while CCL outperforms KD baseline by +0.3 BLEU points. Our approach (+0.6 BLEU) is the most effective one.

3.3 Analysis

In this section, we conducted analytical experiments to better understand what contributes to translation performance gains. Specifically, we investigate whether the PMG 1) enhance the phrasal pattern modeling ability? 2) improve the reordering? and 3) gain better performance with higher quality fine-grained knowledge?

Better Phrasal Pattern Modelling Our method is expected to pay more attention on the bi-lingual phrases, leading to better phrase translation accuracy. To evaluate the accuracy of phrase translations, we calculate the improvement over multiple granularities of n-grams in Table 4, our PMG training "NAT w/ PMG" consistently outperforms the baseline, indicating that our proposed multi-granularity training indeed raise the ability of NAT model on capturing the phrasal patterns.

Better Reordering Ability The SMT-distilled bilingual phrasal information could intuitively inform the NAT model the bi-lingual phrasal boundaries, leading to better reordering ability. We compare the reordering ability of NAT model w/ & w/o PMG training with RIBES[1] (Isozaki et al., 2010), which is designed for measuring the reordering performance for distant language pairs. We cate-

[1] http://www.kecl.ntt.co.jp/icl/lirg/ribes

N-gram	2	3	4	5	6
Δ BLEU	0.5	0.3	0.3	0.2	0.2

Table 4: Improvements of our proposed PMG training strategy on different N-grams against vanilla NAT.

Ratio	10%	35%	50%	100%
Δ BLEU	+0.3	+0.6	+0.7	+0.6

Table 5: Improvement of PMG training strategy on different fine-grained data scales against vanilla NAT.

gorize the test set into several bins according the sentence length and report the BLEU and RIBES scores, simultaneously in Figure 1. As seen, the proposed PMG training strategy could improve the translation (BLEU↑) and reordering performance (RIBES↑), confirming our claim. Our finding is consistent with Ding et al. (2020a), where they explicitly injected the SMT-guided alignment information into the MT models, achieving better performance.

Effect of Fine-Grained Text Quality The acquired fine-grained bilingual knowledge, i.e. word alignments and phrase tables, still have extremely large volumes after filtering. Taking WMT14 En-De for example, there are over 56M phrase pairs after filtering with translation probability threshold 0.05. To make the knowledge being more deterministic, we control the quality of fine-grained text with the third party scorer – BERTScore (Zhang et al., 2020). As illustrated in Table 5, keeping the high quality bilingual knowledge (e.g. 50%) can achieve further improvements, showing the great potential of our approach. We will leave the exploration of high-quality bilingual knowledge for NAT as a future work.

4 Related Works

Non-Autoregressive Translation There still exists a performance gap between AT teacher and its NAT student. To bridge this gap, many studies have been proposed. Ghazvininejad et al. (2019); Gu et al. (2019); Kasai et al. (2020) designed novel model structures to considerably improve the NAT model capacity. Wang et al. (2019); Ran et al. (2021); Ding et al. (2021b); Du et al. (2021) explored to improve the model performance with additional training signals or objectives. Guo et al. (2020b); Su et al. (2021) delivered the knowledge from pretrained language models to the NAT models. Above works improve the NAT at the model level, while we improve NAT at the data level.

Most related to our work, Ding et al. (2021a) proposed data-level strategies, including reverse distillation and bidirectional distillation, to make the most of the parallel data. Differently, we break the sentences into fine-grained granularities to fully exploit the parallel data. Note that our model-agnostic method can be applied to any NAT structures.

Curriculum Learning Our proposed training strategy is a novel technique for NAT by exploiting curriculum learning (CL). Recent works have shown that CL can help the autoregressive translation (AT) models achieve fast convergence and better results (Platanios et al., 2019; Liu et al., 2020b; Zhan et al., 2021; Zhou et al., 2021). However, CL for non-autoregressive translation (NAT) models has not been well studied. Among the few attempts, Guo et al. (2020a); Liu et al. (2020a) respectively investigated "parameter- and task-level" curriculum learning approaches, while we proposed progressive multi-granularity training for NAT at "data-level". To the best of our knowledge, this is the first work to investigate the effects of different granularities of data on NAT models.

5 Conclusion

In this paper, we investigated the translation accuracy of different granularities in NAT, and found that the NAT models are better at dealing with fine-grained bilingual knowledge (e.g. words and phrases). Based on this finding, we proposed a simple progressive multi-granularity training strategy. Experiments show that our approach consistently and significantly improves translation performance across language pairs and model architectures. In-depth analyses indicate that our approach generates better word order and phrase patterns, outperforming typical curriculum learning methods.

Acknowledgments

We are grateful to the anonymous reviewers and the area chair for their insightful comments and suggestions. Xuebo Liu and Derek F. Wong were 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).

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *ICML*.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *ACL*.

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, Xuebo Liu, Derek F. Wong, Dacheng Tao, and Zhaopeng Tu. 2021b. Understanding and improving lexical choice in non-autoregressive translation. In *ICLR*.

Liang Ding, Longyue Wang, and Dacheng Tao. 2020a. Self-attention with cross-lingual position representation. In *ACL*.

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

Cunxiao Du, Zhaopeng Tu, and Jing Jiang. 2021. Order-agnostic cross entropy for non-autoregressive machine translation. In *ICML*.

Michel Galley and Christopher D Manning. 2010. Accurate non-hierarchical phrase-based translation. In *NAACL*.

Murray Gell-Mann and Merritt Ruhlen. 2011. The origin and evolution of word order. *PNAS*.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-predict: Parallel decoding of conditional masked language models. In *EMNLP*.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. 2018. Non-autoregressive neural machine translation. In *ICLR*.

Jiatao Gu, Changhan Wang, and Junbo Zhao. 2019. Levenshtein transformer. In *NeurIPS*.

Junliang Guo, Xu Tan, Linli Xu, Tao Qin, Enhong Chen, and Tie-Yan Liu. 2020a. Fine-tuning by curriculum learning for non-autoregressive neural machine translation. In *AAAI*.

Junliang Guo, Zhirui Zhang, Linli Xu, Hao-Ran Wei, Boxing Chen, and Enhong Chen. 2020b. Incorporating bert into parallel sequence decoding with adapters. In *NeurIPS*.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *EMNLP*.

Jungo Kasai, James Cross, Marjan Ghazvininejad, and Jiatao Gu. 2020. Parallel machine translation with disentangled context transformer. In *ICML*.

Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. In *EMNLP*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Philipp Koehn and Chris Callison-Burch. 2009. *Statistical Machine Translation*. Citeseer.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *NAACL*.

Jason Lee, Elman Mansimov, and Kyunghyun Cho. 2018. Deterministic non-autoregressive neural sequence modeling by iterative refinement. In *EMNLP*.

Jinglin Liu, Yi Ren, Xu Tan, Chen Zhang, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2020a. Task-level curriculum learning for non-autoregressive neural machine translation. In *IJCAI*.

Xuebo Liu, Houtim Lai, Derek F Wong, and Lidia S Chao. 2020b. Norm-based curriculum learning for neural machine translation. In *ACL*.

Yang Liu, Qun Liu, and Shouxun Lin. 2006. Tree-to-string alignment template for statistical machine translation. In *ACL*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*.

Luca Onnis, Heidi R Waterfall, and Shimon Edelman. 2008. Learn locally, act globally: Learning language from variation set cues. *Cognition*, 109(3).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell. 2019. Competence-based curriculum learning for neural machine translation. In *NAACL*.

Qiu Ran, Yankai Lin, Peng Li, and Jie Zhou. 2021. Guiding non-autoregressive neural machine translation decoding with reordering information. In *AAAI*.

Yi Ren, Jinglin Liu, Xu Tan, Zhou Zhao, Sheng Zhao, and Tie-Yan Liu. 2020. A study of non-autoregressive model for sequence generation. In *ACL*.

Chitwan Saharia, William Chan, Saurabh Saxena, and Mohammad Norouzi. 2020. Non-autoregressive machine translation with latent alignments. In *EMNLP*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *ACL*.

Yixuan Su, Deng Cai, Yan Wang, David Vandyke, Simon Baker, Piji Li, and Nigel Collier. 2021. Non-autoregressive text generation with pre-trained language models. In *EACL*.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *ACL*.

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*.

Yiren Wang, Fei Tian, Di He, Tao Qin, ChengXiang Zhai, and Tie-Yan Liu. 2019. Non-autoregressive machine translation with auxiliary regularization. In *AAAI*.

Richard Zens, Franz Josef Och, and Hermann Ney. 2002. Phrase-based statistical machine translation. In *AAAI*.

Runzhe Zhan, Xuebo Liu, Derek F Wong, and Lidia S Chao. 2021. Meta-curriculum learning for domain adaptation in neural machine translation. In *AAAI*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *ICLR*.

Xuan Zhang, Pamela Shapiro, Gaurav Kumar, Paul McNamee, Marine Carpuat, and Kevin Duh. 2019. Curriculum learning for domain adaptation in neural machine translation. In *NAACL*.

Chunting Zhou, Graham Neubig, and Jiatao Gu. 2020. Understanding knowledge distillation in non-autoregressive machine translation. In *ICLR*.

Lei Zhou, Liang Ding, Kevin Duh, Shinji Watanabe, Ryohei Sasano, and Koichi Takeda. 2021. Self-guided curriculum learning for neural machine translation. *arXiv*.