Rejuvenating Low-Frequency Words: Making the Most of Parallel Data in 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

Knowledge distillation (KD) is commonly used to construct synthetic data for training non-autoregressive translation (NAT) models. However, there exists a discrepancy on low-frequency words between the distilled and the original data, leading to more errors on predicting low-frequency words. To alleviate the problem, we directly expose the raw data into NAT by leveraging pretraining. By analyzing directed alignments, we found that KD makes low-frequency source words aligned with targets more deterministically but fails to align sufficient low-frequency words from target to source. Accordingly, we propose reverse KD to rejuvenate more alignments for low-frequency target words. To make the most of authentic and synthetic data, we combine these complementary approaches as a new training strategy for further boosting NAT performance. We conduct experiments on five translation benchmarks over two advanced architectures. Results demonstrate that the proposed approach can significantly and universally improve translation quality by reducing translation errors on low-frequency words. Encouragingly, our approach achieves 28.2 and 33.9 BLEU points on the WMT14 English-German and WMT16 Romanian-English datasets, respectively. Our code, data, and trained models are available at `https://github.com/longyuewangdcu/RLFW-NAT`.

1 Introduction

Recent years have seen a surge of interest in non-autoregressive translation (NAT, Gu et al., 2018), which can improve the decoding efficiency by predicting all tokens independently and simultaneously. The *non-autoregressive factorization* breaks conditional dependencies among output tokens, which prevents a model from properly capturing the highly multimodal distribution of target translations. As a result, the translation quality of NAT models often lags behind that of autoregressive translation (AT, Vaswani et al., 2017) models. To balance the trade-off between decoding speed and translation quality, knowledge distillation (KD) is widely used to construct a new training data for NAT models (Gu et al., 2018). Specifically, target sentences in the distilled training data are generated by an AT teacher, which makes NAT easily acquire more deterministic knowledge and achieve significant improvement (Zhou et al., 2020).

Previous studies have shown that distillation may lose some important information in the original training data, leading to more errors on predicting low-frequency words. To alleviate this problem, Ding et al. (2021b) proposed to augment NAT models the ability to learn lost knowledge from the original data. However, their approach relies on external resources (e.g. word alignment) and human-crafted priors, which limits the applicability of the method to a broader range of tasks and languages. Accordingly, we turn to directly expose the raw data into NAT by leveraging pretraining without intensive modification to model architectures (§2.2). Furthermore, we analyze bilingual links in the distilled data from two alignment directions (i.e. source-to-target and target-to-source). We found that KD makes low-frequency source words aligned with targets more deterministically but fails to align low-frequency words from target to source due to information loss. Inspired by this finding, we propose reverse KD to recall more alignments for low-frequency target words (§2.3). We then concatenate two kinds of distilled data to maintain advantages of deterministic knowledge and low-frequency information. To make the most of authentic and synthetic data, we combine three complementary approaches (i.e. raw pretraining,

[*] 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.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 3431–3441
August 1–6, 2021. ©2021 Association for Computational Linguistics

bidirectional distillation training and KD finetuning) as a new training strategy for further boosting NAT performance (§2.4).

We validated our approach on five translation benchmarks (WMT14 En-De, WMT16 Ro-En, WMT17 Zh-En, WAT17 Ja-En and WMT19 En-De) over two advanced architectures (Mask Predict, Ghazvininejad et al., 2019; Levenshtein Transformer, Gu et al., 2019). Experimental results show that the proposed method consistently improve translation performance over the standard NAT models across languages and advanced NAT architectures. Extensive analyses confirm that the performance improvement indeed comes from the better lexical translation accuracy especially on low-frequency tokens.

Contributions Our main contributions are:

- We show the effectiveness of rejuvenating low-frequency information by pretraining NAT models from raw data.
- We provide a quantitative analysis of bilingual links to demonstrate the necessity to improve low-frequency alignment by leveraging both KD and reverse KD.
- We introduce a simple and effective training recipe to accomplish this goal, which is robustly applicable to several model structures and language pairs.

2 Rejuvenating Low-Frequency Words

2.1 Preliminaries

Non-Autoregressive Translation Given a source sentence $\mathbf{x}$, an AT model generates each target word $\mathbf{y}_t$ conditioned on previously generated ones $\mathbf{y}_{<t}$, leading to high latency on the decoding stage. In contrast, NAT models break this *autoregressive factorization* by producing target words in parallel. Accordingly, the probability of generating $\mathbf{y}$ is computed as:

$$p(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^{T} p(\mathbf{y}_t|\mathbf{x};\theta) \tag{1}$$

where T is the length of the target sequence, and it is usually predicted by a separate conditional distribution. The parameters θ are trained to maximize the likelihood of a set of training examples according to $\mathcal{L}(\theta) = \arg\max_\theta \log p(\mathbf{y}|\mathbf{x};\theta)$. Typically, most NAT models are implemented upon the framework of Transformer (Vaswani et al., 2017).

Knowledge Distillation Gu et al. (2018) pointed out that NAT models suffer from the *multimodality problem*, where the conditional independence assumption prevents a model from properly capturing the highly multimodal distribution of target translations. Thus, the sequence-level knowledge distillation is introduced to reduce the modes of training data by replacing their original target-side samples with sentences generated by an AT teacher (Gu et al., 2018; Zhou et al., 2020; Ren et al., 2020). Formally, the original parallel data *Raw* and the distilled data $\overrightarrow{KD}$ can be defined as follows:

$$\text{Raw} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N} \tag{2}$$

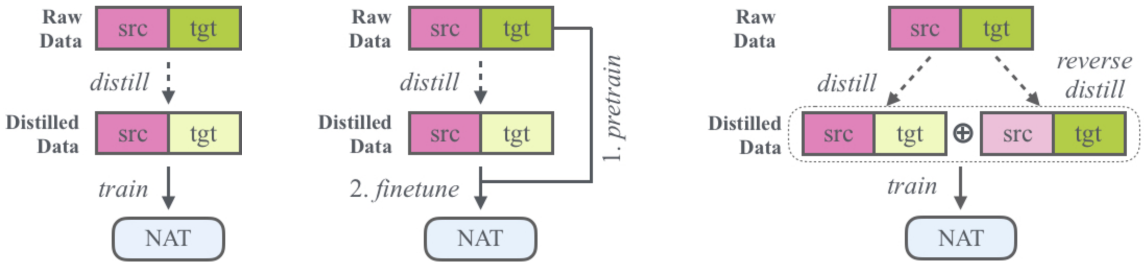
$$\overrightarrow{\text{KD}} = \{(\mathbf{x}_i, f_{s\mapsto t}(\mathbf{x}_i))|\mathbf{x}_i \in \text{Raw}_\text{s}\}_{i=1}^{N} \tag{3}$$

where $f_{s\mapsto t}$ represents an AT-based translation model trained on Raw data for translating text from the source to the target language. N is the total number of sentence pairs in training data. As shown in Figure 1 (a), well-performed NAT models are generally trained on $\overrightarrow{\text{KD}}$ data instead of Raw.

2.2 Pretraining with Raw Data

Motivation Gao et al. (2018) showed that more than 90% of words are lower than 10e-4 frequency in WMT14 En-De dataset. This *token imbalance problem* biases translation models towards overfitting to frequent observations while neglecting those low-frequency observations (Gong et al., 2018; Nguyen and Chiang, 2018; Gu et al., 2020). Thus, the AT teacher $f_{s\mapsto t}$ tends to generate more high-frequency tokens and less low-frequency tokens during constructing distilled data $\overrightarrow{\text{KD}}$.

On the one hand, KD can reduce the modes in training data (i.e. multiple lexical choices for a source word), which lowers the intrinsic uncertainty (Ott et al., 2018) and learning difficulty for NAT (Zhou et al., 2020; Ren et al., 2020), making it easily acquire more deterministic knowledge. On the other hand, KD aggravates the imbalance of high-frequency and low-frequency words in training data and lost some important information originated in raw data. Ding et al. (2021b) revealed the side effect of distilled training data, which cause lexical choice errors for low-frequency words in NAT models. Accordingly, they introduced an extra bilingual data-dependent prior objective to augments NAT models the ability to learn the lost knowledge from raw data. We use their findings as our departure point, but rejuvenate low-frequency



(a) Traditional Training (b) Raw Pretraining (c) Bidirectional Distillation Training

Figure 1: An illustration of different strategies for training NAT models. "distill" and "reverse distill" indicate sequence-level knowledge distillation with forward and backward AT teachers, respectively. The data block in transparent color means source- or target-side data are synthetically generated. Best view in color.

Data	s ↦ t LFW Links			t ↦ s LFW Links		
	R	P	F1	R	P	F1
Raw	66.4	81.9	73.3	72.3	80.6	76.2
$\overrightarrow{\textbf{KD}}$	**73.4**	**89.2**	**80.5**	69.9	79.1	74.2
$\overleftarrow{\textbf{KD}}$	61.2	79.4	69.1	**82.9**	**83.1**	**83.0**

Table 1: Evaluation on aligned links between source- and target-side low-frequency words (LFW). A directed line indicates aligning bilingual words from the source to the target side (s ↦ t) or in an opposite way (t ↦ s). R, P and F1 are recall, precision and F1-score.

Data	**Sentence**
Raw$_\text{S}$	**海克曼** 和 **奥德海姆** 提出 ... 模型
Raw$_\text{T}$	**Hackman** and **Oldham** propose ... model
$\overrightarrow{\textbf{KD}}_\text{T}$	*Heckman* and *Oddheim* propose ... model
$\overleftarrow{\textbf{KD}}_\text{S}$	*哈克曼* 和 *奥尔德姆* 提出 ... 模式

Table 2: An example in different kinds of data. "Raw" means the original data while "$\overrightarrow{\text{KD}}$" and "$\overleftarrow{\text{KD}}$" indicate syntactic data distilled by KD and reverse KD, respectively. The subscript "S" or "T" is short for source- or target-side. The low-frequency words are highlighted with colors and italics are incorrect translations.

words in a more simple and direct way: directly exposing raw data into NAT via pretraining.

Our Approach Many studies have shown that pretraining could transfer the knowledge and data distribution, especially for rare categories, hence improving the model robustness (Hendrycks et al., 2019; Mathis et al., 2021). Here we want to transfer the distribution of lost information, e.g. low-frequency words. As illustrated in Figure 1(b), we propose to first pretrain NAT models on Raw data and then continuously train them on $\overrightarrow{\text{KD}}$ data. The raw data maintain the original distribution especially on low-frequency words. Although it is difficult for NAT to learn high-mode data, the pretraining can acquire general knowledge from authentic data, which may help *better* and *faster* learning further tasks. Thus, we early stop pretraining when the model can achieve 90% of the best performance of raw data in terms of BLEU score (Platanios et al., 2019)[1]. In order to keep the merits of low-modes, we further train the pretrained model on distilled data $\overrightarrow{\text{KD}}$. As it is easy for NAT to learn deterministic knowledge, we finetune the model for the rest steps. For fair comparison, the total training steps of the proposed method are same as the traditional one. In general, we expect that this training recipe can provide a good trade-off between raw and distilled data (i.e. high-modes and complete vs. low-modes and incomplete).

2.3 Bidirectional Distillation Training

Analyzing Bilingual Links in Data KD simplifies the training data by replacing low-frequency target words with high-frequency ones (Zhou et al., 2020). This is able to facilitate easier aligning source words to target ones, resulting in high bilingual coverage (Jiao et al., 2020). Due to the information loss, we argue that KD makes low-frequency target words have fewer opportunities to align with source ones. To verify this, we propose a method to quantitatively analyze bilingual links from two directions, where low-frequency words

[1] In preliminary experiments, we tried another simple strategy: early-stop at fixed step according to the size of training data (e.g. training 70K En-De and early stop at 20K / 30K / 40K, respectively). We found that both strategies achieve similar performance.

are aligned from source to target (s ↦ t) or in an opposite direction (t ↦ s).

The method can be applied to different types of data. Here we take s ↦ t links in Raw data as an example to illustrate the algorithm. Given the WMT14 En-De parallel corpus, we employ an unsupervised word alignment method[2] (Och and Ney, 2003) to produce a word alignment, and then we extract aligned links whose source words are low-frequency (called s ↦ t LFW Links). Second, we randomly select a number of samples from the parallel corpus. For better comparison, the subset should contains the same i in Equation (2) as that of other type of datasets (e.g. i in Equation (3) for $\overrightarrow{\text{KD}}$). Finally, we calculate recall, precision, F1 scores based on low-frequency bilingual links for the subset. Recall (R) represents how many low-frequency source words can be aligned to targets. Precision (P) means how many aligned low-frequency links are correct according to human evaluation. F1 is the harmonic mean between precision and recall. Similarly, we can analyze t ↦ s LFW Links by considering low-frequency targets.

Table 1 shows the results on low-frequency links. Compared with Raw, $\overrightarrow{\text{KD}}$ can recall more s ↦ t LFW links (73.4 vs. 66.4) with more accurate alignment (89.2 vs. 73.3). This demonstrates the effectiveness of KD for NAT models from the bilingual alignment perspective. However, in the t ↦ s direction, there are fewer LFW links (69.9 vs. 72.3) with worse alignment quality (79.1 vs. 80.6) in $\overrightarrow{\text{KD}}$ than those in Raw. This confirms our claim that KD harms NAT models due to the loss of low-frequency target words. Inspired by these findings, it is natural to assume that *reverse KD* exhibits complementary properties. Accordingly, we conduct the same analysis method on $\overleftarrow{\text{KD}}$ data, and found better t ↦ s links but worse s ↦ t links compared with Raw. Take the Zh-En sentence pair in Table 2 for example, $\overrightarrow{\text{KD}}$ retains the source side low-frequency Chinese words "海克曼" (Raw_S) but generates the high-frequency English words "Heckman" instead of the golden "Hackman" ($\overrightarrow{\text{KD}}_\text{T}$). On the other hand, $\overleftarrow{\text{KD}}$ preserves the low-frequency English words "Hackman" (Raw_T) but produces the high-frequency Chinese words "哈克曼" ($\overleftarrow{\text{KD}}_\text{S}$).

Our Approach Based on analysis results, we propose to train NAT models on bidirectional distillation by concatenating two kinds of distilled data. The reverse distillation is to replace the source sentences in the original training data with synthetic ones generated by a backward AT teacher.[3] According to Equation 3, $\overleftarrow{\text{KD}}$ can be formulated as:

$$\overleftarrow{\text{KD}} = \{(\mathbf{y}_i, f_{t\mapsto s}(\mathbf{y}_i))|\mathbf{y}_i \in \text{Raw}_\text{t}\}_{i=1}^{N} \qquad (4)$$

where $f_{t\mapsto s}$ represents an AT-based translation model trained on Raw data for translating text from the target to the source language.

Figure 1(c) illustrates the training strategy. First, we employ both $f_{s\mapsto t}$ and $f_{t\mapsto s}$ AT models to generate $\overrightarrow{\text{KD}}$ and $\overleftarrow{\text{KD}}$ data, respectively. Considering complementarity of two distilled data, we combine $\overrightarrow{\text{KD}}$ and $\overleftarrow{\text{KD}}$ as a new training data for training NAT models. We expect that 1) distilled data can maintain advantages of low-modes; 2) bidirectinoal distillation can recall more LFW links on two directions with better alignment quality, leading to the overall improvements. Besides, Nguyen et al. (2020) claimed that combining different distilled data (generated by various models trained with different seeds) improves data diversification for NMT, and we leave this for future work.

2.4 Combining Both of Them: Low-Frequency Rejuvenation (LFR)

We have proposed two parallel approaches to rejuvenate low-frequency knowledge from authentic (§2.2) and synthetic (§2.3) data, respectively. Intuitively, we combine both of them to further improve the model performance.

From data view, two presented training strategies are: Raw → $\overrightarrow{\text{KD}}$ (Raw Pretraining) and $\overrightarrow{\text{KD}} + \overleftarrow{\text{KD}}$ (Bidirectional Distillation Training). Considering the effectiveness of pretraining (Mathis et al., 2021) and clean finetuning (Wu et al., 2019), we introduce a combined pipeline: Raw $\rightarrow \overrightarrow{\text{KD}} + \overleftarrow{\text{KD}} \rightarrow \overrightarrow{\text{KD}}$ as out best training strategy. There are many possible ways to implement the general idea of combining two approaches. The aim of this paper is not to explore the whole space but simply to show that one fairly straightforward implementation works well and the idea is reasonable. Nonetheless, we compare possible strategies of combination two approaches as well as demonstrate their complementarity in §3.3. While in main experiments (in §3.2), we valid the combination strategy, namely *Low-Frequency Rejuvenation* (LFR).

[2] The FastAlign (Dyer et al., 2013) was employed to build word alignments for the training datasets.

[3] This is different from back-translation (Edunov et al., 2018), which is an alternative to leverage monolingual data.

Model	Iteration	Speed	En-De		Ro-En	
			BLEU	ALF	BLEU	ALF
AT Models						
Transformer-Base (Ro-En Teacher)	n/a	1.0×	27.3	70.5	34.1	73.6
Transformer-Big (En-De Teacher)	n/a	0.8×	29.2	73.0	n/a	n/a
Existing NAT Models						
NAT (Gu et al., 2018)	1.0	2.4×	19.2	n/a	31.4	n/a
Iterative NAT (Lee et al., 2018)	10.0	2.0×	21.6		30.2	
DisCo (Kasai et al., 2020)	4.8	3.2×	26.8		33.3	
Mask-Predict (Ghazvininejad et al., 2019)	10.0	1.5×	27.0		33.3	
Levenshtein (Gu et al., 2019)	2.5	3.5×	27.3		33.3	
Our NAT Models						
Mask-Predict (Ghazvininejad et al., 2019)	10.0	1.5×	27.0	68.4	33.3	70.9
+Low-Frequency Rejuvenation			**27.8**†	72.3	**33.9**†	72.4
Levenshtein (Gu et al., 2019)	2.5	3.5×	27.4	69.2	33.2	71.1
+Low-Frequency Rejuvenation			**28.2**†	72.8	**33.8**†	72.7

Table 3: Comparison with previous work on WMT14 En-De and WMT16 Ro-En. "Iteration" indicates the number of iterative refinement while "Speed" shows the speed-up ratio of decoding. "ALF" is the translation accuracy on low-frequency words. "†" indicates statistically significant difference ($p < 0.05$) from corresponding baselines.

3 Experiment

3.1 Setup

Data Main experiments are conducted on four widely-used translation datasets: WMT14 English-German (En-De, Vaswani et al. 2017), WMT16 Romanian-English (Ro-En, Gu et al. 2018), WMT17 Chinese-English (Zh-En, Hassan et al. 2018), and WAT17 Japanese-English (Ja-En, Morishita et al. 2017), which consist of 4.5M, 0.6M, 20M, and 2M sentence pairs, respectively. We use the same validation and test datasets with previous works for fair comparison. To prove the universality of our approach, we further experiment on different data volumes, which are sampled from WMT19 En-De.[4] The *Small* and *Medium* corpora respectively consist of 1.0M and 4.5M sentence pairs, and *Large* one is the whole dataset which contains 36M sentence pairs. We preprocess all data via BPE (Sennrich et al., 2016) with 32K merge operations. We use tokenized BLEU (Papineni et al., 2002) as the evaluation metric, and *sign-test* (Collins et al., 2005) for statistical significance test. The translation accuracy of low-frequency words is measured by AoLC (Ding et al., 2021b), where word alignments are established based on the widely-used automatic alignment tool GIZA++ (Och and Ney, 2003).

Models We validated our research hypotheses on two state-of-the-art NAT models:

- *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. We followed its optimal settings to keep the iteration number as 10 and length beam as 5.

- *Levenshtein Transformer* (LevT, Gu et al. 2019) that introduces three steps: deletion, placeholder and token prediction. The decoding iterations adaptively depends on certain conditions.

We closely followed previous works to apply sequence-level knowledge distillation to NAT (Kim and Rush, 2016). Specifically, we train both Base and Big Transformer as the *AT teachers*. For Big model, we adopt large batch strategy (i.e. 458K tokens/batch) to optimize the performance. Most NAT tasks employ Transformer-Big as their strong teacher except for Ro-En and *Small* En-De, which are distilled by Transformer-Base.

Training Traditionally, NAT models are usually trained for 300K steps on regular batch size (i.e.

[4] `http://www.statmt.org/wmt19/translation-task.html`

Model	Zh-En		Ja-En	
	BLEU	ALF	BLEU	ALF
AT	25.3	66.2	29.8	70.8
MaskT	24.2	61.5	28.9	66.9
+LFR	25.1†	64.8	29.6†	68.9
LevT	24.4	62.7	29.1	66.8
+LFR	25.1†	65.3	29.7	69.2

Table 4: Performance on other language pairs, including WMT17 Zh-En and WAT17 Ja-En. "$\dagger$" indicates statistically significant difference ($p < 0.05$) from corresponding baselines.

Model	Law	Med.	IT	Kor.	Sub.
AT	41.5	30.8	27.5	8.6	15.4
MaskT	37.3	28.2	24.6	7.3	11.2
+LFR	38.1†	28.8	25.4†	8.9†	14.3†
LevT	37.5	28.4	24.7	7.5	12.4
+LFR	38.5†	29.4†	25.9†	8.4†	14.5†

Table 5: Performance on domain shift setting. Models are trained on WMT14 En-De news domain but evaluated on out-of-domain test sets, including law, medicine, IT, koran and subtitle. "$\dagger$" indicates statistically significant difference ($p < 0.05$) from corresponding baselines.

128K tokens/batch). In this work, we empirically adopt large batch strategy (i.e. 480K tokens/batch) to reduce the training steps for NAT (i.e. 70K). Accordingly, the learning rate warms up to 1×10^{-7} for 10K steps, and then decays for 60k steps with the cosine schedule (Ro-En models only need 4K and 21K, respectively). For regularization, we tune the dropout rate from [0.1, 0.2, 0.3] based on validation performance in each direction, and apply weight decay with 0.01 and label smoothing with ϵ = 0.1. We use Adam optimizer (Kingma and Ba, 2015) to train our models. We followed the common practices (Ghazvininejad et al., 2019; Kasai et al., 2020) to evaluate the performance on an ensemble of top 5 checkpoints to avoid stochasticity.

Note that the total training steps of the proposed approach (in §2.2~2.4) are identical with those of the standard training (in §2.1). Taking the best training strategy (Raw $\rightarrow \overrightarrow{\text{KD}} + \overleftarrow{\text{KD}} \rightarrow \overrightarrow{\text{KD}}$) for example, we empirically set the training step for each stage is 20K, 20K and 30K, respectively. And Ro-En models respectively need 8K, 8K and 9K steps in corresponding training stage.

3.2 Results

Comparison with Previous Work Table 3 lists the results of previous competitive NAT models (Gu et al., 2018; Lee et al., 2018; Kasai et al., 2020; Gu et al., 2019; Ghazvininejad et al., 2019) on the WMT16 Ro-En and WMT14 En-De benchmark. We implemented our approach on top of two advanced NAT models (i.e. Mask-Predict and Levenshtein Transformer). Compared with standard NAT models, our training strategy significantly and consistently improves translation performance (BLEU↑) across different language pairs and NAT models. Besides, the improvements on translation performance are mainly due to a increase of translation accuracy on low-frequency words (ALF↑), which reconfirms our claims. For instance, our method significantly improves the standard Mask-Predict model by +0.8 BLEU score with a substantial +3.6 increase in ALF score. Encouragingly, our approach push the existing NAT models to achieve new SOTA performances (i.e. 28.2 and 33.9 BLEU on En-De and Ro-En, respectively).

It is worth noting that our data-level approaches neither modify model architecture nor add extra training loss, thus do not increase any latency ("Speed"), maintaining the intrinsic advantages of non-autoregressive generation. We must admit that our strategy indeed increase the amount of computing resources due to that we should train $f_{t \mapsto s}$ AT teachers for building $\overleftarrow{\text{KD}}$ data.

Results on Other Language Pairs Table 4 lists the results of NAT models on Zh-En and Ja-En language pairs, which belong to different language families (i.e. Indo-European, Sino-Tibetan and Japonic). Compared with baselines, our method significantly and incrementally improves the translation quality in all cases. For Zh-En, LFR achieves on average +0.8 BLEU improvement over the traditional training, along with increasing on average +3.0% accuracy on low-frequency word translation. For long-distance language pair Ja-En, our method still improves the NAT model by on average +0.7 BLEU point with on average +2.2 ALF. Furthermore, NAT models with the proposed training strategy perform closely to their AT teachers (i.e. 0.2 ΔBLEU). This shows the effectiveness and universality of our method across language pairs.

Model	BLEU		
	1.0M	4.5M	36.0M
AT	25.5	37.6	40.2
MaskT	23.7	35.4	36.8
+LFR	24.3†	36.2†	37.7†

Table 6: Performance on different scale of training data. The small and medium datasets are sampled from the large WMT19 En-De dataset, and evaluations are conducted on the same testset. "†" indicates statistically significant difference ($p < 0.05$) from corresponding baselines.

Results on Domain Shift Scenario The lexical choice must be informed by linguistic knowledge of how the translation model's input data maps onto words in the target domain. Since low-frequency words get lost in traditional NAT models, the problem of lexical choice is more severe under domain shift scenario (i.e. models are trained on one domain but tested on other domains). Thus, we conduct evaluation on WMT14 En-De models over five out-of-domain test sets (Müller et al., 2020), including law, medicine, IT, Koran and movie subtitle domains. As shown in Table 5, standard NAT models suffer large performance drops in terms of BLEU score (i.e. on average -2.9 BLEU over AT model). By observing these outputs, we found a large amount of translation errors on low-frequency words, most of which are domain-specific terminologies. In contrast, our approach improves translation quality (i.e. on average -1.4 BLEU over AT model) by rejuvenating low-frequency words to a certain extent, showing that LFR increases the domain robustness of NAT models.

Results on Different Data Scales To confirm the effectiveness of our method across different data sizes, we further experiment on three En-De datasets at different scale. The small- and medium-scale training data are randomly sampled from WM19 En-De corpus, containing about 1.0M and 4.5M sentence pairs, respectively. The large-scale one is collected from WMT19, which consists of 36M sentence pairs. We report the BLEU scores on same testset `newstest2019` for fair comparison. We employs base model to train the small-scale AT teacher, and big model with large batch strategy (i.e. 458K tokens/batch) to build the AT teachers for medium- and large-scale. As seen in Table 6, our simple training recipe boost performances for NAT models across different size of datasets, especially on large scale (+0.9), showing the robustness and effectiveness of our approach.

Model	BLEU	ALF
Mask-Predict	27.0	68.4
+Raw Data Prior	27.8	72.4
+Low-Frequency	27.8	72.3
+Combination	28.1	72.9

Table 7: Complementary to other work. "Combination" indicates combining "+Raw Data Prior" proposed by Ding et al. (2021b) with our "+Low-Frequency". Experiments are conducted on WMT14 En-De.

Complementary to Related Work Ding et al. (2021b) is relevant to our work, which introduced an extra bilingual data-dependent prior objective to augment NAT models the ability to learn low-frequency words in raw data. Our method is complementary to theirs due to that we only change data and training strategies (model-agnostic). As shown in Table 7, two approaches yield comparable performance in terms of BLEU and ALF. Besides, combination can further improve BLEU as well as ALF scores (i.e. +0.3 and +0.6). This illustrates the complementarity of model-level and data-level approaches on rejuvenating low-frequency knowldege for NAT models.

3.3 Analysis

We conducted extensive analyses to better understand our approach. All results are reported on the Mask-Predict models.

Accuracy of Lexical Choice To understand where the performance gains come from, we conduct fine-grained analysis on lexical choice. We divide "All" tokens into three categories based on their frequency, including "High", "Medium" and "Low". Following Ding et al. (2021b), we measure the accuracy of lexical choice on different frequency of words. Table 8 shows the results. **Takeaway:** *The majority of improvements on translation accuracy is from the low-frequency words, confirming our hypothesis.*

Low-Frequency Words in Output We expect to recall more low-frequency words in translation output. As shown in Table 9, we calculate the ratio of low-frequency words in generated sentences. As seen, KD biases the NAT model towards gen-

Model	En-De				Zh-En				Ja-En			
	All	High	Med.	Low	All	High	Med.	Low	All	High	Med.	Low
MaskT (Raw)	74.3	75.9	74.6	72.5	68.5	71.5	68.3	65.1	73.1	75.5	74.7	69.1
MaskT (KD)	76.3	82.4	78.3	68.4	72.7	81.4	75.2	61.5	75.3	82.8	76.3	66.9
+Raw-Pretrain	77.7	83.1	78.4	71.9	73.4	81.6	75.3	64.1	76.1	83.4	76.7	68.3
+Bi-Distillation	77.9	83.1	78.5	72.3	73.7	81.7	75.3	64.8	76.5	83.5	76.7	68.9

Table 8: Analysis on different frequency words in terms of accuracy of lexical choice. We split "All" words into "High", "Medium" and "Low" categories. Shades of cell color represent differences between ours and KD.

Model	En-De	Zh-En	Ja-En
MaskT (Raw)	10.3%	6.7%	9.4%
MaskT (KD)	7.6%	4.2%	6.9%
+Raw-Pretrain	9.3%	5.6%	8.4%
+Bi-Distillation	**9.7%**	**6.8%**	**8.7%**

Table 9: Ratio of low-frequency target words in output.

#	Strategy	BLEU	ALF
1	Raw	24.1	69.3
2	$\overrightarrow{\text{KD}}$	25.4	66.4
3	Raw+$\overrightarrow{\text{KD}}$	25.6	67.7
4	Raw→$\overrightarrow{\text{KD}}$	**25.9**	68.2
5	Raw+$\overleftarrow{\text{KD}}$+$\overrightarrow{\text{KD}}$	25.7	67.9
6	Raw→$\overleftarrow{\text{KD}}$+$\overrightarrow{\text{KD}}$	25.7	68.3
7	Raw→$\overleftarrow{\text{KD}}$+$\overrightarrow{\text{KD}}$→$\overrightarrow{\text{KD}}$	**26.3**	69.5

Table 10: Performances of different strategies. The models are trained and tested on WMT14 En-De. "A+B" means concatenate A and B while "A→B" indicates pretraining on A and then finetuning on B.

erating high-frequency tokens (*Low freq.*↓) while our method can not only correct this bias (on average +18% and +26% relative changes for *+raw-pretrain* and *+Bi-distillation*), but also enhance translation (BLEU↑ in Table 4). **Takeaway:** *Our method generates translations that contain more low-frequency words.*

Effects of Variant Training Strategies As discussed in §2.4, we carefully investigate alternative training approaches in Table 10. We make the total training step identical to that of vanilla NAT models, and report both BLEU and ALF scores. As seen, all variant strategies perform better than the standard KD method in terms both BLEU and ALF scores, confirming the necessity of our work. **Takeaway:** *1) Pretraining is more effective than combination on utilizing data manipulation strategies; 2) raw data and bidirectional distilled data are complementary to each other; 3) it is indispensable to finetune models on* $\overrightarrow{\textit{KD}}$ *in the last stage.*

Model	All	High	Med.	Low
Training on Raw Data				
AT-Teacher	79.3	84.7	80.2	73.0
AT-Student	76.8	80.2	77.4	72.8
Training on Distilled Data				
AT-Student	77.3	82.5	78.6	70.9
+LFT	78.1	83.2	78.7	72.5

Table 11: Analysis on AT models in term of the accuracy of lexical choice on WMT14 En-De. We split "All" words into "High", "Medium" and "Low" categories.

Our Approach Works for AT Models Although our work is designed for NAT models, we also investigated whether our LFT method works for general cases, e.g. autoregressive models. We used Transformer-BIG as the teacher model. For fair comparison, we leverage the Transformer-BASE as the student model, which shares the same model capacity with NAT student (i.e. MaskT). The result lists in Table 11. As seen, AT models also suffer from the problem of low-frequency words when using knowledge distillation, and our approach also works for them. **Takeaway:** *Our method works well for general cases through rejuvenating more low-frequency words.*

4 Related Work

Low-Frequency Words Benefiting from continuous representation learned from the training data, NMT models have shown the promising performance. However, Koehn and Knowles (2017) point that low-frequency words translation is still one of the key challenges for NMT according to the Zipf's law (Zipf, 1949). For AT models, Arthur et al. (2016) address this problem by integrating a count-based lexicon, and Nguyen and Chiang (2018) propose an additional lexical model, which is jointly trained with the AT model. Recently, Gu et al. (2020) adaptively re-weight the rare words during training. The lexical choice problem is more serious for NAT models, since 1) the lexical choice errors (low-resource words in particular) of AT distillation will propagate to NAT models; and 2) NAT lacks target-side dependencies thus misses necessary target-side context. In this work, we alleviate this problem by solving the first challenge.

Data Manipulation Our work is related to previous studies on manipulating training data for NMT. Bogoychev and Sennrich (2019) show that forward- and backward-translations (FT/ BT) could both boost the model performances, where FT plays the role of domain adaptation and BT makes the translation fluent. Fadaee and Monz (2018) sample the monolingual data with more difficult words (e.g. rare words) to perform BT, achieving significant improvements compared with randomly sampled BT. Nguyen et al. (2020) diversify the data by applying FT and BT multiply times. However, different from AT, the prerequisite of training a well-performed NAT model is to perform KD. We compared with related works in Table 10 and found that our approach consistently outperforms them. Note that all the ablation studies focus on exploiting the parallel data without augmenting additional data.

Non-Autoregressive Translation A variety of approaches have been exploited to bridge the performance gap between NAT and AT models. Some researchers proposed new model architectures (Lee et al., 2018; Ghazvininejad et al., 2019; Gu et al., 2019; Kasai et al., 2020), aided with additional signals (Wang et al., 2019; Ran et al., 2019; Ding et al., 2020), introduced sequential information (Wei et al., 2019; Shao et al., 2019; Guo et al., 2020; Hao et al., 2021), and explored advanced training objectives (Ghazvininejad et al., 2020; Du et al., 2021). Our work is close to the research line on training methods. Ding et al. (2021b) revealed the low-frequency word problem in distilled training data, and introduced an extra Kullback-Leibler divergence term derived by comparing the lexical choice of NAT model and that embedded in the raw data. Ding et al. (2021a) propose a simple and effective training strategy, which progressively feeds different granularity of data into NAT models by leveraging curriculum learning.

5 Conclusion

In this study, we propose simple and effective training strategies to rejuvenate the low-frequency information in the raw data. Experiments show that our approach consistently and significantly improves translation performance across language pairs and model architectures. Notably, domain shift is an extreme scenario to diagnose low-frequency translation, and our method significant improves them. Extensive analyses reveal that our method improves the accuracy of lexical choices for low-frequency source words, recalling more low-frequency words in translations as well, which confirms our claim.

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

Philip Arthur, Graham Neubig, and Satoshi Nakamura. 2016. Incorporating discrete translation lexicons into neural machine translation. In *EMNLP*.

Nikolay Bogoychev and Rico Sennrich. 2019. Domain, translationese and noise in synthetic data for neural machine translation. *ArXiv*.

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. Progressive multi-granularity training for 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, Di Wu, Dacheng Tao, and Zhaopeng Tu. 2020. 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*.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *NAACL*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *EMNLP*.

Marzieh Fadaee and Christof Monz. 2018. Back-translation sampling by targeting difficult words in neural machine translation. In *EMNLP*.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. 2018. Representation degeneration problem in training natural language generation models. In *ICLR*.

Marjan Ghazvininejad, Vladimir Karpukhin, Luke Zettlemoyer, and Omer Levy. 2020. Aligned cross entropy for non-autoregressive machine translation. In *ICML*.

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

Chengyue Gong, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2018. Frage: Frequency-agnostic word representation. *NeurIPS*.

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*.

Shuhao Gu, Jinchao Zhang, Fandong Meng, Yang Feng, Wanying Xie, Jie Zhou, and Dong Yu. 2020. Token-level adaptive training for neural machine translation. In *EMNLP*.

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

Yongchang Hao, Shilin He, Wenxiang Jiao, Zhaopeng Tu, Michael Lyu, and Xing Wang. 2021. Multi-task learning with shared encoder for non-autoregressive machine translation. In *NAACL*.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving human parity on automatic chinese to english news translation. *arXiv*.

Dan Hendrycks, Kimin Lee, and Mantas Mazeika. 2019. Using pre-training can improve model robustness and uncertainty. In *ICML*.

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

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

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 Rebecca Knowles. 2017. Six challenges for neural machine translation. In *WMT*.

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

Alexander Mathis, Thomas Biasi, Steffen Schneider, Mert Yuksekgonul, Byron Rogers, Matthias Bethge, and Mackenzie W Mathis. 2021. Pretraining boosts out-of-domain robustness for pose estimation. In *WACV*.

Makoto Morishita, Jun Suzuki, and Masaaki Nagata. 2017. Ntt neural machine translation systems at wat 2017. In *IJCNLP*.

Mathias Müller, Annette Rios, and Rico Sennrich. 2020. Domain Robustness in Neural Machine Translation. In *AMTA*.

Toan Nguyen and David Chiang. 2018. Improving lexical choice in neural machine translation. In *NAACL*.

Xuan-Phi Nguyen, Joty Shafiq, Kui Wu, and Ai Ti Aw. 2020. Data diversification: A simple strategy for neural machine translation. In *NeurIPS*.

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

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

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. 2019. Guiding non-autoregressive neural machine translation decoding with reordering information. *arXiv*.

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*.

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

Chenze Shao, Jinchao Zhang, Yang Feng, Fandong Meng, and Jie Zhou. 2019. Minimizing the bag-of-ngrams difference for non-autoregressive neural machine translation. In *AAAI*.

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*.

Bingzhen Wei, Mingxuan Wang, Hao Zhou, Junyang Lin, and Xu Sun. 2019. Imitation learning for non-autoregressive neural machine translation. In *ACL*.

Lijun Wu, Yiren Wang, Yingce Xia, QIN Tao, Jianhuang Lai, and Tie-Yan Liu. 2019. Exploiting monolingual data at scale for neural machine translation. In *EMNLP*.

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

George K. Zipf. 1949. Human behavior and the principle of least effort.