Redistributing Low-Frequency Words: Making the Most of Monolingual Data in Non-Autoregressive Translation

Anonymous ACL submission

Abstract

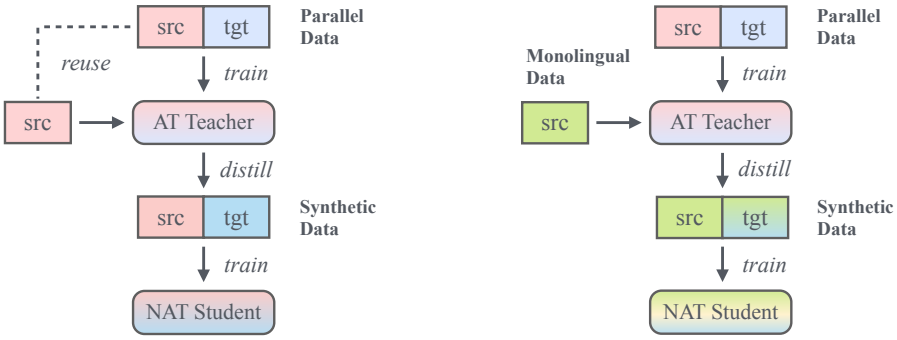
Knowledge distillation (KD) is the preliminary step for training non-autoregressive translation (NAT) models, which eases the training of NAT models at the cost of losing important information for translating low-frequency words. In this work, we provide an appealing alternative for NAT – *monolingual KD*, which trains NAT student on external monolingual data with AT teacher trained on the original bilingual data. Monolingual KD is able to transfer both the knowledge of the original bilingual data (implicitly encoded in the trained AT teacher model) and that of the new monolingual data to the NAT student model. Extensive experiments on eight WMT benchmarks over two advanced NAT models show that monolingual KD consistently outperforms the standard KD by improving low-frequency word translation, without introducing any computational cost. Monolingual KD enjoys desirable expandability, which can be further enhanced (when given more computational budget) by combining with the standard KD, a reverse monolingual KD, or enlarging the scale of monolingual data. Extensive analyses demonstrate that these techniques can be used together profitably to further recall the useful information lost in the standard KD. Encouragingly, combining with standard KD, our approach achieves 30.4 and 34.1 BLEU points on the WMT14 English-German and German-English datasets, respectively. Code, data, and models will be released.

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. However, the *independence assumption* prevents a model from properly capturing the highly multimodal distribution of target translations. In response to this problem, a sequence-level knowledge distillation (KD, Kim and Rush 2016) becomes the preliminary step for training NAT models, which produces more deterministic knowledge by reducing the translation modes of the bilingual data (Zhou et al., 2020).

Although the standard KD on original bilingual data eases the training of NAT models, distillation may lose some important information in the raw training data, leading to *more errors on predicting low-frequency words* (Ding et al., 2021b,a). To remedy this problem, Ding et al. (2021b) augmented NAT models the ability to learn lost knowledge from the raw bilingual data with an additional objective, and Ding et al. (2021a) first pre-trained NAT models on the raw training data and then fine-tuned them on the distilled training data. While previous studies mainly focus on recalling the lost information during the distillation of the original *bilingual* data, in this work we propose to improve the prediction of low-frequency words by redistributing them in the external *monolingual* data, which has the great potential to complement the original bilingual data on the word distribution.

Specifically, we leverage the monolingual data to perform KD (*monolingual KD*, §2.2), and train the NAT student model on the distilled monolingual data (Figure 1b). Monolingual KD provides appealing benefits. Firstly, the monolingual data and bilingual data in machine translation are generally complementary to each other (Zhang and Zong, 2016; Wu et al., 2019; Zhou and Keung, 2020; Siddhant et al., 2020; Jiao et al., 2021). Accordingly, monolingual KD is able to transfer both the knowledge of the bilingual data (implicitly encoded in the trained teacher model) and that of the monolingual data to the NAT student, without introducing additional computational cost. Secondly, the amount of available monolingual data is several orders of magnitude larger than that of bilingual data, which offers monolingual KD the potential to further improve translation performance by exploiting more monolingual data.



(a) Standard KD (b) Monolingual KD

Figure 1: Illustration of (a) standard bilingual data KD and (b) the proposed monolingual KD. The main difference between the two KDs lies in constructing the distilled data by (a) reusing the source side of bilingual data, or (b) introducing a new monolingual data.

Furthermore, we analyze the bilingual links in the bilingual and monolingual distilled data from two alignment directions (i.e. source-to-target and target-to-source). We found that the monolingual KD makes low-frequency source words aligned with targets more deterministically compared to bilingual KD, but both of them fail to align low-frequency words from target to source due to information loss. Starting from this finding, we propose reverse monolingual KD to recall more alignments for low-frequency target words. We then concatenate two kinds of monolingual distilled data (*bidirectional monolingual KD*, §2.3) to maintain advantages of deterministic knowledge and low-frequency information.

We validated our approach on several translation benchmarks across scales (WMT14 En↔De, WMT16 Ro↔En, WMT17 Zh↔En, and WMT19 En↔De) over two advanced NAT models: Mask Predict (Ghazvininejad et al., 2019) and Levenshtein (Gu et al., 2019). Experiments demonstrate the effectiveness and universality of our approach. Specifically, we have the following findings:

- Monolingual KD achieves better performance than the standard KD in all cases, and the proposed bidirectional monolingual KD can further improve performance by a large margin.
- Monolingual KD enjoys appealing expandability: enlarging the scale of monolingual data consistently improves performance until reaching the bottleneck of model capacity.
- Monolingual KD is complementary to the standard KD, and combining them obtains further improvement by alleviating two key issues of NAT, i.e., the multimodality problem and the low-frequency word translation problem.

The paper is an early step in exploring monolingual KD for NAT, which can narrow the performance gap between NAT models and the SOTA AT models. We hope the promising effect of monolingual KD on NAT can draw more interest and can make NAT a common translation framework.

2 Redistributing Low-Frequency Words

2.1 Preliminaries

Non-Autoregressive Translation Recent years have seen a surge of interest in NAT (Gu et al., 2018), which can improve the decoding efficiency by predicting all tokens independently and simultaneously. Specifically, the probability of generating a target sentence $\mathbf{y}$ by given the source sentence $\mathbf{x}$ is computed as $p(\mathbf{y}|\mathbf{x}) = p_L(T|\mathbf{x};\theta)\prod_{t=1}^{T} p(\mathbf{y}_t|\mathbf{x};\theta)$, where T is the length of $\mathbf{y}$, which is predicted by a separate conditional distribution $p_L(\cdot)$. The parameters θ are trained to maximize the likelihood of a set of training examples according to $\mathcal{L}(\theta) = \arg\max_\theta \log p(\mathbf{y}|\mathbf{x};\theta)$. The conditional independence assumption prevents an NAT model from properly capturing the highly multimodal distribution of target translations (*multimodality problem*, Gu et al., 2018). As a result, the translation quality of NAT models often lags behind that of AT models (Vaswani et al., 2017).

Standard Knowledge Distillation Knowledge distillation is the preliminary step for training NAT models by reducing the modes in the original bilingual data, which makes NAT easily acquire more deterministic knowledge and achieve significant

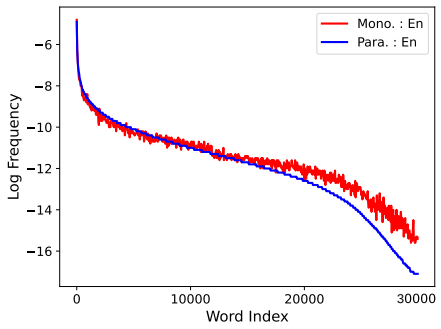


Figure 2: Word distributions of bilingual and monolingual English data on WMT14 En-De training data. Index is ranked by its frequency in bilingual data, where we normalize the frequency and report in log-scale.

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
$\overrightarrow{\text{KD}}_{\text{B}}$	73.4	89.2	80.5	69.9	79.1	74.2
$\overrightarrow{\text{KD}}_{\text{M}}$	75.1	87.7	80.9	70.8	81.4	75.7
$\overleftarrow{\text{KD}}_{\text{M}}$	63.7	80.2	71.0	**81.4**	**86.2**	**83.7**
$\overleftrightarrow{\text{KD}}_{\text{M}}$	**75.7**	**89.6**	**82.1**	80.5	79.4	79.9

Table 1: Evaluation of aligned links between source- and target-side low-frequency words on WMT14 En-De training data. "$\overrightarrow{\text{KD}}$" denotes the standard KD on source-language data, and "$\overleftarrow{\text{KD}}$" denotes reverse KD on target-language data. The subscripts $_{\text{B}}$ and $_{\text{M}}$ represent Bilingual and Monolingual distilled data.

improvement (Zhou et al., 2020). Typically, a sequence-level KD (Kim and Rush, 2016) is employed for NAT training, as shown in Figure 1a.

2.2 Monolingual Knowledge Distillation

Different Distributions of Source Words To empirically reveal the difference on word distribution between bilingual and monolingual data, we visualize the overall word distributions, as plotted in Figure 2. We can observe the significant difference between bilingual and monolingual data in the low-frequency part, which indicates that the words that occur less in the bilingual data are not necessarily low-frequent in the external monolingual data. Starting from the observation, we propose to exploit external monolingual data to offer more useful information for predicting low-frequent words in bilingual data, which are generally lost in the standard knowledge distillation.

Our Approach Researches and competitions have shown that fully exploiting the monolingual data is at the core of achieving better generalization and accuracy for MT systems (Sennrich et al., 2016a; Zhang and Zong, 2016; Barrault et al., 2020). In this work we want to transfer the distribution of lost information, e.g. low-frequency words, from monolingual data to the NAT training. Figure 1b shows the pipeline of our proposed *Monolingual KD* for NAT, which differs from the *Standard KD* at how to construct the distilled data. Instead of reusing the source side of the original bilingual data, monolingual KD performs distillation on newly monolingual data, which eliminates the dependency on the original training data.

Intuitively, the monolingual KD can embed both the knowledge of the original bilingual data (implicitly encoded in the trained teacher model) and that of the newly introduced monolingual data. The comprehensive experiments in the following section provide empirical support for our hypothesis. In addition, the complementarity between the bilingual and monolingual data makes explicitly combining *Standard KD* and *Monlingual KD* can further improve model performance.

2.3 Bidirectional Monolingual KD

Recalling Low-Frequency Target Words KD simplifies the training data by replacing low-frequency target words with high-frequency ones (Zhou et al., 2020; Ding et al., 2021b). This is able to facilitate easier aligning source words to target ones, resulting in high bilingual coverage (Jiao et al., 2020). Inspired by the low-frequency word (LFW) links analysis (Ding et al., 2021a), we borrow this LFW analysis to show the necessity of leveraging both the source- and target-side monolingual data. Concretely, we follow (Ding et al., 2021a) to evaluate the links of low-frequency words aligning from source to target (s ↦ t) with three metrics: 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 in an opposite direction (t ↦ s) by considering the links of low-frequency target words.

Table 1 lists the results. Comparing with the standard $\overrightarrow{\text{KD}}_{\text{B}}$, the forward monolingual KD ($\overrightarrow{\text{KD}}_{\text{M}}$ in Section 2.2) achieves better alignment quality

of s $\mapsto$ t LFW links (F1: 80.9 vs. 80.5) by aligning more low-frequency source words (R: 75.1 vs. 73.4). The backward monolingual KD ($\overleftarrow{\text{KD}}_{\text{M}}$) can complementarily produce better alignment of low-frequency target words (t $\mapsto$ s LFW links). As we expected, combining the two types of distilled data ($\overleftrightarrow{\text{KD}}_{\text{M}}$) can produce better alignments for both low-frequency source (F1: 82.1 vs. 80.5) and target words (F1: 79.9 vs. 74.2).

Our Approach (*Bid. Monolingual KD*) Based on the above observations, we propose to train NAT models on bidirectional monolingual data by concatenating two kinds of distilled data. Like back-translation (Edunov et al., 2018), the reverse monolingual distillation $\overleftarrow{\text{KD}}_{\text{M}}$ is to synthesize the source sentences by a backward AT teacher, which is trained in the reverse direction of the original bilingual data. The mixture of the source-original and target-original synthetic datasets (i.e. $\overleftrightarrow{\text{KD}}_{\text{M}}$) is used to train the final NAT model. We expect that the better alignments of LFW links can lead to overall improvement of translation performance.

3 Experiments

3.1 Experimental Setup

Bilingual Data We conducted experiments on two widely-used NAT benchmarks: WMT14 English-German and WMT16 English-Romanian tasks, which consist of 4.5M and 0.6M sentence pairs respectively. To prove the universality of our approach on large-scale data, we also validated on WMT17 English-Chinese and WMT19 English-German tasks, which consist of 20.6M and 36.8M sentence pairs respectively. We shared the source and target vocabularies, except for En↔Zh data. We split the training data into subword units using byte pair encoding (BPE) (Sennrich et al., 2016c) with 32K merge operations, forming a vocabulary of 37k, 32k, 33k/48k and 44k for WMT14 En↔De, WMT16 En↔Ro, WMT17 En↔Zh and WMT19 En↔De respectively. We used case-sensitive token-BLEU (Papineni et al., 2002) to measure the translation quality (except for En-Zh, we used sacre-BLEU (Post, 2018)), and *sign-test* (Collins et al., 2005) for statistical significance test.

Monolingual Data We closely followed previous works to randomly sample monolingual data from publicly available News Crawl corpus[1] for the WMT tasks (Sennrich et al., 2016b; Wu et al., 2019). We randomly sampled English and German data from News Crawl 2007∼2020, and randomly sampled Romanian data from News Crawl 2015. For Chinese monolingual data, we used News Crawl 2008∼2020, News Commendary v16 and XMU data. For fair comparison, the monolingual data generally has the same size as corresponding bilingual data, as listed in Table 2.

Task	Lang.	Bilingual data		Monolingual Data	
		# Sent.	# Word	# Sent.	# Word
W14	En	4.5M	127.7M	4.5M	138.6M
	De		132.5M		124.0M
W16	En	0.6M	16.1M	0.6M	16.5M
	Ro		16.7M		17.3M
W17	En	20.6M	535.7M	20.6M	591.5M
	Zh		487.6M	18.4M	540.1M
W19	En	36.8M	881.0M	36.8M	937.3M
	De		911.0M		867.6M

Table 2: Data statistics of parallel and monolingual data. For fair comparison, the monolingual data has the same size with the corresponding bilingual data.

Model Training We validated our approach on two state-of-the-art NAT models:

- *MaskPredict* [MaskT, Ghazvininejad et al. 2019] that uses the conditional masked language model (Devlin et al., 2019) to iteratively generate the target sequence from the masked input. We followed its optimal settings to keep the iteration number be 10 and length beam be 5.

- *Levenshtein Transformer* [LevT, Gu et al. 2019] that introduces three steps: deletion, placeholder prediction and token prediction, and the decoding iterations adaptively depends on certain conditions. We followed their setting and reproduced their reported results.

We trained both BASE and BIG Transformer (Vaswani et al., 2017) as the *AT teachers* for both standard and monolingual KD. For BIG models, we adopted *large-batch training* (i.e. 458K tokens/batch) to optimize the performance (Ott et al., 2018). The En↔Ro tasks employed Transformer-BASE as the teacher, and the other tasks used Transformer-BIG as the teacher. We also used *large-batch* (i.e. 480K tokens/batch) to train NAT models with Adam optimizer (Kingma and Ba, 2015). The learning rate warms up to 1×10^{-7}

[1] http://data.statmt.org/news-crawl

Data	MaskT		LevT	
	BLEU	△	BLEU	△
$\overrightarrow{\text{KD}}_{\text{B}}$	25.4	–	25.6	–
$\overrightarrow{\text{KD}}_{\text{M}}$	25.8	+0.4	26.2	+0.6
$\overleftarrow{\text{KD}}_{\text{M}}$	24.9	-0.5	24.5	-1.1
$\overleftrightarrow{\text{KD}}_{\text{M}}$	26.6	+1.2	26.7	+1.1
$\overrightarrow{\text{KD}}_{\text{M}}$+$\overrightarrow{\text{KD}}_{\text{B}}$	26.7	+1.3	26.8	+1.2
$\overleftarrow{\text{KD}}_{\text{M}}$+$\overrightarrow{\text{KD}}_{\text{B}}$	26.6	+1.2	26.5	+0.9
$\overleftrightarrow{\text{KD}}_{\text{M}}$+$\overrightarrow{\text{KD}}_{\text{B}}$	27.1	+1.7	27.3	+1.7

Table 3: BLEU scores of different monolingual distillation strategies. "+$\overrightarrow{\text{KD}}_{\text{B}}$" means concatenating two sets of distilled data for model training, and "△" denotes improvement/decline over $\overrightarrow{\text{KD}}_{\text{B}}$. We used the same AT teacher and trained all models for the same steps.

for 10K steps, and then decays for 60k steps with the cosine schedule (Ro↔En models only need 4K and 21K steps, respectively). 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.

3.2 Ablation Study on Monolingual KD

In this section, we evaluated the impact of different components of the monolingual KD on WMT14 En-De validation sets.

Impact of Distillation Strategy Table 3 lists the results of different distillation strategies. The forward monolingual KD ("$\overrightarrow{\text{KD}}_{\text{M}}$") consistently outperforms its standard counterpart ("$\overrightarrow{\text{KD}}_{\text{B}}$") (i.e. 25.8 vs. 25.4, and 26.2 vs. 25.6), which we attribute to the advantage of monolingual KD on exploiting both the original bilingual data knowledge (implicitly encoded in the trained AT teacher model) and the new monolingual data knowledge. Concatenating forward- and reverse-KD ($\overleftrightarrow{\text{KD}}_{\text{M}}$) can further improve the NAT performance, which is consistent with the findings in Table 1.

We also investigated whether monolingual KD is complementary to standard KD (i.e. "+ $\overrightarrow{\text{KD}}_{\text{B}}$" column). As seen, standard KD consistently improves translation performance across monolingual KD variants. Another interesting finding is that although reverse monolingual KD ($\overleftarrow{\text{KD}}_{\text{M}}$) significantly underperforms its forward counterpart ($\overrightarrow{\text{KD}}_{\text{M}}$) when used alone, they achieve comparable performance when using together with standard KD. We discuss in details how the two KD models complement each other in Section 3.4.

Sampling	$\overleftrightarrow{\text{KD}}_{\text{M}}$		+ $\overrightarrow{\text{KD}}_{\text{B}}$	
	MaskT	LevT	MaskT	LevT
RANDOM	26.6	26.7	27.1	27.3
LOW-FREQ	26.4	26.6	26.9	27.1
LM-SEL	26.9	26.8	27.4	27.5

Table 4: Impact of monolingual data sampling.

Impact of Monolingual Data Sampling Some researchers may doubt that our approach heavily depends on the sampled monolingual data. To dispel the doubt, we investigated whether our model is robust to the selected monolingual data by varying the sampling strategies. Specifically, we conducted experiments on the full set of monolingual data from News Crawl 2007∼2020, which consist of 243M English and 351M German sentences. We compared with two representative approaches that sampled data with different priors: (1) LOW-FREQ samples difficult examples containing low-frequency words (Fadaee and Monz, 2018); (2) LM-SEL selects high quality examples with language model (Moore and Lewis, 2010).

As listed in Table 4, the difference of three sampling strategies w.r.t BLEU is not significant under the significance test $p < 0.05$ (Collins et al., 2005), demonstrating that *our approach is robust to the monolingual data sampling*. For the simplicity and robust applicability of our approach across different scenarios, we used RANDOM sampling as the default strategy in the following experiments.

3.3 Main Results

NAT Benchmarks Table 5 lists the results on the WMT14 En↔De and WMT16 En↔Ro benchmarks. Encouragingly, the conclusions in Section 3.2 hold across language pairs, demonstrating the effectiveness and universality of our approach. We also compared the performance against several previous competitive NAT models. Although the results are not directly comparable since we used additional monolingual data, our approach improves previous SOTA BLEU on the NAT benchmarks. Notably, our data-level approaches neither modify model architecture nor add extra training loss, thus does not increase any latency ("Speed"), maintain-

Model	Iter.	WMT14		WMT16	
		En-De	De-En	En-Ro	Ro-En
AT Models					
Transformer-BASE (En↔Ro Teacher)	n/a	27.3	31.3	33.9	34.1
Transformer-BIG (En↔De Teacher)	n/a	29.2	32.4	-	-
Existing NAT Models with Standard KD					
DisCo (Kasai et al., 2020)	4.8	27.3	31.3	33.2	33.3
Imputer (Saharia et al., 2020)	8.0	28.2	31.8	34.4	34.1
Mask-Predict (Ghazvininejad et al., 2019)	10.0	27.0	30.5	33.1	33.3
+Raw Data Pre-Train (Ding et al., 2021a)		27.8	-	-	33.9
Levenshtein (Gu et al., 2019)	2.5	27.3	-	-	33.3
+Raw Data Pre-Train (Ding et al., 2021a)		28.2	-	-	33.8
Our NAT Models					
Mask-Predict					
+Standard KD	10.0	27.0	31.1	32.9	33.3
+*Mono. KD*		28.2†	31.8	33.6†	33.7
+Standard KD		28.7†	32.3†	33.9†	34.1†
+*Bidirectional Mono. KD*		29.1†	32.6†	34.2†	34.3†
+Standard KD		**30.1**†	**33.7**†	**35.0**†	**35.3**†
Levenshtein					
+Standard KD	2.5	27.3	30.9	32.7	33.2
+*Mono. KD*		28.6†	32.1†	33.5†	33.9
+Standard KD		29.1†	32.6†	34.0†	34.2†
+*Bidirectional Mono. KD*		29.5†	33.6†	34.3†	34.2†
+Standard KD		**30.4**†	**34.1**†	**34.9**†	**35.4**†

Table 5: Comparison with previous work on NAT benchmarks. "Iter." indicates the number of iterative refinement. "†" indicates statistically significant difference ($p < 0.01$) from standard KD.

ing the intrinsic advantages of NAT models. The main side-effect of our approach is the increased training time for training an additional AT teacher model to build distilled data in the reverse direction. Fortunately, we can eliminate the side-effect by using only the monolingual KD ("Mono. KD"), which still consistently outperforms the standard KD without introducing any computation cost.

Larger-Scale WMT Benchmarks To verify the effectiveness of our method across different data sizes, we further experimented on two widely-used large-scale MT benchmarks, i.e. WMT17 En↔Zh and WMT19 En↔De. As listed in Table 6, our bidirectional monolingual KD outperforms standard KD by averagely +1.9 and +2.3 BLEU points on En↔Zh and En↔De datasets, respectively, demonstrating the robustness and effectiveness of our monolingual KD approach. By combining with standard KD, our methods can achieve further +1.8 and +0.9 BLEU improvements.

Model	En-Zh		En-De	
	→	←	→	←
AT Teacher	35.6	24.6	40.2	40.1
MaskT				
+Stand. KD	33.7	23.4	36.8	37.2
+*Mono. KD*	34.5	24.9†	37.4	37.9
+Stand. KD	34.8†	25.1†	38.1†	38.5†
+*Bid. Mono. KD*	35.2†	25.6†	39.2†	39.4†
+Stand. KD	38.2†	25.8†	40.1†	40.5†
LevT				
+Stand. KD	33.9	23.3	37.5	37.7
+*Mono. KD*	34.6	24.6†	38.1	38.4
+Stand. KD	35.1†	24.7†	38.5†	39.1†
+*Bid. Mono. KD*	35.4†	25.5†	39.6†	40.2†
+Stand. KD	38.5†	25.8†	40.5†	40.8†

Table 6: BLEU scores on large-scale WMT17 En↔Zh (20.6M) and WMT19 En↔De (36.8M) data.

Data	All	High	Med.	Low
Raw	3.67	2.41	3.28	6.81
$\overrightarrow{\text{KD}}_{\text{B}}$	1.95	1.68	1.87	4.52
$\overrightarrow{\text{KD}}_{\text{M}}$	1.79	1.66	1.72	4.29
$+\overrightarrow{\text{KD}}_{\text{B}}$	1.77	1.62	1.71	3.95
$\overleftrightarrow{\text{KD}}_{\text{M}}$	1.72	1.52	1.64	4.01
$+\overrightarrow{\text{KD}}_{\text{B}}$	1.64	1.50	1.62	3.69

Table 7: Data complexity of different distillations of WMT14 En-De training data. Word frequencies are estimated on the source sentences of bilingual data.

Data	WMT14 En-De			WMT14 De-En		
	H	M	L	H	M	L
AT Teacher						
Raw Data	84.7	80.2	73.0	85.4	81.1	74.2
NAT Student						
$\overrightarrow{\text{KD}}_{\text{B}}$	82.4	78.2	68.4	83.7	79.6	69.9
$\overrightarrow{\text{KD}}_{\text{M}}$	82.9	78.4	69.5	83.9	80.1	71.2
$+\overrightarrow{\text{KD}}_{\text{B}}$	83.1	78.7	70.8	84.3	80.5	72.1
$\overleftrightarrow{\text{KD}}_{\text{B}}$	84.1	79.1	72.7	85.0	80.9	73.4
$+\overrightarrow{\text{KD}}_{\text{B}}$	84.6	79.7	73.6	85.2	81.4	75.2

Table 8: Accuracy of word translation. Darker color denotes more improvement over standard KD. "H/M/L" represent high/medium//low frequency words, which are estimated on the source sentences of bilingual data.

3.4 Analysis

In this section, we provide some insights into how monolingual KD works. We report the results on WMT14 En-De data using Mask-Predict.

Monolingual KD Reduces Complexity of Training Data by Improving Low-Frequency Word Alignment We first present data-level qualitative analyses to study how monolingual KD complements bilingual KD. Zhou et al. (2020) revealed that standard KD improves NAT models by reducing the complexity of original bilingual data. Along this thread, we used the data complexity metric to measure different distilled datasets. Formally, the translation uncertainty of a source sentence x can be operationalized as conditional entropy:

$$\mathcal{H}(\mathbf{Y}|\mathbf{X}=\mathrm{x}) = -\sum_{y\in Y} p(\mathrm{y}|\mathrm{x}) \log p(\mathrm{y}|\mathrm{x}) \approx \sum_{t=1}^{T_x} \mathcal{H}(y|x=x_t),$$

where T_x denotes the length of the source sentence, x and y represent a word in the source and target vocabularies, respectively.

We run *fast-align* on each parallel corpus to obtain word alignment. For fair comparison, we sampled the subsets (i.e. 4.5M) of "$\overleftrightarrow{\text{KD}}_{\text{M}}$" and "$\overleftrightarrow{\text{KD}}_{\text{M}} + \overrightarrow{\text{KD}}_{\text{B}}$" to perform complexity computation. As seen in Table 7, standard KD significantly reduces the data complexity compared to that of the bilingual data (1.95 vs. 3.67), and monolingual KD reduces even more data complexity. Additionally, the data complexity can be further reduced by combining with standard KD.

Monolingual KD Mainly Improves Low-Frequency Word Translation We first followed Ding et al. 2021b to measure the translation accuracy of words with different frequencies, as shown in Table 8. The improvements over low-frequency words are the major reason for the performance gains, where the monolingual KD and bidirectional monolingual KD outperform the standard KD by averagely +1.2% and +3.9%, respectively. These findings confirm our hypothesis that monolingual KD can improve the translation of low-frequency words by redistributing them in the new monolingual data. Combining with standard KD can further improve the accuracy of translating low-frequency words, which reconfirms our hypothesis on the complementarity between the two KD methods on low-frequency words.

3.5 Further Exploiting Monolingual Data

In this section, we provide some potential directions to further improve NAT performance by making the most of monolingual data.

Exploiting Monolingual Data at Scale One strength of monolingual KD is the potential to exploit more monolingual data to further improve translation performance. To validate our claim, we scaled the size of monolingual data by $\{2\times, 5\times, 10\times\}$, which are randomly sampled from the full set of monolingual data. As shown in Table 9, enlarging the monolingual data consistently improves the BLEU scores, while this trend does not hold when further scaling the monolingual data (i.e. $10\times$). One possible reason is that the limited capacity of NAT-base models cannot fully exploit the large data, which suggests future exploration of larger NAT architectures.

Mono Size	WMT14 En-De MaskT	WMT14 En-De LevT	WMT14 De-En MaskT	WMT14 De-En LevT
Bidirectional Monolingual KD				
1×	29.1	29.5	32.6	33.6
2×	29.7	30.1	33.1	33.9
5×	**30.6**	**30.9**	**33.9**	**34.5**
10×	30.4	30.8	33.3	34.4
Combining with Standard KD				
1×	30.1	30.4	33.7	34.1
2×	30.7	30.9	34.2	34.5
5×	**31.3**	**31.7**	**34.5**	**34.7**
10×	30.9	31.5	34.2	34.6

Table 9: BLEU scores of using monolingual data at scale. We train all models with the same training steps.

Mono. KD	Mono. to Train AT	Mono. to Train NAT	BLEU AT	BLEU NAT
n/a	×	×	29.2	27.0
$\overrightarrow{\text{KD}}_{\text{M}}$	×	✓	29.2	28.7
	✓	×	30.1	27.8
	✓	✓	30.1	28.9
$\overleftrightarrow{\text{KD}}_{\text{M}}$	×	✓	29.2	30.1
	✓	×	31.8	28.2
	✓	✓	31.8	30.5

Table 10: Applying monolingual KD for AT teacher and/or NAT student on WMT14 En-De test set. Raw data (for AT) and $\overleftrightarrow{\text{KD}}_{\text{B}}$ (for NAT) are used by default.

Augmenting AT Teacher with Monolingual KD An alternative to exploit monolingual data is to strength the AT teacher with monolingual KD, as listed in Table 10. Applying monolingual KD for AT teacher is less effective than using it for NAT training, which we attribute to the information loss when transferred from AT teacher to NAT student. Applying monolingual KD to both AT teacher and NAT student can further improve the NAT performance, at the cost of more computational cost.

4 Related Work

Sequence-level KD (Kim and Rush, 2016) is a preliminary step for training NAT models to reduce the intrinsic uncertainty and learning difficulty (Zhou et al., 2020; Ren et al., 2020). Recent studies have revealed that KD reduces the modes (i.e. multiple lexical choices for a source word) in the original data by re-weighting the training examples (Furlanello et al., 2018; Tang et al., 2020), at the cost of losing some important information, leading to more errors on predicting low-frequency words (Ding et al., 2021b). In response to this problem, Ding et al. (2021a) proposed to rejuvenate low-frequency words by pretraining NAT models on the raw bilingual data. In this study, we attempt to solve this problem from a different perspective – rediscovering low-frequency words from external monolingual data, which can simultaneously exploit the knowledge of bilingual data (implicitly encoded in the parameters of AT teacher).

Closely related to our work, Zhou and Keung (2020) improved NAT models by augmenting source-side monolingual data. Their work can be regarded as a special case of our approach (i.e. "*Mono. KD + Standard KD*" in Section 3.3), and our work has several more contributions. Firstly, we demonstrated the effectiveness of using only monolingual KD for NAT models, which can achieve better performance than the standard KD without introducing any computational cost. Secondly, we proposed a novel bidirectional monolingual KD to exploit both the source-side and target-side monolingual data. Finally, we provide insights into how monolingual KD complements the standard KD.

5 Conclusion

In this work, we propose a simple, effective and scalable approach – *monolingual KD* to redistribute the low-frequency words in the bilingual data using external monolingual data. Monolingual KD consistently outperforms the standard KD with more translation accuracy of low-frequency words, which attribute to its strength of exploiting both the knowledge of the original bilingual data (implicitly encoded in the parameters of AT teacher) and that of the new monolingual data.

Monolingual KD enjoys appealing expandability, and can be further enhanced by (1) combining with a reverse monolingual KD to recall more alignments for low-frequency target words; (2) combining with the standard KD to explicitly combine both types of complementary knowledge; (3) enlarging the scale of monolingual data that is cheap to acquire. Our study empirically indicates the potential to make NAT a practical translation system. Future directions include designing advanced monolingual KD techniques and validating on larger-capacity NAT models (e.g. BIG setting).

References

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *WMT*.

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*.

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*.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. 2018. Born again neural networks. In *ICML*.

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*.

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*.

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 *ACL*.

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*.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *ACL*.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *WMT*.

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

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

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. 2016a. Improving neural machine translation models with monolingual data. In *ACL*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Improving neural machine translation models with monolingual data. In *ACL*.

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

Aditya Siddhant, Ankur Bapna, Yuan Cao, Orhan Firat, Mia Xu Chen, Sneha Kudugunta, Naveen Arivazhagan, and Yonghui Wu. 2020. Leveraging monolingual data with self-supervision for multilingual neural machine translation. In *ACL*.

Jiaxi Tang, Rakesh Shivanna, Zhe Zhao, Dong Lin, Anima Singh, Ed H. Chi, and Sagar Jain. 2020. Understanding and improving knowledge distillation. *arXiv*.

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*.

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

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *EMNLP*.

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

Jiawei Zhou and Phillip Keung. 2020. Improving non-autoregressive neural machine translation with monolingual data. In *ACL*.