Multilingual Non-Autoregressive Machine Translation without Knowledge Distillation

Chenyang Huang[*1], **Fei Huang**[*4], **Zaixiang Zheng**[3],
Osmar Zaïane[1], **Hao Zhou**[†2], **Lili Mou**[1]
[1]Dept. of Computing Science, Alberta Machine Intelligence Institute (Amii), University of Alberta
[2]Institute for AI Industry Research (AIR), Tsinghua University
[3]ByteDance Research [4]Damo Academy, Alibaba
chenyangh@ualberta.ca huangfei382@163.com zhengzaixiang@bytedance.com
zhouhao@air.tsinghua.edu.cn zaiane@ualberta.ca doublepower.mou@gmail.com

Abstract

Multilingual neural machine translation (MNMT) aims at using one single model for multiple translation directions. Recent work applies non-autoregressive Transformers to improve the efficiency of MNMT, but requires expensive knowledge distillation (KD) processes. To this end, we propose an M-DAT approach to non-autoregressive multilingual machine translation. Our system leverages the recent advance of the directed acyclic Transformer (DAT), which does not require KD. We further propose a pivot back-translation (PivotBT) approach to improve the generalization to unseen translation directions. Experiments show that our M-DAT achieves state-of-the-art performance in non-autoregressive MNMT.[1]

1 Introduction

Multilingual neural machine translation (MNMT) aims at using a single model for multiple translation directions (Firat et al., 2016). It has attracted the attention of the research community over the years (Bapna and Firat, 2019; Zhang et al., 2021), and has been widely applied in the industry (Johnson et al., 2017). Most state-of-the-art MNMT models are based on the autoregressive Transformer (AT, Vaswani et al., 2017). However, the inference of AT is slow, which results in significant latency in real-world applications (Gu et al., 2018).

Recent work applies the non-autoegressive Transformer (NAT, Gu et al., 2018), which generates target tokens in parallel for efficient inference. However, NAT often generates inconsistent sentences (e.g., with repetitive words). Qian et al. (2021) propose the Glancing Transformer (GLAT), which is trained in a curriculum learning fashion. It tackles the weakness of NAT by focusing less on the training samples that lead to inconsistent generalization.

To accelerate multilingual non-autoregressive translation, Song et al. (2022) propose a Switch-GLAT method, which is based on the Glancing Transformer, and is equipped with back-translation for data augmentation. To the best of our knowledge, Switch-GLAT is currently the only non-autoregressive system for multilingual translation. However, it suffers from two drawbacks. First, Switch-GLAT requires sequence-level knowledge distillation (KD, Kim and Rush, 2016) in every translation direction, which is inconvenient for multilingual tasks. Second, Switch-GLAT is unable to generalize to unseen translation directions (zero-shot translation), which is an essential aspect of multilingual machine translation systems (Johnson et al., 2017; Chen et al., 2017; Gu et al., 2019).

In this work, we propose a multilingual Directed Acyclic Transformer (M-DAT) approach to non-autoregressive multilingual machine translation. Our system leverages the recent directed acyclic Transformer (DAT, Huang et al., 2022b), which does not rely on KD. In addition, we propose a pivot back-translation (PivotBT) approach for the multilingual translation task. Specifically, we back-translate a target sentence to a randomly selected language to obtain an augmented source sentence. The newly generated source sentence and the original target sentence form a synthetic data sample. We observe that if the back-translation direction (e.g., German → Romanian) does not exist in the training set (i.e., zero-shot), the augmented source

[*]Work partially done during an internship at ByteDance.
[†]Work partially done while working at ByteDance.
[1]Our code and training/evaluation scripts are available at https://github.com/MANGA-UOFA/M-DAT

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023, pages 161–170

sentence will be of low quality. Therefore, our proposed PivotBT uses an intermediate language for the back translation (e.g., German → English → Romanian). Our PivotBT is efficient, as the inference of our non-autoregressive model is fast.

We evaluated M-DAT in both supervised and zero-shot translation directions. In the supervised setting, our M-DAT achieves 0.4 higher BLEU scores than the previous state-of-the-art Switch-GLAT, while maintaining fast inference. Moreover, our M-DAT does not require KD, and is convenient to be trained on multilingual datasets. In the zero-shot translation settings, our M-DAT is the first NAT model to effectively generalize to unseen translation directions, and even outperforms a strong autoregressive baseline, which is largely attributed to our proposed PivotBT.

2 Related Work

The non-autoregressive Transformer (NAT, Gu et al., 2018) predicts all target words in parallel to achieve fast inference, and has been applied to various text generation tasks, such as machine translation (Gu and Kong, 2021; Huang et al., 2022a), summarization (Su et al., 2021; Liu et al., 2022a,b), and dialogue generation (Zou et al., 2021; Qi et al., 2021). However, the output quality of NAT models tends to be low (Stern et al., 2019; Ghazvininejad et al., 2019), and as a remedy, the Glancing Transformerthe Glancing Transformer (GLAT, Qian et al., 2021) applies an adaptive training algorithm that allows the model to progressively learn more difficult data samples.

Sequence-level knowledge distillation (KD, Kim and Rush, 2016) is commonly used to improve the translation quality of non-autoregressive models. As shown in Zhou et al. (2020), KD data are less complex compared with the original training set, which is easier for NAT models. However, Ding et al. (2021) find that the KD process tends to miss low-frequency words (e.g., proper nouns), which results in worse translation for NAT models. Therefore, there is a need to remove the KD process for NAT models.

The most related work to ours is Switch-GLAT (Song et al., 2022). It combines the Glancing Transformer and knowledge distillation for multilingual machine translation tasks. In addition, Switch-GLAT is equipped with a back-translation technique to augment training data.

Our system is based on the directed acyclic

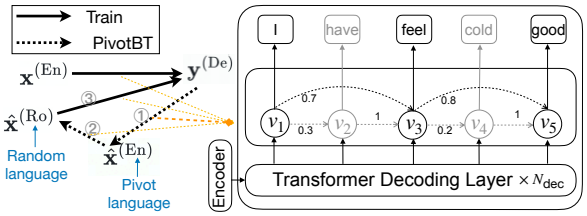

Figure 1: An example of our PivotBT augmenting a German sentence $\mathbf{y}$ to Romanian $\hat{\mathbf{x}}$, where English is used as the pivot language. The training and back-translation steps are accomplished by DAT (Huang et al., 2022b). N_{dec} is the number of decoding layers.

Transformer (DAT, Huang et al., 2022b), which expands the generation canvas to allow multiple plausible translation fragments. Then, DAT selects the fragments by predicting linkages, which eventually form an output sentence. In this way, M-DAT is more capable of handling complex data samples, and does not rely on KD.

We propose PivotBT to augment the training data to improve the generalization of M-DAT. Our PivotBT is inspired by online back-translation (Zhang et al., 2020), which extends the original back-translation (BT, Sennrich et al., 2016) by randomly picking augmented directions. Different from the previous work, our approach applies pivot machine translation (Cheng et al., 2017) to improve the reliability of the back-translation directions that are unseen in the training set.

3 Methodology

Multilingual translation handles multiple translation directions with a single model. Suppose a data sample contains a source sentence $\mathbf{x} = (\mathrm{x}_1, \cdots, \mathrm{x}_{T_\mathrm{x}})$ and a target sentence $\mathbf{y} = (\mathrm{y}_1, \cdots, \mathrm{y}_{T_\mathrm{y}})$, where T_x and T_y denote the lengths. In addition, language tags l_{src} and l_{tgt} are given to indicate the languages of the source and target sentences, respectively. A multilingual machine translation dataset $\mathcal{D}$ can be represented by $\{(\mathbf{x}^{(i)}, l^{(i)}_{\mathrm{src}}, \mathbf{y}^{(i)}, l^{(i)}_{\mathrm{tgt}})\}_{i=1}^{K}$, where K is the size of the dataset.

Our multilingual Directed Acyclic Transformer (M-DAT) has an encoder–decoder architecture. The encoder of M-DAT takes in an input sentence $\mathbf{x}^{(i)}$ and the target language tag $l^{(i)}_{\mathrm{tgt}}$, whereas the decoder predicts the target-language words independently as the translation.

3.1 Directed Acyclic Transformer

To train our system without KD, we adapt the recent directed acyclic Transformer (DAT, Huang et al., 2022b), as it does not use KD but achieves comparable performance to autoregressive models on bilingual machine translation tasks (one direction per model).

In general, DAT expands its output canvas to generate multiple plausible translation fragments. Further, DAT predicts links to select the fragments, which form an output sentence. As seen in Figure 1, DAT predicts extra words and forms the final generation "I feel good", using the predicted links.

Suppose the DAT decoder has S generation steps ($S > T_{\mathrm{y}}$). For each step s within $1 \le s \le S$, DAT makes two predictions: word prediction $p^{(s)}_{\text{word}}(\cdot)$ and link prediction $p^{(s)}_{\text{link}}(\cdot)$.

The word prediction $p^{(s)}_{\text{word}}(\cdot)$ gives the distribution over possible words by mapping DAT's sth decoder state $\boldsymbol{h}_s$ to the probability distribution over the vocabulary, given by

$$p^{(s)}_{\text{word}}(\cdot) = \mathrm{softmax}(\mathbf{W}_{\text{word}}\boldsymbol{h}_s) \tag{1}$$

where $\mathbf{W}_{\text{word}}$ is a learnable matrix.

The link prediction $p^{(s)}_{\text{link}}(\cdot)$ computes the distribution over the subsequent steps of the sth step, determining which follow-up step should be linked to the sth step. Specifically, the link prediction leverages the attention mechanism (Bahdanau et al., 2015), which compares the sth step's hidden state $\boldsymbol{h}_s$ with the states of subsequent generation steps (from $s+1$ to S), given by

$$p^{(s)}_{\text{link}}(\cdot) = \mathrm{softmax}([\boldsymbol{k}_s^\top \boldsymbol{q}_{s+1}; \cdots; \boldsymbol{k}_s^\top \boldsymbol{q}_S]) \tag{2}$$

where $\boldsymbol{k}_s = \mathbf{W}_{\mathrm{k}}\boldsymbol{h}_s$ and $\boldsymbol{q}_s = \mathbf{W}_{\mathrm{q}}\boldsymbol{h}_s$. $\mathbf{W}_{\mathrm{k}}$ and $\mathbf{W}_{\mathrm{q}}$ are learnable matrices. The operation $[;]$ concatenates scalars into a column vector.

Given a reference sequence $\mathbf{y}$ in the training set $\mathcal{D}$, DAT selects T_{y} of all S generation steps to generate the words in $\mathbf{y}$, where the selected steps are connected by predicted links. We denote the indices of the selected steps by $\boldsymbol{a} = (a_1, \cdots, a_{T_{\mathrm{y}}})$, where $1 = a_1 < \cdots < a_{T_{\mathrm{y}}} = S$. We refer to each selection of the steps $\boldsymbol{a}$ as a *path*.

Consider a groundtruth sequence $\mathbf{y}_{1:T_{\mathrm{y}}}$. The joint probability of the sequence, together with some path $\boldsymbol{a}_{1:T_{\mathrm{y}}}$, is

$$p(\mathbf{y}_{1:T_{\mathrm{y}}}, \boldsymbol{a}_{1:T_{\mathrm{y}}}) = \prod_{t=2}^{T_{\mathrm{y}}} p^{(a_{t-1})}_{\text{link}}(a_t) \prod_{t=1}^{T_{\mathrm{y}}} p^{(a_t)}_{\text{word}}(\mathrm{y}_t) \tag{3}$$

where $p^{(a_{t-1})}_{\text{link}}(a_t)$ is the probability that the two generation steps a_{t-1} and a_t are linked up. Specially, a_1 is set to 1, and is not considered as a random variable. $p^{(a_t)}_{\text{word}}(\mathrm{y}_t)$ is the probability of predicting the word y_t at the a_tth generation step.

Finally, the probability of generating the reference sentence $p(\mathbf{y}_{1:T_{\mathrm{y}}})$ is obtained by the marginalization of all possible paths, given by

$$\begin{aligned} p(\mathbf{y}_{1:T_{\mathrm{y}}}) &= \sum_{\boldsymbol{a} \in \Gamma_{S,T_{\mathrm{y}}}} p(\mathbf{y}_{1:T_{\mathrm{y}}}, \boldsymbol{a}_{1:T_{\mathrm{y}}}) \\ &= \sum_{\boldsymbol{a} \in \Gamma_{S,T_{\mathrm{y}}}} \prod_{t=2}^{T_{\mathrm{y}}} p^{(a_{t-1})}_{\text{link}}(a_t) \prod_{t=1}^{T_{\mathrm{y}}} p^{(a_t)}_{\text{word}}(\mathrm{y}_t) \end{aligned} \tag{4}$$

where $\Gamma_{S,T_{\mathrm{y}}} = \{\boldsymbol{a} = (a_1, \cdots, a_{T_{\mathrm{y}}}) | 1 = a_1 < \cdots < a_{T_{\mathrm{y}}} = S\}$ represents all paths of length T_{y}. The computation of (4) is efficient through dynamic programming.[2]

Note that $p^{(s)}_{\text{word}}(\cdot)$ and $p^{(s)}_{\text{link}}(\cdot)$ are independently predicted for different generation steps; thus, DAT is non-autoregressive and is fast in inference.

3.2 Pivot Back-Translation

We propose a pivot back-translation (PivotBT) approach to improve the robustness of M-DAT. Following Zhang et al. (2020) and Song et al. (2022), we augment the training data with back-translation (BT, Sennrich et al., 2016). Specifically, a randomly selected language is chosen for such data augmentation.

We observe that when the back-translation direction is unseen (i.e., zero-shot), the synthesized source sentence will be of low quality, which results in a less meaningful synthetic training sample. To this end, we propose to handle the zero-shot scenario by PivotBT, which uses an intermediate language as a pivot and performs multi-step back-translation.

Given a training sample $(\mathbf{x}, l_{\text{src}}, \mathbf{y}, l_{\text{tgt}})$, we first randomly pick a language l_{aug} from the set of languages in the multilingual training set $\mathcal{D}$. If the back-translation direction $l_{\text{tgt}} \to l_{\text{aug}}$ is in the training set, we directly back-translate $\mathbf{y}$ to $\hat{\mathbf{x}}$ of language l_{aug}. Otherwise, we choose a pivot language l_{pivot} (e.g., English) such that $l_{\text{tgt}} \to l_{\text{pivot}}$ and $l_{\text{pivot}} \to l_{\text{aug}}$ are both in the training set.[3] In this

[2] We refer readers to Huang et al. (2022b).

[3] Most multilingual translation datasets are English-centric, where using English as l_{pivot} guarantees the connection of l_{tgt} and l_{aug}, which is the case of this work. In general, we can use multiple pivot languages to connect l_{tgt} and l_{aug} with multiple back-translation steps.

#	Model Variant	EFD	EFZ	MANY
1	M-AT w/ standard layout	**34.57**	**31.15**	**30.36**
2	M-AT w/ shallow decoder	34.19	30.87	29.20
3	Switch-GLAT*	33.34	29.76	28.47
4	M-DAT w/ lookahead	33.72	30.39	28.69
5	w/ n-gram beam search	33.83	30.55	29.73

Table 1: BLEU scores on three WMT datasets. *Trained with sequence-level knowledge distillation.

Model Variant	IWSLT		Europarl	
	Supervised	0-Shot	Supervised	0-Shot
M-AT w/ standard layout	**30.00**	12.87	**35.79**	15.84
M-AT w/ shallow decoder	29.23	4.95	34.95	8.11
Residual M-AT	29.72	17.67	35.18	26.13
M-DAT w/ lookahead	28.58	18.53	34.83	25.86
w/ n-gram beam search	29.42	**19.35**	35.48	**27.44**

Table 2: BLEU scores on IWSLT and Europarl.

Model Variant	Latency (ms)	Speedup
M-AT w/ standard layout	352.4	1.0×
M-AT w/ shallow decoder	84.2	4.2×
Switch-GLAT	**19.6**	**18.0**×
M-DAT w/ lookahead	21.9	16.1×
w/ n-gram beam search	67.6	5.2×

Table 3: Latency and speedup on WMT-EFD.

way, we are able to obtain an augmented source sentence $\hat{\mathbf{x}}$ by first translating $\mathbf{y}$ to $\hat{\mathbf{x}}_{\text{pivot}}$ of the intermediate language l_{pivot}, and then translating $\hat{\mathbf{x}}_{\text{pivot}}$ to the augmented source sentence $\hat{\mathbf{x}}$ of language l_{aug}. Finally, the newly synthesized sample $(\hat{\mathbf{x}}, l_{\text{aug}}, \mathbf{y}, l_{\text{tgt}})$ is added to the training.

In our PivotBT, the multi-step back-translation is conducted by M-DAT itself. Since M-DAT is fast in inference, the back-translation is also efficient.

We denote the loss of training the real samples in dataset $\mathcal{D}$ by $\mathcal{L}_{\text{real}}$ and that of the synthetic samples by $\mathcal{L}_{\text{PivotBT}}$. The overall training loss of our proposed system is $\mathcal{L} = \mathcal{L}_{\text{real}} + \lambda \mathcal{L}_{\text{PivotBT}}$, where λ is a hyperparameter controlling the strength of the back-translation.

4 Experiments

4.1 Setup

We evaluated M-DAT on five datasets: WMT-EFD, WMT-EFZ, WMT-MANY, IWSLT, and Europarl. The three WMT corpora (Song et al., 2022) only contain supervised directions in the test set, whereas the test sets of IWSLT and Europarl (Liu et al., 2021) contain both supervised directions and unseen directions (zero-shot). The training hyperparameters and evaluate metrics are strictly following Song et al. (2022) and Liu et al. (2021). We provide more details in Appendices A and B.

4.2 Main Results

Supervised Translation. Table 1 summarizes the BLEU scores of three WMT datasets. We evaluated M-DAT with two decoding algorithms: *lookahead* and *n-gram beam search*. The lookahead method directly decodes the generated words in parallel, and jointly maximizes the probability of the next position and predicted tokens, whereas n-gram beam search generates a few candidate sentences and ranks them with an n-gram language model. We observe that the generation quality with n-gram beam search is higher, which is consistent with Huang et al. (2022b).

We also see that M-DAT outperforms Switch-GLAT on average with both lookahead and beam search decoding methods. This makes our M-DAT the state of the art in non-autoregressive multilingual translation. In addition, our system does not rely on KD, which makes it convenient to be trained on multilingual datasets.

To compare M-DAT with the autoregressive multilingual Transformer (M-AT, Johnson et al., 2017), we include two autoregressive Transformer-based variants: 1) the standard layout (Vaswani et al., 2017), which has the same number of encoder layers; and 2) the layout with a shallow decoder, which moves all but one decoding layers to the encoder. The layout with a shallow decoder is suggested by Kasai et al. (2021) as it achieves close results to the standard layout but is faster in inference. We observe that M-DAT is only slightly lower in BLEU scores compared with the M-AT models. This shows our M-DAT, in addition to its efficiency, largely closes the gap between AT and NAT in non-autoregressive multilingual translation tasks.

Zero-Shot Translation. The ability to generalize to unseen translation directions is important for multilingual models. In this setting, we do not compare our model with Switch-GLAT as it fails to achieve reasonable performance.[4] Instead, we compare M-DAT with the M-AT models (Johnson et al., 2017), and include a recent study, Residual M-AT (Liu et al., 2021), which relaxes the residual connections in the Transformer encoder to force the decoder to learn more generalized representations.[5]

[4] As seen in the Table 7 of Song et al. (2022), Switch-GLAT only obtained a 2.34 BLEU on a zero-shot dataset.

[5] The numbers of Residual M-AT are based on our replication, and are close to those in Liu et al. (2021).

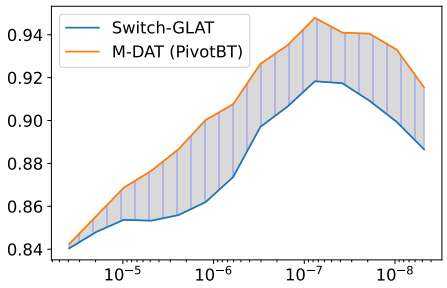

Figure 2: Comparison between M-DAT and Switch-GLAT in the preservation ratio of low-frequency words.

As seen in Table 2, the M-ATs are incapable of zero-shot translation, and are largely outperformed by the Residual M-AT. On the other hand, our M-DAT with the lookahead decoding outperforms Residual M-AT by 0.86 BLEU on the IWSLT dataset, although it is 0.27 lower on the Europarl dataset. With n-gram beam search, the improvement is 1.68 BLEU on IWSLT and 1.31 on Europarl. Our M-DAT is the first non-autoregressive model to outperform a strong AT baseline in zero-shot multilingual translation.[6]

Inference Speed. We compare the inference speed on the test set of WMT-EFD and present the results in Table 3. The batch size is set to 1 to mimic the real-world scenario, where the user requests come one after another (Gu et al., 2018). As seen, M-DAT with lookahead is about 16 times faster than the standard autoregressive baseline, and is about 4 times faster than M-AT with a shallow decoder. Compared with Switch-GLAT, M-DAT with lookahead is about the same efficiency. Admittedly, M-DAT with beam search is slower, but is still 5.2 times faster than the standard M-AT. In general, M-DAT obtains a good speed–quality trade-off.

4.3 Analysis

Low-Frequency Words. We analyze the generated text (on the WMT-EFD test set) of our M-DAT and Switch-GLAT. We followed Ding et al. (2021), and computed the percentage of a word being preserved in the translated sentence (in the corresponding language); then we grouped the words by their frequencies in the dataset.[7] As seen in Figure 2, M-DAT keeps more low-frequency words, which verifies our motivation to develop a non-autoregressive multilingual model without the help of knowledge distillation. In addition to the BLEU scores, this result further provides evidence that our M-DAT has better translation quality than Switch-GLAT.

[6]In our experiments, the autoregressive baseline models are not equipped with PivotBT, as their inefficient inference (as seen in Table 3) makes the training with back-translation impractical. In fact, our PivotBT is another way to take advantage of the fast inference of non-autoregressive models.

[7]We applied the `FastAlign` alignment tool (Dyer et al., 2013) to approximate word preservation.

#	Model Variant	Supervised	Zero-Shot
1	M-DAT (PivotBT)	**29.42**	**19.35**
2	rand-lang & w/o pivot	29.33	18.10
3	src-lang & w/o pivot	29.38	13.78
4	w/o BT	28.55	13.37

Table 4: Ablation study on the IWSLT dataset. The results are generated with n-gram beam search.

Ablation Study. Table 4 presents an ablation study on the pivot back-translation (PivotBT) of M-DAT using the IWSLT dataset. In addition to PivotBT, we consider 3 variants: 1) *M-DAT rand-lang & w/o pivot*, which randomly selects an augmented source language but does not translate through a pivot language; 2) *M-DAT src-lang & w/o pivot*, which directly back-translates the target sentence to the language of the source sentence; and 3) *M-DAT w/o BT*, which does not augment the training data with back-translation.

In the supervised setting, we observe that back-translation improves the performance (Lines 1–3 vs. Line 4), which is consistent with the findings of previous work (Johnson et al., 2017). Among back-translation methods, *src-lang & w/o pivot* performs the worst in the zero-shot setting (Line 3). We conjuncture that this is because only applying back-translation to the source language makes the model focus too much on the supervised directions, which degenerates the generalization to the zero-shot setting. On the other hand, our PivotBT outperforms the direct random back-translation (*rand-lang & w/o pivot*) and the source-language back-translation (*src-lang & w/o pivot*). This confirms that PivotBT provides the model with better-augmented samples for the zero-shot translation.

5 Conclusion

In this work, we propose M-DAT to tackle non-autoregressive multilingual machine translation (MNMT). Our approach leverages the recent directed acyclic Transformer so that we do not need the knowledge distillation process, which is particularly inconvenient in the multilingual translation task. Further, we propose a pivot back-translation method to improve the robustness. Our M-DAT achieves state-of-the-art results on supervised and zero-shot settings for non-autoregressive MNMT.

6 Limitation

One possible limitation is that our M-DAT obtains slightly lower BLEU scores compared with autoregressive models in the supervised setting. However, this is not the drawback of this work, as it is understandable that non-autoregressive models trade quality with more efficiency. Nevertheless, our M-DAT outperforms the previous state-of-the-art NAT approach in both supervised and zero-shot settings, and is easier to be deployed since KD is not required. Our PivotBT, in principle, can also be applied to the training of multilingual autoregressive Transformer (M-AT). However, we do not include M-AT with PivotBT for two reasons: 1) the main focus of this research is on the non-autoregressive Transformer; and 2) the decoding of M-AT is much slower than our M-DAT, which makes the training of M-AT with PivotBT impractical.

Acknowledgments

We would like to thank all reviewers and chairs for their comments. This research was supported in part by the Natural Science Foundation of China under Grant No. 62376133. This research was also supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) under Grant Nos. RGPIN-2020-04440 and RGPIN-2020-04465, the Amii Fellow Program, the Canada CIFAR AI Chair Program, the Alberta Innovates Program, and the Digital Research Alliance of Canada (alliancecan.ca).

References

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Roee Aharoni, Melvin Johnson, and Wolfgang Macherey. 2019. The missing ingredient in zero-shot neural machine translation. *arXiv preprint arXiv:1903.07091*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 1538–1548.

Yun Chen, Yang Liu, Yong Cheng, and Victor O.K. Li. 2017. A teacher-student framework for zero-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1925–1935.

Yong Cheng, Qian Yang, Yang Liu, Maosong Sun, and Wei Xu. 2017. Joint training for pivot-based neural machine translation. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 3974–3980.

Liang Ding, Longyue Wang, Xuebo Liu, Derek F. Wong, Dacheng Tao, and Zhaopeng Tu. 2021. Rejuvenating low-frequency words: Making the most of parallel data in non-autoregressive translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 3431–3441.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM Model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-predict: Parallel decoding of conditional masked language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 6112–6121.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O.K. Li, and Richard Socher. 2018. Non-autoregressive neural machine translation. In *International Conference on Learning Representations*.

Jiatao Gu and Xiang Kong. 2021. Fully non-autoregressive neural machine translation: Tricks of the trade. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP*, pages 120–133.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O.K. Li. 2019. Improved zero-shot neural machine translation via ignoring spurious correlations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1258–1268.

Chenyang Huang, Hao Zhou, Osmar R Zaïane, Lili Mou, and Lei Li. 2022a. Non-autoregressive translation with layer-wise prediction and deep supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10776–10784.

Fei Huang, Hao Zhou, Yang Liu, Hang Li, and Minlie Huang. 2022b. Directed acyclic Transformer for non-autoregressive machine translation. In *International Conference on Machine Learning*.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, pages 339–351.

Jungo Kasai, Nikolaos Pappas, Hao Peng, James Cross, and Noah A. Smith. 2021. Deep encoder, shallow decoder: Reevaluating non-autoregressive machine translation. In *International Conference on Learning Representations*.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327.

Danni Liu, Jan Niehues, James Cross, Francisco Guzmán, and Xian Li. 2021. Improving zero-shot translation by disentangling positional information. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1259–1273.

Puyuan Liu, Chenyang Huang, and Lili Mou. 2022a. Learning non-autoregressive models from search for unsupervised sentence summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7916–7929.

Puyuan Liu, Xiang Zhang, and Lili Mou. 2022b. A character-level length-control algorithm for non-autoregressive sentence summarization. *Advances in Neural Information Processing Systems*, pages 29101–29112.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Weizhen Qi, Yeyun Gong, Jian Jiao, Yu Yan, Weizhu Chen, Dayiheng Liu, Kewen Tang, Houqiang Li, Jiusheng Chen, Ruofei Zhang, et al. 2021. Bang: Bridging autoregressive and non-autoregressive generation with large scale pretraining. In *International Conference on Machine Learning*, pages 8630–8639.

Lihua Qian, Hao Zhou, Yu Bao, Mingxuan Wang, Lin Qiu, Weinan Zhang, Yong Yu, and Lei Li. 2021. Glancing transformer for non-autoregressive neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1993–2003.

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

Zhenqiao Song, Hao Zhou, Lihua Qian, Jingjing Xu, Shanbo Cheng, Mingxuan Wang, and Lei Li. 2022. switch-GLAT: Multilingual parallel machine translation via code-switch decoder. In *International Conference on Learning Representations*.

Mitchell Stern, William Chan, Jamie Kiros, and Jakob Uszkoreit. 2019. Insertion transformer: Flexible sequence generation via insertion operations. In *Proceedings of the International Conference on Machine Learning*, pages 5976–5985.

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 *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 234–243.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

Biao Zhang, Ankur Bapna, Rico Sennrich, and Orhan Firat. 2021. Share or not? Learning to schedule language-specific capacity for multilingual translation. In *International Conference on Learning Representations*.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639.

Chunting Zhou, Jiatao Gu, and Graham Neubig. 2020. Understanding knowledge distillation in non-autoregressive machine translation. In *International Conference on Learning Representations*.

Yicheng Zou, Zhihua Liu, Xingwu Hu, and Qi Zhang. 2021. Thinking clearly, talking fast: Concept-guided non-autoregressive generation for open-domain dialogue systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2215–2226.

	Directions	Languages	Language pairs
WMT-EFZ	6	3	en↔fr, en↔zh
WMT-EFD	6	3	en↔fr, en↔de
WMT-MANY	10	6	en↔fr, en↔zh, en↔de, en↔ro, en↔ru

Table 5: The translation directions of the WMT-EFD, WMT-EFZ, and WMT-MANY datasets.

A Datasets

WMT Datasets. WMT-EFZ, EFD, and MANY are specifically curated for multilingual machine translation; each is a mix of a few general bilingual machine translation corpora.[8] We list the translation directions of the three datasets in Table 5. As seen, both WMT-EFZ and WMT-EFD have 3 languages and 6 translation directions, whereas WMT-MANY has 5 languages and 10 directions.

We strictly followed Song et al. (2022) for data preparation, and we set the vocabulary size as 85K for WMT-EFD and WMT-EFZ, and 95K for WMT-MANY.

IWSLT Dataset. We followed Liu et al. (2021) for the IWSLT dataset, and directly obtained the processed data from their published codebase.[9] The vocabulary size of IWSLT is 19K. The IWSLT test set contains 3 supervised language pairs: en↔ro, en↔it, and en↔nl. Additionally, it contains 3 zero-shot language pairs: ro↔it, ro↔nl, and it↔nl. The training set for each supervised direction includes 145K samples.

Europarl Dataset. We further include the Europarl dataset to evaluate the multilingual capability of the proposed method. We followed Liu et al. (2021) for data preprocessing. The Europarl test set has 16 supervised directions (containing English) and 56 zero-shot directions (not containing English). The translation directions are detailed in Table 8.

B Settings

Evaluation Metrics. We used BLEU (Papineni et al., 2002) to evaluate the translation quality. To ensure a fair comparison with previous work, we applied two BLEU variants. For the WMT datasets, we followed Song et al. (2022) and applied tokenized BLEU. For the IWSLT and Europarl datasets, we followed Liu et al. (2021) and adopted SacreBLEU (Post, 2018).

We evaluated the latency on a single Tesla V100 GPU with a batch size of 1 to mimic the real-world scenario where the users' requests come one by one. Our evaluation scripts are also available from the released code.

WMT and Europarl Datasets. We used the Transformer-base configuration (Vaswani et al., 2017) as the backbone. To train the model, we set the batch size such that it contains 64K tokens, and let the model train for 800K updates. We used the Adam optimizer. The learning rate was warmed up to 5e-4 using 10K updates, and was annealed with the inverse square roots scheduler. The back-translation strength λ was set to 0.5. To balance the sizes of different translation directions, we set the upsampling ratio to 1/3. Since the validation set only contains supervised directions for the IWSLT dataset, we further applied the regularization on the encoder representations (Arivazhagan et al., 2019) to prevent overfitting to the supervised directions.

Following the setting in most non-autoregressive machine translation studies (Gu et al., 2018; Gu and Kong, 2021; Song et al., 2022; Huang et al., 2022a), we evaluated both AT and NAT models by averaging the weights of the best 5 checkpoints, which were selected by their BLEU scores on the validation set.

For the neural architecture, both our M-DAT and the M-AT with the standard layout have 6 encoder layers and 6 decoder layers. The shallow-decoder M-AT has 12 encoder layers and 1 decoder layer.

IWSLT dataset. Most of the settings for IWSLT are the same as those for the WMT and Europarl datasets, but we made some adaptations. Since the IWSLT dataset is smaller, we set the batch size to 32K tokens. Further, we followed Liu et al. (2021), and set the number of encoders and decoders to 5 for our M-DAT and the M-AT with the standard layout. On the other hand, the M-AT with the shallow-decoder layout M-AT has 10 encoder layers and 1 decoder layer.

C Detailed Results

We list the per-direction BLEU scores of the three WMT datasets in Table 6, IWSLT in Table 7, and Europarl in Table 8.

As seen, our M-DAT is only slightly outperformed by M-AT (Johnson et al., 2017), and the gap is small. However, our M-DAT outperforms

[8] The corpora are obtained from the WMT workshop: https://www.statmt.org/wmt17/

[9] https://github.com/nlp-dke/NMTGMinor/tree/master/recipes/zero-shot

Supervised Setting	AVG	en-it	en-nl	en-ro	it-en	nl-en	ro-en
M-AT w/ standard layout	**30.0**	**32.7**	29.0	**24.4**	**34.6**	29.8	**29.5**
M-AT w/ shallow decoder	**30.0**	**32.7**	29.0	**24.4**	**34.6**	29.8	**29.5**
Residual M-AT	29.7	32.6	**29.1**	23.9	33.8	29.6	29.3
M-DAT w/ lookahead	28.6	30.8	27.6	22.3	33.3	29.3	28.2
w/ n-gram beam search	29.4	32.1	28.4	23.0	34.0	**30.0**	29.0
Zero-Shot Setting	AVG	it-en	it-ro	nl-it	nl-ro	ro-it	ro-nl
M-AT	12.87	13.3	12.3	13.4	11.7	14.2	12.3
M-AT	12.87	13.3	12.3	13.4	11.7	14.2	12.3
Residual M-AT	17.67	18.2	17.3	18.3	15.2	19.8	17.2
M-DAT w/ lookahead	18.53	20.0	17.8	19.3	15.3	20.9	17.9
w/ n-gram beam search	**19.35**	**20.6**	**18.6**	**20.3**	**16.0**	**21.8**	**18.8**

Table 6: BLEU scores on the IWSLT dataset.

Model		AVG	en-fr	fr-en	en-de	de-en		AVG	en-fr	fr-en	en-zh	zh-en
M-AT w/ standard layout	**WMT-EFD**	**34.57**	**42.30**	**37.88**	25.85	**32.23**	**WMT-EFZ**	**31.15**	**42.14**	**37.64**	21.02	**23.80**
M-AT w/ shallow decoder		34.19	42.19	37.39	**26.22**	30.96		30.87	42.15	37.92	**21.44**	21.98
Switch-GLAT		33.34	40.81	36.00	25.27	31.29		29.76	40.54	36.48	19.47	22.55
M-DAT w/ lookahead		33.72	41.81	37.69	24.00	31.31		30.39	41.23	36.96	20.46	22.90
w/ n-gram beam search		33.83	41.54	37.84	24.23	31.70		30.55	41.12	37.34	20.87	22.85

WMT-MANY	AVG	en-de	de-en	en-fr	fr-en	en-ro	ro-en	en-ru	ru-en	en-zh	zh-en
M-AT w/ standard layout	**30.36**	24.67	**32.16**	41.67	**38.00**	**32.86**	**35.65**	**24.22**	**30.47**	20.66	23.27
M-AT w/ shallow decoder	29.19	**25.41**	31.38	**41.98**	37.88	30.18	34.16	21.87	26.92	**20.90**	21.27
Switch-GLAT	28.47	24.18	30.49	39.47	36.30	31.93	32.40	24.16	28.33	16.25	21.23
M-DAT w/ lookahead	28.69	23.48	30.30	40.78	36.76	30.77	34.83	20.14	28.30	19.61	21.94
w/ n-gram beam search	29.73	23.91	31.52	41.12	37.63	32.11	35.44	21.45	30.09	20.60	**23.42**

Table 7: BLEU scores on three WMT datasets.

Switch-GLAT in most of the language directions on the WMT datasets.

Moreover, our proposed M-AT outperforms the strong Residual M-AT model (Liu et al., 2021) on all zero-shot translation directions of the IWSLT dataset and of most of the zero-shot directions of the Europarl dataset.

Direction	M-AT w/ standard layout	M-AT w/ shallow decoder	Residual M-AT	M-DAT w/ lookahead	M-DAT w/ n-gram beam search
da-en	**38.5**	37.6	37.9	38.4	38.1
de-en	**36.3**	35.4	35.4	35.8	36
en-da	**36.5**	35.5	36.3	34.7	35.9
en-de	**28.6**	27.7	27.8	26.8	28.1
en-es	**43.0**	42.1	42.2	41.7	42.9
en-fi	21.9	20.5	21.5	19.8	**22**
en-fr	38.8	38	37.8	37.2	**39**
en-it	**33.3**	32.7	32.7	31.7	33.2
en-nl	**30.0**	28.6	29.5	28.8	**30.0**
en-pt	**39.3**	37.9	38.3	37.7	38.8
es-en	**43.2**	43.0	42.6	42.7	42.1
fi-en	**32.7**	31.4	32.3	32.6	**32.7**
fr-en	38.5	38.2	38.1	**38.8**	38.6
it-en	**36.6**	36.3	36.3	36.1	36.0
nl-en	**34.5**	33.3	33.6	34.1	34.3
pt-en	40.9	**41.0**	40.5	40.4	40
Average	**35.79**	34.95	35.18	34.83	35.48
da-de	15.2	9.8	23.85	23.5	**25.1**
da-es	28.1	14.8	31.08	31.7	**33.3**
da-fi	13.3	8.1	17.77	15.9	**18.4**
da-fr	22.0	13.0	28.58	28.6	**30.8**
da-it	18.4	10.6	26.17	25.0	**26.8**
da-nl	18.3	8.8	24.24	25.1	**26.6**
da-pt	24.6	11.0	28.57	28.6	**30.1**
de-da	8.7	5.4	26.58	27.8	**29.8**
de-es	21.8	9.8	29.65	31.1	**32.8**
de-fi	8.9	5.0	**19.02**	15.3	17.5
de-fr	15.7	8.6	29.2	28.5	**30.4**
de-it	15.2	6.8	**26.27**	24.3	26.0
de-nl	17.7	4.9	25.42	25.5	**27.1**
de-pt	15.4	6.7	**28.83**	28.0	29.7
es-da	11.2	5.7	29.02	29.5	**30.7**
es-de	12.0	5.5	**24.85**	23.5	24.6
es-fi	9.0	5.7	**20.28**	17.0	19.0
es-fr	26.2	11.1	33.47	34.2	**35.8**
es-it	19.5	9.0	**31.55**	29.6	31.0
es-nl	13.7	4.9	25.3	25.9	**27.1**
es-pt	23.9	9.8	34.24	35.5	**36.9**
fi-da	9.8	5.8	24.45	24.2	**26.1**
fi-de	9.2	5.9	20.4	19.1	**20.9**
fi-es	15.2	9.8	28.74	27.6	**29.2**
fi-fr	12.9	7.8	25.27	25.2	**27.0**
fi-it	9.3	6.5	**24.37**	21.5	23.1
fi-nl	11.3	5.0	20.95	21.2	**22.9**
fi-pt	12.9	6.4	25.78	24.9	**26.6**
fr-da	9.8	7.4	25.57	26.9	**28.6**
fr-de	12.9	6.9	22.08	22.2	**23.7**
fr-es	26.8	14.2	32.83	35.5	**36.9**
fr-fi	9.9	7.1	15.56	15.4	**17.5**
fr-it	19.5	12.5	28.18	28.8	**30.4**
fr-nl	14.6	6.2	24.33	24.6	**26.5**
fr-pt	20.8	10.1	30.01	32.7	**34.0**
it-da	9.4	4.7	24.62	25.5	**26.8**
it-de	10.5	5.1	21.75	20.5	**21.9**
it-es	22.2	10.1	31.99	33.6	**34.6**
it-fi	8.6	4.7	**17.45**	14.6	16.7
it-fr	22.1	9.9	30.77	31.8	**33.2**
it-nl	12.5	4.5	22.71	23.2	**24.8**
it-pt	16.7	7.9	29.8	30.4	**32.1**
nl-da	14.3	6.8	27.97	26.5	**28.0**
nl-de	13.8	6.9	**25.1**	22.6	24.1
nl-es	21.9	13.6	29.78	29.7	**31.4**
nl-fi	9.0	5.7	**17.19**	14.6	16.2
nl-fr	19.4	10.7	27.12	27.7	**29.4**
nl-it	16.4	8.9	24.79	23.4	**25.1**
nl-pt	18.2	9.2	27.57	26.9	**28.4**
pt-da	11.9	6.1	26.8	28.2	**29.3**
pt-de	11.5	6.3	**24.42**	22.4	23.7
pt-es	28.1	13.2	36.41	37.2	**38.2**
pt-fi	9.8	5.9	17.26	16.4	**18.3**
pt-fr	25.0	12.3	33.03	34.1	**35.0**
pt-it	19.2	9.7	29.96	29.5	**30.3**
pt-nl	12.8	5.4	24.25	25.5	**26.4**
Average	15.84	8.11	26.13	25.86	**27.44**

Table 8: BLEU scores on the Europarl dataset.