Helping the Weak Makes You Strong: Simple Multi-Task Learning Improves Non-Autoregressive Translators

Xinyou Wang[♠] **Zaixiang Zheng**[♡] **Shujian Huang**[♠]
[♠]National Key Laboratory for Novel Software Technology, Nanjing University
[♡]ByteDance AI Lab
wangxinyou@smail.nju.edu.cn, *zhengzaixiang@bytedance.com*
huangsj@nju.edu.cn

Abstract

Recently, non-autoregressive (NAR) neural machine translation models have received increasing attention due to their efficient parallel decoding. However, the probabilistic framework of NAR models necessitates conditional independence assumption on target sequences, falling short of characterizing human language data. This drawback results in less informative learning signals for NAR models under conventional MLE training, thereby yielding unsatisfactory accuracy compared to their autoregressive (AR) counterparts. In this paper, we propose a simple and model-agnostic multi-task learning framework to provide more informative learning signals. During training stage, we introduce a set of sufficiently weak AR decoders that solely rely on the information provided by NAR decoder to make prediction, forcing the NAR decoder to become stronger or else it will be unable to support its weak AR partners. Experiments on WMT and IWSLT datasets show that our approach can consistently improve accuracy of multiple NAR baselines without adding any additional decoding overhead.

1 Introduction

State-of-the-art neural machine translation (NMT) systems are mainly autoregressive (AR) models (Bahdanau et al., 2015; Vaswani et al., 2017), which decompose the joint probability of a sequence of tokens in a left-to-right order, modeling dependencies of each token with its preceding ones. Despite having strong performance, such sequential decoding causes considerable latency, thereby unsatisfactory efficiency.

In contrast, non-autoregressive (NAR) translation models (Gu et al., 2018) permit potentially more efficient parallel decoding. To do so, NAR models necessitate a notorious conditional independence assumption on target sequences as a trade-off. This assumption, however, is probalistically insufficient to describe the highly multi-modal nature of human language data, imposing severe challenges for NAR models in a way of yielding less informative learning signals and gradients under the conventional MLE training. As a result, NAR models often manifest implausible neural representations, especially in the decoder part as the decoder governs the generation, resulting in significant performance sacrifice. To close the accuracy gap, a majority of previous studies aim at improving the modeling of dependencies with more conditional information (Qian et al., 2021; Ghazvininejad et al., 2019). We argue that these research efforts are equivalent to providing better alternative learning signals without changing the NAR models' probabilistic framework. However, most of these methods require a specific modification to the commonly-used Transformer model architecture.

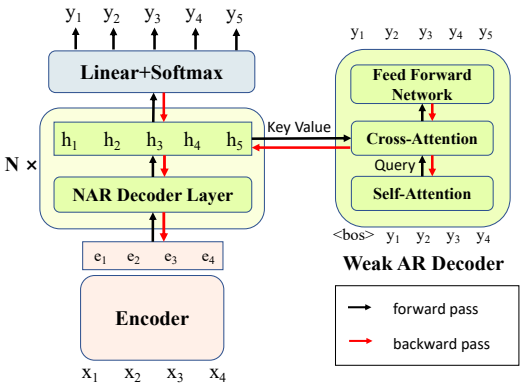

Figure 1: Illustration of our approach, where we introduce a set of auxiliary weak AR decoders, each of which must make its predictions solely relying on the information contained in the NAR decoder hidden states. Thus, the information provided by the NAR decoder must be sufficiently useful for the AR decoders to be capable of predicting the target sequence because the AR decoders are parameterized as weakly as possible, which will in turn let the NAR decoder learn to get stronger.

[1]Code will be released at https://github.com/wxy-nlp/MultiTaskNAT.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 5513 - 5519
December 7-11, 2022

A natural question may arise: can we encourage the NAR decoder to learn from sources of signals that are more informative than that of the conditional independence assumption, in order to better capture target dependencies? It would be more advantageous if it is also modification-free regarding model architectures and could also used with all current NAR systems.

In this paper, we propose a simple multi-task learning framework that introduces auxiliary weak AR decoders to make NAR models stronger. The key idea is that we parameterize the auxiliary AR decoders as weakly as possible and force them to predict target sequences solely based on the information from NAR decoder's hidden representations, such that they can no longer model the target sequence on their own unless the knowledge provided by the NAR decoder is sufficiently useful. As a result, the NAR decoder has no choice but to become stronger so as to support the AR partners that are poorly parameterized. Additionally, our approach is plug-and-play and model-agnostic, and the weak AR decoders that we introduce are discarded during the inference stage, resulting in no additional decoding overhead.

We empirically evaluate its applications to several classes of NAR model, including vanilla NAR Transformer (Gu et al., 2018) and its CTC-based variant (Libovický and Helcl, 2018; Saharia et al., 2020). Experiments on widely-used WMT14 English-to-German, WMT16 English-to-Romanian, and IWSLT14 German-to-English benchmarks show that our approach consistently helps build more accurate NAR models over strong baselines.

2 Preliminary

Neural machine translation (NMT) is formally defined as a conditional probability model $p(\boldsymbol{y}|\boldsymbol{x};\theta)$ parameterized by deep neural networks θ. Given an input sequence $\boldsymbol{x} = (x_1, x_2, \cdots, x_m)$, a neural autoregressive model (Bahdanau et al., 2015; Vaswani et al., 2017) predicts the target sequence $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)$ sequentially based on the conditional distribution, which decomposes $p(\boldsymbol{y}|\boldsymbol{x};\theta)$ by the autoregressive factorization:

$$p_{AR}(\boldsymbol{y}|\boldsymbol{x};\theta) = \prod_{t=1}^{n} p(y_t|\boldsymbol{y}_{<t}, \boldsymbol{x};\theta),$$

where θ is the set of model parameters. Although such factorization achieved great success, its sequential prediction may cause high decoding latency and error accumulation during inference, especially for long sentences.

Non-autoregressive Translation. To solve above problems, Gu et al. (2018) proposed non-autoregressive Transformer based on conditional independence assumption among target tokens, which models $p(\boldsymbol{y}|\boldsymbol{x};\theta)$ in a per-token factorization:

$$p_{NAR}(\boldsymbol{y}|\boldsymbol{x};\theta) = \prod_{t=1}^{n} p(y_t|\boldsymbol{x};\theta).$$

As a result, NAR models can boost up inference by predicting target words simultaneously, thereby improving the efficiency significantly.

However, as noted in Gu et al. (2018), the target-side conditional independence assumption prohibits NAR models from capturing complex dependencies among target tokens, thereby significantly hurting accuracy. To mitigate this, a line of work proposes to modify the training objective (Libovický and Helcl, 2018; Wang et al., 2019; Shao et al., 2020; Ghazvininejad et al., 2020; Qian et al., 2021; Du et al., 2021), while other work uses latent variable to enhance modeling (Kaiser et al., 2018; Shu et al., 2020; Bao et al., 2021, 2022). Besides, several research proposes iterative-based models, which perform iterative refinement of translations based on previous predictions (Lee et al., 2018; Ghazvininejad et al., 2019; Gu et al., 2019; Kasai et al., 2020). The most related work to this paper is Hao et al. (2021), which shows that utilizing an additional AR decoder could help the encoder of NAR models contain more linguistic knowledge.

3 Methodology

In this section, we will dive deep into our simple yet effective multi-task learning framework, including model architecture and training scheme.

Model Architecture. The overall illustration of our approach is depicted in Figure 1. Specifically, for every NAR decoder layer, we introduce an auxiliary weak AR decoder, where each AR decoder is parameterized by one Transformer layer, being as weak as possible. In this case, these AR decoders will no longer capture the underlying structure of target sequences on their own, unless their NAR decoder layers can provide useful neural representations. As a result, the NAR decoder layers can additionally learn from such informative task signals and become stronger to support the weak AR

partners, being forced to capture sufficient context and dependency information.

Training Objective. Our training objective composes two parts for the NAR model of interest and the auxiliary weak AR decoders, respectively. For the NAR part, we keep the original model-specific training objective unaltered. For instance, we apply CTC loss for CTC-based NAR models (Saharia et al., 2020). As for the AR decoders, we apply the cross-entropy loss for training. The final loss is a weighted sum of the two components:

$$\mathcal{L} = \lambda \mathcal{L}_{NAR} + (1 - \lambda) \sum_{i}^{N} \mathcal{L}_{AR}^{(i)},$$

where N is the number of NAR decoder layers, and $\mathcal{L}_{NAR}$ and $\mathcal{L}_{AR}$ represent the NAR loss and AR loss, respectively. The λ is a predefined weight.

Glancing Training. According to previous studies, glancing training (Qian et al., 2021) can considerably improve the translation quality of non-iterative NAR models. We apply glancing training technique to our method. More specifically, we first randomly sample reference tokens as NAR decoder inputs like Qian et al. (2021), and then let the weak AR decoder make predictions based on the NAR decoder hidden states.

Minimizing Training Cost. The major challenge of our method is additional training computational and memory overhead. To this end, we employ two techniques to reduce training costs:

(a) Parameter-sharing of AR decoders. As all AR decoders are homogeneous, we can tie their parameters to reduce the total number of parameters.

(b) Layer dropout for AR decoders. Simultaneously enabling every NAR decoder layer to pair its AR decoder partner is fairly inefficient. To this end, we randomly select half of the AR decoders, instead of all, for multi-task learning.

Both strategies help make the training cost affordable without losing accuracy gains.

Inference. We only use the NAR decoder for inference without any AR decoders. The AR decoder is only used for training. Therefore, our approach has no additional decoding overhead.

4 Experiments

Experimental Settings. We conduct experiments on the most widely used machine translation benchmarks: WMT14 English-German (WMT14 EN-DE, 4.5M translation pairs), WMT16 English-Romanian (WMT16 EN-RO, 610K translation pairs) and IWSLT14 German-English (IWSLT14 DE-EN, 160K translation pairs). We follow Gu and Kong (2021) for data preprocessing and use BLEU (Papineni et al., 2002) as the evaluation metric. To alleviate training difficulties, we use sequence-level knowledge distillation (Hinton et al., 2015) for all datasets to alleviate multi-modality problem as in Gu et al. (2018).

4.1 Main Results

Our approach achieves superior results compared to existing strong NAR systems. Table 1 presents our main results on the benchmarks. As seen, our method significantly improves the translation quality and outperforms other strong baseline models. Besides, when applying the glancing training technique, our method can result in further advancements. Compared with CMLM, which employs iterative decoding, our model can achieve higher performance, while using single-step generation. Hao et al. (2021)'s work is related to ours, which also utilizes a multi-task framework. We reproduce their method on the CTC-based NAR model, and results show that our method can achieve greater improvements. Compared with the strong autoregressive teacher Transformer (Vaswani et al., 2017), our model can further close the performance gap. And when decoding using beam search, our method can outperform Transformer on each dataset.

Our model-agnostic approach can help boost several classes of NAR models. We use Vanilla-NAR (Gu et al., 2018) and CTC (Saharia et al., 2020) models as baselines and apply our multi-task learning approach to each baseline model. The result is shown in Table 2. It can be seen that our method consistently and significantly improves the translation quality for each baseline model and each language pair. This illustrates the generality of our method.

4.2 Analysis

Does AR decoders being weak really matter? Recall that we let AR decoder be sufficiently weak to force NAR decoder to be strong. But how does the capacity of AR decoders affect the efficacy of our approach? We hence conduct experiments with the different number of AR decoder layers. e.g., 1, 3, and 6. As demonstrated in Figure 2, each

Model	WMT14		WMT16		IWSLT14
	EN-DE	DE-EN	EN-RO	RO-EN	DE-EN
Vanilla-NAR (Gu et al., 2018)	17.69	21.47	27.29	29.06	/
CMLM_1 (Ghazvininejad et al., 2019)	18.05	21.83	27.32	28.20	/
Flowseq (Ma et al., 2019)	23.72	28.39	29.73	30.72	27.55
NAR-DCRF (Sun et al., 2019)	23.44	27.22	/	/	27.44
CTC (Saharia et al., 2020)	25.7	28.1	32.2	31.6	/
AXE (Ghazvininejad et al., 2020)	23.5	27.9	30.75	31.54	/
O_AXE (Du et al., 2021)	26.1	30.2	32.4	33.3	/
CNAT (Bao et al., 2021)	25.56	29.36	/	/	31.15
GLAT (Qian et al., 2021)	25.21	29.84	31.19	32.04	/
GLAT+CTC (Qian et al., 2021)	26.39	29.54	32.79	33.84	/
DSLP (Huang et al., 2022)	27.02	31.61	34.17	34.60	/
CMLM_{10} (Ghazvininejad et al., 2019)	27.03	30.53	33.08	33.08	/
CMLM_{10}+MTL (Hao et al., 2021)	**27.98**	31.27	33.80	33.60	/
Transformer (ours)	27.42	31.45	34.11	34.14	35.20
CTC (ours)	26.27	29.60	32.63	33.47	33.91
CTC+MTL (ours)	26.47	30.09	33.35	33.90	34.45
CTC+Our method	26.80	30.36	33.63	34.14	35.13
CTC+Our method & Glancing Training	27.25	30.70	33.88	34.73	35.15
beam search=20	27.75	**31.81**	**34.38**	**35.28**	**36.05**

Table 1: Results of NAR models trained with knowledge distillation on test set of WMT14, WMT16 and IWSLT14. CMLM_k refers to k iterations of decoding.

Model	WMT14		WMT16		IWSLT14
	EN-DE	DE-EN	EN-RO	RO-EN	DE-EN
Vanilla-NAR	17.79	22.02	27.84	29.35	28.32
+ Our method	**21.43**	**25.85**	**29.88**	**30.89**	**32.26**
CTC	26.27	29.60	32.63	33.47	33.91
+ Our method	**26.80**	**30.36**	**33.63**	**34.14**	**35.13**

Table 2: Results of applying our method to different NAR models, showing the generality of our method.

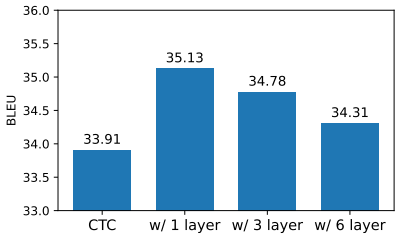

Figure 2: Results on the test of IWSLT14 to analyze the effectiveness of the number of AR decoder layers. We use CTC-based model as baseline, and w/ 1 layer means the AR decoder has 1 layer.

depth AR decoder can bring improvement, but as the number of AR decoder layers increases, the improvement effect for NAR gradually weakens. This verifies our motivation that a weaker AR decoder force NAR decoder to contain more useful information, in turn helping the NAR model.

Ablation study on the training cost optimization. We evaluate the impact of the proposed training

	Param. Sharing	Layer Dropout	training time	#params	BLEU
Ours			31.2h	83.8M	35.15
	✓		31.0h	55.3M	35.10
		✓	20.3h	83.8M	35.07
	✓	✓	**19.4h**	**55.3M**	35.13
CTC (Saharia et al., 2020)			17.3h	50.6M	33.91

Table 3: Study on training cost reduction.

Methods	(0,20]	(20,40]	(40,60]	>60
Transformer	25.58	28.12	27.58	23.42
CTC + Our method	24.77	27.54	27.18	25.07
Gap	-0.81	-0.58	-0.40	**+1.65**

Table 4: Results on the test of WMT14 EN-DE to analyze the performance differences of various target sentence length intervals.

cost reduction strategies. As shown in Table 3, after using the above two techniques, the number of parameters (83.8M vs 55.3M) and training time (31.2h vs 19.4h) are greatly reduced while keeping the model performance almost unchanged.

Our approach helps handle lengthy sentences. To further analyze the performance differences on target sentences of different lengths, we divide the target sentences into buckets of different lengths. As shown in Table 4, as the sentence length in-

Methods	WMT14 EN-DE	IWSLT14 DE-EN
Transformer	0.04%	0.02%
Vanilla-NAR	16.2%	6.94%
+Our method	**6.3%**	**2.90%**
CTC	0.87%	1.41%
+Our method	**0.11%**	**0.18%**

Table 5: Results of repeated token percentage.

Model	WMT14		IWSLT14
	EN-DE	DE-EN	DE-EN
Transformer	27.42	31.45	35.20
Vanilla-NAR	11.02	15.13	17.72
CTC	18.34	23.58	26.77
+ Our method & GLAT	**24.14**	**28.71**	**31.48**

Table 6: Results without knowledge distillation. "GLAT" denotes glancing training.

creases, the performance gap between our model and the Transformer decreases. Remarkably, our model outperforms Transformer when the target sentence length is greater than 60. Longer sentences mean that the model needs to deal with more complex contextual associations. We conjecture that our proposed multi-task training method significantly improves the contextual information contained in the NAR hidden state, and thus has better performance on long sentence translation.

Our approach reduces token repetitions. We also study the rate of repeated tokens as in (Saharia et al., 2020) to see to what extent our approach can tackle the multi-modality problem. Table 5 shows the repetition before and after applying our approach, demonstrating that our method consistently reduces the occurrence of repeated words by a significant margin. Even when equipping CTC alone can alleviate the repetition issue, our approach can give rise to further improvements.

Performance without knowledge distillation. Despite knowledge distillation as a commonly-used workaround, it bounds the performance of NAR models under their AR teacher, along with the extra need to build teacher models. To validate the effectiveness of our method in the raw data scenario, we conduct experiments on the WMT14 and IWSLT14 datasets without knowledge distillation. As shown in Table 6, the baseline CTC model can be significantly enhanced by our approach, further closing the performance gap with the AR model.

Advantages of our method over other multi-task framework. Hao et al. (2021)'s work also utilizes a multi-task framework, and our method can make greater improvements. We attribute this to the location and capacity of our multi-task learning module, i.e. the weak AR decoder. For the location of the AR decoder, we argue that the decoder governs the generation, so placing the AR decoder upon the NAR decoder is supposed to more directly and explicitly improve the generation of NAR, while Hao et al. (2021) is based on the NAR encoder output. For the capacity of the AR decoder, we contend that the AR decoders should be as weak as possible, such that they can no longer model the target sequence on their own unless their NAR decoder layers can provide useful neural representations. In contrast, Hao et al. (2021) do not elaborate on parameterization capacity and use a standard AR decoder.

5 Conclusion

In this paper, we propose a multi-task learning framework for NAR. Along with the training of the weak AR decoder, the NAR hidden state will contain more contextual information, resulting in performance improvement. Experiments on WMT and IWSLT benchmarks show that our method can significantly and consistently improve the translation quality. When using beam search decoding, our CTC-based variant outperforms strong Transformer on all of the benchmarks, while introducing no additional decoding overhead.

Limitations

Our research's potential drawback is that it adds to the training burden. To tackle this problem, we introduce two techniques to reduce training costs. We greatly minimize the number of parameters that should be trained as well as the training time without sacrificing performance. Notably, our method does not introduce additional overhead for inference. Therefore, we can achieve a large performance improvement while maintaining the original fast decoding speed.

Acknowledgements

We would like to thank the anonymous reviewers for their insightful comments. Meanwhile, we also want to thank Yu Bao for his valuable suggestions. Zaixiang Zheng and Shujian Huang are corresponding authors. This work is supported by National Science Foundation of China (No. U1836221, 6217020152).

References

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

Yu Bao, Shujian Huang, Tong Xiao, Dongqi Wang, Xinyu Dai, and Jiajun Chen. 2021. Non-autoregressive translation by learning target categorical codes. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5749–5759, Online. Association for Computational Linguistics.

Yu Bao, Hao Zhou, Shujian Huang, Dongqi Wang, Lihua Qian, Xinyu Dai, Jiajun Chen, and Lei Li. 2022. latent-GLAT: Glancing at latent variables for parallel text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8398–8409, Dublin, Ireland. Association for Computational Linguistics.

Cunxiao Du, Zhaopeng Tu, and Jing Jiang. 2021. Order-agnostic cross entropy for non-autoregressive machine translation. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 2849–2859. PMLR.

Marjan Ghazvininejad, Vladimir Karpukhin, Luke Zettlemoyer, and Omer Levy. 2020. Aligned cross entropy for non-autoregressive machine translation. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 3515–3523. PMLR.

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 (EMNLP-IJCNLP)*, pages 6112–6121, Hong Kong, China. Association for Computational Linguistics.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O. K. Li, and Richard Socher. 2018. Non-autoregressive neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

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 2021*, pages 120–133, Online. Association for Computational Linguistics.

Jiatao Gu, Changhan Wang, and Junbo Zhao. 2019. Levenshtein transformer. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 11179–11189.

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 *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3989–3996, Online. Association for Computational Linguistics.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. 2015. Distilling the knowledge in a neural network. *ArXiv preprint*, abs/1503.02531.

Chenyang Huang, Hao Zhou, Osmar R Zaïane, Lili Mou, and Lei Li. 2022. Non-autoregressive translation with layer-wise prediction and deep supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10776–10784.

Lukasz Kaiser, Samy Bengio, Aurko Roy, Ashish Vaswani, Niki Parmar, Jakob Uszkoreit, and Noam Shazeer. 2018. Fast decoding in sequence models using discrete latent variables. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2395–2404. PMLR.

Jungo Kasai, James Cross, Marjan Ghazvininejad, and Jiatao Gu. 2020. Non-autoregressive machine translation with disentangled context transformer. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 5144–5155. PMLR.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Jason Lee, Elman Mansimov, and Kyunghyun Cho. 2018. Deterministic non-autoregressive neural sequence modeling by iterative refinement. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182, Brussels, Belgium. Association for Computational Linguistics.

Jindřich Libovický and Jindřich Helcl. 2018. End-to-end non-autoregressive neural machine translation with connectionist temporal classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3016–3021, Brussels, Belgium. Association for Computational Linguistics.

Xuezhe Ma, Chunting Zhou, Xian Li, Graham Neubig, and Eduard Hovy. 2019. FlowSeq: Non-autoregressive conditional sequence generation with generative flow. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4282–4292, Hong Kong, China. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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, Online. Association for Computational Linguistics.

Chitwan Saharia, William Chan, Saurabh Saxena, and Mohammad Norouzi. 2020. Non-autoregressive machine translation with latent alignments. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1098–1108, Online. Association for Computational Linguistics.

Chenze Shao, Jinchao Zhang, Yang Feng, Fandong Meng, and Jie Zhou. 2020. Minimizing the bag-of-ngrams difference for non-autoregressive neural machine translation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 198–205. AAAI Press.

Raphael Shu, Jason Lee, Hideki Nakayama, and Kyunghyun Cho. 2020. Latent-variable non-autoregressive neural machine translation with deterministic inference using a delta posterior. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8846–8853. AAAI Press.

Zhiqing Sun, Zhuohan Li, Haoqing Wang, Di He, Zi Lin, and Zhi-Hong Deng. 2019. Fast structured decoding for sequence models. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3011–3020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Yiren Wang, Fei Tian, Di He, Tao Qin, ChengXiang Zhai, and Tie-Yan Liu. 2019. Non-autoregressive machine translation with auxiliary regularization. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 5377–5384. AAAI Press.

Appendix

A Training Hyperparameters

We follow the normal hyperparameters used in NAR works. We design our NAR model with the base setting hyperparameters of Transformer (Vaswani et al., 2017): both the encoder and the decoder has 6 layers, each layers has 8 attention head, and hidden dimension is 512. For the WMT tasks, we train the models with a batch size of 64K tokens and 300K updates. In the case of IWSLT tasks, we use a smaller batch size of 16K tokens, and set the maximum updates to 250K. For regularization, we set the dropout rate to 0.1 for WMT tasks and 0.3 for IWSLT tasks. We use Adam optimizer (Kingma and Ba, 2015) with $\beta = (0.9, 0.999)$. We employ weight decay of 0.01 and label smoothing of 0.1. For the shallow AR decoder, we set the number of decoder layers to 1. We set the hyperparameter λ used in Eq. 1 to 0.5. To obtain robust results, we averaged the last 5 best checkpoints, following Vaswani et al. (2017). All models are implemented on `fairseq` (Ott et al., 2019).