A Survey on Non-Autoregressive Generation for Neural Machine Translation and Beyond

Yisheng Xiao*, Lijun Wu*, Junliang Guo, Juntao Li, Min Zhang, Tao Qin, *Senior Member, IEEE* and Tie-Yan Liu, *Fellow, IEEE*

Abstract—Non-autoregressive (NAR) generation, which is first proposed in neural machine translation (NMT) to speed up inference, has attracted much attention in both machine learning and natural language processing communities. While NAR generation can significantly accelerate inference speed for machine translation, the speedup comes at the cost of sacrificed translation accuracy compared to its counterpart, autoregressive (AR) generation. In recent years, many new models and algorithms have been designed/proposed to bridge the accuracy gap between NAR generation and AR generation. In this paper, we conduct a systematic survey with comparisons and discussions of various non-autoregressive translation (NAT) models from different aspects. Specifically, we categorize the efforts of NAT into several groups, including data manipulation, modeling methods, training criterion, decoding algorithms, and the benefit from pre-trained models. Furthermore, we briefly review other applications of NAR models beyond machine translation, such as grammatical error correction, text summarization, text style transfer, dialogue, semantic parsing, automatic speech recognition, and so on. In addition, we also discuss potential directions for future exploration, including releasing the dependency of KD, reasonable training objectives, pre-training for NAR, and wider applications, etc. We hope this survey can help researchers capture the latest progress in NAR generation, inspire the design of advanced NAR models and algorithms, and enable industry practitioners to choose appropriate solutions for their applications. The web page of this survey is at https://github.com/LitterBrother-Xiao/Overview-of-Non-autoregressive-Applications.

Index Terms—Non-autoregressive, Neural Machine Translation, Transformer, Sequence Generation, Natural Language Processing

1 Introduction

Machine translation [1] is one of the most critical and challenging tasks in natural language processing (NLP), which aims to translate natural language sentences from the source language to the target language. Recently, with the breakthrough of deep learning [2], Neural Machine Translation (NMT) [3], [4], [5], [6], [7], [8], which takes the different neural networks as backbone models, e.g., RNN [3], [9] and CNN [10], [11], has achieved outstanding performances, especially for the self-attention [12] based Transformer [13] models [14], [15]. NMT usually adopts the autoregressive generation (AR) method for translation (AT), which means the target tokens are one-by-one generated in a sequential manner. Therefore, AT is quite time-consuming when generating target sentences, especially for long sentences. To alleviate this problem and accelerate decoding, non-autoregressive generation (NAR) for machine translation (NAT) is first proposed in [16], which can translate/generate all the target tokens in parallel. Therefore, the inference speed is hugely increased, and much attention to NAT/NAR methods has been attracted with impressive progress [17], [18], [19], [20], [21], [22], [23]. However, the translation accuracy is damaged and sacrificed as a result of parallel decoding. Compared with AT, the tokens are generated without internal dependency for NAT models, unlike the AT models where the t-th token has previous $t-1$ contextual token information to help its generation. Hence, the NAT models seriously suffer from lacking target side information to make predictions (e.g., decoding length) and correctly generate target translations. In summary, we attribute the main challenge of NAT models to the *'failure of capturing the target side dependency.'*

To mitigate the above-mentioned challenge, significant efforts have been paid in the past few years from different aspects, e.g., data manipulation [20], [24], modeling methods [18], [25], decoding strategies [26], [27], to better capture the dependency on target side information. Although impressive progress has been achieved and the translation accuracy is greatly improved for NAT models, the translation quality still falls behind their AT counterparts. To continue narrowing the performance gap and facilitating the development of NAT in the future, a solid review of current NAT research is necessary. Therefore, we make the first comprehensive survey of existing non-autoregressive technologies for NMT in this paper. Our review summarizes the core challenge of NAT research and presents various advanced approaches to solve the challenge. Specifically, we introduce the approaches from the following aspects:

- **Data Manipulation.** As a data-driven task, the scale and quality of training data are crucial for NMT tasks. Due to the lack of target dependency for NAT models, lots of methods are proposed to reduce the complexity of the training data to provide an easier training task for them.
- **Modeling.** Various advanced model architectures are proposed to better capture the target dependency, includ-

- *Yisheng Xiao, Juntao Li, and Min Zhang are with Soochow University, Suzhou, China. E-mail: ysxiaoo@stu.suda.edu.cn; ljt@suda.edu.cn; minzhang@suda.edu.cn.*
- *Lijun Wu, Junliang Guo, Tao Qin, and Tie-Yan Liu are with Microsoft Research, Beijing 100080, China. E-mail: lijuwu@microsoft.com; junliangguo@microsoft.com; taoqin@microsoft.com; tyliu@microsoft.com.*

Yisheng Xiao and Lijun Wu contributed equally to this paper. Juntao Li is the corresponding author. *This work is supported by NSFC No. 62206194 and the Beijing Academy of Artificial Intelligence.*

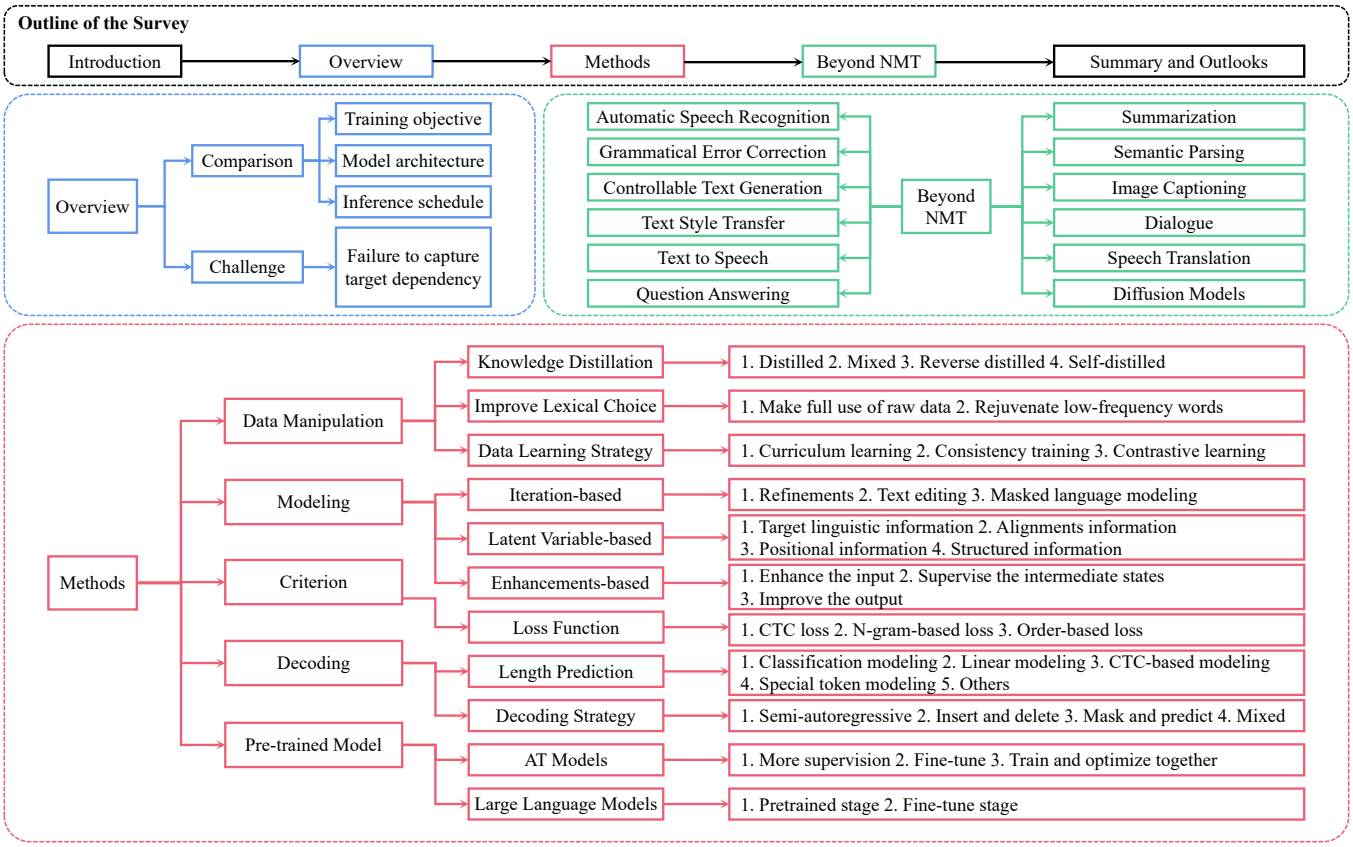


Fig. 1. Outline of the survey. We first review the developments of neural machine translation and non-autoregressive related methods. Then we present an overview of recent AT and NAT models, including a comparison of three different aspects (i.e., training objective, model architecture, and inference schedule). Besides, we also summarize the main challenge that recent NAT models encounter compared with AT models. To solve this challenge, we introduce several widely used methods to help improve the ability of NAT models at different levels, including data manipulation, modeling, criterion, decoding, and benefiting from pre-trained models. Then we present a summary of the above methods for the machine translation task. We also extend this survey to other extensive applications, such as automatic speech recognition, controllable text generation, question answering, image captioning, text summarization, grammatical error correction, text style transfer, dialogue, semantic parsing, text to speech, speech translation and diffusion models. Finally, some open problems and future outlooks are discussed. Best viewed this outline in color.

ing iteration-based methods that can provide partially observed target information, latent variable-based methods that introduce latent variables to learn the target side dependency, and enhancements-based methods that directly provide stronger target side information to the input/output/intermediate states of the decoder.

- **Criterion.** Some works point out that the traditional cross-entropy loss is not optimal for NAT models and propose better criteria to improve the training of NAT models, including Connectionist Temporal Classification (CTC) based, N-gram-based, and order-based loss functions.
- **Decoding.** Decoding algorithm is another decisive factor in NMT models. Upon NAT models, different tricks are proposed to improve the decoding process and provide better translation results.
- **Pre-Trained Model.** Finally, given the strong representation capacity of pre-trained models, it is appealing to utilize them to improve the performance of NAT models. Therefore, lots of methods have been proposed to leverage the information from pre-trained AT models or large-scale language models to help the training of NAT models.

Besides summarizing the improving works for NAT, we also review other applications of NAR methods beyond NMT, such as automatic speech recognition, controllable text generation, question answering, image captioning, text summarization, grammatical error correction, text style transfer, dialogue, semantic parsing, text to speech, speech translation and diffusion models. We further point out the recent trends and possible promising directions for future development, such as releasing the dependency of knowledge distillation, designing more reasonable training objectives, and pre-training for NAR models. We hope this survey paper can provide researchers and engineers with valuable insights to attract more people to either promote the development of NAT/NAR techniques or bring NAT/NAR methods into other fields, such as NAR text generation with large-scale pre-trained language models. Besides, the up-to-the-minute solutions for each NAT problem and thorough analysis of model performance and computational cost are also expected to assist considerable industry practitioners.

The organization of this survey paper is as follows. To begin with, we make a comparison between AT and NAT models from different views, such as their training objectives, model architectures, and inference schedules. Then we analyze the main challenge that the current NAT models encounter in Section 2. We attribute the low quality of NAT models to the failure to well capture the target dependency, and from Section 3 to Section 7, we summarize the efforts

paid for improvement from different aspects, including data manipulations (Section 3), modeling methods (Section 4), training criterion (Section 5), decoding ways (Section 6), and the benefit from pre-trained models (Section 7). In addition, we also give a short summary of current NAT models in Section 8. Next, we investigate the extension of NAR generation approaches to other various applications beyond NMT in Section 9. At last, we conclude this paper and discuss our future outlooks in Section 10. A detailed version of our survey outline is also shown in Figure 1.

2 Overview of AT and NAT Models

In this section, we first give an overall introduction to AT and NAT models. We briefly compare them from several different aspects, including the training objective, model architecture, and inference strategy. Besides, we also analyze the main challenge that NAT models encounter compared with AT.

2.1 Comparison

Both AT and NAT models aim to make a correct sentence translation from a source language to a target language. Due to their unique characteristics, their differences are apparent in training, modeling, and inference. Before a detailed comparison, we first introduce the necessary notations.

Given a dataset $D = \{(X, Y)_i\}_{i=1}^{N}$, where $(X, Y)_i$ refers to a paired sentence data, and N is the size of the dataset. X is the sentence to be translated from source language $\mathcal{X}$ and Y is the ground-truth sentence from target language $\mathcal{Y}$. The goal of NMT models is to learn a mapping function $f(\cdot)$ from the source sentence to the target sentence $f : X \rightarrow Y$ to estimate the unknown conditional distribution $P(Y|X;\theta)$, where θ denotes the parameter set of a network model. We now compare the details of AT and NAT models as below.

Training Objective. (1) For paired sentences (X, Y), where $X = \{x_1, x_2, ..., x_{T_X}\}$ and $Y = \{y_1, y_2, ..., y_{T_Y}\}$, the training objective $\mathcal{L}_{\text{AT}}$ of an autoregressive NMT (AT) model is to maximize the following likelihood:

$$\mathcal{L}_{\text{AT}} = \sum_{t=1}^{T_Y} \log P(y_t|y_{<t}, X; \theta), \tag{1}$$

where y_t is the token to be translated at current time step t and $y_{<t}$ are the tokens predicted in previous $t-1$ decoding steps. From the above equation, we can clearly see that the training of AT models adopts the autoregressive factorization in a left-to-right manner. Note that during training, the ground-truth target tokens are leveraged with the teacher forcing method [13], [28]. In this way, the translation quality is guaranteed with the help of contextual dependencies.

(2) In contrast, the non-autoregressive NMT (NAT) models [16] use the conditional independent factorization for prediction, and the objective is to maximize the likelihood:

$$\mathcal{L}_{\text{NAT}} = \sum_{t=1}^{T} \log P(y_t|X; \theta), \tag{2}$$

notice that T is the length of the target sentence. During training, $T = T_Y$ is the length of the ground-truth target sentence, while in inference, $T = P_L(X)$ which is usually predicted by a length prediction module P_L. Compared with AT models, it is obvious that the conditional tokens $y_{<t}$ are removed for NAT models. Hence, we can do parallel translations without autoregressive dependencies, and the inference speed is greatly improved.

(3) Besides the AT and NAT models, researchers aim to find an intermediate state between current AT and NAT, which can also serve as a universal formulation of both models to achieve a balance between decoding speed and translation quality. For example, Wang *et al.* [17] propose a semi-autoregressive NMT (SAT) model, which keeps the autoregressive property in global but relieves it in local. Shortly speaking, SAT models can produce multiple target tokens in parallel at each decoding step (local non-autoregressive) and dependently generate tokens for the next step (global autoregressive). Mathematically, SAT models aim to maximize the following likelihood:

$$\mathcal{L}_{\text{SAT}} = \sum_{t=1}^{[(T-1)/k]+1} \log P(G_t|G_{<t}, X; \theta), \tag{3}$$

where k denotes the number of the tokes that the SAT models parallelly generate at one time step. G_t is a group of k target tokens at t-th step. $G_{<t}$ is the $t-1$ groups of target tokens generated in the previous $t-1$ decoding steps. Note that if $k = 1$, it equals an AT model, and if $k = T$, it generalizes to a NAT model.

(4) In comparison, iteration-based NAT models share a spirit of mixed autoregressive and non-autoregressive translation, but on the sentence level with a refinement approach. That is, iteration-based NAT models keep the non-autoregressive property in every iteration step and refine the translation results during different iteration steps [18], [29]. The training goal is to maximize:

$$\mathcal{L}_{\text{Iter}} = \sum_{y_t \in Y_{tgt}} \log P(y_t|\hat{Y}, X; \theta), \tag{4}$$

where $\hat{Y}$ indicates the translation result of the last iteration, and Y_{tgt} is the target of this iteration.

In the first iteration, only X is fed into the model, which is the same as NAT models. After that, each iteration takes the translation generated from the last iteration as context for refinement to decode the translation. Generally speaking, NAT models with iterative refinements are viewed as iteration-based NAT models, while models with only one decoding step are viewed as fully NAT models.

Model Architecture. As for model architecture, both AT and NAT models take the encoder and decoder framework for translation. The encoder and decoder can be different neural networks, such as RNN [9], CNN [11], and Transformer [13]. Due to the superior performance of the Transformer network, we focus on the Transformer model for discussion in this survey. The encoder is used to encode the source sentences, while the decoder is utilized for decoding the target sentence. Compared to AT and NAT models, they adopt the same encoder architecture, and the differences are reflected in the decoders to match the specific training objective. (1) Specifically, AT models need to prevent earlier decoding steps from peeking at information from later steps. Therefore, the constraint of an autoregressive factorization of the output distribution is required, and they adopt the strict causal mask by applying a lower triangular matrix in the self-attention module of the conventional Transformer decoder [13]. (2) However, for NAT models, including the iteration-based NAT models, this constraint is no longer necessary, so they adopt the unmasked self-attention over all target tokens [16].

(3) As for SAT models, they adopt a coarse-grained lower triangular matrix as the causal mask, which means that they allow k tokens to peep later information in the same group while keeping the constraint between different groups.

Inference Schedule. When going to the inference stage, the differences are as follows. (1) The AT models predict the target tokens in a one-by-one manner, and the tokens predicted previously are fed back into the decoder to generate the next token. (2) While SAT models predict a group of target tokens at one time, the previously generated groups of tokens are fed into the decoder to generate the next group of tokens, which is the same as the AT models. (3) For iteration-based NAT models, it needs k iterations for inference. The translated results of the previous iteration will be fed into the decoder again for refinements. (4) As for fully NAT models, they generate all predicted target tokens at only one step, which greatly speeds up inference. It is worth noting that AT and SAT models suffer from the gap between training and inference [30], [31], [32]. That is, they utilize ground-truth target tokens during training, while the models can only take previously generated target tokens for inference. This indeed leads to inconsistency between training and inference and hence hurts the performance. In contrast, fully NAT models are free from this trouble, but for iteration-based NAT models, prediction in the previous iteration is adopted for refinements, and this mismatched problem may be more serious. More details about this will be discussed in Section 6.

2.2 The Main Challenge of NAT Models

When achieving parallel decoding, a critical issue of NAT models is that they have no tokens with target information fed into the decoder [16] during training and inference. They can only rely on the source side information, which heavily increases the difficulty for NAT models. Previously, when Gu *et al.* [16] first propose their NAT model, they notice that using nothing or only position embeddings in the first decoder layer results in poor translation performance. To alleviate this problem, they propose an initial module by copying the source tokens as the initialization for the decoder input. However, the source and target sentences from distinct languages are indeed different. This way does not help the decoder since no target information is given.

As a result, missing the target information leads the NAT models *to fail to capture the target dependency of target tokens*, and we attribute the main challenge of low quality for NAT models to this defect. To better understand and further release this problem, we now give specific analysis with examples and also briefly show improvement methods in the following contents.

Understanding the Problem. Since no target information is fed into the decoder, NAT models remove the word dependency of the target sentence completely and generate target tokens entirely depending on the source sentence. Hence, terrible situations can happen to harm the translation quality. (1) First, the conditional independence assumption prevents a model from properly capturing the highly multi-modal distribution of target translations, which is called multi-modality problem [16]. Almost all the NAT models suffer from this trouble. Due to the strong assumption that each target token is predicted independently, if there are several different target sentences that can be viewed as reasonable translations, NAT models are possible to select fragments of each sentence and combine them as a candidate translation. Take an example, when translating `thank you` into German, `Vielen Dank` and `Danke` are both reasonable translations. However, NAT models may generate `Danke Dank`, which is truly unreasonable but should be impossible in AT models. Zhang *et al.* [33] also focus on the multi-modality problem but especially on the syntactic granularity. They first categorize the syntactic multi-modality problem into long-range and short-range types. Then they conduct a systematic study to evaluate the effectiveness of different loss functions for each kind. Finally, they introduce Combined CTC and OAXE (CoCO) loss to alleviate the complicated syntactic multi-modality problem. (2) Over-translation and under-translation [22] are also common translation errors. The issue of over-translation refers to the same word token being successively generated multiple times, leading the same token from different reasonable translations to appear at different positions in the final translation. The under-translation indicates that several necessary tokens in the source sentence are neglected, leading to several tokens missing in the translation results. Take an example, when translating German sentence `es gibt heute viele Farmer mit diesem Ansatz` into English sentence, a reasonable translation can be `there are lots of farmers doing this today`. However, NAT models may miss the word `of` (under-translation) or generate the word `of` twice (over-translation), leading the results to be `there are lots farmers doing this today` or `there are lots of of farmers doing this today`. This seriously harms the translation quality. Instead, if target dependency is given as AT models, the problem of repetitive tokens and missing tokens can be avoided.

2.3 Overview of Improving Methods

As we discussed that NAT models are hard to model the target side dependency, various methods have been proposed to alleviate this problem by reducing the dependency of target tokens at different levels, which hence improves the ability of NAT models. Specifically, these methods include (1) data manipulation, which focuses on the improvements of training data corpus and data learning strategies, (2) improvements on the modeling level, where we first summarize two popular and widely used training frameworks (iteration-based methods and latent variable-based methods) along with various specific implementations of them. Besides, various other enhancements-based methods are introduced for NAT models, (3) improvements on the training criterion, where better criteria compared with traditional cross-entropy loss are proposed to meet the unique characteristics of NAT models, (4) improvements on the decoding level, where we introduce the tremendous progress made on length prediction and decoding strategy, and (5) benefiting from pre-trained models, i.e., guiding NAT models to benefit from their AT counterparts and other large-scale pre-trained language models such as BERT [34]. We plot a figure in the Appendix (Figure 4) to structure these methods better.

In Table 1, we give a summary and overview of different NAT models based on the above improvement category.

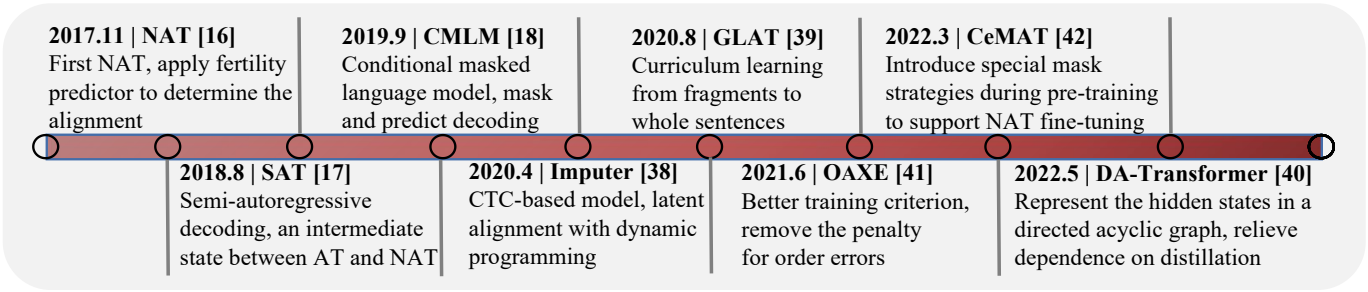


Fig. 2. Representative methods along the development of NAT models.

In each category, we also present the specific sub-topics to better classify the category, along with the representative published works. For example, the decoding strategies can be divided to semi-autoregressive decoding (i.e., SAT [17]), insert and delete (i.e., LevT [35]), mask and predict (i.e., Easy-First [36]), and also mixed decoding (Unified [37]). Besides, for each work, we summarize a short description and list its published place (e.g., ACL, EMNLP), the decoding speed, and the performance on the mostly evaluated dataset WMT14 English→German (EN→DE) for a quick understanding.

Before introducing these methods, in Figure 2, the most important and popular works along the NAT development are shown in the timeline. The NAT is first proposed in November 2017, and the inference speed is hugely improved, but the accuracy is far behind the AT model. After its birth, SAT [17] is proposed to serve as the bridge between AT and NAT models with better translation performance. Other representative works are then introduced, including iteration-based methods: CMLM [18] and Imputer [38], fully NAT models: GLAT [39] and DA-Transformer [40]. These models mainly conduct improvements to the model structure. Besides, improvements based on training criteria are also introduced in OAXE [41]. Recently, CeMAT [42] is proposed to explore the potential of pre-training a non-autoregressive model and then fine-tuning on the translation task. With the rapid growth of NAT models, their performance gap with AT models is narrowing, and the tendency to develop NAT models in real-world systems is increasing. We will elaborate on these methods in the following sections.

3 Data Manipulations

Neural machine translation is a data-driven task, and the performance of the NAT model heavily relies on the volume as well as the quality of the bilingual training data. Therefore, various data manipulation methods are proposed to help the model better capture the target side dependency. In this section, we will introduce these methods from two perspectives: (1) knowledge distillation which aims to reduce the complexity of the training corpus; (2) data learning strategies that help the model better learn and understand the training data. The introduced methods are listed in the "Data Manipulations" category of Table 1.

3.1 Knowledge Distillation

Initially, Knowledge Distillation (KD) [43] is proposed to train a weaker student model with soft prediction distributions generated by a stronger teacher model. Sequence-level knowledge distillation [44] extends it to the sentence level, where a pre-trained teacher model predicts sequences of tokens that are taken as the targets of the student model. When applied to NAT models, a pre-trained AT model is utilized to generate distilled target sentences for all source sentences in the training set, with either greedy decoding or beam search. For example, given a pre-trained AT model θ_{AT} and the training set $D = \{(X, Y)_i\}_{i=1}^{N}$, the distilled target sentences Y' are generated as $Y' \sim P(Y|X; \theta_{\text{AT}})$, where Y' are decoded with various decoding algorithms such as greedy decoding and beam search decoding. Then, we train the NAR models on this distilled training set $D' = \{(X, Y')_i\}_{i=1}^{N}$ with the traditional negative log-likelihood loss:

$$\mathcal{L}_{\text{KD}} = -\sum\nolimits_{(x,y')\in\mathcal{D}'} \log P(y'|x; \theta_{\text{NAT}}), \tag{5}$$

where θ_{NAT} is the parameter set of the NAT model. KD is widely adopted as the distilled corpus is regarded as less noisy and more deterministic than the original one. To investigate the reason behind this, we review related works and give a detailed analysis from two aspects: (1) why is KD effective for NAT models? (2) does there exist drawbacks to current KD methods, and how to solve them?

Understanding Knowledge Distillation. Zhou *et al.* [45] propose two quantitive measures, including the complexity and faithfulness, to analyze the property of distilled data and its correlation with NAT performance. Specifically, they find that while KD generally simplifies the training data by reducing the complexity and increasing the faithfulness, a larger and better teacher model does not always lead to a better student model. Instead, the capacity of the teacher model should be aligned with the student NAT model to achieve the best performance. In addition, Ren *et al.* [46] design a model to measure the target token dependency over the data and find that KD can reduce the dependency when predicting target tokens, which is helpful for the training of NAT. Xu *et al.* [19] find that KD can also reduce the lexical diversity and word reordering degree, which helps the model better learn the alignment between source and target.

Problem and Improvements. Despite the effectiveness, there exist some problems when utilizing KD. Zhou *et al.* [45] find that the capacity of the NAT model should be correlated with the complexity of the distilled dataset. Therefore they propose several methods, including born-again network [47] and mixture-of-experts [48] to adjust the complexity of the dataset w.r.t the model capacity. In addition, after knowledge

distillation, the density estimation of real data may be harmed [49] and the lexical choice may be mistaken [20]. Specifically, Ding *et al.* [20] suppose that the distilled data mainly focuses on the performance of high-frequency words. They propose two evaluation metrics to measure the lexical translation accuracy and conclude that the accuracy of low-frequency words is seriously decreased when the NAT model is trained on distilled datasets. To deal with this problem, Ding *et al.* [50] make full use of raw, distilled, and reverse-distilled data to rejuvenate low-frequency words. Zhou *et al.* [51] add monolingual data for training of the teacher model to enrich the distilled dataset. The effect and improvement of self-distillation are also explored [21]. Shao *et al.* [52] also notice that adopting a single distillation reference from a specific teacher is not optimal for training NAT models and thus propose the diverse distillation with reference selection (DDRS) strategy during training.

3.2 Data Learning Strategies

Aside from constructing informative training datasets, designing suitable learning strategies is another way to improve NAT models. We introduce various data learning strategies in this subsection. Curriculum learning [100] is a machine learning strategy inspired by human learning, which trains the model by feeding training instances in an order (e.g., from easy to hard) instead of randomly. Guo *et al.* [24] introduce the idea of curriculum learning into the training of NAT models by progressively switching the decoder input from AT to NAT to provide a smooth transformation between two training strategies. Liu *et al.* [95] extend this method by designing more fine-grained curriculums. Qian *et al.* [39] propose an adaptive glancing sampling strategy to guide the model to learn from fragments first and then from whole sentences gradually. The ratio of fragments is correlated with the capacity of the model at the current training stage. Bao *et al.* [53] further extend this glancing sampling strategy to a variable-based model. Song *et al.* [101] combine this glancing sampling strategy with a code-switch method for the task of multilingual machine translation. Ding *et al.* [54] divide training data into multiple granularities, such as words, phrases, and sentences, and propose a progressive multi-granularity training strategy to train the model from easy to hard. Apart from curriculum learning, consistency training is an effective method for autoregressive NMT models [102]. For NAT models, Xie *et al.* [55] utilize consistency training to improve the training consistency on different masked sentences. They assumed that the prediction of the same masked position should be consistent in different contexts or with different models. A similar idea is also explored for variational autoencoder-based latent-variable NAT models in recent papers [56], which propose posterior consistency regularization to improve the ability of models. They first apply data augmentation on both source and target sentences twice and then predict the latent variable and regularize these two results. Besides, contrastive learning is also adopted to improve the performance of NAT models [57], which optimizes the similarity of several different representations of the same token in the same sentence, resulting in more informative and robust representations.

4 Modeling

Model structure plays a critical role for NAT models to better capture the target side dependency. This section first introduces two popular frameworks for NAT: iteration-based methods and latent variable-based methods, then we summarize the efforts made on other enhancements-based methods for NAT models. The introduced methods are listed in the "Modeling" category of Table 1. Representative methods are illustrated in Figure 5 of the Appendix.

4.1 Iteration-Based Methods

Iteration-based methods aim to find the trade-off between translation speed and quality. Instead of generating all target tokens in one pass, they learn the conditional distribution over partially observed generated tokens. Lee *et al.* [29] first propose the iterative model, they utilize either the output of the previous iteration or the noised target sentence to initial the decoder input for refinements. Besides, iteration-based models can be divided into the following categories:

Text Editing. These methods learn to generate tokens with different atomic operations. Stern *et al.* [58] propose Insertion Transformer which models both what to insert and where to insert relative to the current slot representations via concatenated outputs. They also introduce several order loss functions for training. Chan [87] introduce KERMIT, which is similar to Insertion Transformer but models the joint data distribution and its decompositions. Welleck *et al.* [103] frame the insertion learning problem as an imitation learning problem, in which a generation policy is learned to mimic the actions of an oracle generation policy. They propose an annealed coaching method and a roll-in and roll-out procedure to learn insertion. Besides, Gu *et al.* [104] propose an insertion-based model with Inferred Generation Order (InDIGO), where the generation orders are modeled as latent variables. InDIGO can automatically infer the generation orders by simultaneously predicting a word and its position to be inserted during inference. Deletion operation is introduced in [35]. They propose Levenshtein Transformer (LevT), which adopts dual policy learning for training and three different classifiers to decide where and how many tokens to insert, whether to delete the tokens and predict the tokens. Furthermore, much progress has been made in exploring more potential of the Levenshtein Transformer. Xu *et al.* [105] view translation as a bilingual synchronization task and propose Edit-LevT to explore the potential in a non-autoregressive manner, which adopts Levenshtein Transformer as a backbone model. Later they also propose TM-LevT [106], where Levenshtein Transformer with an additional initial deletion operation is adopted to detect potential irrelevant words in the translation memory. Niwa *et al.* [107] adopt the nearest neighbor as the initial state of the NAR decoder. They introduce NeighborEdit, which retrieves the nearest neighbor of an input sentence and edits it to generate the output sentence. Lu *et al.* [63] propose a more efficient, flexible, and performance insertion-based model, which introduces an insertion-oriented position encoding method and a better algorithm to determine the parallelization of insertion operations.

Masked Language Modeling. Another line of work leverages the success of masked language modeling, initially

TABLE 1
A brief summary and overview of different NAT models discussed in this paper. The numbers of iteration, decoding speedup, and performance are all copied from their original paper. Specifically, "Performance" denotes the BLEU scores on the WMT 14 EN→DE dataset, and "Speedup" refers to the decoding speedup ratio compared with AT model. Note that "*" indicates training with sequence-level knowledge distillation from a big Transformer. The speedup may not be comparable due to their different hardware conditions, and we list them here just for reference. "#" denotes Findings.

Category	Sub-category	Method	Description	Publication	Iteration	Speedup	Performance
Data Manipulations	Improving KD	MD [51]	Add monolingual data, enrich distillation corpus	ACL 2020	1	-	25.73
		RDP [20]	Raw data prior training, improve lexical choice	ICLR 2020	2.5	3.5x	27.80*
		LRF [50]	Add reverse-distill data,rejuvenate low-frequency word	ACL 2021	2.5	3.5x	28.20*
		SDMRT [21]	Self-distillation mixup, pre-rerank and fine-tune training	ARXIV 2021	10	-	27.72*
		DDRS [52]	Diverse distillation with reference selection	NAACL 2022	1	14.7x	27.60
	Learning Strategies	GLAT [39]	Glancing, learn from fragments to whole sentence	ACL 2021	1	15.3x	25.21
		latent-GLAT [53]	Introduce glancing strategy to discrete latent variables	ACL 2022	1	11.3x	26.64
		PMG [54]	Multi-granularity, from words, phrases, sentences gradually	ACL# 2021	3.5	-	27.80*
		MvSR-NAT [55]	Consistency-based, masked token and model level consistency	TASLP 2022	10	3.6x	27.39*
		LaNMT-C [56]	Consistency training for posterior latent variables	NAACL 2022	2	11.0x	26.02
		CCMLM [57]	Contrastive common mask and contrastive dropout for CMLM	EMNLP# 2022	10	-	27.93*
Modeling	Iteration-based methods	NAT-IR [29]	Denoising autoencoders, iterative refinement	EMNLP 2018	10	1.5x	21.61
		Insertion Transformer [58]	Insert tokens each iteration, like balanced binary trees	ICML 2019	$\approx log_2(N)$	-	27.41
		CMLM [18]	Masked language model trained with uniform mask strategy	EMNLP 2019	10	1.7x	27.03*
		DisCo [36]	More visible subsets to predict masked tokens	ICML 2019	Adaptive	3.5x	27.34*
		SMART [59]	Introduce correction task during inference for CMLM	ARXIV 2020	10	-	27.65
		JM-NAT [60]	Jointly masked strategy, N-gram level masking in decoder	ACL 2019	10	5.7x	27.69*
		Imputer [38]	Combine conditional masking with CTC	EMNLP 2020	8	3.9x	28.20*
		REWRITENAT [26]	Reviewer and locator, locate the error and rewrite	EMNLP 2021	2.7	3.9x	27.83*
		CMLMC [61]	CMLM with reveal-position and correction function	ICLR 2022	10	-	28.37*
		SUNDAE [62]	Step-unrolled denoising autoencoder	ICLR 2022	16	-	28.46*
		INSNET [63]	Insertion-oriented relative position encoding, insert in each layer	NeurIPS 2022	16.1	3.78x	28.05
	Latent variable -based methods	NAT [16]	Fertility predictor, determine the latent alignments	ICLR 2018	1	15.6x	17.35
		FlowSeq [64]	Generative flow, a powerful mathematical framework	EMNLP 2019	1	1.1x	23.72
		PNAT [65]	Positional predictor, model the position of target tokens	ARXIV 2019	1	7.3x	23.05*
		SynST [66]	Parse decoder, autoregressively predict a chunked parse tree	ACL 2019	$N/6$	4.9x	20.74
		LaNMT [67]	Delta Posterior, continuous latent variables	AAAI 2020	1	6.8x	24.20
		ReorderNAT [68]	Reorder the source sentence into a pseudo-translation	AAAI 2021	1	16.1x	22.79
		AligNART [69]	Aligner module, alignment decomposition strategy	EMNLP 2021	1	13.4x	26.40
		CNAT [70]	Categorical codes, without external syntactic parser	NAACL 2021	1	10.4x	25.56*
		SNAT [71]	Incorporate the explicit syntactic and semantic structures	EACL 2021	1	22.6x	24.64*
		Fully NAT [72]	Several tricks to improve the Fully NAT	ACL# 2021	1	16.5x	27.49*
	Other enhancements -based methods	ENAT [23]	Phrase-table lookup, embedding mapping	AAAI 2019	1	25.3x	20.65
		NAT-REG [22]	Similarity and reconstruction regularization	AAAI 2019	1	27.6x	20.65
		LAVA NAT [73]	Vocabulary attention, reorder prediction labels of a word	ARXIV 2020	1	29.3x	25.72
		CCAN [74]	Context-aware cross-attention, local and global contexts	COLING 2020	10	-	27.50
		DSLP [75]	Deep supervision, additional layer-wise predictions	AAAI 2022	1	14.8x	27.02
		DAD [76]	Decoder Input Transformation, backward dependency modeling	ARXIV 2022	1	14.7x	27.51
		DA-Transformer [40]	Represent the hidden states in a directed acyclic graph	ICML 2022	1	13.9x	27.49
		DA-Transformer Viterbi [77]	Adopt viterbi decoding for DA-Transformer	EMNLP# 2022	1	13.2x	26.89
		FA-DAT [78]	Adopt fuzzy alignments for DA-Transformer	ICLR 2023	1	14.0x	27.53
Criterion	Loss function	CTC [38]	Compute and stores partial log-probability	EMNLP 2020	1	18.7x	25.60
		BoN [79]	N-gram level loss, minimize the Bag-of-Ngrams difference	AAAI 2020	1	10.8x	20.90
		AXE [80]	Aligned cross-entropy, a differentiable dynamic program	ICML 2020	1	15.3x	23.53*
		EISL [81]	Compute the n-gram matching differences, more robust	NAACL 2022	1	-	24.17*
		OAXE [41]	Order-agnostic cross-entropy, hungarian algorithm	ICML 2021	1	15.3x	26.10*
		ngram-OAXE [82]	Ngram-based OAXE, allow reordering between ngram phrases	NAACL 2022	1	15.2x	26.50*
		CoCO [33]	Combine CTC and OAXE loss	NAACL 2022	1	14.2x	27.41
		MgMO [83]	Multi-granularity Metric-based Optimization	EMNLP 2022	1	-	26.40
		NMLA [84]	Non-monotonic latent alignments, bipartite matching and n-gram matching	NeurIPS 2022	1	14.7x	27.57
Decoding	Semi-autoregressive decoding	SAT [17]	Generate muti-tokens at one decoding step	EMNLP 2018	$N/2$	1.5x	26.90
		RecoverSAT [85]	Recover segment, recover mistakes of muti-tokens	ACL 2020	$N/2$	2.2x	27.11
		GAD++ [86]	Collaboration of NAT drafting and AT verification	ARXIV 2022	4.0	3.2x	28.89*
	Insert and delete	KERMIT [87]	Model the joint data distribution, adopt bidirectional fine-tuning	ARXIV 2019	$\approx log_2(N)$	-	28.7
		LevT [35]	Insert and delete tokens during each iteration	NeurIPS 2019	Adaptive	4.0x	27.27
	Mask and predict	Mask-Predict [18]	Mask the tokens with low confidence and predict them in the next iteration	EMNLP 2019	10	1.7x	27.03*
		Easy-First [36]	Update tokens at each position with an easy to hard order	ICML 2019	Adaptive	3.5x	27.34*
	Mixed decoding	Unified [87]	Unified approach, conditional permutation language modeling	COLING 2020	10	-	26.35
		Diformer [88]	Directional transformer, directional embedding and self-attention	EAMT 2022	10	-	27.99
		HRT [89]	Generate discontinuous sequences autoregressly and fill in others in parallel	ARXIV 2022	$N/2+1$	-	28.49*
Benefiting from Pre-trained Models	AT models	imitate-NAT [90]	Imitation learning framework with imitate module	ACL 2019	1	18.6x	22.44*
		NAT-HINT [91]	Hints from the hidden state, constrain attention distributions	EMNLP 2019	1	30.2x	21.11
		ENGINE [92]	Energy-based inference, minimize the AT model's energy	ACL 2020	-	-	-
		EM+ODD [93]	Unified framework, dynamically optimize AT and NAT	ICML 2020	1	16.4x	24.54
		FCL-NAT [24]	Curriculum learning from better-trained state of AT model	AAAI 2020	1	28.9x	21.70
		MULTI-TASK NAT [94]	Shared encoder, dynamically mix two training loss	NAACL 2021	10	-	27.98*
		TCT-NAT [95]	Task-level curriculum learning, from AT to SAT, then to NAT	IJCAI 2021	1	27.6x	21.94
		weak MTL [96]	Multitask learning framework, provide more informative learning signals	EMNLP 2022	1	-	27.25
	Pre-trained language models	AB-Net [97]	Take two different BERT models as the encoder and decoder	NeurIPS 2020	-	2.4x	28.69*
		NAG-BERT [98]	Employ bert as a backbone, add a CRF Layer	EACL 2021	-	-	-
		CeMAT [42]	Aligned code-switching and masking, dynamic dual-masking	ACL 2022	10	-	27.20
		XLM-D [99]	Lightweight yet effective decorator, adapt the XMLR model into NAT models	EMNLP 2022	8	2.8x	29.80

proposed by BERT [34]. Ghazvininejad *et al.* [18] extend it to the conditional masked language model (CMLM) by masking and predicting target tokens during training. Unlike the fixed masking ratio in BERT, CMLM adopts a uniform masking strategy to capture the interdependencies of target tokens during training. Based on this model, several follow-up works are proposed, including: (1) jointly masking tokens [60], where the tokens in the source sentences are also masked; (2) introducing self-review mechanism [108], which applies an AR-decoder to help infuse sequential information; (3) predicting more visible subsets [36], instead of only predicting the masked tokens, the method predicts every target token; (4) introducing self-correction task [61], rather than only predicting the masked tokens, which can learn to correct the unreasonable tokens generated by inputting a fully masked sequence.

4.2 Latent Variable-Based Methods.

Utilizing latent variables as part of the model is also a popular method to reduce the target side dependency. Latent variable models maximize the following likelihood:

$$\mathcal{L}_{\text{Lat}} = \sum_{t=1}^{T} \log p(Z|X;\theta)p(y_t|Z,X;\theta), \tag{6}$$

where Z is a specific latent variable. The latent variable-based NAT models first predict a latent variable sequence, where each variable may be a chunk of words or include some other prompt information. Existing works mainly apply latent variables to capture the following information.

Prior Target Linguistic Information. Ma *et al.* [64] utilize a powerful mathematical framework called generative flow.

Variational auto-encoders (VAE) based methods are also applied to model the dependency [109]. Shu *et al.* [67] and Lee *et al.* [49] model the latent variables as spherical Gaussian for every token in the encoder. Bao *et al.* [53] utilize a glancing sampling strategy to optimize latent variables.

Alignments between Source and Target Sentences. Gu *et al.* [16] pre-define the latent variable Z as fertility and use it to determine how many target words every source word is aligned to. Song *et al.* [69] predict the alignment by an aligner module as the latent variable Z.

Position Information of Target Tokens. Bao *et al.* [65] propose PNAT, which depends on the part of an extra positional predictor module to achieve the permutation Z. Ran *et al.* [68] propose ReorderNAT, a novel NAT framework that reorders the source sentence by the target word order to help the decision of word positions.

Syntactic Information of Target Sentence. Syntactic labels represent the sentence structure, which can be utilized to guide the generation and arrangement of target tokens. Akoury *et al.* [66] first introduce syntactic labels as a supervision to help the learning of discrete latent variables. However, the method needs an external syntactic parser to produce the syntactic reference tree, which is effective only in limited scenarios. To release the limitation, Bao *et al.* [70] propose to learn a set of latent codes that act like the syntactic label. Liu *et al.* [71] incorporate the explicit syntactic and semantic structures to improve the ability of NAT models. Specifically, they utilize Part of Speech (POS) and Named Entity Recognition (NER) to introduce these information.

4.3 Other Enhancements-based Methods

In addition to the above two popular frameworks for NAT models, many efforts have been made to improve the ability of capturing the target side dependency for NAT models at different stages, and the corresponding module is also added to their models. We summarize these methods into the following categories.

Enhancing the Input of Decoder. Since copying the source sentence to initial the decoder cannot offer any target information [16], Guo *et al.* [23] propose phrase-table lookup and embedding mapping methods to enhance the input of the decoder, which can feed tokens with some target information into the decoder, then help model learn the training data better. While the used phrase table is trained in advance, embedding mapping drew lessons on the idea of adversarial training and can perform word-level constraints to close the input and target sentence. Zhan *et al.* [76] also focus on the input of the decoder. They propose decoder input transformation, which transforms the decoder input into the target space. Then this can close the input and target side embedding and help capture the target side dependency.

Supervising the Intermediate States. Several works give extra guidance to the decoder module. Firstly, additional attention modules are applied to learn more information. Li *et al.* [73] propose the Vocabulary Attention (VA) mechanism along with the Look-Around (LA) strategy to help the model capture long-term token dependencies of the target sentence. Ding *et al.* [74] propose a context-aware cross-attention module that focuses on both local and global contexts simultaneously and therefore enhances the supervision signal of neighbor tokens as well as the information provided by the source texts. Besides, Huang *et al.* [75] provide layer-wise supervision to the intermediate states of each decoder layer.

Improving the Output of Decoder. For the output of the decoder, Wang *et al.* [22] regularize the learning of the decoder representations by introducing similarity and reconstruction regularizations, where the former aims at avoiding similar hidden states to alleviate the repetitive translation problem, and the latter constraints the results to help address the problem of incomplete translations. Besides, Ran *et al.* [85] propose the RecoverSAT model to recover from repetitive and missing token errors by dynamically determining the length of segments that need to recover and then deleting repetitive segments. Huang [40] propose Directed Acyclic TransfoRmer (DA-Transformer), which represents the hidden states in a Directed Acyclic Graph (DAG), and each path of the DAG denotes a specific translation. This method dramatically helps capture the dependency of target tokens. Recently, more enhancing methods based on DA-Transformer have also been proposed [77], [78]. Shao *et al.* [110] introduce a rephraser to provide a better training target for NAT models. They apply reinforcement learning to obtain a good rephraser and then train NAT models based on the rephraser output.

5 Criterion

In addition to training data and model structure, training criterion is always another decisive factor for the success of neural network models. Most NMT models apply cross-entropy (CE) loss as their training criterion:

$$\mathcal{L}_{\text{CE}} = -\sum_{t=1}^{T} \log P(y_t|X;\theta), \tag{7}$$

where each $P(y_t|X;\theta)$ is calculated conditional independently by the NAT model with parameters θ. However, several researchers have pointed out that the traditional CE loss may not be optimal for NAT models and they propose better criteria to improve the performance of NAT models. This section compares these criteria with traditional CE loss, emphasizes their advantages, and summarizes them into the following categories.

Connectionist Temporal Classification (CTC). CTC based criteria [27] compute and store partial log-probability summations for all prefixes and suffixes of the output sequence by dynamic programming to alleviate the misalignment problem. Libovicky *et al.* [111] and Shu *et al.* [67] also use CTC loss to marginalize all the monotonic alignments between target and predictions, which can be written as

$$\mathcal{L}_{\text{CTC}} = -\sum_{a\in\beta(y)} \prod_{i} p(a_i|x,\theta)), \tag{8}$$

where a is a possible latent alignment, $\beta(y)$ denotes all possible alignments based on the CTC format. Shao *et al.* [84] further explore non-monotonic latent alignments and propose two matching objectives named bipartite matching and n-gram matching to enhance the training of NAT models.

N-Gram-Based. N-gram-based criteria [79] focus on n-gram level relationships. The word-level CE loss encourages NAT to generate the target tokens without considering the global correctness, which aggravates the weakness in capturing target side dependency. Shao *et al.* [79] propose an n-gram

level loss function to minimize the Bag-of-Ngrams (BoN) difference between the model output and the reference sentence. Guo *et al.* [60] also introduce the n-gram-based dependency of target tokens to alleviate the problem of repetitive translations. N-gram-based loss can be written as:

$$\mathcal{L}_{\text{BoN}} = \frac{\text{BoN-}L1}{2(T-n+1)}, \tag{9}$$

where BoN-$L1$ is the L1 distance between the number of n-grams predicted by the NAT model and that in the reference sentence, which can be calculated as:

$$\text{BoN-}L1 = \sum\nolimits_g |\text{BoN}_\theta(g) - \text{BoN}_Y(g)|, \tag{10}$$

where $g = (g_1, g_2, ..., g_n)$ is a possible n-gram set. $\text{BoN}_Y(g) = \sum_{t=0}^{T-n} 1\{y_{t+1:t+n} = g\}$ is the number of occurrences of g in sentence Y. $\text{BoN}_\theta(g)$ denotes the BoN for a NAT model with parameters α, which can be written as:

$$\begin{aligned}\text{BoN}_\theta(g) &= \sum\nolimits_Y P(Y|X,\theta) * \text{BoN}_Y(g) \\ &= \sum\nolimits_Y P(Y|X,\theta) * \sum\nolimits_{t=0}^{T-n} 1\{y_{t+1:t+n} = g\} \\ &= \sum\nolimits_{t=0}^{T-n} \prod\nolimits_{i=1}^{n} P(y_{t+i=g_i|X,\theta})\end{aligned} \tag{11}$$

where X and Y denote the source and target sentences, respectively. Liu *et al.* [81] propose a novel Edit-Invariant Sequence Loss (EISL) which focuses on the n-gram matching to make the model perform more robustly when encountering inconsistent sequence order of source and target. They show that NAT benefits from this loss since the vanilla NAT model is struggling to model flexible generation order.

Order-Based. CE loss is sensitive to any inconsistent alignments between the prediction and target, which leads to penalizing a reasonable translation if it only mismatches the positions of target tokens. To soften the penalty for word order errors, Ghazvininejad *et al.* [80] propose aligned cross-entropy (AXE) loss, which uses a differentiable dynamic programming method to determine loss based on the best possible monotonic alignment between the ground-truth and the model predictions. The AXE loss is calculated as:

$$\mathcal{L}_{\text{AXE}} = -\sum\nolimits_{t=1}^{T} \log P_\alpha(y_t|X;\theta) - \sum\nolimits_{k \notin \theta} P_k(\epsilon), \tag{12}$$

where the first term indicates the aligned cross-entropy loss function between the target tokens and predictions, and the second term penalizes the unaligned predictions. Besides, Du *et al.* [41] further propose the order-agnostic cross-entropy (OAXE) loss, which applies the Hungarian algorithm to find the best possible alignment. The OAXE loss almost removes the penalty for order errors and guides NAT models to focus on lexical matching. Du *et al.* [82] also extend OAXE loss by allowing reordering between n-gram phrases but maintaining a strict match of word order within phrases. Li *et al.* [83] additionally learn to alleviate the constraint of strict match between the hypothesis and the reference tokens. They propose Multi-granularity Metric-based Optimization (MgMO) to collect model behaviors on varied granularity of translation segments and use the feedback for back-propagation.

Given a parallel training sample (X, Y), we can define the alignment between a model prediction $\hat{Y} = \{\hat{y}_1, \hat{y}_2, ..., \hat{y}_{T_{\hat{Y}}}\}$ and a target sentence $Y = \{y_1, y_2, ..., y_{T_Y}\}$ as an ordering of the set of target tokens Y, e.g., $O^i = \{y_{T_Y}, y_1, ..., y_{T_Y-1}\}$ denotes that tokens $\hat{y}_1, \hat{y}_2, ..., \hat{y}_{T_{\hat{Y}}}$ in model prediction $\hat{Y}$ are aligned with tokens $y_{T_Y}, y_1, ..., y_{T_Y-1}$ in target sentence Y respectively. Note that during training, $T_Y = T_{\hat{Y}}$. For each target sentence, we can get $T_Y!$ monotonic alignments. Based on each alignment state O^i, the corresponding CE loss can be calculated as $\mathcal{L}_{O^i} = -\log P(O^i|X;\theta)$. Given all possible alignment states $O = \{O^1, O^2, ..., O^{T_Y!}\}$, the OAXE objective is defined as finding the best alignment O^i to minimize:

$$\mathcal{L}_{\text{OAXE}} = \arg\min_{O^i \in O}(\mathcal{L}_{O^i}) \tag{13}$$

where $-\log P(O^i|X;\theta)$ indicates the CE loss with ordering O^i. The above methods are listed in the "Criterion" category of Table 1 and exemplified in Figure 6 of the Appendix.

6 Decoding

The decoding stage is also crucial for neural machine translation models. Some works try to improve the NAT decoding schedule by applying different tricks. As mentioned in section 2.1, NAT models need to know the target length to guide decoding. And after the length is predicted, different decoding schedules are adopted to improve decoding. In this section, we will introduce various length prediction methods and decoding strategies.

6.1 Length Prediction

In AT models, the beginning and end of decoding are controlled by special tokens, including `[BOS]` (beginning of a sentence) and `[EOS]` (end of a sentence), which implicitly determine the target length during decoding. However, as all target tokens are generated in parallel in NAT models, there is no such special token or target information to guide the termination of decoding. NAT models must know the target length in advance and then generate the content based on it. Therefore, how to predict the correct length of the target sentence is critical for NAT models [112]. Different methods for target length prediction have been proposed.

Length Prediction Mechanism. Length information of target sentence is essential to NAT models as mentioned above. Gu *et al.* [16] propose a fertility predictor to decide how many times the source token will be copied when constructing the decoder input. Then, the sum of fertility numbers could be viewed as the length of the target sentence. Other length prediction methods are also proposed: (1) Classification modeling, which formulates the length prediction as a classification task and utilizes the encoder output to predict the target length or the length difference between the source and target [29], [90]; (2) Linear modeling, Sun *et al.* [25] try to use a linear function such as $T_y = \alpha\ T_x + B$ to directly calculate the target length based on source length; (3) Special token modeling by introducing a special `[LENGTH]` token [18], [39], [92]. Akin to the `[CLS]` token in BERT, the `[LENGTH]` token is usually appended to the encoder input, and the model is trained to predict the length of the target sentence utilizing the hidden output of the `[LENGTH]` token; (4) CTC-based modeling, several models implicitly determine the target length from the word alignment information [38], [111] based on the connectionist temporal classification (CTC) [27] results.

Length Prediction Improvements. Inevitably, there is a deviation between the predicted length and the true length. To release the inherent uncertainty of the data itself, length parallel decoding (LPD) [23], [29] and noise parallel decoding (NPD) [16], [18] are widely utilized during inference. (1) LPD is often used in classification-based models. Once the length T is determined, they choose an LPD window m and then obtain multiple translation results with lengths in the range $[T-m, T+m]$. A pre-trained autoregressive model is then used to score and select the best overall translation. (2) Models that adopt NPD choose the top m lengths with the highest length prediction probability and return the translation candidate with the highest log probabilities on the average of all tokens.

6.2 Decoding Strategy

Fully NAT models adopt only one-step decoding, which can greatly speed up decoding but fail to achieve high-quality translation. As shown in Table 1, the performance of iteration-based models is generally better than that of fully NAT models, indicating that NAT models fail to capture the target side dependency correctly with only one-step decoding.

Semi-Autoregressive Decoding. Semi-autoregressive decoding is adopted for SAT models, which generates multiple target tokens at one decoding step. This decoding manner does not remove the dependency of target tokens completely. Several methods are proposed based on the semi-autoregressive decoding manner, such as: (1) Syntactic labels based [66], which applies a syntactic parser to produce the syntactic reference tree for the tokens in the current decoding step, then a group of tokens with a close syntactic relationship will be generated at one step. (2) Recover mechanism [85], which aims to alleviate the multi-modality problem by introducing a recovered segment. Once a group of tokens is generated, the model will recover from missing and repetitive token errors. (3) Aggressive decoding [86], which first aggressively decodes several tokens as a draft in a non-autoregressive manner and then verifies them in an autoregressive manner. This method can improve the translation quality and lower the latency as the drafting and verification can execute in parallel.

Insert and Delete. Insertion-based decoding methods aim to insert tokens during each decoding step. Many works explore this method based on different generation orders, including uniform [103], random [104], or balanced binary trees [58], [87]. More exploration of insertion orders can be found in [113]. Besides, several advanced methods are proposed: (1) introducing deletion operations [35], which also allows the model to delete the unreasonable tokens during each decoding step; (2) adaptive parallelization of insertions [63], where parallel insertion is conducted between each decoder layer to better improve decoding efficiency.

Mask and Predict. Ghazvininejad *et al.* [18] first propose a mask-predict algorithm. Starting from a fully masked sequence, the model aims to predict the masked tokens during each iteration. Then a fraction of target tokens with low prediction probability will be masked again and fed to the decoder for the next iteration. Many related works try to improve the performance of this decoding method: (1) easy-first policy [36], which modifies the mask prediction algorithm by updating each position with an easy to hard order given the prediction probability of previous iterations; (2) correcting the unmasked tokens [59], [61], where the unmasked tokens are self-corrected during each iteration; (3) substitutive masking strategies [114], where several strategies are proposed to explore the effect of numbers and rules of masking tokens; (4) utilizing the locator module [26], which also focuses on the importance of determining the tokens replaced by `[mask]` tokens in the next iteration and transforms it into a binary classification problem. Inspired by the beam search algorithm for the CTC-based model, Kasner *et al.* [115] apply beam search and employ additional features in its scoring model to improve the fluency of NAT.

Mixed Decoding. Since different types of decoding strategies have been proposed for NAT, several works aim to combine these decoding strategies into a unified model [37], [88]. Tian *et al.* [37] propose a unified approach for machine translation that supports autoregressive, semi-autoregressive, and iterative decoding methods. Once the model is trained, any of the above decoding strategies can be applied by repeatedly determining positions and generating tokens on them. Taking a step further, Wang *et al.* [88] propose a directional Transformer, which models the AR and NAR generation with a unified framework by designing a special attention module. Their model supports four decoding strategies and can dynamically select strategies during each iteration. Wang *et al.* [89] combine the strengths of autoregressive and non-autoregressive translation paradigms well. They propose hybridregressive translation method, which first generates discontinuous sequences autoregressively and then fills in all previously skipped tokens in parallel. We depict typical decoding strategies in Figure 7 of the Appendix and use an example to show their differences in the Appendix Figure 8.

7 Benefiting from Pre-trained Models

To improve the performance of NAT models, various methods are proposed to leverage the information from other strong models, such as their AT counterparts and large-scale pre-trained language models. We will introduce these methods in the following content.

7.1 AT Models

Due to the strong performance of AT models, leveraging AT models to help the NAT model training is appealing. But they differ in model structure and decoding strategy. Therefore, different techniques to benefit the NAT model from their AT counterparts are proposed:

Training with the Supervision of AT Models. Wei *et al.* [90] propose a novel imitation learning framework, introducing a better trained AT demonstrator to supervise each decoding state of the NAT model across different times so that the problem of huge search space can be alleviated. Li *et al.* [91] design two kinds of hints from the hidden representation level to regularize the KL-divergence of the encoder-decoder attention between the AT and NAT models, which can help the training of NAT models. Besides, Tu *et al.* [92] propose an energy-based inference network to minimize the energy of AT model and give several methods for relaxing the energy.

Fine-Tuning from AT Models. Guo *et al.* [24] utilize curriculum learning to fine-tune from a better-trained state of AT

models, and two curricula for the decoder input and mask attention are applied. In addition, Liu *et al.* [95] propose task-level curriculum learning to shift the training strategy from AT to SAT gradually, and finally to NAT.

Training with AT Models Together. Sun *et al.* [93] propose a unified Expectation-Maximization (EM) framework. It optimizes both AT and NAT models jointly, which iteratively updates the AT model based on the output of the NAT model and trains the NAT model with the new output of AT model. Besides, Hao *et al.* [94] propose a model with a shared encoder and separated decoders for AT and NAT models. The training for these two models is controlled by different weights to mix two training losses. A similar idea is also adopted in [96], but they introduce a weak AR decoder module that predicts the tokens entirely depending on the information provided by the NAR decoder to improve the capability of the NAR decoder.

7.2 Pre-Trained Language Models

While large-scale pre-trained language models have been proven effective in autoregressive machine translation [116], [117], efforts are also made for non-autoregressive machine translation. Guo *et al.* [97] incorporate BERT into machine translation based on the mask-predict decoding method, which initializes the encoder and decoder with corresponding pre-trained BERT models, and inserts adapter layers into each layer. Su *et al.* [98] employ BERT as the backbone model and add a CRF output layer for better capturing the target side dependency to improve the performance further. Li *et al.* [42] propose a conditional masked language model with an aligned code-switching masking strategy to enhance the cross-lingual ability. The proposed model can be fine-tuned on both NAT and AT tasks with promising performance. Wang *et al.* [99] present XLM-D, where a lightweight yet effective decorator is adopted to adapt the cross-lingual pretraining model (XLMR) into NAT models.

8 Summary of Non-Autoregressive NMT

All of the above techniques can mitigate the challenge of failing to capture the target side dependency more or less by reducing the reliance of NAT models on target tokens. Since NAT models are essentially data-driven, their performance highly depends on data volume, quality, and learning strategies. Thus, data manipulation methods are almost indispensable for existing NAT works. Various KD methods can reduce the complexity of the training corpus, while data learning strategies can facilitate the understanding and learning of training data. Another critical element in capturing the target side dependency is NAT model structure, e.g., iteration-based methods, latent variables, and various add-ons for the decoder module. In addition to data manipulation and model structure, better training criteria are proposed to make up for the deficiency of cross-entropy loss, e.g., leveraging CTC-based criteria to alleviate the misalignment problem, introducing n-gram-based criteria to capture global context other than word-level correctness, and designing order-based criteria to soften the penalty for reasonable translations but with mismatched tokens at the target positions. Since the differences between AT and NAT models are mainly manifested in the decoder part, different improving skills for the NAT decoding mode are also presented. Typical strategies include performing target length prediction to guide the end of decoding and improving the one-step decoding by keeping part of the target side dependency information in semi-autoregressive decoding, providing partial target information in iterative decoding, and exploring their combinations in mixed decoding. Besides, leveraging the information from other strong models can further improve the performance of NAT models, such as utilizing information from their AT counterparts and large-scale pre-trained language models. To help researchers and engineers select appropriate techniques in applications, we also conduct a brief comparison between existing methods on their effectiveness and inference speed in Appendix based on the contents of Figure 9 and Figure 10.

Recent research has also disclosed the potential problems of NAT models. One recent concern is that the decoding speedup ratio mentioned in Table 1 is measured by L_1^{GPU}, i.e., the translation latency of one input sentence by running the model on a single GPU. With the increase of batch size, the superiority of NAT models against AT alternatives on inference efficiency will be shortened [118]. The speedup reported on different hardware architectures also makes the comparison less persuasive. Moreover, the inference speedup of NAT models is generally measured on AT frameworks with a symmetrical encoder and decoder structure with the same number of transformer layers without considering these more efficient AR models with shallow decoders [119]. Besides the lack of evaluation fairness in latency, other defects in translation quality evaluation also exist. Helcl *et al.* [118] point out that the NAT models produce unusual errors that the widely used BLEU [120] metric does not penalize very heavily, leading to spurious comparability in translation quality between AT and NAT models. It is necessary to introduce more feasible metrics [118], [121] such as CHrF [122], ScareBLEU [123], COMET [124] to obtain a reliable and comprehensive evaluation. Last but not least, fine-grained evaluation of translation quality from different aspects, e.g., word repetition, sentence fluency, lexical choice, sentence consistency, and syntactic accuracy, are underexplored but critical to move NAT forward.

9 Extensive Applications Beyond NMT

After seeing the success of non-autoregressive (NAR) techniques on neural machine translation, these strategies are also widely applied to extensive text generation tasks, semantic parsing [125], [126], text to speech [127], [128], etc. In this section, we will conduct a brief discussion about these works.

9.1 Text Generation

The inference efficiency is not only required for neural machine translation but also indispensable for many other text generation tasks [98], [129], [130]. Existing works of introducing NAR techniques into text generation tasks focus on automatic speech recognition [131], [132], [133], text summarization [134], grammatical error correction [135], [136], dialogue [137], [138]. Resembling the encountered challenge of NAT models in Section 2.2, representative works

of non-autoregressive text generation mainly address the problems of missing target side information and length prediction. According to the involved tasks, we structure these works into different groups, including general-purpose NAR methods and typical models for each specific generation task.

General-Purpose NAR Text Generation. Some works aim to design a general NAR method that can support multiple text generation tasks. Su *et al.* [98] employ BERT as the backbone of a NAR generation model for machine translation, sentence compression, and summarization. They add a CRF output layer on the BERT architecture for non-autoregressive tasks. For length prediction, they adopt two special tokens `[eos]` to dynamically guide the end of the generation. They extend the architecture of BERT to capture the target side dependency better and improve the performance further. For length prediction, they propose a simple and elegant decoding mechanism to help the model determine the target length on-the-fly. Jiang *et al.* [130] propose a new paradigm to adopt pre-trained encoders for NAR text generation tasks. They propose a simple and effective iterative training method, MIx Source and pseudo Target (MIST), for the training stage without introducing extra cost during inference. Yang *et al.* [129] attempt to explore the alternatives for KD in text summarization and story generation. They focus on linguistic structure predicted by a Part-of-Speech (POS) predictor to help alleviate the multimodality problem. Mallinson *et al.* [139] propose EdiT5, which decomposes the generation process into three sub-tasks: tagging, re-ordering, and insertion, where the first two adopt a non-autoregressive manner, but the last one uses an autoregressive decoder. They evaluate the performance in sentence fusion, grammatical error correction, and decontextualization. Qi *et al.* [134] explore to design a large-scale pre-trained model that can support different decoding strategies when applied to downstream tasks. Concretely, they leverage different attention mechanisms during the training stage and fine-tuning strategies to adapt from AR to NAR generation. To verify the effectiveness of their model, they evaluate the proposed method for question generation, summarization, and dialogue generation tasks. Moreover, they further introduce a self-paced mixed distillation method [140] to improve the generation capability of BANG. Agrawal *et al.* [141] propose a framework that adopts an imitation learning algorithm for applying NAR models to editing tasks such as controllable text simplification and abstractive summarization. They introduce a roll-in policy and a controllable curriculum to alleviate the mismatching problem between training and inference. Li *et al.* [142] propose ELMER, which introduces a token-level early exit mechanism into NAR models for the first time. They leverage layer permutation language modeling for pre-training, which can achieve substantial performance improvements on text summarization, question generation, and dialogue generation tasks.

Task-Specific NAR Text Generation. Many other works introduce NAR methods for a specific text generation task.

- ***Automatic Speech Recognition.*** Consistent with neural machine translation, automatic speech recognition (ASR) has benefited dramatically from non-autoregressive models. NAR ASR models can significantly speed up the decoding process but also suffer from lower recognition accuracy due to the failure of capturing target side dependency. The difference reflects in the processing unit, which is a unique characteristic in NAR ASR [131]. The models with token-level processing units need length prediction, while models with frame-level need not. Thus, many NAR methods in neural machine translation cannot be directly used for ASR, but require specific modifications and designs, e.g., Iteration-based [143], [144], Audio-CMLM [145], Imputer [146], Mask-CTC [147], and Insertion-based [58], [87] methods. Besides, knowledge distillation is also an effective skill for NAR ASR [148], [149]. Considering that the most widely used CTC method in NAR ASR is under the assumption that there exists strong conditional independence between different token frame predictions, researchers have made considerable efforts to optimize the vanilla CTC-based model [26], [133], [150], [151], [152], [153], [154], [155], [156], [157]. Simultaneously, similar to the NAT method, the NAR ASR model can also benefit from pre-trained models, e.g., BERT [132], [158], [159]. Besides, Higuchi *et al.* [131] carry out a comparative study on NAR ASR to better understand this task. Recently, error correction for NAR ASR has also been explored [160], [161].
- ***Summarization.*** The summarization task is less subject to target side dependency modeling than neural machine translation since all the target output information is explicitly or implicitly included in the long text input. As a result, NAR methods for the summarization task mainly alleviate the challenge of length prediction. For instance, a Non-Autoregressive Unsupervised Summarization (NAUS) model has been proposed recently [162], which first performs an edit-based search towards a heuristically defined score and then generates a summary as a pseudo-ground-truth. The authors also propose a length-control decoding approach for better target length prediction. Furthermore, Liu *et al.* [163] propose a Non-Autoregressive summarization model with Character-level length Control (NACC), which extends the length control algorithm to character level and achieves significant performance improvements on several datasets.
- ***Grammatical Error Correction.*** Grammatical Error Correction (GEC) is an important NLP task that can automatically detect and correct grammatical errors within a sentence. As most contents of a sentence are correct and unnecessary to be modified for the GEC task, the problem of lacking target side information can be effectively alleviated. Thus, NAR methods are more feasible for this task. Li *et al.* [136] focus on the variable-length correction scenario for Chinese GEC. They employ BERT to initialize the encoder and add a CRF layer on the initialized encoder, augmented by a focal loss penalty strategy to capture the target side dependency. Besides, Straka *et al.* [135] propose a character-based non-autoregressive GEC approach for Czech, German and Russian languages, which focuses on sub-word errors. Shen *et al.* [164] propose a simple yet effective masking strategy to encourage the model to focus on the correct tokens and thus to better understand the sentence.
- ***Dialogue.*** Dialogue generation has achieved remarkable progress in the last few years, and many methods have been proposed to alleviate the notorious problem of diversity [165]. However, due to their autoregressive generation strategy, these dialogue generation models suffer from low inference efficiency for generating informative responses.

Inspired by the advances made in NAT [25], [65], NAR models are adopted in dialogue generation to lower the inference latency, where the response length is predicted in advance. Han *et al.* [137] apply the NAR model to model the bidirectional conditional dependency between contexts (x) and responses (y). They also point out that NAR models can produce more diverse responses. Zou *et al.* [138] propose a concept-guided non-autoregressive method for open-domain response generation, which customizes the Insertion Transformer to complete response and then facilitates a controllable and coherent dialogue. These NAR models for dialogue generation can significantly improve response generation speed. Besides, NAR methods can improve task-oriented dialogue systems by enhancing the spoken language understanding sub-task [166], [167].

- ***Text Style Transfer.*** Autoregressive models have been widely used in unsupervised text style transfer. Despite their success, they suffer from high inference latency and low content preservation problems. Several works explore non-autoregressive (NAR) decoding to alleviate these problems. Ma *et al.* [168] first directly adapt the common training scheme from the AR counterpart in their NAR method and then propose to enhance the NAR decoding from three perspectives: knowledge distillation, contrastive learning, and iterative decoding. They also explore the potential reasons why these methods can narrow the performance gap with AR models. Huang *et al.* [169] point out that the autoregressive manner might generate some irrelevant words with strong styles and ignore part of the source sentence content. They propose a NAR generator for unsupervised text style transfer (NAST), which effectively avoids irrelevant words by alignment information. NAST can dramatically improve transfer performance with efficient decoding speed.
- ***Controllable Text Generation.*** Controllable Text Generation (CTG) is an emerging area in the field of natural language generation [170]. It aims to generate texts that meet certain controllable constraints as humans wish reliably. These constraints are generally task-specific, and CTG can be exploited in various tasks. A few recent works have begun to explore CTG on NAR models. Agrawal *et al.* [171] introduce a non-autoregressive approach for controllable text simplification, where the model iteratively edits an input sequence and incorporates lexical complexity information into the refinement process to generate simplifications. Iso *et al.* [172] propose AutoTemplate for lexically constrained text generation task, which decomposes the generation process into two steps, template generation and lexicalization, by converting the input and output formats. Li *et al.* [173] apply the diffusion model to six fine-grained controllable tasks. Surprisingly, their method doubles the control success rate of prior methods and is competitive with strong baseline methods that require additional training (fine-tuning). Recently, Kumar *et al.* [174] propose MUCOLA, a sampling strategy that flexibly combines pre-trained language models with differentiable constraints. Evaluation with several CTG tasks proves that MUCOLA can achieve excellent performance on toxicity avoidance, sentiment control, and keyword-guided generation.
- ***Image Captioning.*** Image captioning is the task of generating natural language captions for given images. In recent years, many researchers have brought NAR models into image captioning, and their introduced methods have achieved very appealing progress [175], [176], [177]. Motivated by Levenshtein Transformer [35], Wang *et al.* [178] propose TIger for image captioning, which consists of three modules, i.e., Inserter, $\text{Tagger}_{\text{del}}$ and $\text{Tagger}_{\text{add}}$ to achieve explicit caption editing. Besides, Fei *et al.* [179] customize the shared encoder [94], [180] and a NAR decoder for image captioning to improve the modeling capacity of the AR model. Fei *et al.* [181] also introduce a novel uncertainty-aware framework that leverages an Insertion Transformer-based [58] structure to generate image captions from easy to difficult non-autoregressively and an uncertainty-adaptive beam search technique to speed up the decoding further. Chen *et al.* [182] additionally explore the potential of a fully NAR model for image captioning, in which a Discrete Mode Learning (DML) paradigm is employed to alleviate the mode collapse problem.
- ***Question Answering.*** NAR components also serve a vital role in question answering. To solve the exposure bias problem when using AR models in hybrid tabular-textual question answering, Zhang *et al.* [183] propose a non-autoregressive program generation framework, which can generate complete program tuples in parallel and help address the error accumulation issue, and thus can boost both the performance and efficiency. Wang *et al.* [184] propose Knowledge Enhanced Contrastive Prompt-tuning (KECP) for Extractive Question Answering (EQA). They transform the task into a NAR Masked Language Modeling (MLM) generation problem without additional pre-training stages. Experiments on multiple benchmarks demonstrate the effectiveness of the proposed strategy.

9.2 Semantic Parsing

Compared with the non-autoregressive text generation tasks, non-autoregressive semantic parsing relies more on the length prediction mechanism, in which minor differences can lead to entirely different results. Several NAR models applied to semantic parsing are inspired by CMLM [18] but with better length prediction mechanisms. Babu *et al.* [125] study the potential limitations of the original CMLM when applied for semantic parsing and designed a new LightConv Pointer model to improve it, where the target length is computed by a separate module of multiple layers of CNNs with gated linear units. They also use label smoothing to avoid the easy over-fitting in length prediction. During inference, iterative refinement does not bring many benefits to task-oriented semantic parsing, and thus only one step is applied. Shrivastava *et al.* [126] design Span Pointer Networks based on CMLM with a span prediction mechanism to decide the target length. The length module of semantic parsing merely needs frame syntax to perform span prediction, while text generation requires both syntax and semantics.

9.3 Text to Speech

Significant progress has also been made in the non-autoregressive text to speech (NAR TTS) task. Ren *et al.* [127] point out three main problems in autoregressive TTS compared with the non-autoregressive fashion, i.e., the speed of the inference stage is slow, the generated speech

is not robust, and the generated speech is unable to be controlled. Accordingly, they present a model based on Transformer in a non-autoregressive manner to alleviate the above three problems. Besides, one-to-many (O2M) mapping problem is typical in NAR TTS since differences lie in human speaking greatly. Many other NAR TTS models are also proposed to alleviate this problem and improve speech quality. Peng *et al.* [128] propose ParaNet, which extracts attention from the autoregressive TTS model and then re-defines the alignment. Lu *et al.* [185] apply the variational auto-encoder structure to model the alignment information with a latent variable and further use the attention-based soft alignment strategy. Shah *et al.* [186] propose a NAR model by replacing the attention module of the conventional attention-based TTS model with an external duration model for low-resource and highly expressive speech. Besides, a very deep VAE model with residual attention also benefits the NAR TTS [187]. Notice that the above models may need a teacher model to guide their learning. Lee *et al.* [188] propose a bidirectional inference variational auto-encoder to rely less on the teacher model and meanwhile without decreasing the performance. Since over-smoothing is a severe problem that harms the performance of NAR TTS models, many works focus on alleviating this problem. Ren *et al.* [189] summarize these methods into the two categories, i.e., simplify data distributions [190], [191], which provides more conditional input information, and enhance modeling methods [192], [193], which try to enhance the model capacity to fit the complex data distributions. Ren *et al.* [189] combine these two methods to improve the performance of NAR TTS further. The diversity problem of TTS is also explored in recent work. Bae *et al.* [194] propose a variational autoencoder with the hierarchical and multi-scale structure for NAR TTS (HiMuV-TTS) to improve the diversity of generated speech. As most parallel end-to-end TTS models fail to disentangle general prosody features from the speech, Li *et al.* [195] introduce a cross-utterance conditional VAE (CUC-VAE) system to achieve better naturalness and more prosody diversity. Besides, Liu *et al.* [196] propose Controllable and LOssless Non-autoregressive End-to-end TTS (CLONE) to model the general prosody effectively.

9.4 Speech Translation

Much progress has also been made in speech translation along with the development of NAR ASR models mentioned in section 9.1. Many NAR ASR models are applicable for end-to-end speech translation [197] by completing the automatic speech recognition and machine translation stages simultaneously. Since speech translation resembles text translation, effective strategies applied in text translation are also introduced to speech translation. In seeing the success of connectionist temporal classification (CTC) on machine translation [67], Chuang *et al.* [198] propose CTC-based speech-to-text translation model. They construct an auxiliary speech recognition task based on CTC to further improve performance. Inaguma *et al.* [199] propose Orthros to jointly train the NAR and AR decoders on a shared speech encoder, which is similar to sharing encoder structure in machine translation [94]. Besides, a rescoring mechanism is proposed for Orthros [200], in which an auxiliary shallow

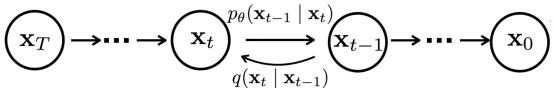


Fig. 3. The forward and reverse process of a diffusion model [205].

AR decoder is introduced to choose the best candidate. On the NAR side, they use CMLM and a CTC-based model as NAR decoders, denoted as Orthros-CMLM and Orthros-CTC, respectively. Such muti-decoder is also widely used for speech translation [201], [202], which is a two-pass decoding method that decomposes the overall task into two sub-tasks, i.e., ASR and machine translation. Inaguma *et al.* [203] propose Fast-MD, where the hidden intermediates are generated in a non-autoregressive manner by a Mask-CTC model. They also introduce a sampling prediction strategy to reduce the mismatched training and testing.

9.5 Diffusion Models

The diffusion model is first proposed in [204], which estimates the data distribution $X_0 \in \mathbb{R}^d$ through a series of latent variables $X_T \cdots X_0$ as Markov chain, with each variable $X_i \in \mathbb{R}^d$ and X_T a Gaussian noise. During training, the diffusion model defines a forward process that constructs the intermediate latent variables $X_1 \cdots X_T$ by incrementally adding Gaussian noise to data X_0 until it turns to approximate a Gaussian at diffusion step T. As shown in Figure 3, $q(X_t|X_{t-1})$ denotes the forward process and $p_\theta(X_{t-1}|X_t)$ refers to the reverse process estimated by p_θ. During inference, diffusion models adopt a non-autoregressive manner to denoise from X_T (a Gaussian) to X_0 (target data). Early research mainly explores their effectiveness on continuous data, such as images and audio generation [205], [206], [207], [208], [209], [210], [211], [212]. Very few works investigate diffusion models with discrete state spaces, such as text generation [212], [213], [214]. Among which, Savinov *et al.* [62] learn from the vanilla diffusion model and propose SUNDAE, a step-unrolled denoising autoencoder that introduces a specific corruption function during training. Unlike previous diffusion models, SUNDAE converges in fewer iterations during inference. More recently, Li *et al.* [173] propose Diffusion-LM, which adopts continuous latent representations and efficient gradient-based methods for controllable text generation. Gong *et al.* [215] further extend the diffusion model to more text generation tasks, which achieves promising performance. Yu *et al.* [216] combine Energy-Based Models (EBMS) and the diffusion model for interpretable text modeling. Reid *et al.* [217] propose DIFFUSER, a new edit-based generative model for text generation based on denoising diffusion models. They also confirm the effectiveness of the diffusion model on machine translation, summarization, and style transfer. There are also existing works that draw the connection between non-autoregressive methods and the diffusion model. In fact, Bert [34] and generative masked language models [18], [218] can also be viewed as diffusion models. They just adopt the different training methods to model the denoising process [209]. Gong *et al.* [215] conduct a theoretical analysis to reveal the connection between their diffusion model DIFFUSEQ and non-autoregressive models. They claim that DIFFUSEQ can be seen as a more generalized form of

an iterative non-autoregressive model. As mentioned in Section 4.1, each denoising process from X_T to X_0 of the diffusion model serves as a decoding iteration.

9.6 Others

In addition to the success of NAR methods on the above-mentioned tasks, many researchers have conducted a pilot study on other scenarios. Information Extraction (IE) also benefits from the non-autoregressive technique. In essence, the facts in plain text are unordered, but the AR models need to predict the following fact conditioned on the previously decoded ones. Yu *et al.* [219] propose a novel non-autoregressive framework, named MacroIE, for OpenIE, which treats IE as a maximal clique discovery problem and predicts the fact set at once to relieve the burden of predicting fact order. For Video Generation (VG), Yu *et al.* [220] propose a dynamics-aware implicit generative adversarial network (DIGAN) for non-autoregressive video generation, which greatly increases inference speed via parallel computing of multiple frames. For Voice Conversion (VC), Hayashi *et al.* [221] extend the FastSpeech2 model in NAR TTS to the voice conversion task and introduce a convolution-augmented Transformer (Conformer). The proposed method can learn both local and global context information of the input sequence and extend variance predictors to variance converters to transpose the prosody components of the source speaker. A fully non-autoregressive many-to-many voice conversion method is also presented in [222], which includes a streaming transformer-based acoustic model and a streaming vocoder. Wang *et al.* [223] propose FastLTS for unconstrained lip-to-speech synthesis. They use a GAN-based vocoder along with adversarial training to improve audio quality and adopt a fully parallelized architecture with a non-autoregressive decoder and vocoder to improve inference efficiency. Besides, self-supervised speech representations are effective in various speech applications. However, existing representation learning methods generally rely on the autoregressive model, leading to low inference efficiency. Liu *et al.* [224] propose Non-Autoregressive Predictive Coding (NPC) to learn speech representations in a non-autoregressive manner by only considering local dependencies of speech, which can significantly improve inference speed. The decoding efficiency of full-line code completion can also benefit from NAR models [225], where a syntax-aware sampling strategy is leveraged to improve the performance. The authors further point out that the dependency on target tokens in code completion is weaker, which is profit for NAR modeling. Barezi *et al.* [226] make an attempt to adopt the NAR model in multi-label learning for extreme classification tasks. Their designed non-autoregressive latent variable model significantly outperforms the autoregressive baselines. Feng *et al.* [227] propose Multi-scale Attention Normalizing Flow(MANF), a novel non-autoregressive deep learning model for time series forecasting tasks. MANF can avoid the influence of cumulative error and meanwhile reduce the time complexity.

10 Conclusion and Outlooks

This paper reviews the development of non-autoregressive methods in neural machine translation and other related tasks. We first summarize the main challenge encountered in NAT research. Then, we structure existing solutions from different perspectives, including data manipulation, modeling, criterion, decoding, and benefiting from pre-trained models, along with a discussion on their effectiveness and inference speed. Besides, we present an overview of the applications of NAR methods in extensive tasks, e.g., summarization, semantic parsing, text to speech, and speech translation. We hope this survey can help researchers and engineers better understand the non-autoregressive techniques and choose suitable strategies for their application tasks.

Although impressive progress has been made on non-autoregressive models, there still exist some open problems:

- KD is the most effective method utilized in NAR models, which depends on pre-training an AR model in advance. However, how to release this condition and improve the performance of NAR models on raw datasets are worthy of further consideration.
- Although iteration-based models have been proposed to help capture the target side contextual dependency in multiple decoding steps and achieved comparable performance with AR models, their speedup w.r.t AR models will diminish when decoding with large batch sizes [118], [119]. Therefore, more attention should be paid to mitigate the challenge mentioned above under the framework of fully NAT models.
- Reasonable training objectives are critical for capturing the target side dependencies for NAR models. Recently, Huang *et al.* [228] point out that simply training NAT models by maximizing the likelihood can lead to an approximation of marginal distributions but drops all dependencies between tokens, and they revisit the previous success (including some advanced criterion introduced in Section 5) in a unified framework. Thus, how to design suitable training objectives is worth further exploration.
- AR models are generally applied to various application scenarios, including bilingual and multilingual, high-resource and low-resource, etc. However, most applications of NAR models are limited to the bilingual scenario until now. Therefore, to expand the impact of NAR models, it is worthy of applying NAR to more application scenarios.
- In recent years, considerable efforts have been made to enhance autoregressive models with powerful pre-training techniques and models, with impressive performance being achieved. However, only very few papers apply these powerful pre-trained models to help NAR models [97], [130], and there is only a preliminary exploration of the pre-training techniques for NAR models [42], [134], [142]. Thus, it is promising to explore pre-training methods for non-autoregressive generation and other related tasks.

References

[1] W. J. Hutchins and H. L. Somers, *An introduction to machine translation*. London: Academic Press, 1992.
[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[3] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *ICLR*, 2015.

[4] K. Cho, B. van Merriënboer, Ç. Gùlçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder–decoder for statistical machine translation," in *EMNLP*, 2014, pp. 1724–1734.

[5] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *NIPS*, 2014, pp. 3104–3112.

[6] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv:1609.08144*, 2016.

[7] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *EMNLP*, 2015, pp. 1412–1421.

[8] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *ACL*, 2016, pp. 1715–1725.

[9] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur, "Recurrent neural network based language model." in *Interspeech*, vol. 2, no. 3, 2010, pp. 1045–1048.

[10] Y. LeCun, Y. Bengio *et al.*, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, pp. 255–258, 1998.

[11] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *ICML*. PMLR, 2017, pp. 1243–1252.

[12] Z. Lin, M. Feng, C. dos Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," in *ICLR*, 2017.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017, pp. 5998–6008.

[14] M. Dehghani, S. Gouws, O. Vinyals, J. Uszkoreit, and L. Kaiser, "Universal transformers," in *ICLR*, 2018.

[15] L. Wu, Y. Wang, Y. Xia, F. Tian, F. Gao, T. Qin, J. Lai, and T.-Y. Liu, "Depth growing for neural machine translation," in *ACL*, 2019, pp. 5558–5563.

[16] J. Gu, J. Bradbury, C. Xiong, V. O. Li, and R. Socher, "Non-autoregressive neural machine translation," in *ICLR*, 2018.

[17] C. Wang, J. Zhang, and H. Chen, "Semi-autoregressive neural machine translation," in *EMNLP*, 2018, pp. 479–488.

[18] M. Ghazvininejad, O. Levy, Y. Liu, and L. Zettlemoyer, "Mask-predict: Parallel decoding of conditional masked language models," in *EMNLP-IJCNLP*, 2019, pp. 6112–6121.

[19] W. Xu, S. Ma, D. Zhang, and M. Carpuat, "How does distilled data complexity impact the quality and confidence of non-autoregressive machine translation?" in *Findings of ACL-IJCNLP*, 2021, pp. 4392–4400.

[20] L. Ding, L. Wang, X. Liu, D. F. Wong, D. Tao, and Z. Tu, "Understanding and improving lexical choice in non-autoregressive translation," in *ICLR*, 2020.

[21] J. Guo, M. Wang, D. Wei, H. Shang, Y. Wang, Z. Li, Z. Yu, Z. Wu, Y. Chen, C. Su *et al.*, "Self-distillation mixup training for non-autoregressive neural machine translation," *arXiv:2112.11640*, 2021.

[22] Y. Wang, F. Tian, D. He, T. Qin, C. Zhai, and T.-Y. Liu, "Non-autoregressive machine translation with auxiliary regularization," in *AAAI*, vol. 33, no. 01, 2019, pp. 5377–5384.

[23] J. Guo, X. Tan, D. He, T. Qin, L. Xu, and T.-Y. Liu, "Non-autoregressive neural machine translation with enhanced decoder input," in *AAAI*, vol. 33, no. 01, 2019, pp. 3723–3730.

[24] J. Guo, X. Tan, L. Xu, T. Qin, E. Chen, and T.-Y. Liu, "Fine-tuning by curriculum learning for non-autoregressive neural machine translation," in *AAAI*, vol. 34, no. 05, 2020, pp. 7839–7846.

[25] Z. Sun, Z. Li, H. Wang, D. He, Z. Lin, and Z.-H. Deng, "Fast structured decoding for sequence models," in *NeurIPS*, vol. 32, 2019, pp. 3016–3026.

[26] X. Geng, X. Feng, and B. Qin, "Learning to rewrite for non-autoregressive neural machine translation," in *EMNLP*, 2021, pp. 3297–3308.

[27] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006, pp. 369–376.

[28] J. F. Kolen and S. C. Kremer, *A field guide to dynamical recurrent networks*. John Wiley & Sons, 2001.

[29] J. Lee, E. Mansimov, and K. Cho, "Deterministic non-autoregressive neural sequence modeling by iterative refinement," in *EMNLP*, 2018, pp. 1173–1182.

[30] W. Zhang, Y. Feng, F. Meng, D. You, and Q. Liu, "Bridging the gap between training and inference for neural machine translation," in *ACL*, 2019, pp. 4334–4343.

[31] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," in *ICLR*, 2016.

[32] L. Wu, F. Tian, T. Qin, J. Lai, and T.-Y. Liu, "A study of reinforcement learning for neural machine translation," in *EMNLP*, 2018, pp. 3612–3621.

[33] K. Zhang, R. Wang, X. Tan, J. Guo, Y. Ren, T. Qin, and T.-Y. Liu, "A study of syntactic multi-modality in non-autoregressive machine translation," in *NAACL-HLT*, 2022, pp. 1747–1757.

[34] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019, pp. 4171–4186.

[35] J. Gu, C. Wang, and J. Zhao, "Levenshtein transformer," in *NeurIPS*, vol. 32, 2019, pp. 11 181–11 191.

[36] J. Kasai, J. Cross, M. Ghazvininejad, and J. Gu, "Non-autoregressive machine translation with disentangled context transformer," in *ICML*. PMLR, 2020, pp. 5144–5155.

[37] C. Tian, Y. Wang, H. Cheng, Y. Lian, and Z. Zhang, "Train once, and decode as you like," in *COLING*, 2020, pp. 280–293.

[38] C. Saharia, W. Chan, S. Saxena, and M. Norouzi, "Non-autoregressive machine translation with latent alignments," in *EMNLP*, 2020, pp. 1098–1108.

[39] L. Qian, H. Zhou, Y. Bao, M. Wang, L. Qiu, W. Zhang, Y. Yu, and L. Li, "Glancing transformer for non-autoregressive neural machine translation," in *ACL-IJCNLP*, 2021, pp. 1993–2003.

[40] F. Huang, H. Zhou, Y. Liu, H. Li, and M. Huang, "Directed acyclic transformer for non-autoregressive machine translation," in *ICML*. PMLR, 2022, pp. 9410–9428.

[41] C. Du, Z. Tu, and J. Jiang, "Order-agnostic cross entropy for non-autoregressive machine translation," in *ICML*. PMLR, 2021, pp. 2849–2859.

[42] P. Li, L. Li, M. Zhang, M. Wu, and Q. Liu, "Universal conditional masked language pre-training for neural machine translation," in *ACL*, 2022, pp. 6379–6391.

[43] G. Hinton, O. Vinyals, J. Dean *et al.*, "Distilling the knowledge in a neural network," *arXiv:1503.02531*, vol. 2, no. 7, 2015.

[44] Y. Kim and A. M. Rush, "Sequence-level knowledge distillation," in *EMNLP*, 2016, pp. 1317–1327.

[45] C. Zhou, J. Gu, and G. Neubig, "Understanding knowledge distillation in non-autoregressive machine translation," in *ICLR*, 2019.

[46] Y. Ren, J. Liu, X. Tan, Z. Zhao, S. Zhao, and T.-Y. Liu, "A study of non-autoregressive model for sequence generation," in *ACL*, 2020, pp. 149–159.

[47] T. Furlanello, Z. Lipton, M. Tschannen, L. Itti, and A. Anandkumar, "Born again neural networks," in *ICML*. PMLR, 2018, pp. 1607–1616.

[48] T. Shen, M. Ott, M. Auli, and M. Ranzato, "Mixture models for diverse machine translation: Tricks of the trade," in *ICML*. PMLR, 2019, pp. 5719–5728.

[49] J. Lee, D. Tran, O. Firat, and K. Cho, "On the discrepancy between density estimation and sequence generation," in *Proceedings of the Fourth Workshop on Structured Prediction for NLP*, 2020, pp. 84–94.

[50] L. Ding, L. Wang, X. Liu, D. F. Wong, D. Tao, and z. Tu, "Rejuvenating low-frequency words: Making the most of parallel data in non-autoregressive translation," in *ACL-IJCNLP*, 2021, pp. 3431–3441.

[51] J. Zhou and P. Keung, "Improving non-autoregressive neural machine translation with monolingual data," in *ACL*, 2020, pp. 1893–1898.

[52] C. Shao, X. Wu, and Y. Feng, "One reference is not enough: Diverse distillation with reference selection for non-autoregressive translation," in *NAACL-HLT*, 2022, pp. 3779–3791.

[53] Y. Bao, H. Zhou, S. Huang, D. Wang, L. Qian, X. Dai, J. Chen, and L. Li, "*latent*-glat: Glancing at latent variables for parallel text generation," in *ACL*, 2022, pp. 8398–8409.

[54] L. Ding, L. Wang, X. Liu, D. F. Wong, D. Tao, and Z. Tu, "Progressive multi-granularity training for non-autoregressive translation," in *Findings of ACL-IJCNLP*, 2021, pp. 2797–2803.

[55] P. Xie, Z. Li, Z. Zhao, J. Liu, and X. Hu, "Mvsr-nat: Multi-view subset regularization for non-autoregressive machine translation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1–10, 2022.

[56] M. Zhu, J. Wang, and C. Yan, "Non-autoregressive neural machine translation with consistency regularization optimized variational framework," in *NAACL-HLT*, 2022, pp. 607–617.

[57] H. Cheng and Z. Zhang, "Con-nat: Contrastive non-autoregressive neural machine translation," in *Findings of EMNLP*, 2022, pp. 6219–6231.

[58] M. Stern, W. Chan, J. Kiros, and J. Uszkoreit, "Insertion transformer: Flexible sequence generation via insertion operations," in *ICML*. PMLR, 2019, pp. 5976–5985.

[59] M. Ghazvininejad, O. Levy, and L. Zettlemoyer, "Semi-autoregressive training improves mask-predict decoding," *arXiv:2001.08785*, 2020.

[60] J. Guo, L. Xu, and E. Chen, "Jointly masked sequence-to-sequence model for non-autoregressive neural machine translation," in *ACL*, 2020, pp. 376–385.

[61] X. S. Huang, F. Perez, and M. Volkovs, "Improving non-autoregressive translation models without distillation," in *ICLR*, 2022.

[62] N. Savinov, J. Chung, M. Binkowski, E. Elsen, and A. van den Oord, "Step-unrolled denoising autoencoders for text generation," in *ICLR*, 2023.

[63] S. Lu, T. Meng, and N. Peng, "Insnet: An efficient, flexible, and performant insertion-based text generation model," in *NeurIPS*, vol. 35, 2022, pp. 7011–7023.

[64] X. Ma, C. Zhou, X. Li, G. Neubig, and E. Hovy, "Flowseq: Non-autoregressive conditional sequence generation with generative flow," in *EMNLP-IJCNLP*, 2019, pp. 4282–4292.

[65] Y. Bao, H. Zhou, J. Feng, M. Wang, S. Huang, J. Chen, and L. Li, "Non-autoregressive transformer by position learning," *arXiv:1911.10677*, 2019.

[66] N. Akoury, K. Krishna, and M. Iyyer, "Syntactically supervised transformers for faster neural machine translation," in *ACL*, 2019, pp. 1269–1281.

[67] R. Shu, J. Lee, H. Nakayama, and K. Cho, "Latent-variable non-autoregressive neural machine translation with deterministic inference using a delta posterior," in *AAAI*, vol. 34, no. 05, 2020, pp. 8846–8853.

[68] Q. Ran, Y. Lin, P. Li, and J. Zhou, "Guiding non-autoregressive neural machine translation decoding with reordering information," in *AAAI*, vol. 35, no. 15, 2021, pp. 13 727–13 735.

[69] J. Song, S. Kim, and S. Yoon, "Alignart: Non-autoregressive neural machine translation by jointly learning to estimate alignment and translate," in *EMNLP*, 2021, pp. 1–14.

[70] Y. Bao, S. Huang, T. Xiao, D. Wang, X. Dai, and J. Chen, "Non-autoregressive translation by learning target categorical codes," in *NAACL-HLT*, 2021, pp. 5749–5759.

[71] Y. Liu, Y. Wan, J. Zhang, W. Zhao, and S. Y. Philip, "Enriching non-autoregressive transformer with syntactic and semantic structures for neural machine translation," in *EACL*, 2021, pp. 1235–1244.

[72] J. Gu and X. Kong, "Fully non-autoregressive neural machine translation: Tricks of the trade," in *Findings of ACL-IJCNLP*, 2021, pp. 120–133.

[73] X. Li, Y. Meng, A. Yuan, F. Wu, and J. Li, "Lava nat: A non-autoregressive translation model with look-around decoding and vocabulary attention," *arXiv:2002.03084*, 2020.

[74] L. Ding, L. Wang, D. Wu, D. Tao, and Z. Tu, "Context-aware cross-attention for non-autoregressive translation," in *COLING*, 2020, pp. 4396–4402.

[75] C. Huang, H. Zhou, O. R. Zaïane, L. Mou, and L. Li, "Non-autoregressive translation with layer-wise prediction and deep supervision," in *AAAI*, vol. 36, no. 10, 2022, pp. 10 776–10 784.

[76] J. Zhan, Q. Chen, B. Chen, W. Wang, Y. Bai, and Y. Gao, "Non-autoregressive translation with dependency-aware decoder," *arXiv:2203.16266*, 2022.

[77] C. Shao, Z. Ma, and Y. Feng, "Viterbi decoding of directed acyclic transformer for non-autoregressive machine translation," in *Findings of EMNLP*, 2022, pp. 4390–4397.

[78] Z. Ma, C. Shao, S. Gui, M. Zhang, and Y. Feng, "Fuzzy alignments in directed acyclic graph for non-autoregressive machine translation," in *ICLR*, 2023.

[79] C. Shao, J. Zhang, Y. Feng, F. Meng, and J. Zhou, "Minimizing the bag-of-ngrams difference for non-autoregressive neural machine translation," in *AAAI*, vol. 34, no. 01, 2020, pp. 198–205.

[80] G. Marjan, V. Karpukhin, L. Zettlemoyer, and O. Levy, "Aligned cross entropy for non-autoregressive machine translation," in *ICML*. PMLR, 2020, pp. 3515–3523.

[81] G. Liu, Z. Yang, T. Tao, X. Liang, J. Bao, Z. Li, X. He, S. Cui, and Z. Hu, "Don't take it literally: An edit-invariant sequence loss for text generation," in *NAACL-HLT*, 2022, pp. 2055–2078.

[82] C. Du, Z. Tu, L. Wang, and J. Jiang, "ngram-oaxe: Phrase-based order-agnostic cross entropy for non-autoregressive machine translation," in *COLING*, 2022, pp. 5035–5045.

[83] Y. Li, L. Cui, Y. Yin, and Y. Zhang, "Multi-granularity optimization for non-autoregressive translation," *arXiv:2210.11017*, 2022.

[84] C. Shao and Y. Feng, "Non-monotonic latent alignments for ctc-based non-autoregressive machine translation," in *NeurIPS*, vol. 35, 2022, pp. 8159–8173.

[85] Q. Ran, Y. Lin, P. Li, and J. Zhou, "Learning to recover from multi-modality errors for non-autoregressive neural machine translation," in *ACL*, 2020, pp. 3059–3069.

[86] H. Xia, T. Ge, F. Wei, and Z. Sui, "Lossless speedup of autoregressive translation with generalized aggressive decoding," *arXiv:2203.16487*, 2022.

[87] W. Chan, N. Kitaev, K. Guu, M. Stern, and J. Uszkoreit, "Kermit: Generative insertion-based modeling for sequences," *arXiv:1906.01604*, 2019.

[88] M. Wang, J. Guo, Y. Wang, D. Wei, H. Shang, Y. Li, C. Su, Y. Chen, M. Zhang, S. Tao *et al.*, "Diformer: Directional transformer for neural machine translation," in *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, 2022, pp. 81–90.

[89] Q. Wang, X. Hu, and M. Chen, "Hybrid-regressive neural machine translation," *arXiv:2210.10416*, 2022.

[90] B. Wei, M. Wang, H. Zhou, J. Lin, and X. Sun, "Imitation learning for non-autoregressive neural machine translation," in *ACL*, 2019, pp. 1304–1312.

[91] Z. Li, Z. Lin, D. He, F. Tian, T. Qin, L. Wang, and T.-Y. Liu, "Hint-based training for non-autoregressive machine translation," in *EMNLP-IJCNLP*, 2019, pp. 5708–5713.

[92] L. Tu, R. Y. Pang, S. Wiseman, and K. Gimpel, "Engine: Energy-based inference networks for non-autoregressive machine translation," in *ACL*, 2020, pp. 2819–2826.

[93] Z. Sun and Y. Yang, "An em approach to non-autoregressive conditional sequence generation," in *ICML*. PMLR, 2020, pp. 9249–9258.

[94] Y. Hao, S. He, W. Jiao, Z. Tu, M. Lyu, and X. Wang, "Multi-task learning with shared encoder for non-autoregressive machine translation," in *NAACL-HLT*, 2021, pp. 3989–3996.

[95] J. Liu, Y. Ren, X. Tan, C. Zhang, T. Qin, Z. Zhao, and T.-Y. Liu, "Task-level curriculum learning for non-autoregressive neural machine translation," in *IJCAI*, 2021, pp. 3861–3867.

[96] X. Wang, Z. Zheng, and S. Huang, "Helping the weak makes you strong: Simple multi-task learning improves non-autoregressive translators," in *EMNLP*, 2022, pp. 5513–5519.

[97] J. Guo, Z. Zhang, L. Xu, H.-R. Wei, B. Chen, and E. Chen, "Incorporating bert into parallel sequence decoding with adapters," in *NeurIPS*, vol. 33, 2020, pp. 10 843–10 854.

[98] Y. Su, D. Cai, Y. Wang, D. Vandyke, S. Baker, P. Li, and N. Collier, "Non-autoregressive text generation with pre-trained language models," in *EACL*, 2021, pp. 234–243.

[99] Y. Wang, S. He, G. Chen, Y. Chen, and D. Jiang, "Xlm-d: Decorate cross-lingual pre-training model as non-autoregressive neural machine translation," in *EMNLP*, 2022, pp. 6934–6946.

[100] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *ICML*, 2009, pp. 41–48.

[101] Z. Song, H. Zhou, L. Qian, J. Xu, S. Cheng, M. Wang, and L. Li, "switch-GLAT: Multilingual parallel machine translation via code-switch decoder," in *ICLR*, 2022.

[102] X. Liang, L. Wu, J. Li, Y. Wang, Q. Meng, T. Qin, W. Chen, M. Zhang, T.-Y. Liu *et al.*, "R-drop: regularized dropout for neural networks," in *NeurIPS*, vol. 34, 2021, pp. 10 890–10 905.

[103] S. Welleck, K. Brantley, H. D. Iii, and K. Cho, "Non-monotonic sequential text generation," in *ICML*. PMLR, 2019, pp. 6716–6726.

[104] J. Gu, Q. Liu, and K. Cho, "Insertion-based decoding with automatically inferred generation order," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 661–676, 2019.

[105] J. Xu, J. Crego, and F. Yvon, "Bilingual synchronization: Restoring translational relationships with editing operations," in *EMNLP*, 2022, pp. 8016–8030.

[106] ——, "Non-autoregressive machine translation with translation memories," *arXiv preprint arXiv:2210.06020*, 2022.

[107] A. Niwa, T. Sho, and O. Naoaki, "Nearest neighbor non-autoregressive text generation," *arXiv:2208.12496*, 2022.

[108] P. Xie, Z. Cui, X. Chen, X. Hu, J. Cui, and B. Wang, "Infusing sequential information into conditional masked translation model with self-review mechanism," in *COLING*, 2020, pp. 15–25.

[109] L. Kaiser, S. Bengio, A. Roy, A. Vaswani, N. Parmar, J. Uszkoreit, and N. Shazeer, "Fast decoding in sequence models using discrete latent variables," in *ICML*. PMLR, 2018, pp. 2390–2399.

[110] C. Shao, J. Zhang, J. Zhou, and Y. Feng, "Rephrasing the reference for non-autoregressive machine translation," *arXiv preprint arXiv:2211.16863*, 2022.

[111] J. Libovickỳ and J. Helcl, "End-to-end non-autoregressive neural machine translation with connectionist temporal classification," in *EMNLP*, 2018, pp. 3016–3021.

[112] M. Wang, G. Jiaxin, Y. Wang, Y. Chen, S. Chang, H. Shang, M. Zhang, S. Tao, and H. Yang, "How length prediction influence the performance of non-autoregressive translation?" in *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 2021, pp. 205–213.

[113] W. Chan, M. Stern, J. Kiros, and J. Uszkoreit, "An empirical study of generation order for machine translation," in *EMNLP*, 2020, pp. 5764–5773.

[114] J. Kreutzer, G. Foster, and C. Cherry, "Inference strategies for machine translation with conditional masking," *arXiv:2010.02352*, 2020.

[115] Z. Kasner, J. Libovickỳ, and J. Helcl, "Improving fluency of non-autoregressive machine translation," *arXiv:2004.03227*, 2020.

[116] J. Zhu, Y. Xia, L. Wu, D. He, T. Qin, W. Zhou, H. Li, and T. Liu, "Incorporating bert into neural machine translation," in *ICLR*, 2020.

[117] Z. Yang, B. Hu, A. Han, S. Huang, and Q. Ju, "Csp: Code-switching pre-training for neural machine translation," in *EMNLP*, 2020, pp. 2624–2636.

[118] J. Helcl, B. Haddow, and A. Birch, "Non-autoregressive machine translation: It's not as fast as it seems," in *NAACL-HLT*, 2022, pp. 1780–1790.

[119] J. Kasai, N. Pappas, H. Peng, J. Cross, and N. Smith, "Deep encoder, shallow decoder: Reevaluating non-autoregressive machine translation," in *ICLR*, 2020.

[120] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.

[121] R. Schmidt, T. Pires, S. Peitz, and J. Lööf, "Non-autoregressive neural machine translation: A call for clarity," in *EMNLP*, 2022, pp. 2785–2799.

[122] M. Popović, "chrf: character n-gram f-score for automatic mt evaluation," in *the Tenth Workshop on SMT*, 2015, pp. 392–395.

[123] M. Post, "A call for clarity in reporting bleu scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers*, 2018, pp. 186–191.

[124] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie, "Comet: A neural framework for mt evaluation," in *EMNLP)*, 2020, pp. 2685–2702.

[125] A. Babu, A. Shrivastava, A. Aghajanyan, A. Aly, A. Fan, and M. Ghazvininejad, "Non-autoregressive semantic parsing for compositional task-oriented dialog," in *NAACL-HLT*, 2021, pp. 2969–2978.

[126] A. Shrivastava, P. Chuang, A. Babu, S. Desai, A. Arora, A. Zotov, and A. Aly, "Span pointer networks for non-autoregressive task-oriented semantic parsing," in *Findings of EMNLP*, 2021, pp. 1873–1886.

[127] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," in *NeurIPS*, vol. 32, 2019.

[128] K. Peng, W. Ping, Z. Song, and K. Zhao, "Non-autoregressive neural text-to-speech," in *ICML*. PMLR, 2020, pp. 7586–7598.

[129] K. Yang, W. Lei, D. Liu, W. Qi, and J. Lv, "Pos-constrained parallel decoding for non-autoregressive generation," in *ACL*, 2021, pp. 5990–6000.

[130] T. Jiang, S. Huang, Z. Zhang, D. Wang, F. Zhuang, F. Wei, H. Huang, L. Zhang, and Q. Zhang, "Improving non-autoregressive generation with mixup training," *arXiv:2110.11115*, 2021.

[131] Y. Higuchi, N. Chen, Y. Fujita, H. Inaguma, T. Komatsu, J. Lee, J. Nozaki, T. Wang, and S. Watanabe, "A comparative study on non-autoregressive modelings for speech-to-text generation," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 47–54.

[132] F.-H. Yu and K.-Y. Chen, "Non-autoregressive transformer-based end-to-end asr using bert," *arXiv:2104.04805*, 2021.

[133] X. Song, Z. Wu, Y. Huang, C. Weng, D. Su, and H. Meng, "Non-autoregressive transformer asr with ctc-enhanced decoder input," in *ICASSP*. IEEE, 2021, pp. 5894–5898.

[134] W. Qi, Y. Gong, J. Jiao, Y. Yan, W. Chen, D. Liu, K. Tang, H. Li, J. Chen, R. Zhang *et al.*, "Bang: Bridging autoregressive and non-autoregressive generation with large scale pretraining," in *ICML*. PMLR, 2021, pp. 8630–8639.

[135] M. Straka, J. Náplava, and J. Straková, "Character transformations for non-autoregressive gec tagging," in *W-NUT*, 2021, pp. 417–422.

[136] P. Li and S. Shi, "Tail-to-tail non-autoregressive sequence prediction for chinese grammatical error correction," in *ACL-IJCNLP*, 2021, pp. 4973–4984.

[137] Q. Han, Y. Meng, F. Wu, and J. Li, "Non-autoregressive neural dialogue generation," *arXiv:2002.04250*, 2020.

[138] Y. Zou, Z. Liu, X. Hu, and Q. Zhang, "Thinking clearly, talking fast: Concept-guided non-autoregressive generation for open-domain dialogue systems," in *EMNLP*, 2021, pp. 2215–2226.

[139] J. Mallinson, J. Adamek, E. Malmi, and A. Severyn, "Edit5: Semi-autoregressive text-editing with t5 warm-start," *arXiv:2205.12209*, 2022.

[140] W. Qi, Y. Gong, Y. Shen, J. Jiao, Y. Yan, H. Li, R. Zhang, W. Chen, and N. Duan, "A self-paced mixed distillation method for non-autoregressive generation," *arXiv:2205.11162*, 2022.

[141] S. Agrawal and M. Carpuat, "An imitation learning curriculum for text editing with non-autoregressive models," in *ACL*, 2022, pp. 7550–7563.

[142] J. Li, T. Tang, W. X. Zhao, J.-Y. Nie, and J.-R. Wen, "ELMER: A non-autoregressive pre-trained language model for efficient and effective text generation," in *EMNLP*, 2022, pp. 1044–1058.

[143] E. A. Chi, J. Salazar, and K. Kirchhoff, "Align-refine: Non-autoregressive speech recognition via iterative realignment," in *NAACL-HLT*, 2021, pp. 1920–1927.

[144] N. Chen, P. Zelasko, L. Moro-Velázquez, J. Villalba, and N. Dehak, "Align-denoise: Single-pass non-autoregressive speech recognition," in *Interspeech*, 2021, pp. 3770–3774.

[145] N. Chen, S. Watanabe, J. Villalba, P. Żelasko, and N. Dehak, "Non-autoregressive transformer for speech recognition," *IEEE Signal Processing Letters*, vol. 28, pp. 121–125, 2020.

[146] W. Chan, C. Saharia, G. Hinton, M. Norouzi, and N. Jaitly, "Imputer: Sequence modelling via imputation and dynamic programming," in *ICML*. PMLR, 2020, pp. 1403–1413.

[147] Y. Higuchi, S. Watanabe, N. Chen, T. Ogawa, and T. Kobayashi, "Mask ctc: Non-autoregressive end-to-end asr with ctc and mask predict," in *Interspeech*, 2020, pp. 3655–3659.

[148] X. Gong, Z. Zhou, and Y. Qian, "Knowledge transfer and distillation from autoregressive to non-autoregressive speech recognition," *arXiv:2207.10600*, 2022.

[149] J. W. Yoon, B. J. Woo, S. Ahn, H. Lee, and N. S. Kim, "Inter-kd: Intermediate knowledge distillation for ctc-based automatic speech recognition," in *SLT*. IEEE, 2023, pp. 280–286.

[150] J. Lee and S. Watanabe, "Intermediate loss regularization for ctc-based speech recognition," in *ICASSP*. IEEE, 2021, pp. 6224–6228.

[151] J. Nozaki and T. Komatsu, "Relaxing the conditional independence assumption of ctc-based asr by conditioning on intermediate predictions," *arXiv:2104.02724*, 2021.

[152] K. Deng, Z. Yang, S. Watanabe, Y. Higuchi, G. Cheng, and P. Zhang, "Improving non-autoregressive end-to-end speech recognition with pre-trained acoustic and language models," in *ICASSP*. IEEE, 2022, pp. 8522–8526.

[153] R. Fan, W. Chu, P. Chang, and J. Xiao, "Cass-nat: Ctc alignment-based single step non-autoregressive transformer for speech recognition," in *ICASSP*. IEEE, 2021, pp. 5889–5893.

[154] Y. Higuchi, H. Inaguma, S. Watanabe, T. Ogawa, and T. Kobayashi, "Improved mask-ctc for non-autoregressive end-to-end asr," in *ICASSP*. IEEE, 2021, pp. 8363–8367.

[155] Y. Yang, Y. Li, and B. Du, "Improving ctc-based asr models with gated interlayer collaboration," *arXiv:2205.12462*, 2022.

[156] K.-H. Lu and K.-Y. Chen, "A context-aware knowledge transferring strategy for ctc-based asr," in *SLT*. IEEE, 2023, pp. 60–67.

[157] S. Dingliwal, M. Sunkara, S. Ronanki, J. Farris, K. Kirchhoff, and S. Bodapati, "Personalization of ctc speech recognition models," in *SLT*. IEEE, 2023, pp. 302–309.

[158] Y. Bai, J. Yi, J. Tao, Z. Tian, Z. Wen, and S. Zhang, "Fast end-to-end speech recognition via non-autoregressive models and cross-modal knowledge transferring from bert," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1897–1911, 2021.

[159] Y. Higuchi, T. Ogawa, T. Kobayashi, and S. Watanabe, "Bectra: Transducer-based end-to-end asr with bert-enhanced encoder," *arXiv:2211.00792*, 2022.

[160] R. Fan, G. Ye, Y. Gaur, and J. Li, "Acoustic-aware non-autoregressive spell correction with mask sample decoding," *arXiv:2210.08665*, 2022.

[161] H. Futami, H. Inaguma, S. Ueno, M. Mimura, S. Sakai, and T. Kawahara, "Non-autoregressive error correction for ctc-based asr with phone-conditioned masked lm," *arXiv:2209.04062*, 2022.

[162] C. H. Puyuan Liu and L. Mou, "Learning non-autoregressive models from search for unsupervised sentence summarization," in *ACL*, 2022, pp. 7916–7929.

[163] P. Liu, X. Zhang, and L. Mou, "A character-level length-control algorithm for non-autoregressive sentence summarization," in *NeurIPS*, vol. 35, 2022, pp. 29 101–29 112.

[164] K. Shen, Y. Leng, X. Tan, S. Tang, Y. Zhang, W. Liu, and E. Lin, "Mask the correct tokens: An embarrassingly simple approach for error correction," *arXiv:2211.13252*, 2022.

[165] I. Kulikov, A. Miller, K. Cho, and J. Weston, "Importance of search and evaluation strategies in neural dialogue modeling," in *INLG*, 2019, pp. 76–87.

[166] L. Cheng, W. Jia, and W. Yang, "An effective non-autoregressive model for spoken language understanding," in *CIKM*, 2021, pp. 241–250.

[167] ——, "Capture salient historical information: A fast and accurate non-autoregressive model for multi-turn spoken language understanding," *ACM Transactions on Information Systems*, vol. 41, no. 2, pp. 1–32, 2022.

[168] Y. Ma and Q. Li, "Exploring non-autoregressive text style transfer," in *EMNLP*, 2021, pp. 9267–9278.

[169] F. Huang, Z. Chen, C. H. Wu, Q. Guo, X. Zhu, and M. Huang, "Nast: A non-autoregressive generator with word alignment for unsupervised text style transfer," in *Findings of ACL-IJCNLP*, 2021, pp. 1577–1590.

[170] H. Zhang, H. Song, S. Li, M. Zhou, and D. Song, "A survey of controllable text generation using transformer-based pre-trained language models," *arXiv:2201.05337*, 2022.

[171] S. Agrawal, W. Xu, and M. Carpuat, "A non-autoregressive edit-based approach to controllable text simplification," in *Findings of ACL-IJCNLP*, 2021, pp. 3757–3769.

[172] H. Iso, "Autotemplate: A simple recipe for lexically constrained text generation," *arXiv:2211.08387*, 2022.

[173] X. Li, J. Thickstun, I. Gulrajani, P. S. Liang, and T. B. Hashimoto, "Diffusion-lm improves controllable text generation," in *NeurIPS*, vol. 35, 2022, pp. 4328–4343.

[174] S. Kumar, B. Paria, and Y. Tsvetkov, "Gradient-based constrained sampling from language models," *arXiv:2205.12558*, 2022.

[175] Z.-c. Fei, "Fast image caption generation with position alignment," *arXiv:1912.06365*, 2019.

[176] J. Gao, X. Meng, S. Wang, X. Li, S. Wang, S. Ma, and W. Gao, "Masked non-autoregressive image captioning," *arXiv:1906.00717*, 2019.

[177] L. Guo, J. Liu, X. Zhu, X. He, J. Jiang, and H. Lu, "Non-autoregressive image captioning with counterfactuals-critical multi-agent learning," in *IJCAI*, 2021, pp. 767–773.

[178] Z. Wang, L. Chen, W. Ma, G. Han, Y. Niu, J. Shao, and J. Xiao, "Explicit image caption editing," in *ECCV*. Springer, 2022, pp. 113–129.

[179] Z. Fei, "Efficient modeling of future context for image captioning," in *ACM MM*, 2022, pp. 5026–5035.

[180] C. Zhou, F. Meng, J. Zhou, M. Zhang, H. Wang, and J. Su, "Confidence based bidirectional global context aware training framework for neural machine translation," in *ACL*, 2022, pp. 2878–2889.

[181] Z. Fei, M. Fan, L. Zhu, J. Huang, X. Wei, and X. Wei, "Uncertainty-aware image captioning," *arXiv:2211.16769*, 2022.

[182] Q. Chen, C. Deng, and Q. Wu, "Learning distinct and representative modes for image captioning," in *NeurIPS*, vol. 35, 2022, pp. 9472–9485.

[183] T. Zhang, H. Xu, J. van Genabith, D. Xiong, and H. Zan, "Napg: Non-autoregressive program generation for hybrid tabular-textual question answering," *arXiv:2211.03462*, 2022.

[184] J. Wang, C. Wang, M. Qiu, Q. Shi, H. Wang, J. Huang, and M. Gao, "Kecp: Knowledge enhanced contrastive prompting for few-shot extractive question answering," *arXiv:2205.03071*, 2022.

[185] H. Lu, Z. Wu, X. Wu, X. Li, S. Kang, X. Liu, and H. Meng, "Vaenar-tts: Variational auto-encoder based non-autoregressive text-to-speech synthesis," *arXiv:2107.03298*, 2021.

[186] R. Shah, K. Pokora, A. Ezzerg, V. Klimkov, G. Huybrechts, B. Putrycz, D. Korzekwa, and T. Merritt, "Non-autoregressive tts with explicit duration modelling for low-resource highly expressive speech," *arXiv:2106.12896*, 2021.

[187] P. Liu, Y. Cao, S. Liu, N. Hu, G. Li, C. Weng, and D. Su, "Vara-tts: Non-autoregressive text-to-speech synthesis based on very deep vae with residual attention," *arXiv:2102.06431*, 2021.

[188] Y. Lee, J. Shin, and K. Jung, "Bidirectional variational inference for non-autoregressive text-to-speech," in *ICLR*, 2020.

[189] Y. Ren, X. Tan, T. Qin, Z. Zhao, and T.-Y. Liu, "Revisiting over-smoothness in text to speech," in *ACL*, 2022, pp. 8197–8213.

[190] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," in *Interspeech*, 2017, pp. 4006–4010.

[191] A. Łańcucki, "Fastpitch: Parallel text-to-speech with pitch prediction," in *ICASSP*. IEEE, 2021, pp. 6588–6592.

[192] H. Guo, H. Lu, X. Wu, and H. Meng, "A multi-scale time-frequency spectrogram discriminator for gan-based non-autoregressive tts," *arXiv:2203.01080*, 2022.

[193] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-tts: A generative flow for text-to-speech via monotonic alignment search," in *NeurIPS*, vol. 33, 2020, pp. 8067–8077.

[194] J. Bae, J. Yang, T. Bak, and Y.-S. Joo, "Hierarchical and Multi-Scale Variational Autoencoder for Diverse and Natural Non-Autoregressive Text-to-Speech," in *Interspeech*, 2022, pp. 813–817.

[195] Y. Li, C. Yu, G. Sun, H. Jiang, F. Sun, W. Zu, Y. Wen, Y. Yang, and J. Wang, "Cross-utterance conditioned vae for non-autoregressive text-to-speech," in *ACL*, 2022, pp. 391–400.

[196] Z. Liu, Q. Tian, C. Hu, X. Liu, M. Wu, Y. Wang, H. Zhao, and Y. Wang, "Controllable and lossless non-autoregressive end-to-end text-to-speech," *arXiv:2207.06088*, 2022.

[197] M. Post, G. Kumar, A. Lopez, D. Karakos, C. Callison-Burch, and S. Khudanpur, "Improved speech-to-text translation with the fisher and callhome spanish-english speech translation corpus," in *SLT Workshop*, 2013.

[198] S.-P. Chuang, Y.-S. Chuang, C.-C. Chang, and H.-Y. Lee, "Investigating the reordering capability in ctc-based non-autoregressive end-to-end speech translation," in *Findings of ACL-IJCNLP*, 2021, pp. 1068–1077.

[199] H. Inaguma, Y. Higuchi, K. Duh, T. Kawahara, and S. Watanabe, "Orthros: Non-autoregressive end-to-end speech translation with dual-decoder," in *ICASSP*. IEEE, 2021, pp. 7503–7507.

[200] ——, "Non-autoregressive end-to-end speech translation with parallel autoregressive rescoring," *arXiv:2109.04411*, 2021.

[201] S. Dalmia, B. Yan, V. Raunak, F. Metze, and S. Watanabe, "Searchable hidden intermediates for end-to-end models of decomposable sequence tasks," in *NAACL-HLT*, 2021.

[202] J. Shi, J. D. Amith, X. Chang, S. Dalmia, B. Yan, and S. Watanabe, "Highland puebla nahuatl speech translation corpus for endangered language documentation," in *AmericasNLP*, 2021, pp. 53–63.

[203] H. Inaguma, S. Dalmia, B. Yan, and S. Watanabe, "Fast-md: Fast multi-decoder end-to-end speech translation with non-autoregressive hidden intermediates," in *ASRU*. IEEE, 2021, pp. 922–929.

[204] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *ICML*. PMLR, 2015, pp. 2256–2265.

[205] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *NeurIPS*, vol. 33, 2020, pp. 6840–6851.

[206] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "Diffwave: A versatile diffusion model for audio synthesis," in *ICLR*, 2020.

[207] G. Mittal, J. Engel, C. Hawthorne, and I. Simon, "Symbolic music generation with diffusion models," *arXiv:2103.16091*, 2021.

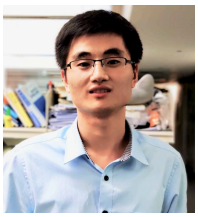
[208] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *ICML*. PMLR, 2021, pp. 8162–8171.

[209] J. Austin, D. D. Johnson, J. Ho, D. Tarlow, and R. van den Berg, "Structured denoising diffusion models in discrete state-spaces," in *NeurIPS*, vol. 34, 2021, pp. 17 981–17 993.

[210] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," in *NeurIPS*, vol. 35, 2022, pp. 36 479–36 494.

[211] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv:2204.06125*, 2022.
[212] A. Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. Mcgrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," in *ICML*. PMLR, 2022, pp. 16 784–16 804.
[213] E. Hoogeboom, D. Nielsen, P. Jaini, P. Forré, and M. Welling, "Argmax flows and multinomial diffusion: Learning categorical distributions," in *NeurIPS*, vol. 34, 2021, pp. 12 454–12 465.
[214] E. Hoogeboom, A. A. Gritsenko, J. Bastings, B. Poole, R. van den Berg, and T. Salimans, "Autoregressive diffusion models," in *ICLR*, 2022.
[215] S. Gong, M. Li, J. Feng, Z. Wu, and L. Kong, "Diffuseq: Sequence to sequence text generation with diffusion models," *arXiv:2210.08933*, 2022.
[216] P. Yu, S. Xie, X. Ma, B. Jia, B. Pang, R. Gao, Y. Zhu, S.-C. Zhu, and Y. N. Wu, "Latent diffusion energy-based model for interpretable text modelling," in *ICML*. PMLR, 2022, pp. 25 702–25 720.
[217] M. Reid, V. J. Hellendoorn, and G. Neubig, "Diffuser: Discrete diffusion via edit-based reconstruction," *arXiv:2210.16886*, 2022.
[218] A. Wang and K. Cho, "Bert has a mouth, and it must speak: Bert as a markov random field language model," in *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, 2019, pp. 30–36.
[219] B. Yu, Y. Wang, T. Liu, H. Zhu, L. Sun, and B. Wang, "Maximal clique based non-autoregressive open information extraction," in *EMNLP*, 2021, pp. 9696–9706.
[220] S. Yu, J. Tack, S. Mo, H. Kim, J. Kim, J.-W. Ha, and J. Shin, "Generating videos with dynamics-aware implicit generative adversarial networks," in *ICLR*, 2021.
[221] T. Hayashi, W.-C. Huang, K. Kobayashi, and T. Toda, "Non-autoregressive sequence-to-sequence voice conversion," in *ICASSP*. IEEE, 2021, pp. 7068–7072.
[222] Z. Chen, H. Miao, and P. Zhang, "Streaming non-autoregressive model for any-to-many voice conversion," *arXiv:2206.07288*, 2022.
[223] Y. Wang and Z. Zhao, "Fastlts: Non-autoregressive end-to-end unconstrained lip-to-speech synthesis," in *ACM MM*, 2022, pp. 5678–5687.
[224] A. H. Liu, Y.-A. Chung, and J. Glass, "Non-autoregressive predictive coding for learning speech representations from local dependencies," *arXiv:2011.00406*, 2020.
[225] F. Liu, Z. Fu, G. Li, Z. Jin, H. Liu, and Y. Hao, "Non-autoregressive model for full-line code completion," *arXiv:2204.09877*, 2022.
[226] E. J. Barezi, I. Calixto, K. Cho, and P. Fung, "A study on the autoregressive and non-autoregressive multi-label learning," *arXiv:2012.01711*, 2020.
[227] S. Feng, K. Xu, J. Wu, P. Wu, F. Lin, and P. Zhao, "Multi-scale attention flow for probabilistic time series forecasting," *arXiv:2205.07493*, 2022.
[228] F. Huang, T. Tao, H. Zhou, L. Li, and M. Huang, "On the learning of non-autoregressive transformers," in *ICML*. PMLR, 2022, pp. 9356–9376.
[229] Y. Oka, K. Sudoh, and S. Nakamura, "Using perturbed length-aware positional encoding for non-autoregressive neural machine translation," *arXiv:2107.13689*, 2021.
[230] X. Kong, Z. Zhang, and E. Hovy, "Incorporating a local translation mechanism into non-autoregressive translation," in *EMNLP*, 2020, pp. 1067–1073.
[231] C. Shao, Y. Feng, J. Zhang, F. Meng, X. Chen, and J. Zhou, "Retrieving sequential information for non-autoregressive neural machine translation," in *ACL*, 2019, pp. 3013–3024.
[232] L. Qin, F. Wei, T. Xie, X. Xu, W. Che, and T. Liu, "Gl-gin: Fast and accurate non-autoregressive model for joint multiple intent detection and slot filling," in *ACL-IJCNLP*, 2021, pp. 178–188.
[233] D. Wu, L. Ding, F. Lu, and J. Xie, "Slotrefine: A fast non-autoregressive model for joint intent detection and slot filling," in *EMNLP*, 2020, pp. 1932–1937.
[234] H. Le, R. Socher, and S. C. Hoi, "Non-autoregressive dialog state tracking," in *ICLR*, 2021.
[235] T. Komatsu, "Non-autoregressive asr with self-conditioned folded encoders," in *ICASSP*. IEEE, 2022, pp. 7427–7431.
[236] T. Wang, Y. Fujita, X. Chang, and S. Watanabe, "Streaming end-to-end asr based on blockwise non-autoregressive models," *arXiv:2107.09428*, 2021.
[237] S. Beliaev and B. Ginsburg, "Talknet 2: Non-autoregressive depth-wise separable convolutional model for speech synthesis with explicit pitch and duration prediction," *arXiv:2104.08189*, 2021.
[238] C.-M. Chien and H.-y. Lee, "Hierarchical prosody modeling for non-autoregressive speech synthesis," in *SLT Workshop*. IEEE, 2021, pp. 446–453.
[239] A. Mohammadshahi and J. Henderson, "Recursive non-autoregressive graph-to-graph transformer for dependency parsing with iterative refinement," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 120–138, 2021.
[240] J. Ho and T. Salimans, "Classifier-free diffusion guidance," in *NeurIPS Workshop*, 2021.
[241] Y. Yin, L. Huang, Y. Liu, and K. Huang, "Diffgar: Model-agnostic restoration from generative artifacts using image-to-image diffusion models," *arXiv:2210.08573*, 2022.
[242] K. Akuzawa, K. Onishi, K. Takiguchi, K. Mametani, and K. Mori, "Conditional deep hierarchical variational autoencoder for voice conversion," in *APSIPA ASC*. IEEE, 2021, pp. 808–813.

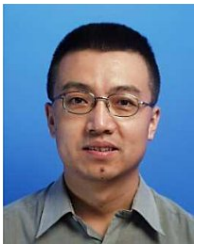
Yisheng Xiao is now a graduate student at the natural language processing laboratory, Soochow University, supervised by Prof. Min Zhang. He received his Bachelor's degree, majoring in software engineering, from the same place in 2021. His research interests lie in Natural Language Processing, especially in neural machine translation, non-autoregressive methods, efficient text generation, and pre-trained language models.



Lijun Wu is a Senior Researcher of Machine Learning Group in Microsoft Research Asia (MSRA). He got his Ph.D. degree from Sun Yat-sen University (SYSU) in 2020, a member of the joint Ph.D. program between SYSU and MSRA. He received MSRA Ph.D. Fellowship in 2018. His researches focus on Deep Learning, Natural Language Processing, Multimodality Learning, and Medical Health.

Junliang Guo is a Researcher of the Machine Learning group in Microsoft Research Asia (MSRA). He received his Ph.D. degree in computer science from the University of Science and Technology of China (USTC) in 2021. His research interests lie in the general area of machine learning, specifically in sequence modeling and representation learning, as well as their applications in natural language processing and heterogeneous types of data such as networks.

Juntao Li is now an associate professor at the Institute of Artificial Intelligence, Soochow University. Before that, he obtained a doctoral degree from Peking University in 2020. He is now working on pre-trained language models, text generation, and dialogue systems.



Min Zhang is a Distinguished Professor at Soochow University (China). He received his bachelor's degree and Ph.D. degree from the Harbin Institute of Technology in 1991 and 1997, respectively. His current research interests include machine translation, natural language processing, and machine learning. He has authored 150 papers in leading journals and conferences and has co-edited 10 books that were published by Springer and IEEE.

Tao Qin received the Ph.D. degree and Bachelor degree both from Tsinghua University. He is a Senior Member of ACM and IEEE and a Senior Principal Research Manager in Machine Learning Group, Microsoft Research Asia. His research interests include machine learning (with the focus on deep learning and reinforcement learning), artificial intelligence (with applications to language understanding and computer vision).

Tie-Yan Liu received the Ph.D. degree and Bachelor degree both from Tsinghua University. He is an Assistant managing director of Microsoft Research Asia, leading the machine learning research area. He is an adjunct/honorary professor at Carnegie Mellon University (CMU). He has published 200+ papers in refereed conferences and journals, e.g., ICML, KDD, NeurIPS, with 30000+ citations. He is a fellow of IEEE and ACM.

APPENDIX

OVERVIEW OF IMPROVING METHODS FOR NAT

To have a clear overview of improving methods for NAT, we show the general framework and the data flow of various NAT models in Figure 4, which contain different components such as the data preparation, NAT encoder, and NAT decoder.

REPRESENTATIVE MODELING METHODS

Figure 5 presents a few representative modeling methods mentioned in Section 4.

CRITERION

Figure 6 compares different loss functions in NAT models.

ILLUSTRATION OF DECODING STRATEGIES

We show several typical decoding strategies in Figure 7 and their detailed decoding process with a specific example in Figure 8, including autoregressive decoding, semi-autoregressive decoding, fully non-autoregressive decoding, mask and predict iterative decoding, and insert and delete iterative decoding.

COMPARISONS BETWEEN EXISTING METHODS

In the paper, we mainly present the NAT works with their performances on the WMT14 English→German translation task. Here we give a broader comparison of WMT14 English↔German, WMT16 English↔Romanian (EN↔RO), and IWSLT14/16 English↔German translation benchmark datasets. The decoding iterations and the speedup ratio compared to AT models are reported in Table 2. Figure 9 plots the BLEU-Speedup curve to demonstrate the correlations between performance and inference speed achieved by representative NAT methods better, and Figure 10 further presents the evolution of BLEU scores on WMT14 English→German translation by the time of Fully NAT and Iterative NAT. Methods in the lower left part of Figure 9, e.g., DisCo [36], NAT [16] can achieve much faster inference speed but at the cost of significant performance decrease, while methods in the upper right part can make a better trade-off between speed-up and performance. A few powerful NAT methods can even achieve comparable and slightly better performance than the strong AT model with a speed advantage. In Figure 10, iteration-based NAT models generally achieve higher BLEU scores than fully NAT methods at the cost of multiple inference time, but their performance gap is rapidly shrinking, e.g., the recent combination of CTC length prediction, latent variable, and extra upsampling module can achieve competitive performance with strong iteration-based NAT methods. It can be expected that fully NAT methods can achieve better performance while maintaining their speed advantage with emerging effective strategies and a suitable combination.

RESOURCES

We collect valuable resources for NAT models with their open-source information, including the paper URL, code address (Github), and deep learning tools. Table 3, Table 4 and Table 5 are the summarized information for the resources of NAT task and other extensive tasks.

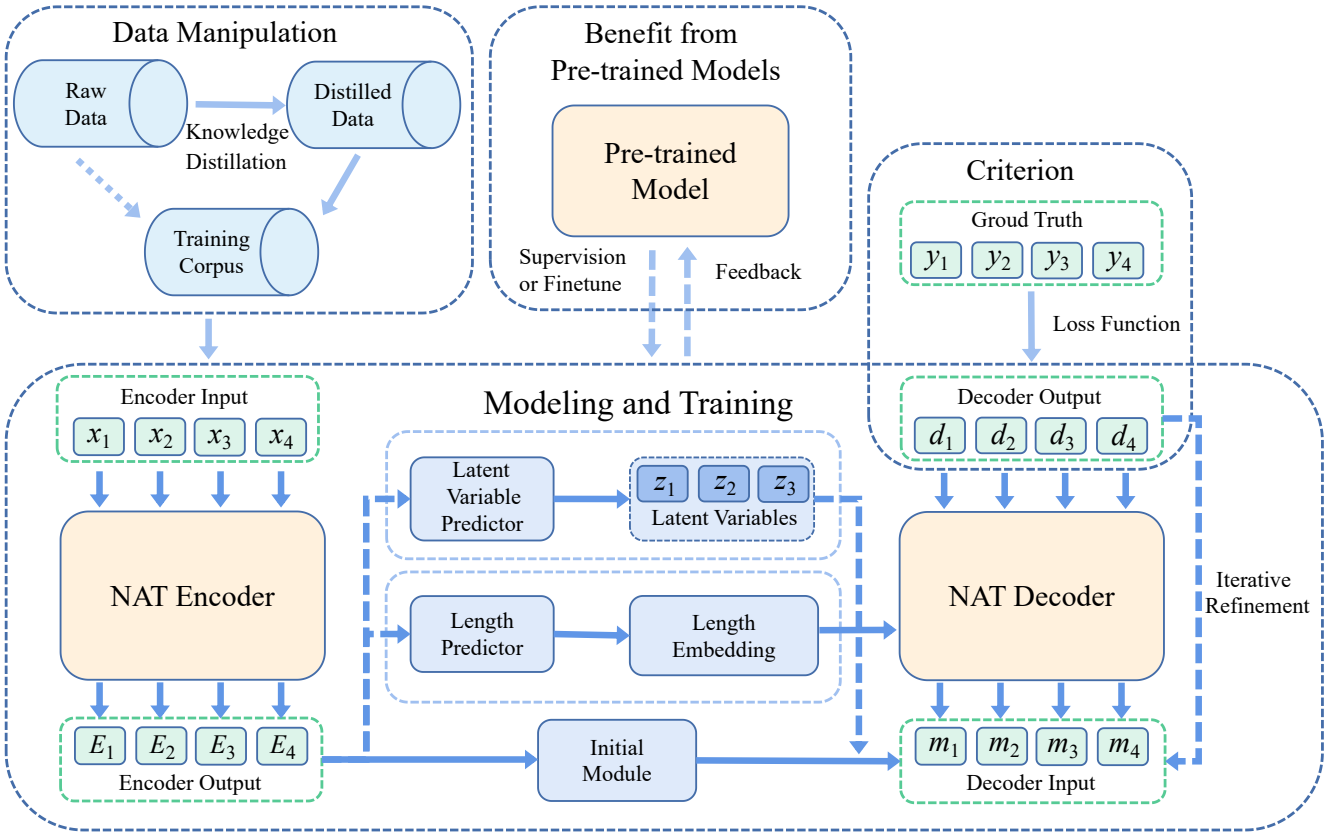
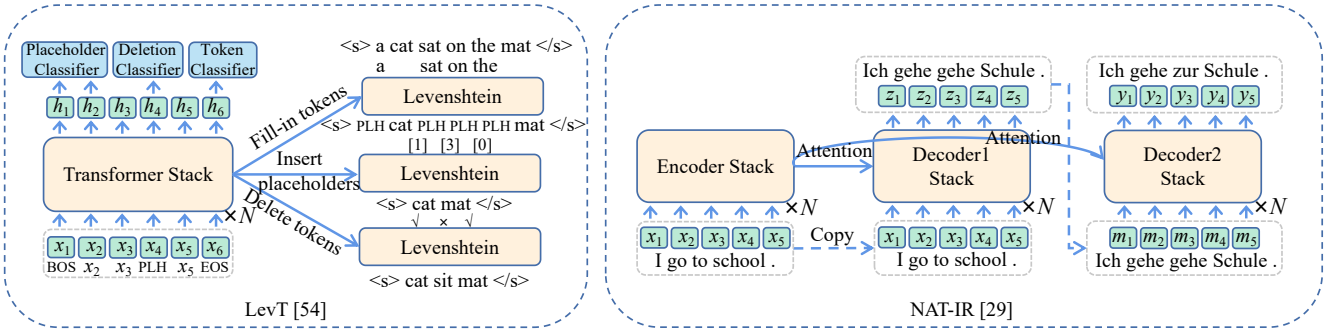
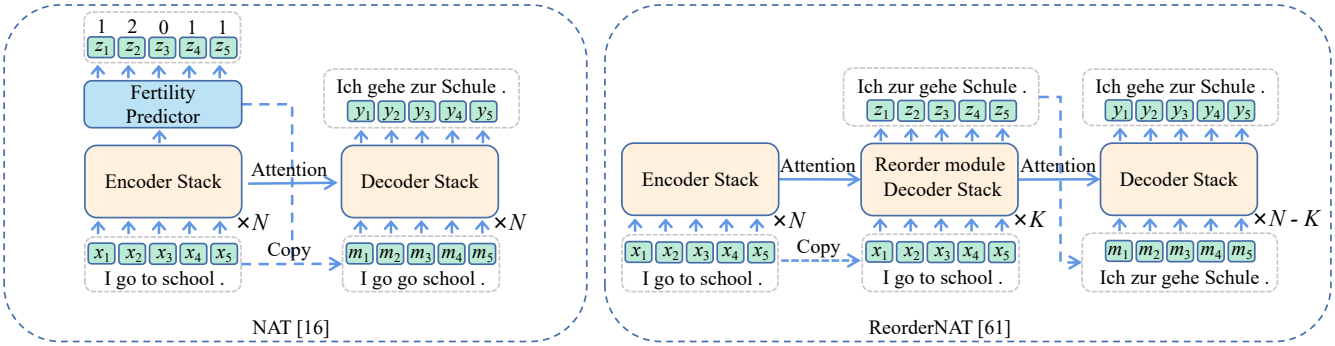


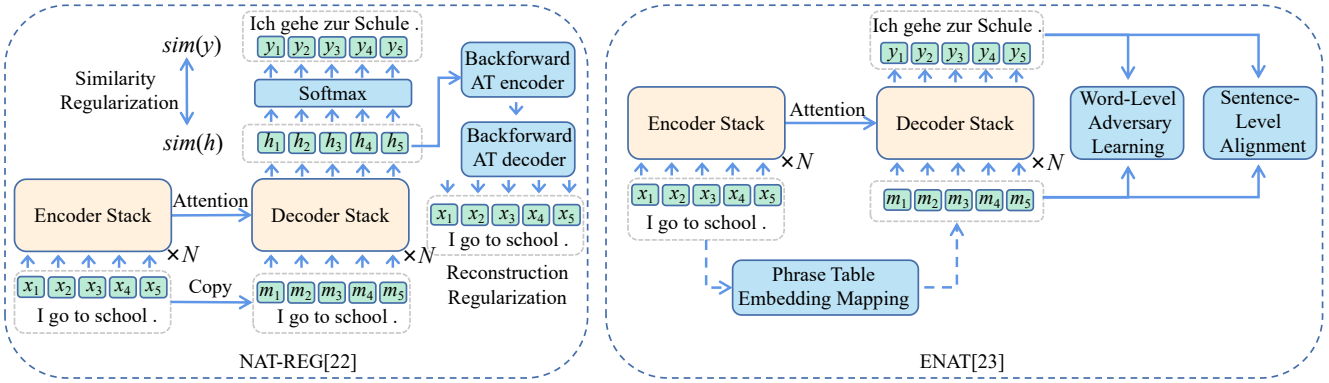
Fig. 4. The overall framework of various components for improving NAT models. The dashed arrow denotes this part is not applied in all NAT models. Different knowledge distillation methods will be applied in Data Preparing part. Length Predictor is involved in most NAT models, and Latent Variable Predictor is applied in latent variable-based models. The initial Module is used to initialize the Decoder Input, such as soft copy, source copy, partial target tokens, etc.



(a) An illustration of two iteration-based methods introduced in Section 4.1.



(b) An illustration of two latent variable-based methods introduced in Section 4.2.



(c) An illustration of two enhancements-based methods introduced in Section 4.3.

Fig. 5. We show the model structures of two iteration-based methods, e.g., two iteration-based methods, NAT-IR [29] and LevT [35]; two latent variable-based methods, e.g., NAT [16] and ReorderNAT [68], where the fertility predictor and reorder module are applied to predict the latent variables; and two enhancements-based model, NAT-REG [22] and ENAT [23], and the corresponding enhancement module is also given.

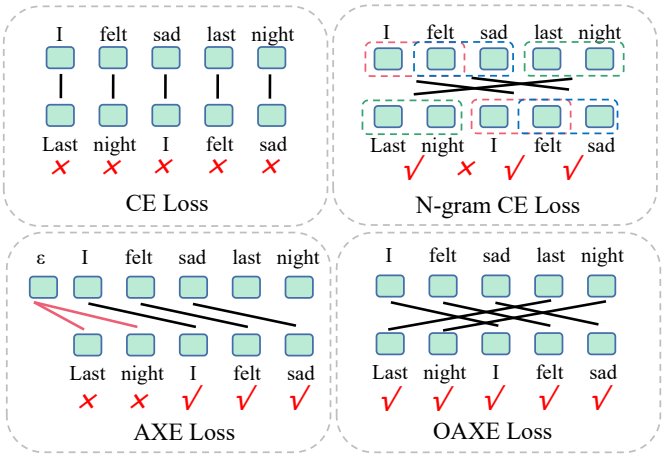


Fig. 6. An illustration of different loss functions. e.g., Model prediction: **Last night I feel sad** and Ground-truth: **I feel sad last night**. Traditional CE loss will give a penalty to all tokens. N-gram CE loss only finds a two-gram **night I** unreasonable. AXE loss finds the best possible monotonic alignment and penalizes unaligned tokens, denoted as ϵ, while OAXE loss removes the order errors and give no penalty to this prediction.

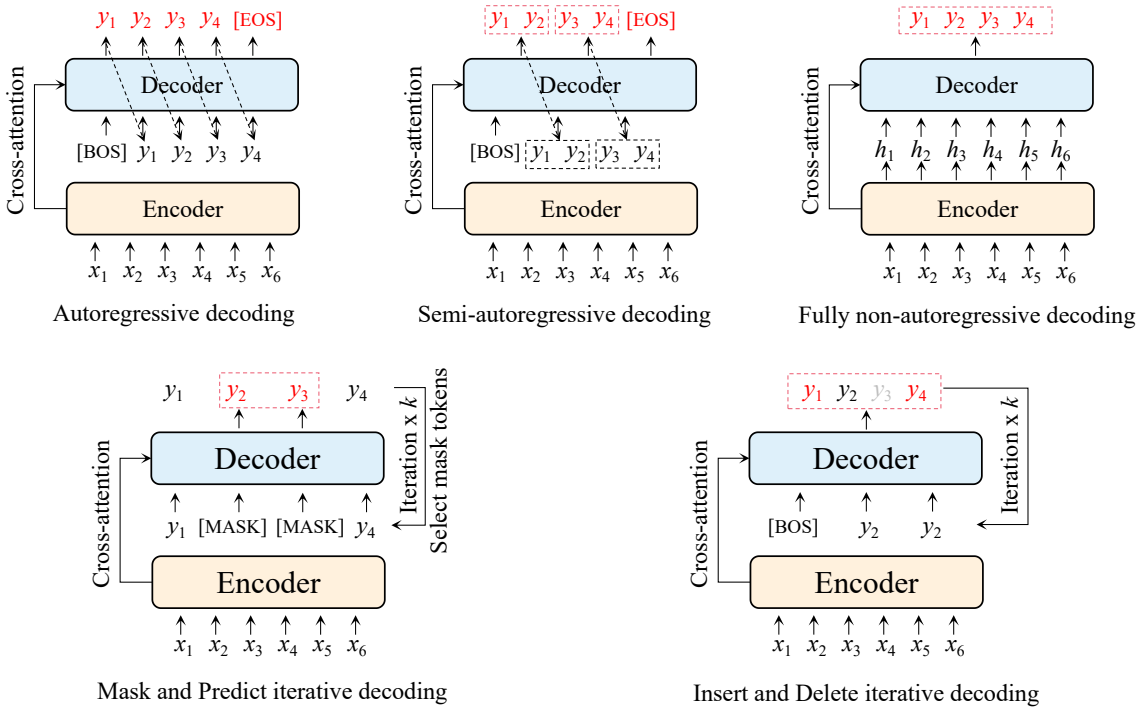


Fig. 7. An illustration of several different decoding strategies introduced in Section 6, e.g., autoregressive and semi-autoregressive decoding methods generate tokens in a left-to-right order, fully non-autoregressive and mask and predict iterative decoding methods predict all/part of target tokens in parallel (tokens marked by red), insert and delete iterative decoding method generate tokens with insertion (tokens marked by red) and deletion (tokens marked by grey) operations.

Src: Wir sind stolz auf wusere Leistung , aber wir wollen jedes Spiel game .
Tgt: We are proud of our performance , but we want to win every game .
Output of different decoding strategies:

Autoregressive decoding:
Iter.1: We are proud of our performance , but we want to win every game .
Iter.2: We are proud of our performance , but we want to win every game .
Iter.3: We are proud of our performance , but we want to win every game .
...
Iter.15: We are proud of our performance , but we want to win every game .

Semi-autoregressive decoding (k=2):
Iter.1: We are proud of our performance , but we want to win every game .
Iter.2: We are proud of our performance , but we want to win every game .
Iter.3: We are proud of our performance , but we want to win every game .
...
Iter.8: We are proud of our performance , but we want to win every game .

Fully non-autoregressive decoding:
Iter.1: We are proud of our performance , but we want to win every game .

Mask and Predict iterative decoding (Iteration=4):
Iter.0: We are of of our perform , and and want to to win win games .
Iter.1: We are of of our perform , but and we want to win win game .
Iter.2: We are of of our performance , but we want want win every game .
Iter.3: We are of of our performance , but we want win every game .
Iter.4: We are proud of our performance , but we want to win every game .

Insert and Delete iterative decoding:
Iter.0: We are proud of our performance , but we want to win every game .
Iter.1: We are proud of our performance , but we want to to win every game .
Iter.2: We are are proud of our performance , but we want to win every game .
Iter.3: We are proud of our performance , but we want to win every game .
Iter.4: We are proud of our performance , but we want to win every game .
Iter.5: We are proud of our performance , but we want to win every game .

Fig. 8. Cases of several different decoding strategies discussed in this paper. Texts marked in yellow denote the content that will be masked and generated in next iteration.

TABLE 2
Performances on popular datasets, i.e., WMT'16 EN↔RO, WMT'14 EN↔DE, and IWSLT'14/16 EN↔DE. "*" indicates training with sequence-level knowledge distillation from a big Transformer; "†" denotes training without sequence-level knowledge distillation; "‡" refers to results on IWSLT'16.

Model	Iteration	Speedup	WMT'14		WMT'16		IWSLT'14/16	
			EN→DE	DE→EN	EN→RO	RO→EN	EN→DE	DE→EN
NAT [16]	1	15.6x	17.69	21.47	27.29	29.06	26.52‡	-
NAT-IR [29]	10	1.5x	21.61	25.48	29.32	30.19	27.11‡	32.31‡
RDP [20]	2.5	3.5x	27.8*	-	-	33.8	-	-
LRF [50]	2.5	3.5x	28.2*	-	-	33.8	-	-
SDMRT [21]	10	-	27.72*	31.65*	33.72	33.94	27.49	-
MD [51]	1	-	25.73	30.18	31.96	33.57	-	-
DDRS [52]	1	14.7x	27.60	31.48	34.60	34.65	-	33.12
LaNMT-C [56]	2	11.0x	26.02	31.23	32.50	-	-	-
CCMLM [57]	10	-	27.93*	31.57*	33.88	34.18	-	-
perLDPE [229]	Adaptive	-	26.3	29.5	-	-	-	-
GLAT [39]	1	15.3x	25.21	29.84	31.19	32.04	-	29.61†
PMG [54]	2.5	3.5x	27.8*	-	-	33.8*	-	-
latent-GLAT [53]	1	11.3x	26.64	29.93	-	-	-	32.47
Insertion Transformer [58]	$\approx log_2(N)$	-	27.41	-	-	-	-	-
LevT [35]	Adaptive	4.0x	27.27	-	-	33.26	-	-
CMLM [18]	10	1.7x	27.03*	30.53*	33.08	33.31	-	-
SMART [59]	10	1.7x	27.65*	31.27*	-	-	-	-
DisCo [36]	Adaptive	3.5x	27.34*	31.31*	33.22	33.25	-	-
JM-NAT [60]	10	5.7x	27.69*	32.24 *	33.52	33.72	-	32.59
AR Deep-Shallow [119]	N	2.5x	28.3*	31.8*	33.8	34.8	-	-
MvSR-NAT [55]	10	3.8x	27.39*	31.18*	33.38	33.56	-	32.55
REWRITENAT [26]	2.3	3.9x	27.83*	31.52*	33.63	34.09	-	-
CMLMC [61]	10	-	28.37*	31.41*	34.57	34.13	28.51	34.78
FlowSeq [64]	1	1.1x	23.72	28.39	29.73	30.72	27.55	-
NART-DCRF [25]	1	10.4x	23.44	27.22	-	-	-	27.44
PNAT [65]	1	7.3x	23.05	27.18	-	-	-	31.23‡
SynST [66]	$N/6$	4.6x	20.74	25.50	-	-	23.82	-
LaNMT [67]	1	6.8x	25.10	-	-	-	-	-
Imputer [38]	8	3.9x	28.2*	31.8*	34.4	34.1	-	-
LAT [230]	4	6.7x	27.35	32.04	32.87	33.26	-	34.08
SUNDAE [62]	16	-	28.46*	32.30 *	-	-	-	-
INSNET [63]	16.1	3.78x	28.05	-	-	33.91	-	-
AligNART [69]	1	13.2x	26.4	30.4	32.5	33.1	-	-
ReorderNAT [68]	1	6.0×	22.79	27.28	29.30	29.50	25.29‡	-
CNAT [70]	1	10.4x	25.56*	29.36*	-	-	-	31.15
SNAT [71]	1	22.6x	24.64*	28.42*	32.87	32.21	-	-
Fully NAT [72]	1	16.5x	27.49	31.39	33.79	34.16	-	-
ENAT [23]	1	25.3x	20.65	23.02	30.08	-	-	24.13
NAT-REG [22]	1	27.6x	20.65	24.77	-	-	23.14‡	23.89
LAVA NAT [73]	1	20.2x	27.94	31.33	-	32.85	-	33.59†
CCAN [74]	10	-	27.5*	-	-	33.7	-	-
DSLP [75]	1	14.8x	27.02	31.61	34.17	34.60	-	-
DAD [76]	1	14.7×	27.51	31.96	34.68	34.98	-	-
DA-Transformer [40]	1	13.9x	27.49	31.37	-	-	-	-
DA-Transformer Viterbi [77]	1	13.2x	26.89	31.10	-	-	-	-
FA-DAT [78]	1	14.0x	27.53	31.37	-	-	-	-
CTC [38]	1	18.6×	25.7	28.10	32.20	31.60	-	-
Reinforce-NAT [231]	1	3.6x	22.27	27.25	30.57	30.83	27.78‡	-
BoN [79]	1	9.6x	20.90	24.61	28.31	29.29	25.72‡	-
AXE [80]	1	15.3x	23.53*	27.90*	30.75	31.54	-	-
OAXE [41]	1	15.3x	26.10*	30.20*	32.40	33.30	-	-
ngram-OAXE [82]	1	15.2x	26.50*	30.50*	-	-	-	-
CoCO [33]	1	14.2x	27.41	31.37	34.32	-	-	-
MgMO [83]	1	-	26.4	30.3	32.9	33.6	-	-
NMLA [84]	1	14.7x	27.57	31.28	33.86	33.94	-	-
SAT [17]	$N/2$	1.5x	26.90	-	-	-	-	-
RecoverSAT [85]	$N/2$	2.1x	27.11	31.67	32.92	33.19	30.78‡	-
GAD++ [86]	4.0	3.2x	28.89*	-	-	-	-	-
Unified [37]	10	-	26.24	-	-	-	-	30.73
Diformer [88]	10	-	27.99	31.68	34.37	33.34	-	-
HRT [89]	$N/2+1$	-	28.49*	32.28*	34.24	34.35	-	-
imitate-NAT [90]	1	18.4x	22.44*	25.67*	28.61	28.90	28.41‡	-
NAT-HINT [91]	1	30.2x	21.11	25.24	-	-	-	25.55
ENGINE [92]	10	-	-	-	-	34.04	-	33.17
EM+ODD [93]	1	16.4x	24.54	27.93	-	-	-	30.69
FCL-NAT [24]	1	28.9x	21.70	25.32	-	-	-	26.62
MULTI-TASK NAT [94]	10	-	27.98*	31.27*	33.80	33.60	-	-
TCT-NAT [95]	1	27.6x	21.94	25.62	-	-	26.01‡	28.16
weak MTL [96]	1	-	27.25	30.70	33.88	34.73	35.15	-
AB-Net [97]	-	2.4x	28.69*	33.57*	-	35.63	-	36.49
NAG-BERT [98]	1	11.3x	-	-	-	-	-	30.45
CeMAT [42]	10	-	27.2	29.9	33.3†	33.0†	26.7†	33.7†
XML-D [99]	8	2.8x	29.80	32.88	35.34	35.50	-	-

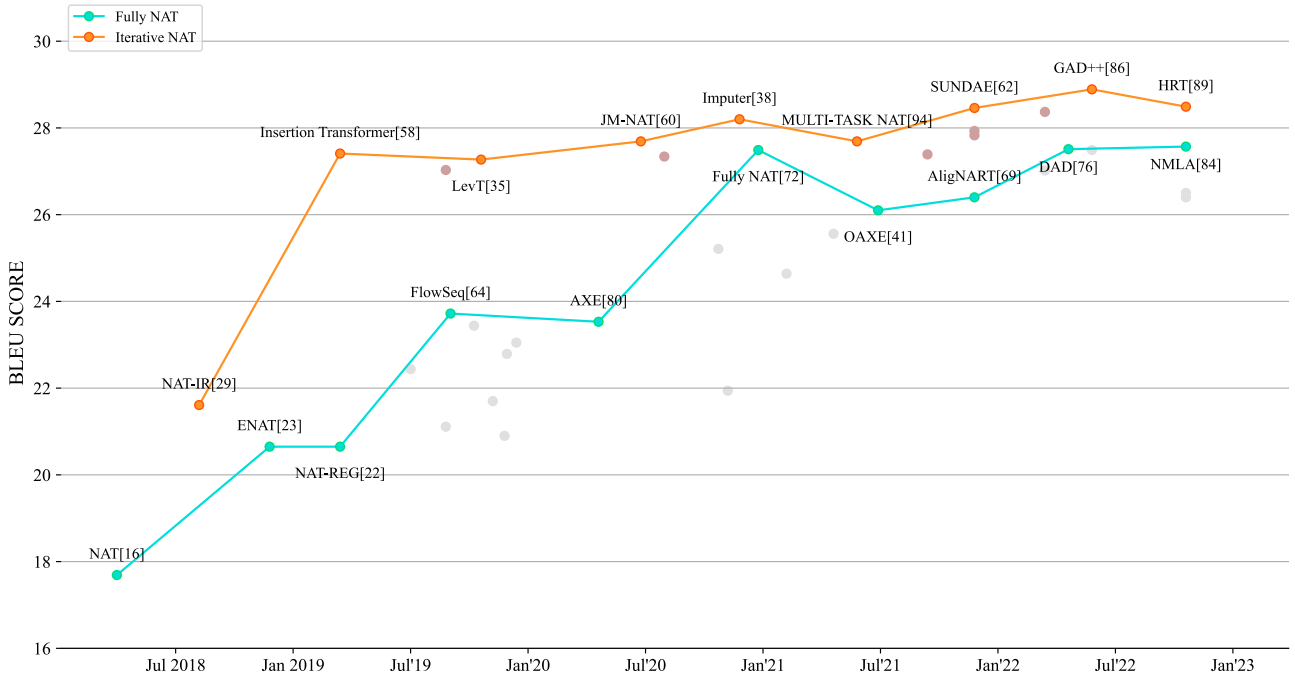


Fig. 9. BLEU v.s. Speedup. BLEU score is reported on the WMT14 EN→DE test set. Speedup is reported for NAT models compared with corresponding AT models. Dotted lines denote the different scores of iterative models achieved with different iterations. Note that the ideal model should appear in the top right-hand corner.

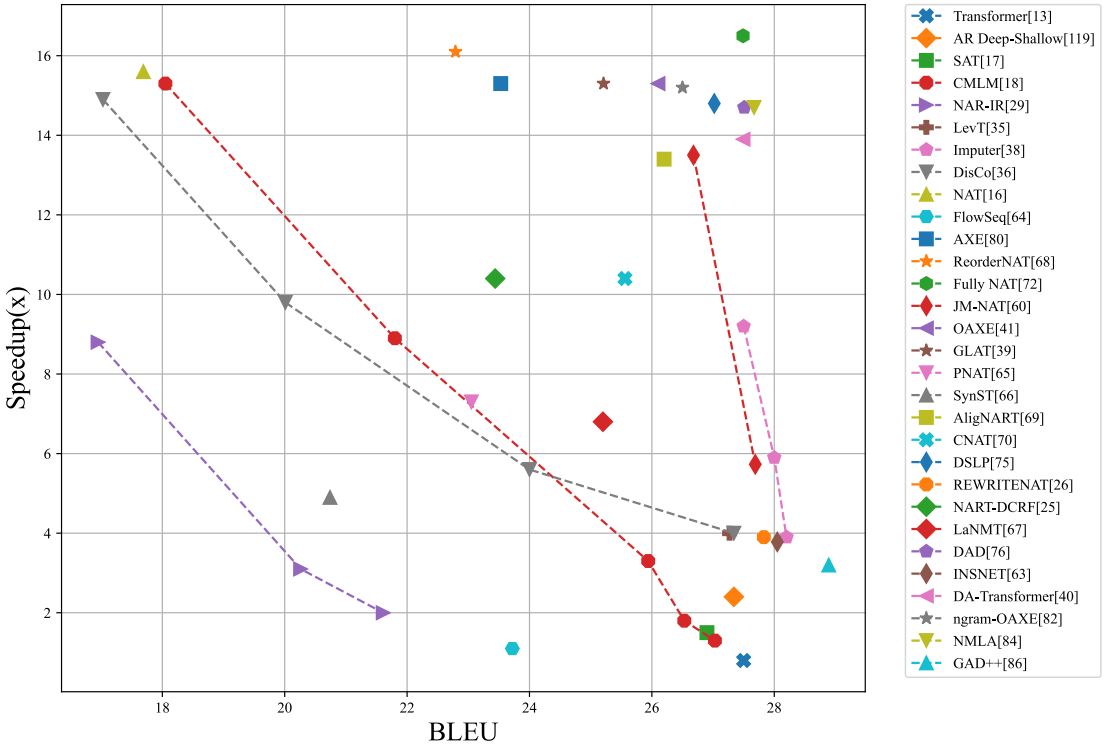


Fig. 10. The evolution of BLEU scores on WMT14 EN→De translation by the time of Fully NAT and Iterative NAT. Note that the performance of iterative NAT models is commonly better than fully NAT models at different stages, but the gap is narrowing with the development of NAT models.

TABLE 3
A collection of NAT published papers and codes.

Method	Paper URL	Code URL	Framework
Machine Translation			
NAT [16]	https://openreview.net/pdf?id=B1l8BtlCb	https://github.com/salesforce/nonauto-nmt	Pytorch
NAT-IR [29]	https://aclanthology.org/D18-1149.pdf	https://github.com/nyu-dl/dl4mt-nonauto	Pytorch
RDP [20]	https://openreview.net/pdf?id=Z1FeSBIX9C	-	-
LRF [30]	https://aclanthology.org/2021.acl-long.266.pdf	https://github.com/longyuewangdcu/RLFW-NAT	To be released
SDMRT [21]	https://arxiv.org/pdf/2112.11640v1.pdf	-	-
MD [51]	https://aclanthology.org/2020.acl-main.171.pdf	-	-
DDRS [52]	https://aclanthology.org/2022.naacl-main.277.pdf	https://github.com/ictnlp/DDRS-NAT	Pytorch/Fairseq
LaNMT-C [56]	https://aclanthology.org/2022.naacl-main.45.pdf	https://github.com/zomux/lanmt	Pytorch
CCMLM [57]	https://aclanthology.org/2022.findings-emnlp.463.pdf	-	-
perLDPE [229]	https://arxiv.org/pdf/2107.13689.pdf	-	-
GLAT [39]	https://aclanthology.org/2021.acl-long.155.pdf	https://github.com/FLC777/GLAT	Pytorch/Fairseq
PMG [54]	https://aclanthology.org/2021.findings-acl.247.pdf	-	-
latent-GLAT [53]	https://aclanthology.org/2022.acl-long.575/	https://github.com/baoy-nlp/Latent-GLAT	Pytorch
Insertion Transformer [58]	http://proceedings.mlr.press/v97/stern19a/stern19a.pdf	https://github.com/pytorch/fairseq	Pytorch/Fairseq
LevT [35]	https://dl.acm.org/doi/pdf/10.5555/3454287.3455290	https://github.com/pytorch/fairseq	Pytorch/Fairseq
CMLM [18]	https://aclanthology.org/D19-1633.pdf	https://github.com/facebookresearch/Mask-Predict	Pytorch/Fairseq
SMART [59]	https://arxiv.org/pdf/2001.08785.pdf	-	-
DisCo [36]	http://proceedings.mlr.press/v119/kasai20a/kasai20a.pdf	https://github.com/facebookresearch/DisCo	Pytorch/Fairseq
JM-NAT [60]	https://aclanthology.org/2020.acl-main.36.pdf	https://github.com/lemmonation/jm-nat	Pytorch/Fairseq
AR Deep-Shallow [119]	https://openreview.net/pdf?id=KptasIaLUpq	https://github.com/jungokasai/deep-shallow	Pytorch/Fairseq
MvSR-NAT [55]	https://ieeexplore.ieee.org/abstract/document/9944912	-	-
REWRITENAT [26]	https://aclanthology.org/2021.emnlp-main.265.pdf	https://github.com/xwgeng/RewriteNAT	Pytorch/Fairseq
CMLMC [61]	https://openreview.net/pdf?id=I2Hw58KHp8O	-	-
FlowSeq [64]	https://aclanthology.org/D19-1437.pdf	https://github.com/XuezheMax/flowseq	Pytorch
NART-DCRF [25]	https://dl.acm.org/doi/pdf/10.5555/3454287.3454558	-	-
PNAT [65]	https://arxiv.org/pdf/1911.10677.pdf	-	-
SynST [66]	https://aclanthology.org/P19-1122.pdf	https://github.com/dojoteef/synst	Pytorch
LaNMT [67]	https://ojs.aaai.org/index.php/AAAI/article/view/6413	https://github.com/zomux/lanmt	Pytorch
Imputer [38]	https://aclanthology.org/2020.emnlp-main.83.pdf	https://github.com/rosinality/imputer-pytorch	Pytorch
LAT [116]	https://aclanthology.org/2020.emnlp-main.79.pdf	https://github.com/shawnkx/NAT-with-Local-AT	Pytorch
AligNAR1 [69]	https://aclanthology.org/2021.emnlp-main.1.pdf	-	-
ReorderNAT [68]	https://ojs.aaai.org/index.php/AAAI/article/view/17618	https://github.com/ranqiu92/ReorderNAT	Pytorch/OpenNMT
CNAT [70]	https://aclanthology.org/2021.naacl-main.458.pdf	https://github.com/baoy-nlp/CNAT	Pytorch
SNAT [71]	https://aclanthology.org/2021.eacl-main.105.pdf	-	-
Fully NAT [72]	https://aclanthology.org/2021.findings-acl.11.pdf	https://github.com/pytorch/fairseq	Pytorch/Fairseq
ENAT [23]	https://ojs.aaai.org/index.php/AAAI/article/view/4257	-	-
NAT-REG [22]	https://ojs.aaai.org/index.php/AAAI/article/view/4476	-	-
LAVA NAT [73]	https://arxiv.org/pdf/2002.03084v1.pdf	-	-
CCAN [74]	https://aclanthology.org/2020.coling-main.389.pdf	-	-
DSLP [75]	https://ojs.aaai.org/index.php/AAAI/article/view/21323	https://github.com/chenyangh/DSLP	Pytorch/Fairseq
DAD [76]	https://arxiv.org/pdf/2203.16266.pdf	https://github.com/zja-nlp/NAT_with_DAD	Pytorch/Fairseq
CTC [27]	https://www.cs.toronto.edu/~graves/icml_2006.pdf	https://github.com/parlance/ctcdecode	C++
Reinforce-NAT [231]	https://aclanthology.org/P19-1288.pdf	https://github.com/ictnlp/RSI-NAT	Pytorch
BoN [79]	https://ojs.aaai.org/index.php/AAAI/article/view/5351	https://github.com/ictnlp/BoN-NAT	Fairseq
AXE [80]	http://proceedings.mlr.press/v119/ghazvininejad20a/ghazvininejad20a.pdf	https://github.com/m3yrin/aligned-cross-entropy	Pytorch
EISL [81]	https://aclanthology.org/2022.naacl-main.150.pdf	https://github.com/guangyliu/EISL	Pytorch/Fairseq
OAXE [41]	http://proceedings.mlr.press/v139/du21c/du21c.pdf	https://github.com/tencent-ailab/ICML21_OAXE	Pytorch/Fairseq
SAT [17]	https://aclanthology.org/D18-1044.pdf	-	-
SUNDAE [62]	https://openreview.net/pdf?id=10GpzBQ1Fg6	https://github.com/vvvm23/sundae	Pytorch
INSNET [63]	https://openreview.net/pdf?id=vsShetzoRG9	-	-
DA-Transformer [40]	https://proceedings.mlr.press/v162/huang22m/huang22m.pdf	https://github.com/thu-coai/DA-Transformer	Pytorch/Fairseq
DA-Transformer Viterbi [77]	https://aclanthology.org/2022.findings-emnlp.322.pdf	-	-
FA-DAT [78]	https://openreview.net/pdf?id=LSz-gQyd0zE	-	-
ngram-OAXE [82]	https://aclanthology.org/2022.coling-1.446.pdf	-	-
CoCO [33]	https://aclanthology.org/2022.naacl-main.126.pdf	-	-
MgMO [83]	https://aclanthology.org/2022.emnlp-main.339.pdf	-	-
NMLA [84]	https://openreview.net/pdf?id=Qvh0SAl7rYzH	https://github.com/ictnlp/NMLA-NAT	Pytorch/Fairseq
GAD++ [86]	https://arxiv.org/pdf/2203.16487v2.pdf	https://github.com/hemingkx/Generalized-Aggressive-Decoding	Pytorch/Fairseq
HRT [89]	https://openreview.net/pdf?id=2NQ8wlmU9a_	-	-
weak MTL [96]	https://aclanthology.org/2022.emnlp-main.371.pdf	https://github.com/wxy-nlp/MultiTaskNAT	-
RecoverSAT [85]	https://aclanthology.org/2020.acl-main.277.pdf	https://github.com/ranqiu92/RecoverSAT	Pytorch/OpenNMT
Unified [37]	https://aclanthology.org/2020.coling-main.25.pdf	-	-
Diformer [88]	https://aclanthology.org/2022.eamt-1.11.pdf	-	-
imitate-NAT [90]	https://aclanthology.org/P19-1125.pdf	-	-
NAT-HINT [91]	https://aclanthology.org/D19-1573.pdf	https://github.com/zhuohan123/hint-nart	Pytorch
ENGINE [92]	https://aclanthology.org/2020.acl-main.251.pdf	https://github.com/lifu-tu/ENGINE	Pytorch/Fairseq
EM+ODD [93]	http://proceedings.mlr.press/v119/sun20c/sun20c.pdf	https://github.com/Edward-Sun/NAT-EM	Pytorch
FCL-NAT [24]	https://ojs.aaai.org/index.php/AAAI/article/view/6289	https://github.com/lemmonation/fcl-nat	Tensorflow/Tensortotensor
MULTI-TASK NAT [94]	https://aclanthology.org/2021.naacl-main.313.pdf	https://github.com/yongchanghao/multi-task-nat	Pytorch/Fairseq
TCT-NAT [95]	https://www.ijcai.org/Proceedings/2020/0534.pdf	-	-
AB-Net [97]	https://dl.acm.org/doi/pdf/10.5555/3495724.3496634	https://github.com/lemmonation/abnet	Pytorch/Fairseq
NAG-BERT [98]	https://aclanthology.org/2021.eacl-main.18.pdf	https://github.com/yxuansu/NAG-BERT	Pytorch/Fairseq
CeMAT [42]	https://aclanthology.org/2022.acl-long.442.pdf	https://github.com/huawei-noah	Pytorch/Fairseq
XML-D [99]	https://aclanthology.org/2022.emnlp-main.466.pdf	-	-

TABLE 4
A collection of NAR related published papers and codes on general-purpose and task specific text generation tasks.

Method	Paper URL	Code URL	Framework
General-Purpose Text Generation			
POSPD [129]	https://aclanthology.org/2021.acl-long.467.pdf	https://github.com/yangkexin/pospd	Pytorch/Fairseq
MIST [130]	https://arxiv.org/pdf/2110.11115v1.pdf	https://github.com/kongds/mist	Pytorch/Fairseq
BANG [134]	http://proceedings.mlr.press/v139/qi21a/qi21a.pdf	https://github.com/microsoft/BANG	-
EDITCL [141]	https://aclanthology.org/2022.acl-long.520.pdf	-	-
NAG-BERT [98]	https://aclanthology.org/2021.eacl-main.18.pdf	https://github.com/yxuansu/NAG-BERT	Pytorch
EDIT5 [139]	https://arxiv.org/pdf/2205.12209.pdf	-	-
ELMER [142]	https://aclanthology.org/2022.emnlp-main.68.pdf	https://github.com/RUCAIBox/ELMER	Pytorch/Fairseq
Summarization			
NAUS [162]	https://aclanthology.org/2022.acl-long.545.pdf	-	-
NACC [163]	https://openreview.net/pdf?id=KXybrIUJnya	https://github.com/MANGA-UOFA/NACC	Pytorch/Fairseq
Dialogue			
CG-nAR [138]	https://aclanthology.org/2021.emnlp-main.169.pdf	https://github.com/rowitzou/cg-nar	Pytorch/Transformers
NonAR+MMI [137]	https://arxiv.org/pdf/2002.04250v2.pdf	-	-
GL-GIN [232]	https://aclanthology.org/2021.acl-long.15.pdf	https://github.com/yizhen20133868/GL-GIN	Pytorch
SlotRefine [233]	https://aclanthology.org/2020.emnlp-main.152.pdf	https://github.com/moore3930/SlotRefine	Tensorflow
NADST [234]	https://openreview.net/pdf?id=H1e_cC4twS	https://github.com/henryhungle/NADST	PyTorch
LR-Transformer [166]	https://dl.acm.org/doi/abs/10.1145/3459637.3482229	-	-
Grammatical Error Correction			
TtT [136]	https://aclanthology.org/2021.acl-long.385.pdf	https://github.com/lipiji/TtT	Pytorch
BERT-GEC [135]	https://aclanthology.org/2021.wnut-1.46.pdf	https://github.com/utal	Pytorch
MaskCorrect [164]	https://arxiv.org/pdf/2211.13252.pdf	-	-
Text Style Transfer			
NAST [169]	https://aclanthology.org/2021.findings-acl.138.pdf	https://github.com/thu-coai/NAST	-
KD+CL+ID [168]	https://aclanthology.org/2021.emnlp-main.730.pdf	https://github.com/sunlight-ym/nar_style_transfer	-
Controllable Text Generation			
PMI [171]	https://aclanthology.org/2021.findings-acl.330.pdf	-	-
MUCOLA [174]	https://aclanthology.org/2022.emnlp-main.144.pdf	https://github.com/Sachin19/mucoco/tree/sampling	Pytorch
Diffusion-LM [173]	https://openreview.net/pdf?id=3s9IrEsjLyk	https://github.com/XiangLi1999/Diffusion-LM	Pytorch/Transformers
AutoTemplate [172]	https://arxiv.org/pdf/2211.08387.pdf	-	-
Image Captioning			
TIger [178]	https://link.springer.com/chapter/10.1007/978-3-031-20059-5_7	https://github.com/baaaad/ECE	Pytorch
FutureCap [179]	https://dl.acm.org/doi/abs/10.1145/3503161.3547840	https://github.com/feizc/Future-Caption	Pytorch/Transformers
Transformer-DML [182]	https://openreview.net/pdf?id=LMuh9bS4tqF	https://github.com/bladewaltz1/ModeCap	Pytorch/Transformers
UAIC [181]	https://arxiv.org/pdf/2211.16769.pdf	-	-
Question Answering			
NAPG [183]	https://arxiv.org/pdf/2211.03462.pdf	-	-
KECP [184]	https://arxiv.org/pdf/2205.03071.pdf	https://github.com/alibaba/EasyNLP	Pytorch/EasyNLP
Automatic Speech Recognition			
NAR CTC/attention [152]	https://ieeexplore.ieee.org/abstract/document/9746316	-	-
S-CFE CTC [235]	https://ieeexplore.ieee.org/abstract/document/9746770	-	-
CASS-NAT [153]	https://ieeexplore.ieee.org/abstract/document/9413429	-	-
DLP [154]	https://ieeexplore.ieee.org/abstract/document/9414198	-	-
CTC-enhanced [133]	https://ieeexplore.ieee.org/abstract/document/9414694	-	-
Align-Refine [143]	https://aclanthology.org/2021.naacl-main.154.pdf	https://github.com/amazon-research/align-refine	To be released
Align-Denoise [144]	http://dx.doi.org/10.21437/Interspeech.2021-1906	https://github.com/bobchennan/espnet/tree	Pytorch/Espnet
LASO [158]	https://ieeexplore.ieee.org/document/9437636	-	-
NAR-BERT-ASR [132]	https://arxiv.org/pdf/2104.04805v1.pdf	-	-
CondChain [192]	https://arxiv.org/pdf/2106.08595v1.pdf	https://github.com/pengchengguo/espnet	Pytorch/Espnet
Streaming NAR [236]	https://arxiv.org/pdf/2107.09428v1.pdf	https://github.com/espnet/espnet	Pytorch/Espnet
Mask-CTC [147]	http://www.interspeech2020.org/uploadfile/pdf/Thu-1-3-7.pdf	https://github.com/espnet/espnet	Pytorch/Espnet
Intermediate CTC [150]	https://ieeexplore.ieee.org/abstract/document/9414594/	https://github.com/espnet/espnet	Pytorch/Espnet
Self-Conditioned CTC [151]	https://arxiv.org/pdf/2104.02724.pdf	https://github.com/espnet/espnet	Pytorch/Espnet
GIC [155]	https://arxiv.org/pdf/2205.12462.pdf	-	-
CAKT [156]	https://ieeexplore.ieee.org/abstract/document/10022825	-	-
Inter-KD [149]	https://ieeexplore.ieee.org/abstract/document/10022581	-	-
BECTRA [159]	https://arxiv.org/pdf/2211.00792.pdf	-	-

TABLE 5
A collection of other NAR related published papers and codes beyond text generation. IE denotes information extraction task, VG denotes video generation task, VC denotes voice conversion task, and SR denotes speech representation task, CC denotes code completion task, PTSF denotes time series forecasting task.

Method	Paper URL	Code URL	Framework
Text to Speech			
BVAE-TTS [188]	https://openreview.net/pdf?id=o3iritJHLfO	https://github.com/LEEYOONHYUNG/BVAE-TTS	Pytorch
GAN-TTS [192]	https://arxiv.org/pdf/2203.01080.pdf	https://github.com/yanggeng1995/GAN-TTS	Pytorch
VARA-TTS [187]	https://arxiv.org/pdf/2102.06431v1.pdf	https://github.com/vara-tts/VARA-TTS	-
Glow-TTS [193]	https://dl.acm.org/doi/pdf/10.5555/3495724.3496400	https://github.com/jaywalnut310/glow-tts	Tensorflow/Tensor2tensor
VAENAR-TTS [185]	https://arxiv.org/pdf/2107.03298v1.pdf	https://github.com/thuhcsi/VAENAR-TTS	Pytorch
ParaNet [128]	http://proceedings.mlr.press/v119/peng20a/peng20a.pdf	https://github.com/ksw0306/WaveVAE	Pytorch
FastSpeech [127]	https://dl.acm.org/doi/pdf/10.5555/3454287.3454572	https://github.com/coqui-ai/TTS	PyTorch/TTS
TalkNet2 [237]	https://arxiv.org/pdf/2104.08189v3.pdf	https://github.com/rishikksh20/TalkNet2-pytorch	-
FastSpeech2 [238]	https://ieeexplore.ieee.org/abstract/document/9383629	https://github.com/ming024/FastSpeech2	Pytorch
HiMuV-TTS [194]	https://arxiv.org/pdf/2204.04004.pdf	-	-
CLONE [196]	https://arxiv.org/pdf/2207.06088.pdf	-	-
CUC-VAE [195]	https://aclanthology.org/2022.acl-long.30.pdf	-	-
Speech translation			
MTL [198]	https://aclanthology.org/2021.findings-acl.92.pdf	https://github.com/voidism/NAR-ST	Pytorch/Espnet
Orthros [199]	https://ieeexplore.ieee.org/abstract/document/9415093	-	-
Orthros-CMLM [200]	https://arxiv.org/pdf/2109.04411v1.pdf	-	-
Fast-MD [203]	https://ieeexplore.ieee.org/abstract/document/9687894	-	-
Semantic Parsing			
Span Pointer [126]	https://aclanthology.org/2021.findings-emnlp.161.pdf	-	-
LightConv Pointer [125]	https://aclanthology.org/2021.naacl-main.236.pdf	https://github.com/facebookresearch/pytext	Pytorch/Pytest
RNGTr [239]	https://aclanthology.org/2021.tacl-1.8.pdf	https://github.com/idiap/g2g-transformer	Pytorch
Diffusion Models			
WaveGrad [206]	https://openreview.net/pdf?id=NsMLjcFaO8O	-	-
DDPM [205]	https://dl.acm.org/doi/pdf/10.5555/3495724.3496298	https://arxiv.org/pdf/2006.11239.pdf	Tensorflow
D3PMs [209]	https://openreview.net/pdf?id=h7-XixPCAL	-	-
Imagen [210]	https://openreview.net/pdf?id=08Yk-n5l2Al	-	-
unCLIP [211]	https://arxiv.org/pdf/2204.06125.pdf	-	-
GLIDE [212]	https://proceedings.mlr.press/v162/nichol22a/nichol22a.pdf	https://github.com/openai/glide-text2im	Pytorch
classifier-free guidance [240]	https://openreview.net/forum?id=qw8AKxfYbI	-	-
LDEBM [216]	https://proceedings.mlr.press/v162/yu22h/yu22h.pdf	-	-
Diffusion-LM [173]	https://openreview.net/pdf?id=3s9IrEsjLyk	https://github.com/XiangLi1999/Diffusion-LM	Pytorch/Transformers
DIFFUSER [217]	https://arxiv.org/pdf/2210.16886.pdf	-	-
DIFFUSEQ [215]	https://arxiv.org/pdf/2210.08933.pdf	https://github.com/Shark-NLP/DiffuSeq	Pytorch/Transformers
DiffGAR [241]	https://arxiv.org/pdf/2210.08573.pdf	-	-
Others			
MacroIE [219](IE)	https://aclanthology.org/2021.emnlp-main.764.pdf	-	-
DIGAN [220](VG)	https://openreview.net/pdf?id=Czsdv-S4-w9	-	-
FastSpeech2-VC [221](VC)	https://ieeexplore.ieee.org/abstract/document/9413973	-	-
CDHVAE [242](VC)	https://ieeexplore.ieee.org/abstract/document/9689369	-	-
NPC [224](SR)	https://arxiv.org/pdf/2011.00406v1.pdf	https://github.com/Alexander-H-Liu/NPC	Pytorch
SANAR [225](CC)	https://arxiv.org/pdf/2204.09877.pdf	-	-
MANF [227](TSF)	https://arxiv.org/pdf/2205.07493.pdf	-	-