What Have We Achieved on Non-autoregressive Translation?

Anonymous ACL submission

Abstract

Recent advances have made non-autoregressive (NAT) translation comparable to autoregressive methods (AT). However, their evaluation using BLEU has been shown to weakly correlate with human annotations. Limited research compares non-autoregressive translation and autoregressive translation comprehensively, leaving uncertainty about the true proximity of NAT to AT. To address this gap, we systematically evaluate four representative NAT methods across various dimensions, including human evaluation. Our empirical results demonstrate that despite narrowing the performance gap, state-of-the-art NAT still underperforms AT under more reliable evaluation metrics. Furthermore, we discover that explicitly modeling dependencies is crucial for generating natural language and generalizing to out-of-distribution sequences.[1]

1 Introduction

Non-autoregressive translation, where the model generates translations in parallel, demonstrates notable decoding speed advantages compared with traditional autoregressive translation (Vaswani et al., 2017) and large language models for translation (OpenAI, 2023). However, it suffers from performance degradation compared to autoregressive counterparts (Gu et al., 2017). The degradation stems from the independence assumption, which ignores the inter-token language dependency on the target side. Various methods are proposed to mitigate the performance gap (Ghazvininejad et al., 2019; Qian et al., 2020; Saharia et al., 2020; Du et al., 2021; Li et al., 2022; Huang et al., 2022b,c).

Although representative methods (Saharia et al., 2020; Li et al., 2022; Huang et al., 2022b,c) have reported comparable translation performance to AT, almost all NAT methods are evaluated under BELU scores (Papineni et al., 2002). Although BLEU has been long adopted, recent work (Freitag et al., 2022) argues that it is not a reasonable metric, considerably underperforming alternative metrics such as COMET (Rei et al., 2020) or large language model evaluation (Kocmi and Federmann, 2023). Limited work has been devoted to a systematic evaluation of advanced NAT against AT, leaving a significant gap in the research literature.

To address this gap, we conduct a comprehensive evaluation of representative NAT methods, aiming to reveal existing limitations and provide insights for future research. Our primary focus is on fully non-autoregressive methods which generate translations in a one-shot manner, achieving the most decoding efficiency advantage. We consider MgMO (Li et al., 2022) for advanced optimization, CTC (Saharia et al., 2020) for modeling latent alignment, and DAT (Huang et al., 2022b) for explicit target-side dependency modeling. CMLM (Ghazvininejad et al., 2019) is adopted as the representative iterative NAT method. All models are tested on representative benchmark datasets under a comprehensive evaluation , including rule-based metrics, model-based metrics and GPT4-based metrics (Kocmi and Federmann, 2023). Moreover, we conduct human evaluation under the MQM framework (Freitag et al., 2021) to gain further insights into the performance of NAT models that may be overshadowed by global automatic evaluations.

Automatic evaluation demonstrates varying degrees of advantage for AT over NAT models. In general, DAT achieves the most competitive performance, followed by MgMO and CTC. Under rule-based evaluation metrics such as BELU and chrf (Popovic, 2015), DAT can achieve comparable or even superior performance compared to AT. However, this competitiveness diminishes when using model-based metrics such as COMET (Rei et al., 2020) or GPT4-based evaluation, under which AT significantly outperforms all NAT models. Fine-grained human evaluation indicates that

[1]We release our resources on https://anonymous.com.

NAT models incorporating explicit dependency modeling (e.g., DAT and CMLM) achieve similar levels of translation fluency with AT, yet suffering various translation accuracy errors. Compared with AT, NAT tends to produce more *grammar* or *punctuation* errors. Models without explicit dependency modeling (MgMO and CTC) suffer the most *mistranslation* and *omission* errors. On the other hand, models with latent alignments (CTC and DAT) are more prone to *spelling* and *addition* errors.

Most of these errors are due to NAT's inadequate dependency modeling. Specifically, DAT's addition errors occur when it generates repeated translations, known as n-gram repetition. This can be easily overlooked by BLEU evaluation, which measures n-gram precision, explaining why DAT performs well in terms of BLEU but not COMET. The n-gram repetition mainly stems from the weak, though explicit, dependency modeling. DAT limits its inter-token dependency within one step using a one-linear-layer attention module for decoding efficiency. In contrast, AT can depend on the entire generation history and encode it with powerful Transformer blocks. To validate our assumption, we train an asymmetric AT with a one-layer decoder and observe similar n-gram repetitions. Furthermore, adding an additional linear layer to the transition attention in DAT effectively reduces the repetition, corroborating our hypothesis.

Apart from translation quality, we compare AT with NAT from the perspective of generalization and robustness. Empirical findings demonstrate that explicit dependency modeling is crucial for generating human-like languages and generalizing to out-of-distribution samples, which NAT methods lack or are still weak at. On the other hand, weak dependency exhibits stronger robustness to input perturbations, as it is less affected by exposure bias (Bengio et al., 2015; Ranzato et al., 2016). Future research on NAT should focus on how to consolidate explicit language dependency while maintaining decoding efficiency.

2 Method

We begin with a brief introduction to autoregressive and non-autoregressive machine translation, before introducing four representative NAT methods.

2.1 Neural Machine Translation

The machine translation task can be formally defined as a sequence-to-sequence generation problem, where the model generates the target language sequence $\mathbf{y} = \{y_1, y_2, ..., y_T\}$ from the target vocabulary $\mathcal{V}$, given the source language sequence $\mathbf{x} = \{x_1, x_2, ..., x_S\}$ based on the conditional probability $p_\theta(\mathbf{y}|\mathbf{x})$ (θ denotes the model parameters).

Autoregressive Translation. Autoregressive neural machine translation factorizes the conditional probability to $\prod_{i=1}^{T} p(y_i|y_1, ..., y_{t-1}, \mathbf{x})$, where the model is trained in a teacher-forcing way with cross-entropy (XE):

$$\mathcal{L}_{\text{AT}} = -\log p(\mathbf{y}|\mathbf{x}) = -\sum_{i=1}^{T} \log p_\theta(y_i|\mathbf{x}, y_{<i}). \quad (1)$$

During inference, the model sequentially generates tokens based on previous predictions.

Non-autoregressive Translation. In contrast, non-autoregressive machine translation (Gu et al., 2017) ignores the dependency between target tokens and factorizes the probability as $\prod_{i=1}^{T} p(y_i|\mathbf{x})$, where tokens at each time step are predicted independently. Vanilla NAT models are optimized with XE loss with target dependency ignored:

$$\mathcal{L}_{\text{NAT}} = -\log p(\mathbf{y}|\mathbf{x}) = -\sum_{i=1}^{T} \log p_\theta(y_i|\mathbf{x}), \quad (2)$$

with an additional loss for length prediction:

$$\mathcal{L}_{\text{length}} = -\log p_\theta(T|\mathbf{x}). \quad (3)$$

Challenges of NAT. The major difficulty of non-autoregressive translation lies in that the decoder side relies solely on the source-side information without any target inputs, e.g., history predictions in AT. Autoregressive models utilize previous token predictions to select the next token from the distribution over the whole vocabulary space:

$$p_\theta(y_i|y_{<i}, \mathbf{x}) = \text{softmax}(\mathbf{W}_{\text{P}}\text{Transformer}(y_{<i}, \mathbf{x}), \quad (4)$$

where $\mathbf{W}_{\text{P}}$ is the vocabulary projection weight. The inter-token dependency involves layers of Transformer blocks. In contrast, NAT models generate translations in a "one-shot" manner, ignoring or weakening the strong language dependency on the target side. As a result, vanilla NAT is not capable of properly modeling the highly multi-modal distribution of target translations, i.e., a source sentence can have multiple valid translations. Various methods aim to alleviate the conditional independence assumption. In this work, we consider four

representative methods: (1) alternative optimization with model architecture unchanged (Li et al., 2022); (2) introducing latent alignments based on an upsampled decoder prediction (Saharia et al., 2020); (3) building shallow but explicit target-side dependency (Huang et al., 2022b); and (4) iterative decoding (Ghazvininejad et al., 2019).

2.2 NAT with Advanced Optimization

Instead of exerting token-by-token cross-entropy supervision, Li et al. (2022) propose multi-granularity optimization (MgMO) to collect multi-granularity feedback on generations sampled from the models and gather them for backpropagation:

$$\mathcal{L}_{\text{MO}} = -\sum_{k=1}^{K} q_\theta(\mathbf{h}^k|\mathbf{x}) R(\mathbf{h}^k, \mathbf{y}^k), \quad (5)$$

where K is the sample space size. $q_\theta(\mathbf{h}^k|\mathbf{x})$ is defined as the normalized probability for each hypothesis $\mathbf{h}^k$:

$$q_\theta(\mathbf{h}^k|\mathbf{x};\alpha) = \frac{\hat{p}_\theta(\mathbf{h}^k|\mathbf{x})^\alpha}{\sum_{\mathbf{h}'\in\mathcal{K}(\mathbf{x})} \hat{p}_\theta(\mathbf{h}'|\mathbf{x})^\alpha}, \quad (6)$$

where $\mathcal{K}(\mathbf{x})$ denotes the sample space and α controls the distribution sharpness. $R(\mathbf{h}, \mathbf{y})$ is a reward function that encourages the generations to be similar with references under various granularity. MgMO requires no architecture modification and thus maintains decoding efficiency.

2.3 NAT with Latent Alignments

Saharia et al. (2020) introduce latent alignment models, e.g., Connectionist Temporal Classification (CTC) (Graves et al., 2006), to mitigate the target-side independence assumption. CTC utilizes a sequence of discrete latent alignment variables to monotonically align the non-autoregressive predictions of the model and target side tokens. The marginal probability over latent alignments $\mathbf{a}$ is derived as:

$$\begin{aligned}\mathcal{L}_{\text{LA}} &= -\log p_\theta(\mathbf{y}|\mathbf{x}) \\ &= -\log \sum_{\mathbf{a}\in\beta(\mathbf{y})} p_\theta(\mathbf{y}|\mathbf{a},\mathbf{x}) p_\theta(\mathbf{a}|\mathbf{x}),\end{aligned} \quad (7)$$

where $\beta(\mathbf{y})$ is a function that returns all possible alignments for a sequence $\mathbf{y}$. Then $\mathbf{a} = \{a_1, \ldots, a_M\}$ is predicted by the decoder output states $\mathbf{H} = \{\mathbf{h}_1, \ldots, \mathbf{h}_M\}$, where $a_i \in \mathcal{V} \cup \{\text{“\_”}\}$. “_” is a special blank token to allow many-to-one and null alignment. For instance, for a target sequence “thank you”, valid alignments $\mathbf{a}$ include “_ thank thank you” and “thank _ you _”. The decoder state length is set as several times the source sequence length to allow long translations. The alignment probability $p_\theta(\mathbf{a}|\mathbf{x})$ is derived by:

$$\begin{aligned}p_\theta(\mathbf{a}|\mathbf{x}) &= \prod_{i=1}^{M} p_\theta(a_i|\mathbf{x}) \\ &= \prod_{i=1}^{M} \text{softmax}(\mathbf{W}_{\text{P}}\mathbf{h}_i).\end{aligned} \quad (8)$$

Since $a_i \in \mathcal{V} \cup \{\text{“\_”}\}$, the posterior probability of $\mathbf{y}$ becomes:

$$p_\theta(\mathbf{y}|\mathbf{a},\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{a} \in \beta(\mathbf{y}) \\ 0 & \text{otherwise.} \end{cases} \quad (9)$$

MgMO and CTC avoid token-by-token CE supervision by introducing segment-level optimization or marginalizing latent alignments. However, they suffer independence assumption in generating tokens (Equation 2) or alignments (Equation 8). Consequently, both MgMO and CTC cannot inherently handle multi-modal problems and heavily rely on techniques such as knowledge distillation (Zhou et al., 2020a) to mitigate this limitation.

2.4 NAT with Explicit Dependency

Huang et al. (2022b) propose directed Acyclic Transformer (DAT) to construct explicit dependencies, by formalizing an alignment as a path in a direct acyclic graph. Similar to CTC, the decoder state length is upsampled to M and $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_M]$ denotes the decoder output hidden states, which are defined as the vertex states. The probability of path $\mathbf{a}$ is redefined as the position transition probability:

$$p_\theta(\mathbf{a}|\mathbf{x}) = \prod_i p_\theta(a_{i+1}|a_i,\mathbf{x}) = \prod_i \mathbf{E}_{a_i,a_{i+1}},$$

where $\mathbf{E} \in \mathbb{R}^{M\times M}$ is the transition matrix normalized by rows. $\mathbf{a} = \{a_1, a_2, \ldots, a_T\}$ is a possible path represented by a sequence of vertex indexes of the vertex states $\mathbf{H}$, i.e., $a_i \in \{1, 2, 3, \ldots, M\}$. Specifically, the transition matrix is obtained by:

$$\mathbf{E} = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}), \quad (10)$$

$$\mathbf{Q} = \mathbf{H}\mathbf{W}_{\text{Q}}, \quad \mathbf{K} = \mathbf{H}\mathbf{W}_{\text{K}},$$

where d is the hidden size, $\mathbf{W}_{\mathrm{Q}}$ and $\mathbf{W}_{\mathrm{K}}$ are learnable matrices. Conditioned on the vertex states in $\mathbf{H}$ and the selected path $\mathbf{a}$, the posterior probability of $\mathbf{y}$ is computed as:

$$p_\theta(\mathbf{y}|\mathbf{a},\mathbf{x}) = \prod_{i=1}^{T} P_\theta(y_i|a_i,\mathbf{x}) = \prod_{i=1}^{T} \mathrm{softmax}(\mathbf{W}_{\mathrm{p}}\mathbf{h}_{a_i}), \quad (11)$$

where $\mathbf{h}_{a_i}$ is the representation of the i-th vertex on the path $\mathbf{a}$.

Different from previous NAT methods, DAT explicitly models token dependencies through vertex transitions. DAT first parallelly predicts a subset of all possible tokens for translating the source sentence and stores it as $\mathbf{H}$, whose size is usually several times (e.g., 8) that of the source sequence. In contrast to Equation 4, the inter-token dependency is a one-step local transition for each vertex $\mathbf{h_i}$, to determine the next token from the rest of the set, i.e., $\{\mathbf{h_{i+1}}, \ldots, \mathbf{h_M}\}$:

$$p_\theta(\mathbf{y}) = \prod_{i=1}^{T} p_\theta(y_{a_i}|y_{a_{i-1}}), \quad (12)$$

$$p_\theta(y_{a_i}|y_{a_{i-1}}) = \mathrm{softmax}(\mathbf{W}_{\mathrm{P}}\mathbf{h}_{\mathrm{argmax}(\mathbf{E}_{a_{i-1},a_i})}), \quad (13)$$

where y_{a_i} is the predicted token of the i-th vertex on the path $\mathbf{a}$[2]. The explicit though weak dependency modelled by one-layer linear weights $\mathbf{W}_{\mathrm{Q}}$ and $\mathbf{W}_{\mathrm{K}}$ alleviate the necessity of knowledge distillation, yet suffering n-gram repeating issues (discussed in Section 4.3).

2.5 NAT with Iterative Refinement

The iterative NAT model (Ghazvininejad et al., 2019) is typically trained with conditional masked language modeling (CMLM) to build inter-token dependencies:

$$\mathcal{L}_{CMLM} = -\sum_{y_t \in \mathcal{Y}(\mathbf{y})} log p_\theta(y_t|\Omega(\mathbf{y},\mathcal{Y}(\mathbf{y})),\mathbf{x}), \quad (14)$$

where $\mathcal{Y}(\mathbf{y})$ is a randomly selected subset of target tokens and Ω denotes a function that masks a selected set of tokens in $\mathcal{Y}(\mathbf{y})$. During decoding, starting from a sequence of initiative tokens, e.g., "<unk>", CMLM models iteratively refine translations from previous iterations to generate target language sequences.

[2]We omit conditional dependency on $\mathbf{x}$ for simplicity.

3 Experiment and Setup

Datasets and Models. We conduct experiments on WMT16 En⇒Ro and WMT21 De⇒En with 4 representative NAT methods apart from the vanilla NAT and AT. For knowledge distillation, We train an autoregressive model on the raw data as the teacher model to generate the distilled dataset. Details can be found in Appendix C.

Evaluation. For translation quality, we adopt four commonly used metrics, which include two rule-based metrics, i.e., BLEU score(Papineni et al., 2002) and chrf (Popovic, 2015), and two model-based metrics, i.e., COMET (Rei et al., 2020) and BLEURT (Sellam et al., 2020). Specifically, for COMET, we utilize the wmt22-comet-da model (Rei et al., 2022), and for BLEURT, the BLEURT-20 model (Pu et al., 2021) is employed. Kocmi and Federmann (2023) propose a GPT-based metric, namely GEMBA, to evaluate translation quality, and demonstrate state-of-the-art correlation with human labels. We adopt GEMBA-GPT4-DA based on GPT-4 (OpenAI, 2023) as an advanced evaluation metric. For human evaluation, we follow (Freitag et al., 2021), an evaluation methodology based on the Multidimensional Quality Metrics (MQM) framework, which provides a hierarchical analysis of translation errors. Human evaluation details can be found in Appendix D.

4 Translation Quality

4.1 Automatic Evaluation

The automatic evaluation results on WMT16 En⇒Ro and WMT21 De⇒En are presented in Table 1. DAT obtains the most competitive performance compared with the AT counterpart across all automatic metrics, followed by MgMO and CTC. MgMO and CTC achieve stronger performance than the representative iterative method, CMLM, when considering COMET and GEMBA which have shown better correlation with human annotation (Rei et al., 2020; Kocmi and Federmann, 2023). Notably, MgMO obtains comparable performance with CTC, without modifying model architecture.

Reliance on Knowledge Distillation. In both translation directions, all fully non-autoregressive methods except DAT and CMLM suffer more from training without distillation. Typically, the vanilla NAT models suffer a decrease of more than 7 BLEU points without KD. For strong NAT methods such as MgMO and CTC, on WMT21 De⇒En, the

Model	BLEU↑	chrf↑	COMET↑	BLEURT↑	GEMBA↑	Speed↑
WMT16 En⇒Ro						
w/o Knowledge Distillation						
AT (Vaswani et al., 2017)	**34.39**†	**58.48**	**78.90**†	**69.89**†	**86.18**†	1.0×
NAT (Gu et al., 2017)	23.75	50.72	65.78	53.91	67.29	15.9×
MgMO (Li et al., 2022)	30.97	56.65	73.54	63.19	80.04	14.9×
CTC (Saharia et al., 2020)	32.73	57.77	74.26	63.99	81.16	14.5×
DAT (Huang et al., 2022b)	33.18	57.35	76.14	66.72	83.72	13.8×
CMLM (Ghazvininejad et al., 2019)	31.97	56.78	74.11	63.34	78.72	2.7×
w/ Knowledge Distillation						
AT (Vaswani et al., 2017)	**33.92**	**58.45**	**78.49**†	**69.26**†	**86.22**†	1.0×
NAT (Gu et al., 2017)	30.97	56.52	72.60	62.12	77.47	15.8×
MgMO (Li et al., 2022)	32.86	57.40	75.52	65.36	82.73	14.9×
CTC (Saharia et al., 2020)	33.28	58.28	75.54	65.71	82.94	14.5×
DAT (Huang et al., 2022b)	33.25	57.89	76.59	67.01	84.27	13.7×
CMLM (Ghazvininejad et al., 2019)	32.71	56.76	72.36	63.42	76.67	2.7×
WMT21 De⇒En						
w/o Knowledge Distillation						
AT (Vaswani et al., 2017)	**31.89**	**60.25**	**84.26**†	**71.94**†	**92.91**†	1.0×
NAT (Gu et al., 2017)	16.85	43.46	55.87	44.80	36.94	15.5×
MgMO (Li et al., 2022)	28.89	58.11	77.76	64.08	83.51	13.8×
CTC (Saharia et al., 2020)	27.35	56.53	75.38	61.79	79.14	13.5×
DAT (Huang et al., 2022b)	31.69	59.60	81.12	69.00	88.29	13.1×
CMLM (Ghazvininejad et al., 2019)	29.36	57.79	76.39	65.00	82.07	2.4×
w/ Knowledge Distillation						
AT (Vaswani et al., 2017)	32.04	**60.85**	**84.72**†	**72.53**†	**93.39**†	1.0×
NAT (Gu et al., 2017)	27.55	56.56	75.50	62.72	76.89	15.2×
MgMO (Li et al., 2022)	30.32	59.36	81.15	67.76	89.17	13.8×
CTC (Saharia et al., 2020)	30.52	59.83	80.06	67.24	86.91	13.4×
DAT (Huang et al., 2022b)	**32.26**	60.80	83.32	71.44	92.05	13.1×
CMLM (Ghazvininejad et al., 2019)	30.25	58.40	77.14	65.63	82.98	2.4×

Table 1: Automatic evaluation results of different translation models on WMT16 En⇒Ro and WMT21 De⇒En, considering both raw data and distillation data settings. We encompass a wide range of metrics including rule-based metrics (BLEU and chrf), model-based metrics (COMET and BLEURT) and LLM-based metrics (GEMBA). Bold numbers represent the best performance and underlined numbers denote the top 3 performance. † denotes translation quality of AT is significantly better than all other NAT models with a $p < 0.01$ (Koehn, 2004).

BLEU scores decrease by more than 2 and 3 points, respectively. On the contrary, the performance of DAT and CMLM is as similarly affected as the AT counterpart, due to explicit dependency modeling similar to AT. In the subsequent sections, we utilize knowledge distillation by default to analyze NAT models in the best-performing setting.

Evaluation Metrics. We consider a set of representative metrics to comprehensively compare NAT methods with AT. We perform significance tests on all pairs of NAT models and their AT counterparts across all metrics. Except for DAT, current NAT methods significantly underperform AT methods in various evaluation metrics including rule-based (BLEU and chrf), model-based (COMET and BLEURT), and GPT4-based metrics (GEMBA), particularly in the raw data setting. A notable observation is that DAT models are more competitive with AT models when evaluated using rule-based metrics, which assess the similarity between generated text and references. In contrast, AT models outperform DAT models significantly under model-based metrics or GPT4 evaluation (GEMBA). These metrics evaluate translation quality by measuring semantic similarity between two sentences based on parametric knowledge. To gain a deeper understanding of this phenomenon, we conduct human evaluation using a systematic and fine-grained framework, i.e., MQM (Freitag et al., 2021), to further compare NAT with AT.

4.2 Human Evaluation

The evaluation results, obtained by averaging the error counts from three translators, are presented in Table 2. We omit human evaluation on the vanilla

Model	MQM↓	FLC. Err↓	ACC. Err↓	NON. Err↓
AT	176.67	34.33	142.00	0.33
MgMO	301.33	52.00	240.00	9.33
CTC	360.67	63.33	280.67	16.67
DAT	229.33	38.00	183.67	0.33
CMLM	375.67	47.33	153.33	175.00

Table 2: Human evaluation results under MQM framework. MQM denotes weighted error counts of three major error types: fluency (FLC.), accuracy (ACC.) and non-translation (NON.).

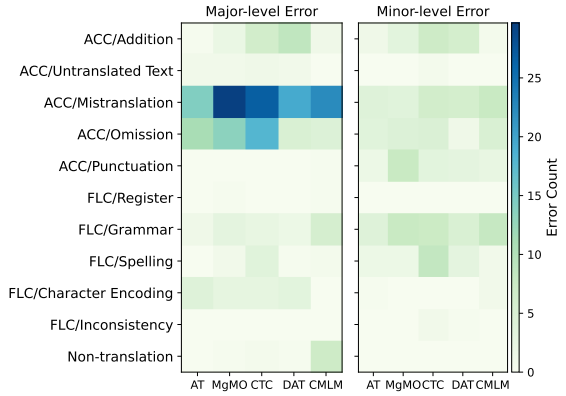

Figure 1: Heatmap visualization of MQM evaluation: darker colours indicate larger error counts for certain error types. The left side presents major-level errors while the right side shows minor-level errors.

NAT due to its poor performance under automatic evaluation. The performance ranking of human evaluation aligns with the automatic evaluation: AT performs the best, followed by DAT, MgMO, CTC, and CMLM. Models with explicit dependency modeling (AT, DAT and CMLM) generate more fluent translations than those without (MgMO and CTC). , with fewer fluency errors. Despite comparable fluency to AT, DAT exhibits low translation accuracy. All NAT methods, particularly CMLM, generate non-translations in certain cases.

A fine-grained error visualization is presented in Figure 1. The *mistranslation* error type at the major level has the highest proportion among all models, with the models lacking explicit dependency (MgMO and CTC) producing the most errors. AT performs generally better than NAT except for a considerable number of *omission* errors. In contrast, NAT models tend to generate translations with additional or duplicated content (*addition*), particularly CTC and DAT which increase decoder length to model latent alignments. These two models also exhibit more *spelling* errors. Compared to AT, NAT models tend to produce more *punctuation* errors and *grammar* errors. Similar to AT, MgMO and CTC translations also frequently lack partial source content (*omission*).

Ref.	AT	NAT	MgMO	CTC	DAT	CMLM
0.00	0.50	27.64	16.47	1.85	0.00	14.52
0.00	0.10	31.10	23.50	1.60	0.00	12.60

Table 3: Uni-gram repetition ratios on WMT16 En⇒Ro (first row) and WMT21 De⇒En (second row). The term "Ref." refers to the reference translation

We explore human annotations to understand typical patterns. Regarding *omission* errors, AT often exhibits incomplete generation at the sentence's end. On the other hand, MgMO and CTC frequently omit content throughout the entire sentence, such as missing adjectives or verbs. The NAT's *grammar* errors primarily stem from incorrect verb tense and singular/plural usage, resulting from its limited language dependency modeling. The case study indicates that, for CTC, the major *addition* errors are attributed to generating words with *spelling* errors, which are regarded as irrelevant content by annotators. For DAT, these *addition* errors stem from **n-gram repetition**, where the model generates a repeated segment from the previous context. For example, "By *the beginning of November*, there are seven races until *the beginning of November*." To give an intuitive representation, we present several cases for the aforementioned error types in Appendix G. All these patterns can be attributed to inadequate language dependency modeling with limited or redundant decoding length.

4.3 Effects of Explicit Dependency

Repetition Ratio. We first examine token repetition ratio (Zhou et al., 2020a; Ghazvininejad et al., 2020; Du et al., 2021) in model translations, which is the ratio of generations with repeated tokens, e.g., "He is is a lawyer". The results are shown in Table 3. We can observe that models without latent alignment modeling (NAT, MgMO and CMLM) suffer severe token repetition during generation.

N-gram Repetition. Besides consecutive uni-gram repetition, a more subtle phenomenon is non-adjacent n-gram repetition. Such a repetition can be overlooked under traditional metrics such as BLEU score, which only calculates the n-gram precision of the generations. Consequently, translations that contain n-gram repetition may even achieve higher BLEU scores. This could explain why DAT performs better than AT under rule-based metrics but

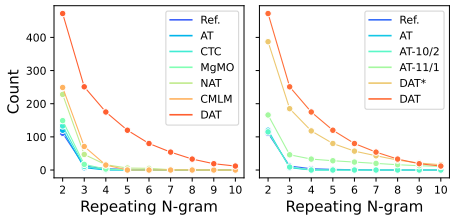

Figure 2: N-gram repetition of different models (WMT21 De⇒En), where the x-axis represents the size of the n-gram and the y-axis represents the count.

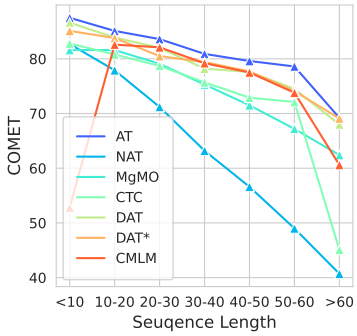

Figure 3: Translation quality (COMET) w.r.t. source sequence length on WMT21 De⇒En.

not under model-based or GPT4-based metrics. We collect the n-gram repetition count for each model, as shown in the left part of Figure 2. We can observe that DAT demonstrates a stronger tendency to generate repeating n-grams with higher counts across various n-gram granularity (2 to 10), which aligns with a substantial number of *addition* errors in human evaluation.

Enhancing Dependency Modeling. DAT utilizes one-linear-layer attention modules to model local vertex transitions. Such explicit dependency modeling can be limited when dealing with long sequence generation. For example, consider the sentence "By *the beginning of November*, there are seven races until *the beginning of November*." In this case, both the beginning and the end of the sentence are valid positions for the temporal prepositional phrase "the beginning of November" In DAT, at the vertex state corresponding to the token "races," only information from that current vertex state is used to determine the index of the next vertex state using one-linear-layer attention layers. In contrast, AT considers all previously generated tokens and utilizes Transformer decoder layers to determine the next token. Under weak dependency modeling in DAT, early generations can be ignored and repeated phrases can be falsely pointed to (e.g., "the beginning of November"). To validate this assumption, we train an asymmetrical AT model with a shallow decoder to simulate weak dependency modeling, and a deep encoder to guarantee model size. As shown in Figure 2 (right-side), AT-11/1 (11-layer encoder and 1-layer decoder) also tends to generate repeated n-grams, and adding one decoder layer (AT-10/2) mitigates this issue. Nevertheless, AT-11/1 performs better than DAT as it relies on the entire generation history rather than just considering the current token. To alleviate this issue without influencing decoding efficiency, we introduce an additional linear layer for both $\mathbf{Q}$ and $\mathbf{K}$ to strengthen the token transition modeling. This refined model is referred as DAT* (Appendix H). With 0.7% additional parameters compared to DAT alone, we observe that DAT* exhibits less n-gram repetition while maintaining decoding speed. We present a case study of how DAT* alleviates n-gram repetition in Appendix G. Nevertheless, DAT* is limited in a one-step local transition foundation and cannot fundamentally resolve n-gram repetition.

5 Generalization and Robustness

Length Generalization. Figure 3 illustrates that all models experience a decline in performance as the length of the source sequence increases, albeit at varying rates. AT surpasses all NAT methods in length generalization. Models incorporating explicit dependency (e.g., AT, DAT, and CMLM) exhibit slower degradation compared to others. Notably, CTC and CMLM experience severe performance drops on sequences longer than 60.

Cross-domain Generalization. Figure 4 illustrates the cross-domain performance averaged across 5 domains. Models with explicit dependency, such as AT and DAT, achieve high cross-domain performance. On the other hand, CTC and CMLM demonstrate substantial degradation in performance when tested on out-of-domain datasets This is due to CTC models generating spelling errors and CMLM models propagating errors from early steps. These issues are further exacerbated in cross-domain testsets that contain more terminologies, leading to subpar performance. The complete results can be found in Appendix I.

Compositional Generalization. We measure compositional generalization on GoGnition (Li et al., 2021) which evaluates the ability to translate unseen phrases of simple and known semantic

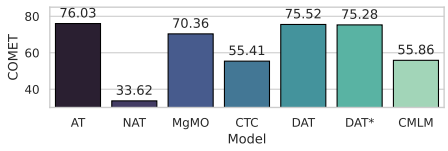

Figure 4: Average cross-domain performance (COMET) of WMT21 De⇒En models on out-of-domain testsets.

Model	AT	DAT	DAT*
Instance-level CTER↓	28.42%	43.66%	42.52%
Aggregate-level CTER↓	62.88%	79.49%	79.12%

Table 4: Compositional generalization performance.

units. The results are shown in Table 4 [3]. Instance-level CTER and Aggregate-level CTER denote the compound translation error rates of translating novel compounds. Despite the narrowing gap in in-domain and out-of-domain testsets, we observe a significant difference in compositional generalization between DAT and AT. This discrepancy is reflected in higher error rates, indicating a disparity in dependency modeling capabilities.

Robustness to Input Perturbations. Finally, we explore models' robustness to different input perturbations, including random replacement, deletion and permutation (Details in Appendix C), with results shown in Figure 5. In contrast to previous findings that suggest explicit modeling provides advantages, models without explicit incorporation of modeling (e.g., MgMO and CTC) are less affected by input noises. This can be because explicit dependency generation may introduce exposure bias (Bengio et al., 2015; Ranzato et al., 2016), where errors occurring at early time steps (AT and DAT) or iterative steps (CMLM) can accumulate and propagate into future predictions, making them susceptible to input perturbations. For complete results, please refer to Appendix J.

To the best of our knowledge, this is the first comparison of NAT and AT in terms of generalization and robustness. In addition to the disparity in translation performance on benchmark datasets, inadequate language dependency modeling causes NAT methods to significantly lag behind AT. However, this weak dependency does provide an advantage in resisting input perturbations.

[3]We only evaluate AT and DAT as they do not rely on knowledge distillation.

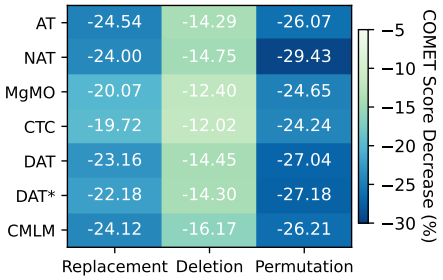

Figure 5: Translation performance (COMET) decreases (%) on noisy testsets of WMT21 De⇒En, with darker colours indicating greater degradation.

6 Related Work

We discuss several representative NAT methods in Section 2. A more detailed discussion on NAT advances is presented in Appendix A. Different from surveys (Xiao et al., 2023; Li et al., 2023) that conduct a comprehensive survey on recent NAT advances, we focus on comparing NAT with AT comprehensively. Our work is also related to previous work analyzing neural machine translation (Appendix B). Zhou et al. (2020a) find that knowledge distillation boosts NAT performance by reducing data complexity. Agrawal et al. (2022) discuss knowledge transfer in the context of multilingual NAT. Huang et al. (2022a) understand the learning process of NAT both theoretically and empirically. Differently, we focus on systematically comparing common NAT techniques with their AT counterparts in a systematic manner to showcase existing performance gaps for future research.

7 Conclusion

We compared representative NAT methods with AT under a comprehensive evaluation that encompasses a set of evaluation dimensions, including human evaluation. Our research aims to fill in the research gap of the real competitiveness of NAT to AT. Both automatic and human evaluations indicated that despite the narrowing gap, NAT methods underperform AT, with varying error patterns such as translation omission, spelling errors and n-gram repetitions. Our empirical results and analyses demonstrated that explicit dependency modeling is crucial for generating human-like languages, although strong dependence can suffer explore bias. Future research on NAT should focus on how to consolidate explicit language dependency while maintaining decoding efficiency.

Limitations

We systematically evaluate NAT and AT, highlighting performance gaps for future research. However, there are limitations: Firstly, we assess state-of-the-art NAT models using research-oriented datasets (WMT, OOD, CG), which mainly consist of English-centric text with a formal style and limited topic range. Secondly, each NAT model is annotated with only 100 samples. This may not cover all potential error types. Finally, we focus primarily on fully non-autoregressive methods due to their superior decoding efficiency. Our results are also limited to training-from-scratch methods; extending conclusions to large language models is left for future work.

Ethical Considerations

We honor the ACL Code of Ethics. No private data or non-public information is used in this work. For human annotation, we hired three annotators who have degrees in English Linguistics or Applied Linguistics. Before formal annotation, annotators were asked to annotate 100 samples randomly extracted from the dataset, and based on average annotation time we set a fair salary (i.e., 32 dollars per hour) for them. During their training annotation process, they were paid as well. The annotation does not involve any personally sensitive information. The annotation strictly follows the annotation guide of MQM (Freitag et al., 2021), with details presented in Appendix D. We adhere to the terms of companies offering commercial LLM APIs and express our gratitude to all global collaborators for their assistance in utilizing these APIs.

References

Sweta Agrawal, Julia Kreutzer, and Colin Cherry. 2022. Exploring the benefits and limitations of multilinguality for non-autoregressive machine translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 177–187, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yu Bao, Shujian Huang, Tong Xiao, Dongqi Wang, Xinyu Dai, and Jiajun Chen. 2021. Non-autoregressive translation by learning target categorical codes. In *NAACL-HLT*, pages 5749–5759. Association for Computational Linguistics.

Yu Bao, Hao Zhou, Shujian Huang, Dongqi Wang, Lihua Qian, Xinyu Dai, Jiajun Chen, and Lei Li. 2022. GLAT: glancing at latent variables for parallel text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8398–8409. Association for Computational Linguistics.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1171–1179.

Liang Ding, Longyue Wang, Xuebo Liu, Derek F. Wong, Dacheng Tao, and Zhaopeng Tu. 2021a. Progressive multi-granularity training for non-autoregressive translation. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 2797–2803. Association for Computational Linguistics.

Liang Ding, Longyue Wang, Xuebo Liu, Derek F. Wong, Dacheng Tao, and Zhaopeng Tu. 2021b. Rejuvenating low-frequency words: Making the most of parallel data in non-autoregressive translation. In *ACL/IJCNLP*, pages 3431–3441. Association for Computational Linguistics.

Liang Ding, Longyue Wang, Xuebo Liu, Derek F. Wong, Dacheng Tao, and Zhaopeng Tu. 2021c. Understanding and improving lexical choice in non-autoregressive translation. In *ICLR*. OpenReview.net.

Cunxiao Du, Zhaopeng Tu, and Jing Jiang. 2021. Order-agnostic cross entropy for non-autoregressive machine translation. In *Proc. of ICML*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Javier Ferrando, Gerard I. Gállego, Belen Alastruey, Carlos Escolano, and Marta R. Costa-jussà. 2022. Towards opening the black box of neural machine translation: Source and target interpretations of the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 8756–8769. Association for Computational Linguistics.

Markus Freitag, George F. Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Trans. Assoc. Comput. Linguistics*, 9:1460–1474.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop

using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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

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

Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006*, volume 148 of *ACM International Conference Proceeding Series*, pages 369–376. ACM.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O. K. Li, and Richard Socher. 2017. Non-autoregressive neural machine translation. *CoRR*, abs/1711.02281.

Jiatao Gu, Changhan Wang, and Junbo Zhao. 2019. Levenshtein transformer. In *NeurIPS*, pages 11179–11189.

Nuno Miguel Guerreiro, Elena Voita, and André F. T. Martins. 2023. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 1059–1075. Association for Computational Linguistics.

Junliang Guo, Xu Tan, Linli Xu, Tao Qin, Enhong Chen, and Tie-Yan Liu. 2020. Fine-tuning by curriculum learning for non-autoregressive neural machine translation. In *AAAI*, pages 7839–7846. AAAI Press.

Fei Huang, Tianhua Tao, Hao Zhou, Lei Li, and Minlie Huang. 2022a. On the learning of non-autoregressive transformers. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9356–9376. PMLR.

Fei Huang, Hao Zhou, Yang Liu, Hang Li, and Minlie Huang. 2022b. Directed acyclic transformer for non-autoregressive machine translation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 9410–9428. PMLR.

Xiao Shi Huang, Felipe Pérez, and Maksims Volkovs. 2022c. Improving non-autoregressive translation models without distillation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Lukasz Kaiser, Samy Bengio, Aurko Roy, Ashish Vaswani, Niki Parmar, Jakob Uszkoreit, and Noam Shazeer. 2018. Fast decoding in sequence models using discrete latent variables. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 2395–2404. PMLR.

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

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation, EAMT 2023, Tampere, Finland, 12-15 June 2023*, pages 193–203. European Association for Machine Translation.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proc. of EMNLP*, pages 388–395.

Feng Li, Jingxian Chen, and Xuejun Zhang. 2023. A survey of non-autoregressive neural machine translation. *Electronics*, 12(13).

Yafu Li, Leyang Cui, Yongjing Yin, and Yue Zhang. 2022. Multi-granularity optimization for non-autoregressive translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5073–5084, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yafu Li, Yongjing Yin, Yulong Chen, and Yue Zhang. 2021. On compositional generalization of neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4767–4780. Association for Computational Linguistics.

Jindrich Libovický and Jindrich Helcl. 2018. End-to-end non-autoregressive neural machine translation with connectionist temporal classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3016–3021. Association for Computational Linguistics.

Jinglin Liu, Yi Ren, Xu Tan, Chen Zhang, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2021. Task-level curriculum learning for non-autoregressive neural machine translation. IJCAI'20, pages 3833–3839.

Xuezhe Ma, Chunting Zhou, X. Li, Graham Neubig, and E. Hovy. 2019. Flowseq: Non-autoregressive conditional sequence generation with generative flow. In *EMNLP-IJCNLP*.

Mathias Müller, Annette Rios, and Rico Sennrich. 2020. Domain robustness in neural machine translation. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas, AMTA 2020, Virtual, October 6-9, 2020*, pages 151–164. Association for Machine Translation in the Americas.

Mathias Müller, Annette Rios, and Rico Sennrich. 2020. Domain robustness in neural machine translation. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (AMTA 2020)*, pages 151–164, Virtual. Association for Machine Translation in the Americas.

Xing Niu, Prashant Mathur, Georgiana Dinu, and Yaser Al-Onaizan. 2020. Evaluating robustness to input perturbations for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8538–8544. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Maja Popovic. 2015. chrf: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation, WMT@EMNLP 2015, 17-18 September 2015, Lisbon, Portugal*, pages 392–395. The Association for Computer Linguistics.

Amy Pu, Hyung Won Chung, Ankur P Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for mt. In *Proceedings of EMNLP*.

L. Qian, Hao Zhou, Yu Bao, Mingxuan Wang, Lin Qiu, W. Zhang, Y. Yu, and Lei Li. 2020. Glancing transformer for non-autoregressive neural machine translation. *ArXiv*, abs/2008.07905.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Chitwan Saharia, William Chan, Saurabh Saxena, and Mohammad Norouzi. 2020. Non-autoregressive machine translation with latent alignments. In *EMNLP*.

Robin M. Schmidt, Telmo Pires, Stephan Peitz, and Jonas Lööf. 2022. Non-autoregressive neural machine translation: A call for clarity. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2785–2799. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of ACL*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Chenze Shao, Yang Feng, Jinchao Zhang, Fandong Meng, Xilin Chen, and Jie Zhou. 2019. Retrieving sequential information for non-autoregressive neural machine translation. In *ACL*, pages 3013–3024. Association for Computational Linguistics.

Chenze Shao, Jinchao Zhang, Yang Feng, Fandong Meng, and Jie Zhou. 2020. Minimizing the bag-of-ngrams difference for non-autoregressive neural machine translation. In *AAAI*.

Zhiqing Sun, Zhuohan Li, Haoqing Wang, Zi Lin, Di He, and Zhi-Hong Deng. 2019. Fast structured decoding for sequence models. In *NeurIPS*.

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

Elena Voita, Rico Sennrich, and Ivan Titov. 2021. Language modeling, lexical translation, reordering: The training process of NMT through the lens of classical SMT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 8478–8491. Association for Computational Linguistics.

Yisheng Xiao, Lijun Wu, Junliang Guo, Juntao Li, Min Zhang, Tao Qin, and Tie-Yan Liu. 2023. A survey on non-autoregressive generation for neural machine translation and beyond. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(10):11407–11427.

Weijia Xu, Sweta Agrawal, Eleftheria Briakou, Marianna J. Martindale, and Marine Carpuat. 2023. Understanding and detecting hallucinations in neural machine translation via model introspection. *CoRR*, abs/2301.07779.

Jianhao Yan, Chenming Wu, Fandong Meng, and Jie Zhou. 2022. Digging errors in NMT: evaluating and understanding model errors from partial hypothesis space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 12067–12085. Association for Computational Linguistics.

Chunting Zhou, Jiatao Gu, and Graham Neubig. 2020a. Understanding knowledge distillation in non-autoregressive machine translation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Chunting Zhou, Jiatao Gu, and Graham Neubig. 2020b. Understanding knowledge distillation in non-autoregressive machine translation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A Recent Advances in NAT

Various techniques have been proposed to address the performance limitations of NAT. Guo et al. (2020); Liu et al. (2021); Ding et al. (2021a); Qian et al. (2020) devise dedicated training curriculums to reduce the learning difficulty of NAT models, whereas Zhou et al. (2020b); Ding et al. (2021c,b) propose improved distillation training. Latent variable modeling has received significant attention in enhancing NAT performance (Libovický and Helcl, 2018; Kaiser et al., 2018; Ma et al., 2019; Saharia et al., 2020; Bao et al., 2022, 2021). Typically, Huang et al. (2022b) explicitly models target dependency as paths in a directed acyclic graph. Another line of research focuses on enhancing the cross-entropy loss or alternating to metric-based objectives (Sun et al., 2019; Shao et al., 2020; Ghazvininejad et al., 2020; Du et al., 2021; Shao et al., 2019; Li et al., 2022). In contrast to fully non-autoregressive methods mentioned earlier, another approach decomposes one-shot generation into multiple iterative non-autoregressive generations (Gu et al., 2019; Ghazvininejad et al., 2019; Huang et al., 2022c). Schmidt et al. (2022) align common NAT techniques and compare translation quality and speed implications under uniform environments. Despite claiming improved performance and comparability with autoregressive models (AT), these approaches are limited in their evaluation using rule-based metrics like BLEU score (Papineni et al., 2002), which demonstrates poor correlation with human preference (Rei et al., 2020; Freitag et al., 2022).

B Analysis Research in NMT

Voita et al. (2021) interpret NMT's learning process during training, and Ferrando et al. (2022); Yan et al. (2022) interpret and analyze model predictions during inference. Müller et al. (2020) study NMT generalization ability to novel domains, whereas Li et al. (2021) demonstrate that NMT's weak compositional generalization capability. Additional metrics proposed by Niu et al. (2020) quantify the effects of input perturbations. Hallucination, which refers to the generation of unrelated outputs by the model, has also been extensively studied (Xu et al., 2023; Guerreiro et al., 2023).

C Experiment Setup

Datasets To evaluate general translation performance, we choose WMT16 En⇒Ro, a widely used benchmark dataset for non-autoregressive translation. In addition, we select a large-scale benchmark dataset, i.e., WMT21 De⇒En, which consists of 101.35M parallel sentences and is further filtered to 88.66M. We apply BPE (Sennrich et al., 2016) on the concatenated training sets with 32,000 operations. Knowledge distillation is commonly used for training NAT models (Gu et al., 2017; Sun et al., 2019; Ghazvininejad et al., 2019, 2020). We train an autoregressive Transformer base model on the raw dataset as the teacher model and use it to generate the distilled dataset. To assess cross-domain translation, we employ the test sets from (Müller et al., 2020), which encompass test instances from 5 domains: medical, IT, koran, law, and subtitles, and we evaluate the models (trained on WMT21 De⇔En) on these test sets. For compositional generalization, we utilize CoGnition (Li et al., 2021) with its original data configurations. Following the approach in (Edunov et al., 2018), we measure model robustness on the WMT21 De⇒En testset by introducing three types of input noise: (1) word deletion with a probability of 0.1; (2) word replacement with "<unk>" with a probability of 0.1; (3) word swapping within a range of 3 words with a probability of 0.1.

Model Settings We adhere to the best-performing model configuration outlined in the corresponding papers (Vaswani et al., 2017; Gu et al., 2017; Saharia et al., 2020; Li et al., 2022; Huang et al., 2022b; Ghazvininejad et al., 2019). For all models, we utilize Transformer with a Transformer_Base configuration: both the encoder and decoder comprise 6 layers with 8 attention heads. The hidden dimension is set to 512, while the feedforward layer dimension is set to 2,048. The model is trained using Adam (Kingma and Ba, 2015) optimizer. We apply a weight decay of 0.01 and label smoothing of 0.1. The learning rate initially increases to $5 \cdot 10^{-4}$ within the first 10K steps and subsequently decays exponentially. All results are based on models trained on the KD dataset unless otherwise stated. For inference, we present results obtained through beam search with a beam size of 5. In the case of iterative models such as CMLM, we set the number of iterative steps as 10. We utilized 4 NVIDIA V100 GPUs for our computations, dedicating two days for the CTC process and five days for DA. Other methods were executed within one day each.

Severity	Category	Weight
Major	Non-translation	25
	all others	5
Minor	Fluency/Punctuation	0.1
	all others	1

Table 5: MQM error weighting (Freitag et al., 2021).

D Human Annotation

We follow Freitag et al. (2021), an evaluation methodology based on the Multidimensional Quality Metrics (MQM) framework, which provides a hierarchical analysis of translation errors. We adopt two common error hierarchy categories: *Accuracy* and *Fluency*. *Accuracy* covers fine-grained 4 error sub-types such as *Addition*, *Omission*, *Mistranslation* and *Untranslated Text*, whereas *Fluency* covers *Punctuation*, *Spelling*, *Grammar*, *Register*, *Inconsistency* and *Character Encoding*. Translations that are too badly garbled to permit error classification are classified as *Non-translation*. In addition to the error type, each error is also annotated with a severity label: minor and major. We follow the error weighting in Freitag et al. (2021) to compute the weighted error counts for each system. Annotation details are presented in Appendix E. We hire three expert translators to conduct side-by-side human evaluations on the 5 German-English translation models, i.e., AT, NAT, MgMO, CTC, DA and CMLM. We randomly sample 100 translations from the WMT21 De⇒En testset and ask translators to annotate translation errors for each instance following the MQM annotation guideline. We average the error counts from the 3 annotators as human evaluation results. For conducting human annotation, we hired three annotators who have degrees in English Linguistics or Applied Linguistics. Before formal annotation, annotators were asked to annotate 100 sampled translations from 5 systems, and based on average annotation time we set a fair salary (i.e., 32 dollars per hour) for them. During their training annotation process, they were paid as well.

E MQM Annotation

We present the details of the error type description in Table 6, the error severity description in Table 7 and error weights in Table 5.

F Human Evaluation Results

The annotation results (average from 3 translators) are presented in Table 8.

G Case Study

We present a case study to showcase the n-gram repetition phenomenon in Table 9. We present several cases to showcase the *spelling* errors of CTC and DAT in Table 10. A case study of *omission* errors is shown in Table 11. A case study of *grammar* and *punctuation* errors is shown in Table 12. A case study of how DAT* alleviates n-gram repetition is presented in Table 13.

H Model Details of DAT*

To strengthen the inter-token dependency of DAT, we increase the depth of the transition model by encoding $\mathbf{Q}$ in Equation 10 with an additional linear layer:

$$\mathbf{Q}^* = \mathrm{ReLU}(\mathbf{Q})\mathbf{W}^*_{\mathrm{Q}}, \tag{15}$$

where ReLU is the rectified linear unit activation function. The same applies to $\mathbf{K}$. We refer to this model as DAT*.

I Cross-domain Performance

The complete cross-domain performance on 5 De⇒En out-of-domain testsets are presented in Table 14.

Compositional generalization in NMT refers to the model's generality to translate compounds (e.g., phrases) of known semantic units (e.g., words). We test AT and DAT on the CoGnition dataset since they do not rely on knowledge distillation, and present the results in Table 4. As shown, DAT underperforms the AT counterpart in compositional generalization by a considerable margin, due to its weak dependency modeling. DAT*

J Robustness to Noisy Input

The translation performance on the WMT21 De⇒En testsets with different types of noises are shown in Table 15, where "None" denotes the performance on the original testset without noise.

Error Category		Description
Accuracy	Addition	Translation includes information not present in the source or repeated content.
	Omission	Translation is missing content from the source.
	Mistranslation	Translation does not accurately represent the source.
	Untranslated text	Source text has been left untranslated.
Fluency	Punctuation	Incorrect punctuation (for locale or style).
	Spelling	Incorrect spelling or capitalization.
	Grammar	Problems with grammar, other than orthography.
	Register	Wrong grammatical register (eg, inappropriately informal pronouns).
	Inconsistency	Internal inconsistency (not related to terminology).
	Character encoding	Characters are garbled due to incorrect encoding.
Non-translation		Impossible to reliably characterize the 5 most severe errors.

Table 6: MQM hierarchy (Freitag et al., 2021).

Severity	Description
Major	Errors that may confuse or mislead the reader due to significant change in meaning or because they appear in a visible or important part of the content.
Minor	Errors that don't lead to loss of meaning and wouldn't confuse or mislead the reader but would be noticed, would decrease stylistic quality, fluency or clarity, or would make the content less appealing.

Table 7: MQM severity levels (Freitag et al., 2021).

	AT		MGMO		CTC		DA		CMLM	
	Maj.	**Min.**	**Maj.**	**Min.**	**Maj.**	**Min.**	**Maj.**	**Min.**	**Maj.**	**Min.**
ACC/Addition	0.33	1.33	2.00	3.33	6.33	7.00	8.67	6.00	1.33	0.67
ACC/Untranslated Text	1.00	0	1.00	0.00	1.33	0.67	1.00	0.00	0.00	0.00
ACC/Mistranslation	14.33	4.00	29.67	3.67	26.67	6.33	19.33	6.00	22.33	7.67
ACC/Omission	11.00	3.67	13.33	4.33	18.00	4.67	5.00	1.33	4.33	5.00
ACC/Punctuation	0.00	1.67	0.00	7.67	0.00	3.33	0.00	3.00	0.33	2.33
FLC/Register	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.33	0.00
FLC/Grammar	1.33	4.00	3.00	7.67	2.33	7.33	1.67	5.00	5.67	8.00
FLC/Spelling	0.00	1.67	1.00	1.67	3.67	8.33	0.33	3.00	0.67	1.00
FLC/Character Encoding	4.00	0.33	2.67	0.00	2.67	0.00	3.33	0.00	0.00	1.00
FLC/Inconsistency	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.33	0.00	0.00
Non-Translation	0.00	0.00	0.33	0.00	0.67	0.00	0.33	0.00	7.00	0.00

Table 8: Human Evaluation Results - Error Counts by Type (Averaged from Three Translators' Annotations).

	Case 1
Source Sentence	In Sachen Kindergarten- respektive Krippenplätzen hat sie bereits Kontakt mit einer örtlichen Einrichtung aufgenommen.
Reference Sentence	Regarding kindergarten respectively nursery places she has already established contact with the local facilities.
DAT Translation	She has already made contact with a local institution in terms of kindergarten and crib places, she has already made contact with a local institution .
DAT* Translation	In terms of kindergarten or crib places, she has already contacted a local institution.
	Case 2
Source Sentence	31 Spieler begrüßte er an der Säbener Straße, darunter auch die neuen Akteure um Edel-Einkauf Leroy Sané, der erstmals nach seinem Wechsel von Manchester City alle neuen Kollegen auf dem Platz traf.
Reference Sentence	He greeted 31 players at the Säbener Straße, among them the new players around special purchase Leroy Sané who met all new colleagues on the field for the first time after his transfer from Manchester City.
DAT Translation	He welcomed 31 players on Säbener Straße, including the new players for fine shopping Leroy Sané, who met all new colleagues on the square for the first time after his move from Manchester City , met all the new colleagues on the pitch for the first time after his move from Manchester City .
DAT* Translation	He welcomed 31 players on Säbener Straße, including the new players around fine shopping Leroy Sané, who met all new colleagues on the pitch for the first time after his move from Manchester City.
	Case 3
Source Sentence	In der Stadt Oakland in Kalifornien wurde ein Gerichtsgebäude in Brand gesteckt.
Reference Sentence	A courthouse was set on fire in Oakland, California.
DAT Translation	In the city of Oakland, California , a courthouse was set on fire in the city of Oakland, California .
DAT* Translation	A courthouse was set on fire in the city of Oakland, California.
	Case 4
Source Sentence	Die Windkraftwerke auf der deutschen Nordsee haben in den ersten sechs Monaten des Jahres 11,51 Terawattstunden Strom in das Netz eingespeist.
Reference Sentence	The wind power plants of the German North Sea delivered 11.51 terawatt hours electricity to the net in the first six months of the year.
DAT Translation	In the first six months of the year , the wind power plants on the German North Sea fed 11.51 terawatt hours of electricity into the grid in the first six months of the year .
DAT* Translation	The wind power plants on the German North Sea fed 11.51 terawatt hours of electricity into the grid in the first six months of the year.
	Case 5
Source Sentence	Bis Anfang November stehen sieben Rennen an.
Reference Sentence	Until the beginning of November seven races are planned.
DAT Translation	By the beginning of November , there are seven races until the beginning of November .
DAT* Translation	There are seven races until the beginning of November.

Table 9: A case study of n-gram repeating of DAT models, comparing with DAT* which enhances dependency modeling by adding a linear layer. The text in the grey background denotes the repeated segment.

	Case 1
Source Sentence	Sie steckten vor einem Jugendgefängnis Bauwagen in Brand, die Polizei setzte Blendgranaten und Pfefferspray ein.
Reference Sentence	They set construction trailers on fire in front of a youth detention center, the police used stun grenades and pepper spray.
CTC Translation	They set construction fire in front of a youth prison, the police used glgrenades (glare grenades) and pepper spray.
	Case 2
Source Sentence	Es gebe aber keine Anhaltspunkte, dass die Anzahl von illegalen Autorennen tatsächlich steige.
Reference Sentence	However, there are no real indications that the number of illegal car races does in fact increase.
CTC Translation	However, there is no indicevidence (indication/evidence) that the number of illegal car races is actually increasing.
	Case 3
Source Sentence	Auf der A81 registriert die Polizei sogar mehr Rennen als auf jeder anderen Bundesautobahn.
Reference Sentence	The police registers even more races on the A81 than on any other federal autobahn.
CTC Translation	On the A81, the police registregister (register) even more races than on any other federal highway.
	Case 4
Source Sentence	Im 24-Stunden-Vergleich wurden in Wien 60 Corona-Neuinfektionen gemeldet - in Niederösterreich gab es 22 Neuinfektionen.
Reference Sentence	In a 24 hour comparison 60 Corona new infections were reported in Vienna - in Lower Austria there were 22 new infections.
DAT Translation	In a 24-hour comparison, 60 corona (Corona) new infections were reported in Vienna - in Lower Austria there were 22 new infections.
	Case 5
Source Sentence	"Ich denke es ist uns gelungen, Rakoczy-Flair zu verbreiten", sagt Kurdirektorin Sylvie Thormann.
Reference Sentence	“I think we still succeeded in spreading Rakoczy flair,” said the Kurstadt director, Sylvie Thormann.
DAT Translation	“I think we have succeeded in spreading rakoczy (Rakoczy) flair,“ says Prime Director Sylvie Thormann

Table 10: A case study of spelling errors of CTC and DAT. The text in the gray background indicates segments with spelling errors, followed by the correct spelling enclosed in brackets.

	Case 1
Source Sentence	In diesem Jahr sind die Fluten besonders schlimm, was Wissenschaflter auf den Klimawandel zurückführen.
Reference Sentence	The floods were especially bad this year, which scientists have connected to climate change.
CTC Translation	This year, the floods are particularly bad, which scientists (have connected) to climate change.
	Case 2
Source Sentence	Den Punkterekord im englischen Fußball verpasste Coach Jürgen Klopp mit seinem Team nur knapp.
Reference Sentence	Coach Jürgen Klopp with his team only narrowly missed the points record in English soccer.
CTC Translation	Coach Jürgen Klopp narrowly missed the (points) and his team in English football.
	Case 3
Source Sentence	Zuletzt hatten Thole/Wickler im September des vergangenen Jahres beim World Tour Final in Rom gespielt.
Reference Sentence	Thole/Wickler recently played in the World Tour Final in Rome in September of last year.
CTC Translation	Thole/Wickler last (year) played at the World Tour Final in Rome (in) September.
	Case 4
Source Sentence	Acht Filme drehte sie mit dem Herzensbrecher.
Reference Sentence	She filmed eight films with the heart breaker.
MgMO Translation	She filmed eight films with the (heart) breaker.
	Case 5
Source Sentence	Frankfurt/Main - Der siebenmalige Zeitfahrweltmeister Tony Martin kann sich durchaus vorstellen, seine Radsport-Karriere fortzusetzen.
Reference Sentence	Frankfurt/Main - The seven-time time trial specialist Tony Martin can clearly picture continuing his bicycling career.
MgMO Translation	Frankfurt/Main - The seven-time time-trial world champion Tony Martin can (clearly) imagine continuing his cycling career.

Table 11: A case study of omission errors of CTC and MgMO. The text indicated within brackets highlights the segments missed by models.

	Case 1
Source Sentence	Auf der A81 registriert die Polizei sogar mehr Rennen als auf jeder anderen Bundesautobahn.
Reference Sentence	The police register even more races on the A81 than on any other federal autobahn.
DA Translation	On the A81, the police registered (register) even more races than on any other federal motorway.
	Case 2
Source Sentence	Auch in der amerikanischen Metropole Seattle lieferten sich Demonstranten am Samstag Zusammenstöße mit der Polizei.
Reference Sentence	In the American metropolis of Seattle demonstrators also ran into clashes with police on Saturday.
CTC Translation	In the American metropolis of Seattle, demonstrators also clashes (clash/clashed) with the police on Saturday.
MgMO Translation	In the American metropolis of Seattle, demonstrators also clashes (clash/clashed) with the police on Saturday.
	Case 3
Source Sentence	Die Polizei war seit dem frühen Abend mit zahlreichen Beamten im Einsatz, im gesamten Stadtgebiet war ein größeres Polizeiaufgebot zu sehen.
Reference Sentence	The police was in use with numerous officers since the early evening, a major police detachment was observed in the entire city area.
CMLM Translation	The police have been working (worked) since the early evening with numerous officials, with a larger police squad throughout the city.
	Case 4
Source Sentence	Zuletzt hielten sich noch einige Dutzend Menschen auf dem Platz auf, verließen ihn jedoch vor Beginn der Sperrstunde um 1 Uhr.
Reference Sentence	Until last, some dozens of people were still present at the place, however, they also left before beginning of the curfew at 1 a.m.
CTC Translation	Finally, a few dozen people stayed on the square, but left it before the start of the curfew at 1 o clock (o'clock).
	Case 5
Source Sentence	Wie die Polizei mitteilt, kam es danach wieder zu Auseinandersetzungen zwischen den beiden Personen.
Reference Sentence	Another scuffle followed between the two persons, according to the police.
MgMO Translation	As the police say, there were clashes between the two people (.)

Table 12: A case study of grammar and punctuation errors. The text in the gray background indicates segments with errors, followed by the correct format enclosed in brackets.

	Case 1
Source Sentence	Bis Anfang November stehen sieben Rennen an.
Reference Sentence	Until the beginning of November seven races are planned.
DAT Vertex Predictions	<BOS> By The Seven There Seven races are been races By **the** As **of** early **beginning of the beginning of of November** , there there will there will be be been are been are seven seven event@@ seven races races have seven seven the races p@@ races are place have run are been planned in p@@ ending scheduled until **the** run scheduled p@@ ending up by **beginning the** early **beginning beginning of** early **of November November beginning November** . <EOS>
DAT Translation	By the beginning of November , there are seven races until the beginning of November .
DAT* Vertex Predictions	<BOS> There are Seven Seven There By seven races By are The are been seven races scheduled As until **the beginning** by **the** early early **beginning of November** early **November of November** , there will are are be have seven been up be are seven seven appear@@ seven races races have seven are races run are been scheduled p@@ place until play be scheduled take place until **the beginning beginning beginning of** in early **of** early early **November** . <EOS>
DAT* Translation	There are seven races until the beginning of November .
	Case 2
Source Sentence	Bei der Kollision fliegen Hand- und Fußbremshebel weg.
Reference Sentence	When they collided hand and foot brake pedals break off.
DAT Vertex Predictions	<BOS> **During** The Hand Flying Hand@@ brake and In case **During the** case **During the** event of **the col@@ col@@** Col@@ ding **col@@ col@@ sion li@@ sion col@@ li@@ li@@ sion , sion** , **the** is **li@@** des leaves fly , the Hand le@@ es Hand@@ away Hand held of Hand of hand hand hand@@ wr@@ held hand and hand le@@ hand le@@ ver ver and ver foot le@@ ver ver and foot b@@ le brake foot foot brake foot bra@@ k@@ king brake brake brake le@@ vers fly le@@ le@@ le@@ le@@ vers vers vers are vers are fly fly fly fle@@ vers fly fly flying fly from the away f@@ away away away **during the** event **col@@ li@@ ding col@@ sion col@@ li@@ sion sion** <EOS>
DAT Translation	During the collision , hand and foot brake levers fly away during the collision .
DAT* Vertex Predictions	<BOS> Hand@@ -@@ Hand Hand@@ Flying In brake -@@ **During** and The foot **col@@** le@@ vers away Col@@ **during the** event of **the col@@ li@@** ding **col@@ col@@ col@@ col@@ sion li@@ sion li@@ sion** , **li@@** breaks involves session **li@@ li@@ sion** there , fly brake re@@ moves fly away of the Hand@@ Hand vers by Hand hand hand@@ held of hand hand hand le@@ - and hand brake brake hand le@@ and vers and and foot F@@ foot foot oot and foot foot under@@ king brake brake brake brake le@@ le@@ bra@@ vers arms vers are fly col@@ le@@ vers vers fly are fly fly flying f@@ fly ail away **during** away **the col@@** A@@ way away away **during the col@@ col@@ li@@ sion li@@** way <EOS>
DAT* Translation	During the collision , hand and foot brake levers fly away.

Table 13: A case study of vertex predictions of DAT and DAT* models. The text in the grey background denotes the repeated segment in DAT. Tokens in bold denote the set of related vertex predictions that construct the phrase "the begging of November". generating repeated n-grams via finding a better vertex transition path, due to its stronger dependency modelling.

Model	IT	Koran	Law	Medical	Subtitles	Average
AT	**78.29**	**62.08**	**85.44**	**79.25**	**75.09**	**76.03**
NAT	32.97	33.10	34.53	32.61	34.90	33.62
MgMO	72.39	57.89	79.00	73.27	69.24	70.36
CTC	50.39	46.66	49.71	57.20	73.10	55.41
DAT	77.86	61.15	85.07	78.78	74.76	75.52
CMLM	61.67	38.30	71.94	59.97	47.42	55.86
DAT*	77.57	61.33	84.05	78.78	74.65	75.28

Table 14: Cross-domian translation performance (COMET). Bold numbers represent the best performance.

Model	None	Replace	Delete	Permutation
AT	84.72	63.93 (-20.79)	72.61 (-12.11)	62.63 (-22.09)
NAT	75.50	57.38 (-18.12)	64.36 (-11.14)	53.28 (-22.22)
CMLM	77.14	58.53 (-18.61)	64.67 (-12.47)	56.92 (-20.22)
CTC	80.06	64.27 (-15.79)	70.44 (-9.62)	60.65 (-19.41)
MgMO	81.15	64.86 (-16.29)	71.09 (-10.06)	61.15 (-20.00)
DAT	83.32	64.02 (-19.30)	71.28 (-12.04)	60.79 (-22.53)
DAT*	83.20	64.75 (-18.45)	71.30 (-11.90)	60.59 (-22.61)

Table 15: Results of translation performance (COMET) on noisy testsets of WMT21 De⇒En.