Imitation Game is Not Optimal: Alleviating Autoregressive Bias in Non-Autoregressive Transformers

Anonymous ACL submission

Abstract

Non-autoregressive Transformer (NART) models predict tokens independently, which presents challenges in capturing token dependencies. Previous approaches have incorporated the Autoregressive (AR) token dependency into the NART models, leading to a discrepancy known as AR exposure bias during the training and decoding processes of NART models, adversely affecting generation quality. We propose two novel approaches that facilitate the recovery of future context information, thereby mitigating AR exposure bias. First, Bidirectional Contextual Knowledge Distillation (BCKD) leverages AR teacher models to distill bidirectional token correlation information, enhancing via data augmentation. Second, the Bidirectional Contextual Transformer (BC-Transformer) captures global contextual information through its innovative graph architecture. Experiments demonstrate that our BC-Transformer achieves translation quality comparable to that of the Autoregressive Transformer (ART) while maintaining the superior generation speed of the DA-Transformer. When both proposed methods are incorporated, NART models significantly outperform ART models ($p < 0.03$). Further analysis reveals that the BC-Transformer surpasses AR baseline models in the translation of long sentences.[1]

1 Introduction

The Autoregressive Transformer (ART, Vaswani et al., 2017) model has shown remarkable effectiveness across multiple NLP tasks, including Machine Translation (Bao et al., 2021), Question Answering (Nassiri and Akhloufi, 2023), and Pre-trained language models (Lewis et al., 2019). Despite their advantages, ART models face inherent challenges, such as high inference latency and exposure bias (Ranzato et al., 2015). Non-Autoregressive Transformer (NART, Gu et al., 2017) models address some of these limitations by offering accelerated generation speeds through their parallel decoding mechanism. However, the conditional independence assumption (Gu and Kong, 2020) undermines their ability to capture token dependencies effectively (Zhou et al., 2019a), which is crucial for maintaining the high generation quality of ART models. Efforts to break these limitations have led to strategies that leverage AR information to guide NART models (Wei et al., 2019; Liu et al., 2020; Guo et al., 2020; Li et al., 2019). The Gu et al.'s (2017) applied conventional sequence-level knowledge distillation (CKD, Kim and Rush, 2016) to NART models, and the Directed Acyclic Transformer (DA-Transformer, Huang et al., 2022b) introduced a Directed Acyclic Graph (DAG) into the NART framework.

Nevertheless, directly mimicking ART models may propagate AR exposure bias into the NAR decoder, limiting the potential of NART models to excel beyond ART models. We refer to this phenomenon as "Autoregressive Bias" (AR Bias). Through theoretical and empirical analyses, we demonstrate that AR Bias originates from CKD and AR Structure Coordination.

To address the shortcomings associated with AR Bias and to boost the generation quality of NART models, we introduce Bidirectional Contextual Knowledge Distillation (BCKD) and the Bidirectional Contextual Transformer (BC-Transformer), which incorporate future context information into both CKD and the DA-Transformer framework. The BCKD approach mitigates AR Bias using bidirectional teacher models to harness token dependencies from both left-to-right (L2R) and right-to-left (R2L) orientations. In contrast, the BC-Transformer employs a Bidirectional Contextual Graph (BCG) rather than a DAG, capturing bidirectional token dependencies and enhancing the model's search and aggregation capabilities. Em-

[1] Codes will be released upon the acceptance of this paper.

pirical results demonstrate that the BC-Transformer achieves translation results comparable to ART models. Additionally, BCKD enhances the generation quality of the BC-Transformer, significantly outperforming ART models ($P < 0.03$) while maintaining the rapid inference speed of fully-NART models. In this paper, we delineate our primary contributions as follows:

- We pinpoint and empirically validate the phenomenon of "Autoregressive Bias" (AR Bias) within NART models, which can be transferred from ART models through CKD and AR structure coordination.
- To counteract AR Bias, we introduce future context information to NART models using BCKD and the BC-Transformer, superseding traditional CKD and the state-of-the-art DA-Transformer.
- Our experimental results confirm the effectiveness of our methods, indicating that NART models can surpass ART models by leveraging bidirectional contextual information. Analysis proves the significant reduction of AR Bias afforded by our proposed methods.

2 Tracking "Autoregressive Bias" in Non-Autoregressive Transformers

A variety of methodologies have been advanced for integrating the AR Factor (f_{AR}) into NART models to elevate generation quality. The relationship can be formalized as:

$$\theta_{\text{NART}} = F(Y, X, f_{\text{AR}}) \tag{1}$$

One such approach, the CKD utilizes the NART student model to assimilate knowledge from an AR teacher model. Alternatively, AR structure coordination incorporates the f_{AR} by melding the AR structure directly into the NART model framework to capture the conditional relationship R between token Y_i and its preceding tokens $Y_{j\in(1,i-1)}$:

$$f_{\text{AR}} = R(Y_i|X, Y_{j\in(1,i-1)}, \theta) \tag{2}$$

This section is devoted to the theoretical examination and the conception of analytical experiments aimed at discerning the extent of "AR Bias" within NART models.

2.1 Model-Induced Autoregressive Bias

The DA-Transformer (Huang et al., 2022b) employs a directed acyclic graph (DAG) $G = \{E, V\}$ to model the sentence probability. Given a target sentence $Y = \{Y_1, Y_2, ..., Y_n\}$, each path in the graph $A = \{a_1, a_2, ..., a_n\}$ composes a candidate sentence. The edges $e \in E$ represent the transition probability between vertices V, capturing the correlation between adjacent target tokens Y_{i-1} and Y_i. Throughout the DAG's training and inference process, the model simultaneously generates candidate tokens $\dot{Y}$ and transition probabilities E, exploring multiple paths to estimate sentence probability.

The DAG formalizes sentence probability by intermediate searching states $S = \{S_1, S_2, ..., S_{n-1}\}$, where each state corresponds to the determination of all tokens and transitions preceding Y_i.

$$\begin{aligned} P(S_i) &= P(\dot{Y}_1) \prod_{j=2}^{i} E_{i-1,i} \times P(\dot{Y}_i) \\ &= P(S_{i-1}) \times E_{i-1,i} \times P(\dot{Y}_i) \end{aligned} \tag{3}$$

Here, The S_{i-1} takes the role of "previous output tokens" as in an ART model, $\dot{Y}_i$ represents word-level generation probability, and E signifies the AR correlation between target tokens:

$$\begin{aligned} P(S_{i-1}) &= P(Y_{j\in(1,i-1)}) \\ P(\dot{Y}_i) &= P(X_i, \theta) \\ E_{i-1,i} &= P(Y_{i-1}, Y_i, \theta) = P(X, \theta) \end{aligned} \tag{4}$$

We then reformulate the sentence probability as:

$$\begin{aligned} P_{\text{DAG}}(Y) &= \prod_{i=1}^{n} P(S_i|X, S_{i-1}, \theta) \\ &= \prod_{i=1}^{n} P(Y_i|X, Y_{j,j\in(1,i-1)}, \theta) \end{aligned} \tag{5}$$

From the derivation above, we conclude that the L2R edges enable the DAG to capture directional dependencies relationship R among adjacent target tokens, resulting in the DA-Transformer learning AR token dependency information:

$$f_{\text{AR}} = R(Y_i|X, Y_{j,j\in(1,i-1)}, \theta) = E \tag{6}$$

This approach introduces the AR factor f_{AR} into the DA-Transformer, introducing AR Bias into model edge confidence. To substantiate our hypothesis, we train two DA-Transformers with diverse edge directions—the forward DA ($\overrightarrow{\text{DA}}$) and

the backward DA ($\overleftarrow{\text{DA}}$), and find a clear correlation between transition probability E and edge direction through visualization (see Figure 3).

2.2 Data-Induced Autoregressive Bias

ART models captures the dependency relationship between adjacent target tokens via next-token prediction tasks. However, they inherently lack future context information. The CKD facilitates NART models to assimilate from ART models by minimizing the Kullback–Leibler (KL) divergence (Kim and Rush, 2016) between ART and NART models.

$$\begin{aligned} \theta_{\text{ART}} &= P(Y_i|X, Y_{j,j\in(1,i-1)}, \theta) \\ f_{\text{AR}} &= \text{KL}(\theta_{\text{ART}}|\theta_{\text{NART}}) \end{aligned} \tag{7}$$

Ding et al. (2020) demonstrated that NART models acquire the ART lexical distribution through CKD. However, due to the inherent AR exposure bias, NART models are typically unable to learn future context dependencies directly from raw data.

To investigate the perpetuation of AR Bias through CKD, we adopted the methodology outlined by Zhou et al. (2019b) to scrutinize the correlation between the decoding direction of the ART teacher model, the quality of CKD data, and the generative quality of the NART student model. For a parallel corpus C, we trained two teacher models with opposite decoding directions: the forward M_T ($\overrightarrow{M}_T$) and backward M_T ($\overleftarrow{M}_T$) which generate two versions of the KD corpus respectively. Subsequently, we trained two student models independently on these corpora—the L2R student model ($\overrightarrow{M}_S$) and the R2L student model ($\overleftarrow{M}_S$).

Through the analysis to evaluate the CKD data accuracy and the student model's translation quality, it was observed that the L2R data exhibits higher correctness at the beginning of the sentences. Conversely, the R2L data demonstrated increased correctness towards the sentence endings (see Figure 4). This phenomenon is mirrored in the student models, resulting in an imbalanced performance enhancement through CKD (refer to Table 2). These findings confirm the transmission of AR Bias from ART models to NART models via the CKD.

3 Alleviation of Autoregressive Bias

Drawing on the insights from Zhou et al. (2019b), regarding AR Bias, we introduce the "Bidirectional Contextual Factor" f_{BC} as a substitute for the "AR Factor" f_{AR} to mitigate AR Bias. The f_{BC} encapsulates the interdependence between the token y_i and all other tokens y_j, where $j \neq i$ thereby endowing NART models with future contextual information.

$$\begin{aligned} \theta_{\text{NART}} &= F(Y, X, f_{\text{BC}}) \\ f_{\text{BC}} &= R(Y_i|X, Y_{j,(j\neq i)}, \theta) \end{aligned} \tag{8}$$

3.1 Bidirectional Contextual Transformer

The BC-Transformer (Shown in Figure 1) employs the BCG $G_{\text{BC}} = \{\overrightarrow{E}, \overleftarrow{E}, V\}$ as opposed to the DAG $G_{\text{DA}} = \{\overrightarrow{E}, V\}$, to extract contextual information for NART models. The BC-Transformer formulates the sentence probability as follows:

$$\begin{aligned} P_{\text{BC}}(Y) &= \sum_{A\in\Gamma} P_\theta(A|X)P_\theta(Y|A, X) \\ A &= \{\overrightarrow{A}, \overleftarrow{A}\} \end{aligned} \tag{9}$$

where $\overrightarrow{A}$ denotes paths interconnected via L2R directional edges $\overrightarrow{E}$, and $\overleftarrow{A}$ signifies directional paths linked through R2L edges $\overleftarrow{E}$. Γ encompasses all potential bidirectional paths composed of $\overrightarrow{A}$ and $\overleftarrow{A}$. The G_{BC} comprises two sets of edges: the L2R edges $\overrightarrow{e} \in \overrightarrow{E}$ and the R2L edge $\overleftarrow{e} \in \overleftarrow{E}$.

$$\begin{aligned} G_{BC} &= \{\overrightarrow{E}, \overleftarrow{E}\, V\} \\ \overrightarrow{E} &= P(Y_{i-1}, Y_i) \\ \overleftarrow{E} &= P(Y_{i+1}, Y_i) \end{aligned} \tag{10}$$

The edges E delineate the directional transition probabilities between adjacent target tokens, calculated via the attention mechenism. Where d is the hidden size, W_Q and W_K are learnable weights.

$$\begin{aligned} E &= \text{softmax}(\frac{QK^T}{\sqrt{d}}) \\ Q &= VW_Q, K = VW_K \end{aligned} \tag{11}$$

For a candidate token $\dot{Y}_i$, the BCG facilitates the model in acquiring the comprehension of sentence construction through two directional search states.

$$\begin{aligned} P(\overrightarrow{S}_i) &= P(\overrightarrow{S}_{i-1}) \times \overrightarrow{E}_{i-1,i} \times P(\dot{Y}_i) \\ P(\overleftarrow{S}_i) &= P(\overleftarrow{S}_{i+1}) \times \overleftarrow{E}_{i+1,i} \times P(\dot{Y}_i) \end{aligned} \tag{12}$$

The L2R state $\overrightarrow{S}_i$ embeds the antecedent contextual information, while the R2L state $\overleftarrow{S}_i$ embeds the subsequent contextual information.

$$\begin{aligned} P(\overrightarrow{S}_{i-1}) &= P(Y_{j\in(1,i-1)}, \theta) \\ P(\overleftarrow{S}_{i+1}) &= P(Y_{j\in(i+1,n)}, \theta) \end{aligned} \tag{13}$$

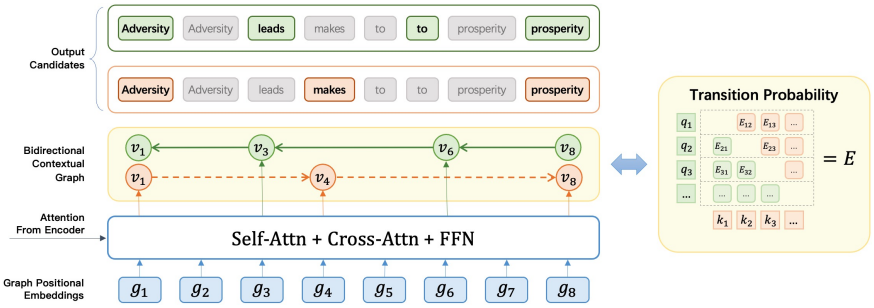

Figure 1: Overview of BC-Transformer. The model utilizes the BCG $G = \{\overrightarrow{E}, \overleftarrow{E}, V\}$ to establish candidate paths from both directions, enhancing contextual information acquisition. Output candidates encompass search results from both directions, facilitating improved generation quality.

This configuration aids the model in acquiring bidirectional contextual information.

$$P_{BC}(Y) = \prod_{i=1}^{n} P(Y_i|X, Y_{j,j<i}, Y_{k,k>i}, \theta) = \prod_{i=1}^{n} P(Y_i|X, Y_{j,j\neq i}, \theta) \quad (14)$$

The BC-Transformer employs edges in both directions to search for candidate sentences during model inference, namely Bidirectional Ensemble Search (BES), and use the directed paths to formalize the candidate probabilities. The beam-search candidate list is partitioned into two identical segments to accommodate two sets of candidates, with the most probable one selected as the model output.

$$P_\theta(Y|\overrightarrow{A}, X) = \prod_{i=2}^{n} \overrightarrow{e}_{a_{i-1},a_i} \times P(Y_i)$$
$$P_\theta(Y|\overleftarrow{A}, X) = \prod_{i=1}^{n-1} \overleftarrow{e}_{a_{i+1},a_i} \times P(Y_i)$$
$$Y = \text{BeamSearch}((V, \overleftarrow{E}) \bigcup (V, \overrightarrow{E})) \quad (15)$$

Since the token prediction and edge probabilities are calculated simultaneously, the BC-Transformer maintains a high degree of decoding parallelism akin to the DA-Transformer.

3.2 Bidirectional Contextual Knowledge Distillation

The BCKD (see Figure 2) introduces the bidirectional context information through the integration of two directional AR teacher models: the L2R model $\overrightarrow{M}$ learns and generates the target tokens from left to right, and the R2L model $\overleftarrow{M}$ learns and generates the target tokens from right to left.

$$f_{\text{BC}} = \text{KL}(\overleftrightarrow{\theta}_{\text{ART}}|\theta_{\text{NART}})$$
$$f_{\text{BC}} = \text{KL}(\overleftarrow{\theta}_{\text{ART}}, \overrightarrow{\theta}_{\text{ART}}, |\theta_{\text{NART}}) \quad (16)$$

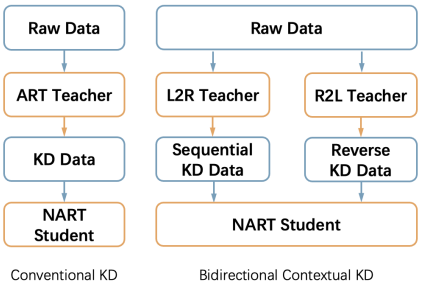

Figure 2: Comparison between Conventional KD (CKD) and Bidirectional Contextual KD (BCKD). The BCKD incorporates two AR teachers, the L2R and R2L models, which enables the NART student model to acquire contextual information from both directions.

For a sentence pair $\{\overrightarrow{S}, \overrightarrow{R}\}$, the R2L model can be effectively implemented via training and inference on the reversed corpus $\{\overleftarrow{S}, \overleftarrow{R}\}$. After model training, the L2R and R2L models $\overrightarrow{M}$ and $\overleftarrow{M}$ generate the corresponding KD corpus, $\overrightarrow{C}_{\text{KD}}$ and $\overleftarrow{C}_{\text{KD}}$. We aggregate them and maintain consistency in word order for the NART model training.

$$\overrightarrow{\theta}_{\text{ART}} = R(Y_i|X, Y_{j\in(1,i-1)}, \theta)$$
$$\overleftarrow{\theta}_{\text{ART}} = R(Y_i|X, Y_{j\in(i+1,n)}, \theta)$$
$$f_{\text{BC}} = \text{KL}(\overleftarrow{\theta}_{\text{ART}}, \overrightarrow{\theta}_{\text{ART}}, |\theta_{\text{NART}}) \quad (17)$$

The L2R and R2L models contain the token correlation information R between token Y_i and its preceding and successor tokens. The combination of both corpora effectively complements each other, yielding sufficient bidirectional contextual information within the BCKD Corpus $C_{\text{BCKD}} = \text{Mix}(\overrightarrow{C}_{\text{KD}}, \overleftarrow{C}_{\text{KD}})$. We use the C_{BCKD} to replace the CKD corpus C_{CKD} for NART model training.

Model	Iter	WMT14		WMT16		ΔBLEU	Speedup
		En-De	De-En	En-Ro	Ro-En		
Transformer (base)*	M	27.27	31.75	33.74	33.96	-	1.0x
Transformer (big) (Vaswani et al., 2017)*	M	28.62	32.29	-	-	-	0.84x
GLAT (wo/KD) (Qian et al., 2021)*	1	19.16	26.86	28.59	29.04	-5.77	15.3x
DA-T (wo/KD) (Huang et al., 2022b)*	1	27.16	30.76	32.83	33.73	-0.56	7.1x
PCFG-NART (wo/KD) (Gui et al., 2023)	1	27.02	31.29	32.72	33.07	-0.66	14.2x
CMLM (w/KD) (Ghazvininejad et al., 2019)	10	27.03	30.53	33.08	33.31	-0.69	2.2x
GLAT (w/KD) (Qian et al., 2021)*	1	25.08	29.82	31.32	32.06	-2.11	15.3x
DA-T (w/KD) (Huang et al., 2022b)*	1	27.65	32.09	33.33	34.00	+0.08	7.1x
BC-T (wo/KD)*	1	27.33	31.61	33.66	34.29	+0.04	7.0x
GLAT (w/BCKD)*	1	25.93	30.75	32.43	33.26	-1.08	15.3x
BC-T (w/BCKD)*	1	**28.60**	**32.59**	**34.41**	**35.00**	+0.97	7.0x

Table 1: Results on WMT14 En-De/De-En and WMT16 En-Ro/Ro-En benchmarks. The ΔBLEU shows the difference between the models and the Transformer (base) baseline. The $*$ represents our implementation. BC-T achieves comparable performance with ART; BCKD enhances the translation quality of both BC-T and GLAT.

4 Experiment

4.1 Experiment Setup

We analyze performance on two language pairs: the WMT14 English-German (En-De) (4.5M), and the WMT16 English-Romanian (En-Ro) datasets (610K).[2] We evaluate both translation directions for each dataset. To guarantee the comparability of our results, we adopt the preprocessing methodologies used for the WMT14 En-De from (Huang et al., 2022b) and for WMT16 En-Ro from (Huang et al., 2022a). We employ the DA-Transformer (Huang et al., 2022b) and Glancing Transformer (GLAT, Qian et al., 2021) as our baseline model to evaluate the effectiveness of the BC-Transformer and BCKD. We maintain the same parameter settings for the BC-Transformer as reported in (Huang et al., 2022b), including an upsampling scale of 8, a batch size of 64K tokens, a dropout rate of 0.1, and a maximum update of 30K. For the WMT16 En-Ro experiments, we adjust the dropout rate to 0.3. During model inference, we configure the beam size 200 and set β to 1.1, with α from 1.0 to 1.4.

For KD in the WMT16 En-Ro setting, we employ the Transformer (base) model. In the WMT14 En-De context, we use the Transformer (big) as the teacher model, for both the DA-Transformer and BC-Transformer, in line with the approach described by Huang et al. (2022b). For the GLAT, we adhere to the guidelines of Qian et al. (2021) and use the Transformer (base) model as the teacher model. We select the tokenized BLEU score (Papineni et al., 2002) and the COMET Score (Rei et al., 2020) as our evaluation metrics.[3]

4.2 Main Results

Our principal findings are summarized in Table 1. The BC-Transformer consistently outperforms the DA-Transformer across all four translation directions, thereby emphasizing the significance of incorporating future context information into NART models. Moreover, our proposed BCKD method significantly enhances the translation quality of both the BC-Transformer and the GLAT, achieving an increase of 1 BLEU point, which establishes its superiority over existing CKD approaches.

We highlight the advantages of our methodologies in two key areas: first, a uniform enhancement in generation quality across distinct translation directions and second, the attainment of this improvement without sacrificing generation speed. Furthermore, according to the BLEU score assessment, the BC-Transformer attains a translation quality comparable to ART models without KD, evidenced by a marginal increase of 0.03 BLEU points. When integrating the BC-Transformer with BCKD, the NART model significantly outstrips the performance of the ART model ($p < 0.03$), which underscores the efficacy of NART models in leveraging bidirectional contextual information.

5 Analysis of Autoregressive Bias Alleviation

5.1 Model-Induced Bias Alleviation

To validate the mitigation of AR Bias through BCG training, we compare the performance of the DA-

[2]We provide WMT17 experiment results in section A.3.

[3]The COMET Scores are presented in the section A.6.

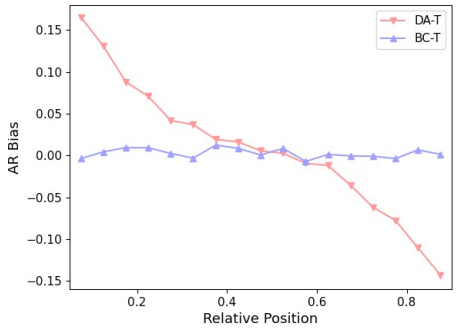

Figure 3: Comparison of AR Bias between DA-Transformer and BC-Transformer. The figure shows a clear correlation between link confidence and relative positions in the DA-Transformer, while the BC-Transformer decreases the AR Bias, demonstrating an almost horizontal line.

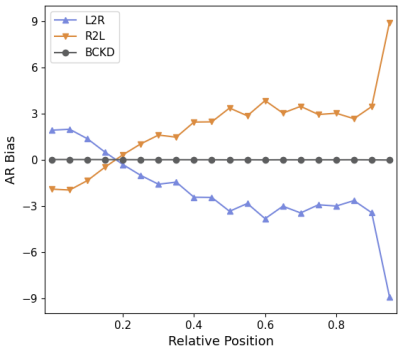

Figure 4: AR Bias visualization of the L2R/R2L/BCKD data on WMT16 Ro-En dataset. Teacher models with diverse inference directions generate biased KD data with imbalanced quality in different relative positions.

Teacher Models	Left ACC	Right ACC
Raw (W/o KD)	68.3%	72.4%
L2R ($\overrightarrow{M}_T$)	69.7%	73.2%
R2L ($\overleftarrow{M}_T$)	68.9%	73.5%
BCKD ($\overleftarrow{M}_T$,$\overrightarrow{M}_T$)	69.7%	73.7%

Table 2: First & last 4 token prediction accuracy of GLAT models with ART teachers ($\overrightarrow{M}_T$ and $\overleftarrow{M}_T$).

Transformer and the BC-Transformer. We train two DA-Transformers with distinct link directions L2R and R2L and implement lookahead decoding for the three models ($\overrightarrow{\text{DAT}}$, $\overleftarrow{\text{DAT}}$, BCT). For each target token y_i, we visualize the L2R and R2L link transition probability $\overrightarrow{P}_i$ and $\overleftarrow{P}_i$. We quantify the AR Bias P_{bias} through the disparity between these transition probabilities.

$$\begin{aligned} \overleftarrow{P}_i &= \overleftarrow{e}_{i+1,i} \\ \overrightarrow{P}_i &= \overrightarrow{e}_{i-1,i} \\ P_{\text{bias}} &= \overrightarrow{P}_i - \overleftarrow{P}_i \end{aligned} \tag{18}$$

Figure 3 exhibits the visualization of AR Bias P_{bias} in each relative positions observed in DAT and BCT. The figure substantiates the reduction of AR Bias by BCG training. The BC-Transformer demonstrates an almost horizontal line in the visualization, indicating negligible differences in link transition probabilities between the two search directions across all relative positions, which suggests a significant decrease in AR Bias.

5.2 Data-Induced Bias Mitigation

We conducted a two-stage analysis involving data quality assessment and model output evaluation to quantify the AR Bias present in the WMT16 Ro-En KD data. To quantify the data-induced AR Bias, token accuracy at each relative position was measured. In the KD dataset, each token Y_i is considered correctly translated if it is found in the reference sentence Y_{raw}.

Following the approach of Huang et al. (2023) we present the translation accuracy $\text{A}(Y_i)$ within 20 relative position buckets. In an effort to control for the variation in translation difficulty, we average the token accuracy from two directional KD corpus $\overrightarrow{Y}_i$ and $\overleftarrow{Y}_i$ as the translation difficulty $D(Y_i)$. We quantify the AR Bias as the difference between the token accuracy A and translation difficulty D.

$$\begin{aligned} \text{A}(Y_i) &= \text{count}(Y_i \in Y_{\text{raw}})/\text{count}(Y_i) \\ D(Y_i) &= \text{Average}(\text{A}(\overrightarrow{Y}_i), \text{A}(\overleftarrow{Y}_i)) \\ \text{Bias}(Y_i) &= \text{A}(Y_i) - \text{D}(Y_i) \end{aligned} \tag{19}$$

The Figure 4 led to two principal observations: 1) The accuracy of the KD data is influenced by the decoding direction used by the teacher model; 2) BCKD effectively mitigates the AR Bias found in conventional KD data.

To determine the impact of BCKD on the student model's capabilities, we employed the method set forth by Zhou et al. (2019b) to compute the token prediction accuracy for the GLAT model enhanced with BCKD (refer to Table 2). Our findings indicate that KD results in an uneven enhancement of token prediction accuracy across the student model. The L2R student model shows a 1.4% increase in accuracy for prefix tokens and a 0.8% increase for suffix tokens, while the R2L student model exhibits a 0.6% improvement for prefix tokens and a 1.1% improvement for suffix tokens. Notably, BCKD, which integrates insights from both teacher models,

Stretegy	Ro-En		En-Ro	
	DA-T	BC-T	DA-T	BC-T
L2R	33.73	34.23	32.83	33.55
R2L	33.90	34.20	33.46	33.56
Bidir	-	**34.29**	-	**33.66**

Table 3: Translation Quality of WMT16 Ro-En / En-Ro with three inference strategies. The BCG training constructs a better-quality graph, and the BES utilizes search results from both directions.

Teacher Model	**En-Ro**	**Ro-En**	ΔBLEU
Raw (W/o KD)	28.59	29.04	-
Single ($\overrightarrow{M}_T$)	31.32	32.06	2.87
Dual ($\overrightarrow{M}_T$,$\overrightarrow{M}'_T$)	32.21	32.75	3.66
Bidir ($\overleftarrow{M}_T$,$\overrightarrow{M}_T$)	**32.43**	**33.26**	**4.03**

Table 4: Translation Quality of GLAT distilled from multiple teacher models on WMT16 En-Ro and Ro-En, ΔBLEU denotes the improvement over the Raw model.

enables the student models to demonstrate superior performance across all token types.

6 Ablation Study

6.1 Search Direction in BC-Transformer

Our proposed BC-Transformer supports two search directions: L2R and R2L. Additionally, the BES strategy combines the search hypotheses from both directions. To evaluate the effectiveness of our BES approach, we compare the BLEU scores of model's outputs with the two individual search directions as well as the ensemble search strategy. We also provide results from the DA-Transformer. The results are detailed in Table 3.

BC-Training Substantially Enhances Graph Quality. By employing directional search on a BCG, there is a notable improvement in translation quality over the DAG, which accentuates the value of BC-Training and the pivotal role that future contextual information plays in NART training.

Bidirectional Ensemble Inference Elevates Generation Quality. The bidirectional ensemble inference surpasses unidirectional inference, which underscores the benefits of combining bidirectional hypotheses. This demonstrates the significance of contextual consideration in enhancing the robustness and accuracy of translation results.

6.2 Teacher Model Inference Direction

The BCKD approach utilizes a pair of models as teacher models. To isolate and understand the effect of the number of teacher models employed, we conduct a controlled experiment, which deviates from the standard practice of training two teachers with opposite inference directions, the L2R ($\overrightarrow{M}_T$) and R2L ($\overleftarrow{M}_T$). Instead, we train two AR teachers with identical inference directions but initiated with different random seeds: L2R ($\overrightarrow{M}_T$) and L2R ($\overrightarrow{M}'_T$). Following this setup, we deploy the GLAT to evaluate the impact of these two KD strategies on the translation tasks of WMT16 En-Ro and Ro-En.

The results, as shown in Table 4, indicate that while increasing the number of teacher models does improve the translation performance of the student model, the BCKD strategy significantly enhances the NART model even more. This enhancement is indicative of the added value that bidirectional contextual knowledge imparts to the distillation process, bolstering the student model's translation capabilities beyond the contribution of additional teacher model guidance alone.

6.3 Quality and Latency Tradeoff

In the realm of fully NART models, the BC-Transformer does not bring additional inference overhead than the DA-Transformer. Moreover, the BC-Transformer, with its multiple search directions, yields enhanced translation outcomes utilizing a larger beam size. To explore the efficiency and quality tradeoff, we conducted an ablation study on the BC-Transformer using the WMT14 De-En dataset (Figure 5).

Larger Beam Size for Improved Performance. Our study reveals that the DA-Transformer reaches peak performance at a beam size of 100, it experiences a subsequent decline in generation quality as the beam size further increases. In contrast, the BC-Transformer exhibits improved performance as the beam size grows, displaying a more scalable relationship between beam size and translation quality.

Superior Quality at Equivalent Beam Sizes. Remarkably, the BC-Transformer surpasses the DA-Transformer by a significant margin at a minimal beam size of 2. At this size, the BC-Transformer performs a bidirectional ensemble lookahead generation, emphasizing its ability to achieve enhanced translation performance without compromising on latency.

BCKD Boosts Performance Across Beam Sizes. The BC-Transformer outshines the ART baseline for beam sizes from 2 to 400 with the BCKD. Demonstrating consistent superiority, even

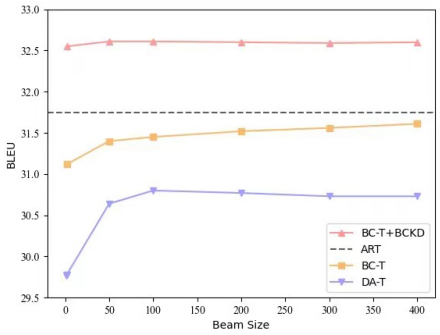

Figure 5: BLEU Score comparison of BC-T, DA-T under different beam size in WMT14 De-En. BC-T acquires comparable result with ART, while BC-T + BCKD outperforms ART under low beam size.

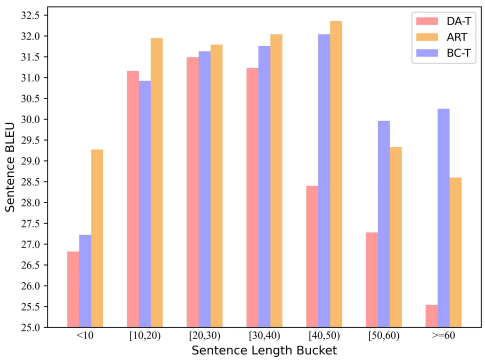

Figure 6: BLEU Score under Sentence Length Buckets. BC-T acquires significant improvements over DA-T and even outperforms ARTs in long sentences (Len>50).

at the lowest beam size where bidirectional lookahead decoding is employed, the BC-Transformer systematically exceeds the performance of the AR baseline. This underlines the effectiveness of BCKD in bolstering the BC-Transformer's translation accuracy irrespective of beam size.

6.4 Translation Quality Analysis

In this section, we delve into a comprehensive analysis of translation quality concerning our methodologies: the BC-Transformer and the BC-Transformer enhanced with BCKD. These approaches are compared against the DA-Transformer at the sentence level, and the following are the observed outcomes:

Superiority in Lengthy Sentences. Upon evaluating sentence-level BLEU scores across different sentence length buckets, our analysis uncovers that the BC-Transformer consistently outperforms the DA-Transformer across all length categories. Noteworthy, the BC-Transformer demonstrates significant improvements, especially with longer sentences. This highlights the superior ability of the BC-Transformer to handle complex sentences and maintain quality in lengthier text passages.

Reduction in Error Accumulation. Through a detailed case study[4], we observe that the DA-Transformer, akin to ART models, is prone to error accumulation. In this common pitfall, an error in the initial token generation cascades, deteriorating the quality of the entire sentence. However, the BC-Transformer shows resilience against error propagation, thereby showcasing its robust generation capabilities and ensuring more reliable translations even faced with challenging inputs.

7 Related Work

Exposure Bias in ART The concept of exposure bias (Ranzato et al., 2015) poses a fundamental challenge for ART models. Several strategies have been proposed to mitigate this issue in ART models. For example, Zhou et al. (2019b) and Tan et al. (2019) have explored the integration of future context information into the translation process. Despite the potential of these modifications to enhance translation quality, there is a trade-off: such adjustments to AR models typically result in increased inference overhead (Tan et al., 2019) or doubled inference latency (Zhou et al., 2019b).

AR-Assisted NART Models. A different line of research has focused on enhancing NART models by utilizing knowledge from their AR counterparts. Li et al. (2019) have promoted the concept of NART models learning from the interior parameters of ART models. Guo et al. (2020) have introduced a curriculum learning approach to refine NAR fine-tuning processes. Further, Hao et al. (2020) and Wang et al. (2022) have presented multitask learning methods for both AR and NAR generation tasks to effectively transfer knowledge and reduce the gap between the two paradigms.

8 Conclusion

This paper introduces the BC-Transformer and BCKD to mitigate AR Bias in NART models. The BC-Transformer enriches the DA-Transformer with a bidirectional structure for improved context capture, while BCKD incorporates a reverse KD model to enhance context awareness further. Our experiments show that these methods significantly elevate NART model performance, outpacing ART baselines and indicating a promising avenue for future quality enhancements in NART systems.

[4]We show translation cases in Appendix (Table 5).

9 Limitations

Our proposed BC-Transformer shares the nature of fully-NART and Mask-Predict models. In this work, we follow the framework of the DA-Transformer to propose the fully-NART BC-Transformer model and verifies the BCKD is effective for the CMLM model (see Section A.4). However, we did not expand the BC-Transformer to the iterative refinement model to further increase the generation quality. We leave it to the future work.

Although the NART model acquires better translation quality on our BCKD dataset, our proposed BCKD still increases the data quantity. In future work, we suggest better data argumentation methods to generate high-quality and unbiased training data within the data quantity budget.

Our proposed methods remain on the scale of the Transformer model for NAR Machine Translation tasks. Previous work has shown the potential of the large-scale NART models (Wang et al., 2023), or NART models with pretraining, such as BANG (Qi et al., 2021), MIST (Jiang et al., 2021) and PreDAT (Huang et al., 2023). We are excited to expand the BC-Transformer to larger-scale pretrained or large language models (Ye et al., 2023) to boost the NAR generation models in the future.

References

Guangsheng Bao, Yue Zhang, Zhiyang Teng, Boxing Chen, and Weihua Luo. 2021. G-transformer for document-level machine translation. *arXiv preprint arXiv:2105.14761*.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in neural information processing systems*, 28.

Liang Ding, Longyue Wang, Xuebo Liu, Derek F Wong, Dacheng Tao, and Zhaopeng Tu. 2020. Understanding and improving lexical choice in non-autoregressive translation. In *International Conference on Learning Representations*.

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

Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. 2017. Non-autoregressive neural machine translation. *arXiv preprint arXiv:1711.02281*.

Jiatao Gu and Xiang Kong. 2020. Fully non-autoregressive neural machine translation: Tricks of the trade. *arXiv preprint arXiv:2012.15833*.

Shangtong Gui, Chenze Shao, Zhengrui Ma, Xishan Zhang, Yunji Chen, and Yang Feng. 2023. Non-autoregressive machine translation with probabilistic context-free grammar. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7839–7846.

Yongchang Hao, Shilin He, Wenxiang Jiao, Zhaopeng Tu, Michael Lyu, and Xing Wang. 2020. Multi-task learning with shared encoder for non-autoregressive machine translation. *arXiv preprint arXiv:2010.12868*.

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

Fei Huang, Pei Ke, and Minlie Huang. 2023. Directed acyclic transformer pre-training for high-quality non-autoregressive text generation. *arXiv preprint arXiv:2304.11791*.

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*, pages 9410–9428. PMLR.

Ting Jiang, Shaohan Huang, Zihan Zhang, Deqing Wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Liangjie Zhang, and Qi Zhang. 2021. Improving non-autoregressive generation with mixup training. *arXiv preprint arXiv:2110.11115*.

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Zhuohan Li, Zi Lin, Di He, Fei Tian, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2019. Hint-based training for non-autoregressive machine translation. *arXiv preprint arXiv:1909.06708*.

Jinglin Liu, Yi Ren, Xu Tan, Chen Zhang, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2020. Task-level curriculum learning for non-autoregressive neural machine translation. *arXiv preprint arXiv:2007.08772*.

Lemao Liu, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2016. Agreement on target-bidirectional neural machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 411–416.

Khalid Nassiri and Moulay Akhloufi. 2023. Transformer models used for text-based question answering systems. *Applied Intelligence*, 53(9):10602–10635.

Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, Xinyi Wang, and John Wieting. 2019. compare-mt: A tool for holistic comparison of language generation systems. *arXiv preprint arXiv:1903.07926*.

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

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

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

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

Dmitriy Serdyuk, Nan Rosemary Ke, Alessandro Sordoni, Adam Trischler, Chris Pal, and Yoshua Bengio. 2017. Twin networks: Matching the future for sequence generation. *arXiv preprint arXiv:1708.06742*.

Xu Tan, Yingce Xia, Lijun Wu, and Tao Qin. 2019. Efficient bidirectional neural machine translation. *arXiv preprint arXiv:1908.09329*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xinyou Wang, Zaixiang Zheng, and Shujian Huang. 2022. Helping the weak makes you strong: Simple multi-task learning improves non-autoregressive translators. *arXiv preprint arXiv:2211.06075*.

Zhihao Wang, Longyue Wang, Jinsong Su, Junfeng Yao, and Zhaopeng Tu. 2023. Revisiting non-autoregressive translation at scale. *arXiv preprint arXiv:2305.16155*.

Bingzhen Wei, Mingxuan Wang, Hao Zhou, Junyang Lin, Jun Xie, and Xu Sun. 2019. Imitation learning for non-autoregressive neural machine translation. *arXiv preprint arXiv:1906.02041*.

Jiasheng Ye, Zaixiang Zheng, Yu Bao, Lihua Qian, and Quanquan Gu. 2023. Diffusion language models can perform many tasks with scaling and instruction-finetuning. *arXiv preprint arXiv:2308.12219*.

Zhirui Zhang, Shuangzhi Wu, Shujie Liu, Mu Li, Ming Zhou, and Tong Xu. 2019. Regularizing neural machine translation by target-bidirectional agreement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 443–450.

Chunting Zhou, Graham Neubig, and Jiatao Gu. 2019a. Understanding knowledge distillation in non-autoregressive machine translation. *arXiv preprint arXiv:1911.02727*.

Long Zhou, Jiajun Zhang, and Chengqing Zong. 2019b. Synchronous bidirectional neural machine translation. *Transactions of the Association for Computational Linguistics*, 7:91–105.

A Appendix

A.1 Calculation Detail of Bidirectional Contextual Transformer

This section describes the calculation detail of the BC-Transformer.

The BC-Decoder receives the graph positional embeddings $G = \{g_1, g_2, ..., g_L\}$ as input, calculates the vertex states V through the Transformer Blocks, then predicts target token candidate $\dot{Y}_i$, with the vertex state V and the learnable weight W_P.

$$[v_1, ..., v_L] = \text{Transformer-Blocks}(g_1, ..., g_L)$$
$$\dot{Y} = \text{softmax}(W_P v_i) \quad (20)$$

Equation 11 refers to the self-attention mechanism for the transition probability construction. The transition matrix is obtained by:

$$E = \text{softmax}(\frac{QK^T}{\sqrt{d}})$$
$$Q = VW_Q, K = VW_K \quad (21)$$

where d is the hidden size, to enable bidirectional contextual information sharing, we employ the same set of learnable weights W_Q and W_K on transition probability calculation of both directions.

A.2 Experiment Settings

We train the BC-Transformer on 4 $\times$ A100 (40G) GPUs. We follow the parameter settings of (Huang et al., 2022b), and show all our parameters on Table 6. For WMT16 En-Ro and Ro-En, we set the dropout rate 0.3. We select the checkpoint via the BES decoding of the validation set, and average the best five checkpoint for model generation.

We employ the ensemble beam search for the model generation with beam size 200, decode beta $\beta = 1.1$ and alpha $\alpha = \{1.0, 1.1, 1.2, 1.3, 1.4\}$.

A.3 Experiment Result on WMT17

To evaluate the capability of our proposed method in the large-scale corpus, we conduct the supplement experiment on WMT17 Zh-En/En-Zh. We compare the quality of the translation of the DA-Transformer (Huang et al., 2022b) with CKD and our proposed BC-Transformer with BCKD.

We follow Huang et al. (2022b) to use the same parameter setting as WMT14 En-De in Table 6 for model training. For BCKD and CKD, we set the update steps for 40k. We use tokenized BLEU for the Zh-En and sacreBLEU for En-Zh language pairs for model evaluation. The experiment results are shown in Table 7. BC-Transformer and BCKD consistently outperform the baseline model DA-Transformer and CKD.

A.4 Effect of BCKD on Iterative Refinement Models

To further explore the effect of BCKD on iterative refinement models, we train the CMLM model (Ghazvininejad et al., 2019) on our CKD and BCKD dataset, including WMT14 En-De, De-En, WMT16 En-Ro, and Ro-En. Due to the increased data quantity, the CMLM did not converge on the BCKD corpus with the original settings. We set the maximum update to 60k steps for CKD and BCKD on En-De and De-En translation pairs, follow the other settings of Ghazvininejad et al. (2019). Experiment results are shown in Table 8. The proposed BCKD consistently boosts the CMLM model's translation quality in all four directions.

A.5 Link Probability Visualization of BC-T and DA-T

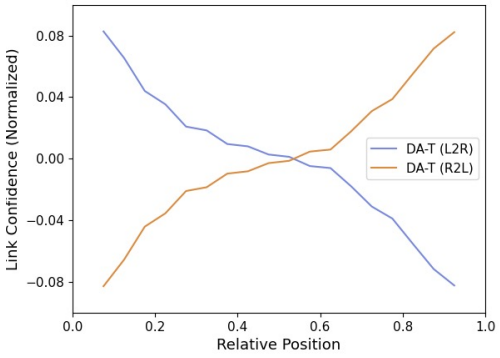

Figure 7: Link transition probability DA-Transformer. The figure clearly shows the decline of model confidence with the increase of previous output tokens.

In this section, we describe the calculation detail of link visualization to compare the DA-T and BC-T for AR Bias alleviation verification. We train two DA-Transformers with diverse link directions (L2R and R2L) and perform lookahead decoding on the three models ($\overleftarrow{DAT}$,$\overrightarrow{DAT}$,BCT).

For each target token y_i, we visualize the link transition probability e. We average the transition probability of two transition directions to model

WMT14 De-En Translation Case	
Source:	Schmucklose Nebenzimmer sind nicht optimal , das Ambiente sollte besonders sein .
Reference:	Plain adjoining rooms are not ideal , the ambience should be special .
DA-T:	Sorless side rooms is not good and the room should be good .
BC-T:	Jewellous side rooms are not optimal , the ambience should be special .

Table 5: Translation Cases Generated by BC-Transformer and DA-Transformer. BC-Transformer shows more advantages in the prevention of error accumulation and grammar coherence.

Parameter	De-En/En-De	Ro-En/En-Ro
upsample scale	8.0	8.0
glat probability	0.5:0.1	0.5:0.1
optimizer	adam	adam
adam betas	0.9,0.999	0.9,0.999
label smoothing	0.01	0.01
weight decay	0.01	0.01
dropout	0.1	0.3
lr scheduler	inverse-sqrt	inverse-sqrt
warmup updates	10000	10000
clip norm	0.1	0.1
learning rate	0.0005	0.0005
warmup init lr	1e-07	1e-07
stop min lr	1e-09	1e-09
max tokens	64k	64k
max update	300000	300000

Table 6: BC-T Training Parameters for WMT14 De-En / En-De & WMT16 Ro-En/En-Ro. We follow the settings as DA-Transformer (Huang et al., 2022b). For WMT16 Ro-En/En-Ro, we set the dropout rate to 0.3.

Model	WMT17	
	Zh-En	En-Zh
DA-T (wo/KD)	24.22	34.21
BC-T (wo/KD)	24.99	34.30
DA-T (w/KD)	24.90	34.35
BC-T (w/BCKD)	25.18	34.70

Table 7: Experiment result on WMT17 Zh-En/En-Zh corpus. We compare the translation quality with DA-T (Huang et al., 2022b). The BC-T with BCKD shows consistent improvement over DA-T and CKD.

the translation difficulty factor T_i.

$$\begin{aligned} \overleftarrow{P}_i &= e_{i+1,i} \\ \overrightarrow{P}_i &= e_{i-1,i} \\ T_i &= (\overleftarrow{P}_i + \overrightarrow{P}_i)/2 \end{aligned} \tag{22}$$

We visualize the normalized link transition probability $\overrightarrow{P}'_i$ as:

$$\begin{aligned} \overleftarrow{P}'_i &= \overleftarrow{P}_i - T_i \\ \overrightarrow{P}'_i &= \overrightarrow{P}_i - T_i \end{aligned} \tag{23}$$

For ensemble inference of BC-Transformer, we

Model	WMT14		WMT16	
	En-De	De-En	En-Ro	Ro-En
CMLM (w/KD)	27.58	30.78	32.52	33.17
CMLM (w/BCKD)	27.68	31.27	33.55	34.15

Table 8: Experiment result of CMLM model on WMT14 En-De/De-En and WMT16 En-Ro/Ro-En corpus. The BCKD consistently boosts the translation quality of the CMLM model..

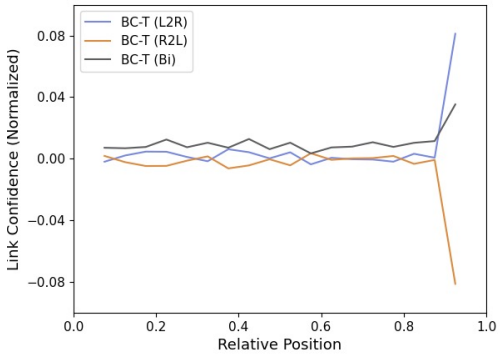

Figure 8: Link Transition probability of BC-Transformer. There is no apparent difference between L2R and R2L confidence, showing the alleviation of AR Bias.

model the normalized transition probability $\overleftrightarrow{P}'_i$ as:

$$\overleftrightarrow{P}'_i = \begin{cases} \overleftarrow{P}_i - T_i, \text{where } e_i \in \overleftarrow{E} \\ \overrightarrow{P}_i - T_i, \text{where } e_i \in \overrightarrow{E} \end{cases} \tag{24}$$

We show the normalized transition probability of $\overrightarrow{DAT}$,$\overleftarrow{DAT}$ in figure 7, BC-T in figure 8. Figure 7 validates the existence of AR Bias in DAT, while Figure 8 shows that the BCG training significantly mitigates the AR Bias of the DAG.

A.6 Supplement of Experiment Result: COMET Score

We also used the COMET score (Rei et al., 2020) to evaluate the effectiveness of our proposed methods: BC-Transformer and BCKD. We use the "wmt22-comet-da" as the evaluation model. The experiment results are shown in Table 9. Our proposed BC-Transformer shows superiority over the DA-

Transformer on WMT14 De-En and WMT16 Ro-En/ En-Ro. Moreover, our proposed BCKD significantly boosts the generation quality of GLAT and BC-Transformer.

A.7 Supplement of Experiment Result: Significance Test

We provide the significance test comparing our proposed BC-Transformer, the baseline model DA-Transformer, and the ART model (see Table 10). We use the CompareMT (Neubig et al., 2019) as the evaluation tool; we set bootstrap 1000 and probability threshold 0.05.

Our proposed BC-Transformer significantly outperforms the DA-Transformer on De-En, En-Ro, and Ro-En. The proposed BC-Transformer has no significant difference with ART in En-De, De-En, and En-Ro, and outperforms ART in Ro-En. Moreover, with BCKD, the BC-Transformer ourperforms ART significantly in all translation directions.

A.8 Elaboration of the Exposure Bias

The term of Exposure Bias was initially proposed by (Bengio et al., 2015), which refers to the discrepancy between training and inference of the Recurrent Neural Networks (RNN). Serdyuk et al. (2017) pinpointed that the RNN trained with teacher forcing struggles with long-range dependency and introduced future context information via a reverse RNN network. Liu et al. (2016) further pinpointed that RNN suffers from a fundamental issue of generating unbalanced tokens, resulting in the suffixes of its outputs being typically worse than the prefixes, which is due to the fact that later predictions directly depend on the previous predictions. Zhang et al. (2019) used the term 'exposure bias' to describe this problem, pinpointing the reason for the bias because of the AR structure of current Neural Machine Translation systems. In this paper, we use the term exposure bias from Zhang et al. (2019) to clearly describe the bias.

A.9 More Cases Comparing BC-Transformer and ART Model

In this section, we provide more cases to show the potential of NAR models on generation quality utilizing the removal of AR Bias. To make a fair comparison, we compare the BC-Transformer and ART models without KD. Table 11 shows that even without KD, the BC-Transformer can still overcome the error accumulation of ART.

Model	WMT14		WMT16	
	En-De	De-En	En-Ro	Ro-En
GLAT (wo/KD) (Qian et al., 2021)*	0.606	0.709	0.694	0.717
DA-T (wo/KD) (Huang et al., 2022b)*	0.683	0.758	0.754	0.761
GLAT (w/KD) (Qian et al., 2021)*	0.680	0.754	0.732	0.745
DA-T (w/KD) (Huang et al., 2022b)*	0.716	0.780	0.760	0.763
BC-T (wo/KD)	0.679	0.762	0.764	0.765
GLAT (w/BCKD)	0.694	0.763	0.742	0.757
BC-T (w/BCKD)	**0.723**	**0.783**	**0.77**	**0.775**

Table 9: COMET Scores on WMT14 En-De/De-En and WMT16 En-Ro/Ro-En benchmarks. Our Proposed BC-Transformer achieves an observable margin over the baseline DA-Transformer on WMT14 De-En and WMT16 En-Ro/Ro-En; BCKD significantly improves the generation quality of both GLAT and BC-Transformer.

Model	WMT14				WMT16			
	En-De		De-En		En-Ro		Ro-En	
	Result	p	Result	p	Result	p	Result	p
BC-T vs ART	-	0.489	-	0.332	-	0.134	>	0.049
BC-T (BCKD) vs ART	>	0.000	>	0.003	>	0.031	>	0.012
DA-T vs ART	-	0.338	<	0.001	<	0.000	-	0.497
BC-T vs DA-T	-	0.391	>	0.002	>	0.000	>	0.029

Table 10: Significant test with BC-T, DA-T, and ART. $>$ indicates the left system wins, $<$ indicates the right system wins. - indicates that the two systems have no significant difference. The BC-T has no significant difference with ART in En-De, De-En, and En-Ro, and outperforms ART in Ro-En. BC-T (BCKD) outperforms ART significantly.

	WMT14 De-En Case 1
Source:	Nach diesen verschiedenen Ausführungen teilten Revierleiter Christoph Wehle und Life @-@ Projektmanagerin Cornelia Bischoff die Helfer in Gruppen ein .
Reference:	After the various statements , forest ranger Christoph Wehle and Life Project manager Cornelia Bischoff divided the helpers into groups .
AT:	After these various presentations , the relief workers were divided into groups by Christoph Wehle , head of the district and life project manager Cornelia Bischoff .
BC-T:	After these various statements , district manager Christoph Wehle and Life project manager Cornelia Bischoff divided helpers into groups .
	WMT14 De-En Case 2
Source:	Aus jedem Clan beziehungsweise Team dürfen dann zwei am Einzelwettbewerb teilnehmen .
Reference:	Two members of each "clan" or team can then take part in the individual competition .
AT:	Two from each clan and team are allowed to participate in the competition .
BC-T:	From each " clan " or team , two can then take part in the individual competition .

Table 11: Translation Cases Generated by BC-T and ART. BC-T can overcome the translation error accumulation of ART models, especially in long sentences.