An EM Approach to Non-autoregressive Conditional Sequence Generation Appendix

A. Non-autoregressive Transformers are Universal Approximators of Sequence-to-Sequence Functions

A.1. Problem Definition

A non-autoregressive Transformer (Vaswani et al., 2017) a Transformer encoder and a non-autoregressive Transformer decoder. More concretely, both encoder and decoder consist of two types of layers: the multi-head attention layer Attn and the token-wise feed-forward layer FF, with both layers having a skip connection[1] (He et al., 2016). The encoder block in the non-autoregressive Transformer $\mathrm{t}_{\mathrm{enc}}$ maps an input $\boldsymbol{X} \in \mathbb{R}^{d \times n}$ consisting of d-dimensional embeddings of n tokens to an output $\mathrm{t}_{\mathrm{enc}}(\boldsymbol{X}) \in \mathbb{R}^{d \times m}$. It consists of a self-attention layer and a feed-forward layer. The decoder block in the non-autoregressive Transformer $\mathrm{t}_{\mathrm{dec}}$ maps an input $\boldsymbol{Y} \in \mathbb{R}^{d \times m}$ consisting of d-dimensional embeddings of m tokens and a context $\boldsymbol{X} \in \mathbb{R}^{d \times n}$ consisting of d-dimensional embeddings of n tokens to an output $\mathrm{t}_{\mathrm{dec}}(\boldsymbol{X}, \boldsymbol{Y}) \in \mathbb{R}^{d \times m}$. It consists of a self-attention layer, a encoder-decoder attention layer, and a feed-forward layer:

$$\mathrm{Attn}(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{Y} + \sum_{i=1}^{h} \boldsymbol{W}_O^i \boldsymbol{W}_V^i \boldsymbol{X} \cdot \sigma\left[(\boldsymbol{W}_Q^i \boldsymbol{X})^T (\boldsymbol{W}_K^i \boldsymbol{Y})\right], \tag{1}$$

$$\mathrm{FF}(\boldsymbol{Y}) = \boldsymbol{Y} + \boldsymbol{W}_2 \cdot \mathrm{ReLU}(\boldsymbol{W}_1 \cdot \boldsymbol{Y}), \tag{2}$$

$$\mathrm{t}_{\mathrm{enc}}(\boldsymbol{X}) = \mathrm{FF}(\mathrm{Attn}_{\mathrm{enc-self}}(\boldsymbol{X}, \boldsymbol{X})), \tag{3}$$

$$\mathrm{t}_{\mathrm{dec}}(\boldsymbol{X}, \boldsymbol{Y}) = \mathrm{FF}(\mathrm{Attn}_{\mathrm{enc-dec}}(\boldsymbol{X}, \mathrm{Attn}_{\mathrm{dec-self}}(\boldsymbol{Y}, \boldsymbol{Y}))), \tag{4}$$

where $\boldsymbol{W}_O^i \in \mathbb{R}^{d \times k}$, $\boldsymbol{W}_V^i, \boldsymbol{W}_K^i, \boldsymbol{W}_Q^i \in \mathbb{R}^{k \times d}$, $\boldsymbol{W}_2 \in \mathbb{R}^{d \times r}$, and $\boldsymbol{W}_1 \in \mathbb{R}^{r \times d}$ are learnable parameters. σ is the softmax function. Following Yun et al. (2020), we also do not use layer normalization (Ba et al., 2016) in the setup of our analysis.

The family of the Transformer encoders is $\mathbb{R}^{d \times n} \to \mathbb{R}^{d \times n}$ functions and can be defined as:

$$\mathcal{T}_{\mathrm{enc}}^{h,k,r} := \left\{ h : \mathbb{R}^{d \times n} \to \mathbb{R}^{d \times n} \,\middle|\, \begin{aligned} \boldsymbol{X}^0 &= \boldsymbol{X} \\ \boldsymbol{X}^i &= \mathrm{t}_{\mathrm{enc}}^{h,k,r}(\boldsymbol{X}^{i-1}) \\ h(\boldsymbol{X}) &= \boldsymbol{X}^M \end{aligned} \right\}, \tag{5}$$

where $\mathrm{t}_{\mathrm{enc}}^{h,k,r}$ denotes a Transformer encoder block defined by an attention layer with h heads of size k each, and a feed-forward layer with r hidden nodes. M is the number of stacked blocks.

Similarly, the family of the non-autoregressive Transformer decoders is $\mathbb{R}^{d \times (n+m)} \to \mathbb{R}^{d \times m}$ functions and can be defined as:

$$\mathcal{T}_{\mathrm{dec}}^{h,k,r} := \left\{ h : \mathbb{R}^{d \times (n+m)} \to \mathbb{R}^{d \times m} \,\middle|\, \begin{aligned} \boldsymbol{Y}^0 &= \boldsymbol{Y} \\ \boldsymbol{Y}^i &= \mathrm{t}_{\mathrm{dec}}^{h,k,r}(\boldsymbol{X}, \boldsymbol{Y}^{i-1}) \\ h(\boldsymbol{X}, \boldsymbol{Y}) &= \boldsymbol{Y}^N \end{aligned} \right\}, \tag{6}$$

where $\mathrm{t}_{\mathrm{dec}}^{h,k,r}$ denotes a Transformer decoder block defined by attention layers with h heads of size k each and a feed-forward layer with r hidden nodes. N is the number of stacked blocks.

Finally, the family of non-autoregressive Transformers is $\mathbb{R}^{d \times n} \to \mathbb{R}^{d \times m}$ functions and can be defined as:

$$\mathcal{T}^{h,k,r} := \left\{ g(\boldsymbol{X}) = h_2(h_1(\boldsymbol{X} + \boldsymbol{E}_1), \boldsymbol{E}_2) \,\middle|\, h_1 \in \mathcal{T}_{\mathrm{enc}}^{h,k,r} \text{ and } h_2 \in \mathcal{T}_{\mathrm{dec}}^{h,k,r} \right\}, \tag{7}$$

[1] The bias $\boldsymbol{b}$ is omitted for all matrix multiplication operations for brevity.

where $\boldsymbol{E}_1 \in \mathbb{R}^{d\times n}$ and $\boldsymbol{E}_2 \in \mathbb{R}^{d\times m}$ are the trainable positional embeddings.

A.2. Transformer Encoders are Universal Approximators of Sequence-to-Sequence Functions (Yun et al., 2020)

Recently, Yun et al. (2020) show that the Transformer encoders equipped with positional embeddings are universal approximators of all continuous $\mathbb{R}^{d\times n} \to \mathbb{R}^{d\times n}$ functions that map a compact domain in $\mathbb{R}^{d\times n}$ to $\mathbb{R}^{d\times n}$.

We first describe the results in Yun et al. (2020). Let us start by defining the target function class $\mathcal{F}_{\text{enc}}$, which consists of all continuous sequence-to-sequence functions with compact support that map a compact domain in $\mathbb{R}^{d\times n}$ to $\mathbb{R}^{d\times n}$. Here continuity is defined with respect to any entry-wise ℓ^p norm, $1 \le p < \infty$. Given two functions $f_1, f_2 : \mathbb{R}^{d\times n} \to \mathbb{R}^{d\times n}$, for $1 \le p < \infty$, we define a distance between them as

$$d_p(f_1, f_2) := \left(\int \| f_1(\boldsymbol{X}) - f_2(\boldsymbol{X}) \|_p^p d\boldsymbol{X} \right)^{1/p}. \tag{8}$$

The Transformer encoders with positional embeddings is defined as:

$$\mathcal{T}^{h,k,r}_{\mathrm{P-enc}} := \left\{ h_{\mathrm{P}}\boldsymbol{X} = h(\boldsymbol{X} + \boldsymbol{E}) | h \in \mathcal{T}^{h,k,r}_{\text{enc}} \text{ and } \boldsymbol{E} \in \mathbb{R}^{d\times n} \right\}, \tag{9}$$

where $\boldsymbol{E}$ is learnable positional embeddings. The following result shows that a Transformer encoder with positional embeddings with a constant number of heads h, head size k, and hidden layer of size r can approximate any function in $\mathcal{F}_{\text{enc}}$:

Theorem A.1. *Let $1 \le p < \infty$ and $\epsilon > 0$, then for any given $f \in \mathcal{F}_{\text{enc}}$, there exists a Transformer encoder $h \in \mathcal{T}^{2,1,4}_{\mathrm{P-enc}}$ such that we have $d_p(f, h) \le \epsilon$.*

We provide the sketch of the proof in Yun et al. (2020) here. Without loss of generality, we can assume that the compact support of f is contained in $[0, 1]^{d\times n}$. The proof of Theorem A.1 can be achieved in the following three steps:

Step 1: Approximate $\mathcal{F}_{\text{enc}}$ with piece-wise constant functions. We first use (a variant of) the classical result that any continuous function can be approximated up to arbitrary accuracy by piece-wise constant functions. For $\delta > 0$, we define the following class of piece-wise constant functions:

$$\overline{\mathcal{F}}_{\text{enc}}(\delta) := \left\{ f : \boldsymbol{X} \mapsto \sum\nolimits_{\boldsymbol{L}\in\mathbb{G}_\delta} \boldsymbol{A}_{\boldsymbol{L}} \mathbb{1}\{\boldsymbol{X} \in \mathbb{G}_{\boldsymbol{L}}\} \middle| \boldsymbol{A}_{\boldsymbol{L}} \in \mathbb{R}^{d\times n} \right\}, \tag{10}$$

where $\mathbb{G}_\delta := \{0, \delta, \dots, 1-\delta\}^{d\times n}$ and, for a grid point $\boldsymbol{L} \in \mathbb{G}_\delta$, $\mathbb{S}_{\boldsymbol{L}} := \prod_{j=1}^{d} \prod_{k=1}^{n} [L_{j,k}, L_{j,k} + \delta) \subset [0, 1]^{d\times n}$ denotes the associated cube of width δ. Let $\overline{f} \in \overline{\mathcal{F}}_{\text{enc}}(\delta)$ be such that $d_p(f, \overline{f}) \le \epsilon/3$.

Step 2: Approximate $\overline{\mathcal{F}}_{\text{enc}}(\delta)$ with *modified* Transformer encoders. We then consider a slightly modified architecture for Transformer networks, where the softmax operator $\sigma[\cdot]$ and $\mathrm{ReLU}(\cdot)$ are replaced by the hardmax operator $\sigma_{\mathrm{H}}[\cdot]$ and an activation function $\phi \in \Phi$, respectively. Here, the set of allowed activations Φ consists of all piece-wise linear functions with at most three pieces, where at least one piece is constant. Let $\overline{\mathcal{T}}^{h,k,r}_{enc}$ denote the function class corresponding to the sequence-to-sequence functions defined by the modified Transformer encoders. The following result establishes that the modified Transformer encoders in $\overline{\mathcal{T}}^{2,1,1}_{enc}$ can closely approximate functions in $\overline{\mathcal{F}}_{\text{enc}}(\delta)$.

Proposition A.1. *For each $\overline{f} \in \overline{\mathcal{F}}_{\text{enc}}(\delta)$ and $1 \le p < \infty$, $\exists\, \overline{g} \in \overline{\mathcal{T}}^{2,1,1}_{\text{enc}}$ such that $d_p(\overline{f}, \overline{g}) = O(\delta^{d/p})$.*

Step 3: Approximate *modified* Transformer encoders with (original) Transformer encoders. Finally, we show that $\overline{g} \in \overline{\mathcal{T}}^{2,1,1}$ can be approximated by $\mathcal{T}^{2,1,4}$. Let $g \in \mathcal{T}^{2,1,4}$ be such that $d_p(\overline{g}, g) \le \epsilon/3$.

Theorem A.1 now follows from these three steps, because we have

$$d_p(f, g) \le d_p(f, \overline{f}) + d_p(\overline{f}, \overline{g}) + d_p(\overline{g}, g) \le 2\epsilon/3 + O(\delta^{d/p}). \tag{11}$$

Choosing δ small enough ensures that $d_p(f, g) \le \epsilon$.

A.3. Proof Sketch of Proposition A.1 (Yun et al., 2020)

Especially, the proof of Proposition A.1 is decomposed into three steps:

Sub-step 1: Quantization by feed-forward layers Given an input $\boldsymbol{X} \in \mathbb{R}^{d\times n}$, a series of feed-forward layers in the modified Transformer encoder can quantize $\boldsymbol{X}$ to an element $\boldsymbol{L}$ on the extended grid $\mathbb{G}_\delta^+ := \{-\delta^{-nd}, 0, \delta, \dots, 1-\delta\}^{d\times n}$.

Sub-step 2: Contextual mapping by self-attention layers Next, a series of self-attention layers in the modified Transformer encoder can take the input $\boldsymbol{L}$ and implement a *contextual mapping* $q : \mathbb{G}_\delta \to \mathbb{R}^n$ such that, for $\boldsymbol{L}$ and $\boldsymbol{L}'$ that are not permutation of each other, all the elements in $q(\boldsymbol{L})$ and $q(\boldsymbol{L}')$ are distinct.

Sub-step 3: Function value mapping by feed-forward layers Finally, a series of feed-forward layers in the modified Transformer encoder can map elements of the contextual embedding $q(\boldsymbol{L})$ to the desired output value of $\overline{f} \in \overline{\mathcal{F}}_{\text{enc}}$ at the input $\boldsymbol{X}$.

A.4. Non-autoregressive Transformers are Universal Approximators of Sequence-to-Sequence Functions

In this paper, we take a further step and show that the non-autoregressive Transformers are universal approximators of all continuous $\mathbb{R}^{d\times n} \to \mathbb{R}^{d\times m}$ functions that map a compact domain in $\mathbb{R}^{d\times n}$ to $\mathbb{R}^{d\times m}$, where n and m can be different.

We start with describing the formal form of Theorem 4.1 in the main text. In the non-autoregressive conditional sequence generation problem, the target function class $\mathcal{F}_{s2s}$ becomes the set of all continuous sequence-to-sequence functions with compact support that map a compact domain in $\mathbb{R}^{d\times n}$ to $\mathbb{R}^{d\times m}$, where n and m can be different. Given two functions $f_1, f_2 : \mathbb{R}^{d\times n} \to \mathbb{R}^{d\times m}$, for $1 \le p < \infty$, similar to Eq. 8, we define a distance between them as

$$d_p(f_1, f_2) := \left(\int \|f_1(\boldsymbol{X}) - f_2(\boldsymbol{X})\|_p^p d\boldsymbol{X} \right)^{1/p}. \tag{12}$$

With the definition of non-autoregressive Transformers in Eq. 7, we have the following result:

Theorem A.2. *Let $1 \le p < \infty$ and $\epsilon > 0$, then for any given $f \in \mathcal{F}_{\text{s2s}}$, there exists a non-autoregressive Transformer $g \in \mathcal{T}^{2,1,4}$ such that we have $d_p(f, g) \le \epsilon$.*

The proof of Theorem A.2 can be done in a similar way as Theorem A.1. Especially, the step 1 and step 3 in the proof of Theorem A.1 can be seamlessly used here. We refer the readers to Yun et al. (2020) for the detailed proof of these two steps.

The step 2 in the proof Theorem A.2 is a bit different. Basically, we need to prove the following result:

Proposition A.2. *For each $\overline{f} \in \overline{\mathcal{F}}_{\text{s2s}}(\delta)$ and $1 \le p < \infty$, $\exists\, \overline{g} \in \overline{\mathcal{T}}^{2,1,1}$ such that $d_p(\overline{f}, \overline{g}) = O(\delta^{d/p})$.*

where $\overline{\mathcal{F}}_{\text{s2s}}(\delta)$ and $\overline{\mathcal{T}}^{2,1,1}$ are defined in a similar way as $\overline{\mathcal{F}}_{\text{enc}}(\delta)$ and $\overline{\mathcal{T}}_{\text{enc}}^{2,1,1}$, respectively. Similar to the proof of Proposition A.1, the proof of Proposition A.2 can be decomposed into three steps:

Sub-step 1*: Quantization by feed-forward layers in the encoder Given an input $\boldsymbol{X} \in \mathbb{R}^{d\times n}$, a series of feed-forward layers in *the encoder of the modified non-autoregressive Transformer* can quantize $\boldsymbol{X}$ to an element $\boldsymbol{L}$ on the extended grid $\mathbb{G}_\delta^+ := \{-\delta^{-nd}, 0, \delta, \dots, 1-\delta\}^{d\times n}$.

Sub-step 2*: Contextual mapping by attention layers in the encoder and the decoder Next, a series of attention layers in *the encoder and decoder of the modified non-autoregressive Transformer* can take the input $\boldsymbol{L}$ and implement a *contextual mapping* $q : \mathbb{G}_\delta \to \mathbb{R}^m$ such that, for $\boldsymbol{L}$ and $\boldsymbol{L}'$ that are not permutation of each other, all the elements in $q(\boldsymbol{L})$ and $q(\boldsymbol{L}')$ are distinct.

Sub-step 3*: Function value mapping by feed-forward layers in the decoder Finally, a series of feed-forward layers in *the decoder of the modified non-autoregressive Transformer* can map elements of the contextual embedding $q(\boldsymbol{L})$ to the desired output value of $\overline{f} \in \overline{\mathcal{F}}_{\text{s2s}}$ at the input $\boldsymbol{X}$.

Since Sub-step 1* and Sub-step 3* are exactly the same as Sub-step 1 and Sub-step 3 in the proof of Proposition A.1, we only provide the proof of Sub-step 2*. We refer the readers to Yun et al. (2020) for the detailed proof of these two sub-steps.

A.5. Proof of Sub-step 2* in Proposition A.2

Without loss of generality, we can assume that the compact support of f is contained in $[0,1]^{d\times n}$. Following Yun et al. (2020), we choose

$$\boldsymbol{E}_1 = \begin{bmatrix} 0 & 1 & 2 & \cdots & n-1 \\ 0 & 1 & 2 & \cdots & n-1 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 1 & 2 & \cdots & n-1 \end{bmatrix}.$$

and

$$\boldsymbol{E}_2 = \begin{bmatrix} 0 & 1 & 2 & \cdots & m-1 \\ 0 & 1 & 2 & \cdots & m-1 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 1 & 2 & \cdots & m-1 \end{bmatrix}.$$

By Sub-step 1*, we quantize any input $\boldsymbol{X} + \boldsymbol{E}_1$ to its quantized version with the feed-forward layers in the Transformer encoder. We call this quantized version $\boldsymbol{L}$:

$$\boldsymbol{L} \in [0:\delta:1-\delta]^d \times [1:\delta:2-\delta]^d \times \cdots \times [n-1:\delta:n-\delta]^d.$$

We do not need to quantize $\boldsymbol{E}_2$ in our Sub-step 1* with the feed-forward layers in the Transformer decoder because $\boldsymbol{E}_2$ is already quantized.

As done in Lemma 6 in Yun et al. (2020), we define $\boldsymbol{u} := (1, \delta^{-1}, \ldots, \delta^{-d+1})$ and $l_j := \boldsymbol{u}^T \boldsymbol{L}_{:,j}$, for all $j \in [n]$. Next, following the construction in Appendix C.2 in Yun et al. (2020), with $n(1/\delta)^d$ self-attention layers in the Transformer encoder, we can get $\widetilde{l}_1 < \widetilde{l}_2 < \cdots < \widetilde{l}_n$ such that the map from $\mathcal{L}$ to $\widetilde{l}_n$ is one-to-one. In addition, $\widetilde{l}_n$ is bounded by $n\delta^{-(n+1)d}$.

Finally, in a similar way as Appendix B.5.1 in Yun et al. (2020), we add an extra encoder-decoder attention layer with attention part $n\delta^{-(n+1)d-1}\Psi(\cdot;0)$. This layer shifts all the layers in the Transformer decoder by $n\delta^{-(n+1)d-1}\widetilde{l}_n$. We define the output of this layer as $g_c(\boldsymbol{L})$. In this way, we ensure that different contexts $\boldsymbol{L}$ are mapped to distinct numbers in $\boldsymbol{u}^T g_c(\boldsymbol{L})$, thus implementing a contextual mapping.

B. Proof of Proposition 5.1

We prove this proposition by contradiction. Assuming that the k-th likely label in position i is chosen by the CRF algorithm and $k > 3$, we consider the following two cases:

Case 1: $i = 0$ is the first position or $i = n-1$ is the last position. Without loss of generality, we can assume $i = 0$. The first and second labels in the current decoding is denoted by l_0^* and l_1^*. We also denote the top 2 labels in position 0 as $l_{0,1}$ and $l_{0,2}$. If $l_{0,1} = l_1^*$, we can set l_0^* to be $l_{0,2}$, which construct a label sequence with higher likelihood in the CRF model. Otherwise, we can set l_0^* to be $l_{0,1}$. In both cases, $k > 3$ is not the optimal solution.

Case 2: i is the neither the first position nor the last position. We denote the labels on the position before and after i as l_{i-1}^* and l_{i+1}^*. We denote the j-th likely label on the position i as $l_{i,j}$. In this case, we will always find such a $j \leq 3$ that $l_{i,j} \neq l_{i-1}^*$ and $l_{i,j} \neq l_{i+1}^*$. Therefore, $k > 3$ is still not the optimal solution. □

C. Illustration of Different Decoding Approaches

Fig. 1 shows how the proposed optimal de-duplicated decoding method solves the sub-optimal decoding problem of the post de-duplication method.

D. Model Settings

The Non-Autoregressive Transformer model (NAT) is developed using the general encoder-decoder framework which is the same as the Autoregressive Transformer (AT) model. Fig. 2 shows the architectures of NAT and AT. We use a simplified

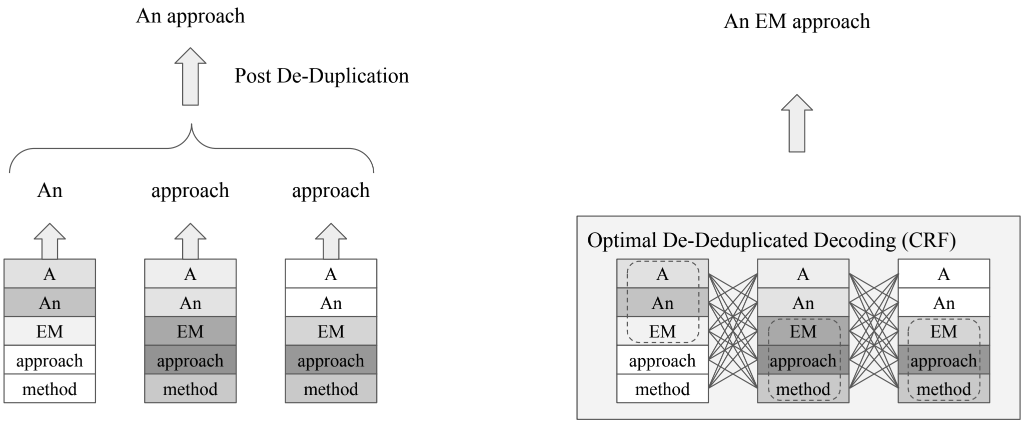


Figure 1. Illustration of different decoding methods. Darker boxes represent higher likelihood. The argmax decoding method produces "An approach appraoch" as the result, which contains word duplication. The empirical post de-duplication method can solve the word duplication problem, but after collapsing, the length of the target sequence is changed. This will cause a discrepancy between the predicted target length and the actual sequence length and thus make the final output sub-optimal. The proposed Optimal De-Duplicated (ODD) decoding can produce the optimal prediction in the CRF framework. Note that OOD decoding only needs to consider the top-3 labels for each position in the forward-backward algorithm, which is very efficient.

version of the NAT model in Gu et al. (2017), that is, we do not copy the source embeddings as the input of the Transformer decoder and do not use the positional attention proposed in Gu et al. (2017). The input to our Transformer decoder is simply the padding symbols. More details about the description about the architectures can be found in Vaswani et al. (2017); Gu et al. (2017).

We use four model settings in our experiments, including `toy`, `small`, `base`, and `large`. The detailed configurations of these four model settings can be found in Tab. 1.

E. Analysis of Training Examples for the NAR Model

In Tab. 2, we present randomly picked examples from the training data for the NAR model on the WMT14 De-En task. We can find that the proposed EM algorithm constantly changes the training examples for the NAR model.

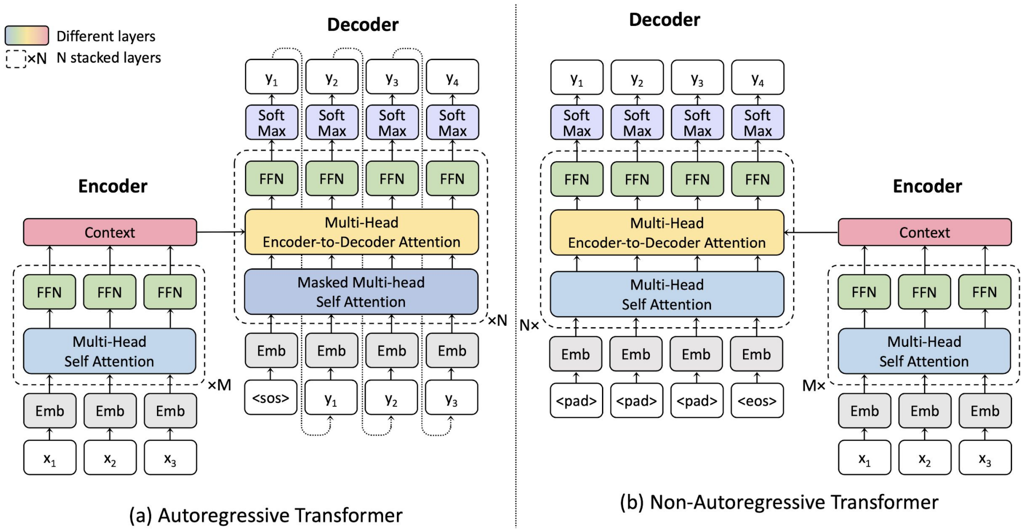
F. Analysis of Translation Results

In Tab. 3, we present randomly picked translation outputs from the test set of WMT14 De-En. We have the following observations:

- The proposed OOD decoding method preserves the original predicted length of tokens, which avoid the sub-optimal problem of the post de-duplication method.
- The proposed EM algorithm can effectively jointly optimize both the AR model and the NAR model. During EM iterations, the multi-modality in the AR models is reduced, while the translation quality of the NAR models is improved.

Table 1. Transformer model settings

	encoder-layer	decoder-layer	hidden-size	filter-size	num-heads
`toy`	3	3	256	1024	4
`small`	5	5	256	1024	4
`base`	6	6	512	2048	8
`large`	6	6	1024	4096	16



(a) Autoregressive Transformer

(b) Non-Autoregressive Transformer

Figure 2. The architectures of Autoregressive Transformer and Non-autoregressive Transformer used in this paper.

References

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Gu, J., Bradbury, J., Xiong, C., Li, V. O., and Socher, R. Non-autoregressive neural machine translation. *arXiv preprint arXiv:1711.02281*, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Yun, C., Bhojanapalli, S., Rawat, A. S., Reddi, S., and Kumar, S. Are transformers universal approximators of sequence-to-sequence functions? In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ByxRM0Ntvr.

Source	, um den Korb zu verkleinern (bis 75 %) und in die Ecke zu schieben .
Ground Truth	to resize the fragment (by 75 %) and move it to the lower right corner .
Iteration 0	to reduce the basket (up to 75 %) and move it to the corner .
Iteration 1	to reduce the basket (up to 75 %) and push it into the corner .
Iteration 2	to reduce the basket (up to 75 %) and put it into the corner .
Source	In den Interviews betonten viele Männer , dass ihre Erwerbsabweichung ihre Karriere behindere .
Ground Truth	In the interviews , many men emphasized that their employment deviation has hindered their careers .
Iteration 0	In the interviews , many men emphasized that their divorce in employment hinders their careers .
Iteration 1	In the interviews , many men stressed that their deviation from employment hindered their careers .
Iteration 2	In the interviews , many men stressed that their deviation in employment hinders their careers .
Source	Um einen Pferd gesund und munter zu halten , müssen Sie seine physischen Bedürfnisse beachten .
Ground Truth	To keep your horse well , healthy and content you must satisfy its physical needs .
Iteration 0	In order to keep a horse healthy and healthy , you must take into account its physical needs .
Iteration 1	In order to keep a horse healthy and cheerful , you must take into account its physical needs .
Iteration 2	In order to keep a horse healthy and cheerful , you must take into account your physical needs .
Source	Der Effekt von &apos; Eiskältefalle &apos; wird nun bei erlittenem Schaden abgebrochen .
Ground Truth	Freezing Trap now breaks on damage .
Iteration 0	Ice Cat Trap effect will now be discarded if damage is dealt .
Iteration 1	The effect of Ice Cage Trap will now be aborted in case of damage suffered .
Iteration 2	The effect of ice cold trap is now aborted in the event of damage suffered .
Source	Wir haben in der Europäischen Union Möglichkeiten , wirksam gegen die Arbeitslosigkeit vorzugehen und zwar so , daß man da , wo es am nötigsten ist , auch etwas davon spürt .
Ground Truth	We have the opportunity , in the EU , to do something that will have a positive effect on unemployment , characterised by taking action where there is the greatest need .
Iteration 0	We in the European Union have the means to combat unemployment effectively , in such a way that we feel something of it where it is most necessary .
Iteration 1	We in the European Union have opportunities to take effective action against unemployment , in such a way that we can feel something about it where it is most necessary .
Iteration 2	We in the European Union have opportunities to take effective action against unemployment , in such a way that we also feel something of it where it is most necessary .
Source	In der Altstadt sind die Gassen so eng und verwinkelt , dass ein Auto nur mühsam vorankommt .
Ground Truth	Kaneo Settlement - Start the walk to Kaneo from St. Sophia church .
Iteration 0	In the old town , the streets are so narrow and winding that a car can only progress with difficulty .
Iteration 1	In the old town , the streets are so narrow and winding that a car is progressing hard .
Iteration 2	In the old town , the streets are so narrow and winding that a car is only progressing laboriously .
Source	Sie müssen sich nur einmal die weltweit steigende Anzahl und Häufigkeit von Naturkatastrophen ansehen , um die Folgen der Klimaveränderung zu erkennen .
Ground Truth	They only need to look at the increasing number and frequency of natural disasters worldwide to see its impact .
Iteration 0	You only have to look at the increasing number and frequency of natural disasters worldwide to see the consequences of climate change .
Iteration 1	They only have to look at the increasing number and frequency of natural disasters around the world to see the consequences of climate change .
Iteration 2	They only have to look at the increasing number and frequency of natural disasters around the world in order to identify the consequences of climate change .

Table 2. Examples in the training data for the NAR model on the WMT14 De-En task.

Appendix

Source	Sie ist die Tochter von Peter Tunks , einem ehemaligen Spieler der australischen Rubgy @-@ Liga , der sich an das Außenministerium in Canberra mit der Bitte um Hilfe für seine Tochter gewandt hat .
Ground Truth	She is the daughter of former Australian league player Peter Tunks , who has appealed to the Department of Foreign Affairs in Canberra to assist his daughter .
AR model - Iter 0	She is the daughter of Peter Tunks , a former player of the Australian Rubgy League , who has addressed to the Ministry of Foreign Affairs in Canberra asking for help for his daughter .
AR model - Iter 2	She is the daughter of Peter Tunks , a former player of the Australian Rubgy League who addressed the Ministry of Foreign Affairs in Canberra asking for help for his daughter .
NAR model - Iter 0	She is the daughter of Peter Tunks , a former player of the Australian Rubgy League , who addressed the Ministry of Foreign Affairs Canberberra asking help help his daughter .
NAR model - Iter 0 w/ post de-duplication	She is the daughter of Peter Tunks , a former player of the Australian Rubgy League , who addressed the Ministry of Foreign Affairs Canberra asking help his daughter .
NAR model - Iter 2	She is the daughter of Peter Tunks , a former player of the Australian Rubgy League ague who addressed the Ministry of Foreign Affairs in Canberra ra asking for help to his daughter .
NAR model - Iter 2 w/ post de-duplication	She is the daughter of Peter Tunks , a former player of the Australian Rubgy League ague who addressed the Ministry of Foreign Affairs in Canberra asking for help to his daughter .
NAR model - Iter 2 w/ ODD decoding	She is the daughter of Peter Tunks , a former player of the Australian Rubgy League ague who addressed the Ministry of Foreign Affairs in Canberra for asking for help to his daughter .
Source	In australischen Berichten war zu lesen , dass sie in der Zwischenzeit im Ferienort Krabi in Südthailand Urlaub macht .
Ground Truth	Reports in Australia said that in the meantime , she was holidaying at the resort area of Krabi in Southern Thailand .
AR model - Iter 0	In Australian reports it was read that in the meantime it is making a holiday in the holiday resort of Krabi in South thailand .
AR model - Iter 2	Australian reports read that , in the meantime , it is a holiday in the resort of Krabi in southern Thailand .
NAR model - Iter 0	Australian reports have that , in the meantime , they is a holiday holiday southern southern in the Krabi of Krabi .
NAR model - Iter 0 w/ post de-duplication	Australian reports have that , in the meantime , they is a holiday southern in the Krabi of Krabi .
NAR model - Iter 2	Australian reports read that , in the meantime , it is a holiday holiday the resort of Kraresort in southern Thailand .
NAR model - Iter 2 w/ post de-duplication	Australian reports read that , in the meantime , it is a holiday the resort of Kraresort in southern Thailand .
NAR model - Iter 2 w/ ODD decoding	Australian reports read that , in the meantime , it is a holiday in the resort of Kraresort in southern Thailand .

Table 3. Examples of translation outputs on the WMT14 De-En task. We do not apply rescoring to the NAR model's outputs.