Motivation: Non-autoregressive sequence models were proposed to reduce the inference time. However, these models assume that the decoding process of each token is conditionally independent from others. Such a generation process makes the output sentence inconsistent, and thus the learned non-autoregressive models could only achieve inferior accuracy compared to their autoregressive counterparts.

Solution: To improve the decoding consistency and reduce the inference cost at the same time, we propose to incorporate a structured inference module into the non-autoregressive models. Specifically, we design an efficient approximation for Conditional Random Fields (CRF) for non-autoregressive sequence models, and further propose a dynamic transition technique to model positional contexts in the CRF.

Multimodality Problem: Although existing work on non-autoregressive models could tackle the same problem by simplifying autoregressive sequence models with a CRF via the autoregressive-to-CRF transition, this can no longer be realized with non-autoregressive models. Therefore, we propose to incorporate the autoregressive approximation into the non-autoregressive model to provide an effective approach to deal with the large vocabulary in machine translation models.

Structured Decoding: Autoregressive sequence models are based on a chain of conditional probabilities with a left-to-right causal structure:

\[ p(y|x) = \prod_{t=1}^{T} p(y_t|x_1:t), \]

Non-autoregressive sequence models were proposed to alleviate the inference latency by removing the sequential dependencies within the target sentence:

\[ p(y|x) = p(T_t|x) \prod_{t=1}^{T} p(y_t|x_t) \]

To tackle the multimodality problem, we incorporate a structured inference module in the non-autoregressive decoder to directly model the multimodal distribution of target sequences. The probability of the target sequence is globally normalized:

\[ p(y|x) = p(T_t|x) \cdot \text{softmax} \left( \sum_{t=1}^{T} \theta_{t-1:t}^{(m)} y_{t-1:t} | x_t \right). \]

TL; DR: We improve non-autoregressive sequence models with a CRF and provide an effective approach to deal with the large vocabulary in machine translation models.

Experimental Results: Table 1: Cases on IWSLT14 En-De. Compared to their ART counterparts, NART models suffer from severe decoding inconsistency problem, which can be solved by CRF-based structured decoding.

Table 2: Performance of BLEU scores on WMT14 En-De/De-En and IWSLT14 En-De tasks. The number in the parentheses denotes the performance gap between NART models and their ART teachers. "/" denotes that the results are not reported. LSTMBased are results from [2, 27]. CNN-based results are from [28]. Transformer [1] results are based on our own reproduction.