Fast Structured Decoding for Sequence Models

Zhiqing Sun[1,*], Zhuohan Li[2,*], Haoqing Wang[3], Di He[3], Zi Lin[3], Zhi-Hong Deng[3]
[1]Carnegie Mellon University [2]University of California, Berkeley [3]Peking University

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.

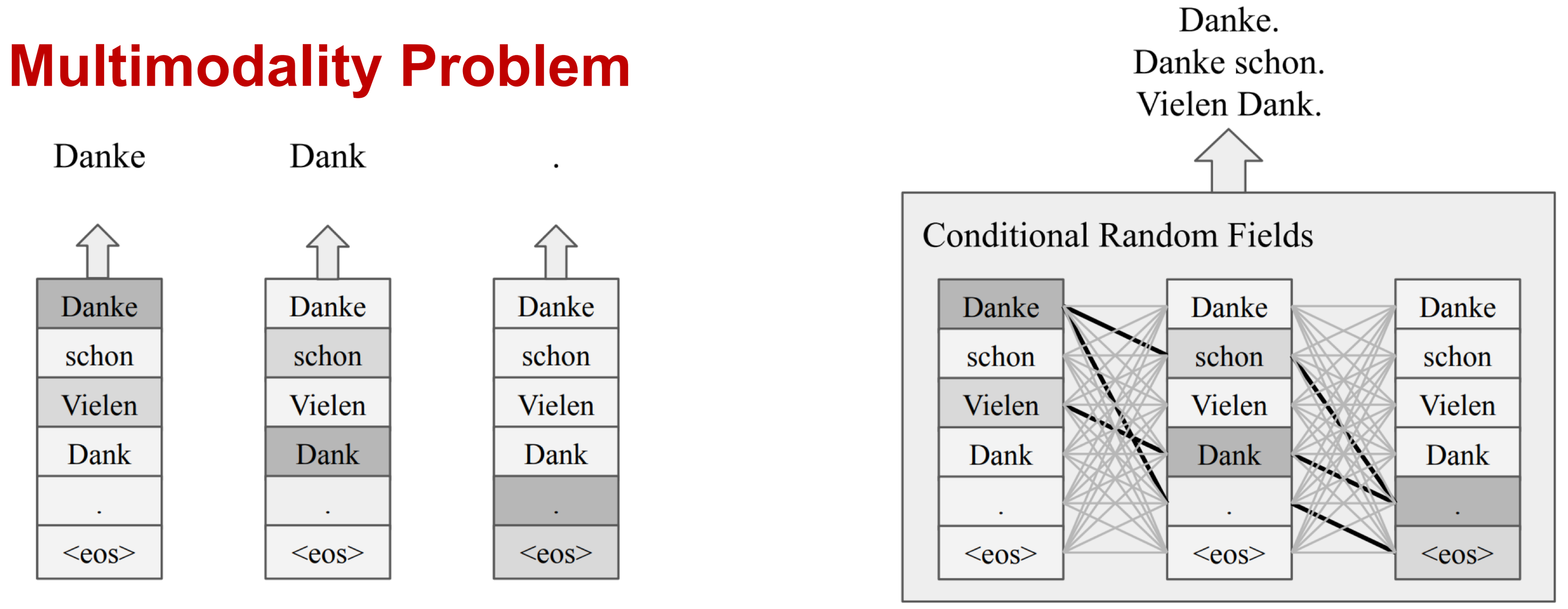
Motivation

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



Structured Decoding

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

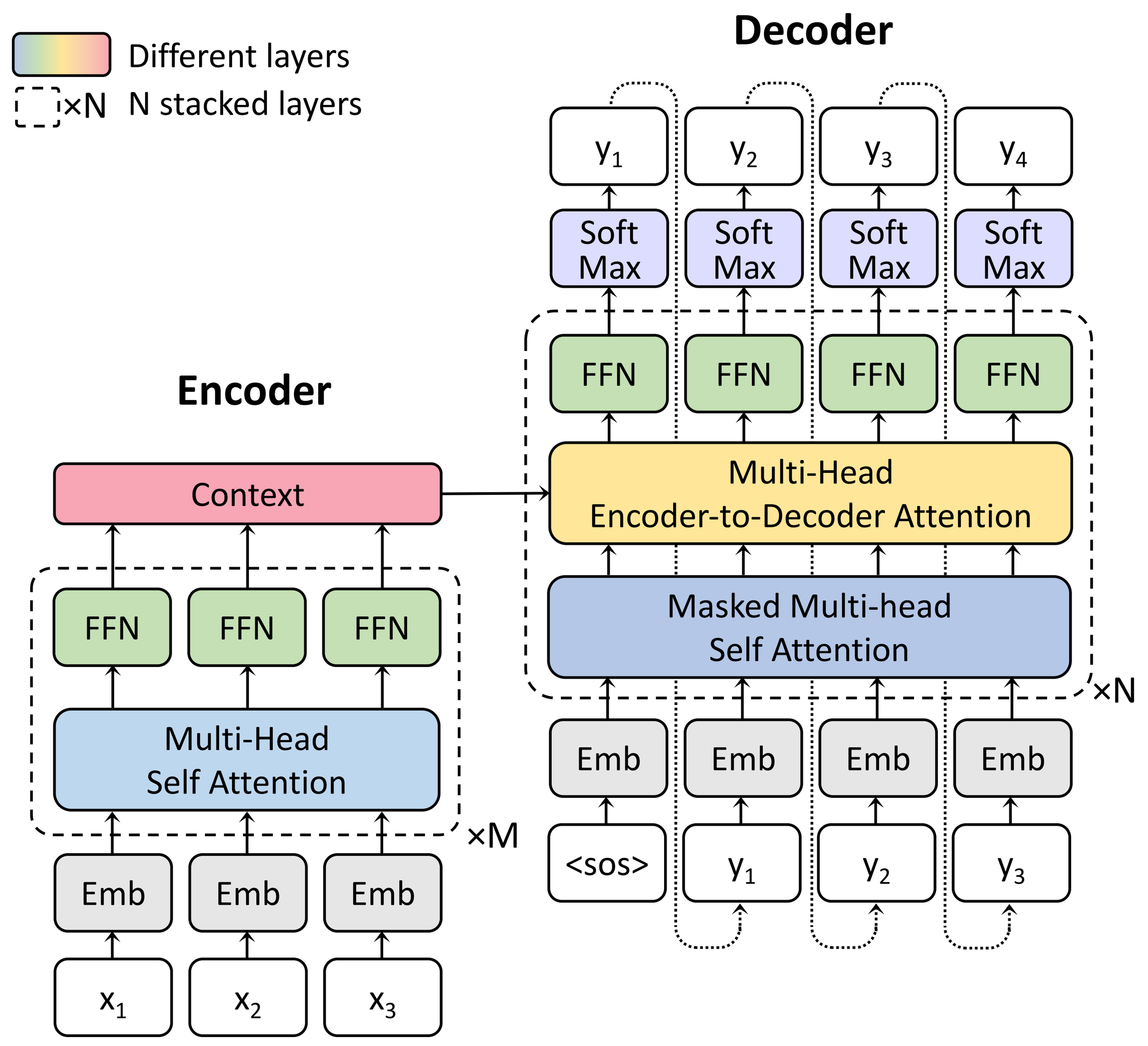
$$p(y|x) = \prod_{i=1}^{T'} p(y_i|y_{<i}, x),$$

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

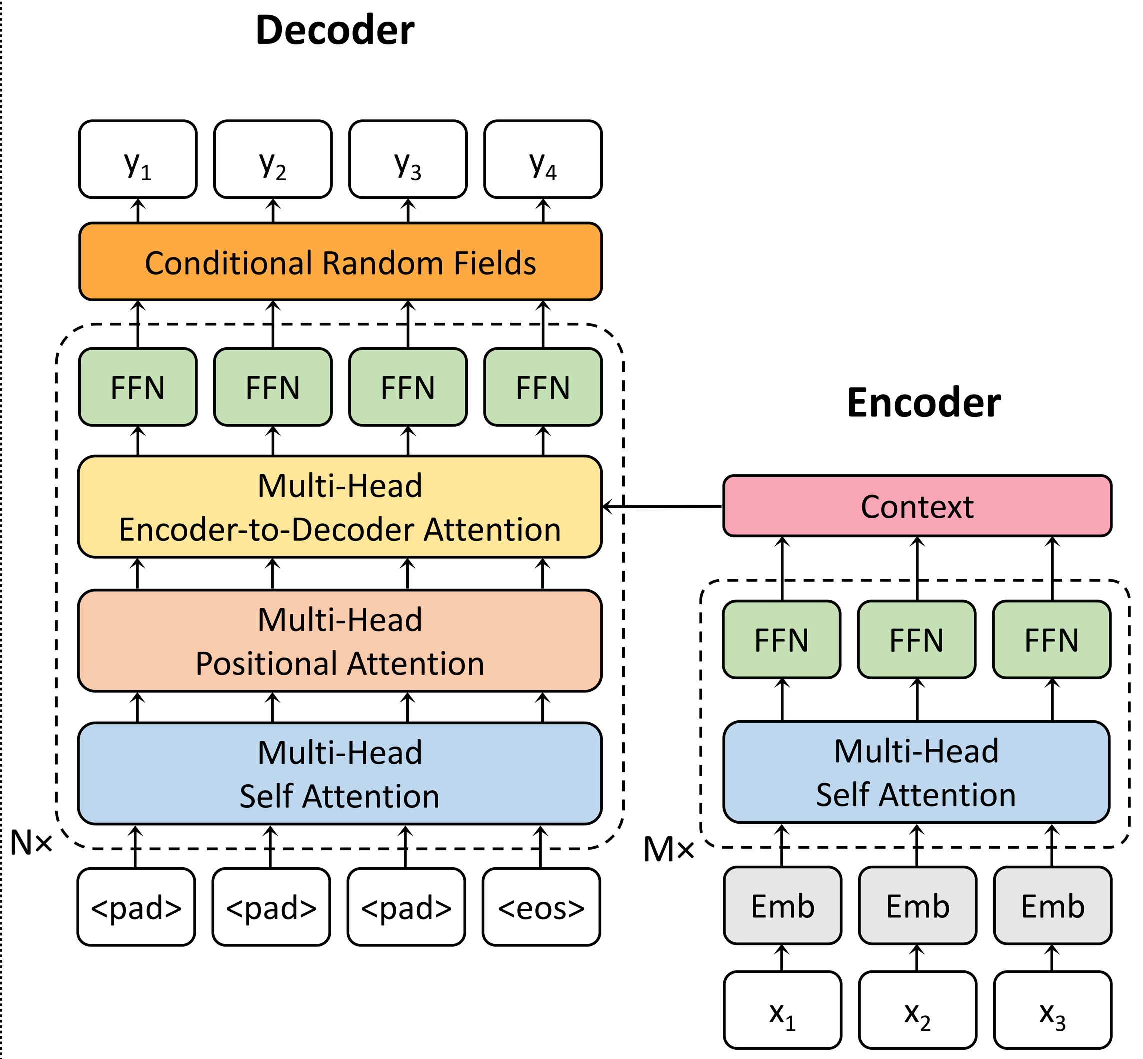
$$p(y|x) = p(T'|x) \cdot \prod_{i=1}^{T'} p(y_i|x)$$

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 sentence is globally normalized:

$$p(y|x) = p(T'|x) \cdot \text{softmax}\left(\sum_{i=2}^{T'} \theta_{i-1,i}(y_{i-1}, y_i \mid x)\right)$$



Autoregressive Transformer



Non-Autoregressive Transformer with Conditional Random Fields

Conditional Random Fields

$$P(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{i=1}^{n} s(y_i, x, i) + \sum_{i=2}^{n} t(y_{i-1}, y_i, x, i)\right)$$

$$t(y_{i-1}, y_i, x, i) \ = \ M_{y_{i-1}, y_i}$$

Low-rank approximation for transition matrix M

$$M = E_1 E_2^T$$

Beam approximation for CRF

For each position i, we heuristically truncate all $|V|$ candidates to a pre-defined beam size k. We keep k candidates with the highest label scores $s(\cdot, x, i)$ for each position i, and accordingly crop the transition matrix between each pair of $i-1$ and i.

Dynamic CRF transition

$$M^i_{dynamic} = f([h_{i-1}, h_i]),$$

$$M^i = E_1 M^i_{dynamic} E_2^T,$$

$$t(y_{i-1}, y_i, x, i) = M^i_{y_{i-1}, y_i},$$

Latency of CRF decoding

Unlike vanilla non-autoregressive decoding, the CRF decoding can no longer be parallelized. However, due to our beam approximation, the computation of linear-chain CRF $O(nk^2)$ is in theory still much faster than autoregressive decoding.

Thanks Jiatao Gu! Our model is in `fairseq` now: https://tinyurl.com/structured-nart

SCAN ME

Experimental Results

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

Source:	jeden morgen fliegen sie 240 kilometer zur farm .
Target:	every morning , they fly 240 miles into the farm .
ART:	every morning , they fly 240 miles to the farm .
NART:	every morning , you fly 240 miles to every morning .
NART-CRF:	every morning , they fly 240 miles to the farm .
Source:	ich weiß , dass wir es können , und soweit es mich betrifft ist das etwas , was die welt jetzt braucht .
Target:	i know that we can , and as far as i 'm concerned , that 's something the world needs right now .
ART:	i know that we can , and as far as i 'm concerned , that 's something that the world needs now
NART:	i know that we can it , , as as as as it it it is , it 's something that the world needs now .
NART-CRF:	i know we can do it , and as far as i 'm concerned that 's something that the world needs now .

Table 2: Performance of BLEU score on WMT14 En-De/De-En and IWSLT14 De-En tasks. The number in the parentheses denotes the performance gap between NART models and their ART teachers. "/" denotes that the results are not reported. LSTM-based results are from [2, 27]; CNN-based results are from [5, 28]; Transformer [1] results are based on our own reproduction.[6]

Models	WMT14 En-De	WMT14 De-En	IWSLT14 De-En	Latency	Speedup
Autoregressive models					
LSTM-based [2]	24.60	/	28.53	/	/
CNN-based [5]	26.43	/	32.84	/	/
Transformer [1] (beam size = 4)	27.41	31.29	33.26	$387ms^\ddagger$	1.00×
Non-autoregressive models					
FT [6]	17.69 (5.76)	21.47 (5.55)	/	$39ms^\dagger$	$15.6\times^\dagger$
FT [6] (rescoring 10)	18.66 (4.79)	22.41 (4.61)	/	$79ms^\dagger$	$7.68\times^\dagger$
FT [6] (rescoring 100)	19.17 (4.28)	23.20 (3.82)	/	$257ms^\dagger$	$2.36\times^\dagger$
IR [9] (adaptive refinement)	21.54 (3.03)	25.43 (3.04)	/	/	$2.39\times^\dagger$
Non-autoregressive models (Ours)					
NART	20.27 (7.14)	22.02 (9.27)	23.04 (10.22)	$26ms^\ddagger$	$14.9\times^\ddagger$
NART (rescoring 9)	24.22 (3.19)	26.21 (5.08)	26.79 (6.47)	$50ms^\ddagger$	$7.74\times^\ddagger$
NART (rescoring 19)	24.99 (2.42)	26.60 (4.69)	27.36 (5.90)	$74ms^\ddagger$	$5.22\times^\ddagger$
NART-CRF	23.32 (4.09)	25.75 (5.54)	26.39 (6.87)	$35ms^\ddagger$	$11.1\times^\ddagger$
NART-CRF (rescoring 9)	26.04 (1.37)	28.88 (2.41)	29.21 (4.05)	$60ms^\ddagger$	$6.45\times^\ddagger$
NART-CRF (rescoring 19)	26.68 (0.73)	29.26 (2.03)	29.55 (3.71)	$87ms^\ddagger$	$4.45\times^\ddagger$
NART-DCRF	**23.44 (3.97)**	**27.22 (4.07)**	**27.44 (5.82)**	$37ms^\ddagger$	$10.4\times^\ddagger$
NART-DCRF (rescoring 9)	**26.07 (1.34)**	**29.68 (1.61)**	**29.99 (3.27)**	$63ms^\ddagger$	$6.14\times^\ddagger$
NART-DCRF (rescoring 19)	**26.80 (0.61)**	**30.04 (1.25)**	**30.36 (2.90)**	$88ms^\ddagger$	$4.39\times^\ddagger$

Table 3: BLEU scores of beam approxiamtion ablation study on WMT En-De.

CRF beam size k	1	2	4	8	16	32	64	128	256
NART-CRF	15.10	20.67	22.54	23.04	23.22	23.26	23.32	23.33	**23.38**
NART-CRF (resocring 9)	19.61	23.93	25.48	25.86	25.93	26.01	26.04	**26.09**	26.08
NART-CRF (resocring 19)	20.02	25.00	26.28	26.56	26.57	26.65	26.68	**26.71**	26.66