



Tree-Structured Non-Autoregressive Decoding for Sequence-to-Sequence Text Generation



Pengyu Ji Yufei Liu Xiang Hu Kewei Tu

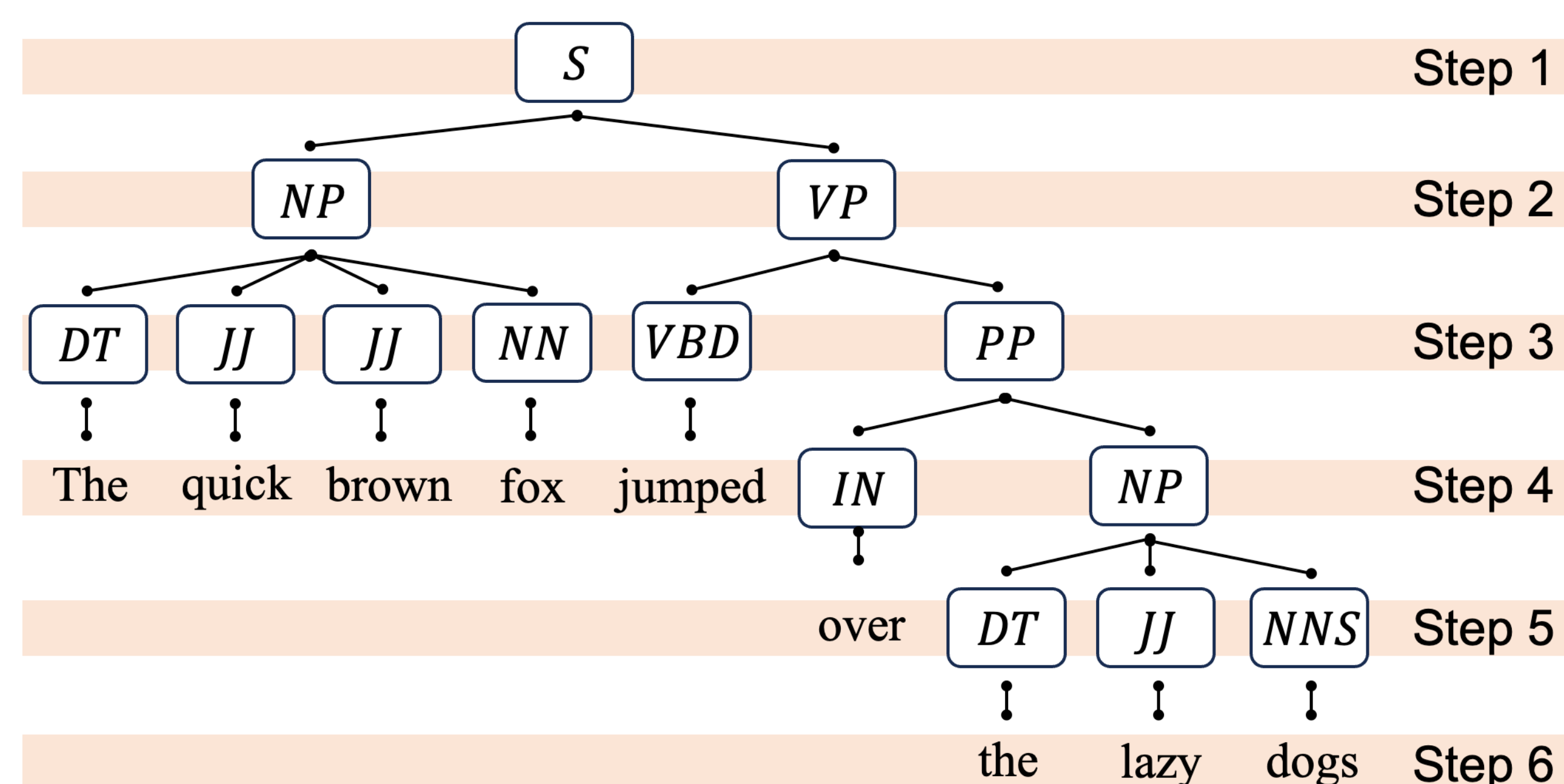
School of Information Science and Technology, ShanghaiTech University

Ant Group

① Overview

- Autoregressive (AT): high latency / good quality
- Non-Autoregressive (NAT): faster inference / degraded quality
- TNAD: a new alternative to AT and NAT, in the trade-off between efficiency and quality
 - Generate a sentence by top-down layer-wise expansion of its constituency parse tree
 - Generate all elements within a layer in parallel

② Generation Illustration



③ Generation Process

Each layer is expanded in one step by two operations:
Branching Prediction (BP) & Node Generation (NG).

BP: Predict branching factors for non-terminals

- If only 1 child node, identify it as a token; otherwise, identify them as non-terminals

NG: Predict labels for all new nodes

- Prediction based on node types (token / non-terminal)

④ Model Details

To incorporate tree structure information:

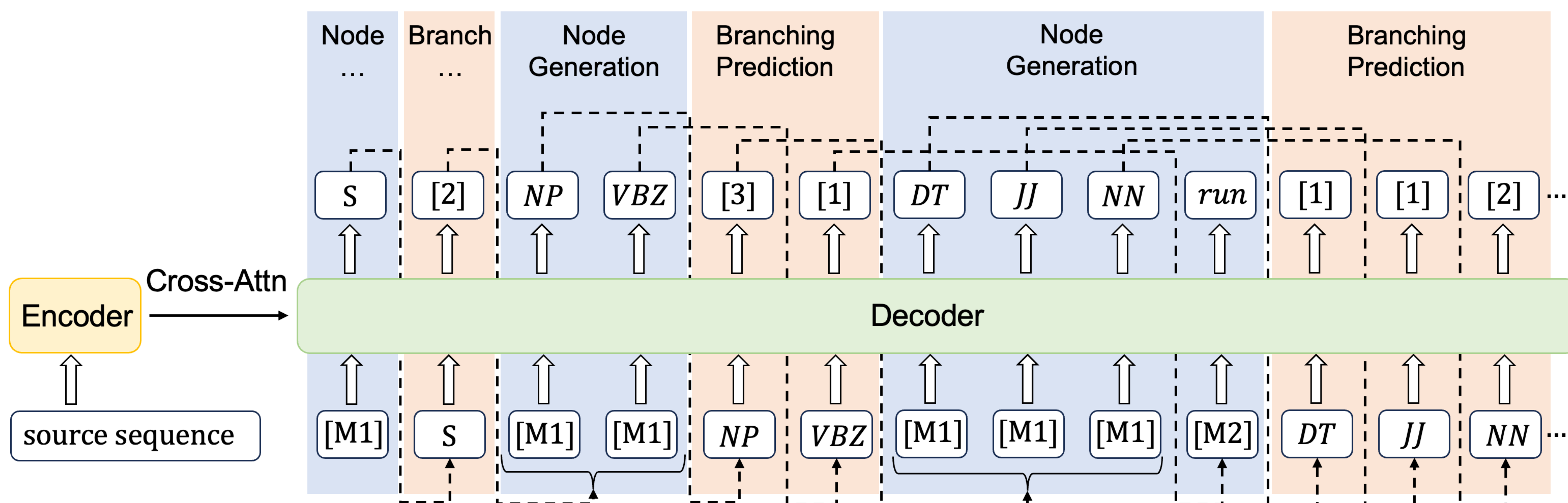
Attention Mask & Attention Bias

Attention Mask: Layer-wise causal attention mask

- Standard causal mask + full attention within the same **BP / NG** block

Attention Bias: Node-distance-based attention bias

- Penalize attention scores by tree nodes distance



⑤ Results

Machine Translation

Models	IWSLT14 De-En			IWSLT14 En-De			WMT16 Ro-En		
	BLEU	Speedup	Iter	BLEU	Speedup	Iter	BLEU	Speedup	Iter
AT	35.64	1.0×	23.25	27.37	1.0×	23.98	33.33	1.0×	28.05
Vanilla-NAT	21.71	7.48×	1	13.07	8.50×	1	24.53	9.35×	1
TNAD	33.53	1.24×	18.31	23.37	2.19×	11.62	32.61	1.35×	20.18
For Reference (Orthogonal to TNAD)									
DAT	32.99	5.71×	1	23.14	5.68×	1	32.25	6.13×	1

Paraphrase Generation

Models	BLEU	ROUGE-1/2/L/avg	Speedup	Iter
AT	16.3	52.1 / 27.3 / 47.7 / 42.4	1.0×	12.1
Vanilla-NAT	10.6	47.8 / 20.7 / 43.0 / 37.2	4.2×	1
TNAD	12.3	48.0 / 22.4 / 44.5 / 38.3	1.2×	7.9
For Reference (Orthogonal to TNAD)				
DAT	12.1	47.5 / 23.2 / 43.2 / 38.0	2.6×	1

IWSLT14

Models	De-En	En-De
Trivial Tree	20.35	12.25
No Label	30.72	20.79
No ALiBi	32.16	22.26
TNAD	33.53	23.37

- Better performance than NAT (higher BLEU / ROUGE score)
- Faster inference than AT (higher speedup / fewer Iter)

- Ablation verifies importance of tree structure (Trivial Tree), tree-node labels (No Label) and attention bias (No ALiBi)