

Sequence Modeling with Unconstrained Generation Order

D. Emelianenko, E. Voita, P. Serdyukov | NeurIPS 2019



TL;DR

- Sequence model that can **insert** tokens at an arbitrary position;
- Implicitly learn the most convenient decoding order from data;
- Competitive results on Machine Translation, Im2Latex and Image Captioning tasks

left-to-right	mixed	right-to-left
a cat sat on a mat .	a cat sat on a mat .	a cat sat on a mat .
a cat sat on a mat .	a cat sat on a mat .	a cat sat on a mat .
a cat sat on a mat .	a cat sat on a mat .	a cat sat on a mat .
a cat sat on a mat .	a cat sat on a mat .	a cat sat on a mat .
a cat sat on a mat .	a cat sat on a mat .	a cat sat on a mat .
a cat sat on a mat .	a cat sat on a mat .	a cat sat on a mat .
a cat sat on a mat .	a cat sat on a mat .	a cat sat on a mat .

Learning to Insert

Main idea: generate a sequence by inserting elements at arbitrary positions

Hence, we generate a sequence of insertions: $\tau = (\tau_0, \tau_1, \tau_2, \dots, \tau_T)$

Insertion: $\tau_t = (pos_t, token_t)$ insert $token_t$ at position $pos_t \in [0; t]$

Model: predict next insertion into partial output formed by previous inserts

$$p(\tau|X, \theta) = \prod_t p(\tau_t|X, \tilde{Y}(\tau_{0:t-1}), \theta)$$

X – input (e.g. source sequence in MT)

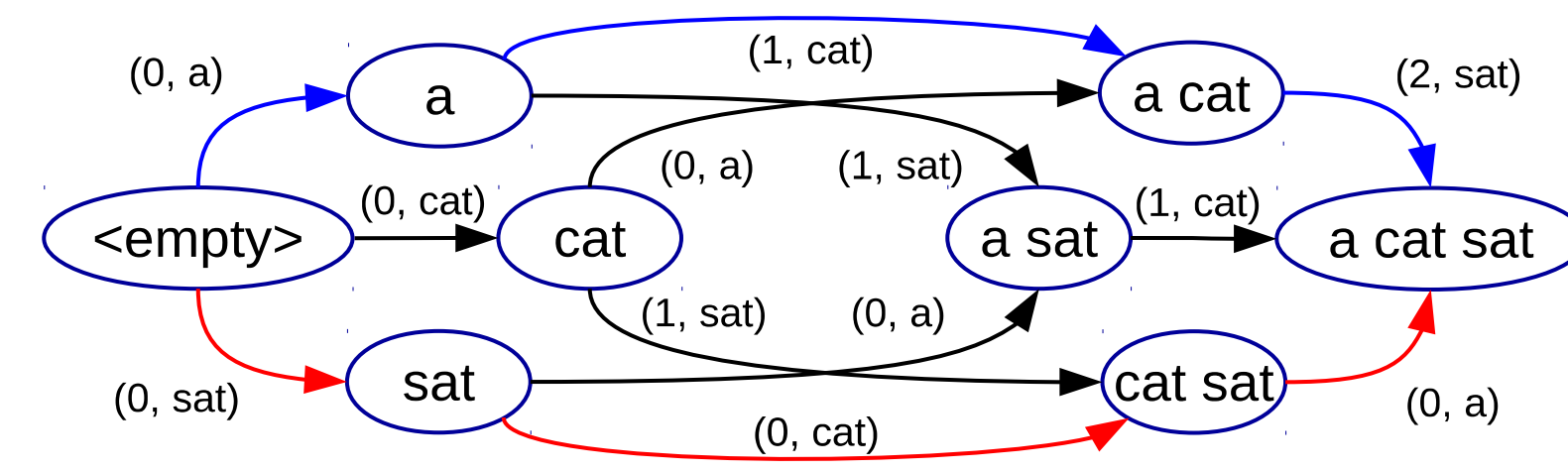
$\tilde{Y}(\tau_{0:t-1})$ – partial output sequence

Log-likelihood: $L = \sum_{\{X, Y\} \in D} \log p(Y|X, \theta) = \sum_{\{X, Y\} \in D} \log \sum_{\tau \in T^*(Y)} \prod_t p(\tau_t|X, \tilde{Y}(\tau_{0:t-1}), \theta)$ (1)

Lower bound: $L \geq \sum_{\{X, Y\} \in D} E_{\tau \sim p(\tau|X, \tau \in T^*(Y), \theta)} \log \prod_t p(\tau_t|X, \tilde{Y}(\tau_{0:t-1}), \theta)$ (2)

Training algorithm:

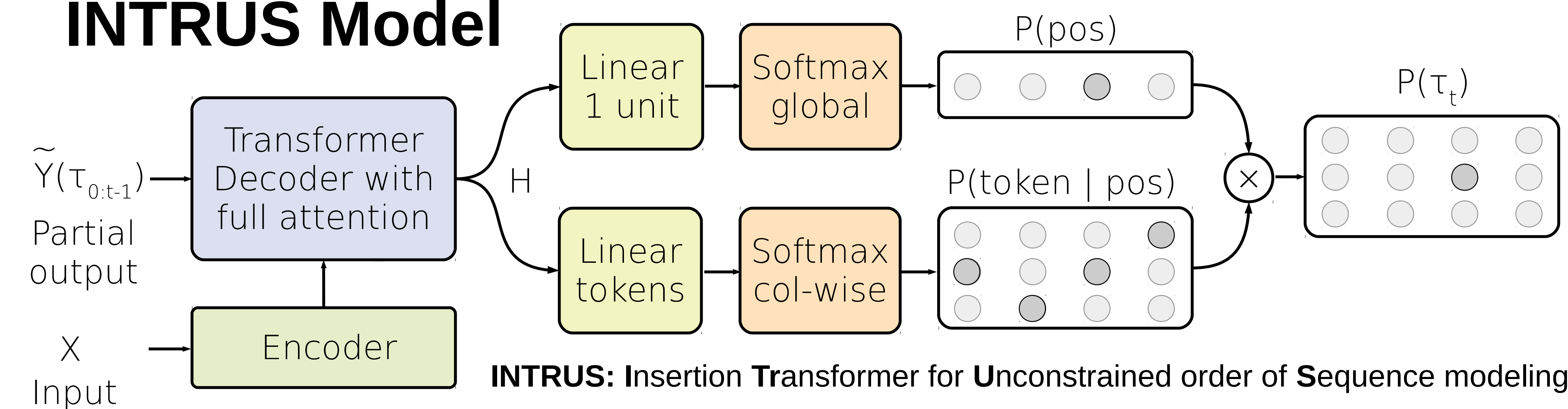
- 1) Sample mini-batch of sequences (X, Y) ;
- 2) Generate a trajectory τ that produces Y proportionally to model probabilities ;
- 3) Maximize lower bound (2) by backprop ;



All trajectories that produce “a cat sat”.
Left-to-right is blue and right-to-left is red.

Pre-training: at step 2, sample generation order at random, ignoring model probabilities

INTRUS Model



INTRUS: Insertion Transformer for Unconstrained order of Sequence modeling

Learned Decoding Orders:

and death is part of everyday life .	the study was conducted among 900 children .	some stories tell about honor and courage .
and death is part of everyday life .	the study was conducted among 900 children .	some stories tell about honor and courage .
and death is part of everyday life .	the study was conducted among 900 children .	some stories tell about honor and courage .
and death is part of everyday life .	the study was conducted among 900 children .	some stories tell about honor and courage .
and death is part of everyday life .	the study was conducted among 900 children .	some stories tell about honor and courage .
and death is part of everyday life .	the study was conducted among 900 children .	some stories tell about honor and courage .
and death is part of everyday life .	the study was conducted among 900 children .	some stories tell about honor and courage .
and death is part of everyday life .	the study was conducted among 900 children .	some stories tell about honor and courage .

a surfer is walking towards the ocean waves .	a seagull that is sitting on a raft in the water .	three people are riding on the back of an elephant .
a surfer is walking towards the ocean waves .	a seagull that is sitting on a raft in the water .	three people are riding on the back of an elephant .
a surfer is walking towards the ocean waves .	a seagull that is sitting on a raft in the water .	three people are riding on the back of an elephant .
a surfer is walking towards the ocean waves .	a seagull that is sitting on a raft in the water .	three people are riding on the back of an elephant .
a surfer is walking towards the ocean waves .	a seagull that is sitting on a raft in the water .	three people are riding on the back of an elephant .
a surfer is walking towards the ocean waves .	a seagull that is sitting on a raft in the water .	three people are riding on the back of an elephant .
a surfer is walking towards the ocean waves .	a seagull that is sitting on a raft in the water .	three people are riding on the back of an elephant .
a surfer is walking towards the ocean waves .	a seagull that is sitting on a raft in the water .	three people are riding on the back of an elephant .
a surfer is walking towards the ocean waves .	a seagull that is sitting on a raft in the water .	three people are riding on the back of an elephant .
a surfer is walking towards the ocean waves .	a seagull that is sitting on a raft in the water .	three people are riding on the back of an elephant .

Experiments

Model	En-Ru	Ru-En	En-Ja	En-Ar	En-De	De-En	Im2Latex	MSCOCO	
	BLEU							BLEU	CIDEr
Left-to-right	31.6	35.3	47.9	12.0	28.04	33.17	89.5	18.0	56.1
Right-to-left	-	-	48.6	11.5	-	-	-	-	-
INTRUS	33.2*	36.4*	50.3*	12.2	28.36*	33.08	90.3*	25.6*	81.0*

Tasks:

- Machine Translation WMT En-Ru, ASPEC En-Ja IWSLT En-De, En-Ar
- ImageToLatex-140K
- MSCOCO Image captioning

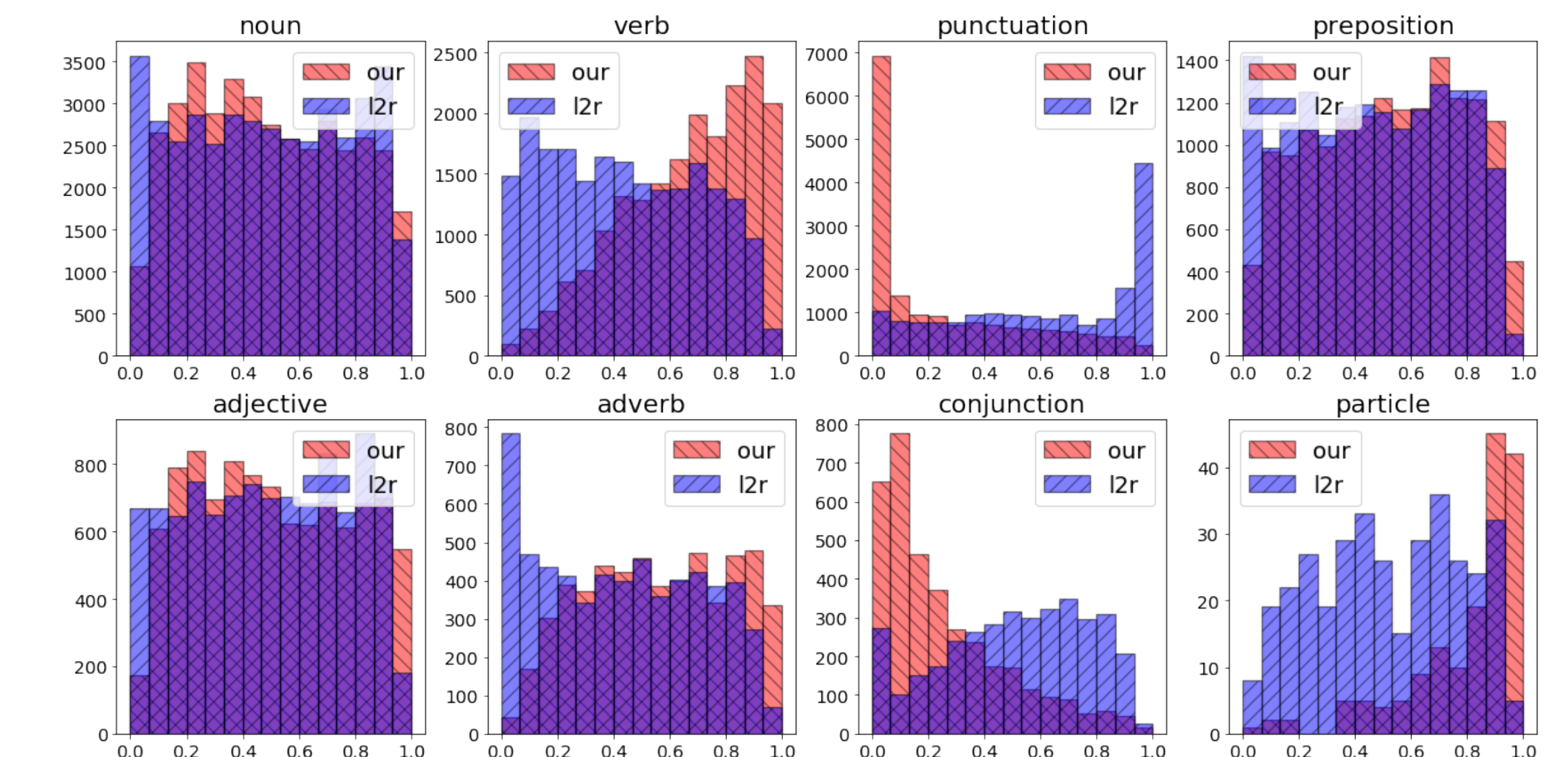
Ablation Analysis:

- Sampling lower bound is better than argmax
- Pre-training greatly affects performance

Training strategy	INTRUS	Argmax	Pretraining left-to-right	No pre-training	Only pretraining		Baseline left-to-right
BLEU	27.5	26.6	26.3	27.1	24.6	25.5	25.8

Ablation analysis on En-Ru translation task

Analysis



Step on which different POS tags appear in sequences generated by INTRUS (our) and Transformer (l2r)

Links:

Code: https://github.com/TIXFeniks/neurips2019_intrus

Yandex Research: <https://research.yandex.com>

