大连理工大学
DALIAN UNIVERSITY OF TECHNOLOGY

LREC-COLING
2024

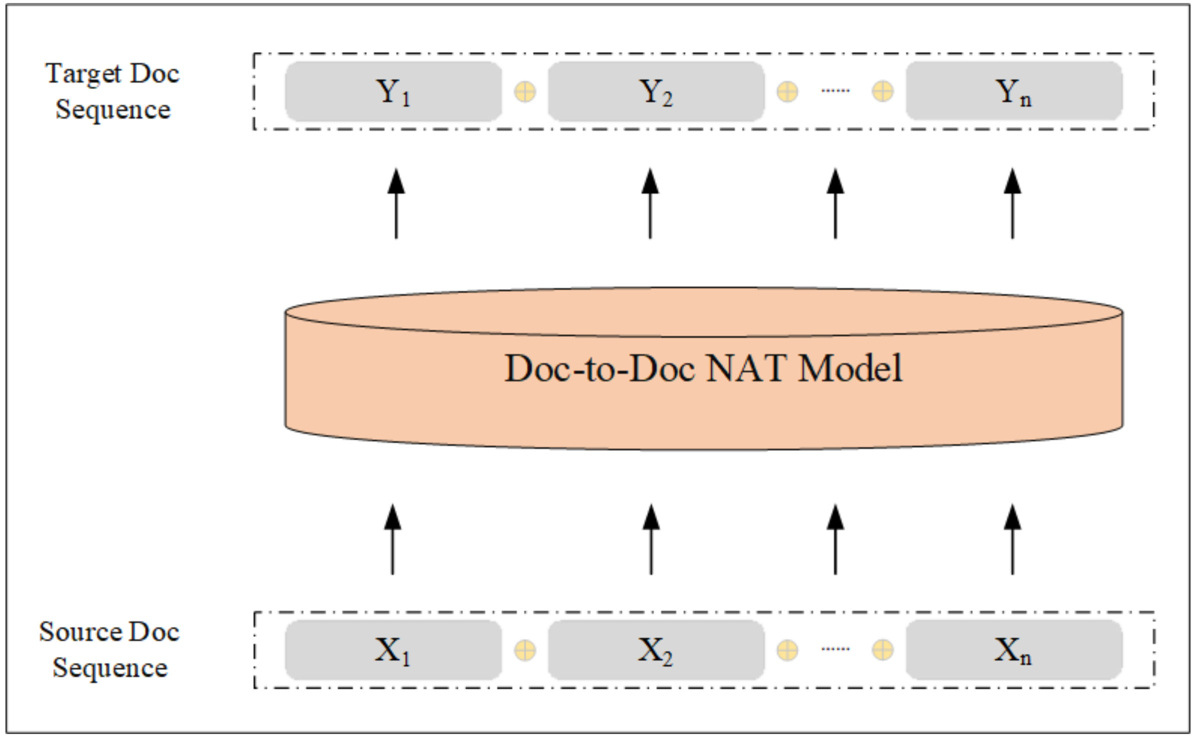
Context-Aware Non-Autoregressive Document-Level Translation with Sentence-Aligned Connectionist Temporal Classification

Hao Yu[1], Kaiyu Huang[2]*, Anqi Zhao[1], Junpeng Liu[1], Degen Huang[1]

[1]School of Computer Science, Dalian University of Technology
[2]School of Computer Science, Beijing Jiaotong University

Task Statement

- **Doc-to-Doc Non-Autoregressive Translation**
  - **Long Sequence Modeling**: Implement non-autoregressive modeling in doc-to-doc scenarios
  - **Model Acceleration**: Impove the Doc-to-Doc model inference speed



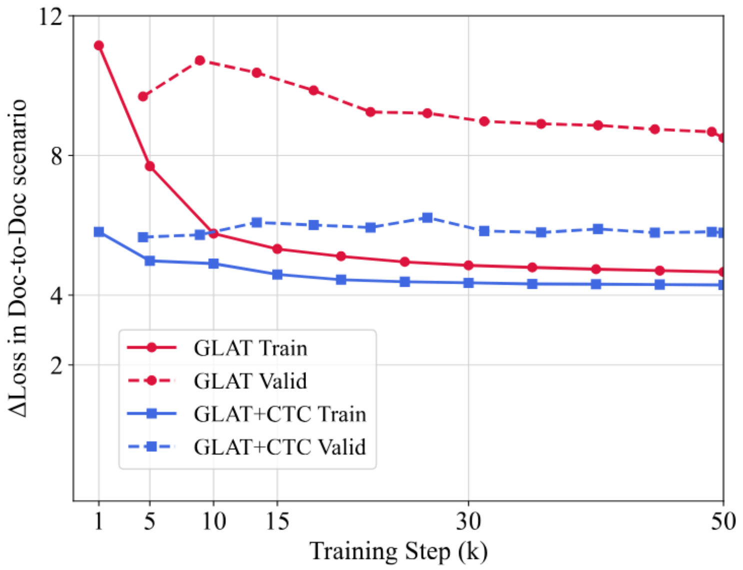
Challenge

- **Doc-to-Doc AT Model**
  - Decoding speed slowly: The AT model needs to be decoded step-by-step.
  - The accumulation of errors: Exposure bias exists in the training process.

- **Doc-to-Doc NAT Model Training Failed**
  - Excessively large search space of decoding path
    - CTC –based method: The search space increases quadratically with the length of the source sequence.
  - Excessively large attention hypothesis space
    - Attention-based method: The hypothesis space increases quadratically with the length of the source sequence.

Motivation

- **Decoding Path Space Pruning**
  - Assume the source document sequence is aligned with the sentences in the target document sequence.
  - Prune the decoding path space in CTC-based model by fixed the position of each sentence start/end token in target document sequence.



- **Attention Hypothesis Space Sparsity**
  - Apply local bias to each sentence of the document in the attention layer.
  - Introducing additional start/end token to encode sentence-level information, and applying context bias to learn document-level contextual information

Method

- **Sentence-Aligned CTC method**

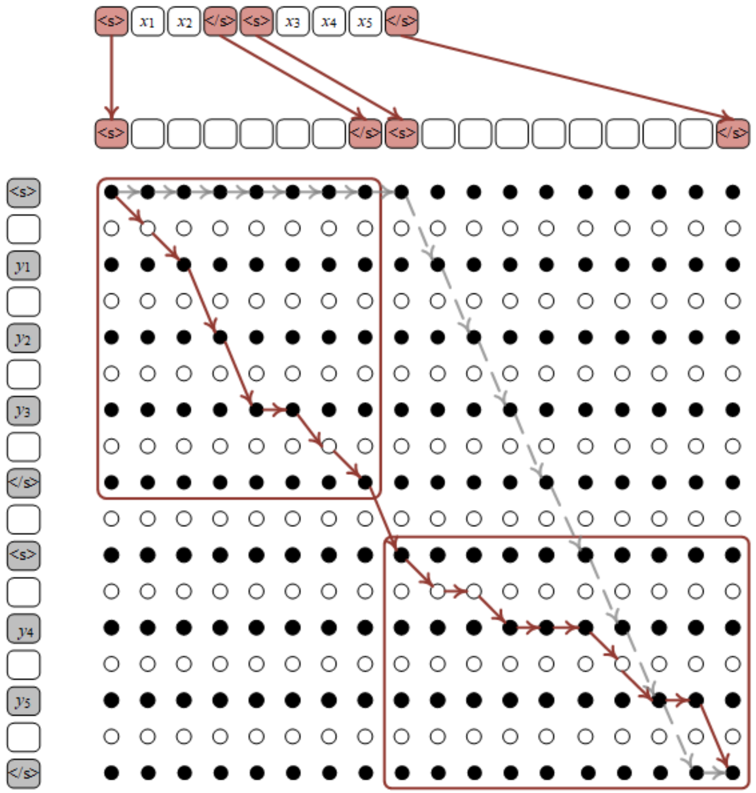
  - **Source/Target document sequence**

$$X = B + X_1 + E + \ldots + B + X_n + E$$
$$Y = B + Y_1 + E + \ldots + B + Y_n + E$$

  - **Calculates the position of B/E token**

$$IndexB = \{I_i = 1 \; if \; X[(i-1)/2] = B \; else \; 0\}\big|_{i=1}^{2|X|}$$
$$IndexE = \{I_i = 1 \; if \; X[i/2] = E \; else \; 0\}\big|_{i=1}^{2|X|}$$

Sentence-Aligned CTC



Method

- **Context-Aware Architecture**

  - **Group tag/Category tag**

$$G_Q = \{g_p = t \; if \; Q_p \in sent_t^Q\}|_{p=1}^{|Q|}$$
$$G_K = \{g_p = t \; if \; K_p \in sent_t^K\}|_{p=1}^{|K|}$$

$$C_Q = \{c_p = 1 \;\; if \; Q_p \in \{B, E\} \; else \; 2\}|_{p=1}^{|Q|}$$
$$C_K = \{c_p = 1 \;\; if \; K_p \in \{B, E\} \; else \; 2\}|_{p=1}^{|K|}$$

  - **Local/Context Attention**

$$LocalMask_{ij} \propto 1 \;\; if \; (G_Q[i] = G_K[j]) \quad else \;\; 0|_{i,j=1,1}^{|Q|,|K|}$$
$$\mathrm{LocalAttention}(Q, K, V) = \mathrm{Softmax}(\frac{QK^T}{\sqrt{d_K}} + LocalMask \cdot \gamma)V$$

$$ContextMask_{ij} \propto 1 \;\; if \; (G_Q[i] = G_K[j] \; or \; C_Q[i] = 1) \quad else \; 0|_{i \in \{1:|Q|\} \; j \in \{1:|K|\}}$$
$$\mathrm{ContextAttention}(Q, K, V) = \mathrm{Softmax}(\frac{QK^T}{\sqrt{d_K}} + ContextMask \cdot \gamma)V$$

Method

- **Context-Aware Architecture**

  - **Hierarchical Attention Structure**

    The top two layers of the model apply context attention

    Other layers at the bottom of the model apply local attention

  - **Sentence-Aligned Softcopy**

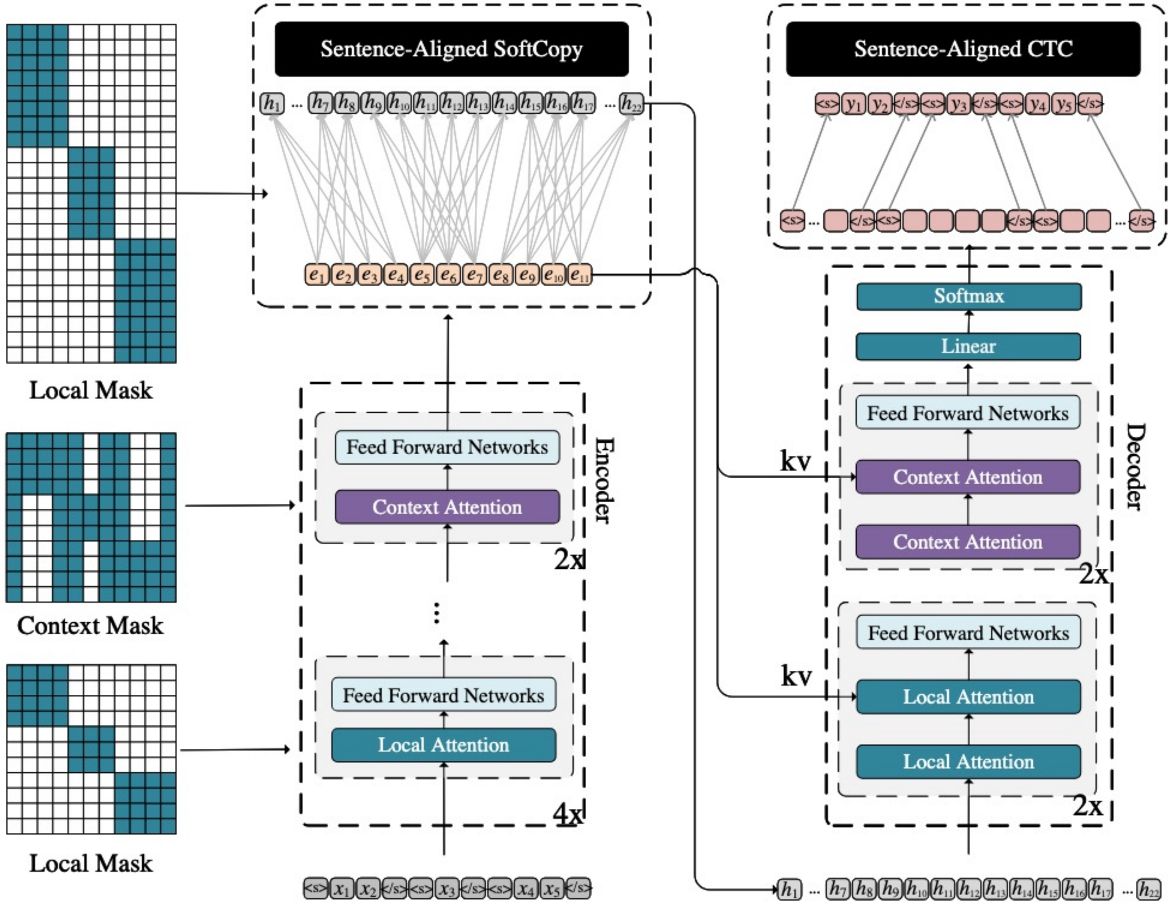
$$A = \begin{pmatrix} \alpha_{1,1} & \cdots & \alpha_{1,s} \\ \vdots & \ddots & \vdots \\ \alpha_{t,1} & \cdots & \alpha_{t,s} \end{pmatrix}$$

$$\alpha_{ij} \propto \exp[-(i - j \cdot \frac{s}{t})^2]$$

$$H = \mathrm{Softmax}(A + LocalMask \cdot \gamma)E$$

Method

- **Overall framework**



Experiment

- **Results on En-De benchmark datasets**

Method	Data	TED		News		Europarl	
		s-BLEU	d-BLEU	s-BLEU	d-BLEU	s-BLEU	d-BLEU
autoregressive translation							
SENTNMT (2017)	raw	23.10	-	22.40	-	29.40	-
HAN (2018)	raw	24.58	-	25.03	-	28.60	-
SAN (2019)	raw	24.42	-	24.84	-	29.75	-
Hybrid Context (2020)	raw	25.10	-	24.91	-	30.40	-
Flat-Transformer (2020)	raw	24.87	-	23.55	-	30.09	-
G-Trans (randinit) (2021)	raw	23.53	25.84	23.55	25.23	32.18	33.87
G-Trans (finetune) (2021)	raw	25.12	27.17	25.52	27.11	32.39	34.08
Disco2NMT (2022)	raw	24.60	-	23.25	-	29.36	-
SENTNMT (2017) †	raw	25.00	27.32	25.26	26.78	31.50	33.19
G-Trans (randinit) (2021) †	raw	23.84	26.14	23.44	25.00	31.95	33.65
G-Trans (finetune) (2021) †	raw	24.98	27.17	25.50	27.09	32.54	34.22
non-autoregressive translation							
GLAT (2021)†	sent-KD	-	0.00	-	0.00	-	0.94
GLAT+CTC (2021)†	sent-KD	-	8.05	-	0.00	-	0.00
GLAT-Latent (2022)†	sent-KD	-	0.75	-	0.93	-	16.77
CASA	sent-KD	**24.24**	**26.45**	23.25	24.72	29.50	31.07
CASA-Latent	sent-KD	24.04	26.28	**23.78**	**25.92**	**29.75**	**31.33**
CASA	doc-KD(finetune)	**24.16**	**26.24**	**23.47**	**25.00**	29.49	31.12
CASA-Latent	doc-KD(finetune)	23.88	26.00	23.09	24.68	**29.85**	**31.44**
CASA	raw	22.44	24.61	**19.16**	**20.55**	25.47	27.06
CASA-Latent	raw	**22.50**	**24.78**	18.55	19.94	**26.31**	**27.85**

- Our method implements non-autoregressive modeling in document-to-document translation scenarios;
- Our method achieves competitive performance with the document-level AT baseline on TED, and News datasets

Experiment

- **Results on Model Acceleration**

	One Instance				Fully GPU Memory			
	TED	News	Europarl	Avg.	TED	News	Europarl	Avg.
	autoregressive translation on raw data							
SENTNMT (2017) †	1.37x	1.36x	1.34x	1.36x	8.03x	8.40x	7.16x	7.86x
G-Trans(randinit) (2021) †	1.00x	1.00x	1.00x	1.00x	1.00x	1.00x	1.00x	1.00x
Shadow(8+4)	1.27x	1.24x	1.23x	1.25x	1.09x	1.12x	1.14x	1.12x
Shadow(10+2)	1.91x	1.91x	1.87x	1.90x	1.31x	1.39x	1.33x	1.34x
2to2	0.97x	0.90x	0.93x	0.93x	3.15x	3.19x	2.78x	3.04x
	non-autoregressive translation on sentence-level KD data							
CASA-Latent	30.27x	29.90x	29.74x	29.97x	14.19x	20.85x	15.01x	16.68x
CASA	**46.67x**	**44.21x**	**47.15x**	**46.01x**	**25.14x**	**32.33x**	**23.00x**	**26.82x**

- Our method can significantly accelerate model decoding in document-to-document translation scenarios;
- Compared with deep encoder + shallow decoder and sentence-level parallel context-aware method, the acceleration effect of our method is more significant.

Experiment

- **Results on discourse phenomena**

Method	Data	Deixis	E_vp	E_infl	L_coh
	autoregressive translation				
SENTNMT (2017) †	raw	50.00	26.20	51.60	45.87
CADec (2019b)	raw	81.60	80.00	72.20	58.10
DocRepair (2019a)	raw	91.80	75.20	86.40	80.60
LSTM-Trans (2020)	raw	90.50	81.00	80.60	73.90
D-LM(PMI) (2021)	raw	96.80	90.60	75.80	97.80
G-Trans (randinit) (2021) †	raw	85.36	76.00	76.00	58.00
G-Trans (finetune) (2021) †	raw	74.48	25.20	50.80	45.87
	non-autoregressive translation				
CASA	raw	50.00	33.80	55.20	45.87
CASA-Latent	raw	50.00	38.40	55.00	45.87
CASA	sent-KD	50.00	19.40	50.40	45.87
CASA-Latent	sent-KD	50.00	21.00	51.00	45.87
CASA	doc-KD(randinit)	50.00	**51.80**	59.40	45.87
CASA-Latent	doc-KD(randinit)	50.00	49.60	**60.00**	45.87
CASA	doc-KD(finetune)	**50.60**	36.20	47.80	**46.13**
CASA-Latent	doc-KD(finetune)	50.48	32.80	47.60	45.87

- We evaluate the discourse modeling capabilities of document-level non-autoregressive methods and find that better results are achieved on document-level distillation datasets
- Compared with the AT baseline system, the performance gap of the discourse phenomenon is still relatively large.

Thanks for Your Listening

E-mail Address: yuhao_dlut@mail.dlut.edu.cn