switch-GLAT: Multilingual Parallel Machine Translation Via Code-Switch Decoder

ICLR 2022

Zhenqiao Song, Hao Zhou, Lihua Qian, Jingjing Xu, Shanbo Cheng, Mingxuan Wang, Lei Li

BYTEDANCE AI LAB

Reporter: Xiachong Feng

Authors

Zhenqiao Song*
Fudan University
Bytedance AI Lab

Hao Zhou*
Research Associate Professor
Insititute for AI Industry Research,
Tsinghua University

Lihua Qian
ByteDance AI Lab

Jingjing Xu
post-doc UCSB

Shanbo Cheng

Shanbo Cheng
research scientist/manager working at ByteDance AI Lab

Lei Li
Assistant Professor UCSB

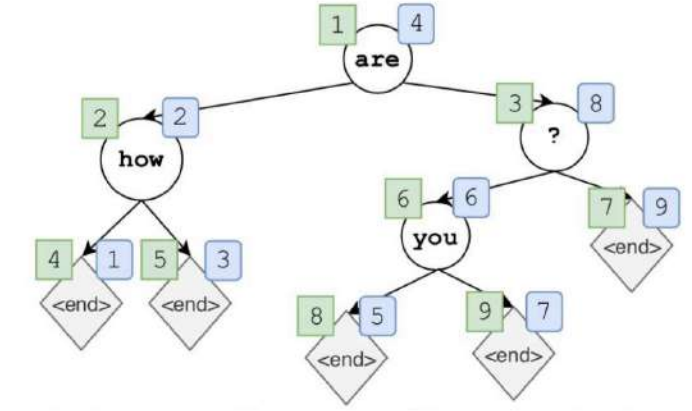
Non-Autoregressive Decoding

Conclusion

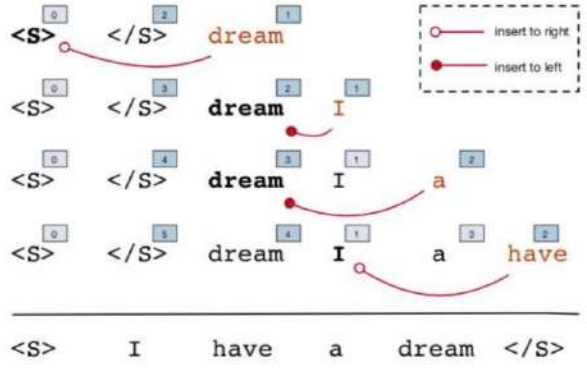
Parallel generation:

t	Canvas	Insertions
0	[]	(ate, 0)
1	[ate]	(friends, 0), (together, 1)
2	[friends, ate, together]	(three, 0), (lunch, 2)
3	[three, friends, ate, lunch, together]	(⟨EOS⟩, 5)

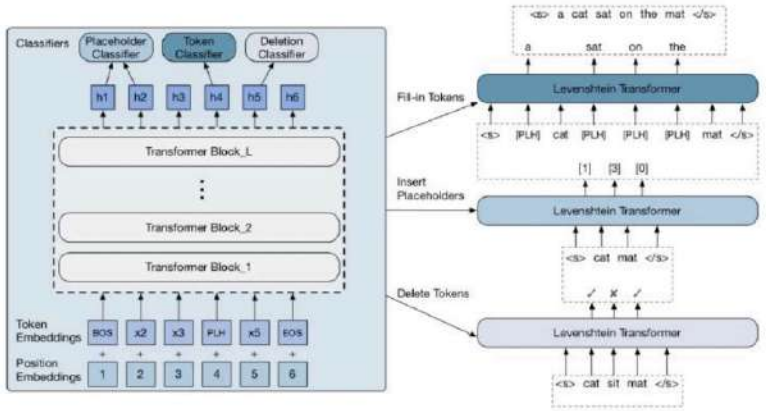
Insertion transformer



Non-Monotonic



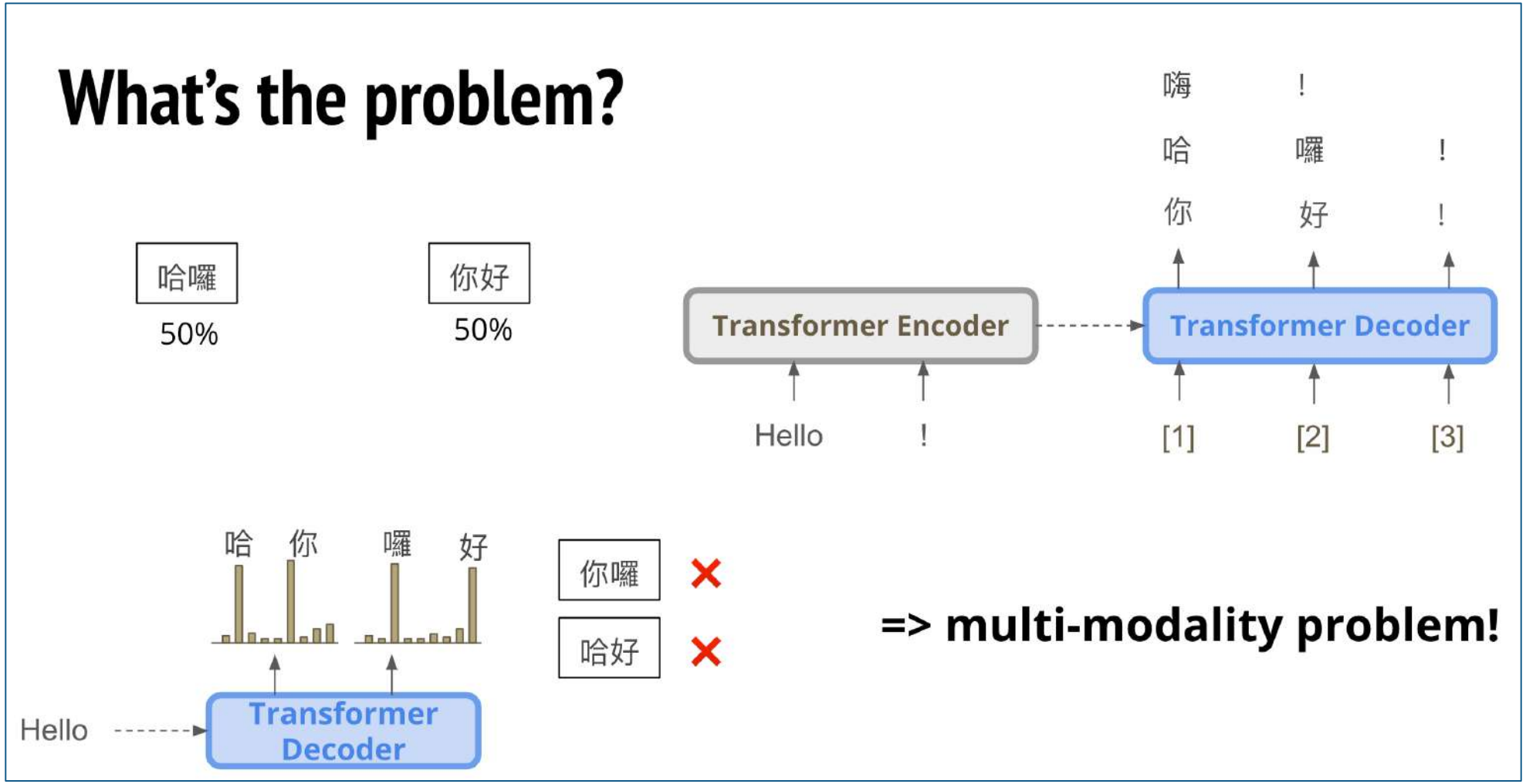
InDIGO



Levenshtein

http://xcfeng.net/res/presentation/Non-Autoregressive%20Decoding.pdf

Non-Autoregressive Decoding



https://docs.google.com/presentation/d/1lnXSxPL3hZQ_OHyV_Lksz3UcOmPyiDqC1eg29Q5IAr0/edit#slide=id.g84c46a9f82_0_0

GLAT

- latent-GLAT: Glancing at Latent Variables for Parallel Text Generation

  *Yu Bao, **Hao Zhou**, Shujian Huang, Dongqi Wang, Lihua Qian, Xinyu Dai, Jiajun Chen, and Lei Li

  In *the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, May, 2022.

  Bib paper code abstract

- switch-GLAT: Multilingual Parallel Machine Translation via Code-switch Decoder

  *Zhenqiao Song, **Hao Zhou**, Lihua Qian, Jingjing Xu, Shanbo Cheng, Mingxuan Wang, and Lei Li

  In *International Conference on Learning Representations (ICLR)*, April, 2022.

  Bib paper abstract

- Non-Autoregressive Translation with Layer-Wise Prediction and Deep Supervision

  *Chenyang Huang, **Hao Zhou**, Osmar Zaiane, Lili Mou, and Lei Li

  In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, February, 2022.

  Bib paper code abstract

- Duplex Sequence-to-Sequence Learning for Reversible Machine Translation

  *Zaixiang Zheng, **Hao Zhou**, Shujian Huang, Jiajun Chen, Jingjing Xu, and Lei Li

  In *the 35th Conference on Neural Information Processing Systems (NeurIPS)*, December, 2021.

  Bib paper abstract

- The Volctrans GLAT System: Non-autoregressive Translation Meets WMT21

  *Lihua Qian, Yi Zhou, Zaixiang Zheng, Yaoming Zhu, Zehui Lin, Jiangtao Feng, Shanbo Cheng, Lei Li, Mingxuan Wang, and **Hao Zhou**

  In *Sixth Conference on Machine Translation (WMT21)*, November, 2021.

  Bib paper code abstract

- Glancing Transformer for Non-Autoregressive Neural Machine Translation

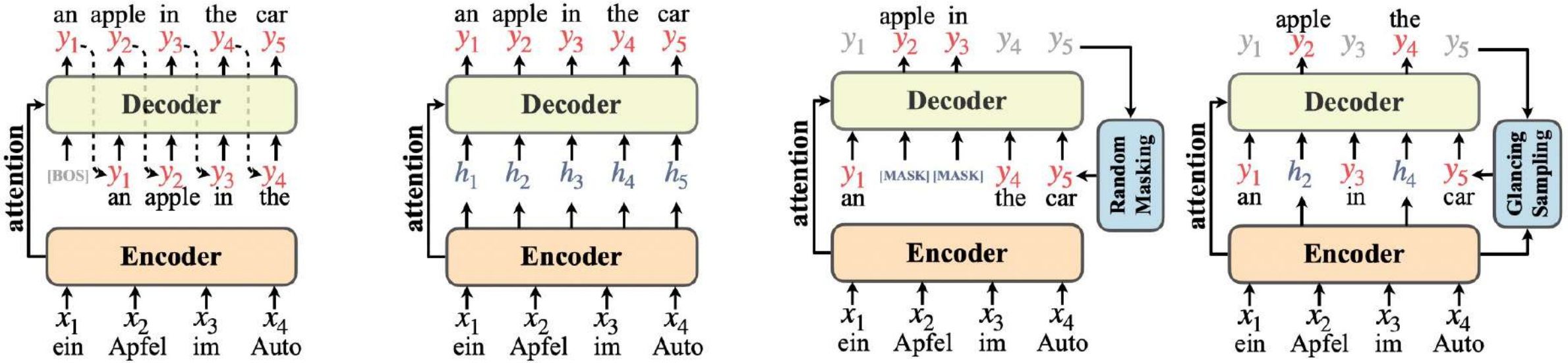
  *Lihua Qian, **Hao Zhou**, Yu Bao, Mingxuan Wang, Lin Qiu, Weinan Zhang, Yong Yu, and Lei Li

  In *the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, July, 2021.

  Bib paper code abstract

字节跳动火山翻译团队的并行生成翻译系统 GLAT 拿下了 WMT2021 De-En/En-De 的双料冠军。

Glancing Transformer



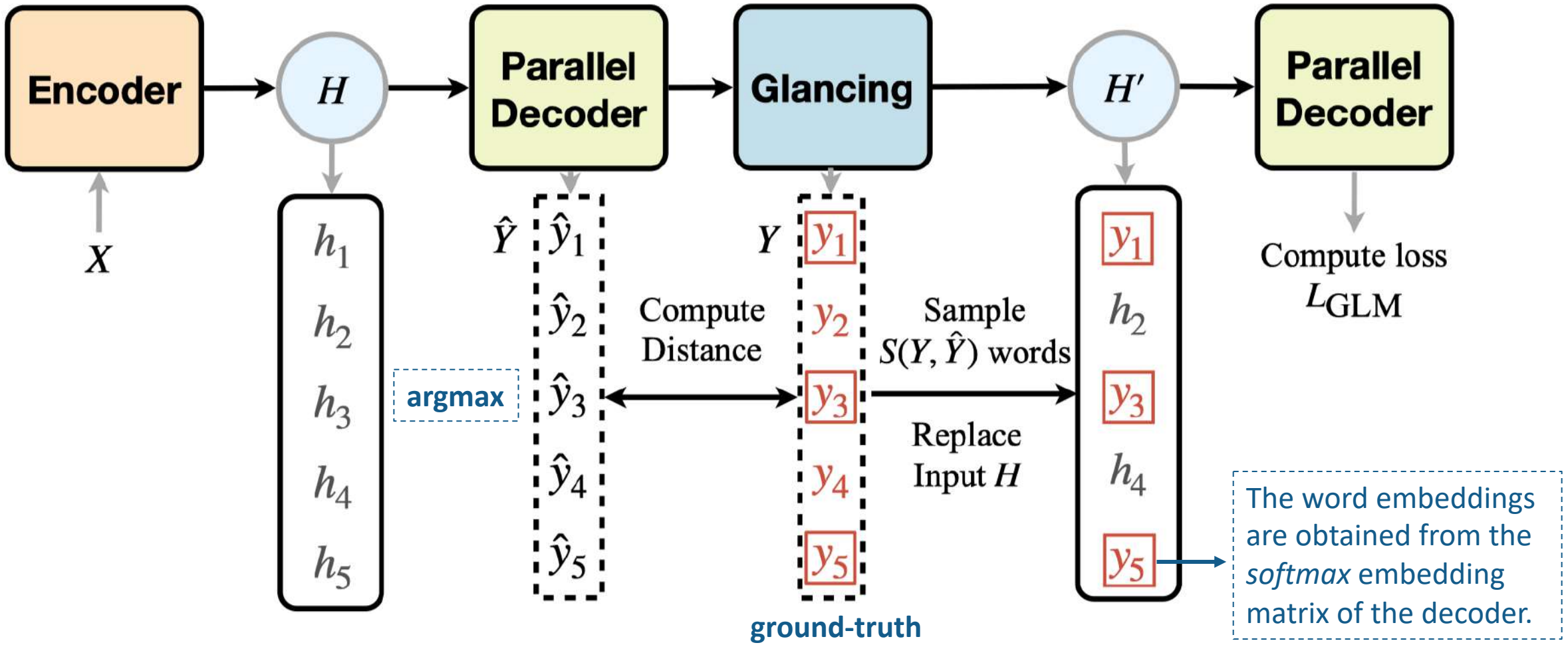
$$\mathcal{L}_{\text{AT}} = \log P(Y|X;\theta) = \sum_{t=1}^{T} \log p(y_t|y_{<t}, X;\theta)$$

$$\mathcal{L}_{\text{NAT}} = \sum_{t=1}^{T} \log P(y_t|X;\theta)$$

$$\mathcal{L}_{\text{MLM}} = \sum_{y_t \in \mathbb{RM}(Y)} \log p\Big(y_t|\Phi\big(Y, \mathbb{RM}(Y)\big), X;\theta\Big)$$

$$\mathcal{L}_{\text{GLM}} = \sum_{y_t \in \overline{\mathbb{GS}(Y,\hat{Y})}} \log p(y_t|\mathbb{GS}(Y,\hat{Y}), X;\theta)$$

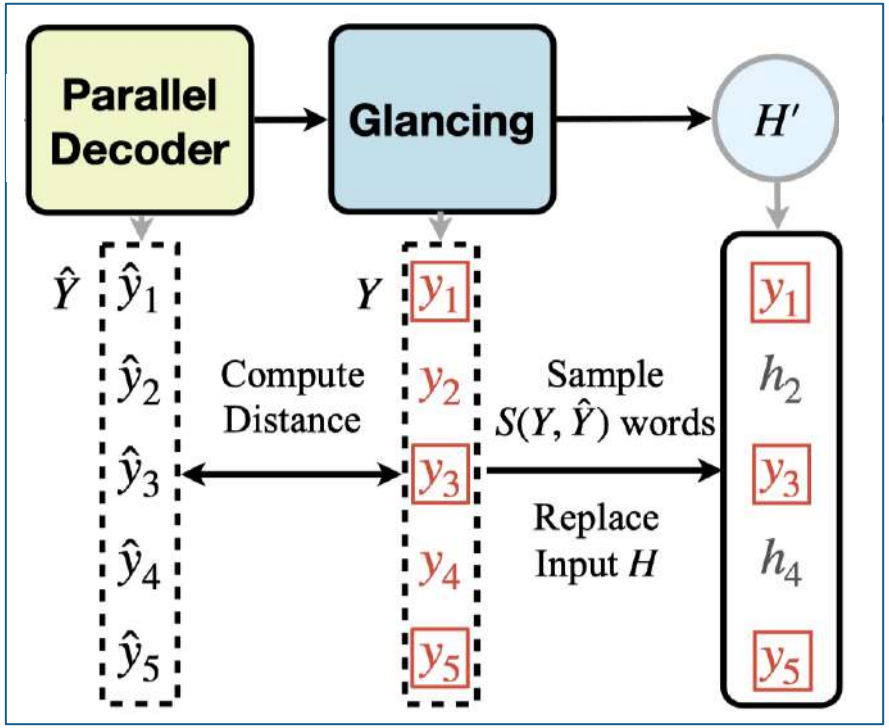
Glancing Transformer: Training



Note that only the second decoding will update the model parameters

Glancing Sampling Strategy

- Selected tokens provide **“correct”** information from the ground-truth target
- Adaptive sampling strategy guides the model to first learn the generation of **fragments** and then gradually turn to the whole **sentences**.
- As the model gets better progressively, the **sampling strategy will sample fewer** words to enable the model to learn the parallel generation of the whole sentence.



Glancing Sampling Strategy

1. Deciding a sampling number S
2. **Randomly** selecting S words from the **reference**.
    1. The random reference word selection is simple and yields good performance empirically.

$$\mathbb{GS}(Y, \hat{Y}) = \text{Random}(Y, S(Y, \hat{Y}))$$

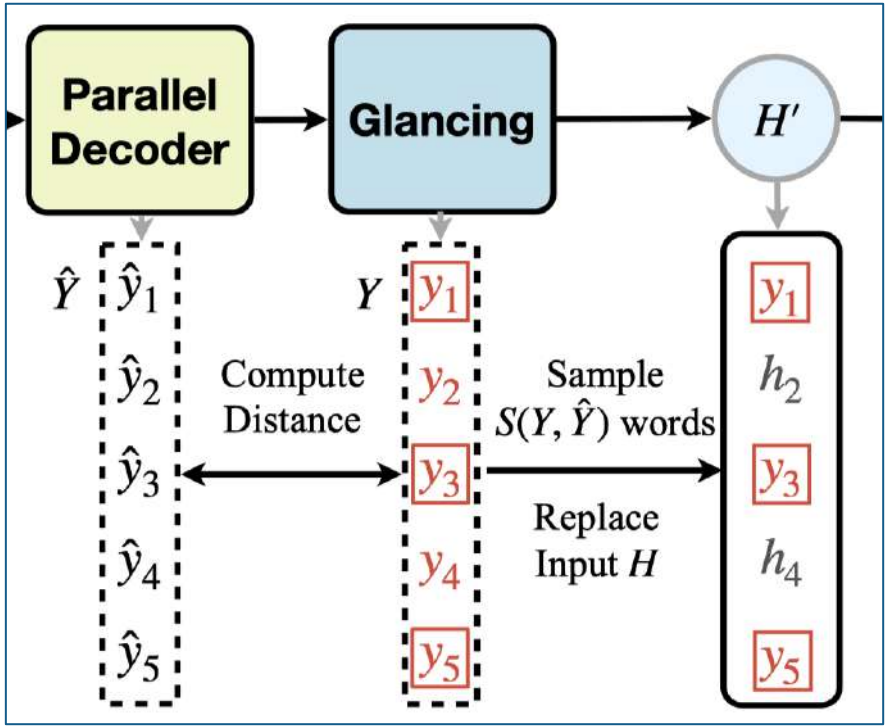
$$S(Y, \hat{Y}) = \lambda \cdot d(Y, \hat{Y})$$

λ: sampling ratio

d:metric for measuring the differences between Y and $\hat{Y}$

Hamming distance

$$d(Y, \hat{Y}) = \sum_{t=1}^{T} (y_t \neq \hat{y}_t)$$



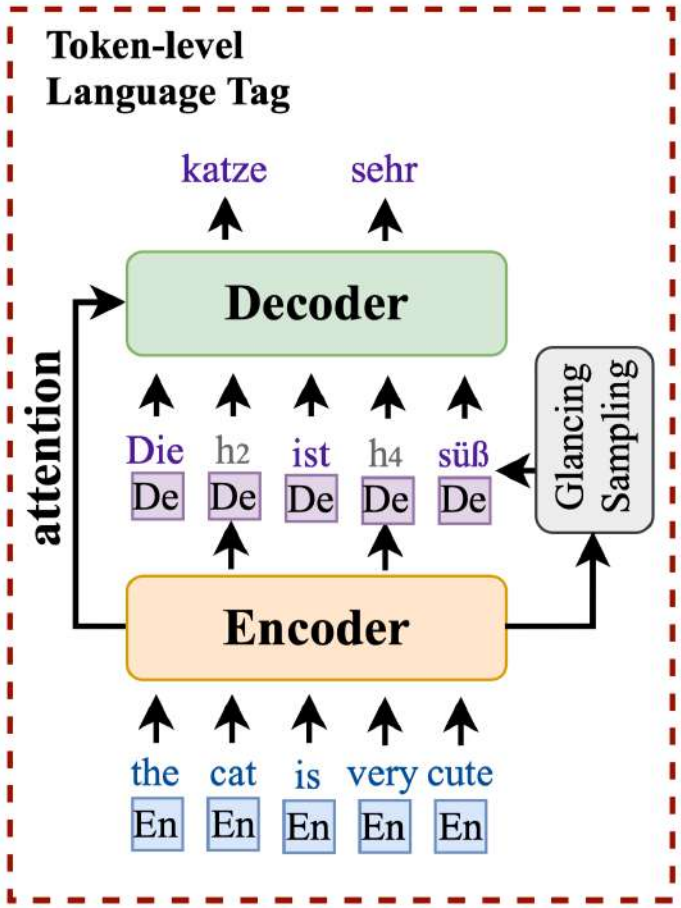
$\boldsymbol{d(Y, \hat{Y})}$: the sampling number can be decided adaptively considering the current trained model's prediction capability.

Glancing Transformer: Inference

- We need to decide the **output lengths** before decoding.

- An additional [LENGTH] token is added to the source input, and the encoder output for the [LENGTH] token is used to predict the length.

switch-GLAT: Code-Switch Decoder



$$L_{\text{multi}} = \sum_{\mathbb{D}^l \in \mathbb{D}} \sum_{(X_j^l, Y_j^l) \in \mathbb{D}^l} \{L_{\text{tag}}(Y_j^l | X_j^l; \theta_M) + L_{\text{len}}^l(j)\}$$

$$L_{\text{len}}^l(j) = -P(L_j^l) \log \hat{P}(L_j^l | [F_e(X_j^l; \theta_M); E_{\text{src}}; E_{\text{tgt}}]; \theta_M)$$

$$L_{\text{tag}}(Y_j^l | X_j^l; \theta_M) = - \sum_{y_t \in \overline{\mathbb{GS}(Y_j^l, \hat{Y}_j^l)}} \log P(y_t | \mathbb{GS}(Y_j^l, \hat{Y}_j^l), X_j^l, \text{src}, \text{tgt}; \theta_M)$$

$$\hat{Y}_j^l = F_d(\tilde{H}_d^0, F_e(X_j^l, \text{src}; \theta), \text{tgt}; \theta_M)$$

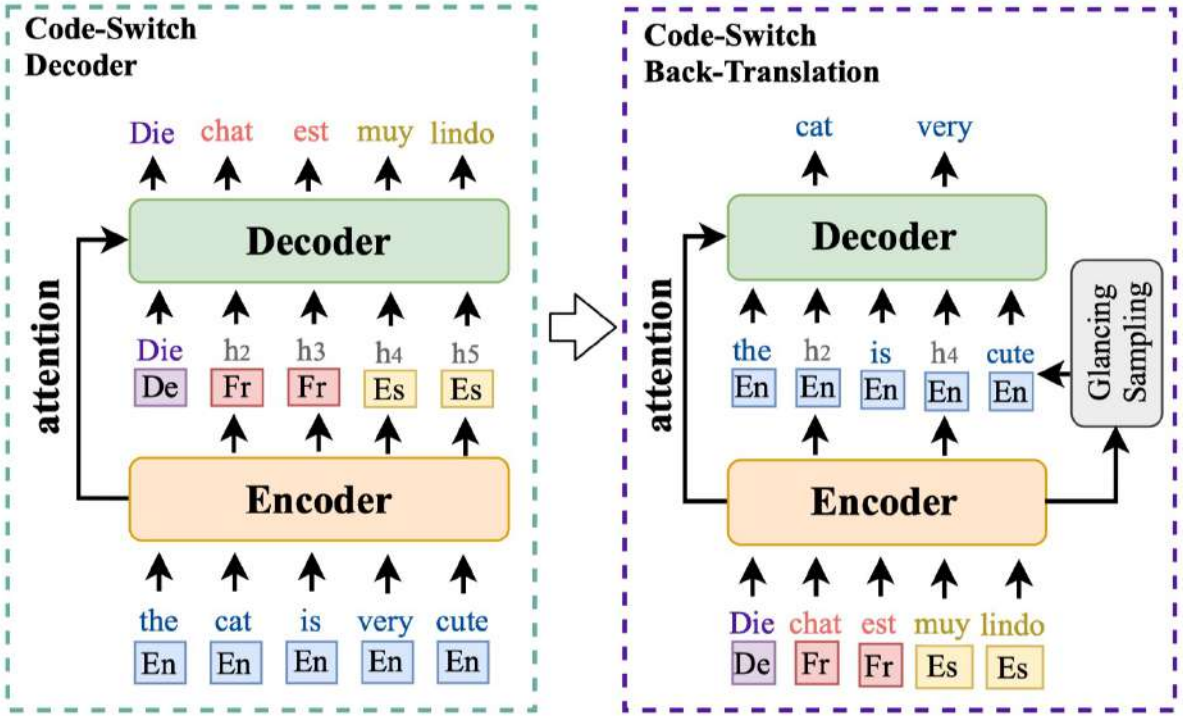
$$\tilde{f}_i^0 = f_i^0 + E_{\text{src}}; \tilde{f}_i^K = f_i^K + E_{\text{src}}$$

$$\tilde{h}_j^0 = h_j^0 + E_{\text{tgt}}; \tilde{h}_j^K = h_j^K + E_{\text{tgt}}$$

Switch-GLAT can translate between different languages using the indicative language tags, of which the decoder is called code-switch decoder.

switch-GLAT: Code-Switch Back-Translation

Through this process, abundant code-switched sentences can be generated, which helps to learn better-aligned cross-lingual representations.



code-switch back-translation corpus

$$\mathbb{D}_C = \{(\hat{Y}_i, X_i)\}_{i=1}^{S}$$

$$L_{\text{csbt}} = \sum_{(\hat{Y}_i, X_i) \in \mathbb{D}_C} L_{\text{tag}}(X_i | \hat{Y}_i; \theta_M)$$

Loss

$$L = L_{\text{multi}} + \lambda \cdot L_{\text{csbt}}$$

$$\lambda(t) = f(t) + \lambda(0), \quad f(t) = \begin{cases} 0, & t < E, \\ 1, & t \geq E \end{cases}$$

Experiments

(1) **WMT-EDF:** We collect 4 language pairs from WMT-14 English (En) ↔ German (De) and English (En) ↔ French (Fr). All three languages belong to Indo-European language family and are relatively **close** on linguistics.

(2) **WMT-EFZ:** We also collect 4 language pairs from WMT-14 English (En) ↔ French (Fr) and WMT-17 English (En) ↔ Chinese (Zh), which are **distant** languages on linguistics and their relationships are more difficult to learn.

(3) **WMT-many:** We also gather 10 language pairs from WMT-14 English (En) ↔ German (De), English (En) ↔ French (Fr), WMT-16 English (En) ↔ Russian (Ru), English (En) ↔ Romanian (Ro) and WMT-17 English (En) ↔ Chinese (Zh) to test switch-GLAT on more **diverse** language pairs.

Result

Models	WMT-EDF						WMT-EFZ				
	En-De	De-En	En-Fr	Fr-En	Avg	Speed	En-Fr	Fr-En	En-Zh	Zh-En	Avg
Bilingual models											
Transformer	27.77	31.55	38.80	37.35	**33.86**	1.3×	38.80	37.35	23.60	24.05	**30.95**
GLAT	26.09	30.53	38.62	34.44	32.42	6.3×	38.62	34.44	21.05	22.89	29.25
Multilingual models											
M-Transformer	25.84	31.57	38.52	36.03	32.99	1.0×	38.06	35.07	20.76	22.19	29.02
CLSR	23.51	31.29	38.58	34.67	32.01	0.9×	37.39	35.62	20.23	21.13	28.59
Adapter	22.19	29.56	40.72	35.88	32.08	0.9×	40.13	35.02	19.87	21.29	29.08
MNAT	13.82	21.89	24.31	25.28	21.32	5.9×	19.46	20.18	7.89	7.35	13.72
switch-GLAT	25.27	31.29	40.81	36.00	**33.34**	6.2×	40.54	36.48	19.47	22.55	**29.76**
– w/o glancing	17.76	23.28	25.91	29.15	24.02	6.0×	21.28	21.97	8.62	8.02	14.97
– w/o CSBT	24.29	29.03	36.45	34.09	30.97	6.2×	35.17	33.65	18.32	20.37	26.87
– w/o CCS	24.72	29.51	37.32	34.17	31.43	6.1×	35.96	33.61	18.45	20.83	27.21

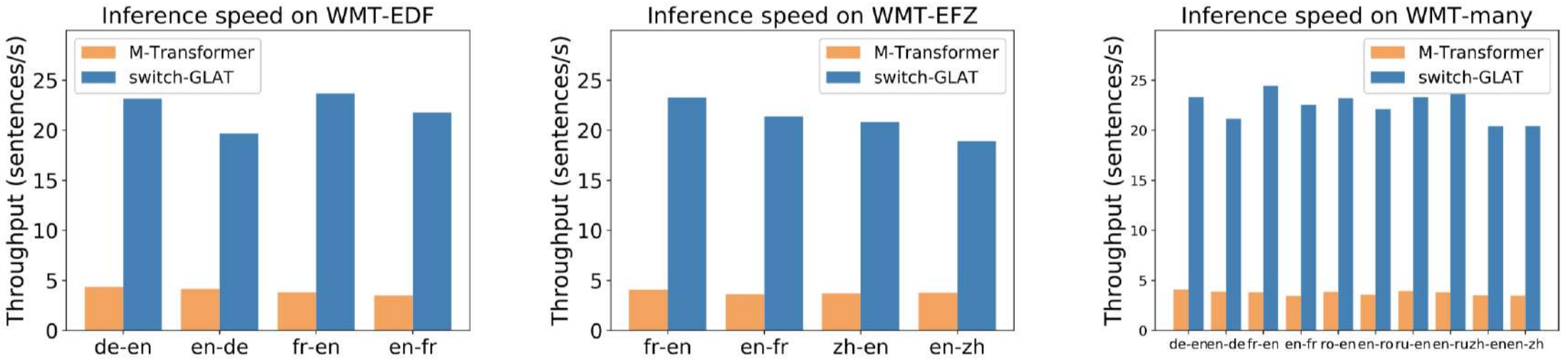
Table 1: Translation performance (BLEU) on WMT-EDF/EFZ[1]. Avg means the average BLEU score.

Result

Models	WMT-many										
	En-De	De-En	En-Fr	Fr-En	En-Ro	Ro-En	En-Ru	Ru-En	En-Zh	Zh-En	Avg
Bilingual models											
Transformer	27.77	31.55	38.80	37.35	33.01	33.59	28.22	29.89	23.60	24.05	**30.78**
GLAT	26.09	30.53	38.62	34.44	31.83	32.59	25.42	28.13	21.05	22.89	29.20
Multilingual models											
M-Transformer	23.14	29.38	35.19	34.07	34.36	35.26	24.63	29.14	16.63	20.17	28.19
CLSR	21.57	30.01	34.07	32.72	30.61	35.29	19.13	30.27	16.28	20.19	27.01
Adapter	23.26	29.87	38.79	40.03	30.42	32.01	22.87	26.38	18.29	20.31	28.22
MNAT	8.33	13.86	12.07	19.21	12.89	21.36	7.66	15.43	4.85	6.01	12.16
switch-GLAT	24.18	30.49	39.47	36.30	31.93	32.40	24.16	28.33	16.25	21.23	**28.47**
– w/o glancing	6.83	16.29	11.08	20.33	13.57	23.34	7.57	18.25	7.31	6.74	13.13
– w/o CSBT	22.48	27.91	31.62	33.55	29.89	32.41	22.43	26.48	15.33	18.32	26.04
– w/o CCS	22.25	28.13	34.01	33.62	30.16	33.07	21.68	26.53	14.92	19.07	26.34

Table 2: Translation performance (BLEU) on WMT-many[2]. Avg means the average score.

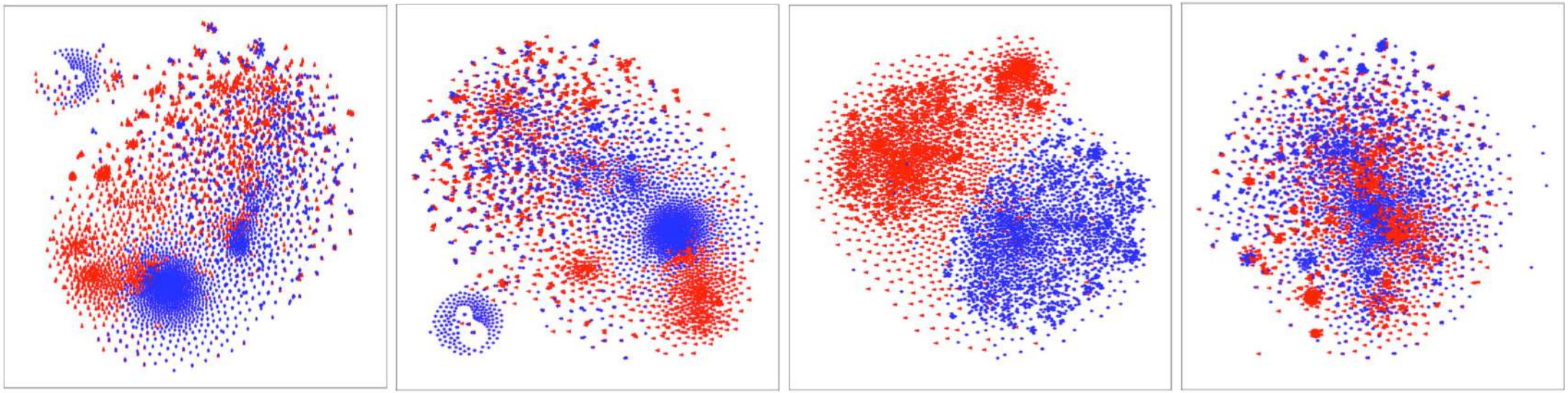
Result



(a) WMT-EDF (b) WMT-EFZ (c) WMT-many

Figure 3: Inference speed evaluated on throughputs.

Result



(a) Transformer (b) GLAT (c) M-Transformer (d) *switch*-GLAT

Figure 4: Representations learned by (a) Transformer (Vaswani et al., 2017), (b) GLAT (Qian et al., 2020), (c) M-Transformer and (d) *switch*-GLAT, projected to 2D.

Gold word pairs from Open Multilingual WordNet (OMW) datasets
English words are displayed in blue color and German words in red.

Result

- Cross-lingual word induction performance to see how well the similar words from different languages are close to each other in the learned vector space
- Word Induction: golden word pairs are extracted from the OMW datasets
- Sentence Retrieval: Tatoeba dataset, Cosine similarity is leveraged to search the nearest neighbour

Models	Word Induction					Sentence Retrieval				
	En-De	De-En	En-Fr	Fr-En	Avg	En-De	De-En	En-Fr	Fr-En	Avg
M-Transformer	30.2	31.7	32.7	38.9	33.3	23.6	22.3	24.5	23.8	23.5
CLSR	24.9	25.8	37.5	36.2	31.1	19.6	18.3	26.8	27.9	23.2
Adapter	32.8	34.5	40.1	42.8	37.5	34.6	28.7	26.9	35.2	31.4
MNAT	24.1	23.3	32.4	33.6	28.3	13.8	17.6	18.2	16.3	16.5
switch-GLAT	33.8	36.2	41.9	46.3	**39.5**	34.8	36.3	36.9	37.2	**36.3**
– w/o glancing	24.3	24.6	32.9	33.7	28.9	18.2	22.8	19.2	20.3	20.1
– w/o CSBT	28.2	30.1	31.9	37.8	32.0	17.6	21.4	23.3	25.3	21.8
– w/o CCS	30.5	29.7	33.6	35.8	32.4	18.8	22.9	22.7	21.6	21.5

Table 3: Results of quality analyses. Avg means the average accuracy.

Conclusion

- switch-GLAT, a non-autoregressive multilingual neural machine translation model
- The multilingual translation performance and cross-lingual representations can both be improved
- Parallel decoder enables a highly efficient inference

Thanks~