Thirty-Third AAAI Conference on Artificial Intelligence,
January 27 – February 1, 2019,
Honolulu, Hawaii, USA.

Non-Autoregressive Neural Machine Translation with Enhanced Decoder Input

Junliang Guo[†*], **Xu Tan**[‡], **Di He**[§], **Tao Qin**[‡], **Linli Xu**[†] and **Tie-Yan Liu**[‡]
[†]Anhui Province Key Laboratory of Big Data Analysis and Application,
School of Computer Science and Technology, University of Science and Technology of China
[‡]Microsoft Research
[§]Key Laboratory of Machine Perception (MOE), School of EECS, Peking University
[†]guojunll@mail.ustc.edu.cn, linlixu@ustc.edu.cn, [‡]{xuta,taoqin,tyliu}@microsoft.com, [§]di_he@pku.edu.cn

Reporter: Junliang Guo
Date: 27 Jan, 2019

Outline

1 Introduction

2 Enhanced Non-Autoregressive Transformer

3 Experiments

4 Conclusion

Introduction

➢ Autoregressive Machine Translation

- Generate a target sequence **word by word from left to right**

$$y_t = \mathbb{D}(y_{1:t-1}, \mathbb{E}(x))$$

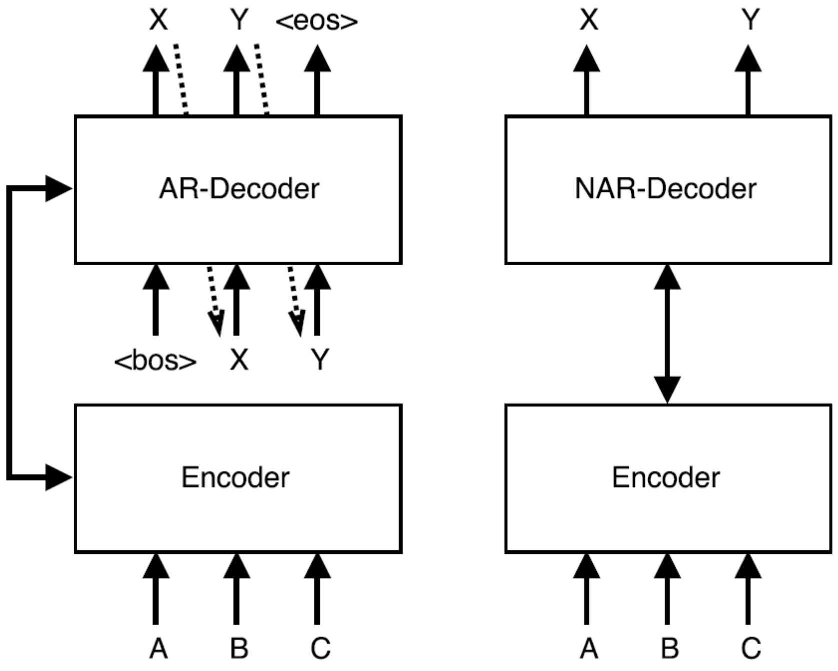
- A natural bottleneck for the inference speed

➢ Non-Autoregressive Machine Translation

- Generate all target tokens **independently and simultaneously**

$$y_t = \mathbb{D}(z, \mathbb{E}(x))$$

where z is the decoder input that is generated independent with y



Gu et.al., ICLR 2018

Non-Autoregressive Machine Translation

- Given the decoder input $z = (z_1, \ldots, z_{T_y})$, the generation of y is defined as:

Target length

Parameters of encoder

$$P(y|x,z) = \prod_{t=1}^{T_y} P(y_t|z,x) = \prod_{t=1}^{T_y} P(y_t|z, \mathbb{E}(x;\theta_{\text{enc}});\theta_{\text{dec}})$$

Parameters of decoder

- Negative log-likelihood loss function

$$L_{\text{neg}}(x,y;\theta_{\text{enc}},\theta_{\text{dec}}) = -\sum_{t=1}^{T_y} \log P(y_t|z,x)$$

Models		Training	Inference
AT models	RNNs based	×	×
	CNNs based	√	×
	Self-Attention based	√	×
NAT models		√	√

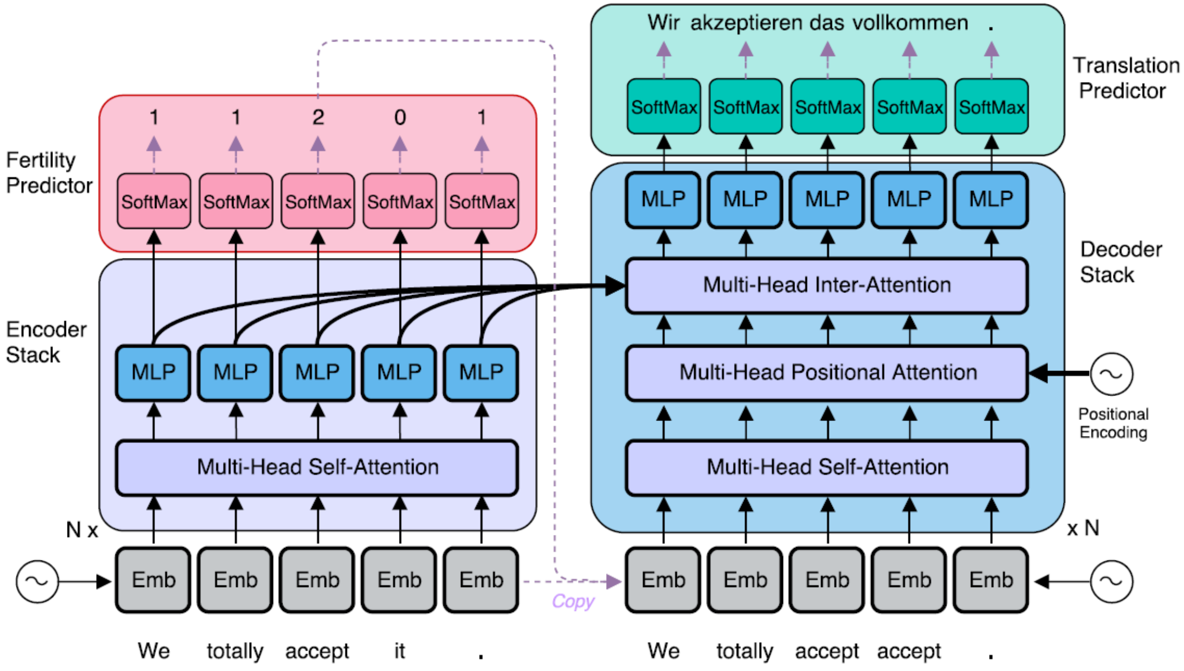
NART (ICLR18)

➢ Takes a copy of source sentence, which is guided by a fertility predictor, as the decoder input

- No target-side information is provided

- Decoder has to handle a harder cross-language task

- Inferior accuracy, e.g., poor on long sentences, missing/duplicating words



Gu et.al., ICLR 2018

Outline

1 Introduction

2 Enhanced Non-Autoregressive Transformer

3 Experiments

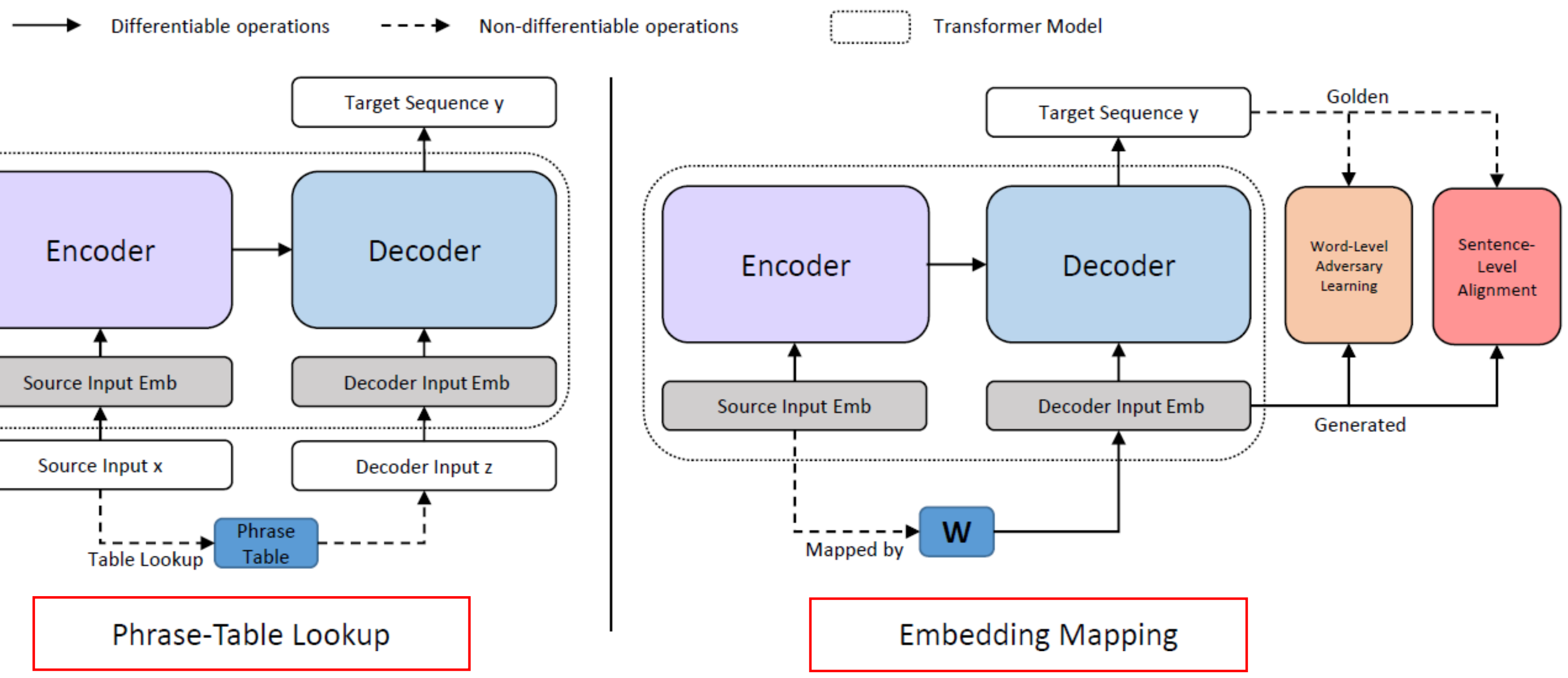
4 Conclusion

Methodology

- We aim to make the decoder input contains target-side information

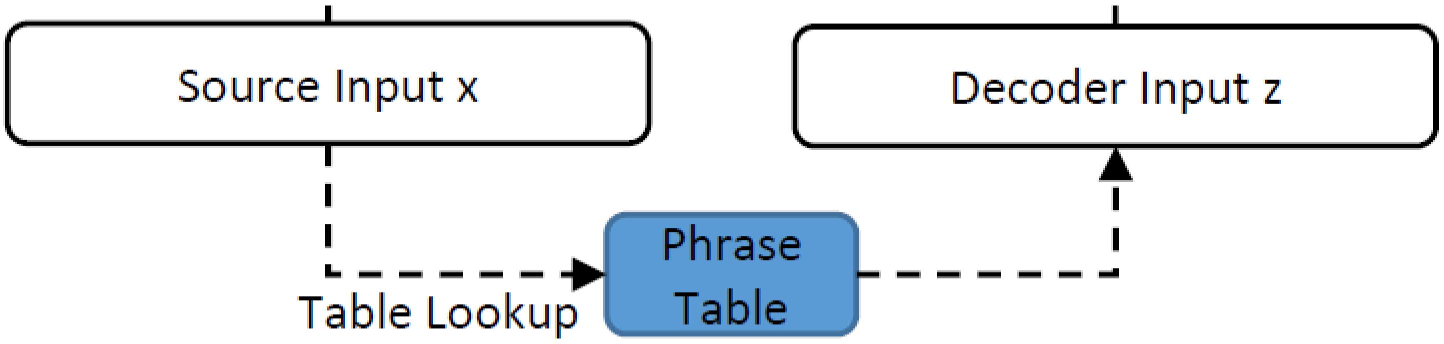
$$y_t = \mathbb{D}(\hat{y}, \mathbb{E}(x))$$

- Two variants:



Phrase-Table Lookup

➢ A straightforward idea is to feed target tokens as decoder input



Pre-train a phrase table on the training set by Moses

with negligible latency

Segment and translate x greedily by phrase-table lookup

Embedding Mapping

- Several shortages of Phrase-Table Lookup model
  - The quality of phrase table depends on the quality of dataset
  - It cannot update its quality autonomous cause it is not end-to-end trained
- We explore to provide target-side information in **embedding space**, instead of explicitly in **token space**

Embedding Mapping

- We use a linear mapping f_G to map the source embedding matrix E_x into the target space:

$$E_{\tilde{z}} = f_G(E_x; W) = E_x W,$$

- Propose two loss functions to ensure a plausible mapping W can be learned

Embedding Mapping

Sentence-level alignment:

$$L_{\text{align}}(x, y) = \|f_G(e(x)) - e(y)\|_2$$

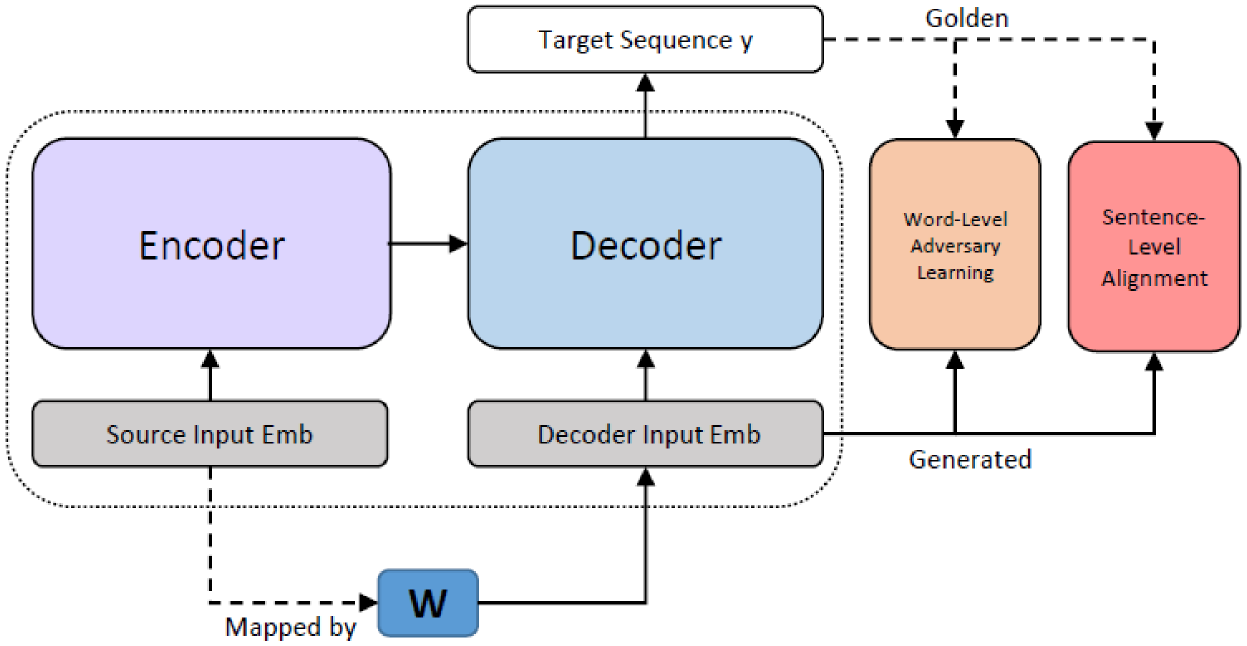
$$e(x) = \frac{1}{T_x}\sum_{i=1}^{T_x} e(x_i)$$

Word-level adversary learning:

$$L_{\text{adv}}(x, y) = \min_{W} \max_{\theta_D} V_{\text{word}}(f_G, f_D)$$

$$V_{\text{word}}(f_G, f_D) = \mathbb{E}_{e(y_i)\sim E_y}[\log f_D(e(y_i))] + \mathbb{E}_{e(x_j)\sim E_x}[\log(1 - f_D(f_G(e(x_j))))]$$

To make the embedding of each token of the decoder input and the target cannot be distinguished by the discriminator D



Embedding Mapping

- The final loss function comes to:

$$\min_{\Theta} \max_{\theta_D} L(x,y) = L_{\text{neg}}(x,y;\theta_{\text{enc}},\theta_{\text{dec}}) + \mu L_{\text{align}}(x,y;W) + \lambda L_{\text{adv}}(x,y;\theta_D,W)$$

1 Introduction

2 Enhanced Non-Autoregressive Transformer

3 Experiments

4 Conclusion

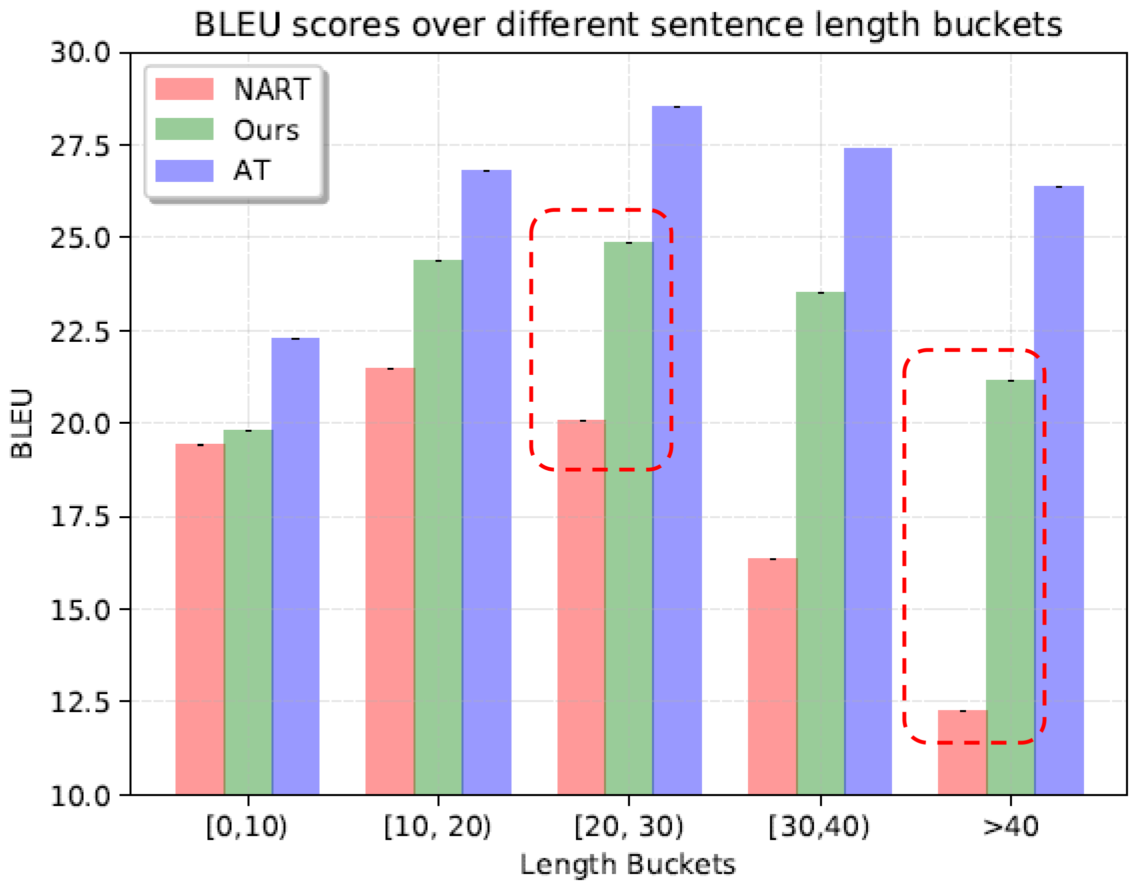
Settings

- We evaluate our model on three datasets:
  - IWSLT14 De-En: 153k training pairs, deep small model
  - WMT14 En-De: 4.5M training pairs, base model
  - WMT16 En-Ro: 2.9M training pairs, base model
- Baselines:
  - NART (Gu et al., ICLR 2018)
  - Latent Transformer (LT) (Kaiser et al., ICML 2018)
  - Iterative Refinement NAT (IR-NAT) (Lee et al., EMNLP 2018)

Translation Accuracy

Models	WMT14 En–De	WMT14 De–En	WMT16 En–Ro	IWSLT14 De–En	Latency	Speedup
LSTM-based S2S (Wu et al. 2016)	24.60	/	/	28.53†	/	/
Transformer (Vaswani et al. 2017)	27.41†	31.29†	35.61†	32.55†	607 ms	1.00×
LT (Kaiser et al. 2018)	19.80	/	/	/	105 ms	5.78×
LT (rescoring 10 candidates)	21.00	/	/	/	/	/
LT (rescoring 100 candidates)	22.50	/	/	/	/	/
NART (Gu et al. 2017)	17.69	21.47	27.29	22.95†	39 ms	15.6×
NART (rescoring 10 candidates)	18.66	22.41	29.02	25.05†	79 ms	7.68×
NART (rescoring 100 candidates)	19.17	23.20	29.79	/	257 ms	2.36×
IR-NAT (Lee, Mansimov, and Cho 2018)	21.54	25.42	29.66	/	254† ms	2.39×
Phrase-Table Lookup	6.03	11.24	9.16	15.69	/	/
ENAT Phrase-Table Lookup	20.26	23.23	29.85	25.09	25 ms	24.3×
ENAT Phrase-Table Lookup (rescoring 9 candidates)	23.22	**26.67**	34.04	**28.60**	50 ms	12.1×
ENAT Embedding Mapping	20.65	23.02	30.08	24.13	**24** ms	**25.3**×
ENAT Embedding Mapping (rescoring 9 candidates)	**24.28**	26.10	**34.51**	27.30	49 ms	12.4×

Comparison in Length Buckets



- NART performs worse on longer sentences
- We achieve more accuracy improvements on these sentences by feeding enhanced decoder input

Case Study

Source:	hier ist ein foto, das ich am nrdlichen ende der baffin-inseln aufnahm, als ich mit inuits auf die narwhal-jagd ging. und dieser mann, olaya, erzhlte mir eine wunderbare geschichte seines grovaters.
Target:	this is a photograph i took at the northern tip of baffin island when i went narwhal hunting with some inuit people, and this man, olayuk, told me a marvelous story of his grandfather.
Teacher:	here's a photograph i took up at the northern end of the fin islands when i went to the narwhal hunt, and this man, olaya, told me a wonderful story of his grandfather.
NART:	here's a photograph that i took up the north end of of the baffin fin when i with iuits went to the narwhal hunt, and this guy guy, ollaya. & lt; em & gt; & lt; / em & gt;
PT:	so here's a photo which i the northern end the detected when i was sitting on on the went. and this man , told me a wonderful story his's.
ENAT Phrase:	here's a photograph i took up at the end of the baffin islands i went to the nnarwhal hunting hunt, and this man, olaaya told me a wonderful story of his grandfather.
ENAT Embedding:	here's a photograph that i took on the north of the end of the baffin islands, when i went to nuits on the narhal hunt, and this man, olaya, told me a wonderful story of his grandfather.

Source:	ich freue mich auf die gesprche mit ihnen allen!
Target:	i look forward to talking with all of you.
Teacher:	i'm happy to talk to you all!
NART:	i'm looking to the talking to to you you.
PT:	i look forward to the conversations with you all!
ENAT Phrase:	i'm looking forward to the conversations with all of you.
ENAT Embedding:	i'm looking forward to the conversations to all of you.

Ablation Study

Approach	Decoder Input	NAT Result
Word-Table Lookup	3.54	19.16
Phrase-Table Lookup	**6.03**	**20.33**

We conduct a **weaker word-to-word** translation to compare with the **phrase-to-phrase** translation to demonstrate the impact of phrase-table quality to the translation accuracy

L_{align}	L_{adv}	BLEU score
√	√	24.13
	√	23.53
√		23.74

The ablation study among the proposed two loss functions of embedding mapping: **sentence-level** alignment and **word-level** adversary learning

1 Introduction

2 Enhanced Non-Autoregressive Transformer

3 Experiments

4 Conclusion

Conclusion

- We demonstrate that the inferior accuracy of non-autoregressive machine translation models comes from the weak target-side information carried in the input to decoder

- We propose two different models to enhance the target-side information in the decoder input, through Phrase-Table Lookup and Embedding Mapping

- We conduct extensive experiments on benchmark datasets to demonstrate the efficacy of proposed models

Q & A

Thanks!