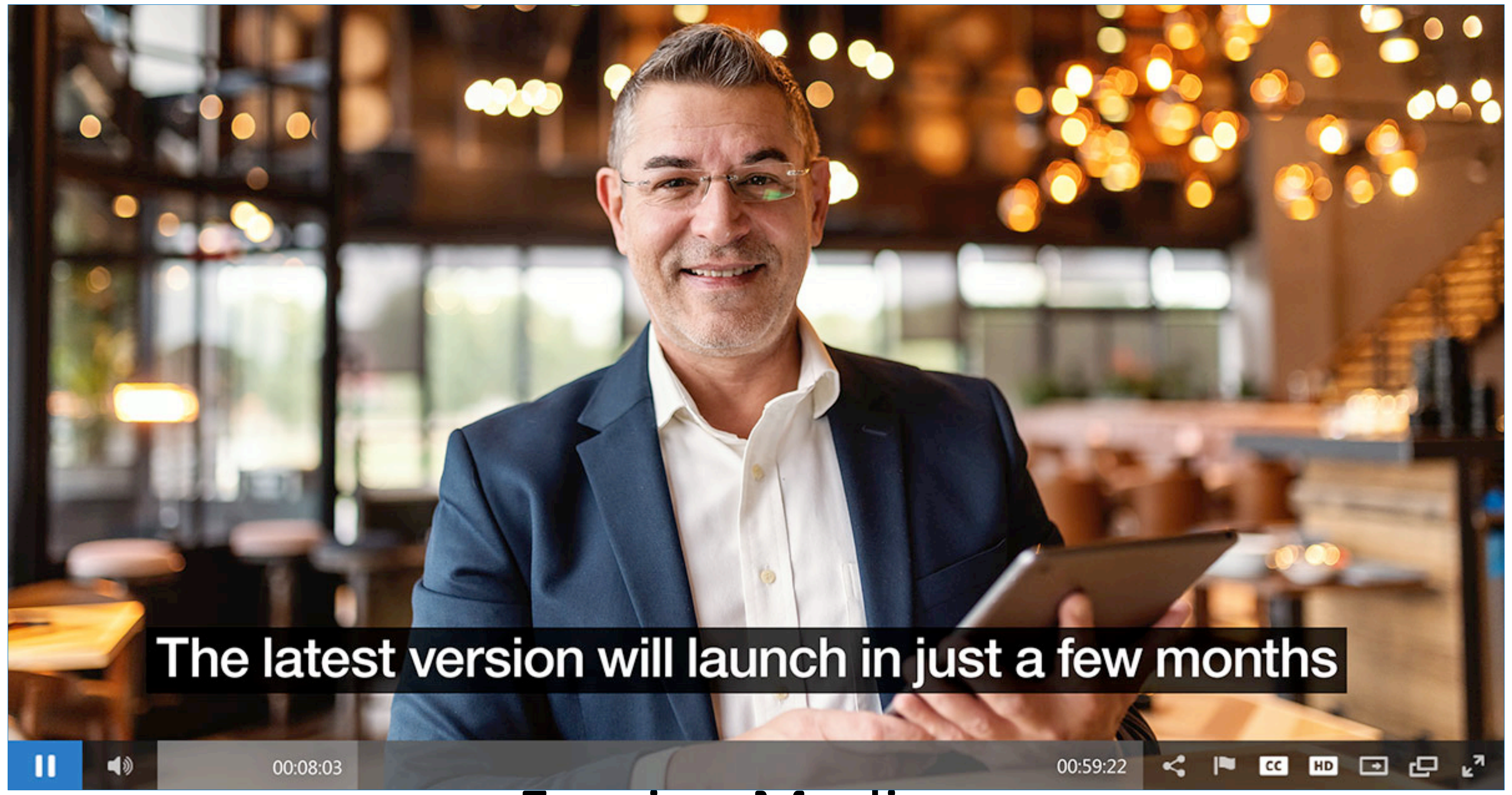
Efficient Machine Translation

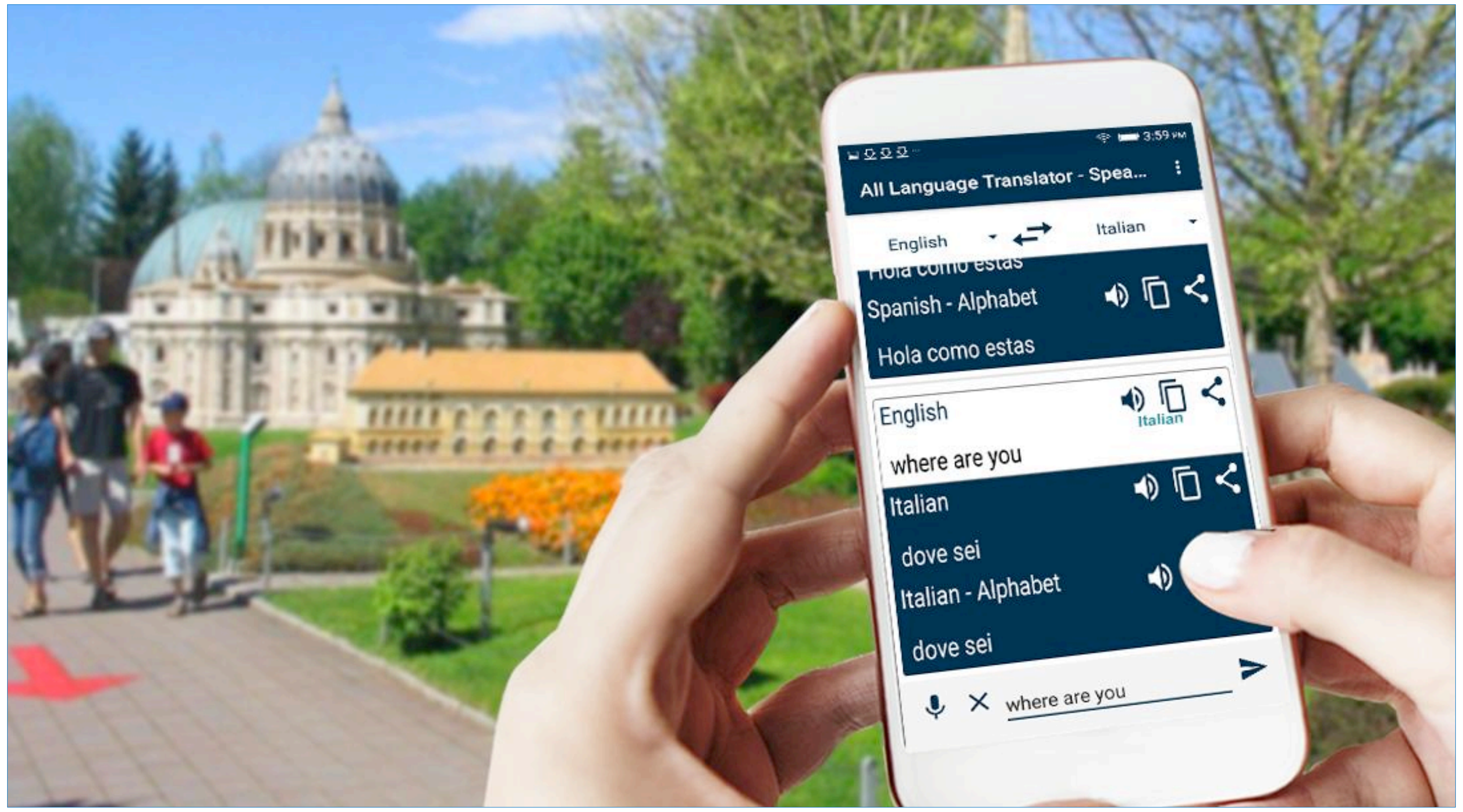
Lei Li
University of California Santa Barbara
2021/10/10

Cross Language Barrier with Machine Translation



Foreign Media

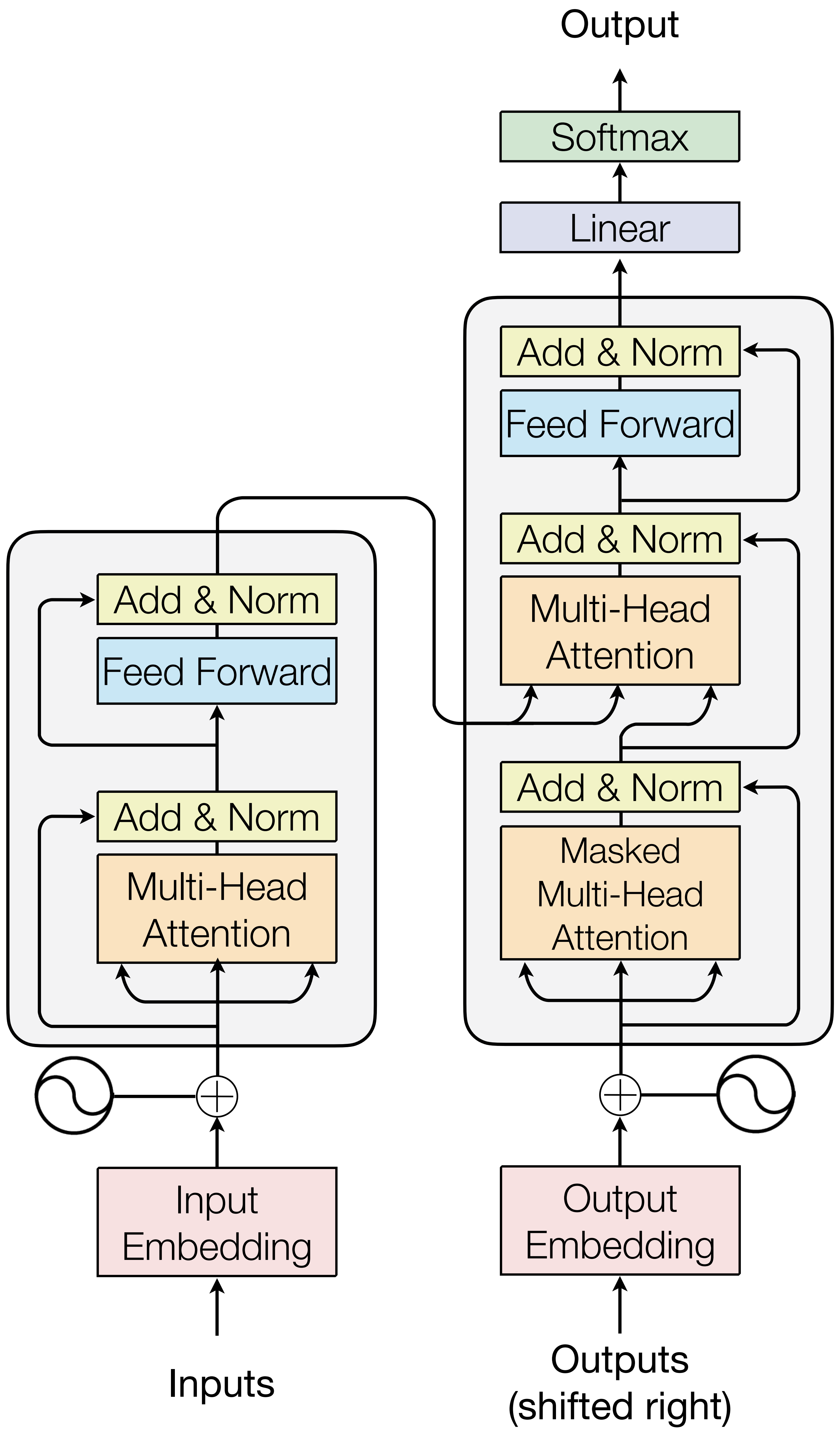
Global Conferences



Tourism

International Trade

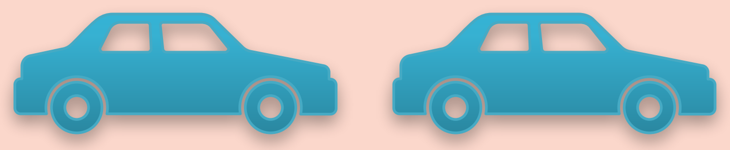
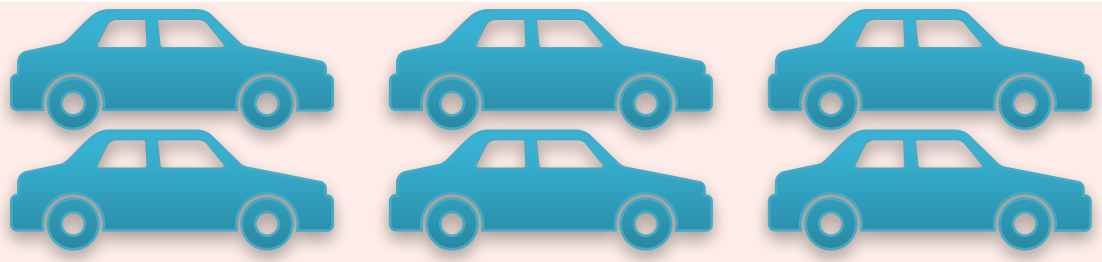
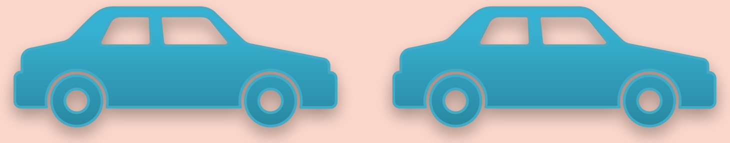
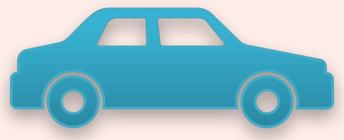
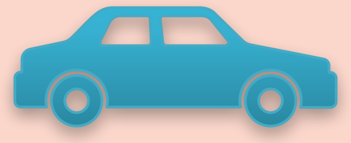
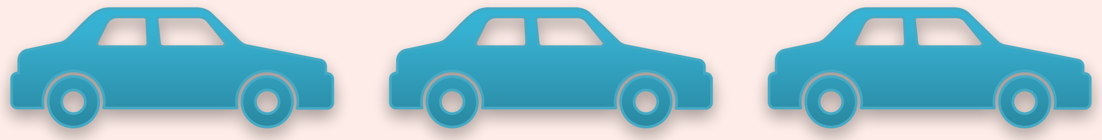
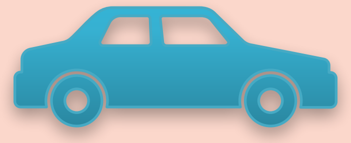
Neural Machine Translation



- Transformer as commonly used backbone architecture for MT.
- 50 - 100m parameters
- Huge computation: 670 GPU hours for training [Vaswani et al 2017].

Training NMT gets more expensive!

- ~~Attention~~ GPU is all you need

model	Size (M)	Total Time (GPU hr)	Train Once (GPU hr)	Infer (ms)	Carbon Footprint (car year)
mRASP (EMNLP20)	60	38k	384		
mRASP2 (ACL 21)	450	128k			
LaSS (ACL 21)	60	41k	384		
LUT (AAAI 21)	144	22k	72	150	
COSTT (AAAI 21)	55	22k	72	140	
Chimera (ACL 21)	165	59k	320	160	
XSTNet (InterSpeech21)	152	24k	240	140	

Affordable and Green MT

- Training NMT models are computationally expensive.
- How to speed up MT training, and inference?
- How to reduce energy consumptions during MT training?

Outline

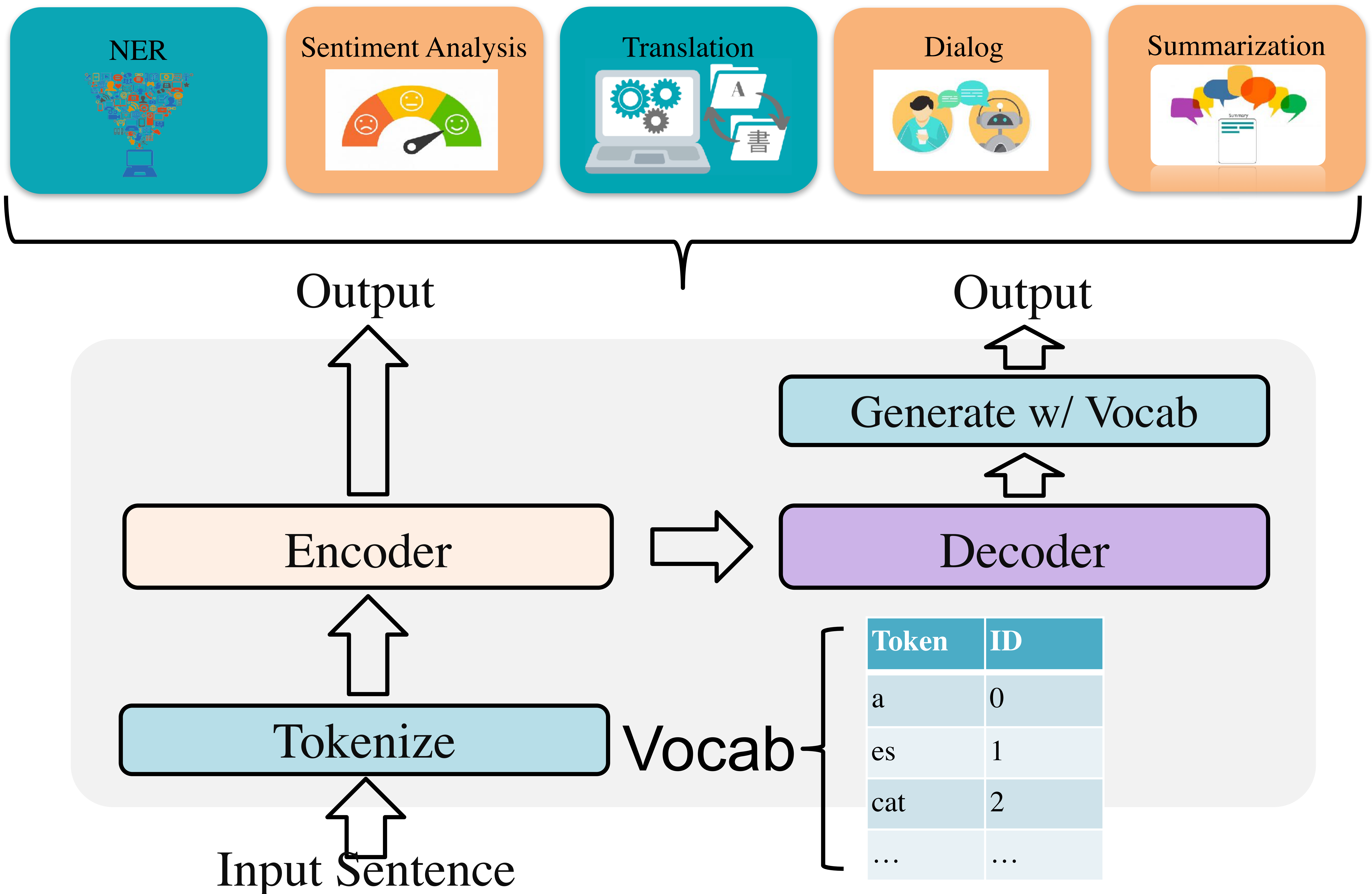
1. Algorithm: Learning Compact Vocabulary for NMT
   – Small vocabulary with improved performance at 100x faster!
2. Model: Parallel Generation
   – Translate at equal or better quality with 10x speedup!
3. Computing: Hardware Acceleration for training and inference
   – Faster than Tensorflow & Pytorch at 14x speedup!

Vocabulary Learning via Optimal Transport for Neural Machine Translation




joint w/ Jingjing Xu[1] Hao Zhou[1] Chun Gan[1] Zaixiang Zheng[1]

Vocabulary is Fundamental and Important



Token	ID
a	0
es	1
cat	2
…	…

Which Vocabulary is Better?

Word level

The	most	eager	is	Oregon	which	is	enlisting	5,000	drivers	in	the	country

Char level

T	h	e	m	o	s	t	e	a	g	e	r	i	s	O	r	e	g	…

Sub-word level

The	most	e	ager	is	O	reg	on	which	is	en	listing	5,000	drivers	in	the	country

Sub-word vocabulary is the dominant choice

* With normal-size data

Why is Sub-word (BPE) superior? Theoretically

- Information theory:
  - Compress the message into compact representation
  - fewest bits to represent both sentence and vocabulary
  - Char-level vocab ==> text sequence will be long
  - Word-level vocab ==> vocab will be large and still OOV
- Entropy:
  - how much information in each token
- Intuition:
  - Reduced entropy (bits-per-char) ==> Better Vocab
  - Even better vocab?

Information-theoretic Vocabulary Evaluation

- Normalized Entropy
  - Information-per-char (IPC)

$$\mathscr{H}(v) = -\frac{1}{l_v}\sum_{i \in v} P(i) log P(i)$$

  - It represents Semantic-information-per-char
    - Smaller IPC is better. Easy to differentiate (therefore easy to generate)

Token	count
a	200
e	90
c	30
t	30
s	90

$\mathscr{H}(v) = 1.37$

VS

Token	count
a	100
aes	90
cat	30

$\mathscr{H}(v) = 0.14$ 😁

Which vocabulary is better?

Sub-word level vocabulary with 1K tokens (BPE-1K)

The	most	e	ag	er	is	O	reg	on	which	is	en	li	st	ing	5	0	00	d	ri	ver	s	in	the	coun	Tr	y

Sub-word level vocabulary with 10K tokens (BPE-10K)

The	most	e	ager	is	O	reg	on	which	is	en	listin g	5,000	dr i	vers	in	the	country

Sub-word level vocabulary with 30K tokens (BPE-30K)

The	most	e	ager	is	O	reg	on	which	is	en	listing	5,000	drivers	in	the	country

From the perspective of size, BPE-1K seems to be better
but longer sequence

* With normal-size data

Which Vocabulary is Better?

Sub-word level vocabulary with 1K tokens (BPE-1K)

The	most	e	ag	er	is	O	reg	on	which	is	en	li	st	ing	5	0	00	d	ri	ver	s	in	the	coun	Tr	y

Sub-word level vocabulary with 10K tokens (BPE-10K)

The	most	e	ager	is	O	reg	on	which	is	en	listin g	5,000	dr i	vers	in	the	country

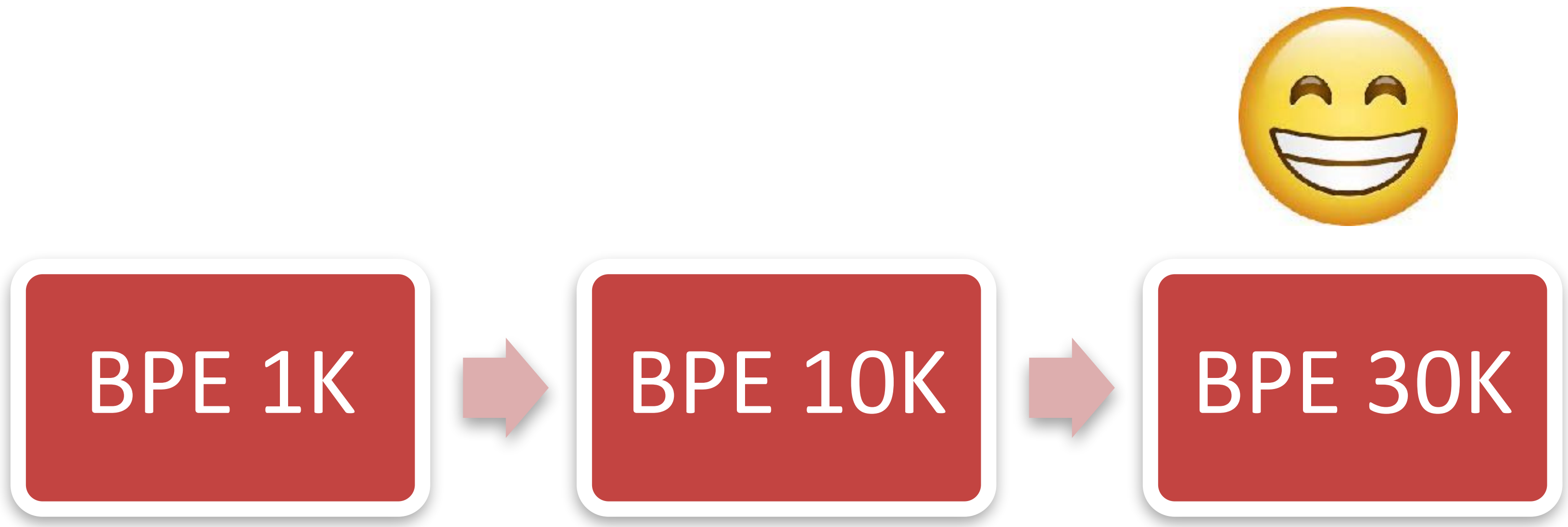
Sub-word level vocabulary with 30K tokens (BPE-30K)

The	most	e	ager	is	O	reg	on	which	is	en	listing	5,000	drivers	in	the	country

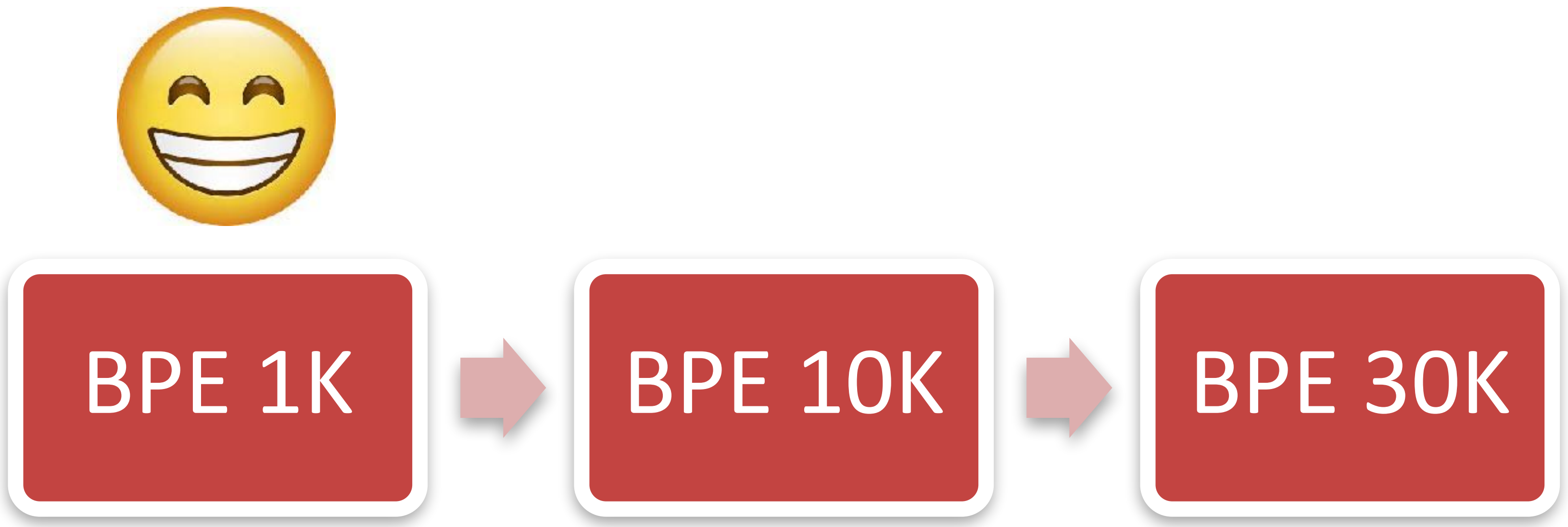
From the perspective of entropy, BPE-30K seems to be better

* With normal-size data

Evaluating Vocabulary Quality is Expensive



IPC



Size

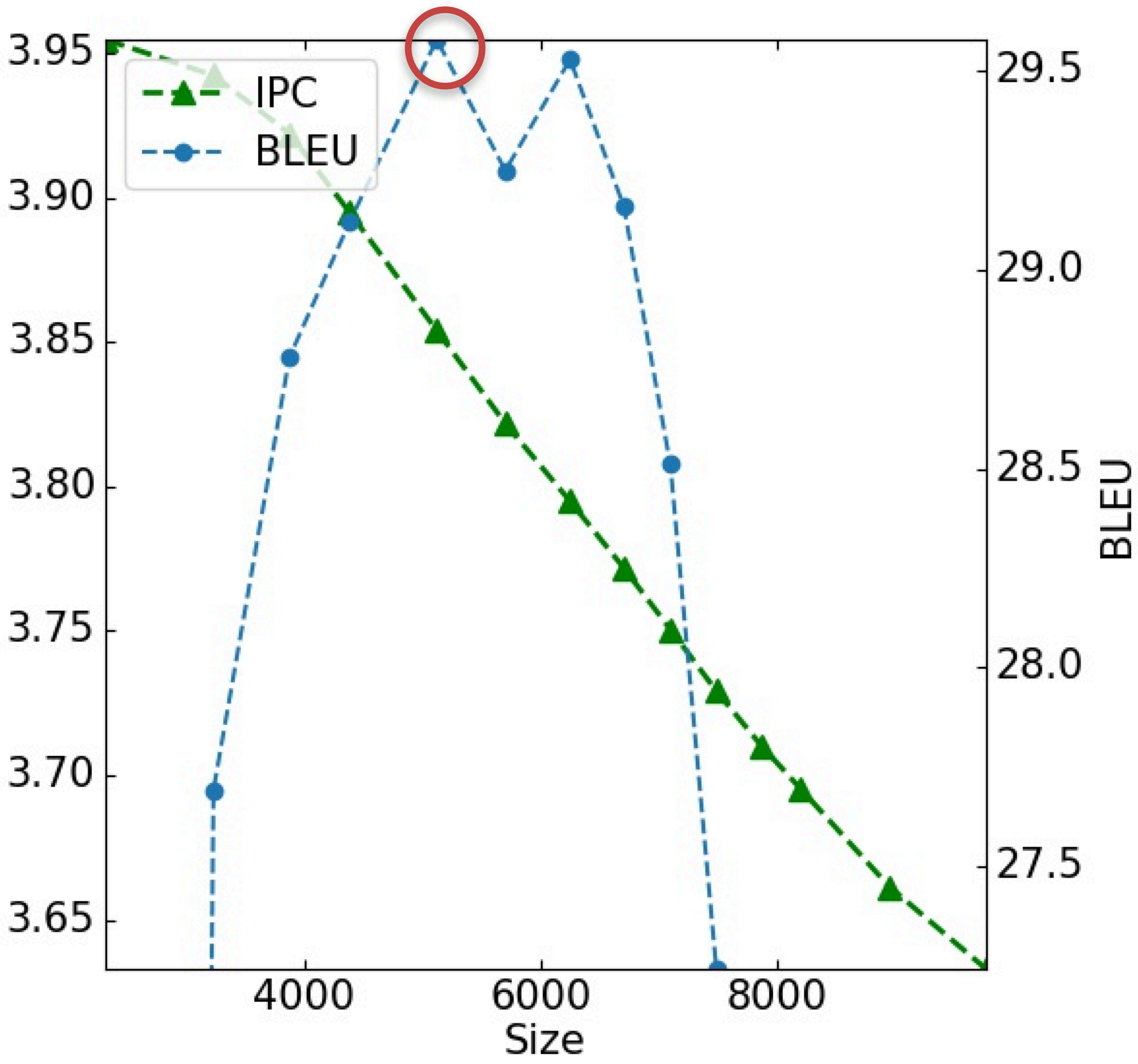
Full training and testing are required to find the optimal vocabulary!

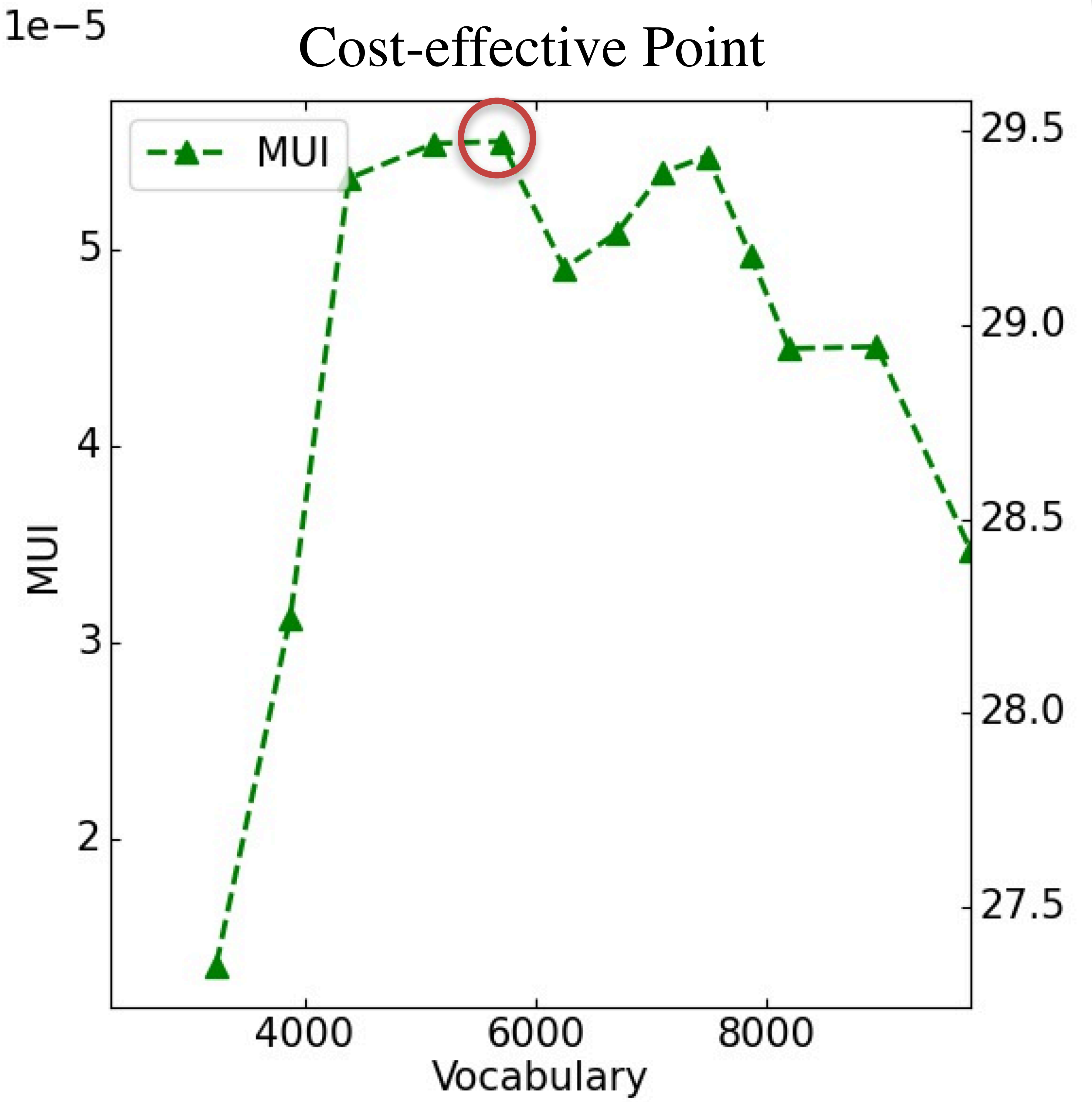
Trading IPC with Size

- Value: **IPC** 
- Cost: **size** 
- Marginal utility of information for Vocabulary (MUV)
  - Negative **gradients** of IPC to size
  - How many value does each unit-of-cost bring?

Xu, Zhou, Gan, Zheng, Li. Vocabulary Learning via Optimal Transport for Neural Machine Translation. ACL 2021.

MUV is good indicator for MT performance



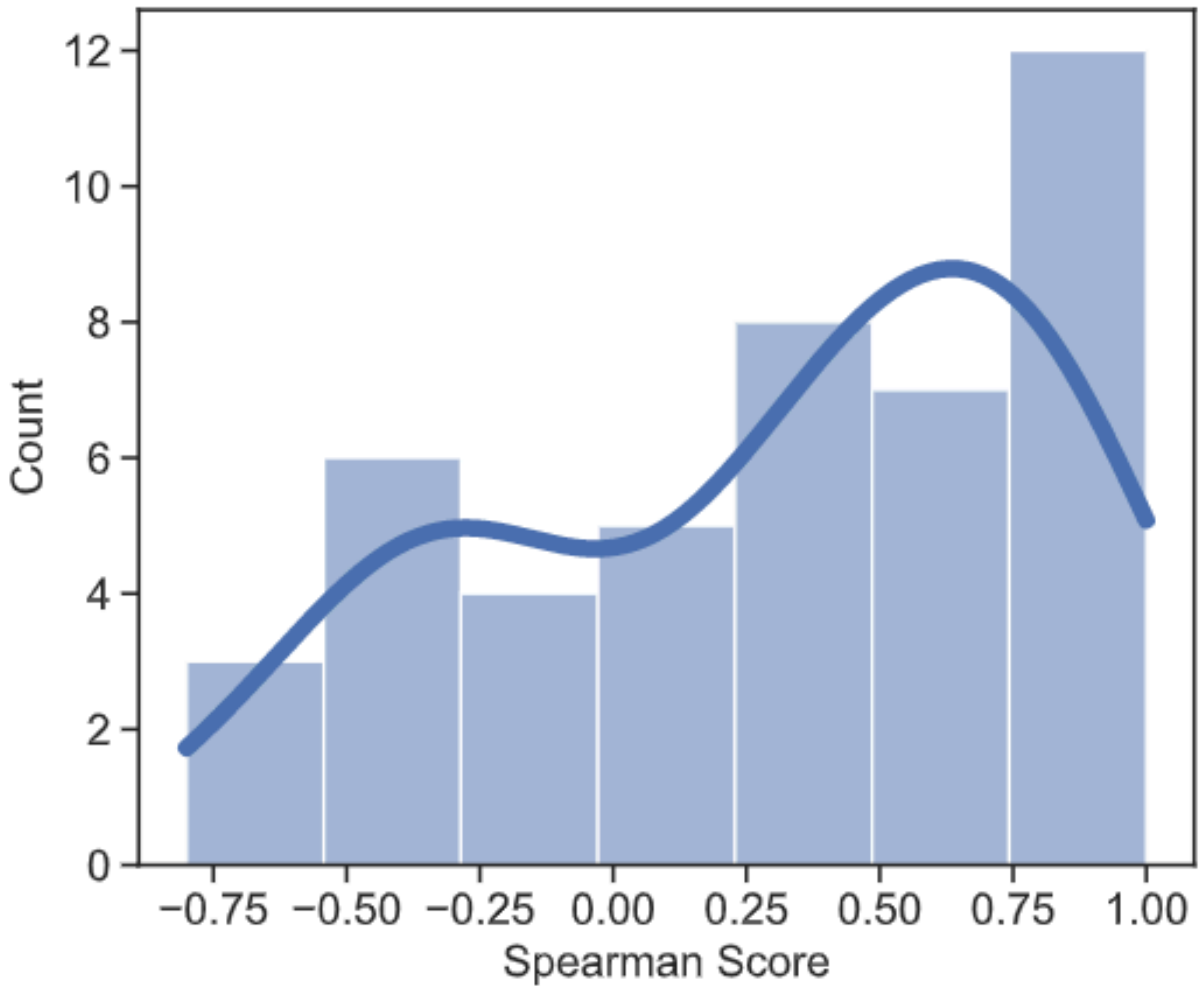


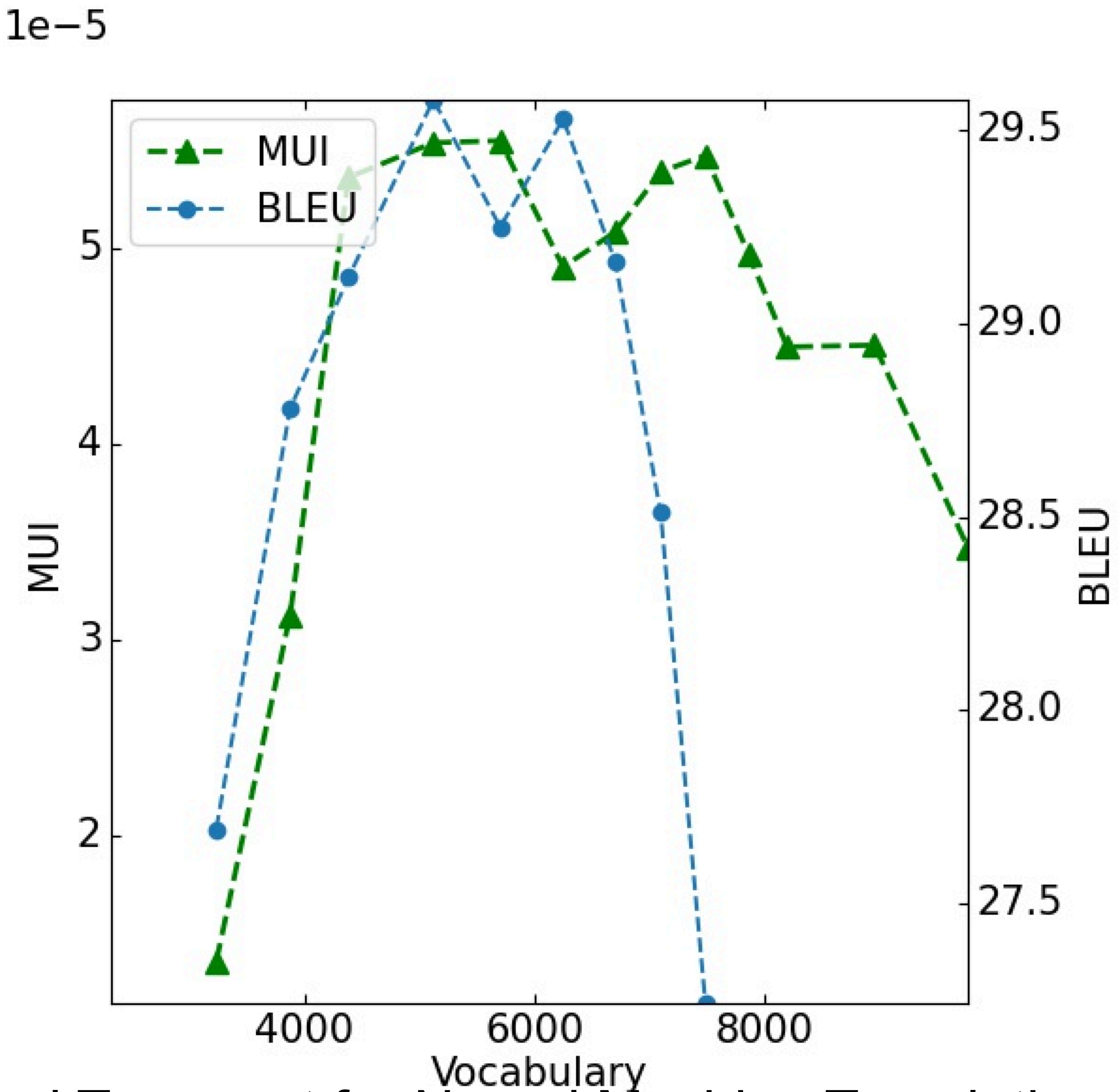
- Cost-effective point in MUV curve (maximum MUV)
  - ==> best BLEU

Xu, Zhou, Gan, Zheng, Li. Vocabulary Learning via Optimal Transport for Neural Machine Translation. ACL 2021.

MUV Indicates MT Performance

- MUV and BLEU are **correlated** on two-thirds of tasks
- **A good coarse-grained evaluation metric!**





Xu, Zhou, Gan, Zheng, Li. Vocabulary Learning via Optimal Transport for Neural Machine Translation. ACL 2021.

Maximizing Marginal Utility of Vocab

- Goal: finding the optimal vocabulary

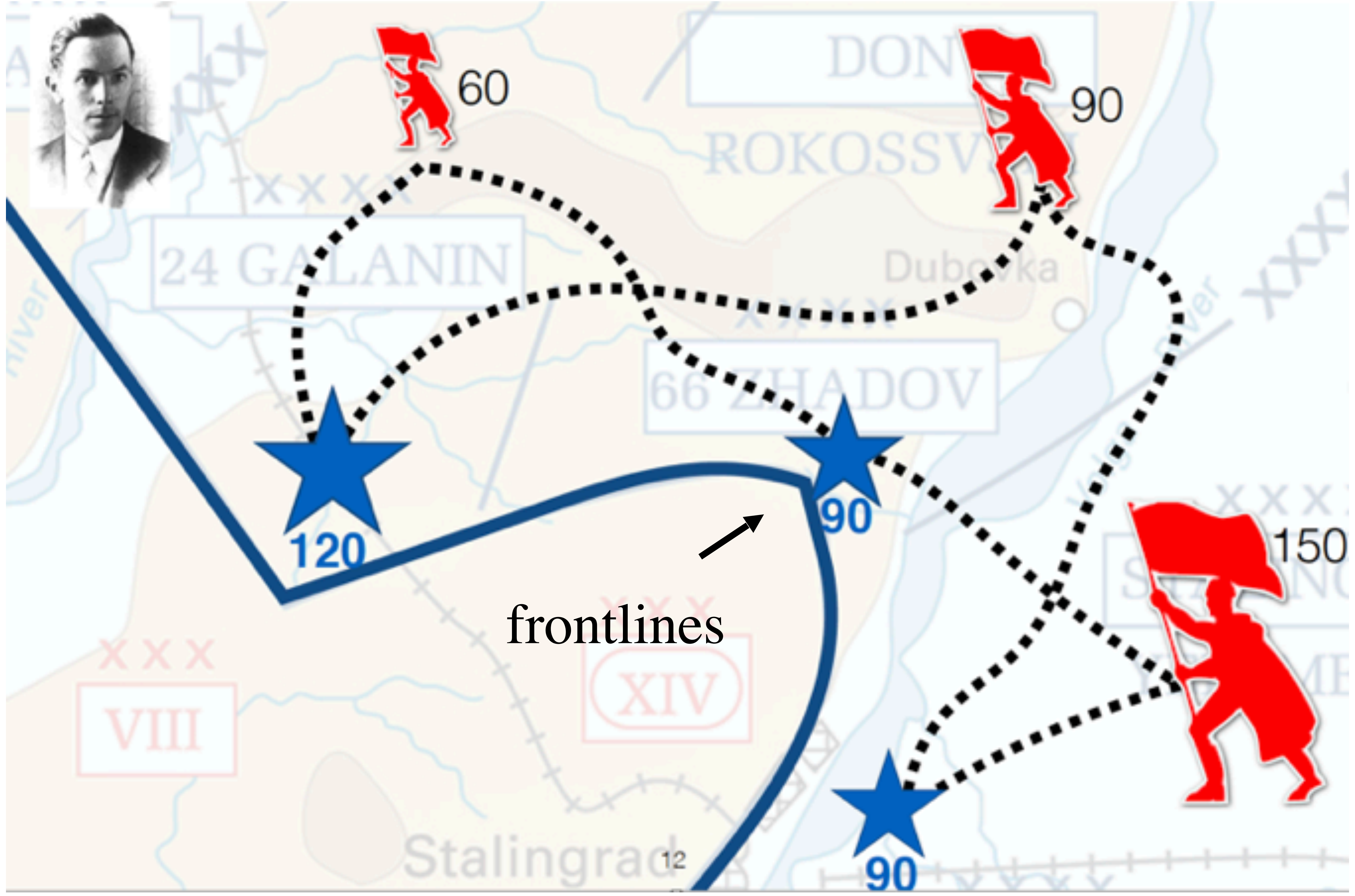
Finding the optimal vocabulary ⟺ Maximizing MUV

- Naive solution:
  – Exhaustive Search for vocabulary with max MUI
- How to search over a huge discrete space?

Xu, Zhou, Gan, Zheng, Li. Vocabulary Learning via Optimal Transport for Neural Machine Translation. ACL 2021.

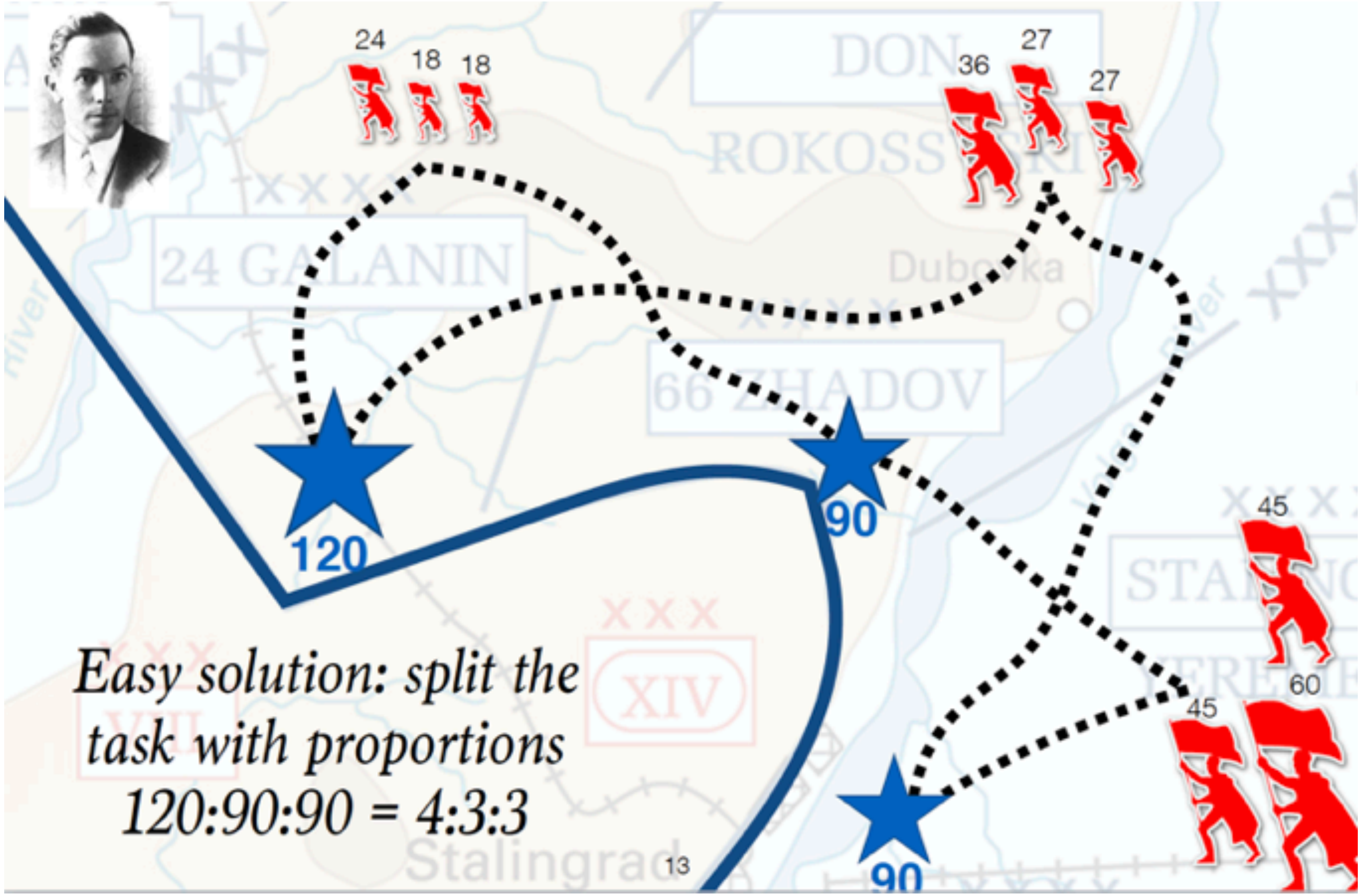
Problem Reduction

- Best BLEU ==> Max MUV ==> Optimal Transport

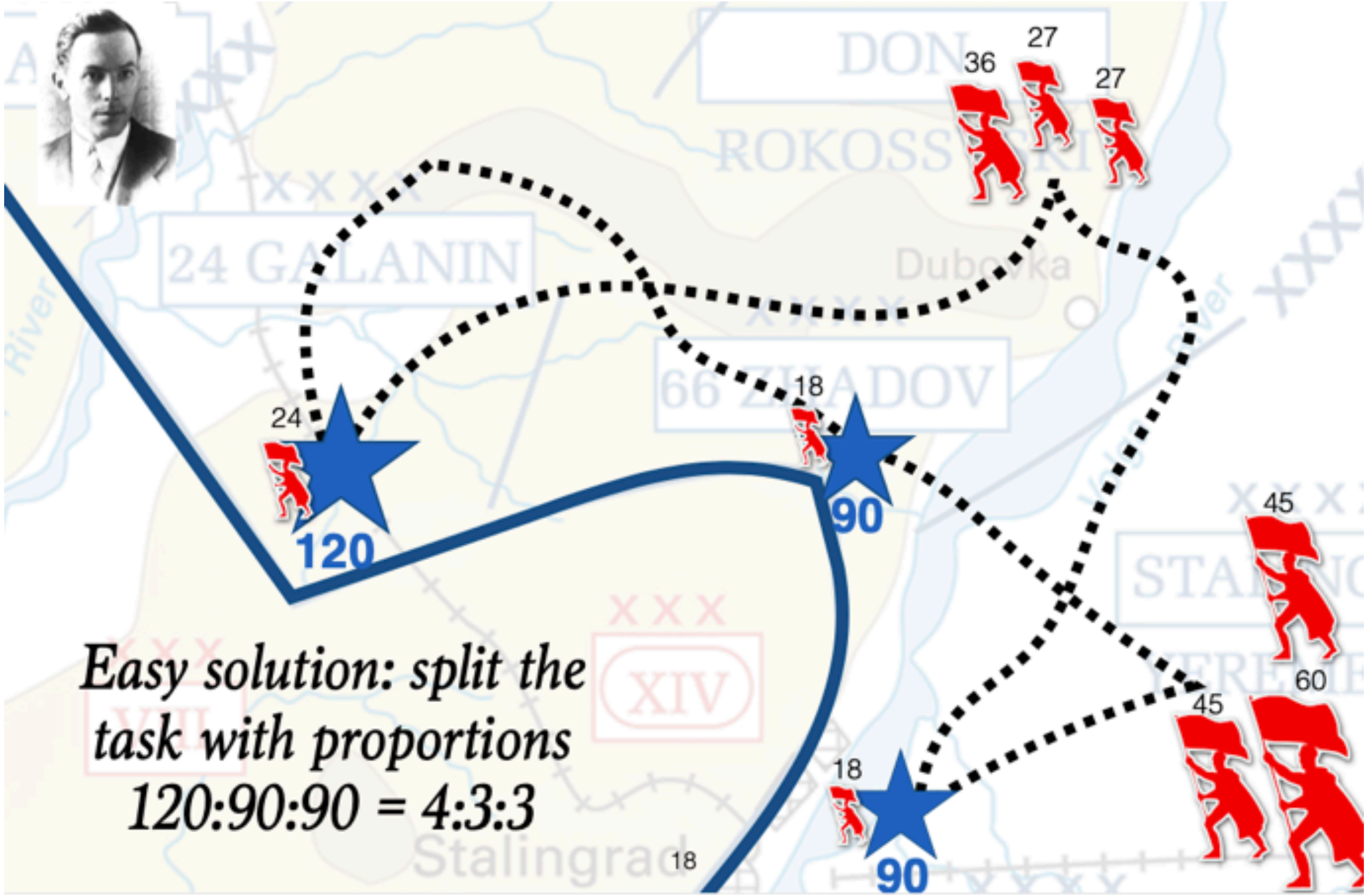


Min cost to Transport soldiers from bases to frontlines

Optimal Transport



Optimal Transport



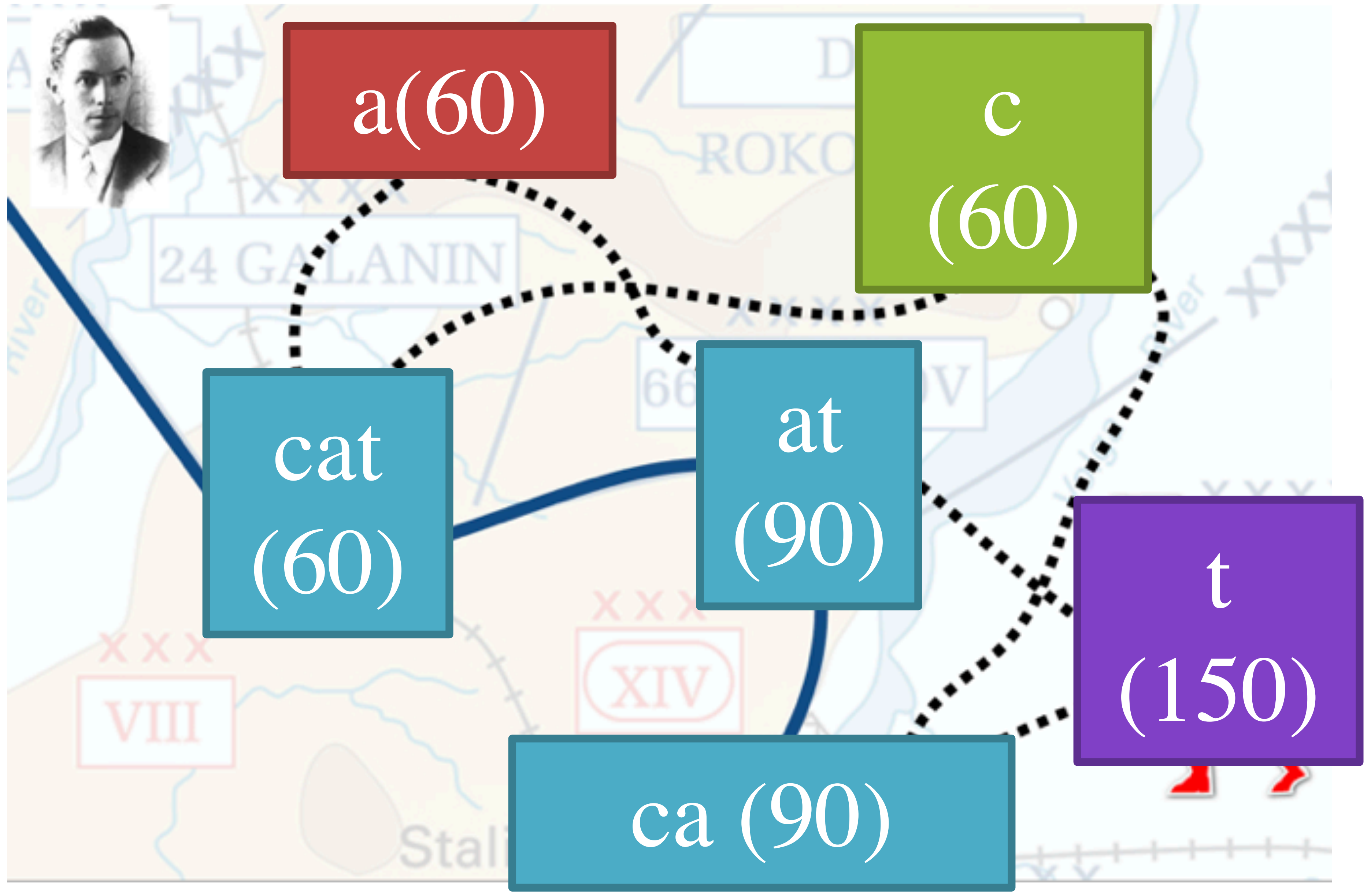
Vocabulary building as Transportation

- Adding one new token means:
  - Transport character frequency to token frequency
  - a b c d e a b d c e f: 2 a 2 b
  - ab c d e ab d c e f: 0 a 0b and 2 ab

Xu, Zhou, Gan, Zheng, Li. Vocabulary Learning via Optimal Transport for Neural Machine Translation. ACL 2021.

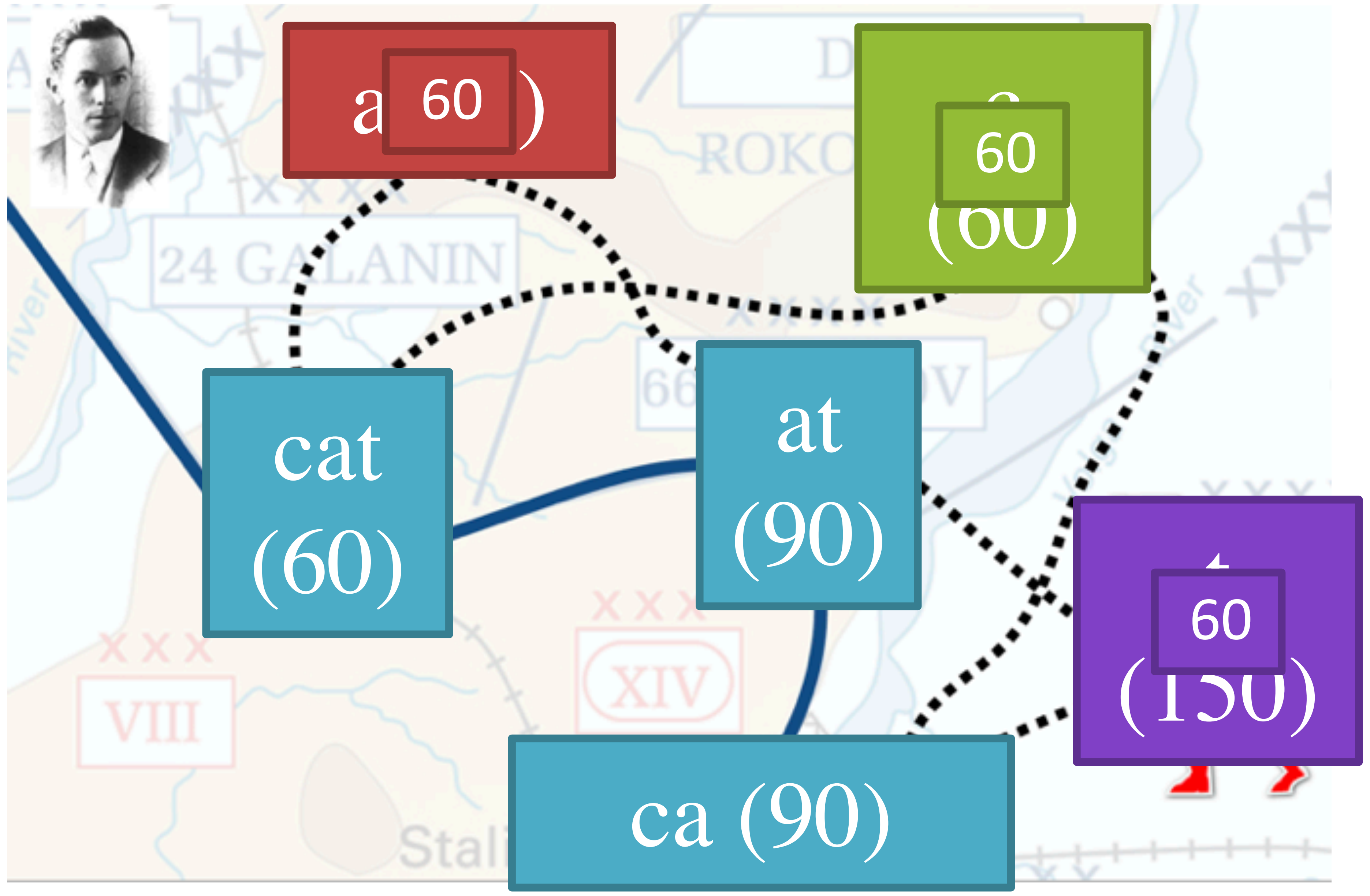
VOLT Formulation

Transport chars to tokens



VOLT Formulation

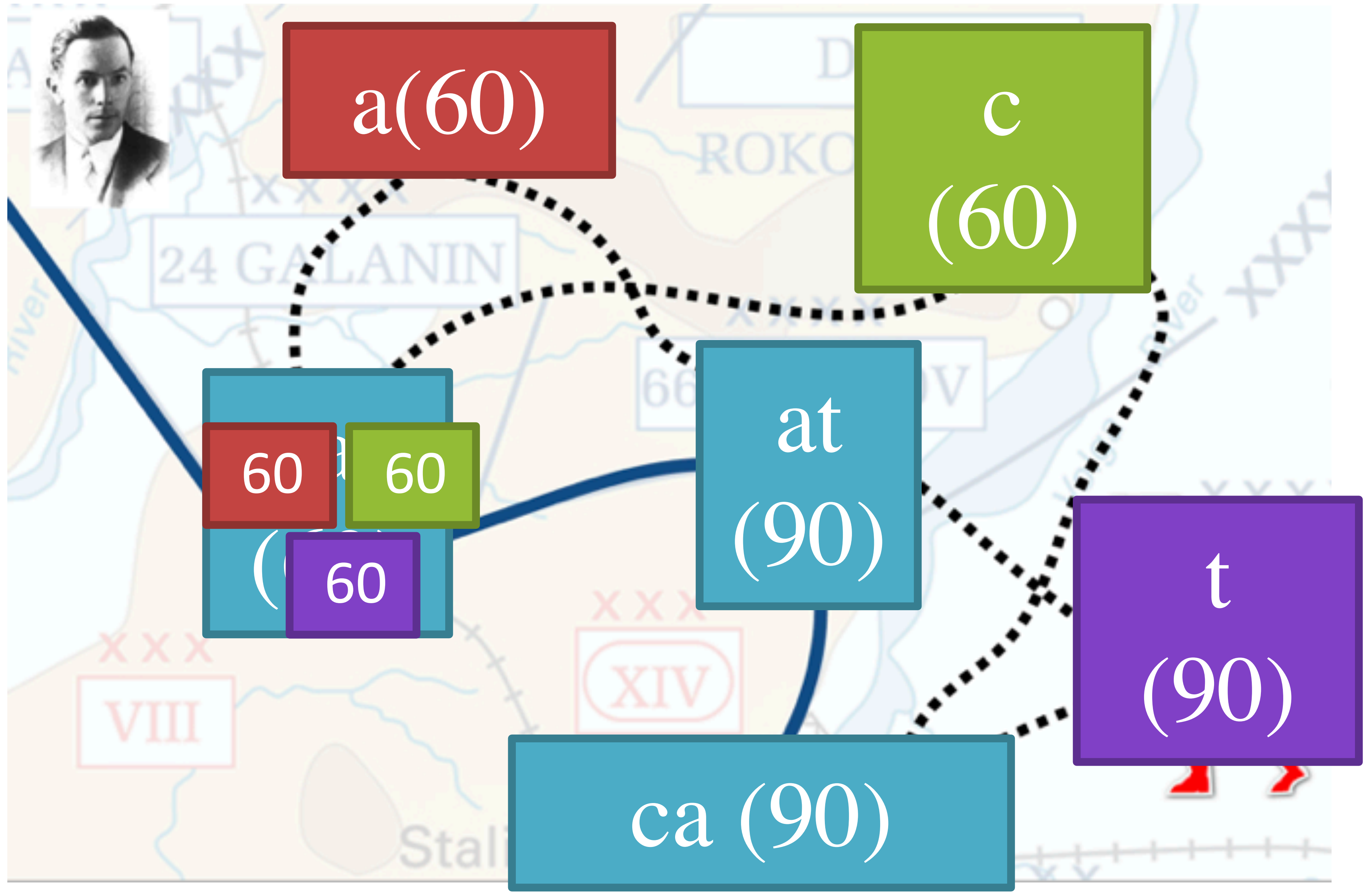
Not all tokens can get chars



Xu, Zhou, Gan, Zheng, Li. Vocabulary Learning via Optimal Transport for Neural Machine Translation. ACL 2021.

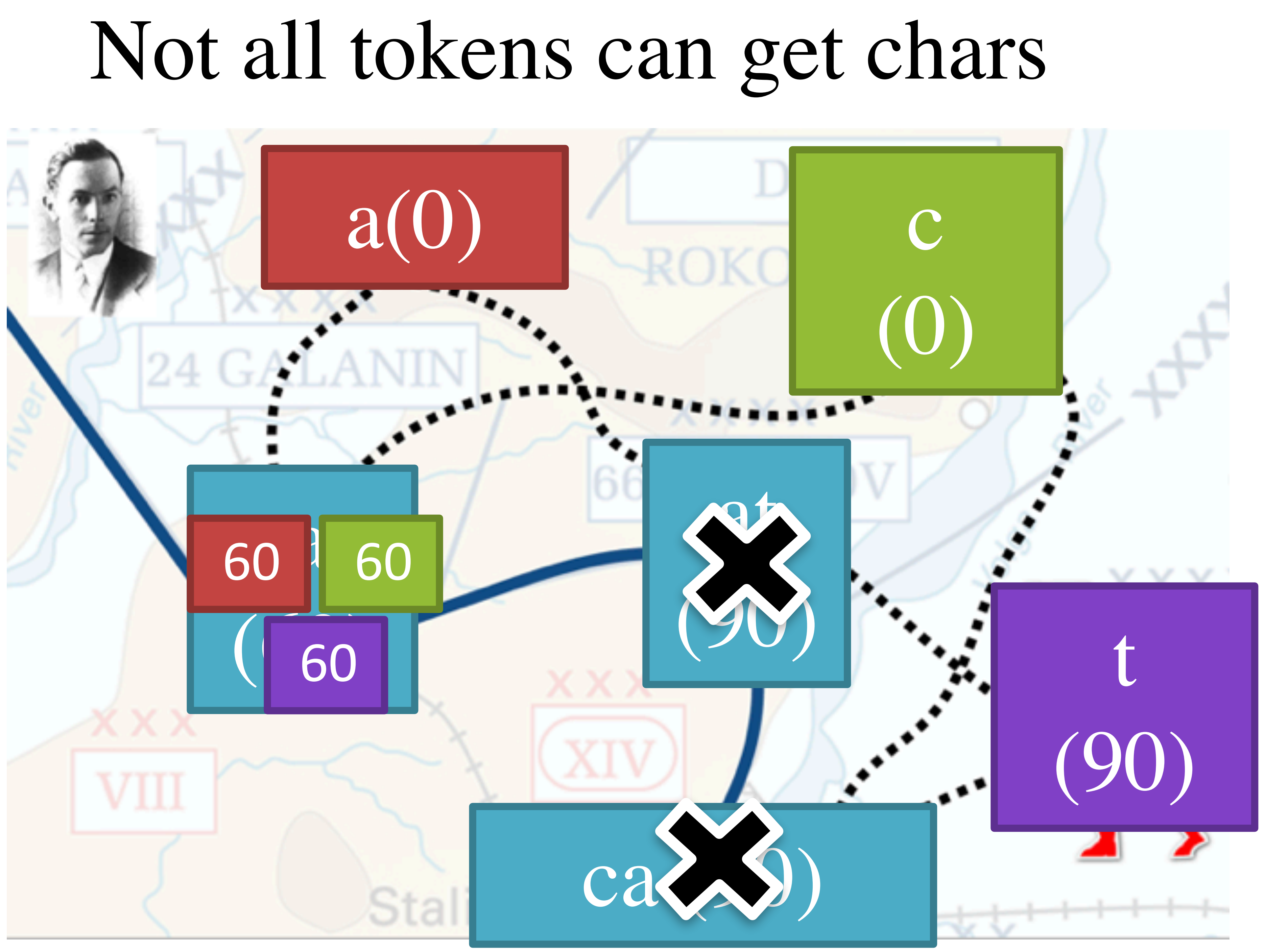
VOLT Formulation

Not all tokens can get chars



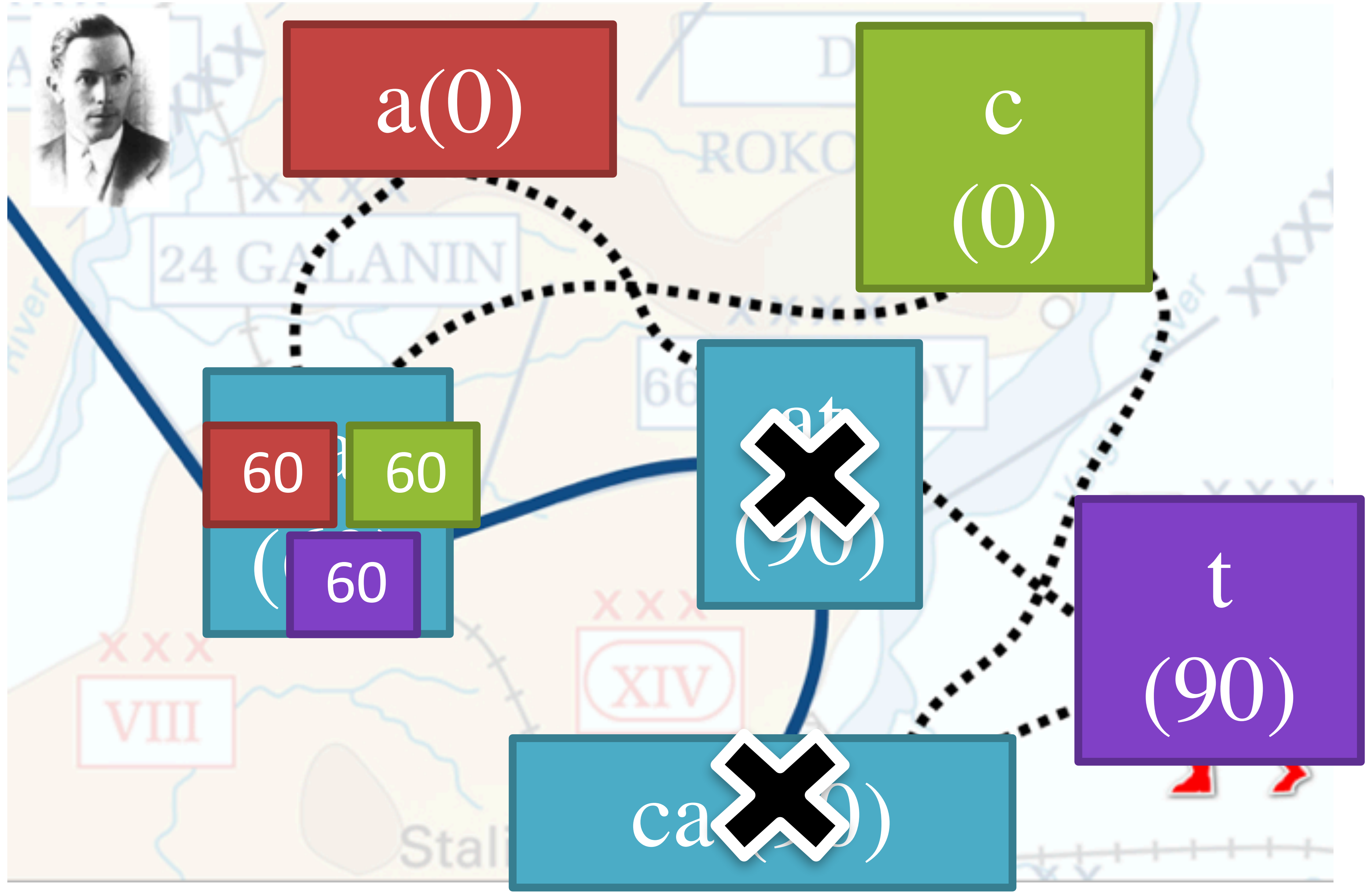
Xu, Zhou, Gan, Zheng, Li. Vocabulary Learning via Optimal Transport for Neural Machine Translation. ACL 2021.

VOLT Formulation



Xu, Zhou, Gan, Zheng, Li. Vocabulary Learning via Optimal Transport for Neural Machine Translation. ACL 2021.

Each Transportation Defines a Vocabulary



Xu, Zhou, Gan, Zheng, Li. Vocabulary Learning via Optimal Transport for Neural Machine Translation. ACL 2021.

Reducing MUV Optimization to OT

- The vocabulary with the maximum MUV
  - Maximum gap between IPC of a vocabulary (with size t) and that of a smaller vocabulary (with size <t)
  - $\max -(H(V_{t+1}) - H(V_t))$
- Intractable, instead to maximize upper-bound of gap $(H(V_{t+1}) - H(V_t))$
- ==> $\max(\max H(V_{t+1}) - \max H(V_t))$
- Finding max $H(V_t)$ ==> Optimal Transport

Xu, Zhou, Gan, Zheng, Li. Vocabulary Learning via Optimal Transport for Neural Machine Translation. ACL 2021.

Optimization

- Find the transportation matrix (=vocab) with lowest cost (-MUV)

Constraints

$$\forall j \in \{a,b,c\}, \sum_{i \in \{ab,bc,a\}} P_{i,j} = P_j$$

$$\forall i \in \{ab,bc,a\}, \sum_{j \in \{a,b,c\}} P_{i,j} - P_i \leq \epsilon$$

Problem

$$\min_{\text{all } P} C(P)$$

Cost Function

$$C(P) = -H(P) + \sum_{\substack{j \in \{a,b,c\}, \\ i \in \{ab,bc,a\}}} P_{i,j} D_{i,j}$$

Transportation matrix P

	cat	at	tea
a	20	10	0
c	20	0	0
e	0	0	0
t	20	10	0

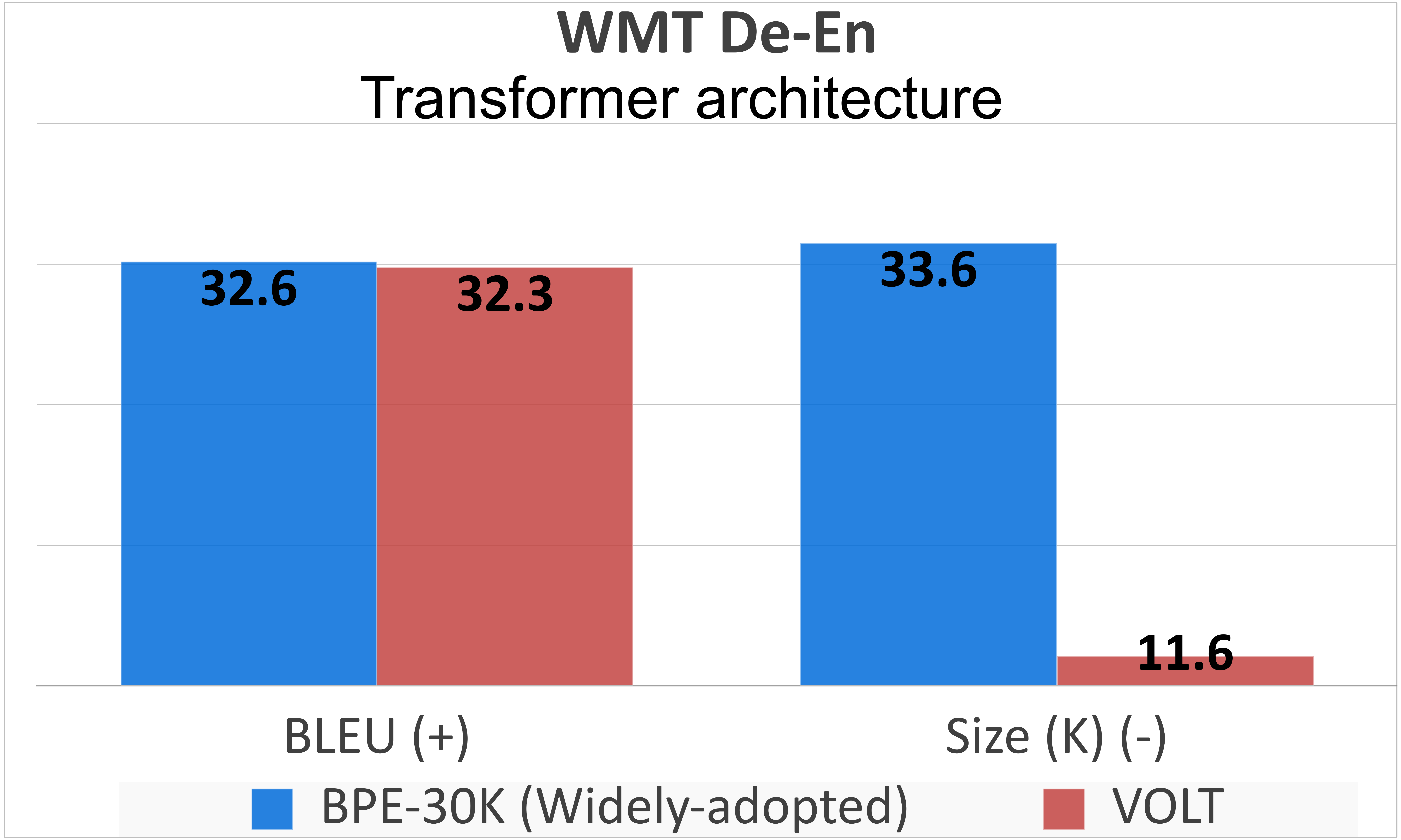
Cost matrix D

	cat	at	tea
a	1	1	1
c	1	∞	∞
e	∞	∞	1
t	1	1	∞

- Sinkhorn Algorithm [Gabriel Peyré et. al]

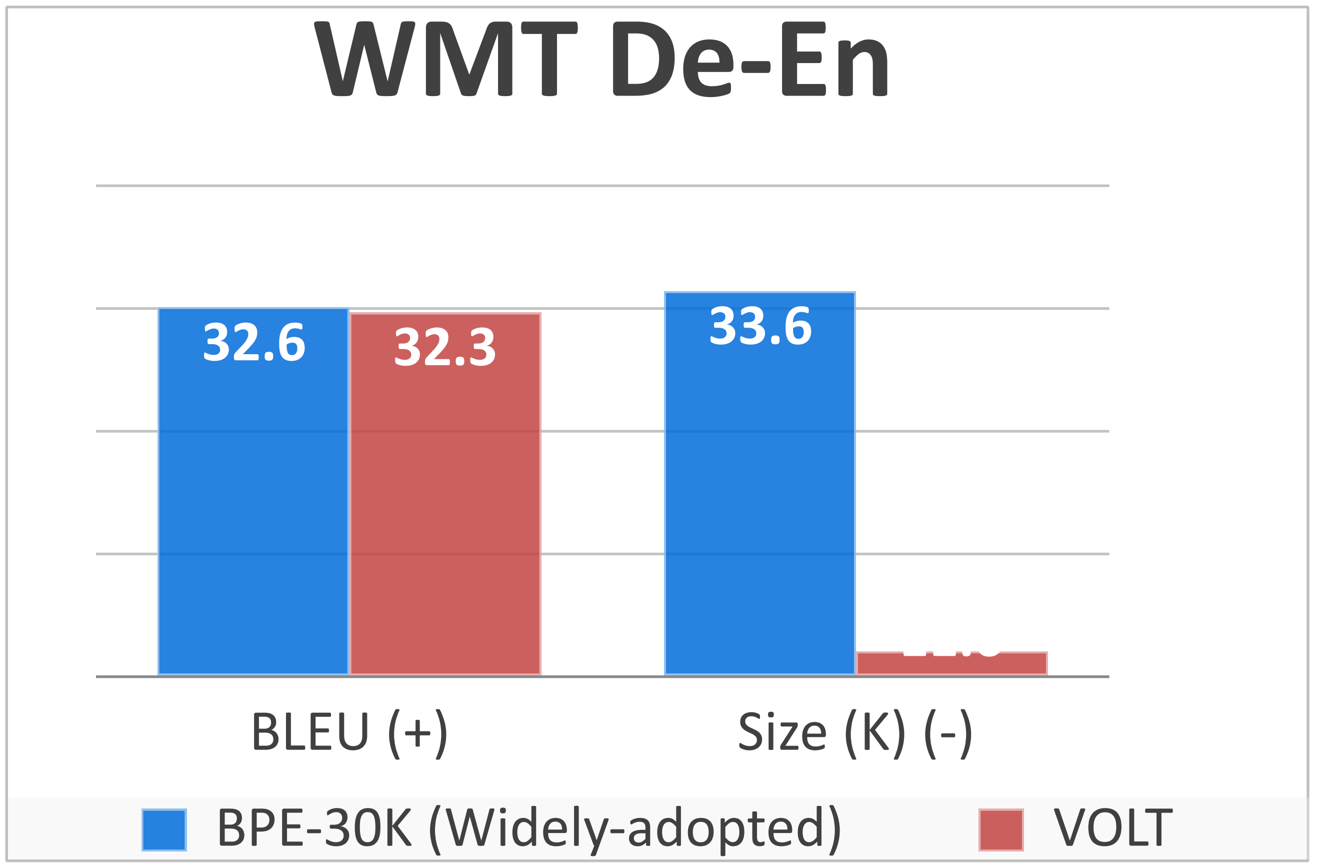
Xu, Zhou, Gan, Zheng, Li. Vocabulary Learning via Optimal Transport for Neural Machine Translation. ACL 2021.

VOLT finds better vocabulary on Bilingual MT



Xu, Zhou, Gan, Zheng, Li. Vocabulary Learning via Optimal Transport for Neural Machine Translation. ACL 2021.

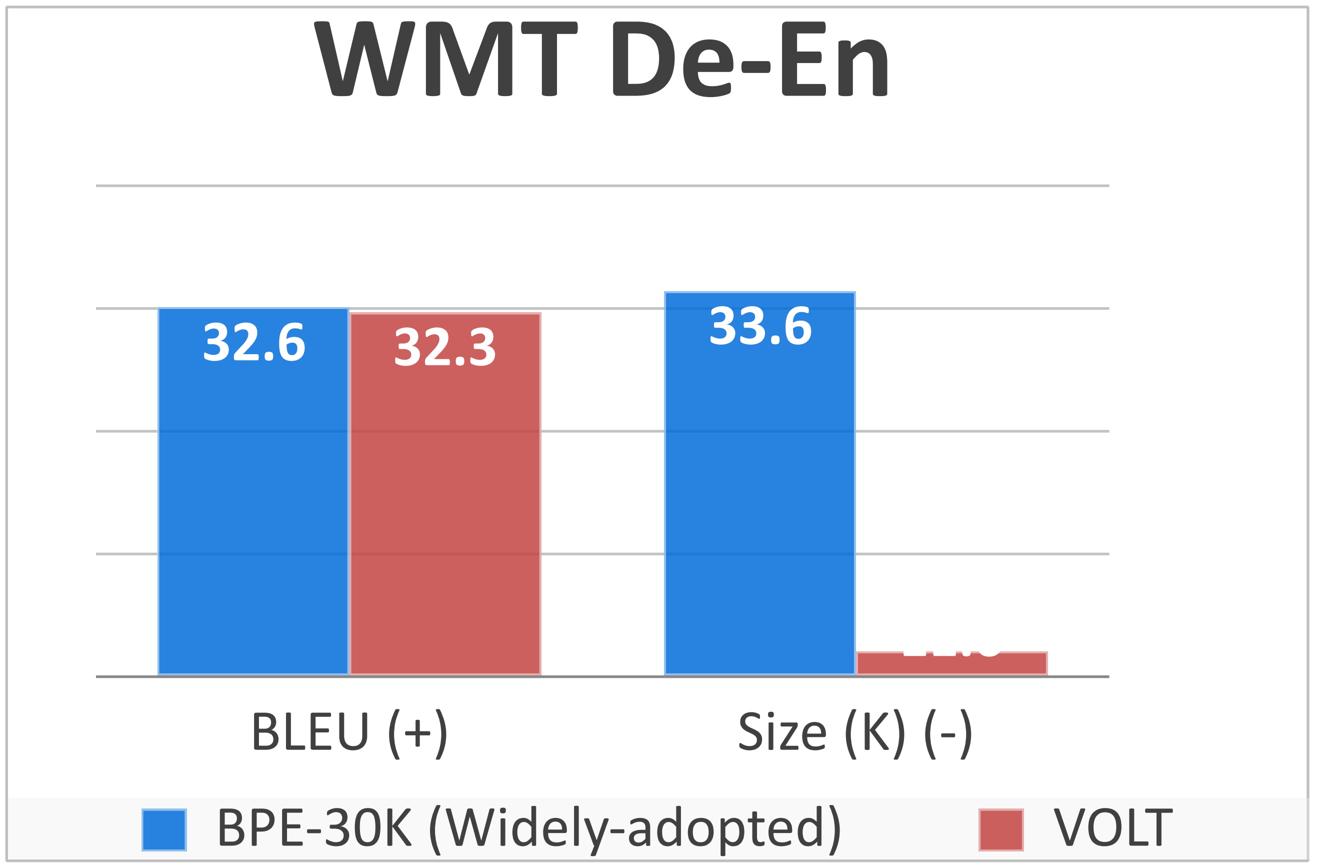
VOLT finds better vocabulary on Bilingual MT

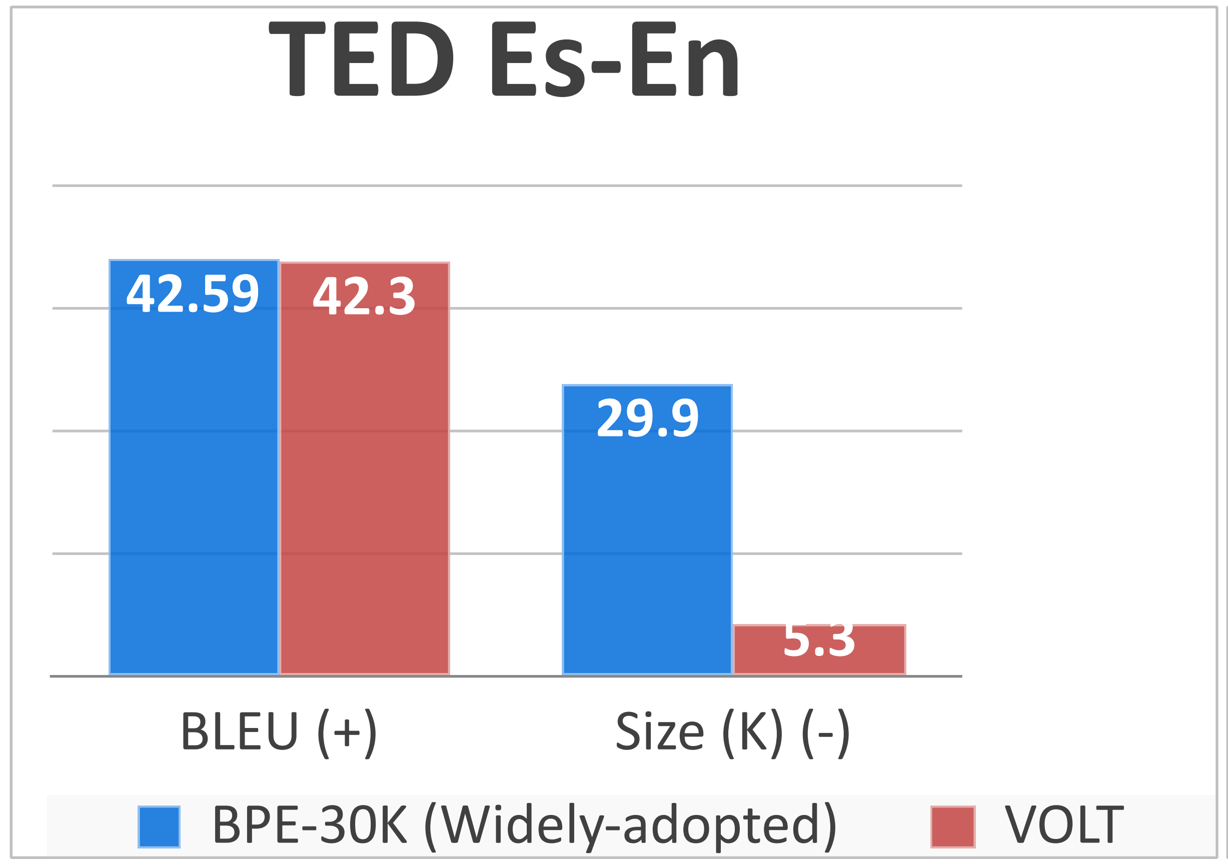


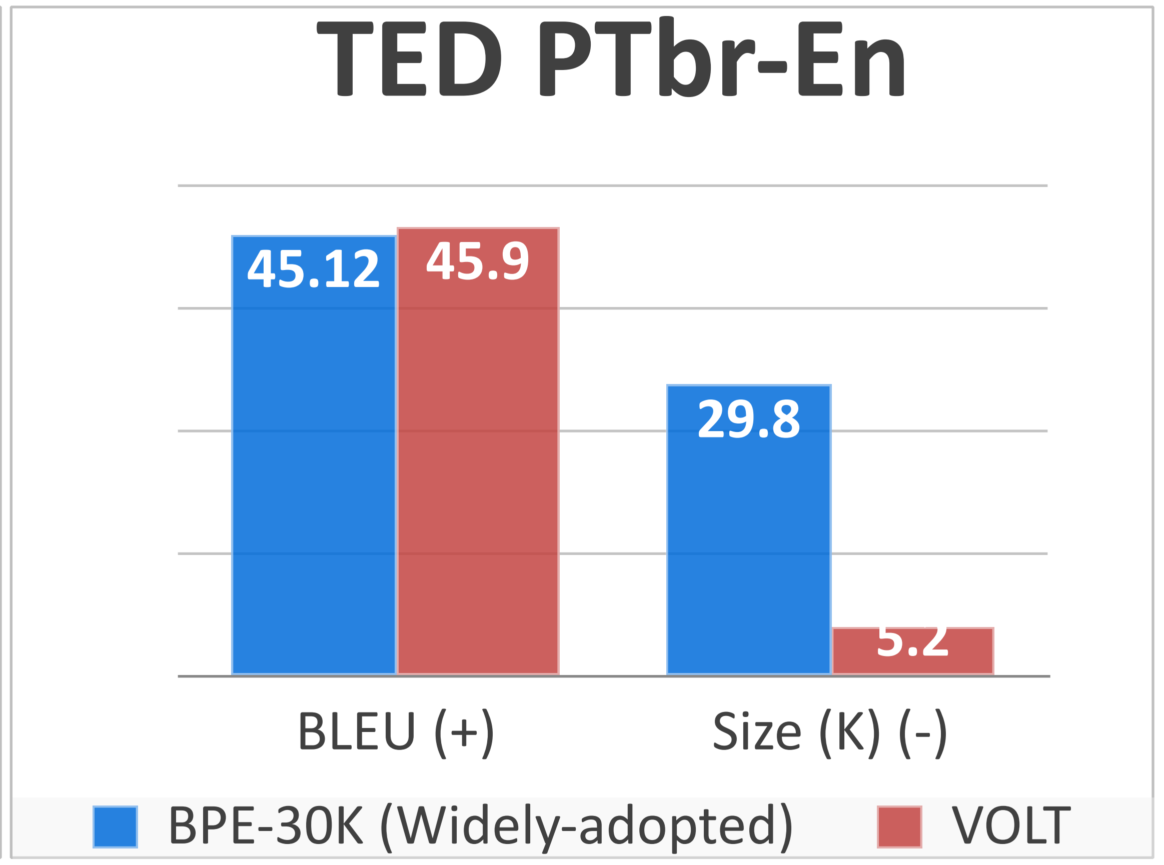
Xu, Zhou, Gan, Zheng, Li. Vocabulary Learning via Optimal Transport for Neural Machine Translation. ACL 2021.

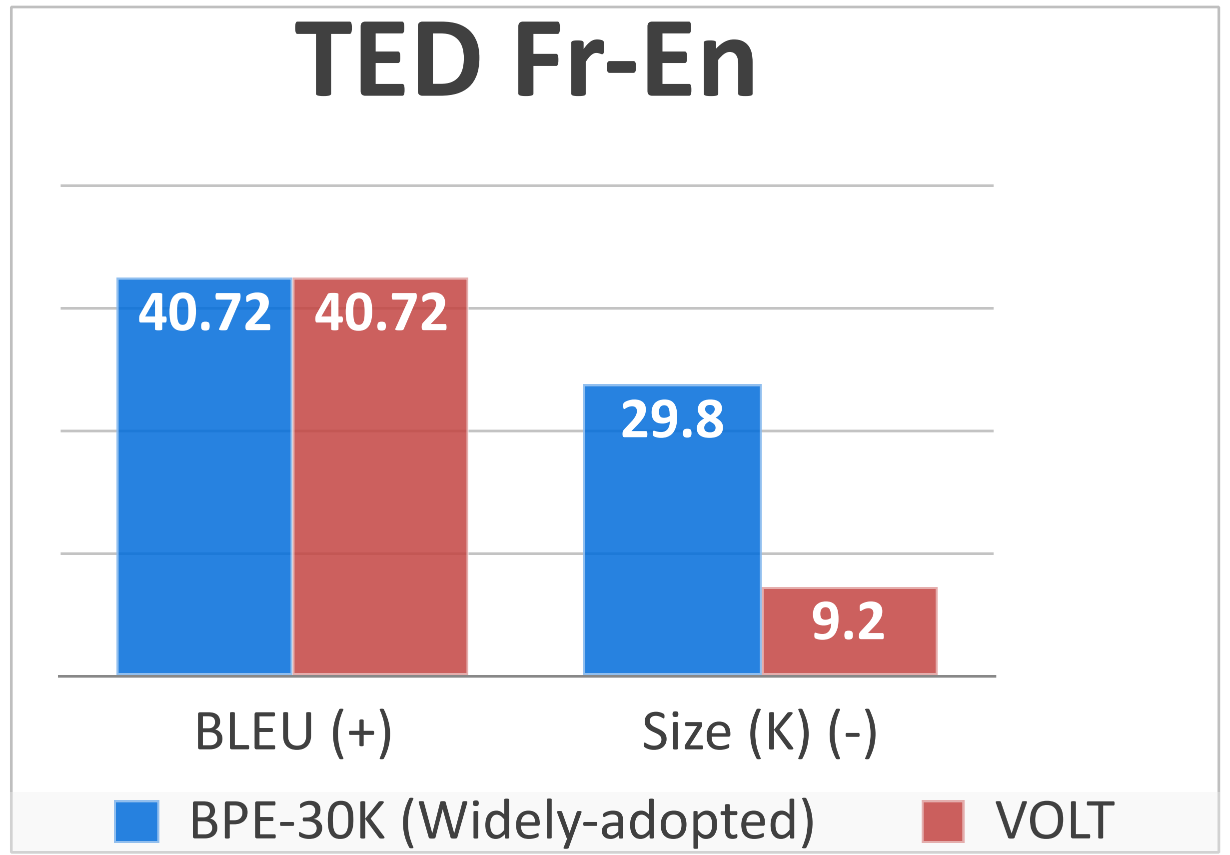
VOLT finds better vocabulary on Bilingual MT

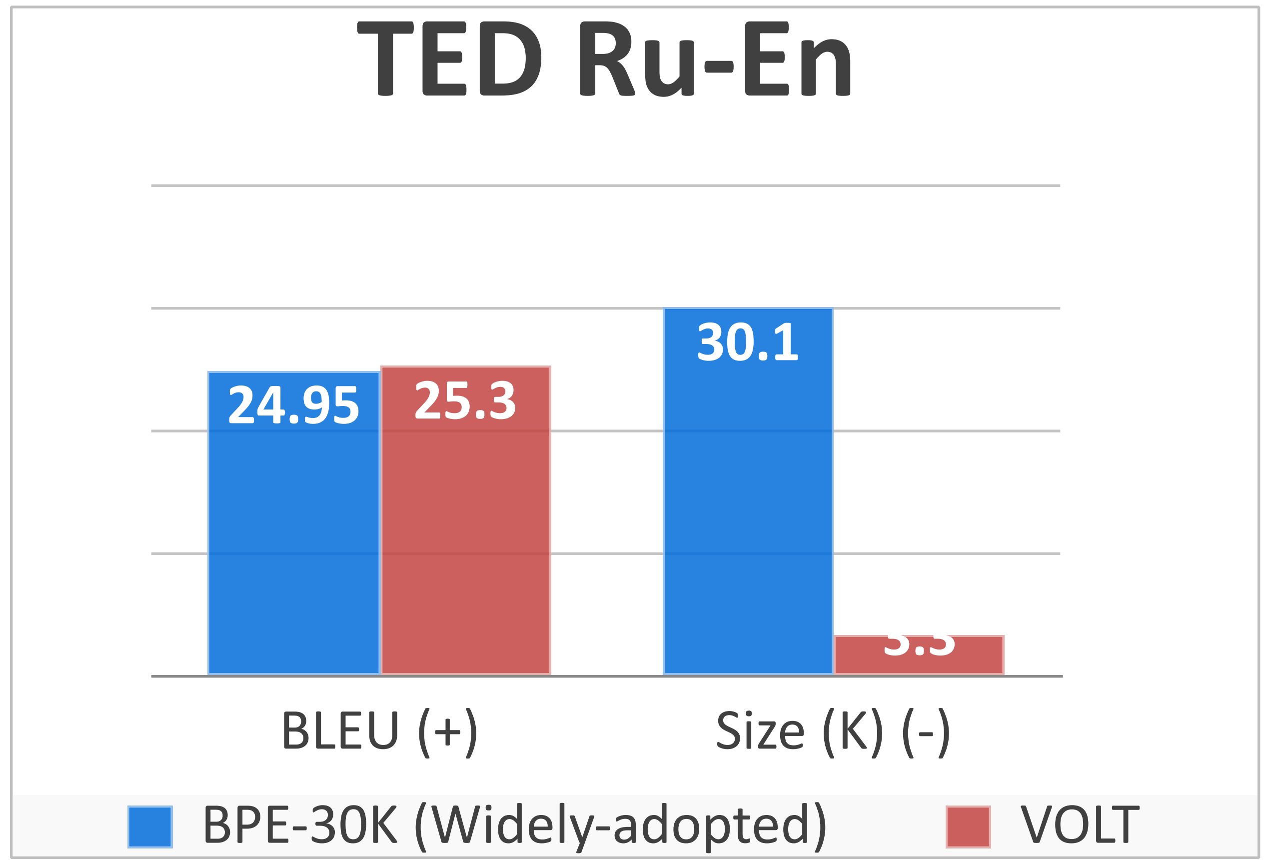
Transformer architecture

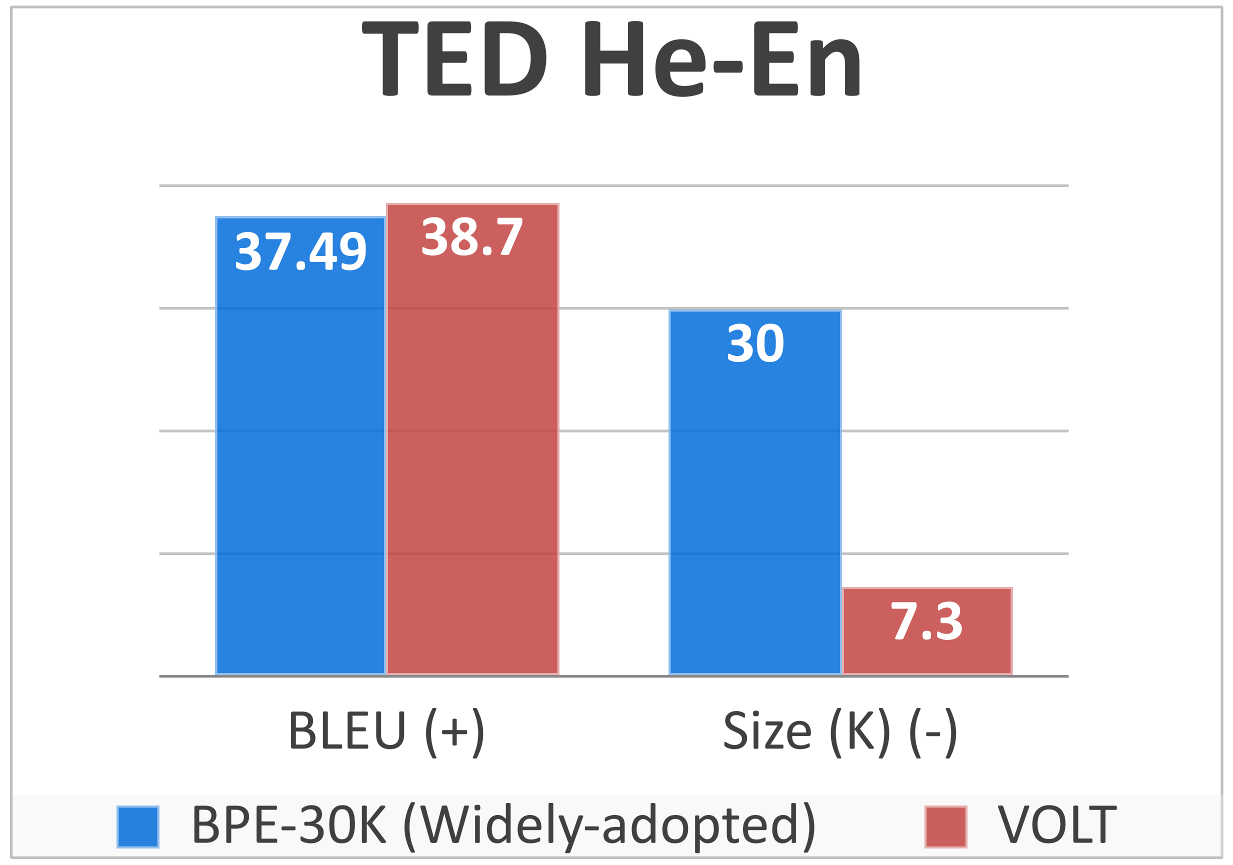


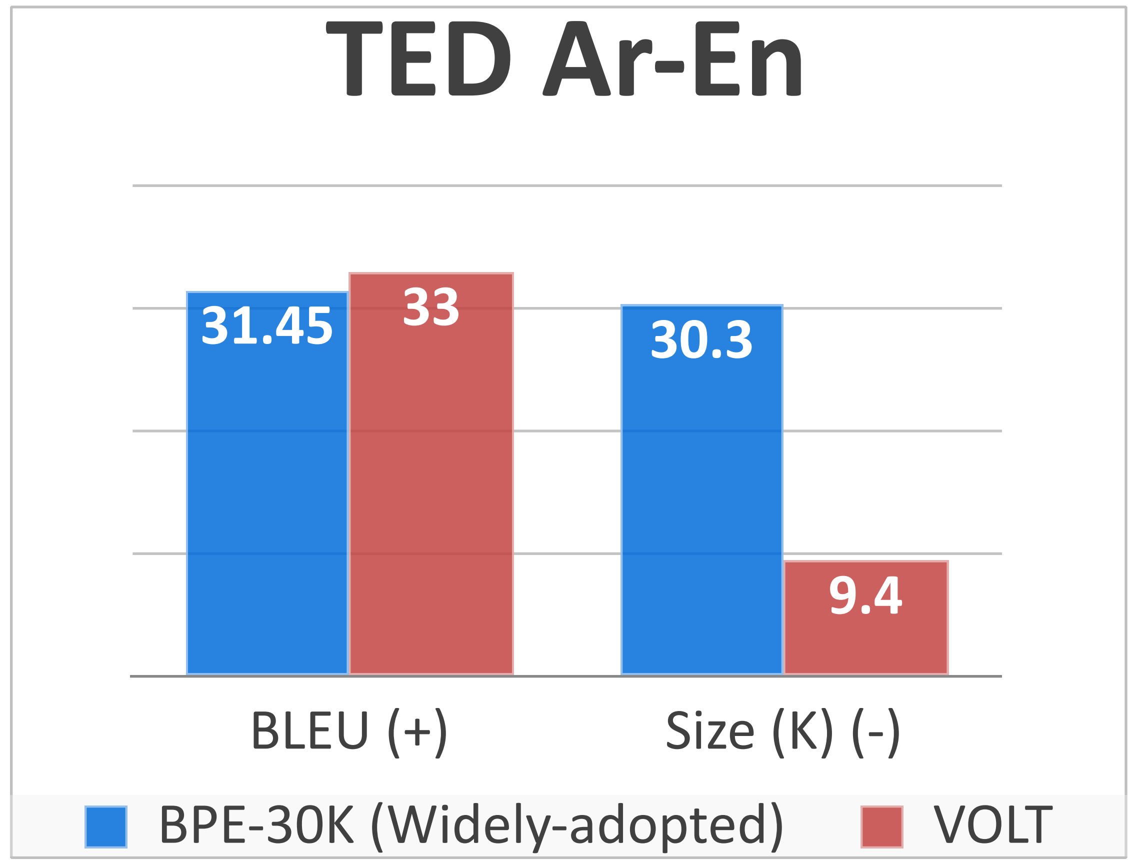


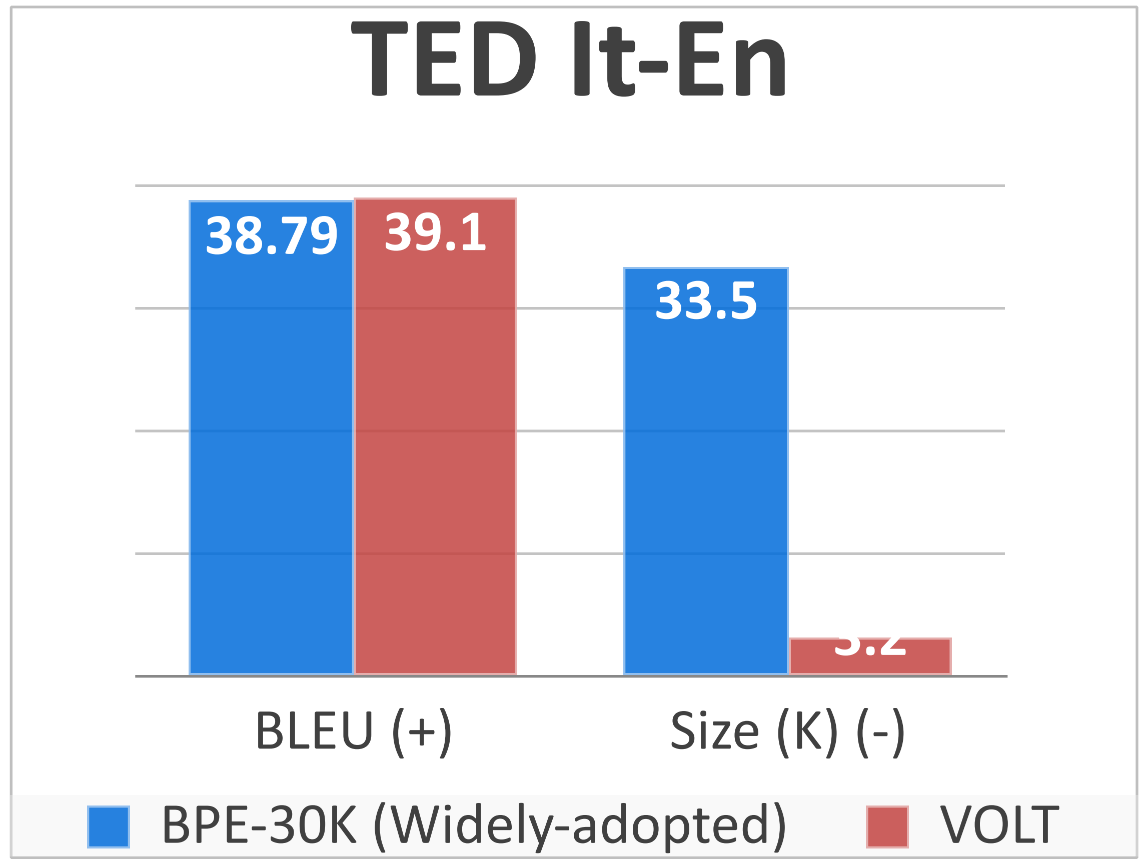






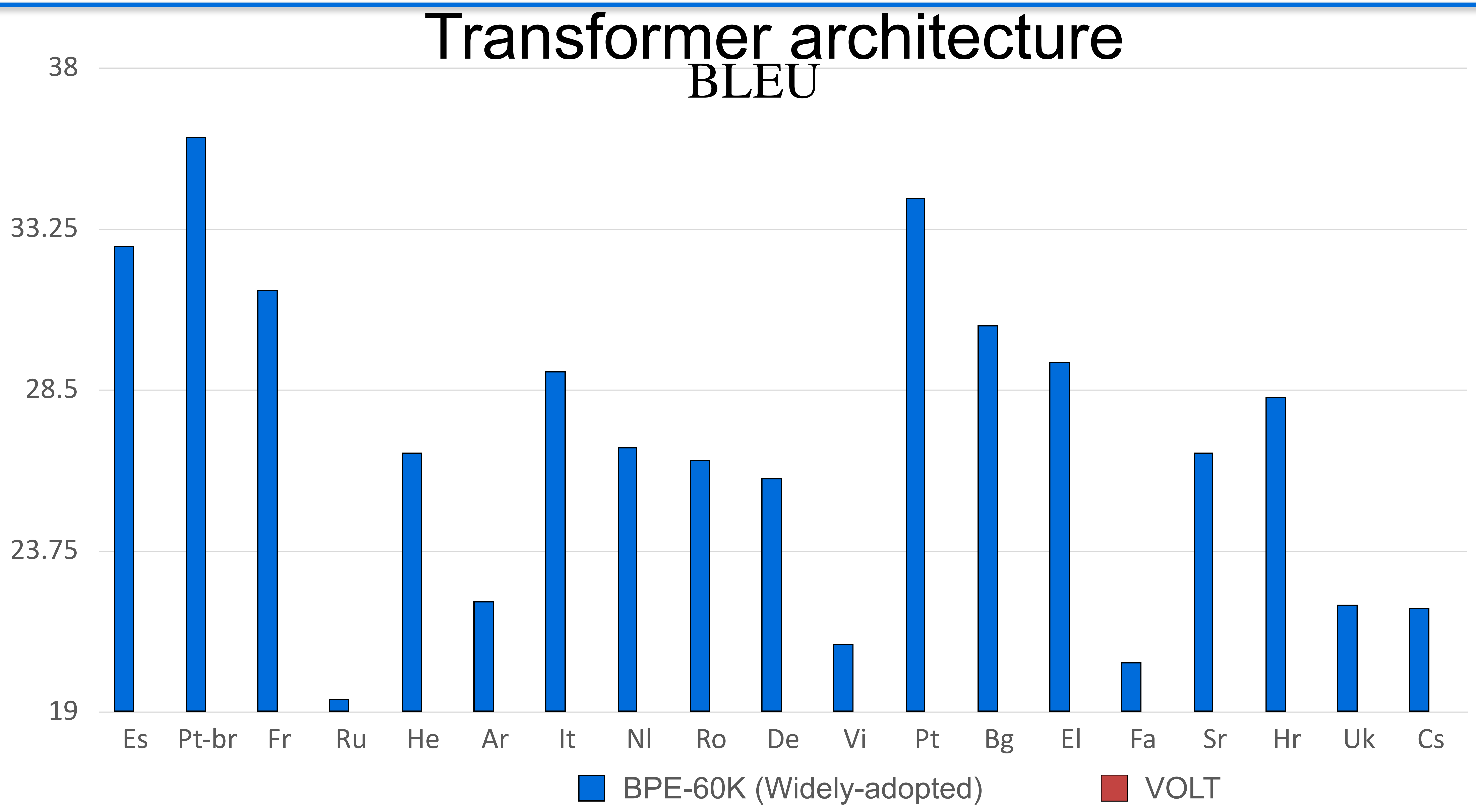






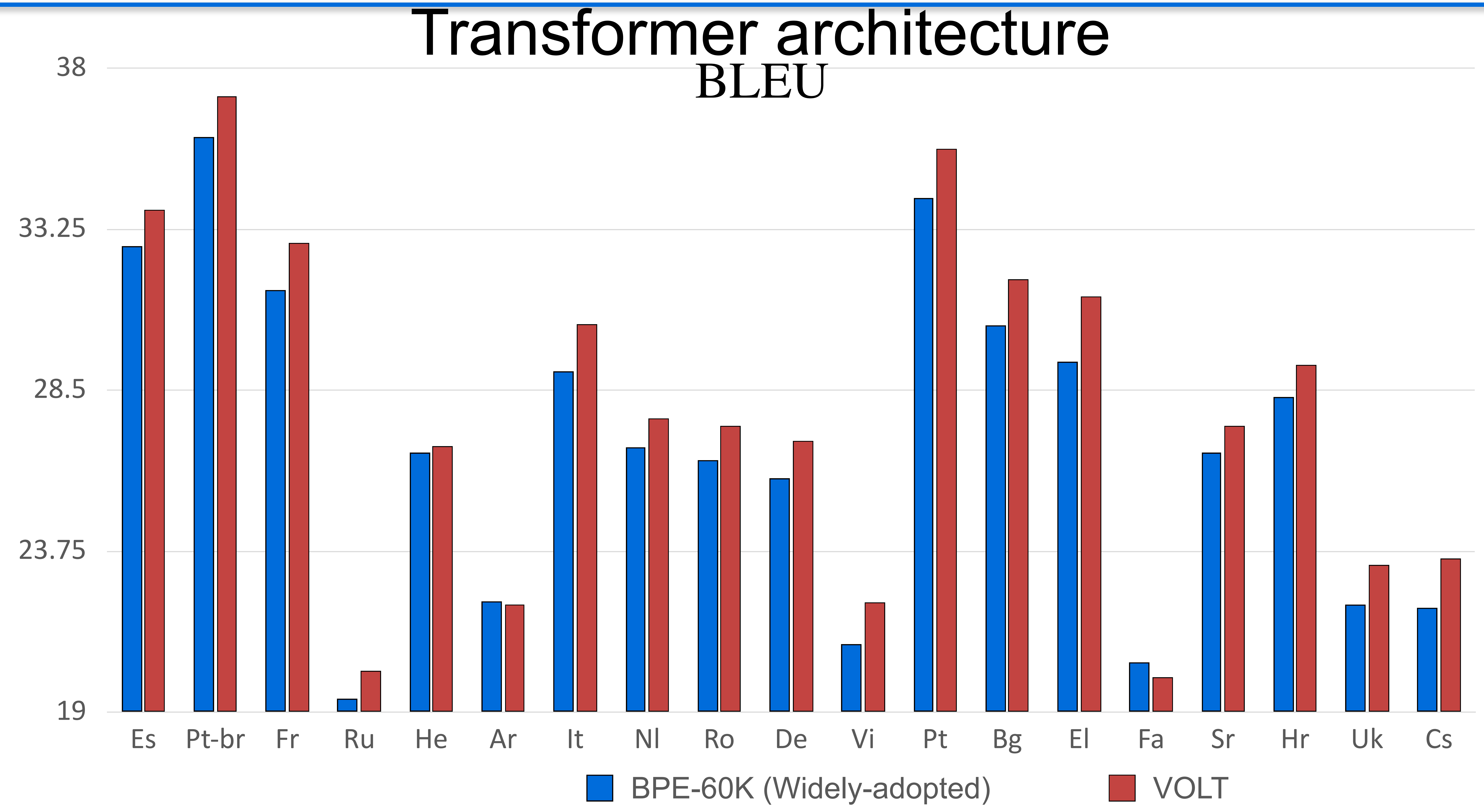
Xu, Zhou, Gan, Zheng, Li. Vocabulary Learning via Optimal Transport for Neural Machine Translation. ACL 2021.

VOLT Finds Better Vocabulary on Multilingual MT



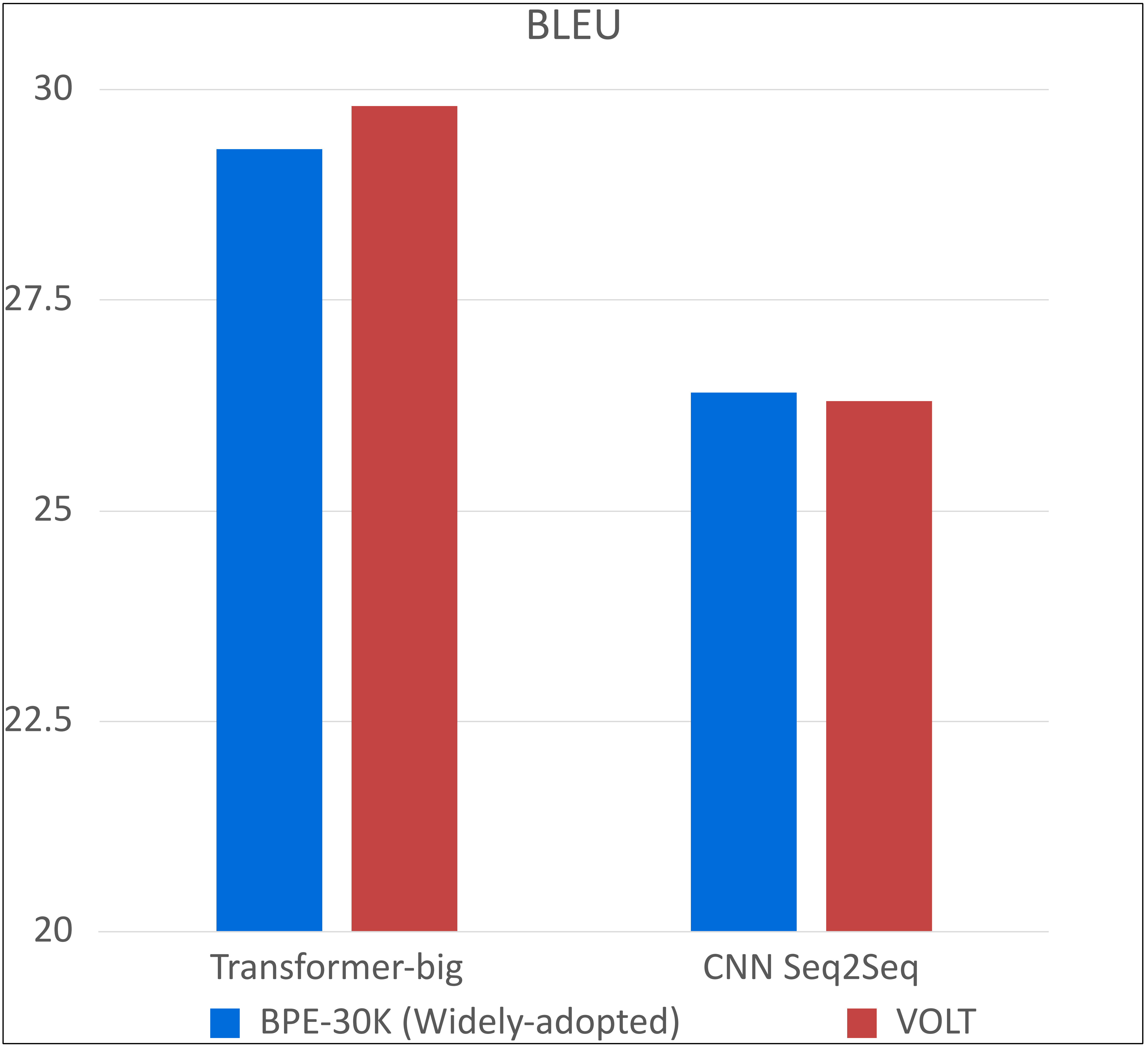
Xu, Zhou, Gan, Zheng, Li. Vocabulary Learning via Optimal Transport for Neural Machine Translation. ACL 2021.

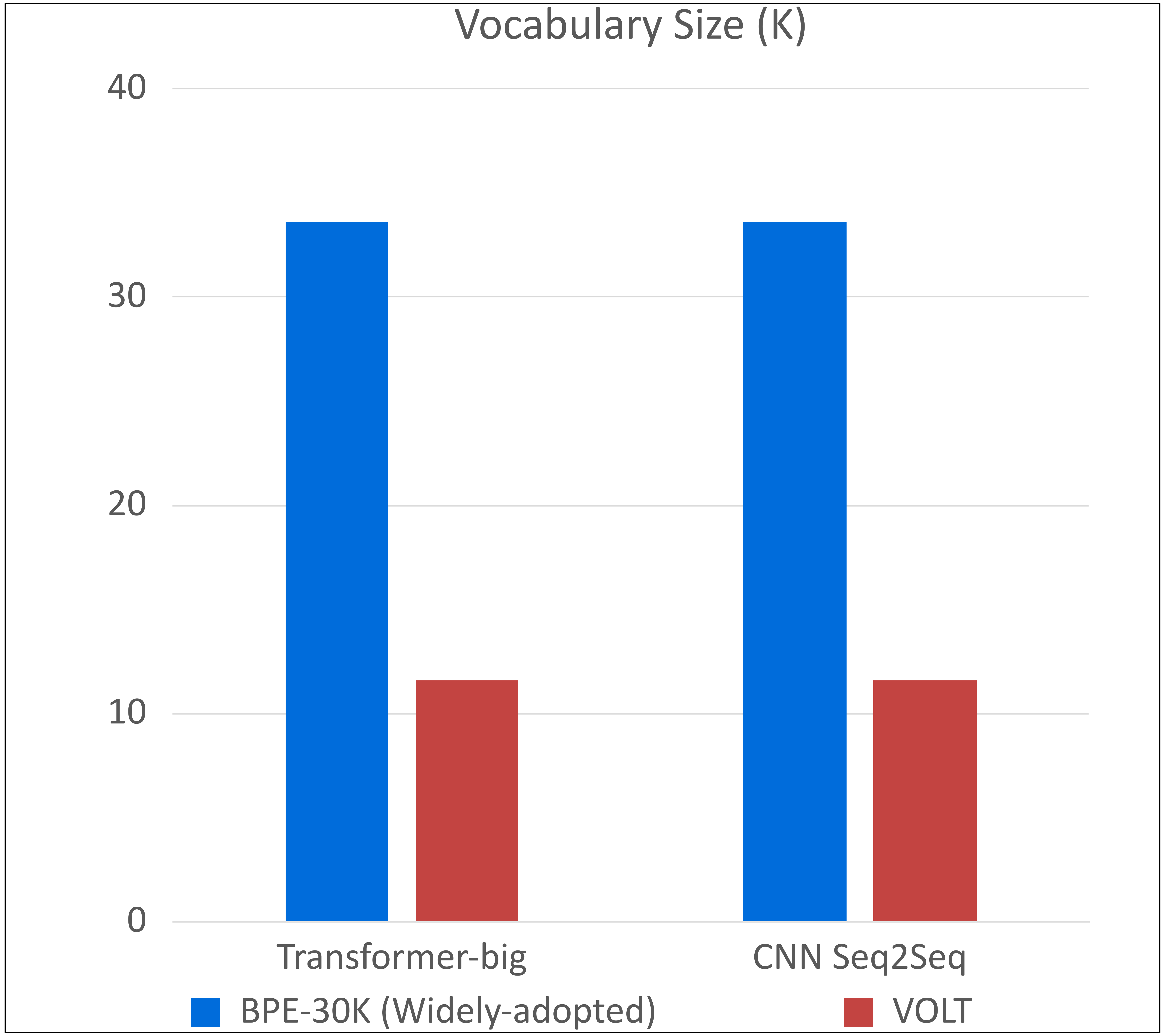
VOLT Finds Better Vocabulary on Multilingual MT



Xu, Zhou, Gan, Zheng, Li. Vocabulary Learning via Optimal Transport for Neural Machine Translation. ACL 2021.

VOLT Generalizes Well to Other Architectures

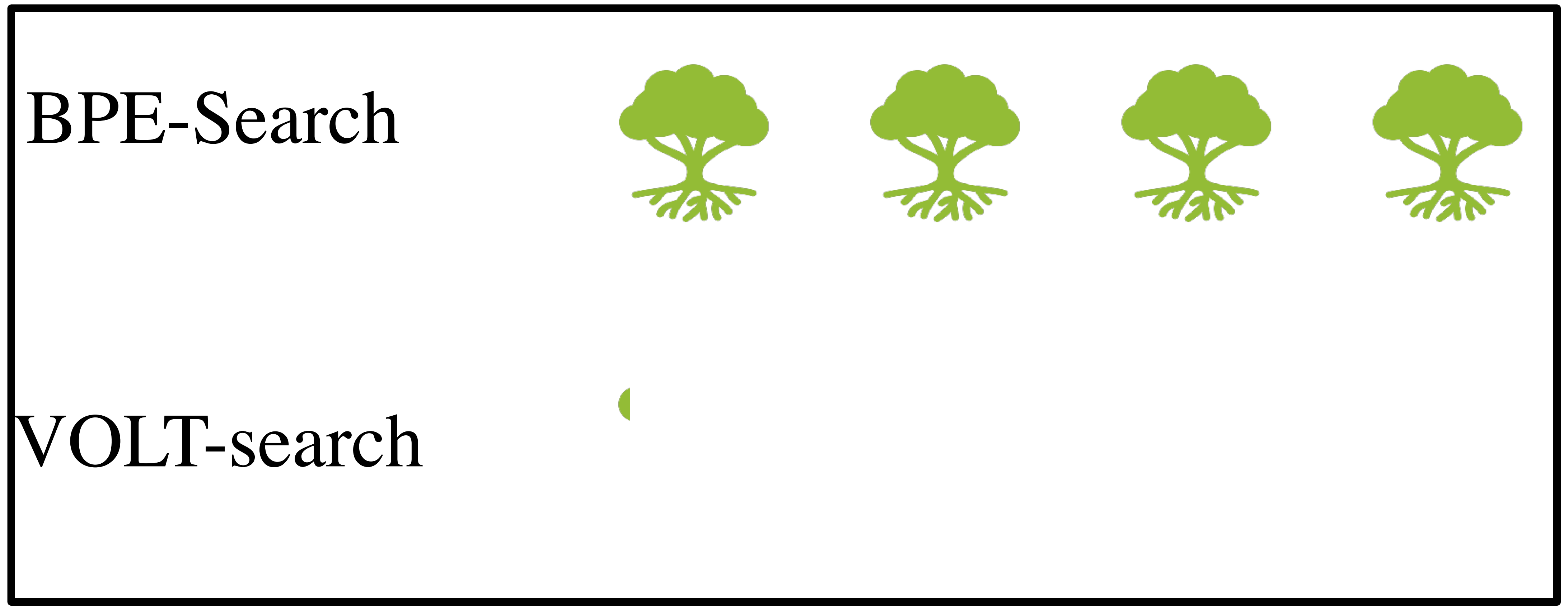




Xu, Zhou, Gan, Zheng, Li. Vocabulary Learning via Optimal Transport for Neural Machine Translation. ACL 2021.

VOLT: A Green Vocabulary Learning Solution

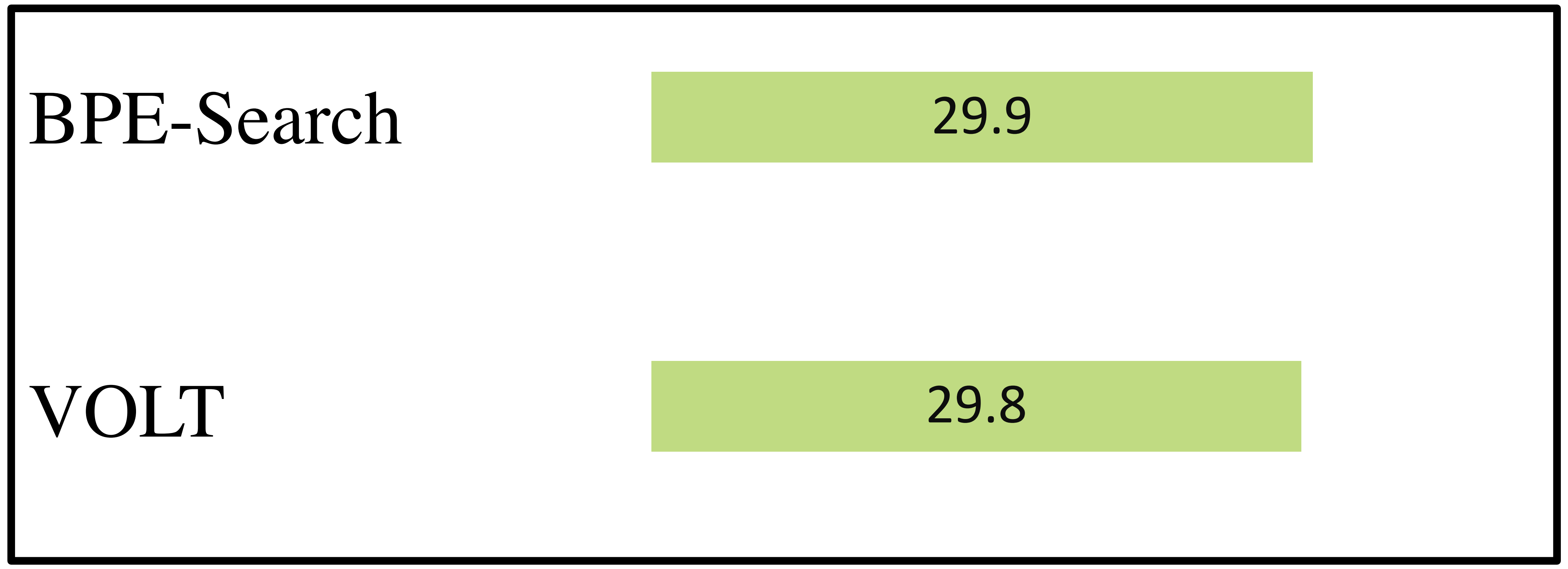
Carbon Emission



384 GPU hours

0.5 CPU hours

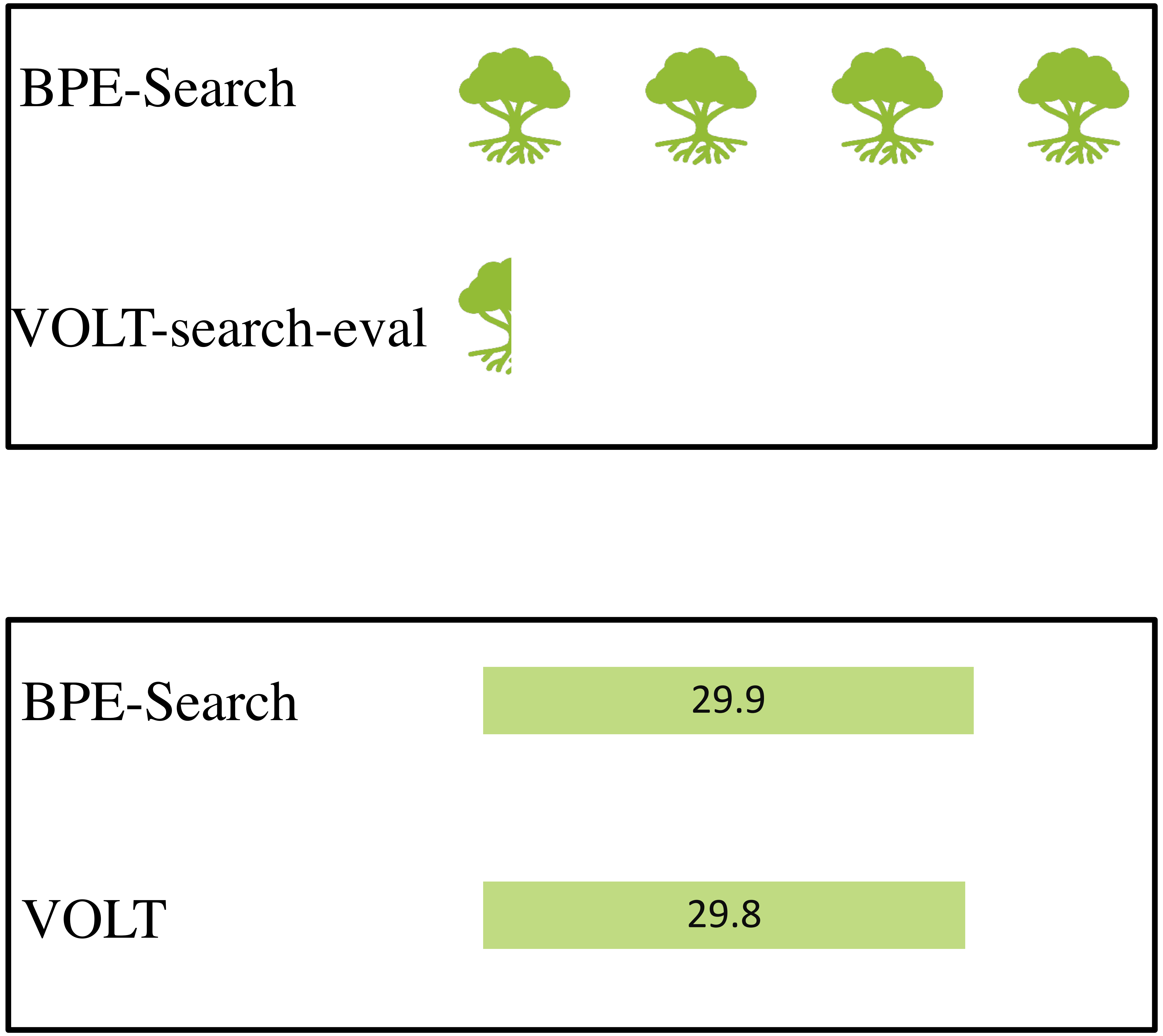
BLEU



Xu, Zhou, Gan, Zheng, Li. Vocabulary Learning via Optimal Transport for Neural Machine Translation. ACL 2021.

VOLT: A Green Vocabulary Learning Solution

Carbon Emission



BLEU

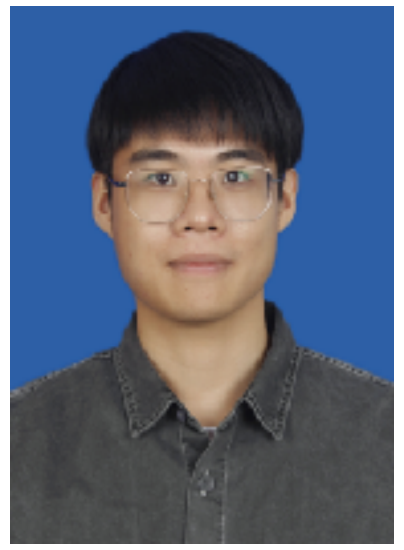
384 GPU hours

0.5 CPU hours + 30GPU hours

How to reduce this?

Glancing Transformer for Non-autoregressive Neural Machine Translation

Joint w/ Lihua Qian, Hao Zhou, Yu Bao, Mingxuan Wang, Lin Qiu, Weinan Zhang, Yong Yu







Transformer is Autoregressive

- Autoregressive models generate sentences sequentially

很多树 $\longrightarrow$ a lot of trees

- The conditional probability is factorized successively

$$p(Y|X;\theta) = \prod_{t=1}^{T} p(y_t|y_{<t}, X;\theta)$$

- Human-style translation is slow. Machine does not have to mimic human!

Wild idea: Parallel Generation?

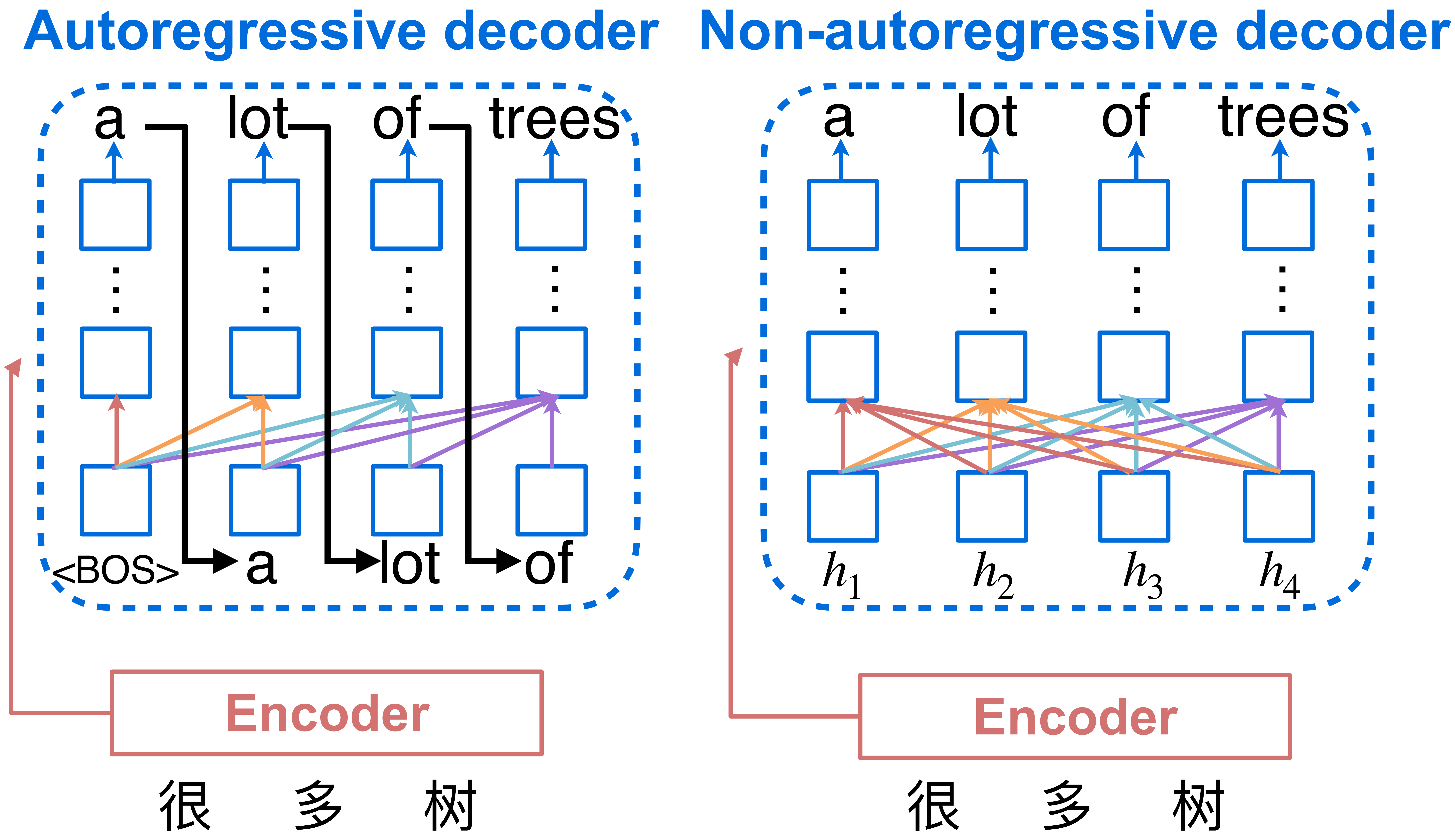
- Non-autoregressive models generate all the tokens in parallel

很多树 $\Rightarrow$ a lot of trees

- Conditional independence assumption

$$p(Y|X;\theta) = \prod_{t=1}^{T} p(y_t|X;\theta)$$

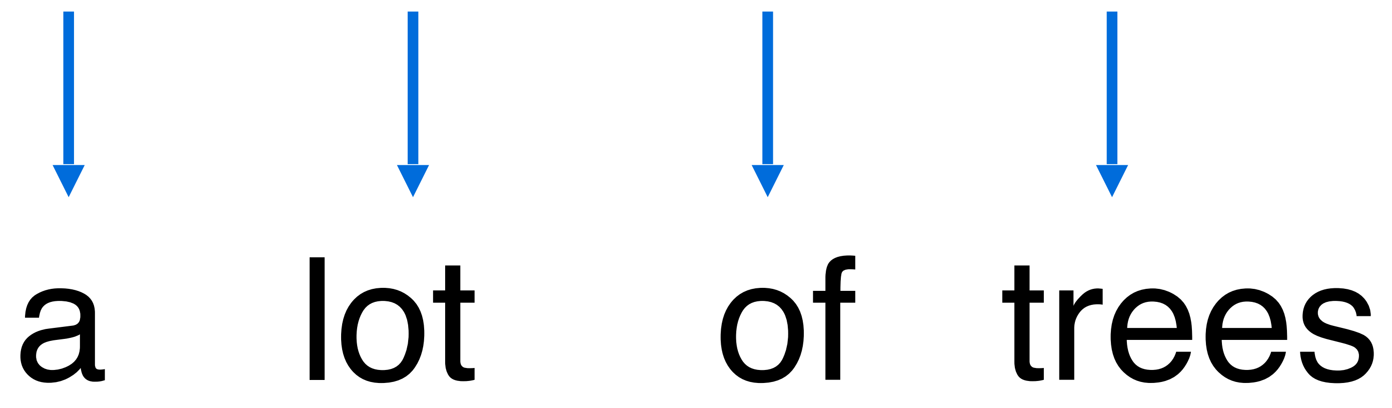
Model architecture



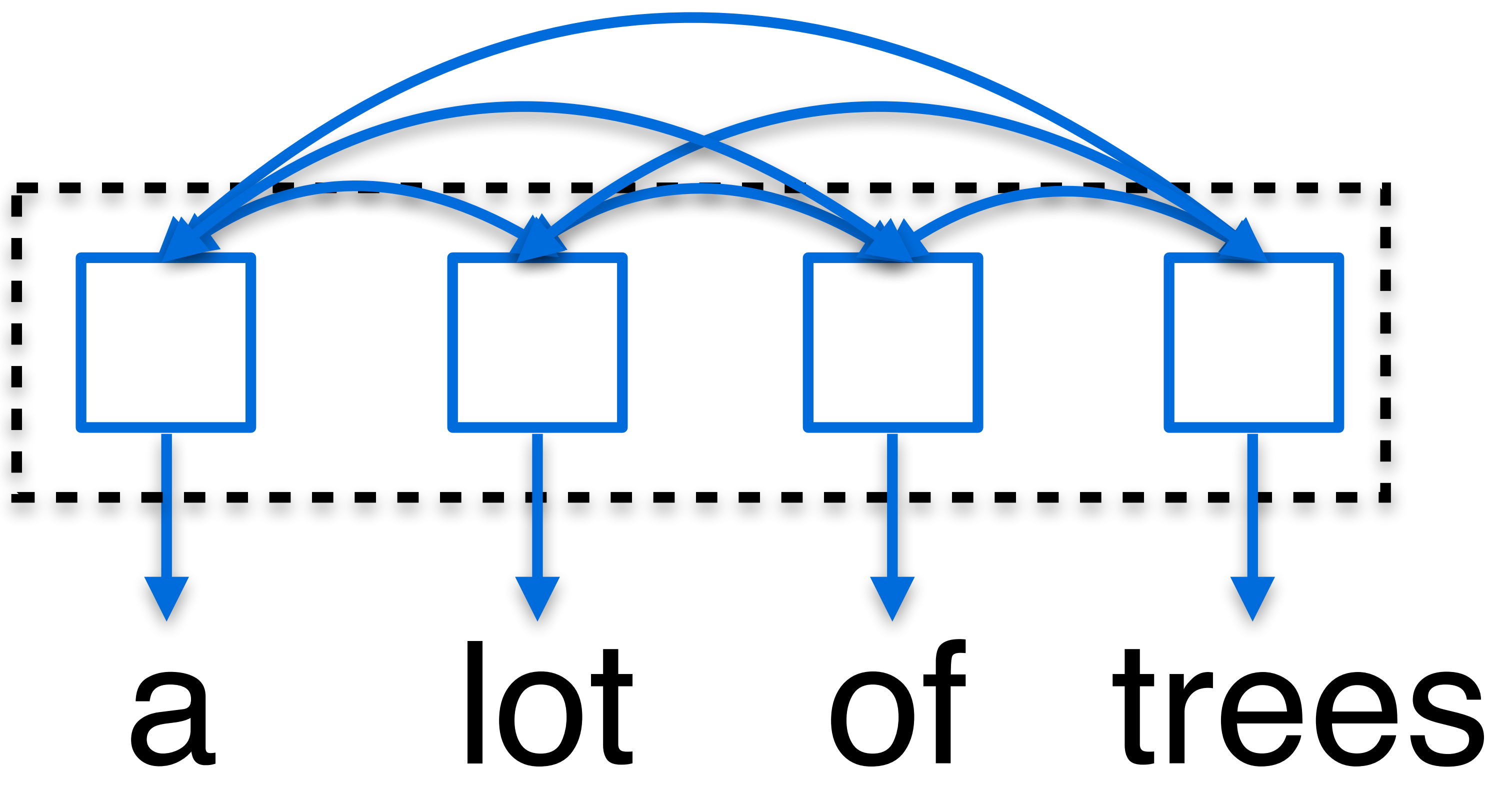
Gu et al, NAT, ICLR 2018

Why Non-autoregressive?

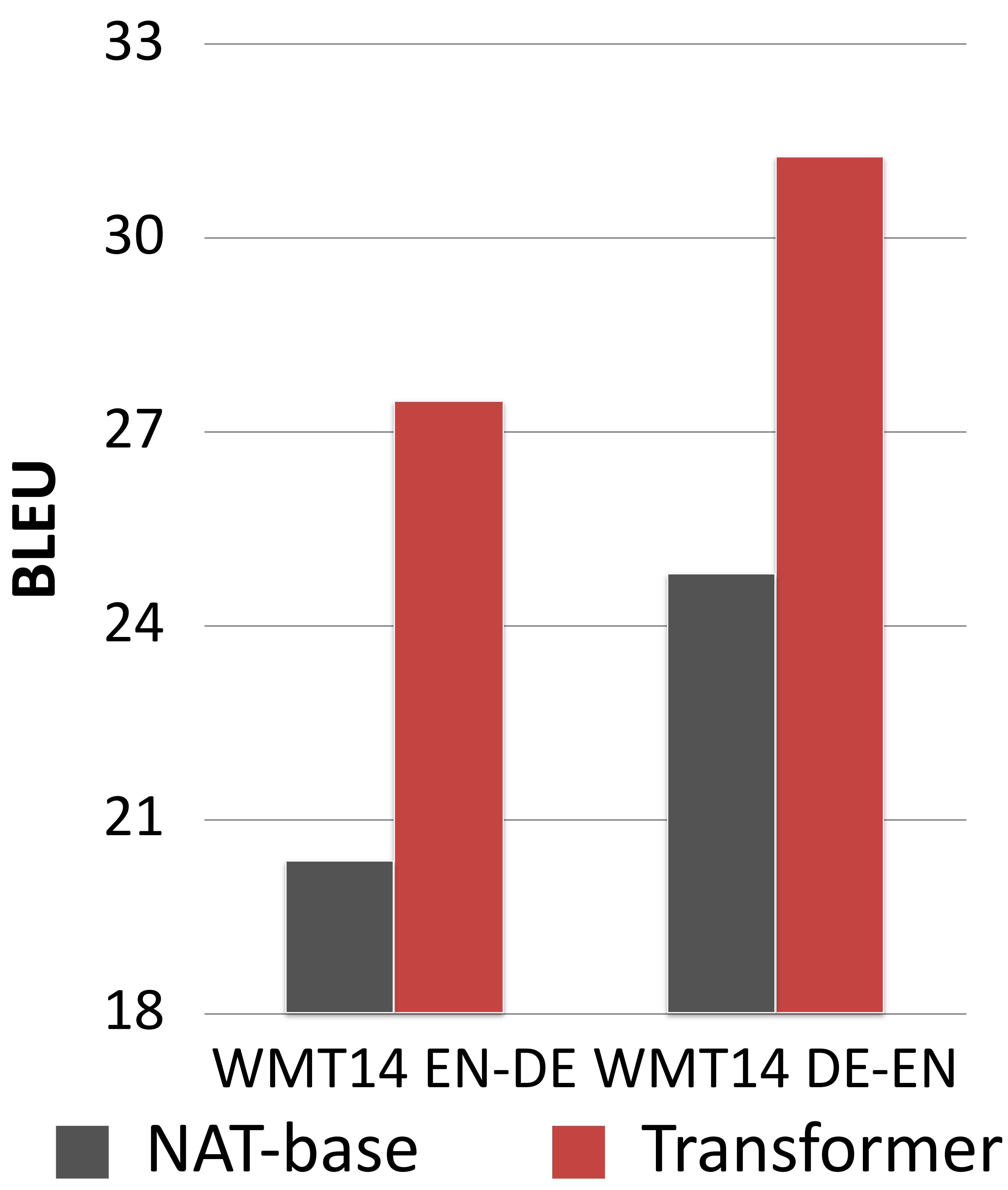
1. Faster decoding in non-autoregressive translation (NAT)



2. Capturing bidirectional context for generation



Challenge: Inferior Quality of NAT



- One input -> multiple target

很多树 → a lot of trees

很多树 → a great many trees

- Inconsistency problem in parallel generation

great many

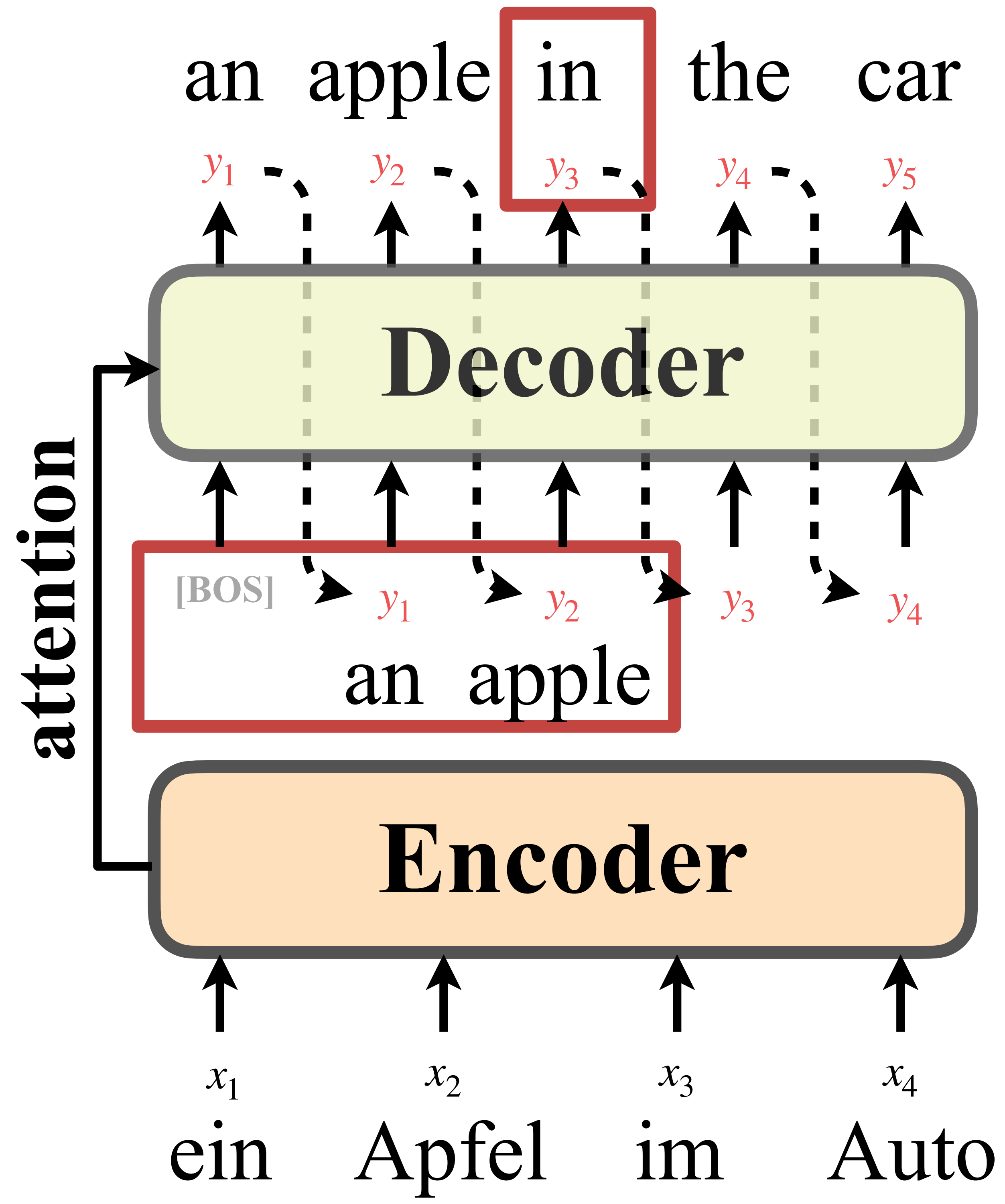
lot of

很多树 → a great of trees

Key Intuition: Word interdependency

- Learning word interdependency in the target sentence is crucial for generating fluent sentences
- Non-autoregressive models lack a effective way of dependency learning

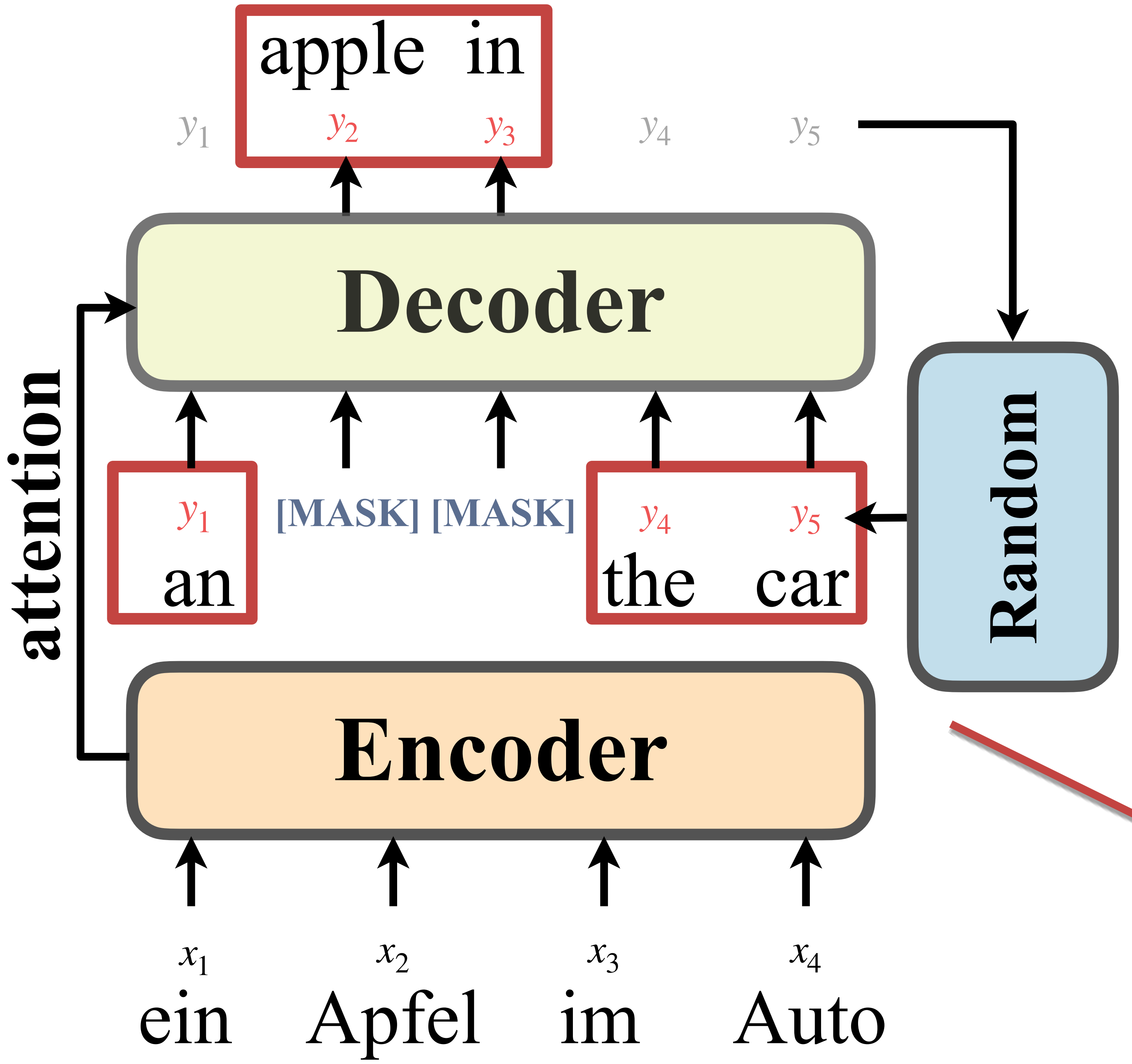
Learning Word Interdependency



Autoregressive models

- predict the next tokens conditioned on the input target tokens (left-to-right)

Learning Word Interdependency



Iterative-NAT

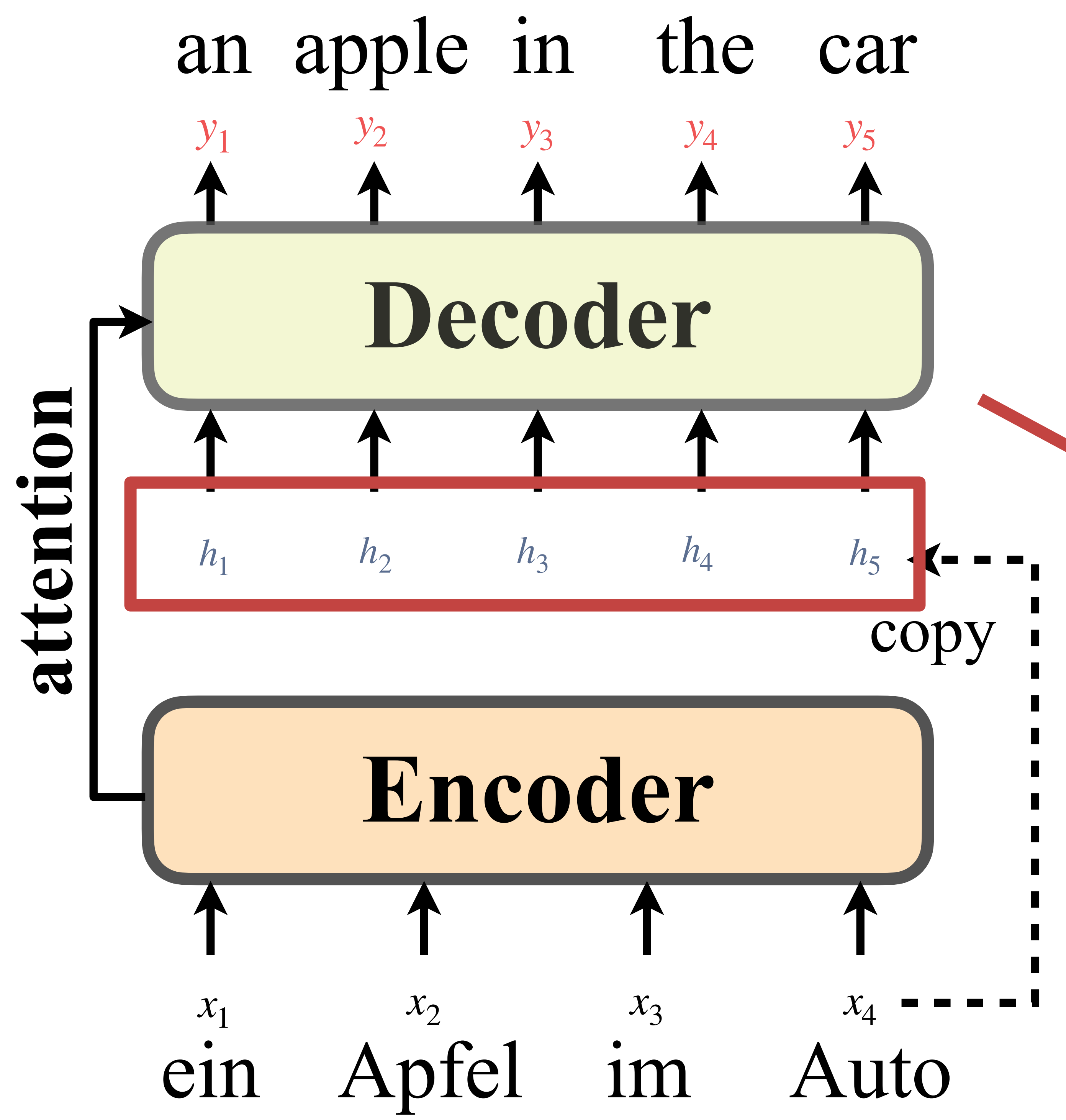
- predict the randomly masked tokens based on unmasked tokens

rely on multiple decoding iterations, therefore does not gain speedup!

Lee et al. Deterministic Non-Autoregressive Neural Sequence Modeling by Iterative Refinement. EMNLP 2018.

Ghazvininejad et al. Mask-Predict: Parallel Decoding of Conditional Masked Language Models. EMNLP 2019.

New Idea for Dependency learning



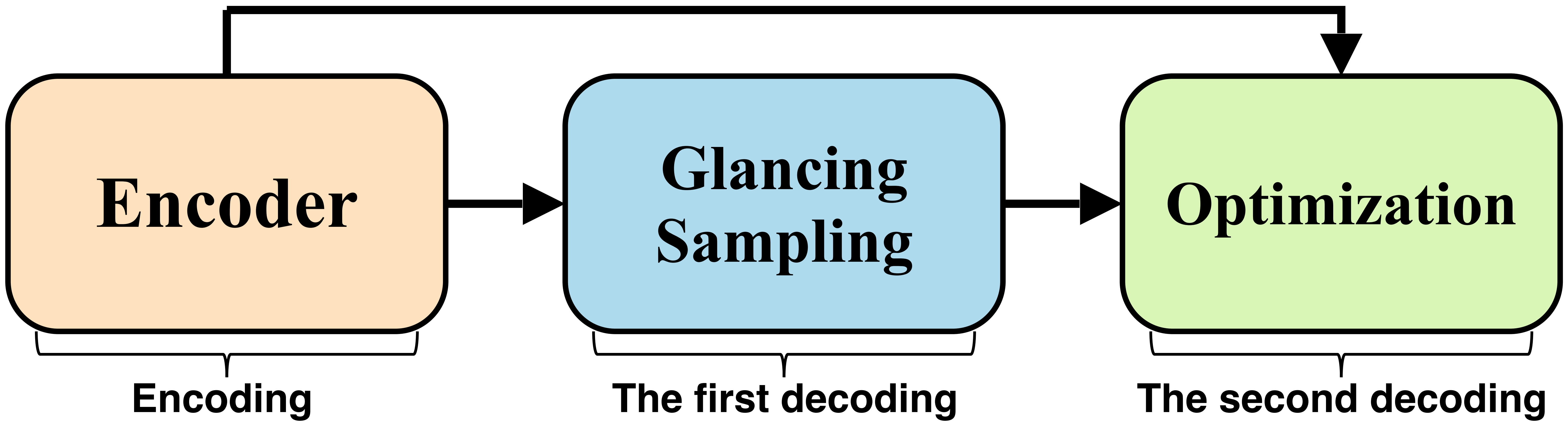
$$L_\theta = -\log p(Y|X;\theta)$$

Lack explicit target word interdependency learning

- Glancing Language Model (GLM)
  - A gradual training method
  - Learning word interdependency for **single-pass** parallel generation

Qian et al. Glancing Transformer for Non-autoregressive Neural Machine Translation. ACL 2021.

Glancing Language Model



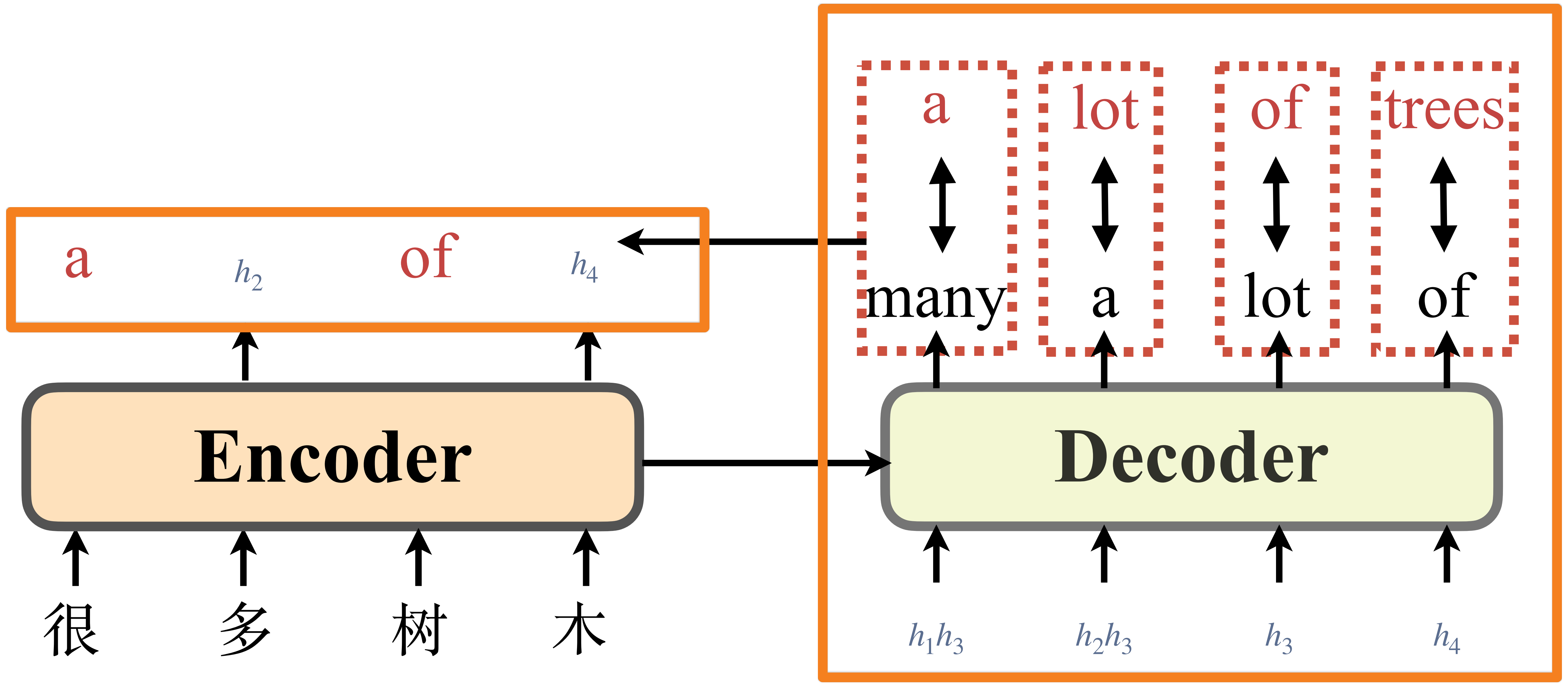
only one-pass decoding in inference

- Perform two decoding during training

1. Glancing Sampling (the first decoding):
   - Based on the prediction, replace part of the decoder inputs with sampled target words
2. Optimization (the second decoding):
   - Learn to predict the remaining words with the replaced decoder inputs

Qian et al. Glancing Transformer for Non-autoregressive Neural Machine Translation. ACL 2021.

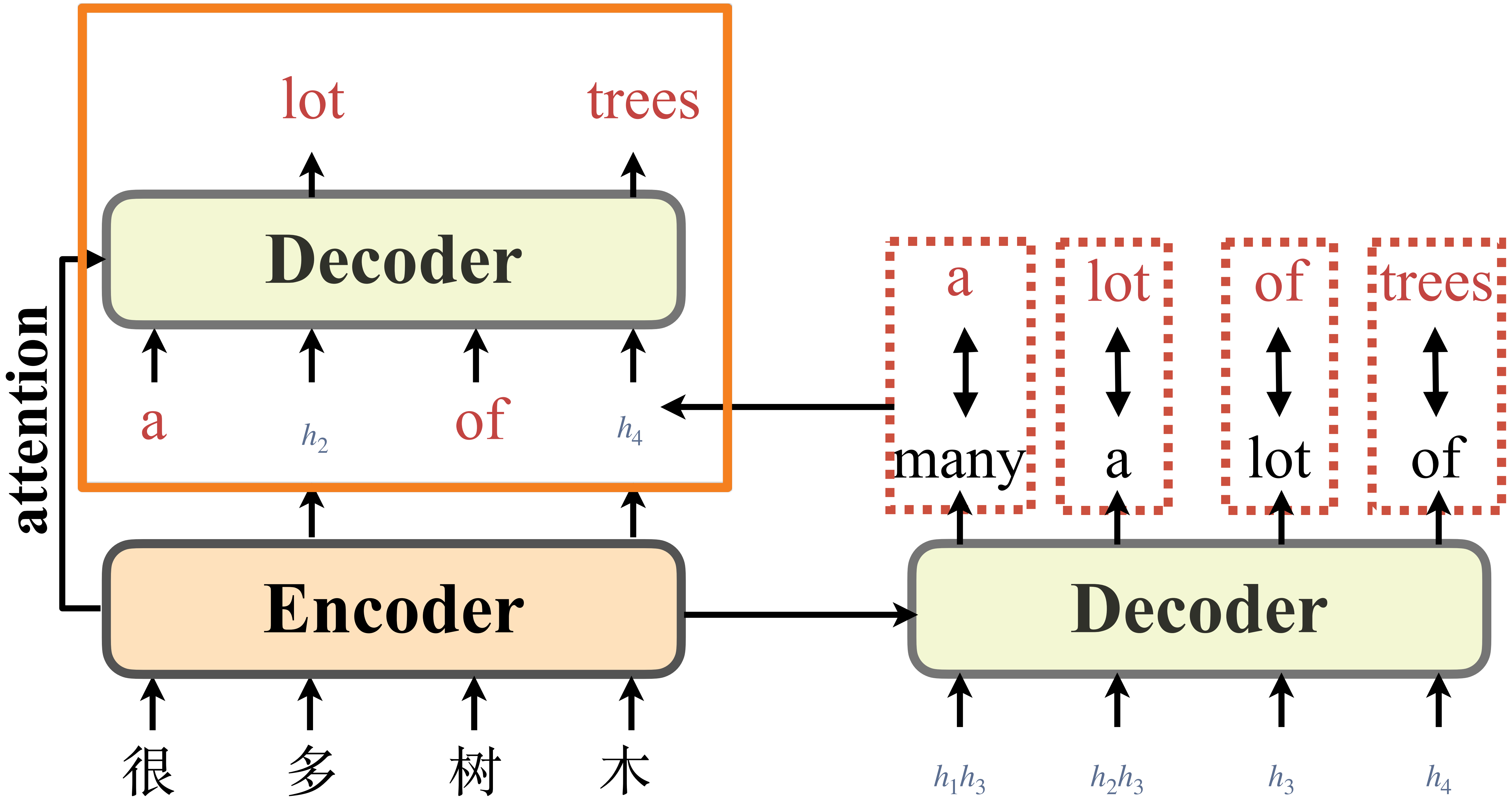
Glancing: Learning Dependency Gradually



- Based on the prediction, replace part of the decoder inputs with sampled target words

Qian et al. Glancing Transformer for Non-autoregressive Neural Machine Translation. ACL 2021.

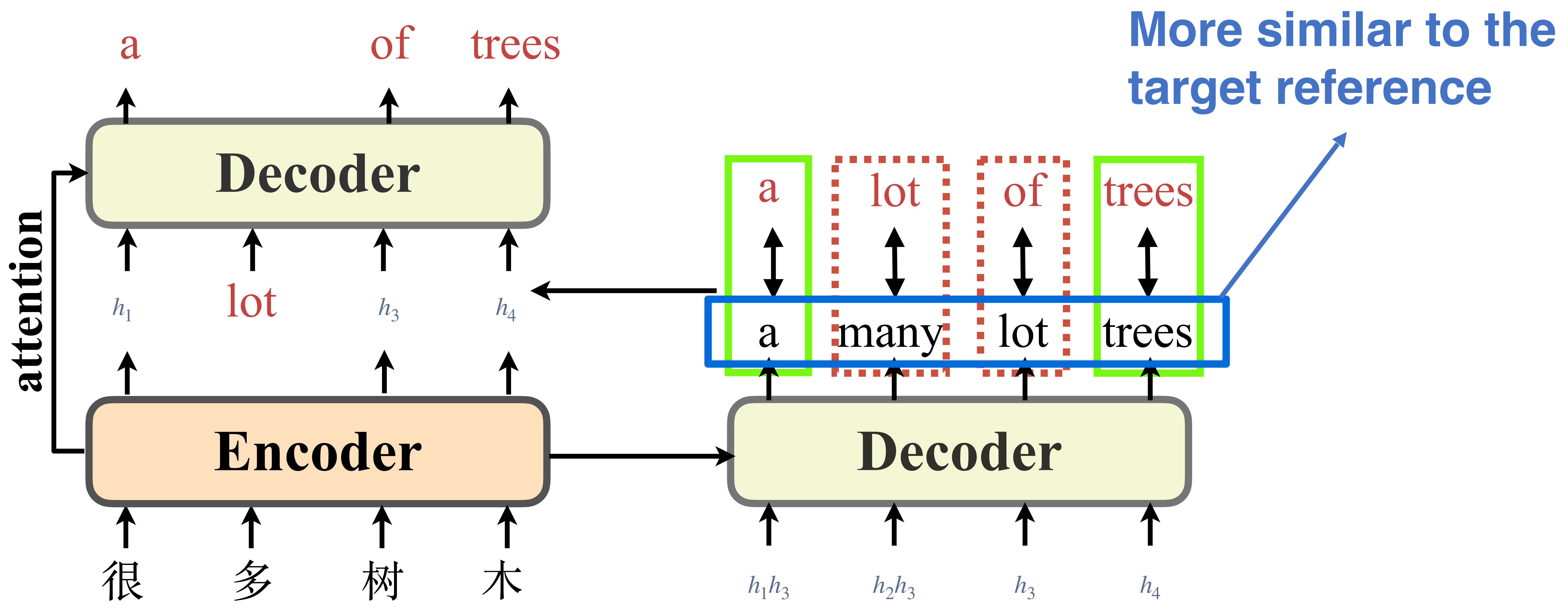
Glancing: Learning Dependency Gradually



- Learn to predict the remaining words with the replaced decoder inputs

Qian et al. Glancing Transformer for Non-autoregressive Neural Machine Translation. ACL 2021.

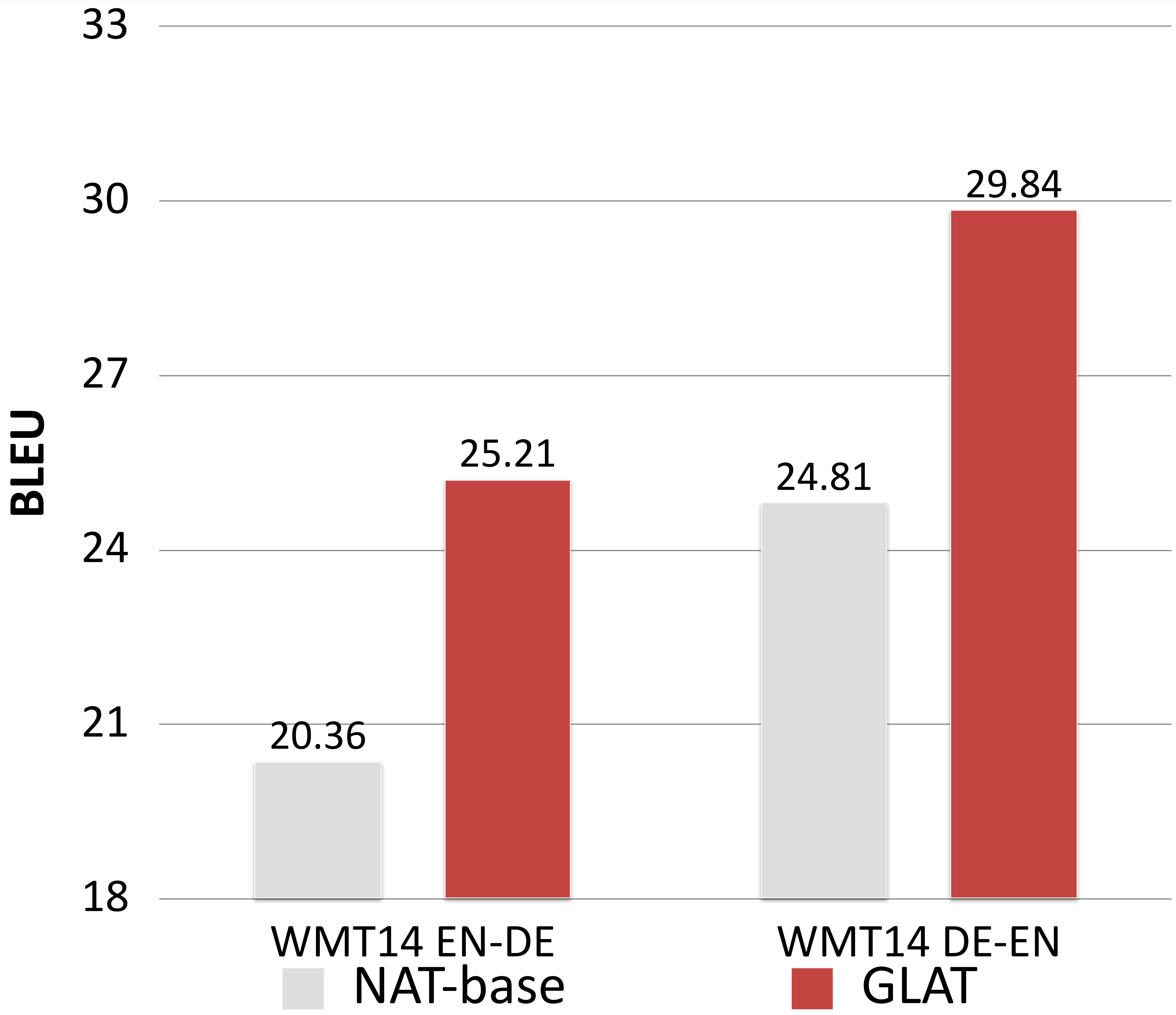
Glancing: Learning Dependency Gradually



During training, the sampling number of target words decreases gradually.

Qian et al. Glancing Transformer for Non-autoregressive Neural Machine Translation. ACL 2021.

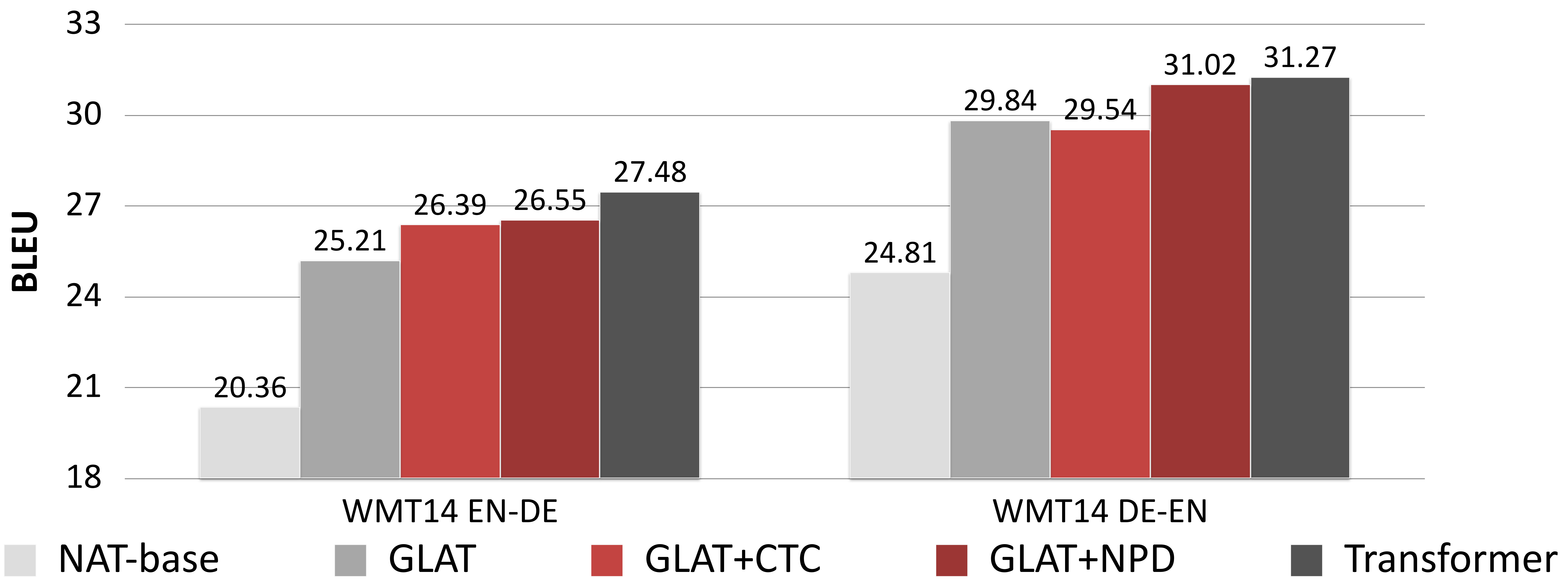
GLAT boosts Translation Quality significantly!



+ 5 BLEU!

Qian et al. Glancing Transformer for Non-autoregressive Neural Machine Translation. ACL 2021.

GLAT approaches Transformer quality!



- GLAT achieves high quality translation while keeping high inference speed-up (8x~15x)

Qian et al. Glancing Transformer for Non-autoregressive Neural Machine Translation. ACL 2021.

GLAT in Real Competition

GLAT achieve the Top score in WMT21 En-De and De-En!
The first NAT system to do so!

newstest2021.de-en test set (de-en)

#	Name	BLEU
1	**Anonymous submission #1276**	35.0
2	**Anonymous submission #1284**	35.0
3	Anonymous submission #1304	34.9
4	Anonymous submission #1117	34.9
5	Anonymous submission #1258	34.9
6	**Anonymous submission #1124**	34.9
7	Anonymous submission #543	34.8
8	**Anonymous submission #963**	34.8
9	Anonymous submission #861	34.7
10	Anonymous submission #738	34.7

BLEU and ChrF are sacreBLEU scores. Systems in **bold face** are your submissior validation errors denoted by -1.0 score.

newstest2021.en-de test set (en-de)

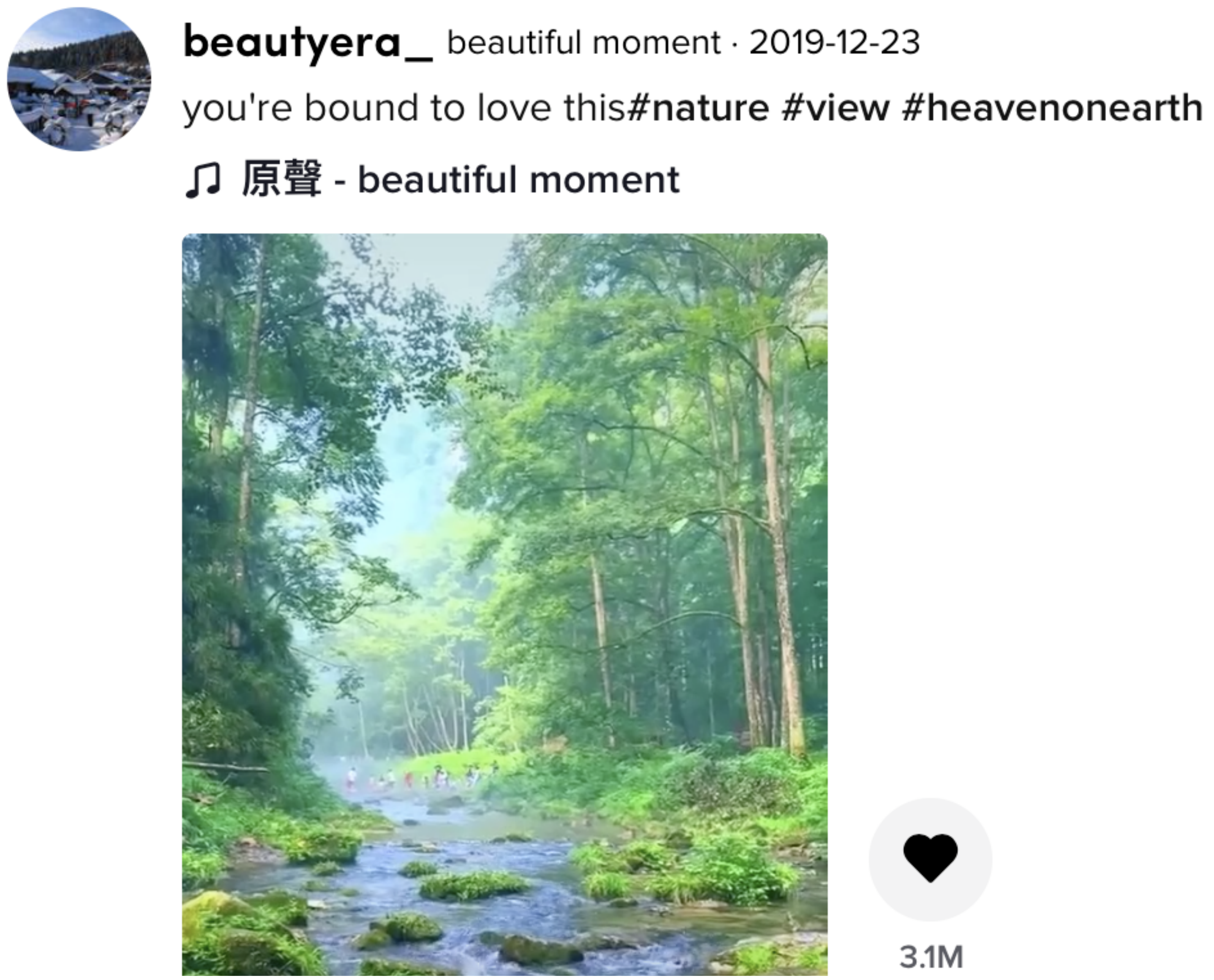
#	Name	BLEU
1	**Anonymous submission #1265**	31.3
2	Anonymous submission #1303	31.3
3	Anonymous submission #1291	31.3
4	Anonymous submission #804	31.3
5	Anonymous submission #368	31.3
6	Anonymous submission #1168	31.3
7	**Anonymous submission #1251**	31.2
8	**Anonymous submission #986**	31.2
9	Anonymous submission #1310	31.2
10	**Anonymous submission #1243**	31.2

BLEU and ChrF are sacreBLEU scores. Systems in **bold face** are your submissio validation errors denoted by -1.0 score.

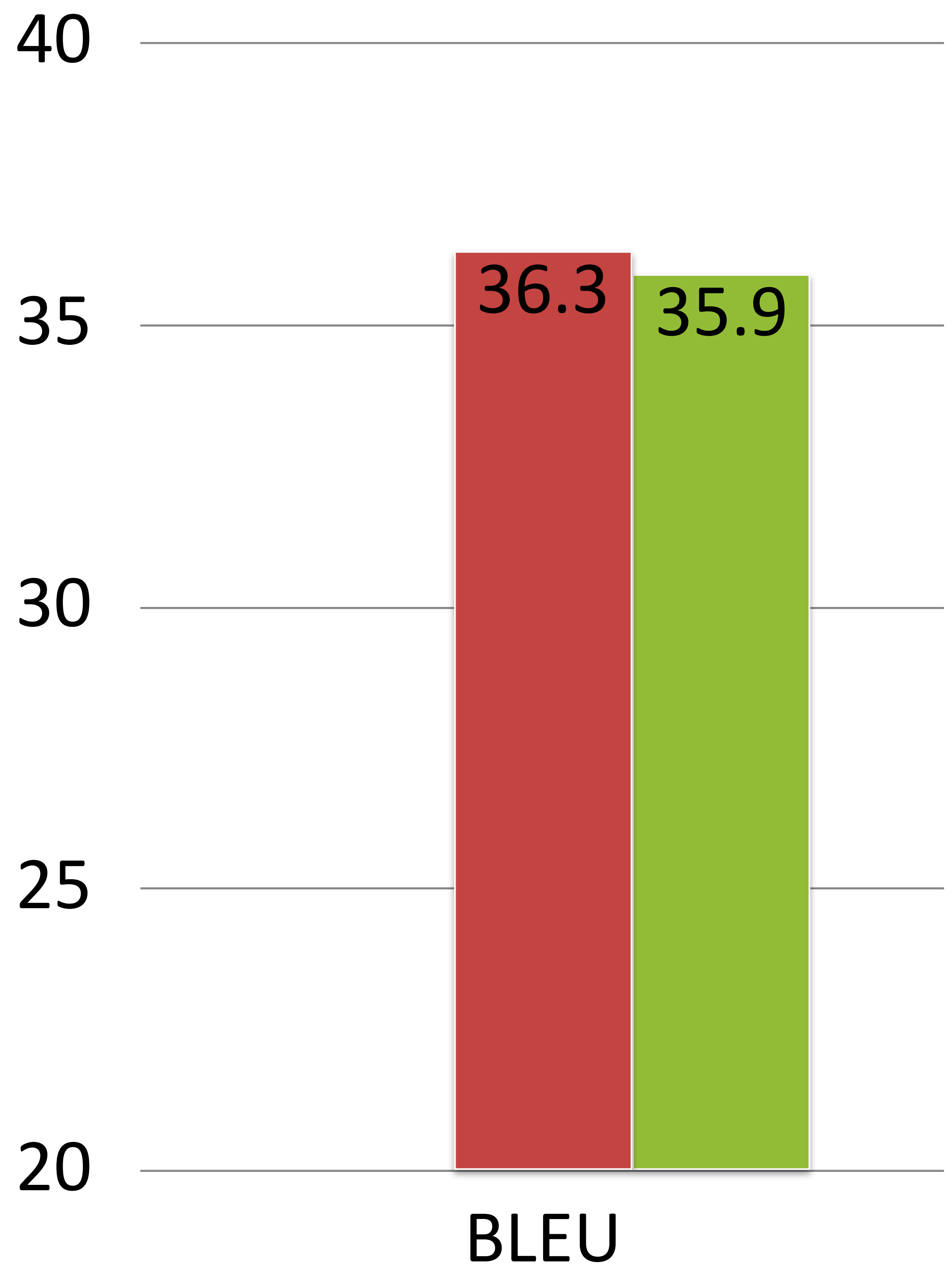
Qian et al. The Volctrans GLAT System: Non-autoregressive Translation Meets WMT21. 2021.

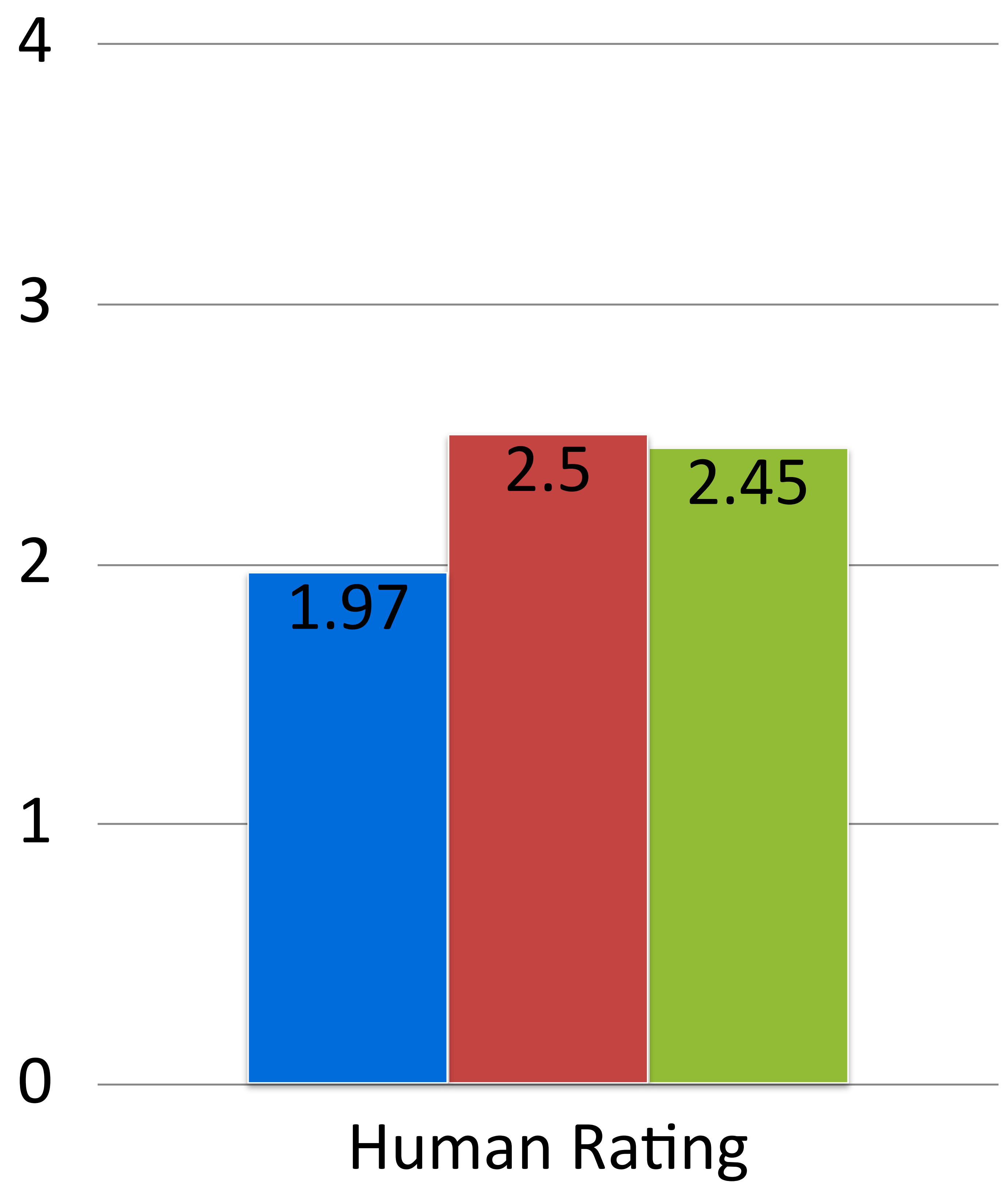
GLAT is the first production NAT system!

- Already deployed online in VolcTrans and serving English-Japanese



Tiktok caption translation



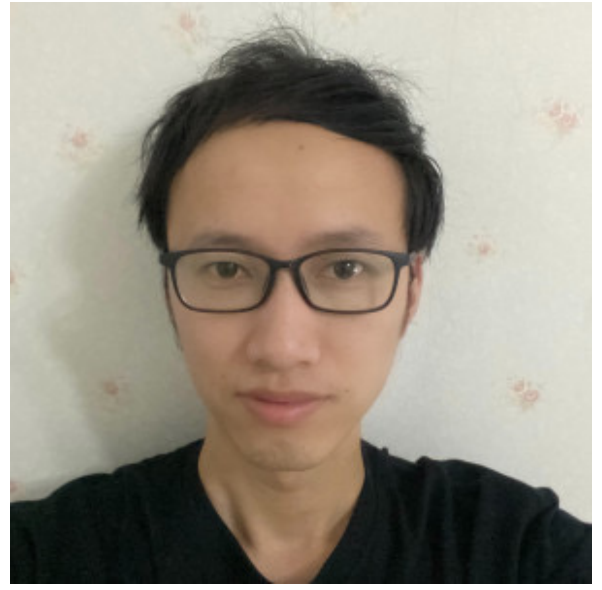


火山翻译
Volctrans

LightSeq: A High Performance Library for Transformers

Joint w/ Xiaohui Wang, Ying Xiong, Yang Wei, Xian Qian, Mingxuan Wang
and community contributors





Need for Hardware Acceleration

- What about Transformer computing?
  - Transformers are still widely used in many sequence processing and generation tasks.
- Large number of parameters cause the high latency in training and inference.
- Current computation libraries are insufficient.

LightSeq: A high-performance library

- **Efficient**
  - LightSeq achieves up to 14x speedup compared with TensorFlow and PyTorch.
- **Functional**
  - LightSeq supports more architecture variants and different search algorithms.
- **Convenient**
  - LightSeq is easy to use without any code modification.
  - Seemless porting from Tensorflow, Pytorch, Hugginface, Fairseq

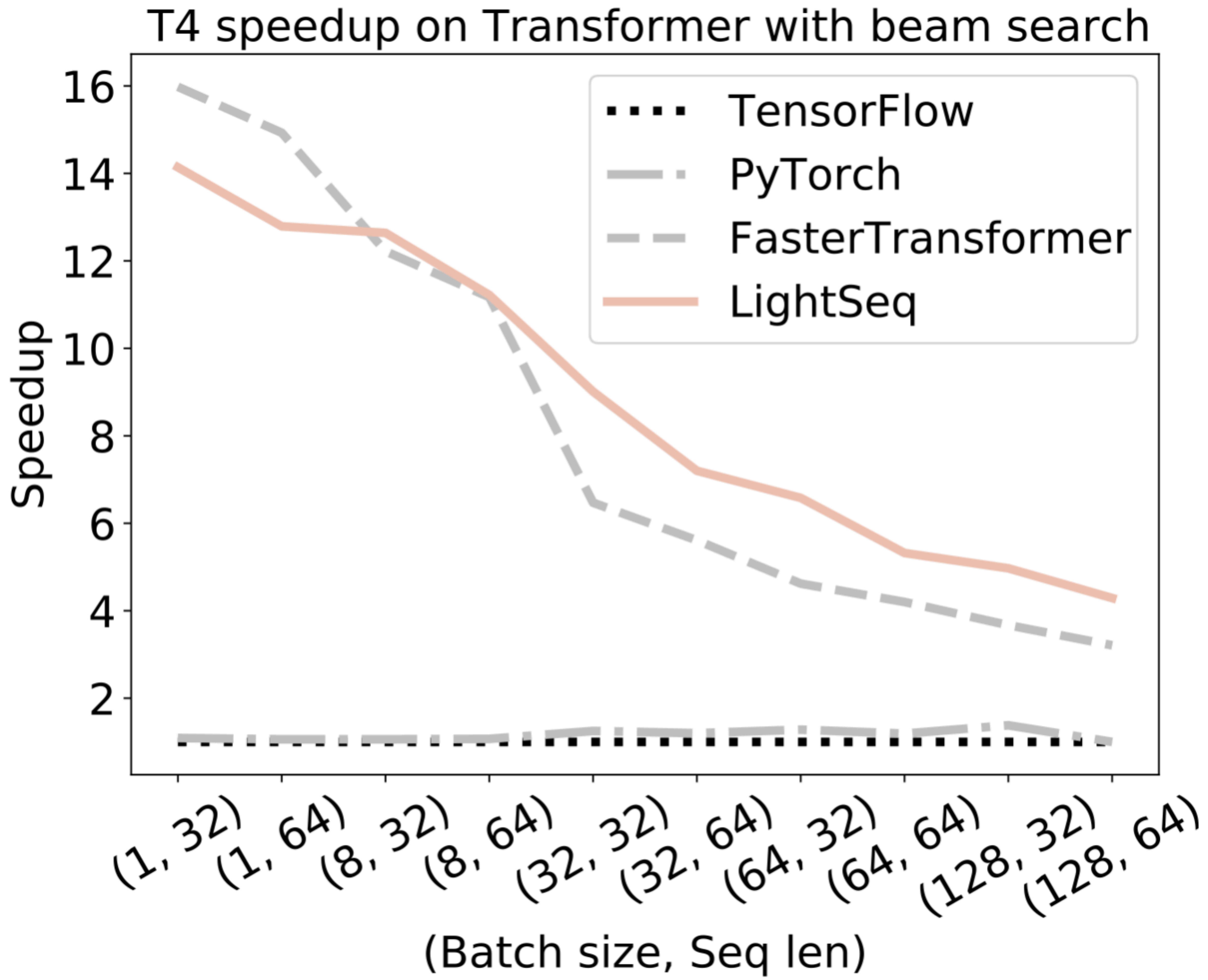
Improve GPU Occupation

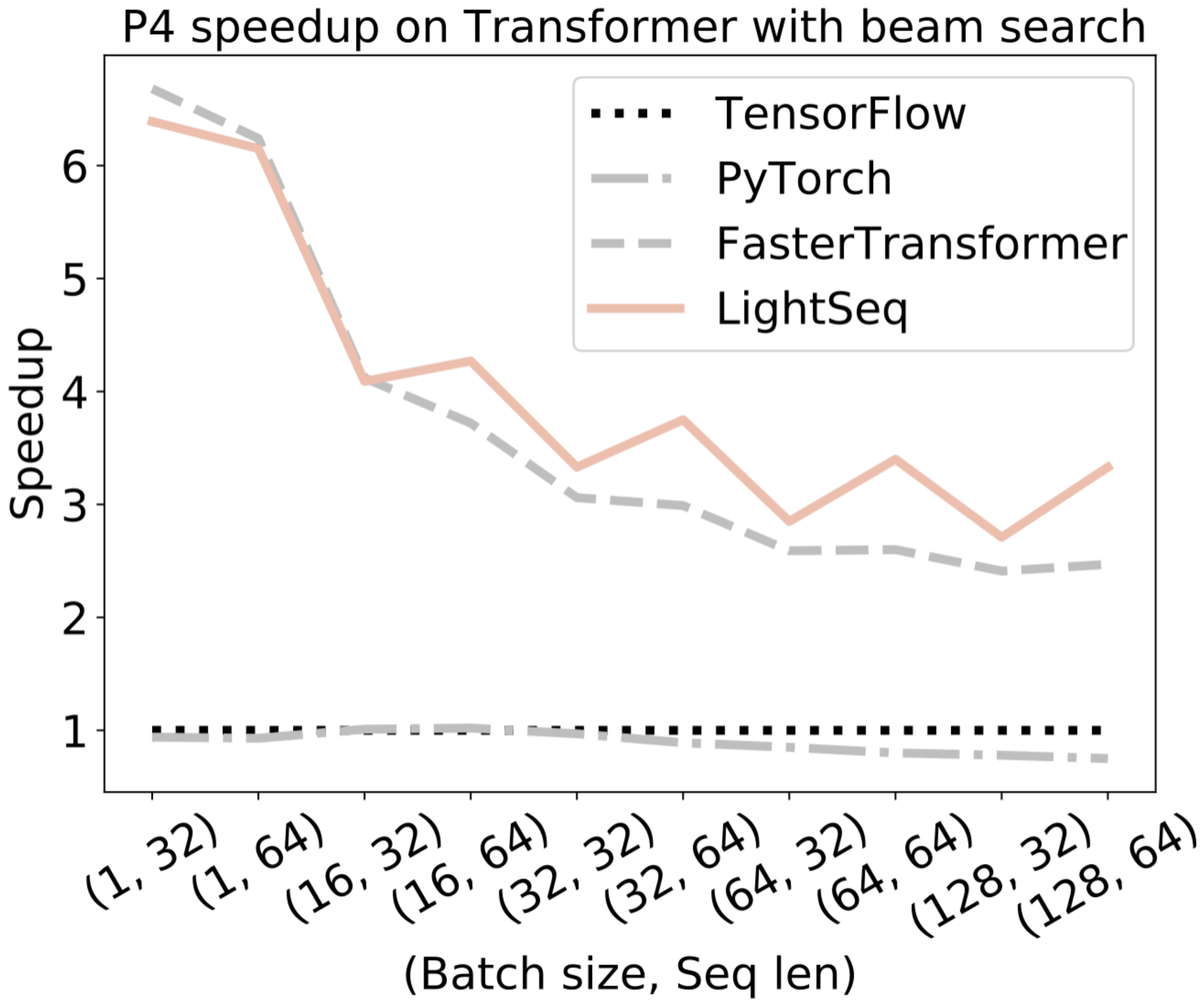
- LightSeq greatly reduces the proportion of kernels other than GEMM.

Wang, Xiong, Wei, Wang, Li. LightSeq: A High Performance Inference Library for Transformers. NAACL 2021.

Speedup for Machine Translation

- LightSeq outperforms others in most cases, especially in large batch size.

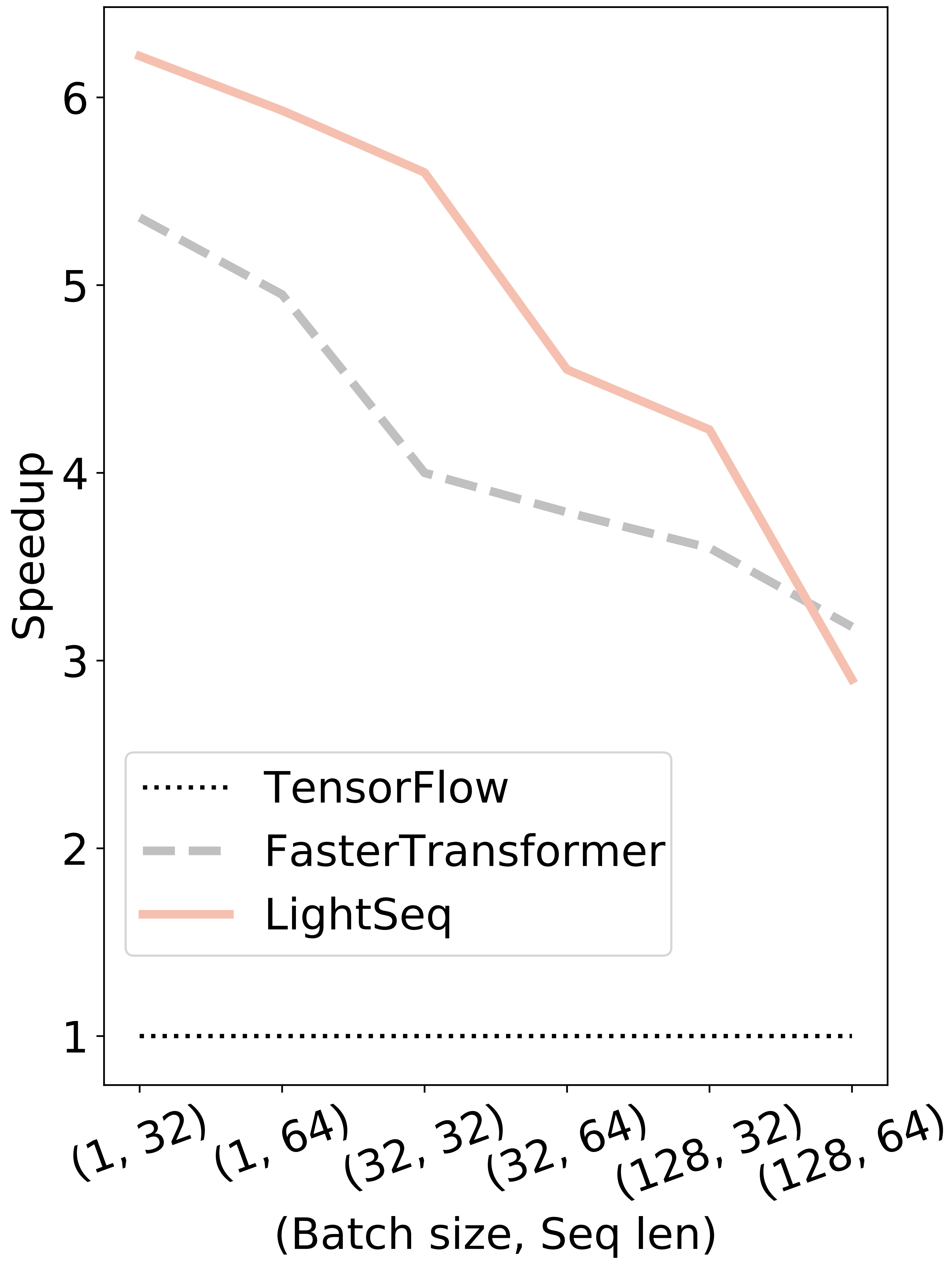




Wang, Xiong, Wei, Wang, Li. LightSeq: A High Performance Inference Library for Transformers. NAACL 2021.

Faster Text Generation w/ LightSeq

- LightSeq outperforms others in most cases



Wang, Xiong, Wei, Wang, Li. LightSeq: A High Performance Inference Library for Transformers. NAACL 2021.

Summary for Efficient MT

- Algorithm: VOLT
  - Learning Compact Vocabulary for NMT
  - Small vocabulary with improved performance at 100x faster!
  - Green solution: 30mins on only one cpu.
- Model: GLAT
  - Parallel Generation really works for the first time!
  - Translate at equal or better quality with 10x speedup!
  - Deployed in production
- Computing: LightSeq
  - Hardware Acceleration for training and inference
  - 14x faster than Tensorflow & Pytorch!

Towards Green MT

- Many challenges remaining!
- Propose new metric: Best value MT
  - GFlops or carbon footprint for model development
- Hardware acceleration for GLAT and other NAT?
- Low-end hardware?
- Taming the model size?

Thanks!

Contact: lilei@cs.ucsb.edu

- Code:
  - VOLT: https://github.com/Jingjing-NLP/VOLT
  - GLAT: https://github.com/FLC777/GLAT
- Open Source Library

LightSeq

Transformer fast training and inference lib

NeurST

Speech and Text Translation Toolkit

CCMT 2021/10/9

13:55–14:20	报告2 预训练时代的机器翻译 王明轩
14:20–15:10	机器翻译前沿趋势Panel（嘉宾: 王瑞 王明轩 李军辉 孟凡东 王强）
15:25–15:40	机器翻译多媒体领域的实践和探索 刘坚（字节跳动）

CCMT 2021/10/10

10:48-11:50	Panel 2端到端语音翻译的研究与应用（主持人：黄辉; 嘉宾：张家俊 刘树杰 熊德意 何中军 王明轩）
15:40-15:55	报告1 绿色词表学习: ACL论文背后的故事 许晶晶
16:10-16:25	报告3 敢想+敢拼：记一次用并行生成参加WMT的经历 周浩