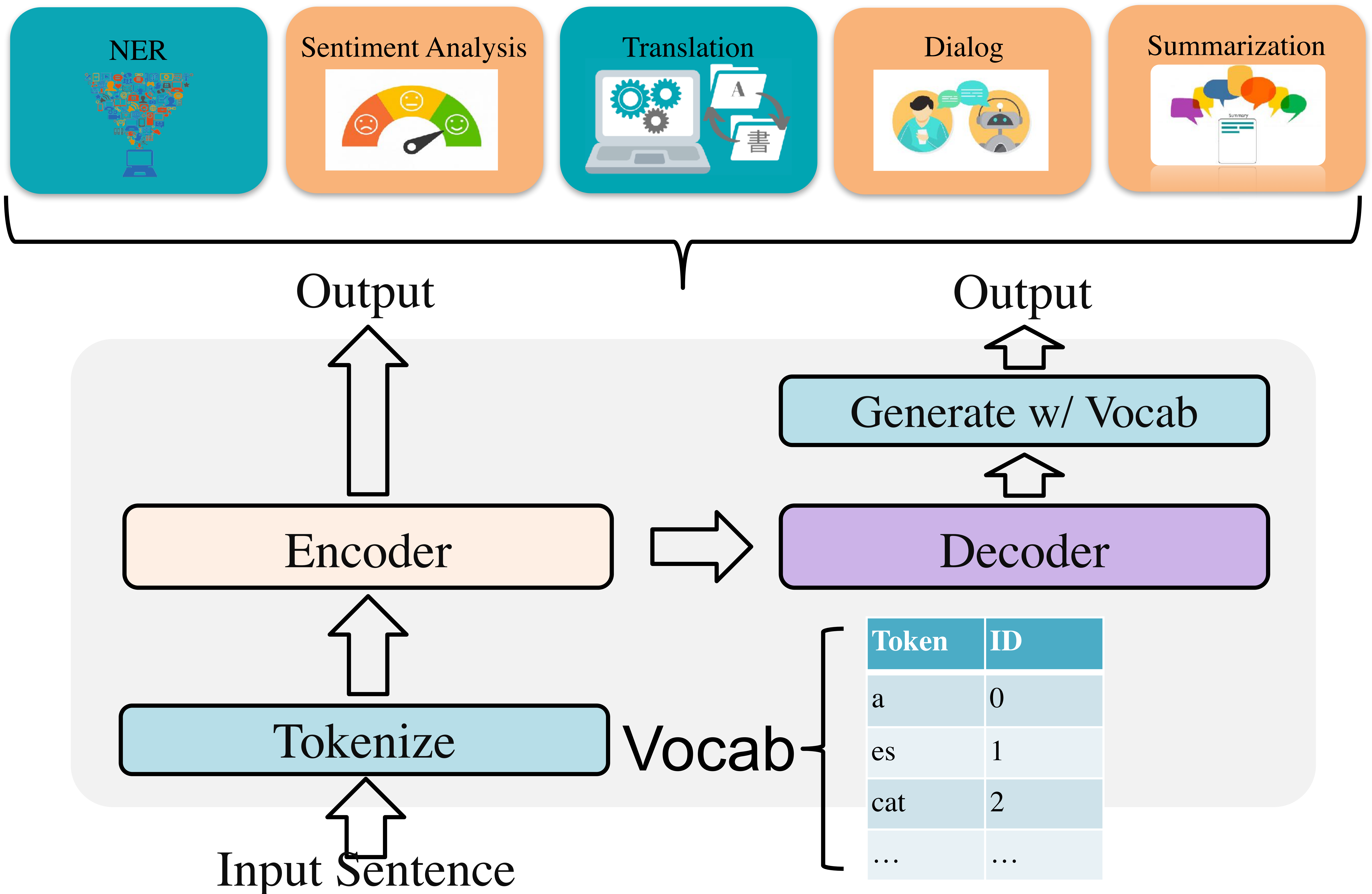
291K
Deep Learning for Machine Translation
Advanced Vocabulary Learning

Lei Li
UCSB

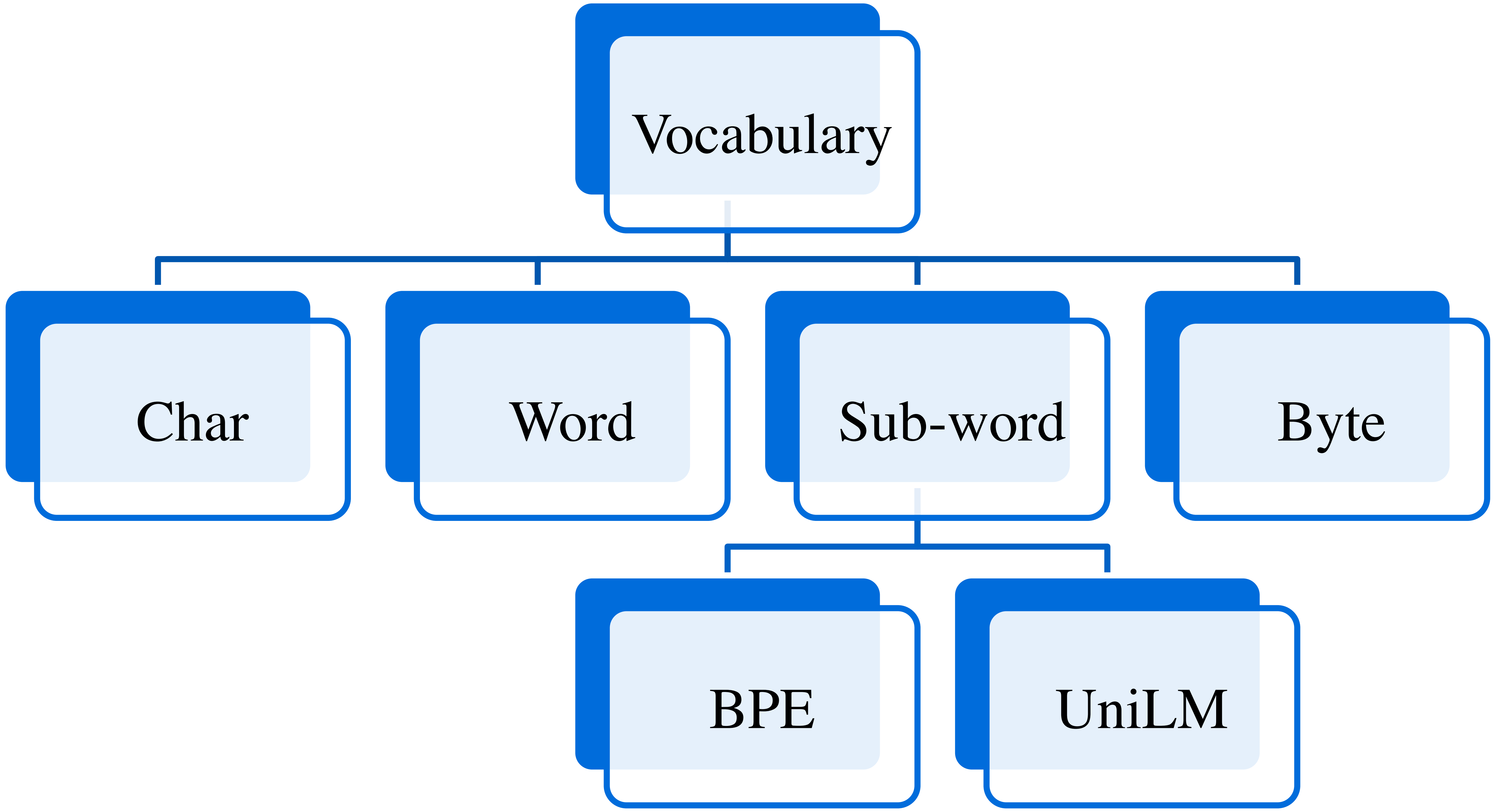
11/24/2021

Vocabulary is Fundamental and Important



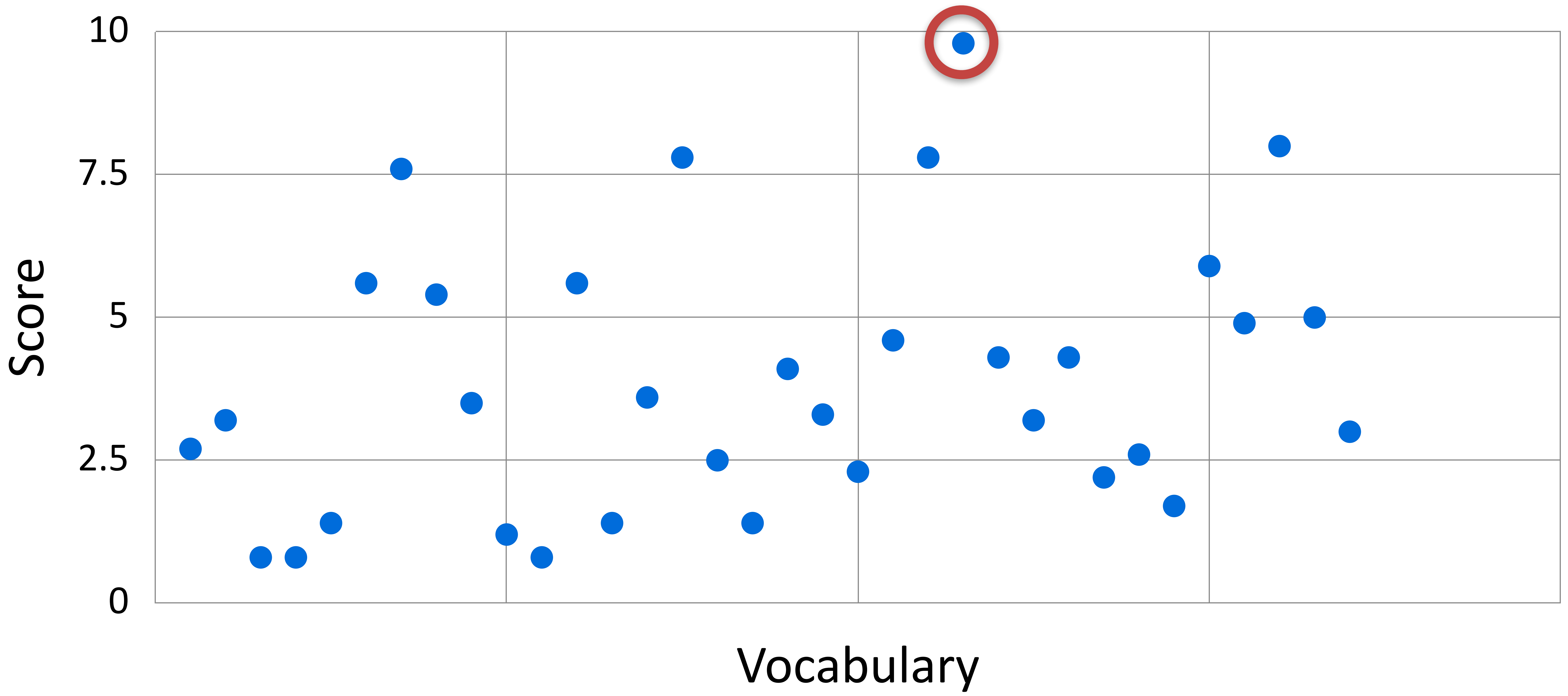
Methods to Construct Vocabulary

How to construct the optimal vocabulary?



How to find the optimal vocabulary?

- Q1: How to efficiently evaluate vocabularies?
- Q2: How to efficiently find the optimal one?



Q1: How to evaluate vocabulary?

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)

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

  - 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? From Size

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? From information?

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

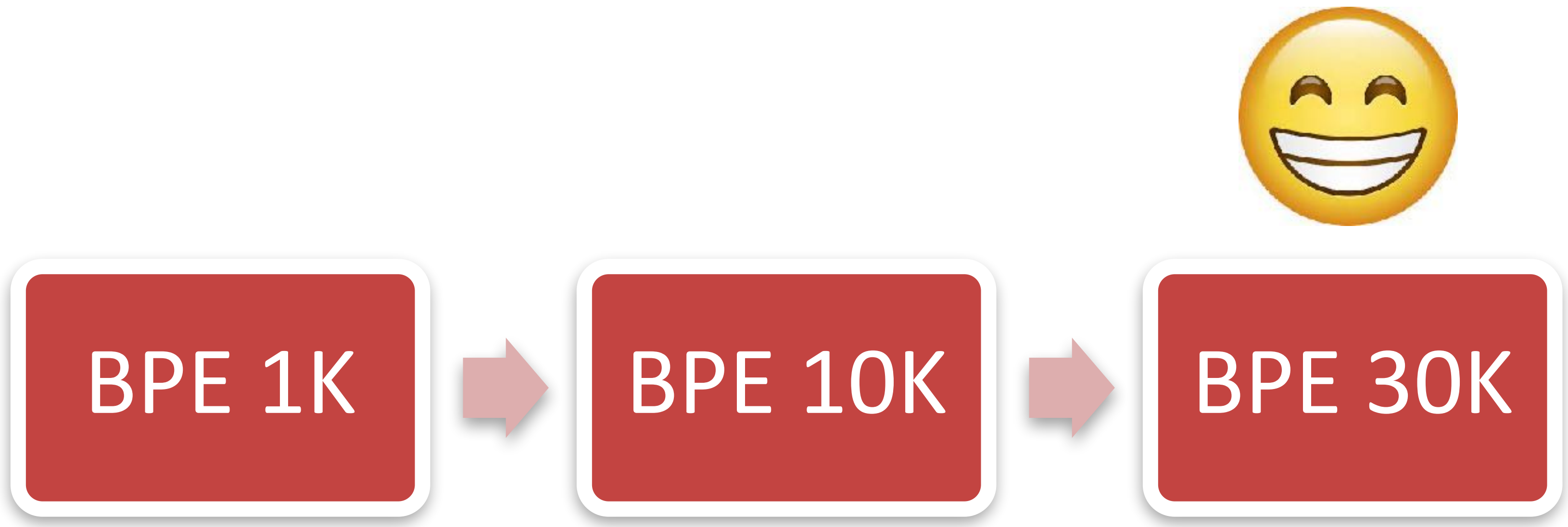
From the perspective of entropy, BPE-30K seems to be better

* With normal-size data

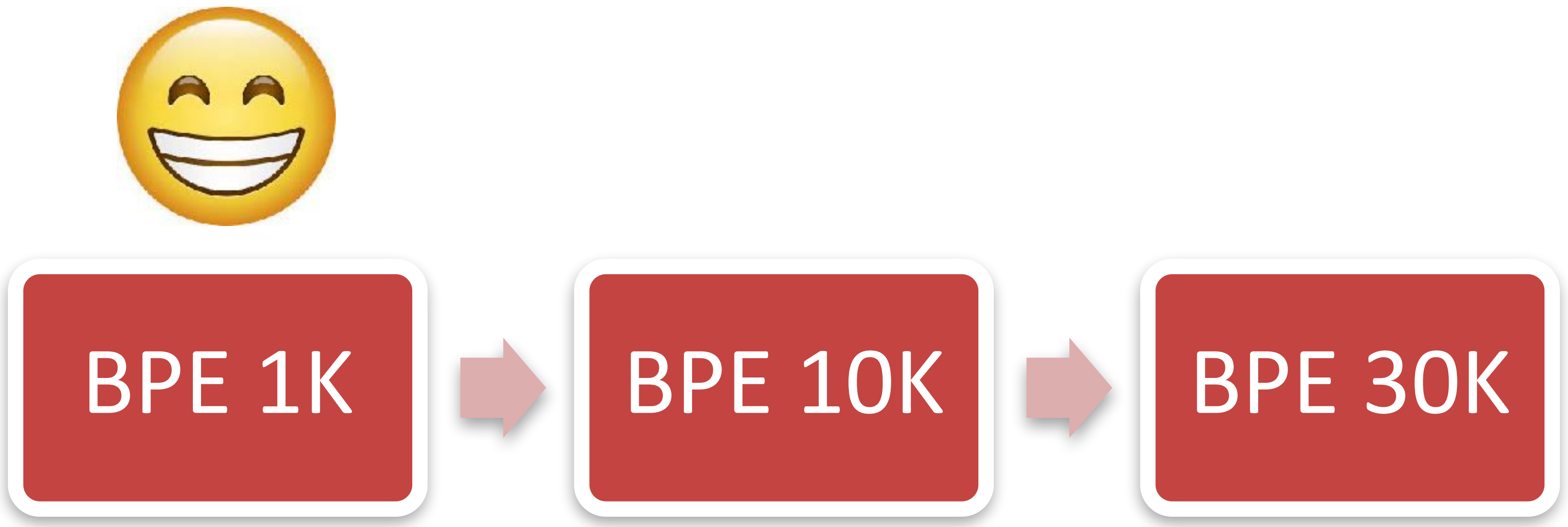
Evaluating Vocabulary Quality is Expensive

Which one is better?

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



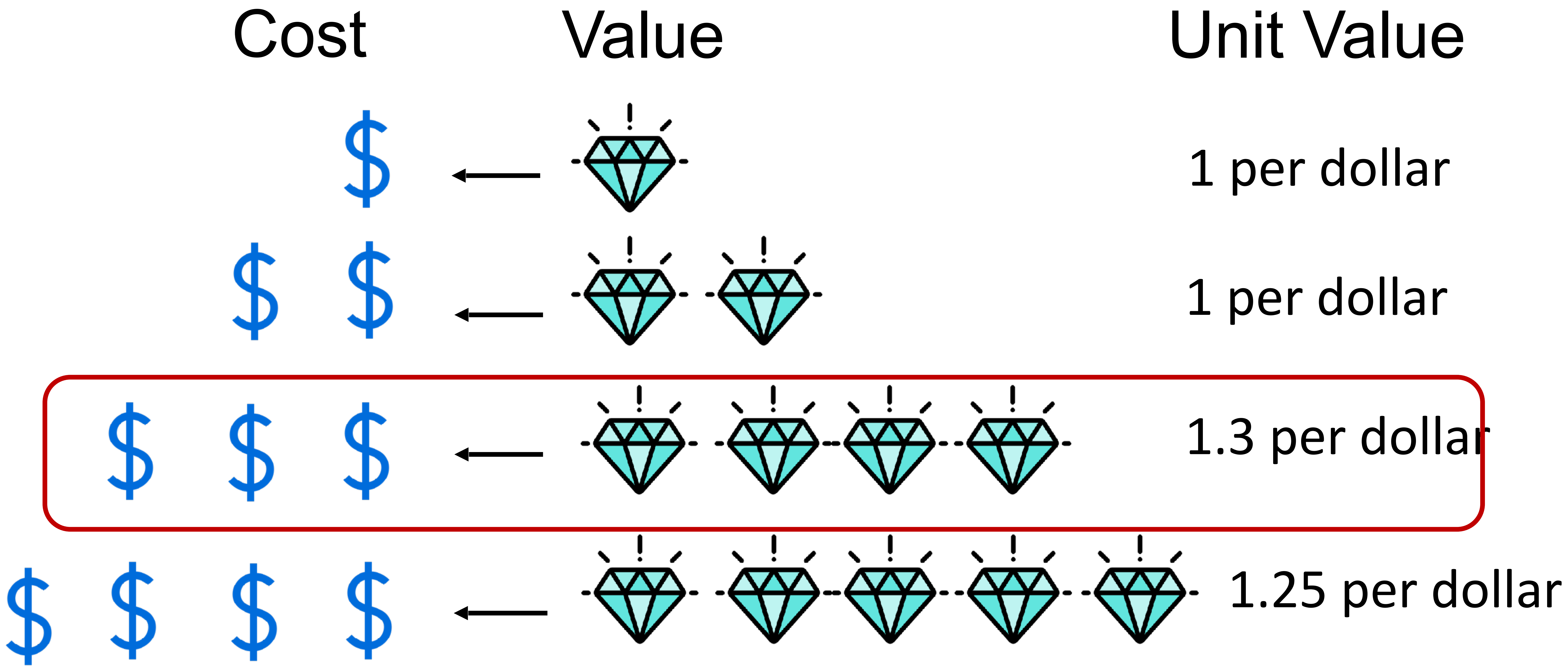
IPC



Size

An analogy: Buying Good with Money

- Value: 
- Cost: 



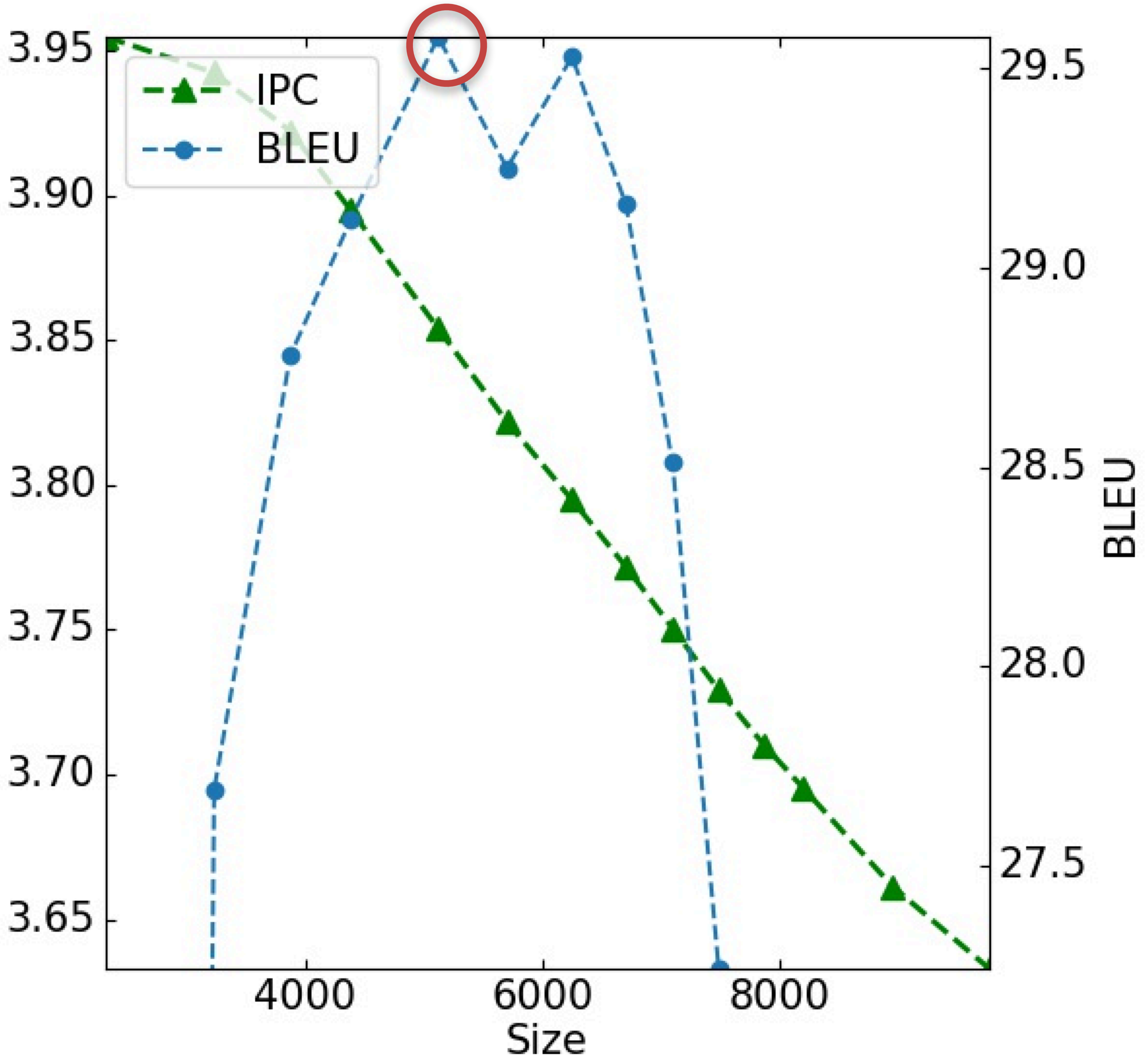
Xu, Zhou, Gan, Zheng, Li. Vocabulary Learning via Optimal Transport for Neural Machine Translation. ACL 2021.

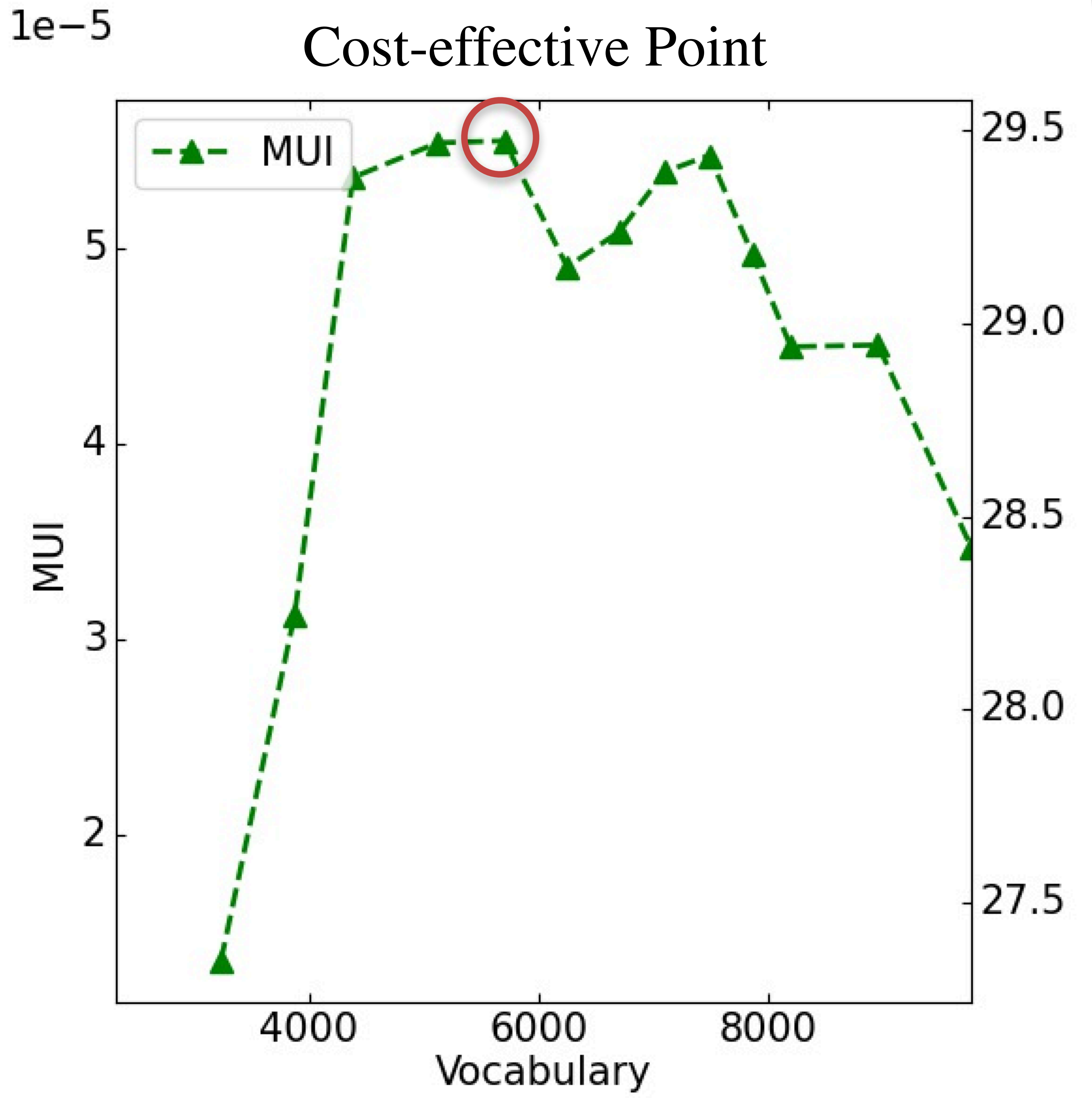
Utility of Information for Adding Tokens

- Value: **IPC** 
- Cost: **size** 
- Marginal utility of information for Vocabulary (MUV)
  - How many value does each unit-of-cost bring?
  - $M_{v_k \to v_{k+m}} = -\dfrac{H(v_k) - H(v_{k+m})}{m}$
  - Negative **gradients** of IPC to size

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

MUV is good indicator for MT performance



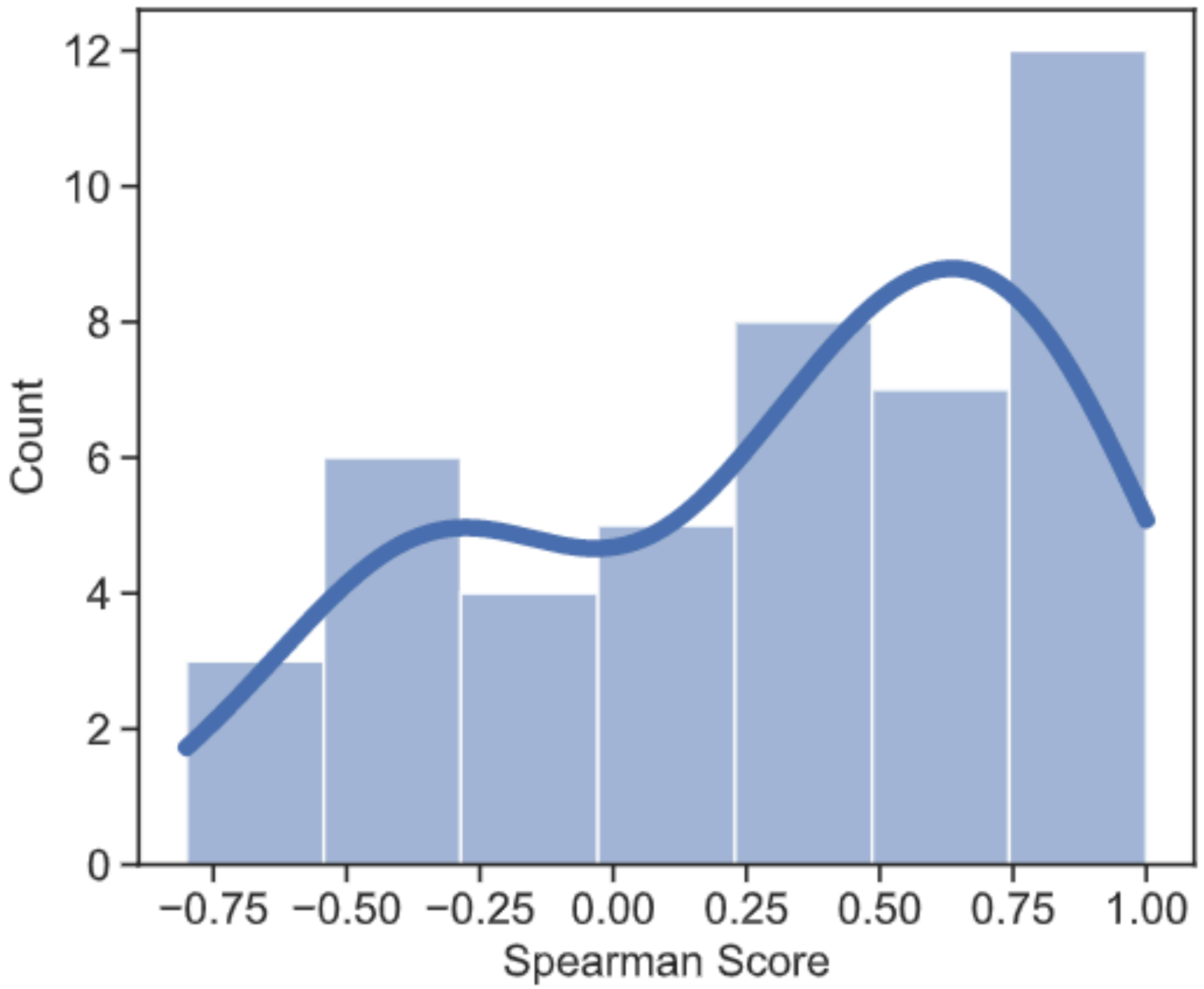


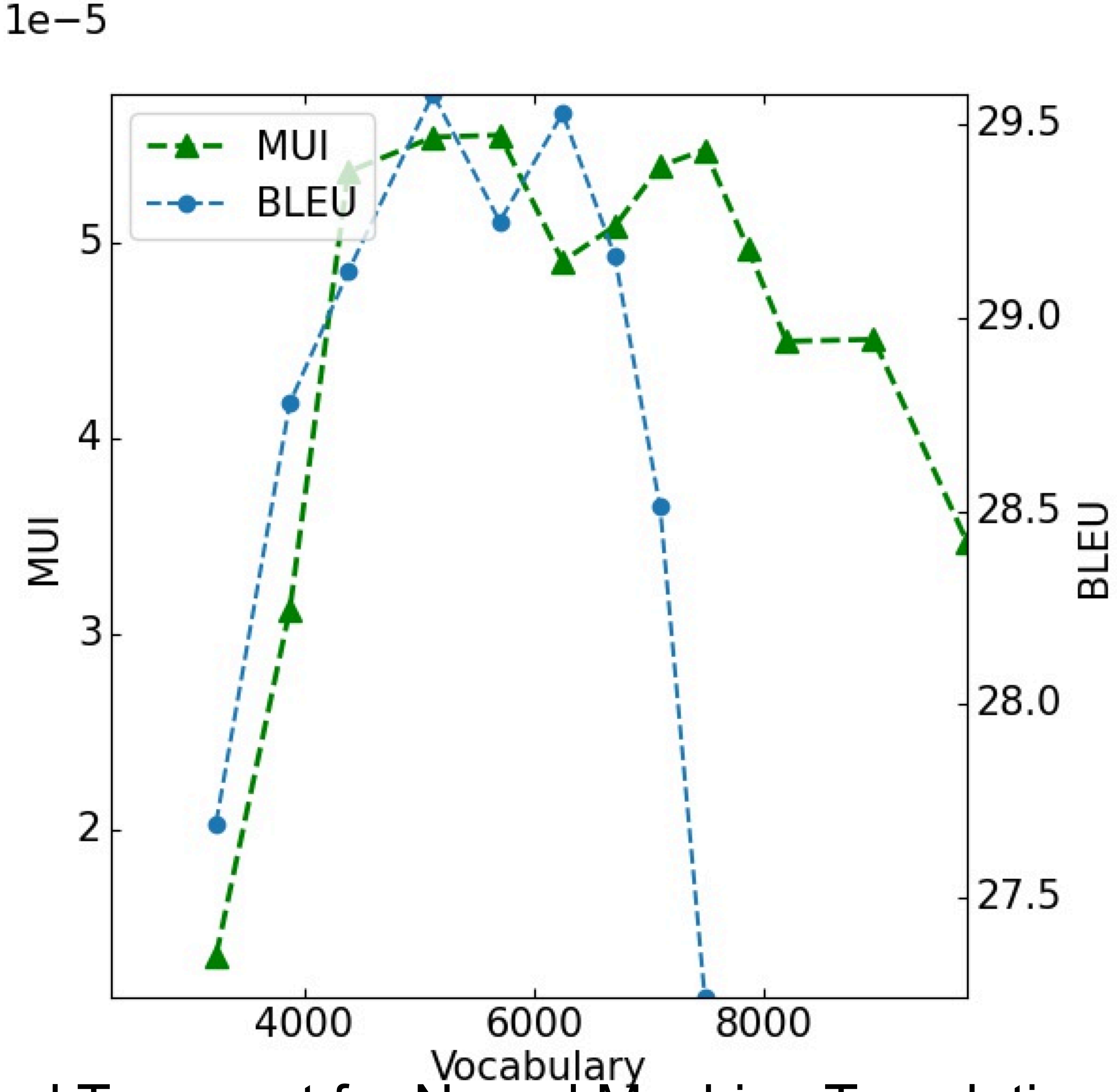
- Cost-effective point in MUV curve (maximum MUV)
  - ==> best BLEU

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.

Problem Reduction: Maximizing Marginal Utility of Vocab

- Goal: finding the optimal vocabulary

Finding the optimal vocabulary ⟺ Maximizing MUV

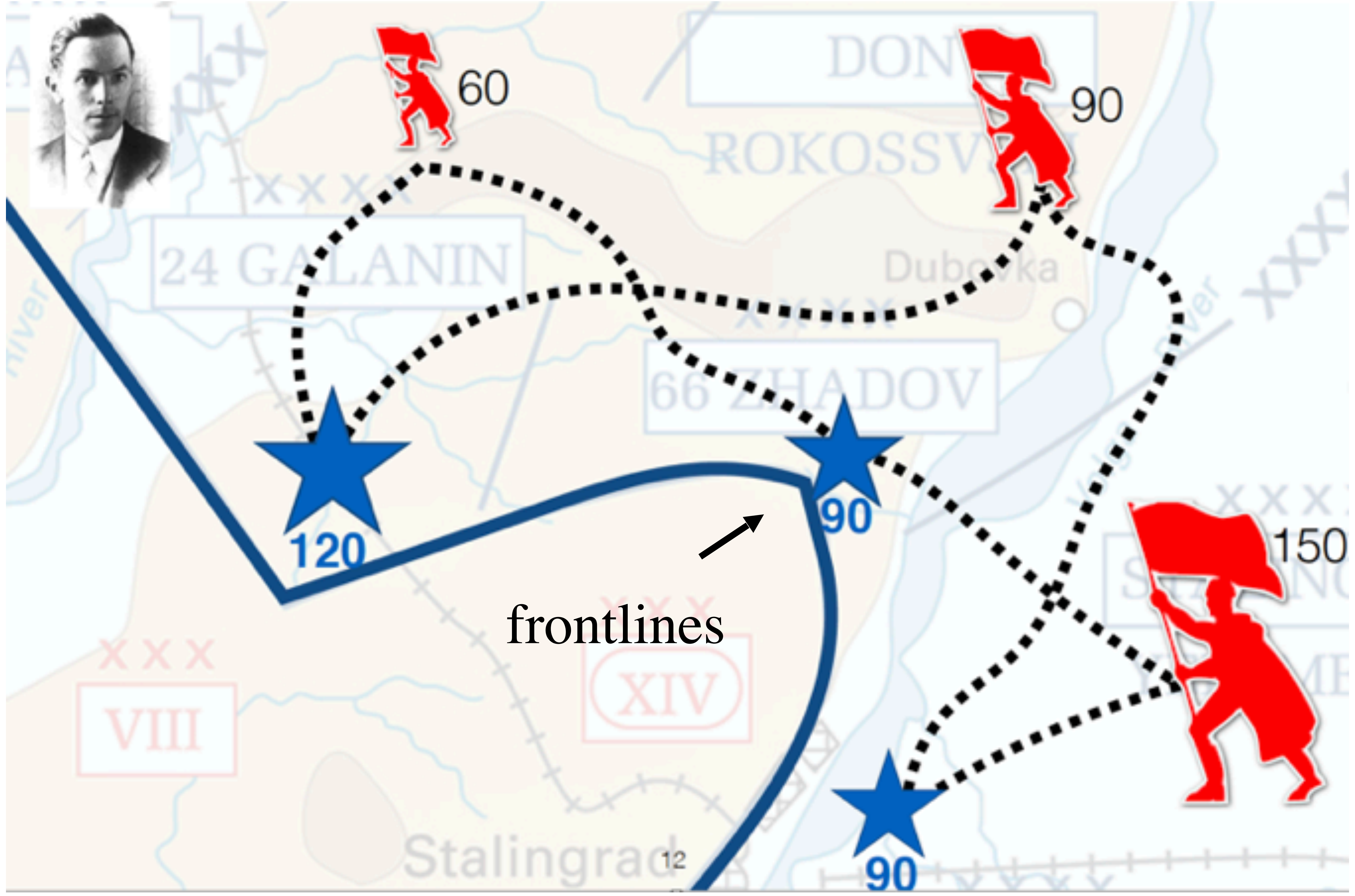
- Naive solution: MUI-Search
  - Exhaustive Search for vocabulary
  - Evaluate MUI for each and find max MUI
- How to search over a huge discrete space?

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

Q2: How to find the optimal vocabulary with the maximum MUI?

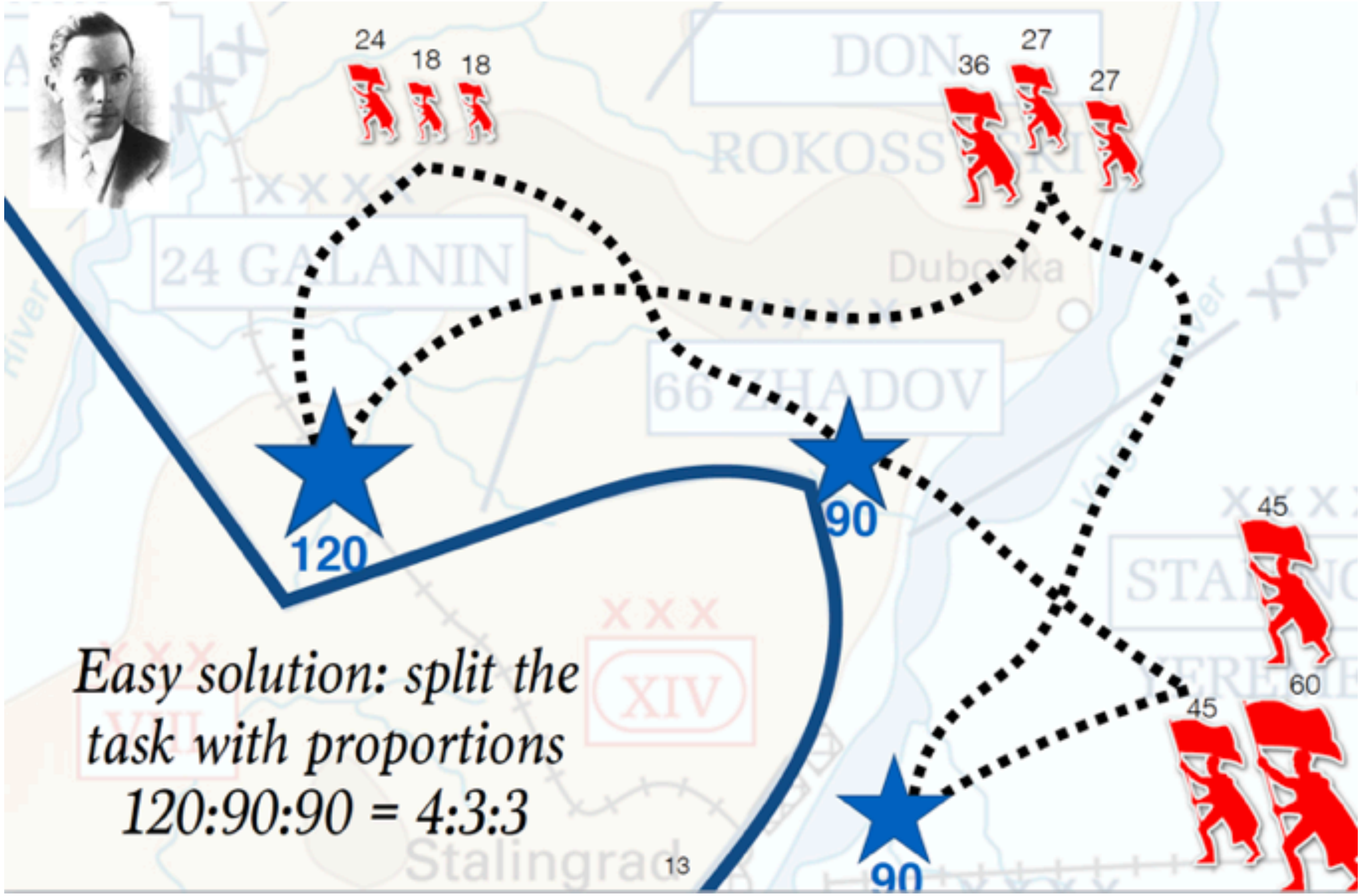
Problem Reduction

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

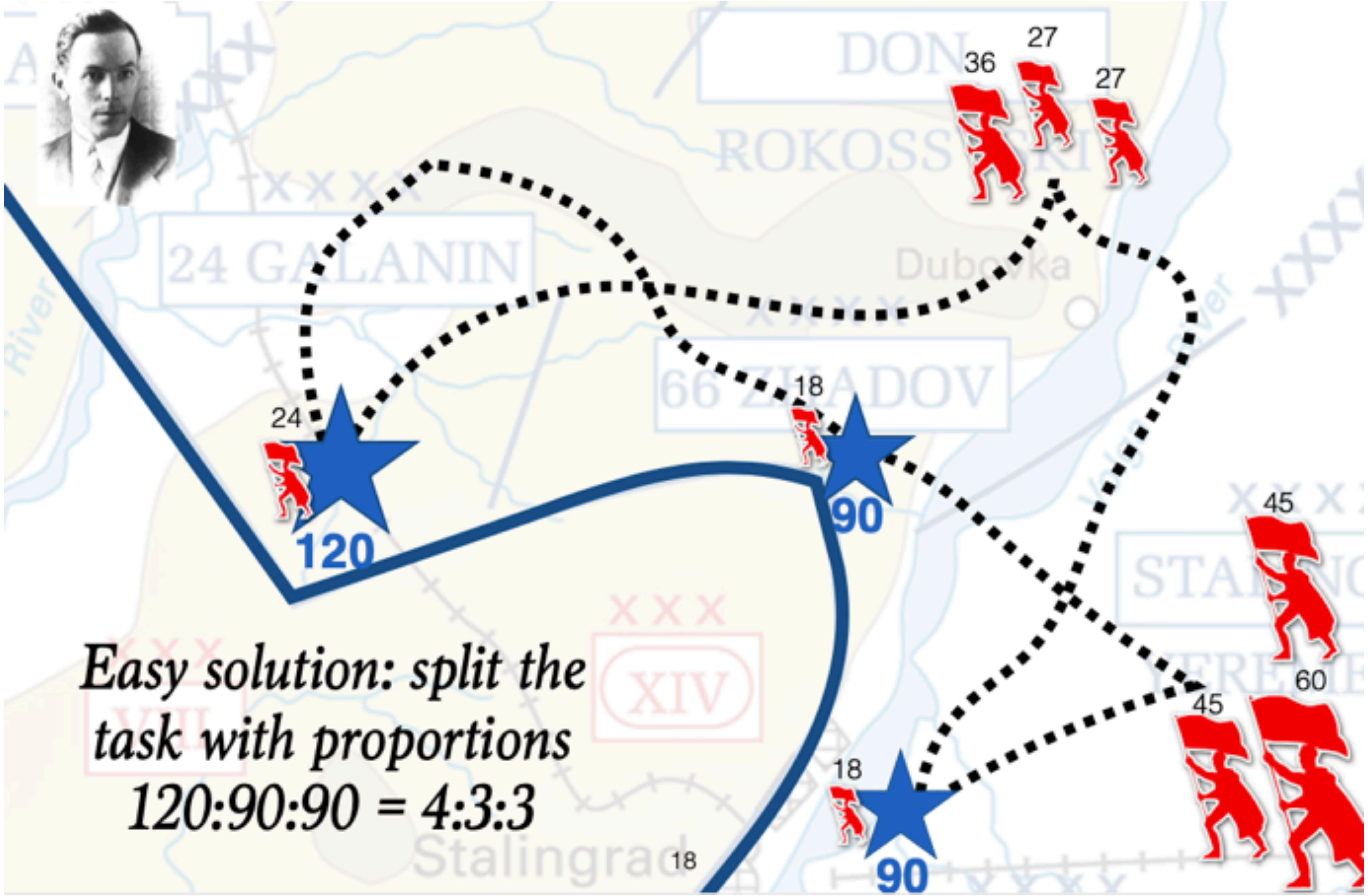


Min cost to Transport soldiers from bases to frontlines

Optimal Transport

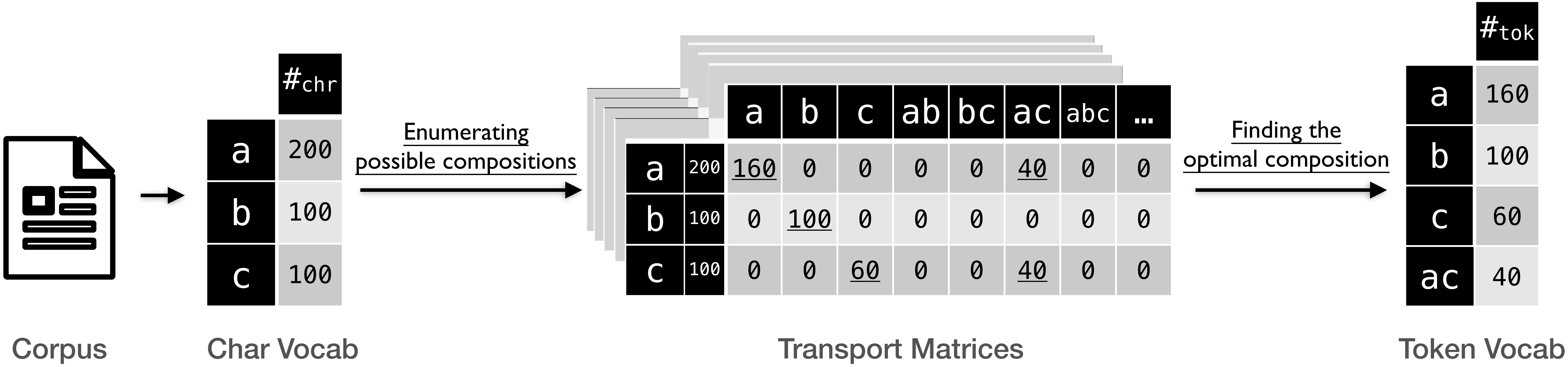


Optimal Transport



Vocabulary building as Transportation of Token Frequency

- Adding one new token means:
  - Transport character frequency to token frequency



	#chr
a	200
b	100
c	100

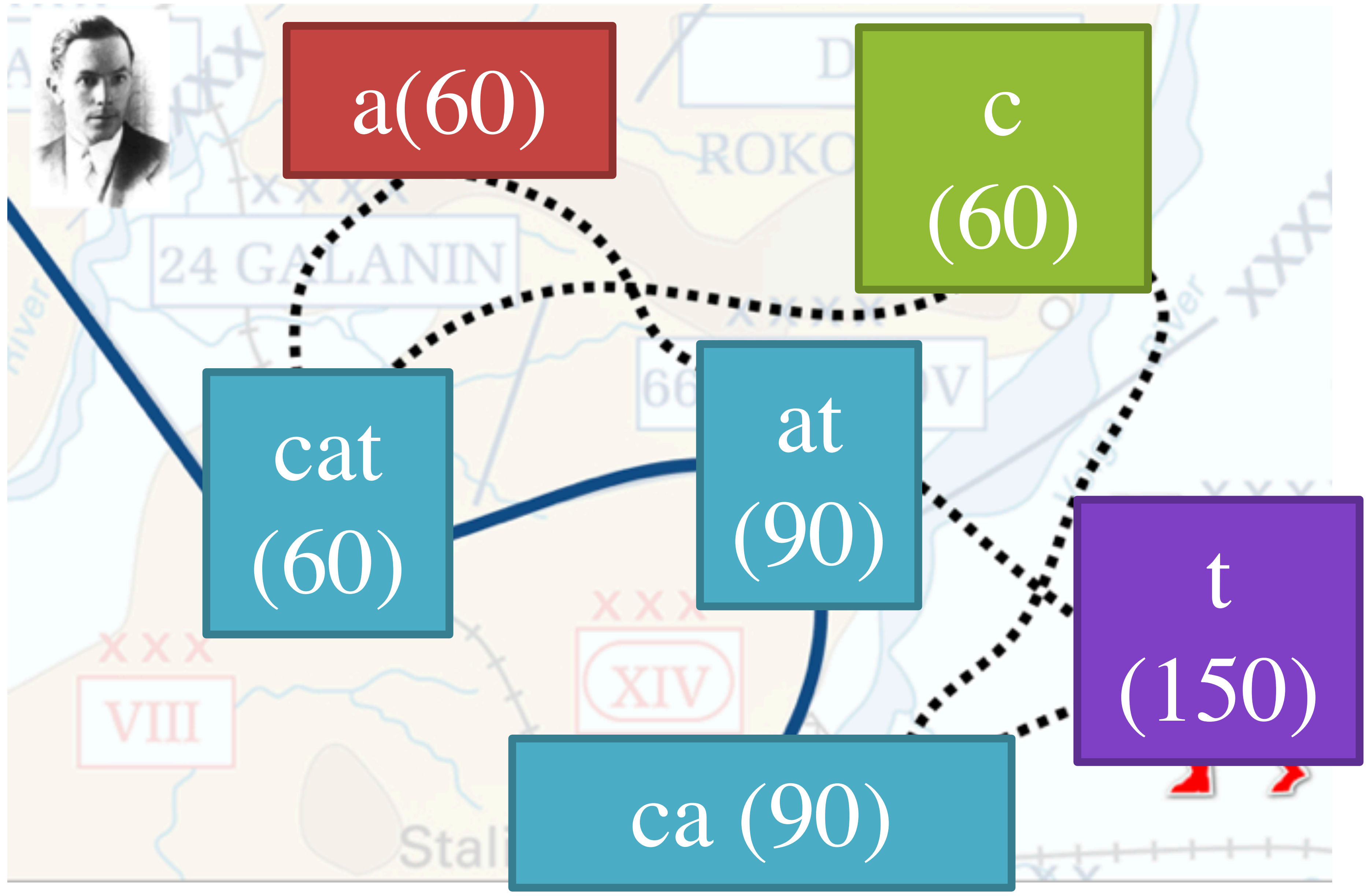
		a	b	c	ab	bc	ac	abc	…
a	200	160	0	0	0	0	40	0	0
b	100	0	100	0	0	0	0	0	0
c	100	0	0	60	0	0	40	0	0

	#tok
a	160
b	100
c	60
ac	40

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

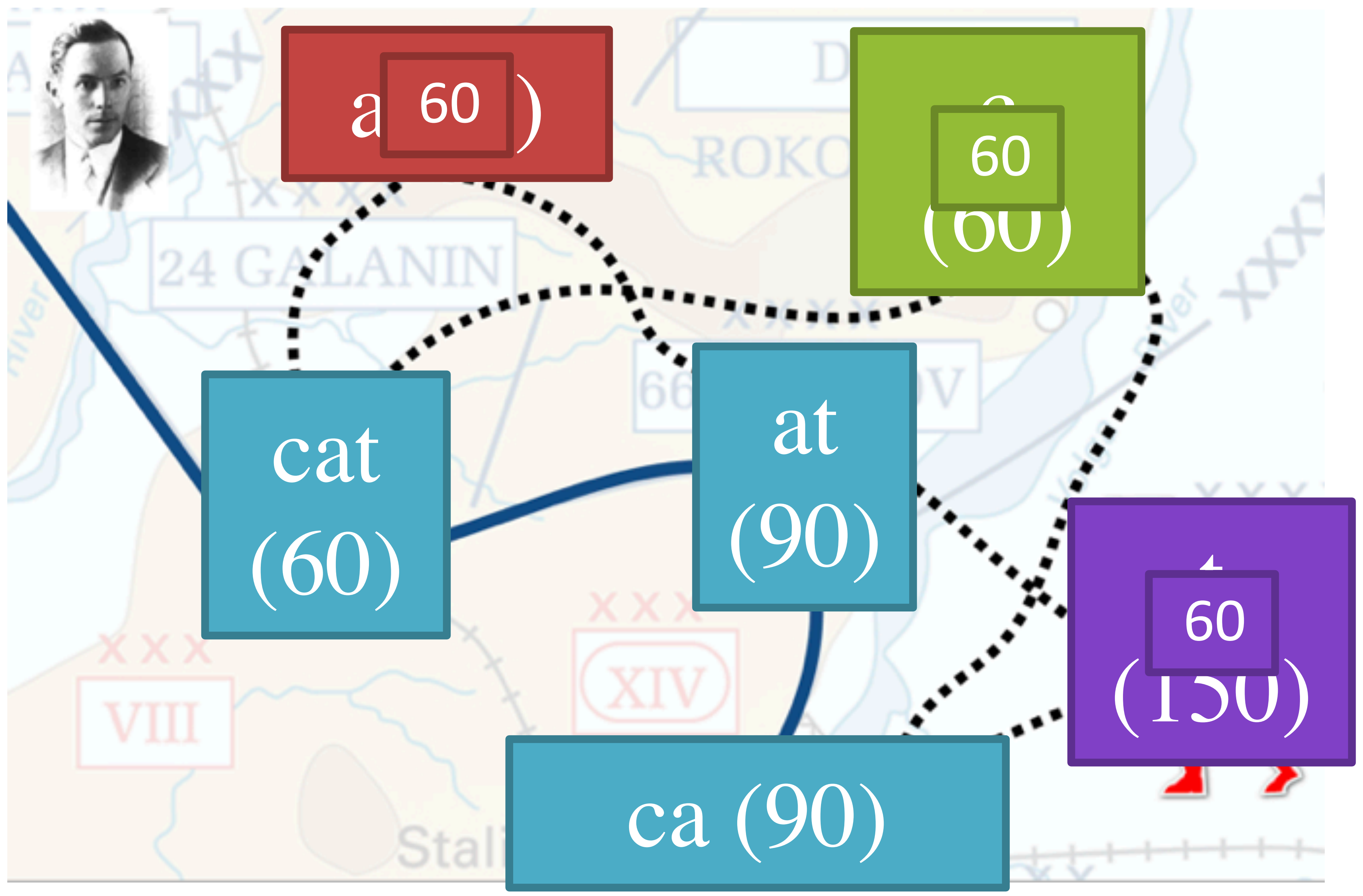
VOLT Formulation

Transport chars to tokens



VOLT Formulation

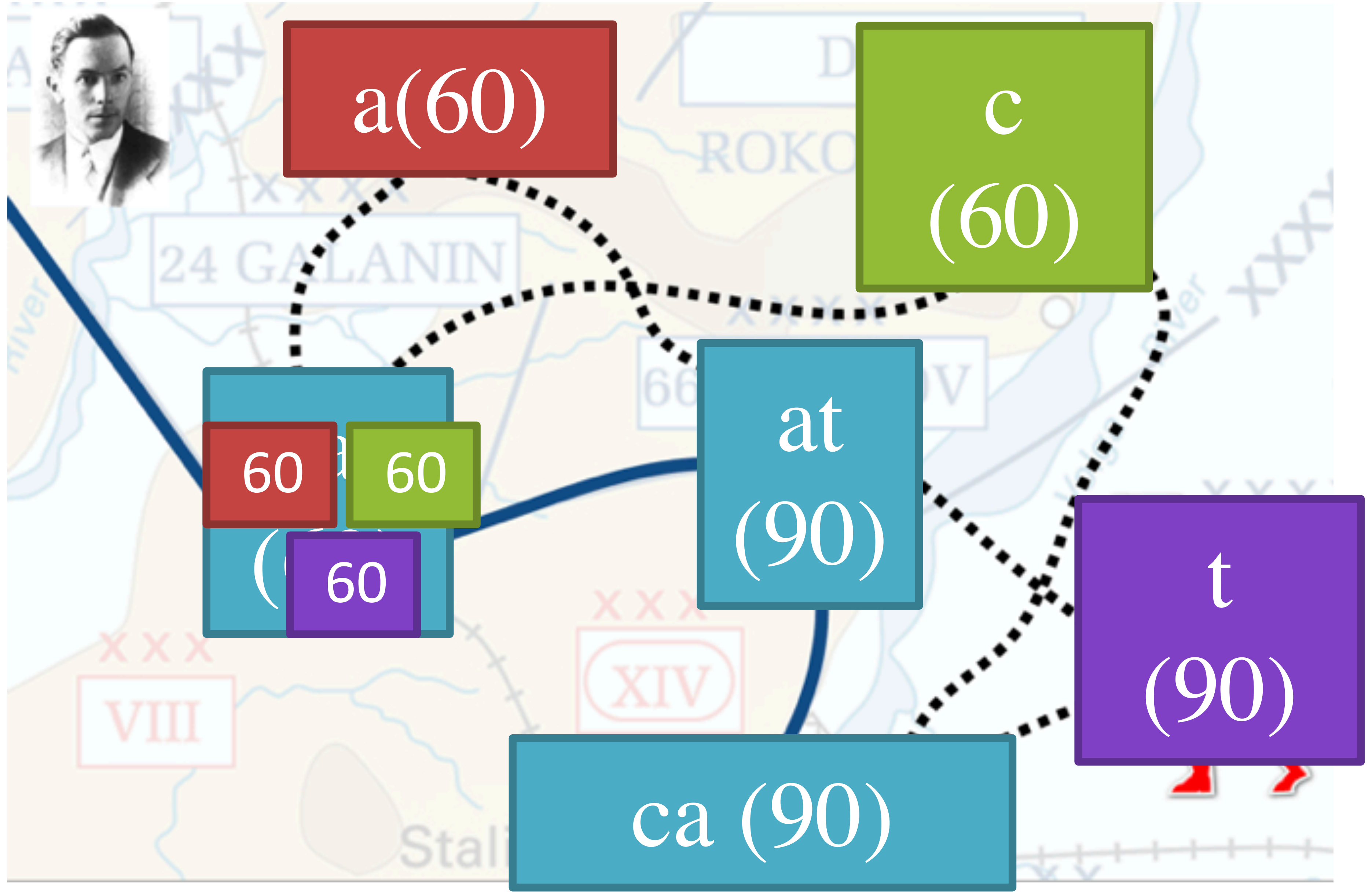
Not all tokens can get chars



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

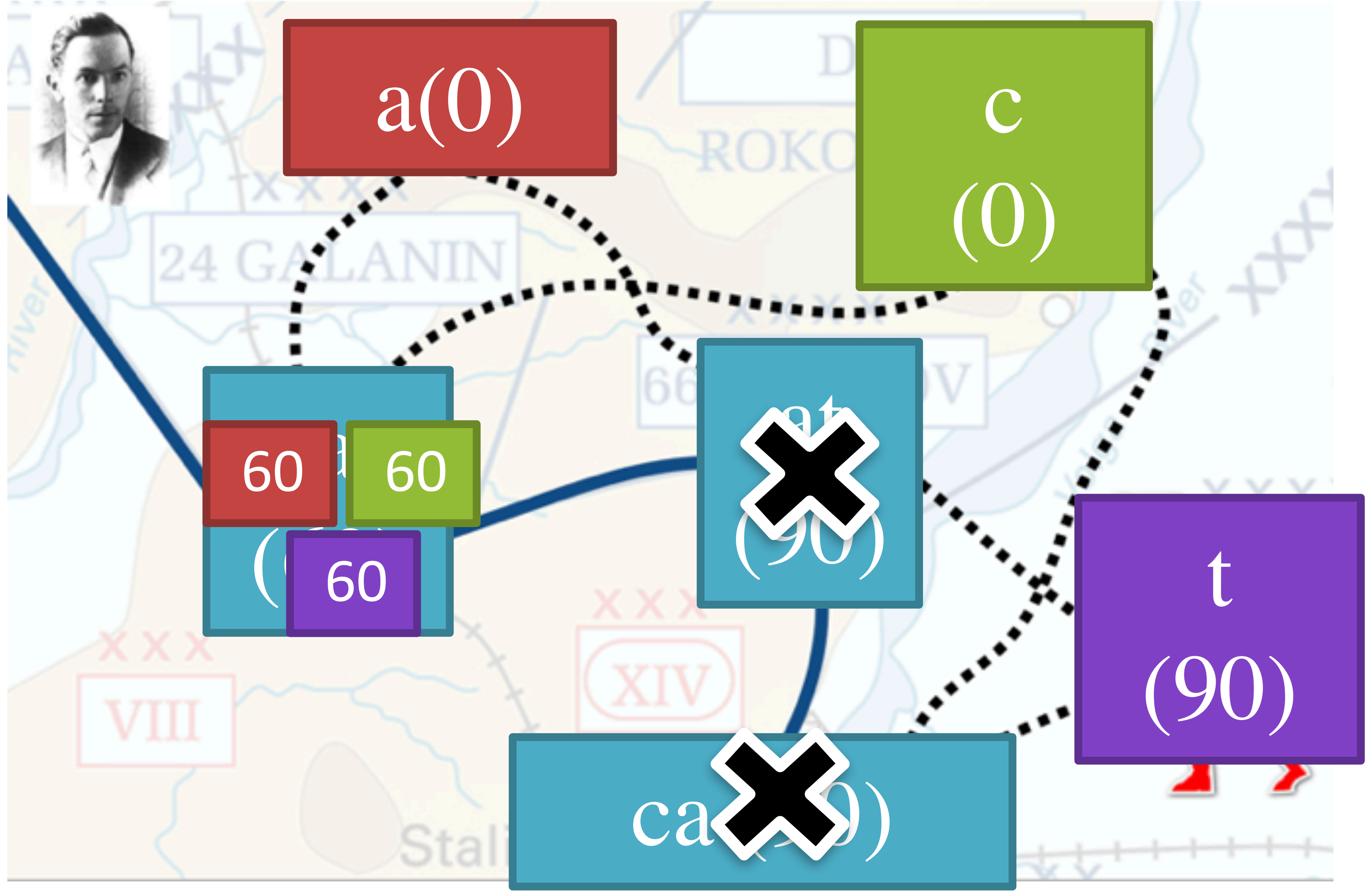
VOLT Formulation

Not all tokens can get chars



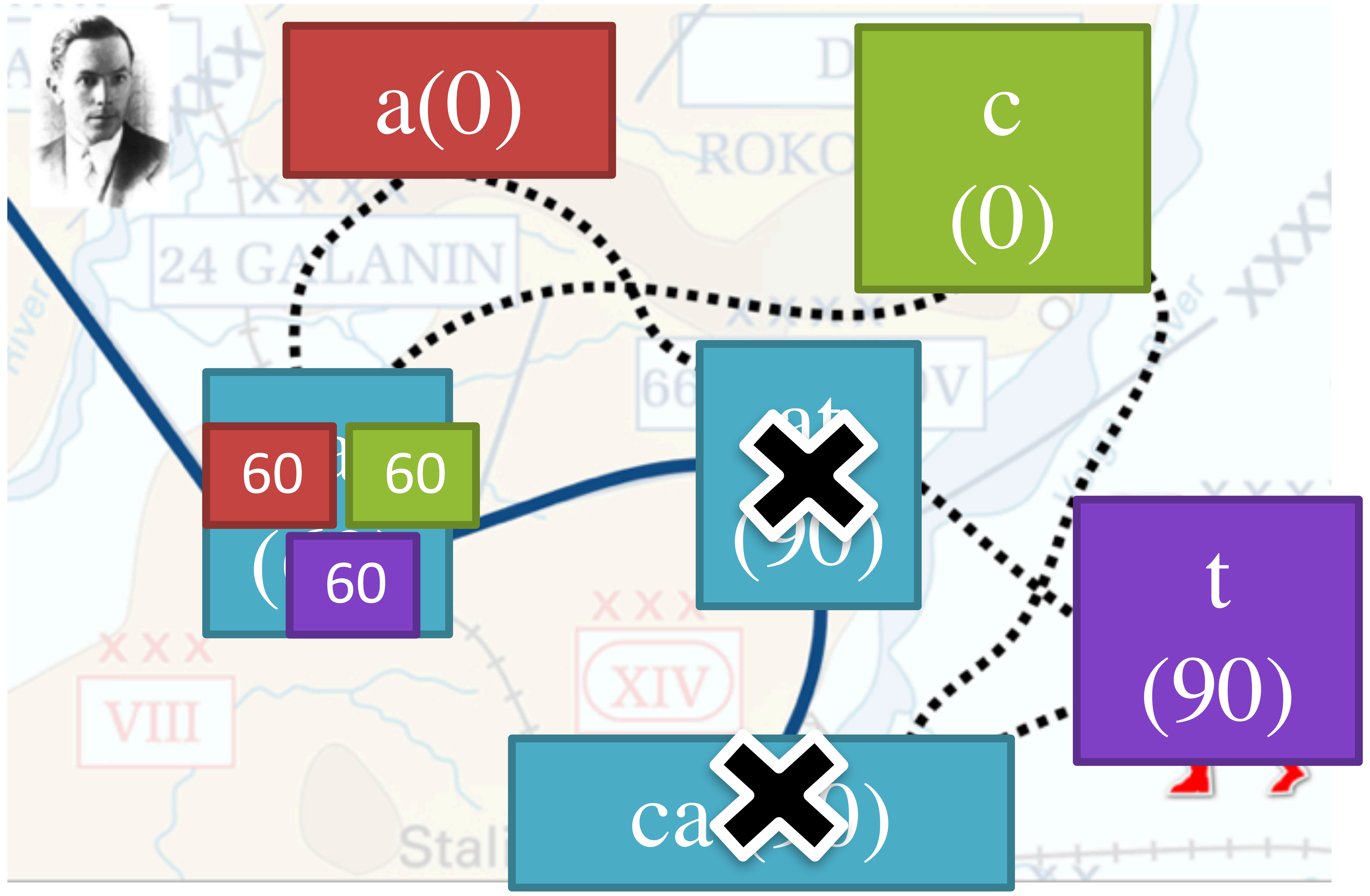
VOLT Formulation

Not all tokens can get chars



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 lower-bound
- ==> $\max_t(\max H(V_t) - \max H(V_{t+1}))$
- Finding $\max_v H(v)$ ==> Optimal Transport

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

Finding the Transportation Matrix

- 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

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

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]

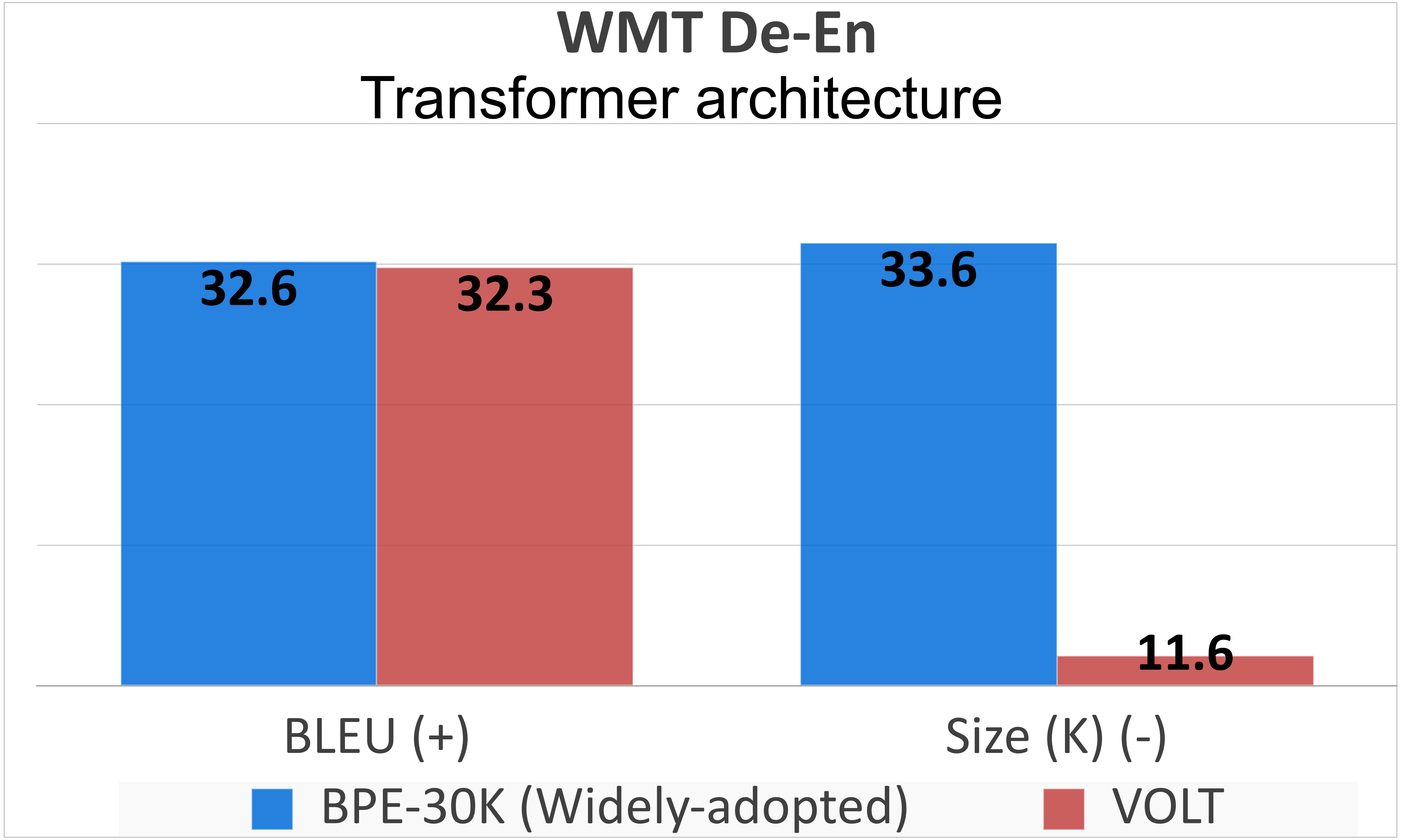
Xu, Zhou, Gan, Zheng, Li. Vocabulary Learning via Optimal Transport for Neural Machine Translation. ACL 2021.

Encoding and Decoding with VOLT

- VOLT uses a greedy strategy to encode text with a constructed sub-word level vocabulary similar to BPE.

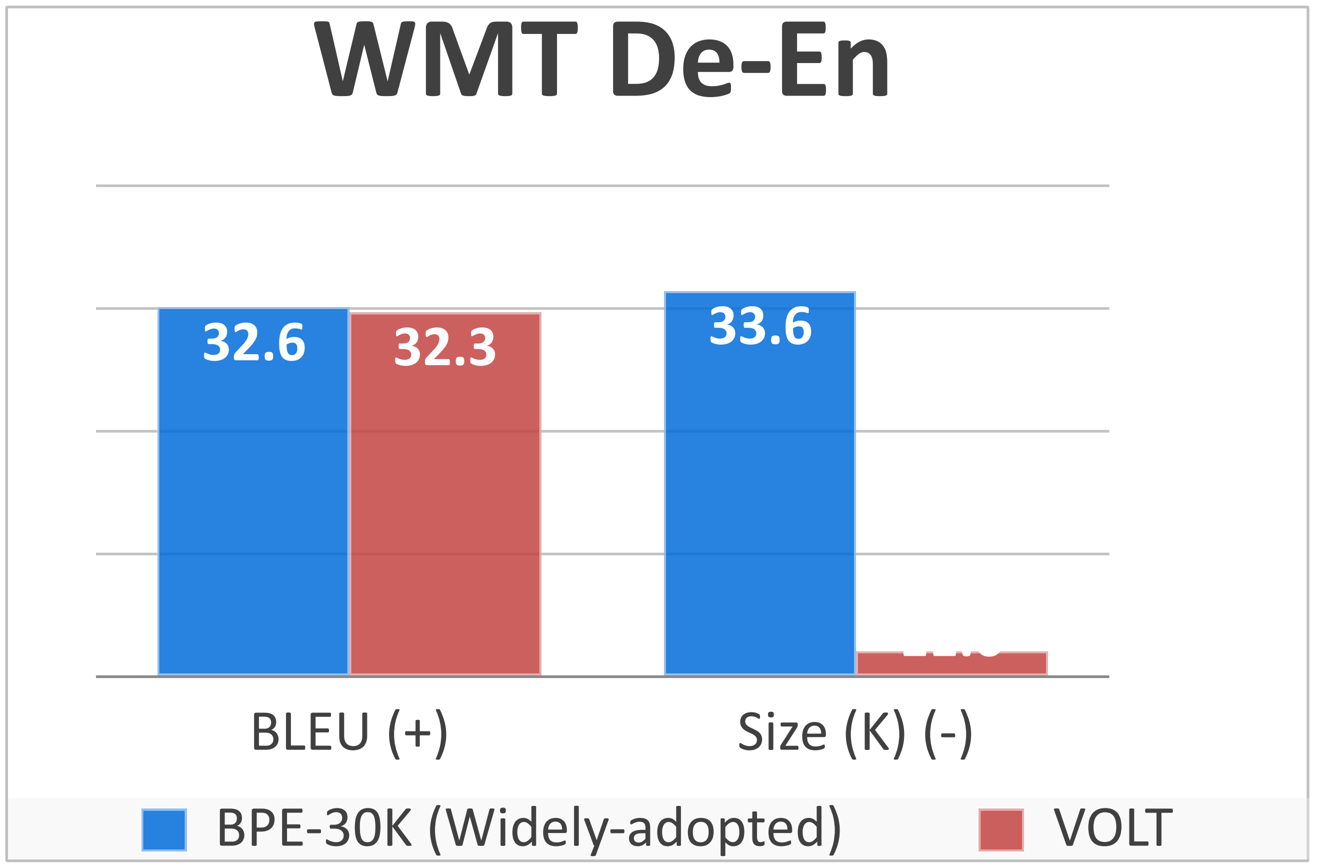
- The vocabulary includes all basic characters.
  - To encode text, it first splits sentences into character-level tokens.
  - Then, we merge two consecutive tokens into one token if the merged one is in the vocabulary.
  - This process keeps running until no tokens can be merged.
  - Out-of-vocabulary tokens will be split into smaller tokens.

VOLT finds better vocabulary on Bilingual MT



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

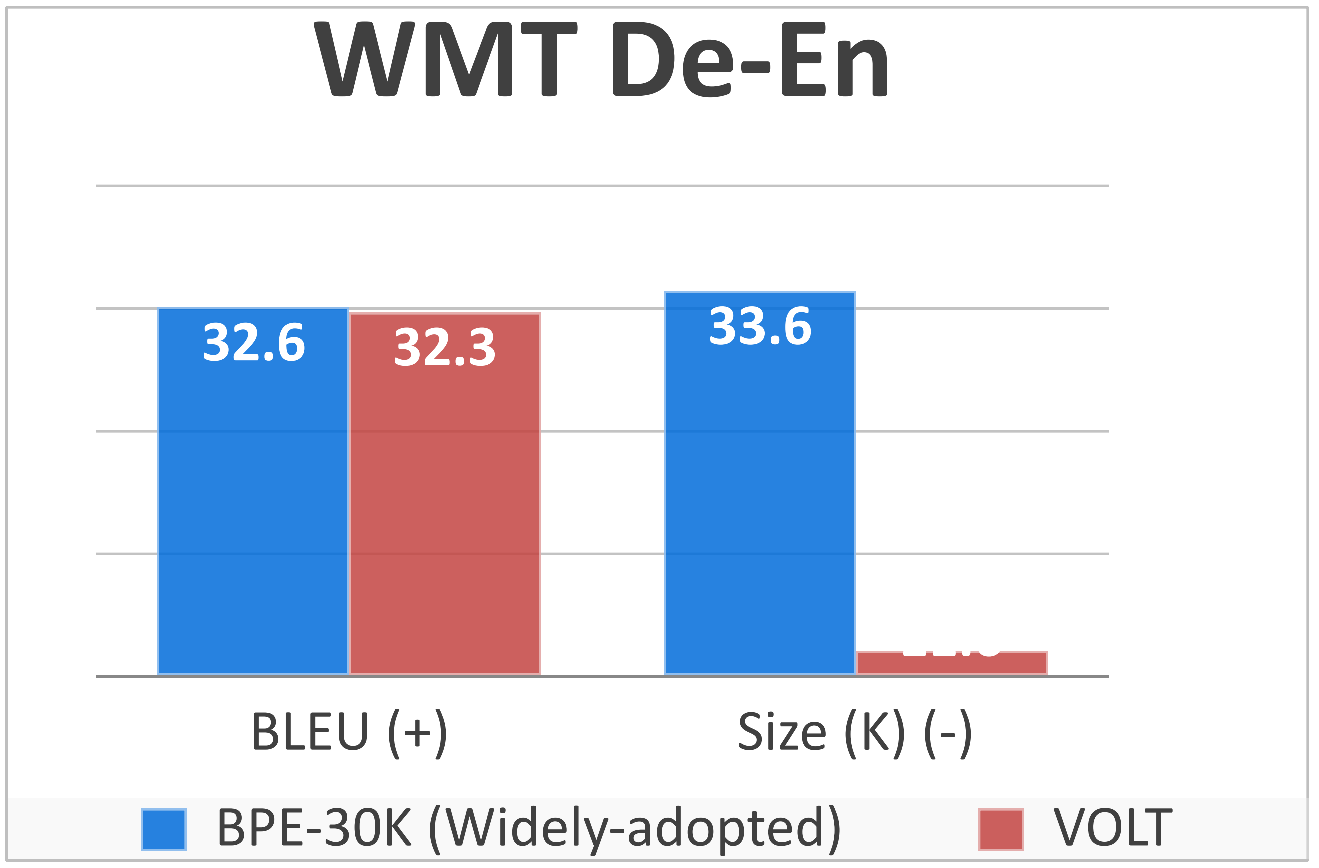
VOLT finds better vocabulary on Bilingual MT

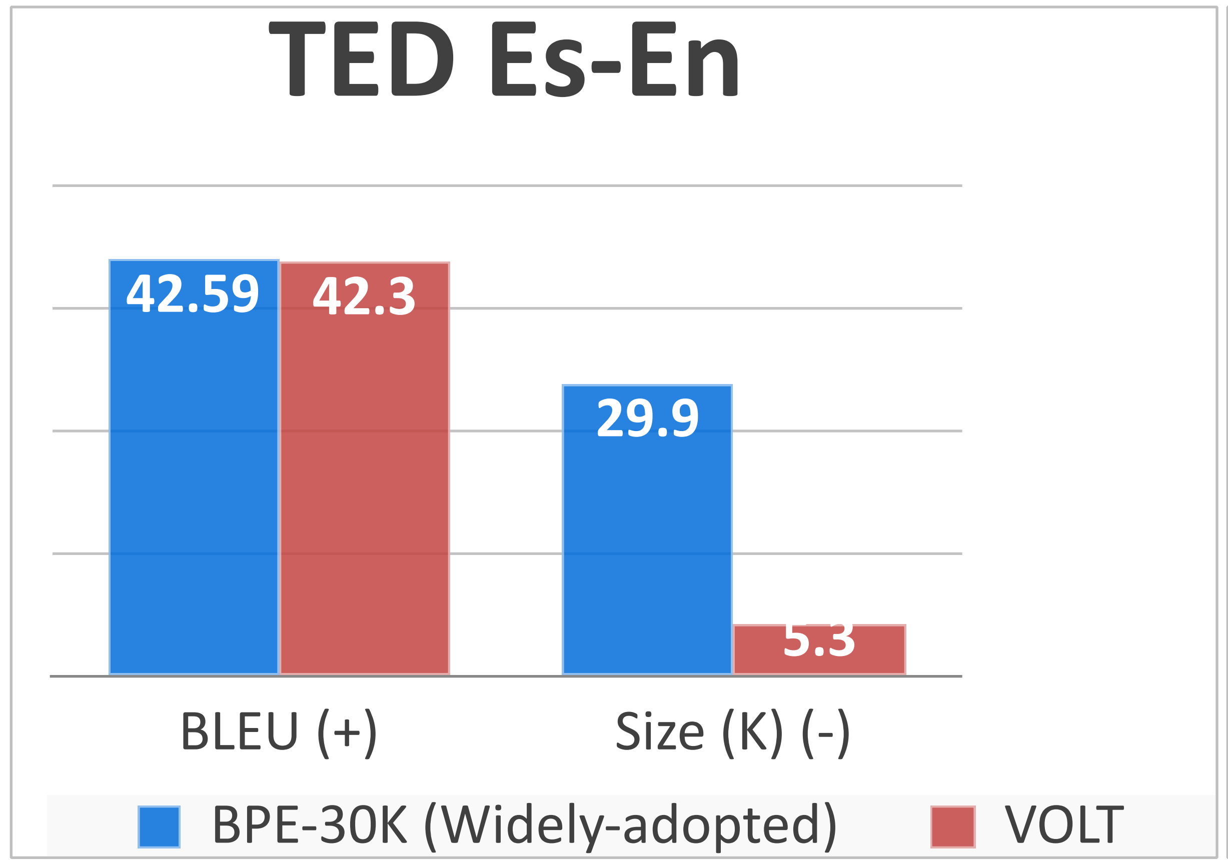


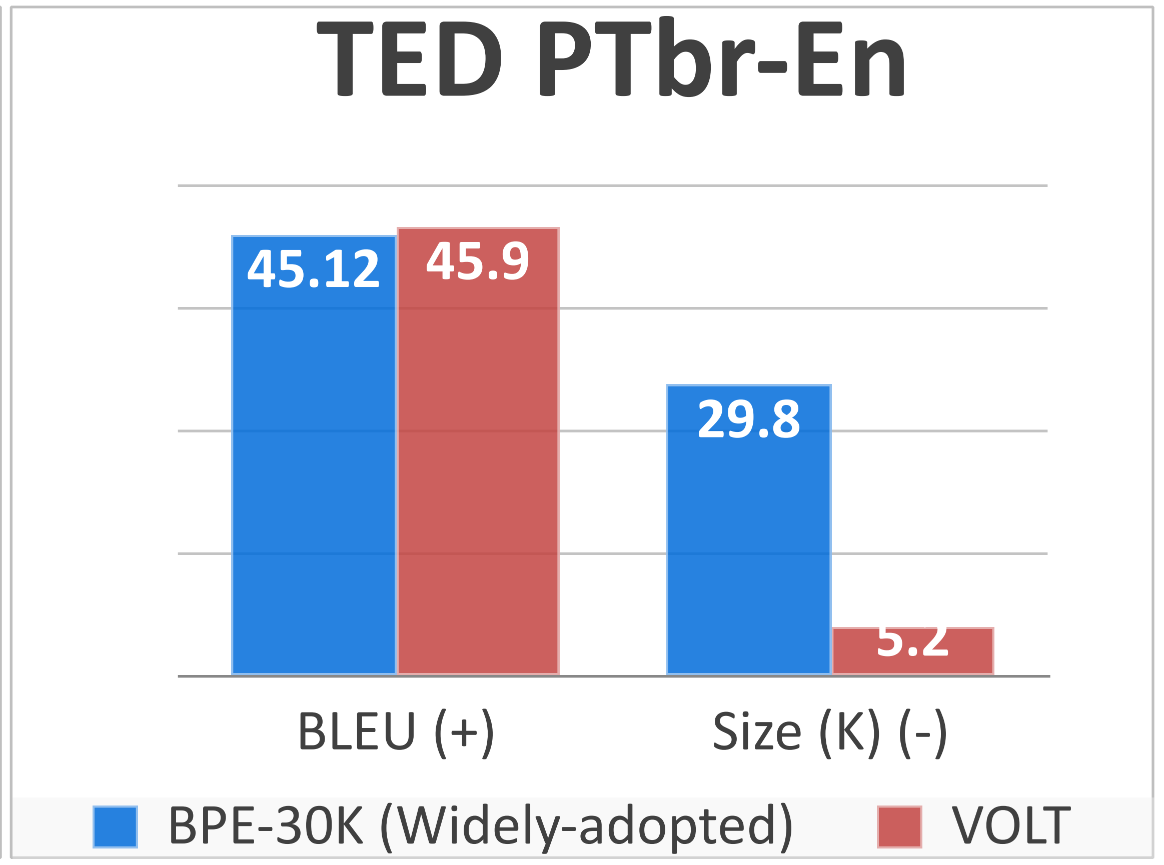
Xu, Zhou, Gan, Zheng, Li. Vocabulary Learning via Optimal Transport for Neural Machine Translation. ACL 2021.

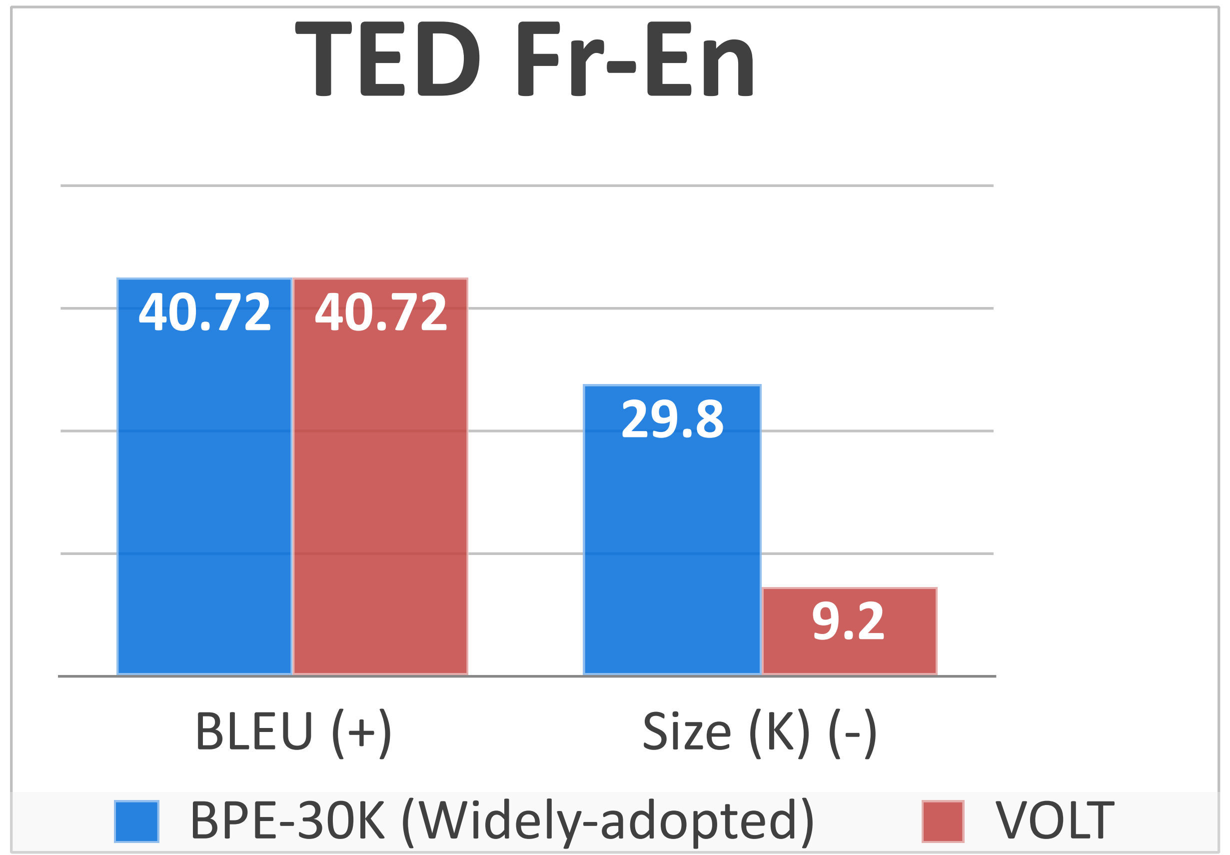
VOLT finds better vocabulary on Bilingual MT

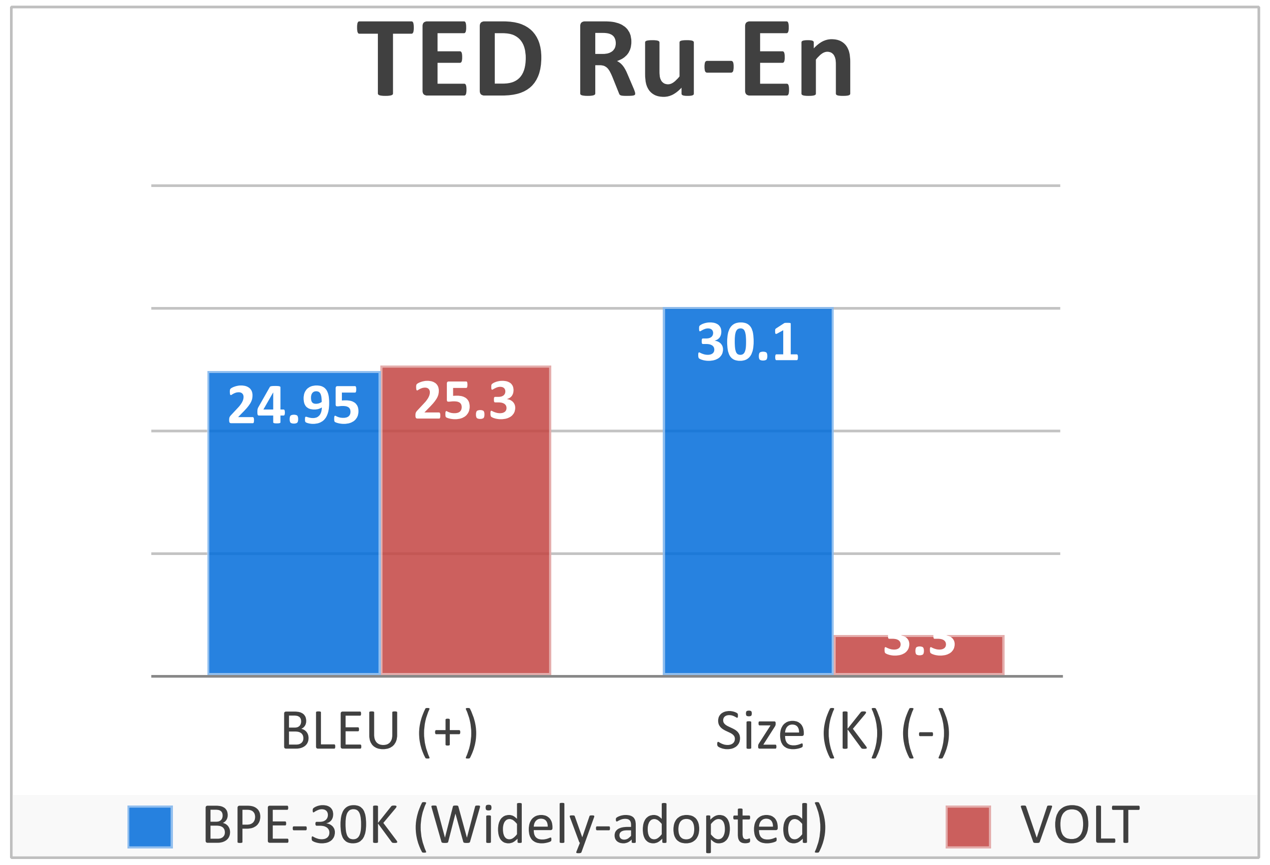
Transformer architecture

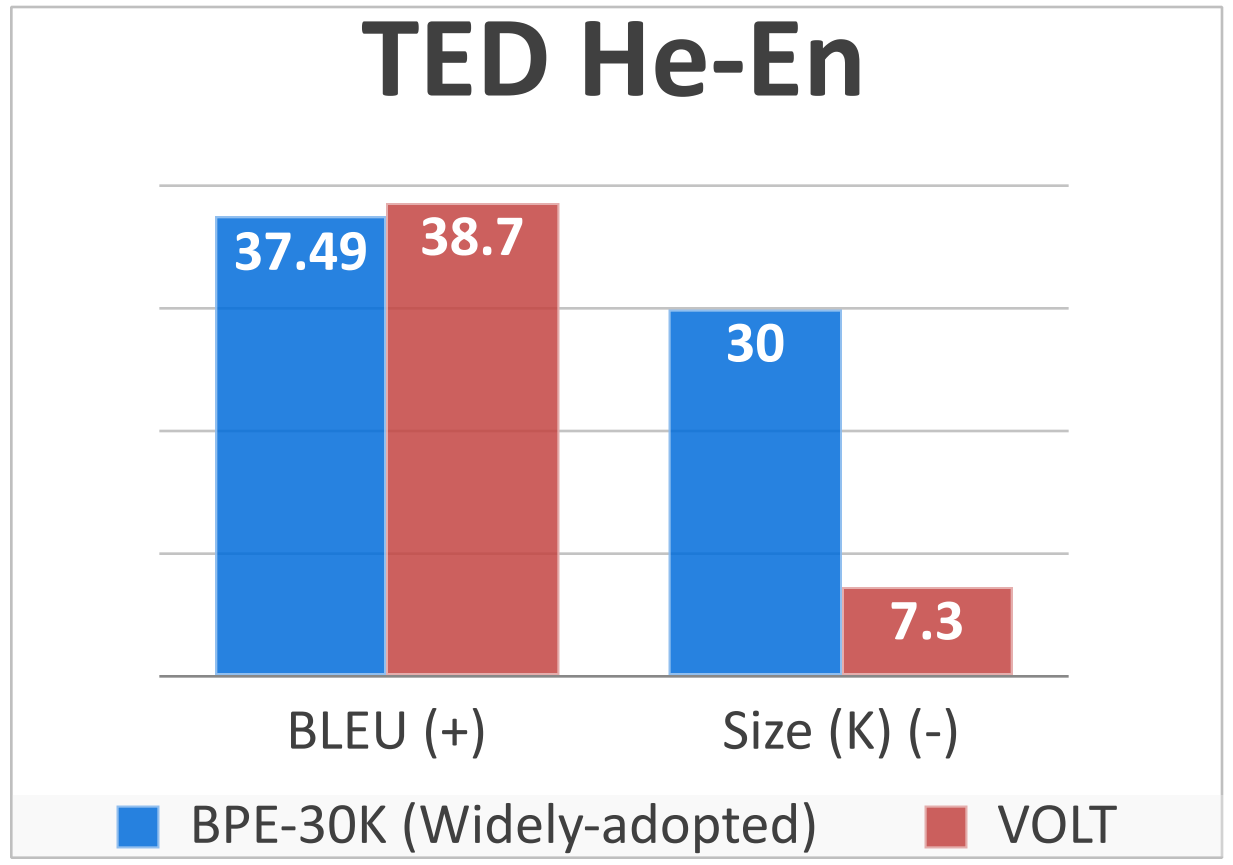


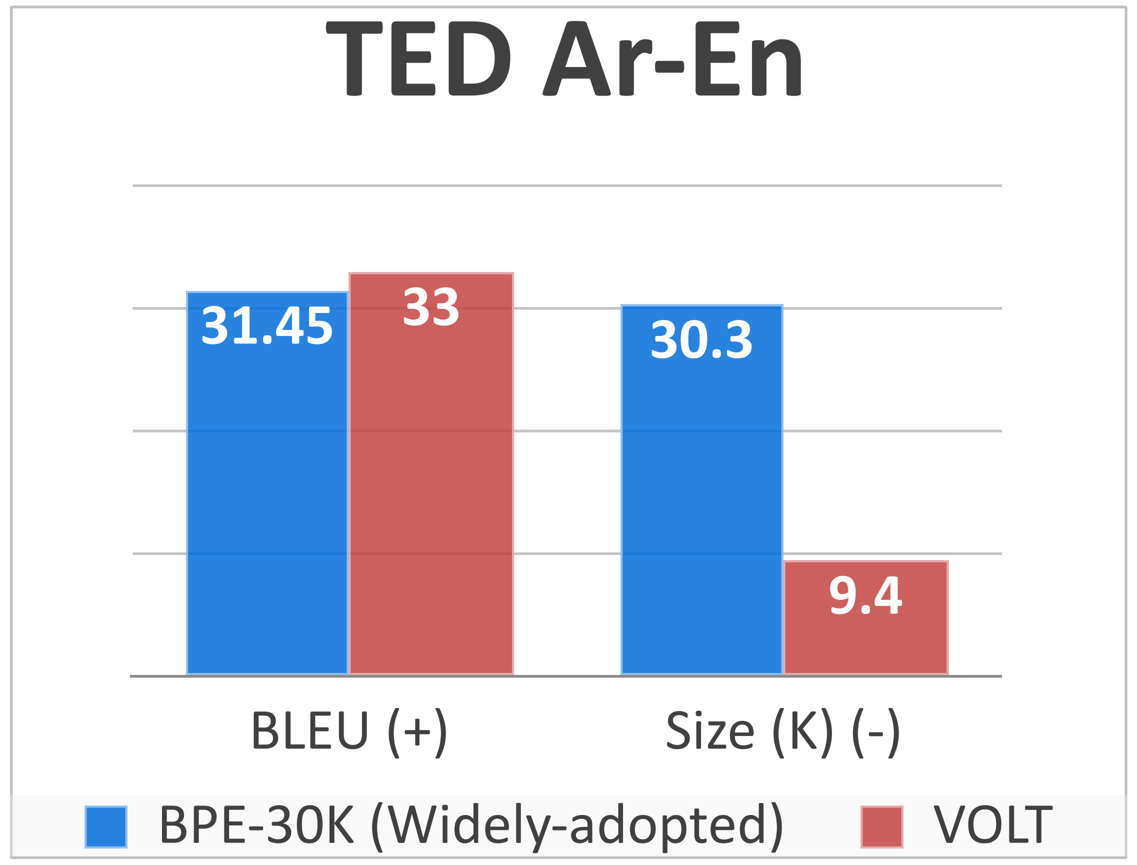


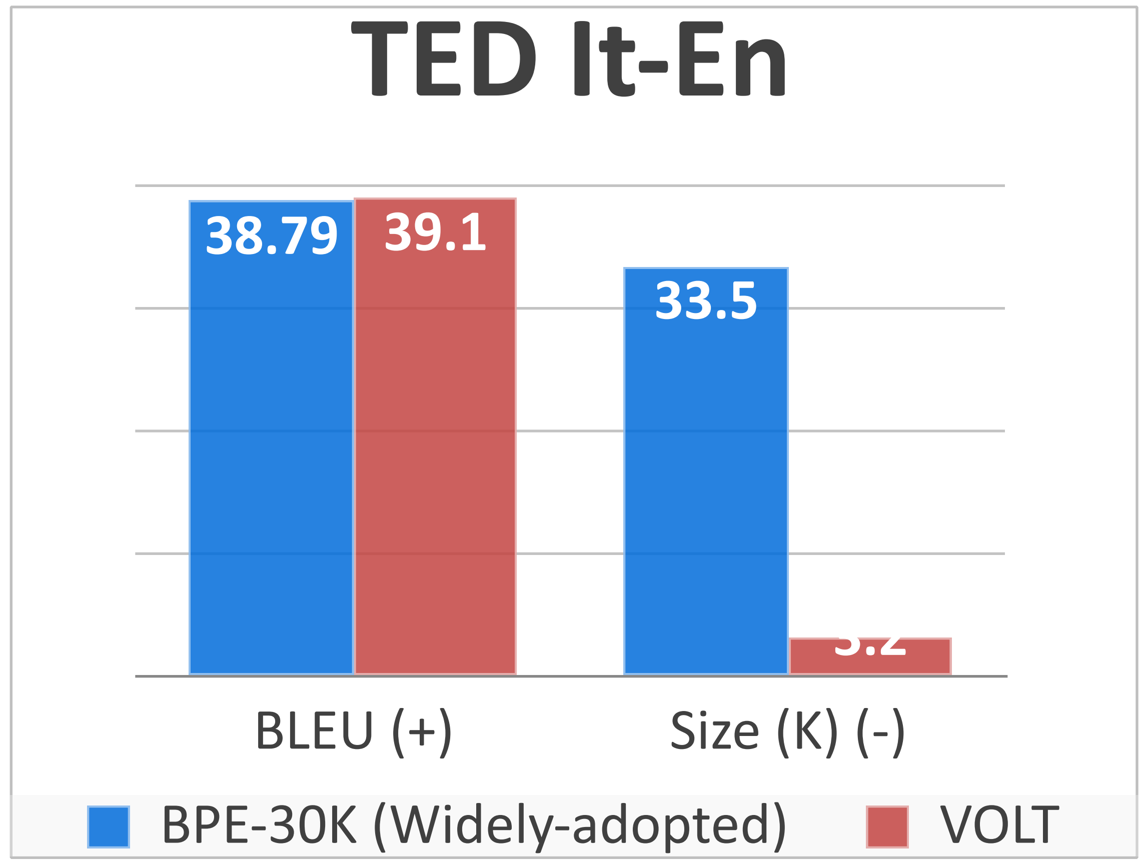












Xu, Zhou, Gan, Zheng, Li. Vocabulary Learning via Optimal Transport for Neural Machine Translation. ACL 2021.

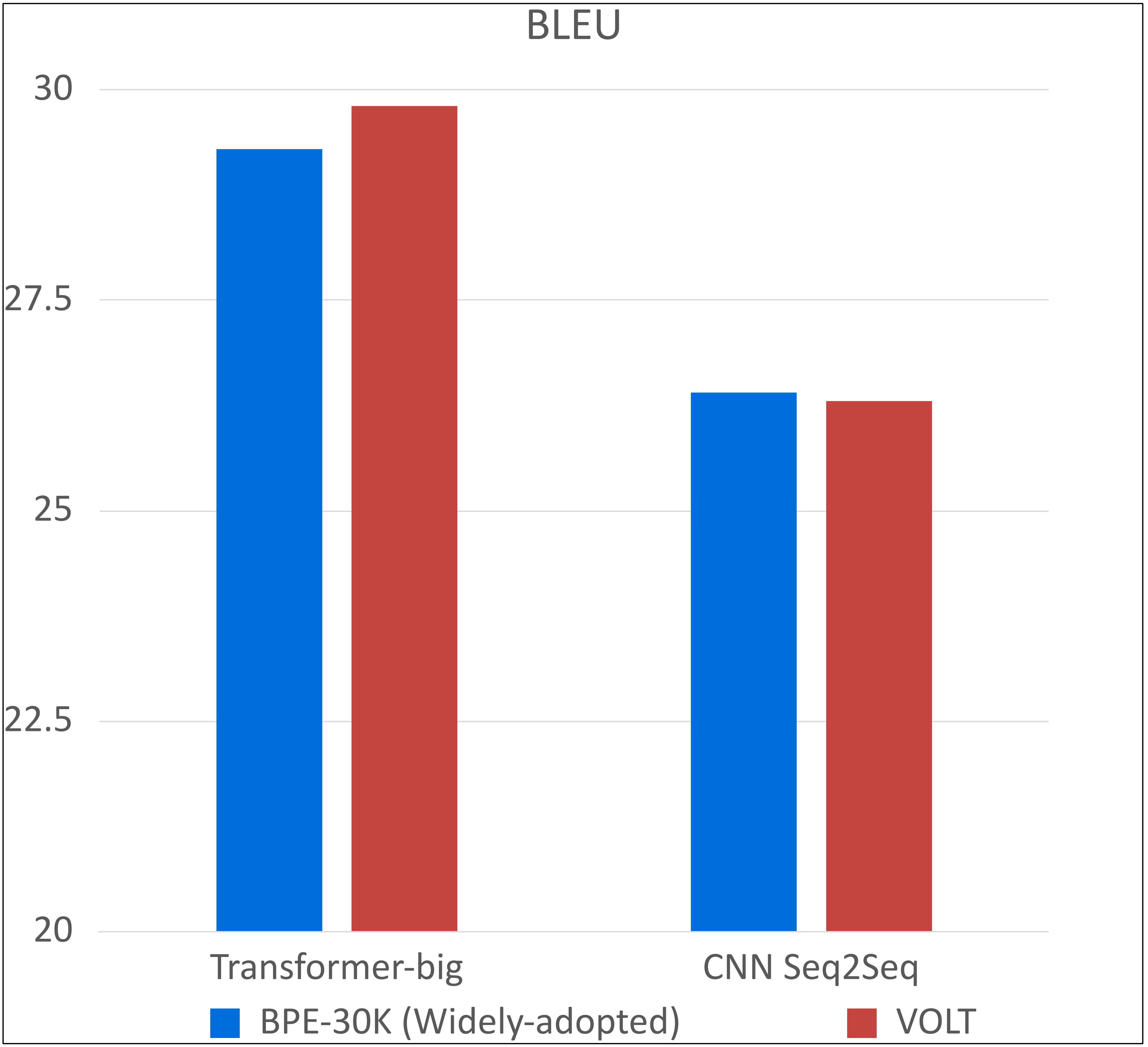
VOLT Finds Better Vocabulary on Multilingual MT

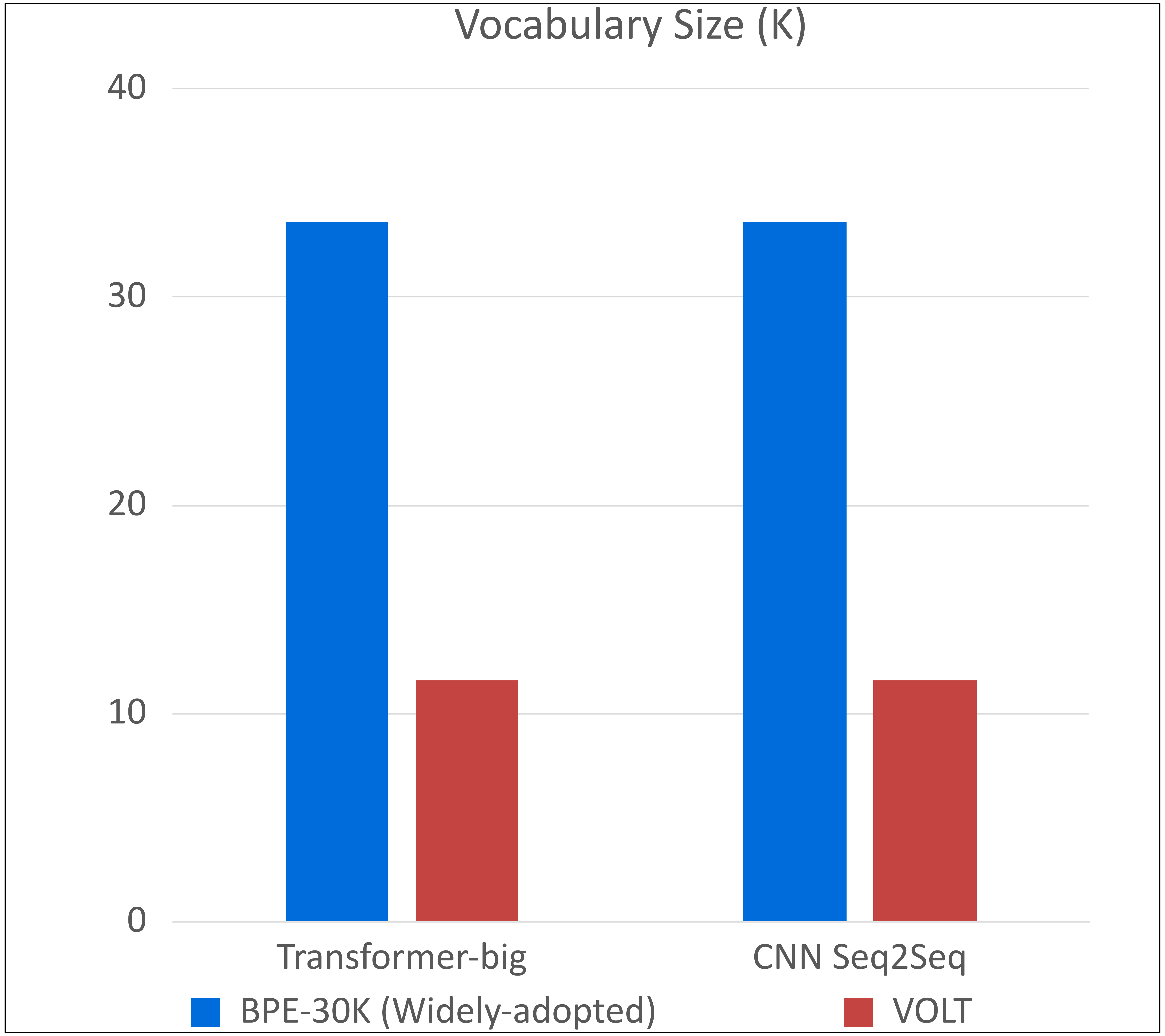
Xu, Zhou, Gan, Zheng, Li. Vocabulary Learning via Optimal Transport for Neural Machine Translation. ACL 2021.

VOLT Finds Better Vocabulary on Multilingual MT

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

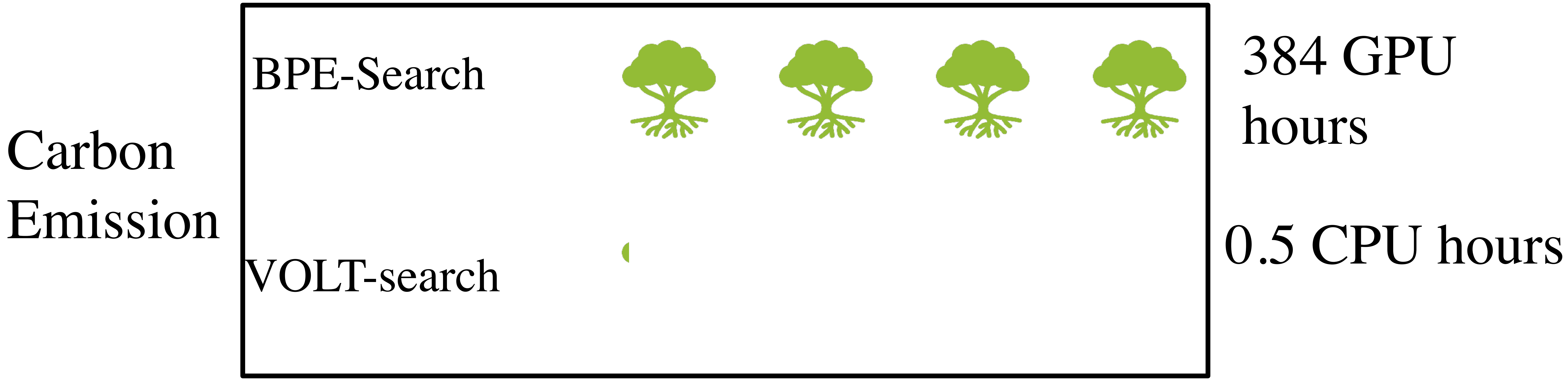
VOLT Generalizes Well to Other Architectures

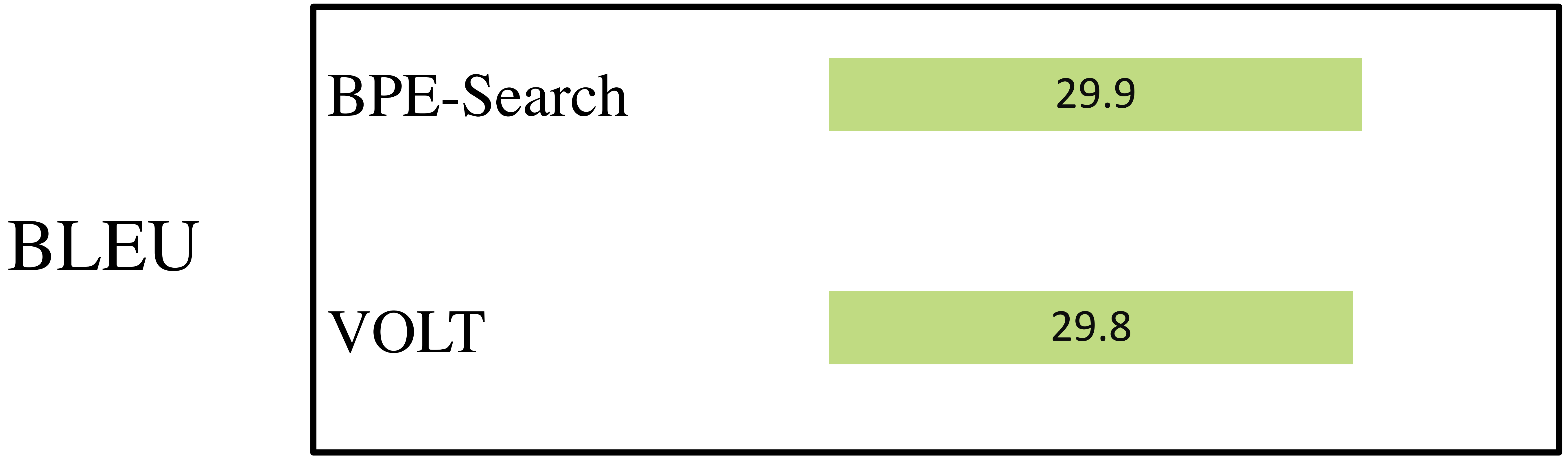




Xu, Zhou, Gan, Zheng, Li. Vocabulary Learning via Optimal Transport for Neural Machine Translation. ACL 2021.

VOLT is green!

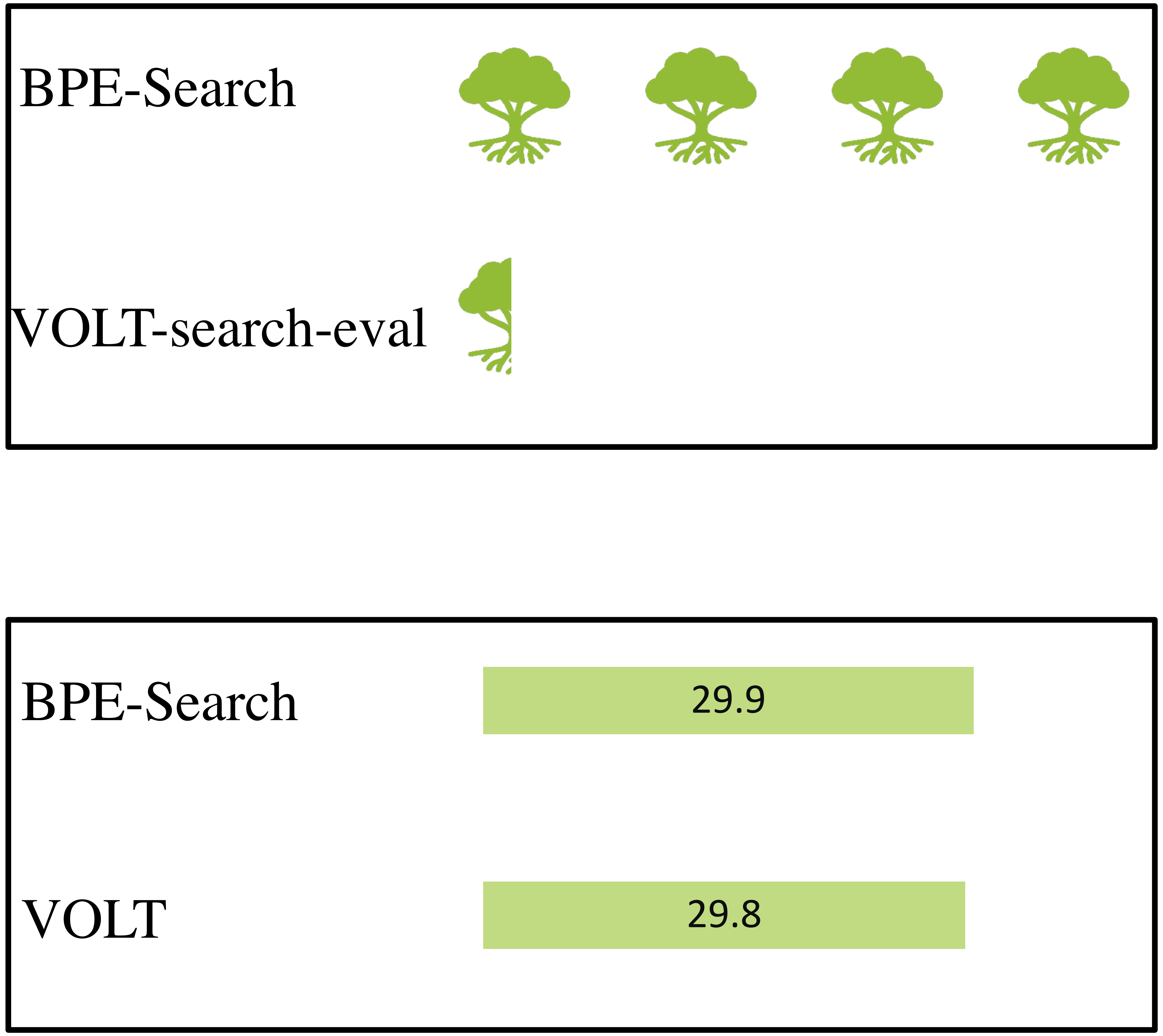




Xu, Zhou, Gan, Zheng, Li. Vocabulary Learning via Optimal Transport for Neural Machine Translation. ACL 2021.

Still need to perform one full training

Carbon Emission



384 GPU hours

0.5 CPU hours + 30GPU hours

How to reduce this?

BLEU

Conclusion

- How to evaluate vocabularies **without trial training**?
  - Better vocabulary should have less information-per-char (IPC)
  - Better vocabulary should have smaller size
  - MUV metric
- How to **efficiently** find the optimal vocabulary?
  - Reduce to OT
  - A green vocabulary learning solution

Code and Blog

- Codes and data are available at:
  - https://github.com/Jingjing-NLP/VOLT
- If you have more questions on paper details, please see our latest paper blog at:
  - https://jingjing-nlp.github.io/volt-blog/

Language Presentation

Read List

- Xu, Zhou, Gan, Zheng, Li. Vocabulary Learning via Optimal Transport for Neural Machine Translation. ACL 2021. (ACL best paper)