CS11-737 Multilingual NLP

Vocabulary Learning

Lei Li

https://lileicc.github.io/course/11737mnlp23fa/

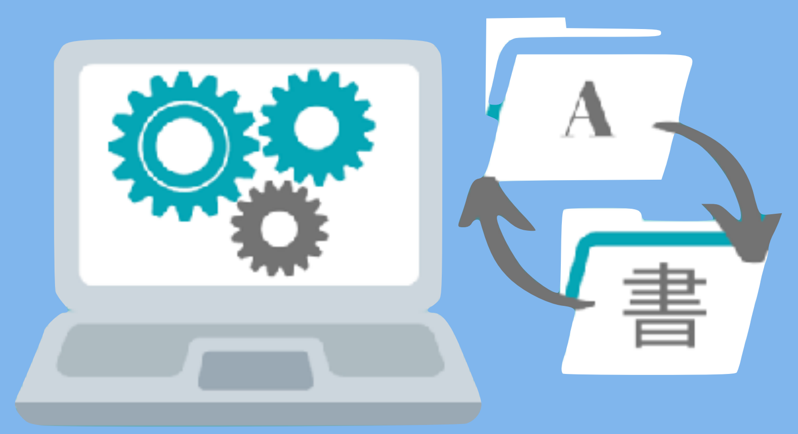
Carnegie Mellon University
Language Technologies Institute

Vocabulary is Fundamental and Important

NER

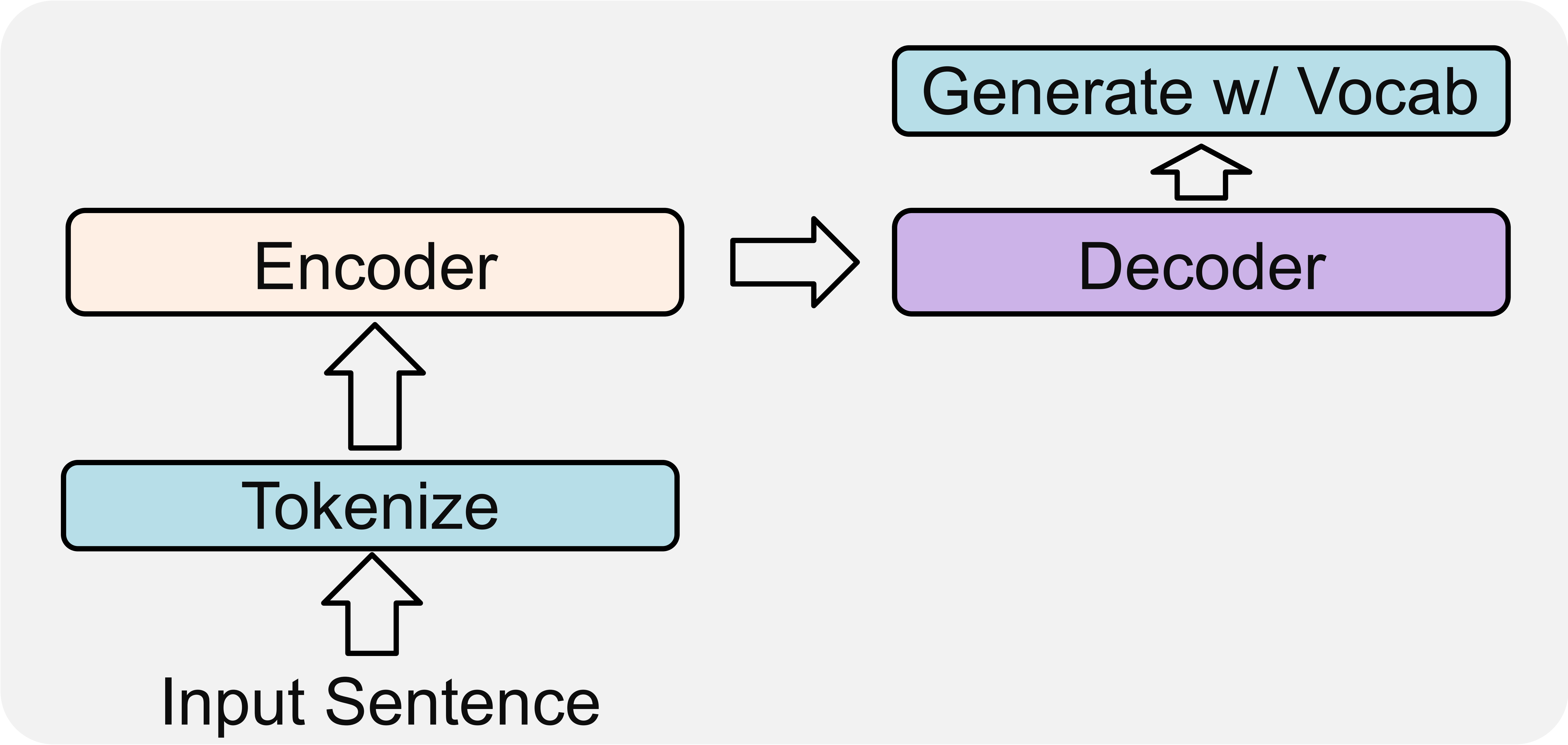
Sentiment Analysis

Translation



Dialog

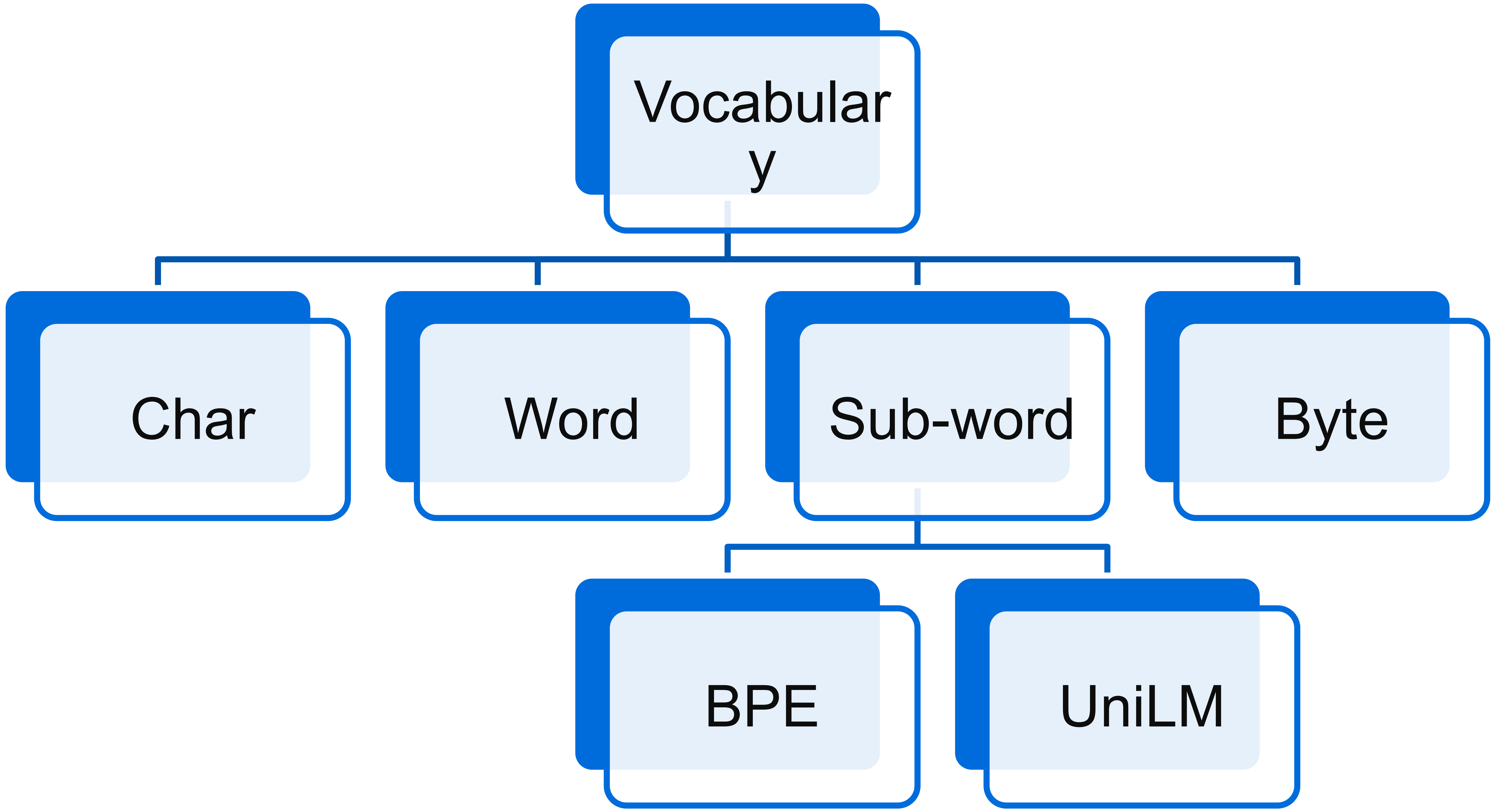
Summarization



Vocab

Token	ID
a	0
es	1
cat	2
…	…

Methods to Construct Vocabulary



Word level

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

Vocabulary

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 | o | n | _ | …

Sub-word level

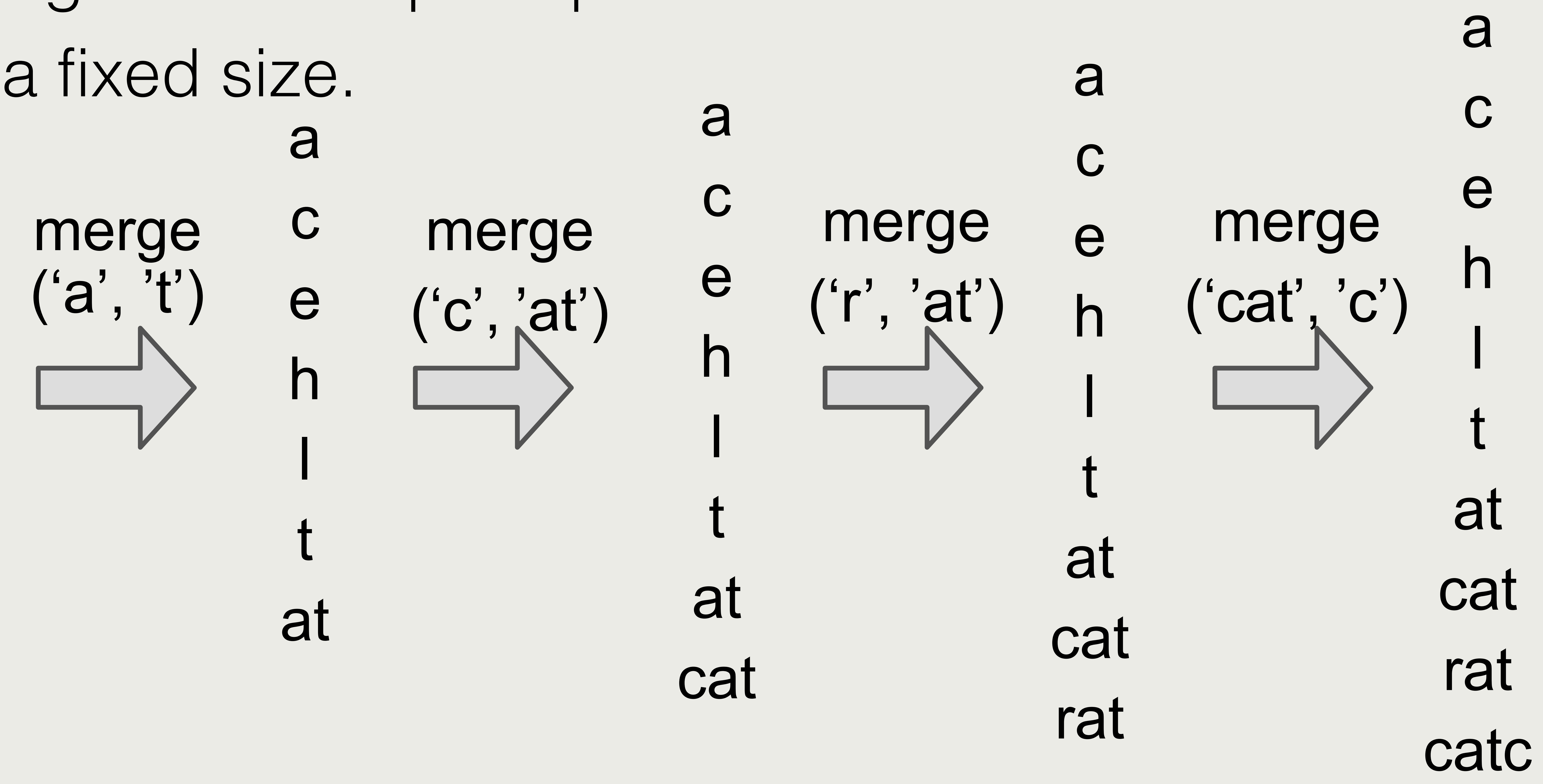
The | most | eager | is | O | re | go | n | which | is | en | list | ing | 5 | , | 000 | driver | s | in | the | country

Sub-word vocabulary is the dominant choice

Recap Sub-word: Byte-Pair-Encoding

- Byte-Pair-Encoding (BPE)
  - starting from chars
  - repeatedly, merge most frequent pairs to form new tokens
  - until reaching a fixed size.

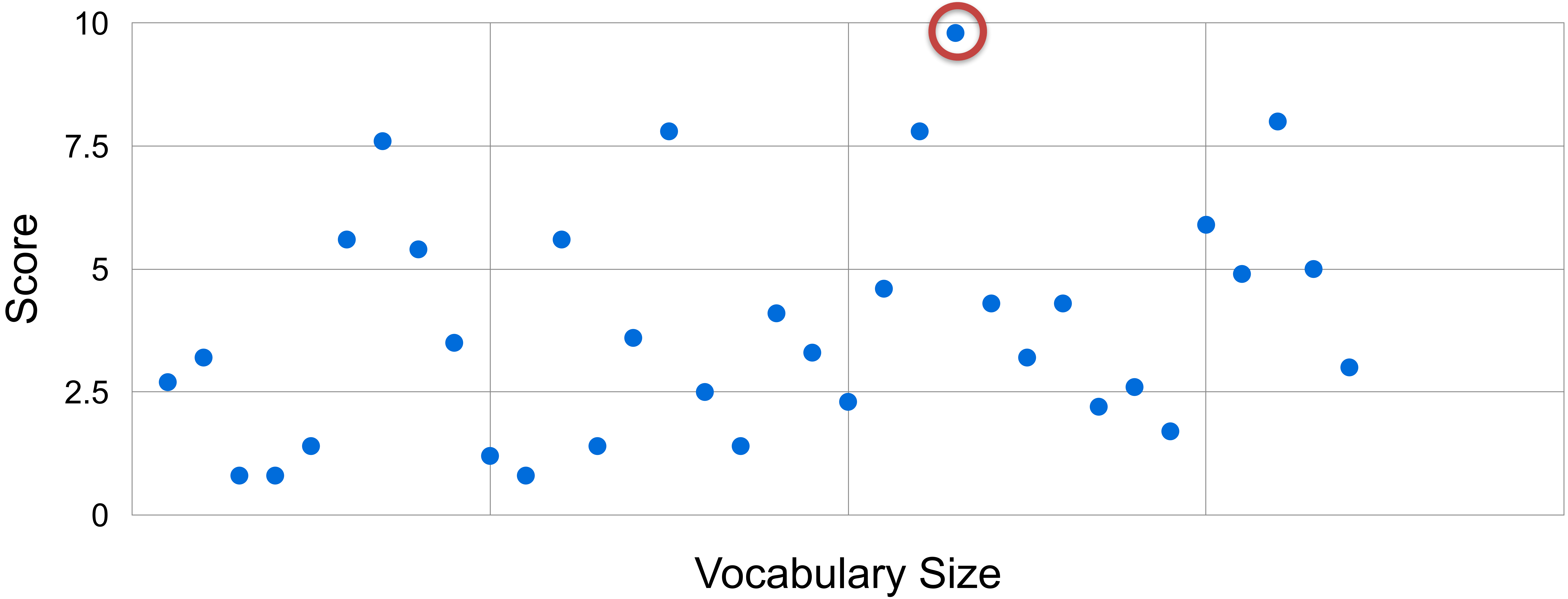
raw word	freq.
cat	90
catch	50
rat	80
rattle	40



Neural Machine Translation of Rare Words with Subword Units. Sennrich et al. ACL 2016

Finding the Optimal Vocabulary

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



2021 ACL Best Paper Award 🏆

Proposed Solution: Vocabulary Learning via Optimal Transport for Neural Machine Translation

Jingjing Xu Hao Zhou Chun Gan Zaixiang Zheng Lei Li





Q1
How to evaluate vocabulary?

Challenge: Finding Optimal Vocabulary

- Vocabulary is a tuning hyperparameter
- On different task and corpus, the best vocabulary is different
- Existing method: BPE-search
  - Computational expensive: 384 hours on GPU for MT (De-En)
- Challenging due to the huge search space

BPE-Search

1. Enumerating choices of vocabulary (BPE 1k, 2k, 3k, …, 100k,)
2. Evaluating quality through full training and testing.
3. Pick the best one based on translation performance (BLEU score)

Vocab 1k tokens | Vocab 10k tokens | Vocab 50k tokens

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?

Our New Observation on Sub-words

- Normalized Entropy (modified based on Information Entropy) $\mathscr{H}(v) = -\frac{1}{l_v}\sum_{i\in v} P(i)logP(i)$ ← token prob.

l_v: average number of chars for v's all tokens

- It measures semantic-information-per-char
  - Smalle… e. Less ambiguity and eas… te

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

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

Token	count
a	100
aes	90
cat	30

$\mathscr{H}(v) = 0.14$

😁

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 | age r | is | O | reg | o n | which | is | e n | listin g | 5,000 | dr i | ver s | in | the | country

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

The | most | e | age r | is | O | reg | o n | which | is | e n | listing | 5,000 | drivers | in | the | country

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

* With normal-size data

A Dilemma in Selecting the Best Vocabulary

Numerous possible vocabularies at the sub-word level.

Normalized Entropy

Vocab 1k tokens

Vocab 10k tokens

Vocab 30k tokens

Size

Which one leads to better MT performance?

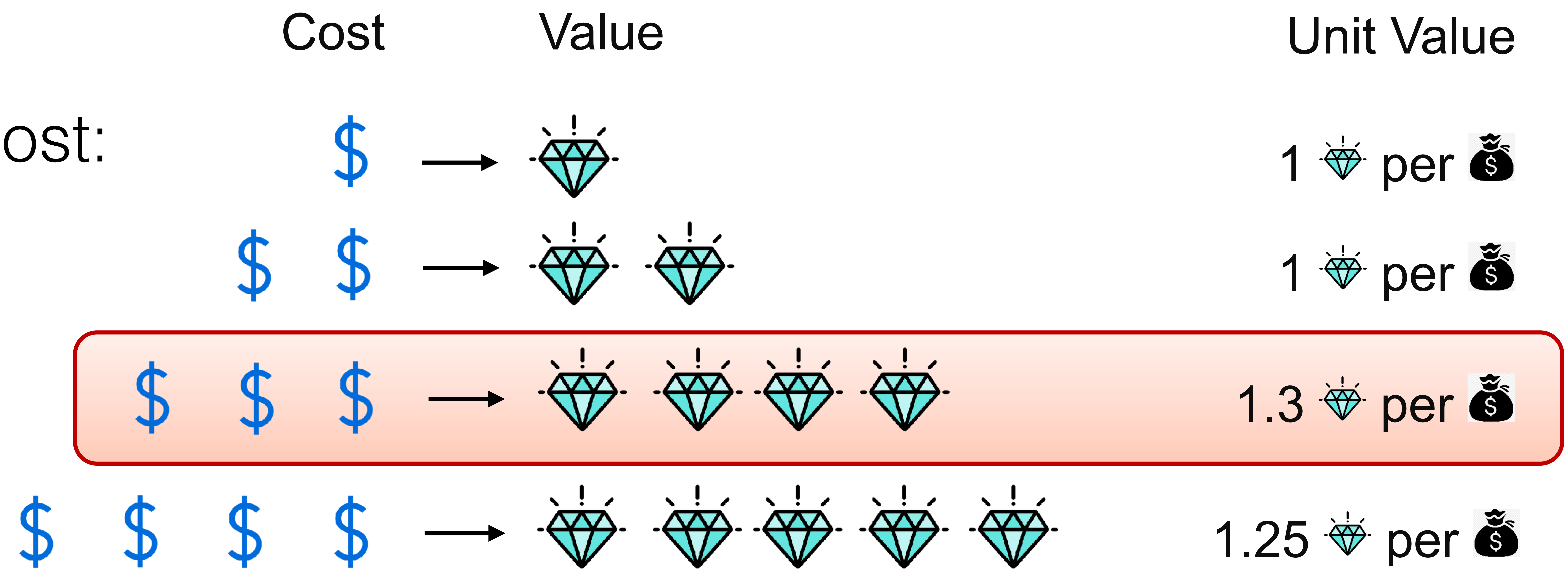
Repeated full training and testing are required to find the optimal vocabulary!

(BPE-Search)

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

An Analogy: Buying Products with Money

- Value:
- Cost:



Optimal when marginal utility is maximized!

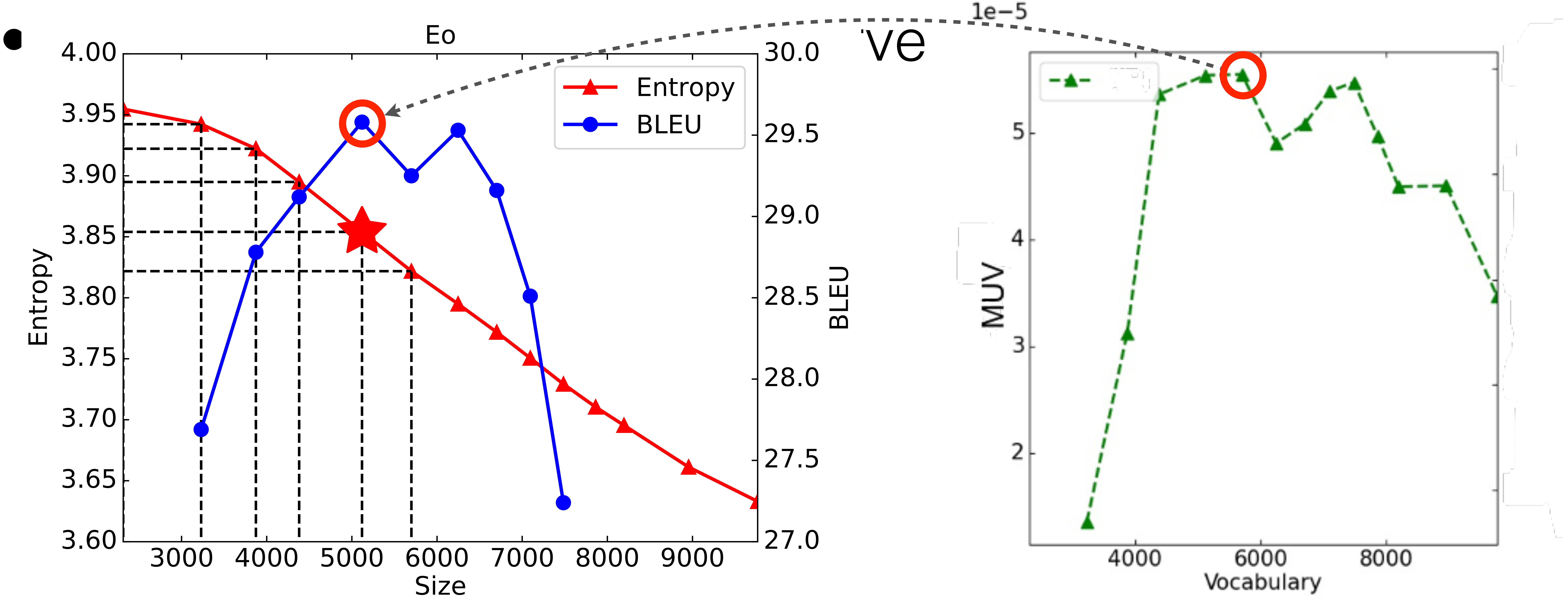
Proposed VOLT: Utility of Information for Adding Tokens

- Value: Normalized Entropy
- Cost: Size
- Marginal Utility of information for Vocabulary (MUV)

$$M_{v_k \to v_{k+m}} = -\frac{H(v_k) - H(v_{k+m})}{m}$$

  - Negative **gradients** of normalized entropy to size
  - How much value each token brings

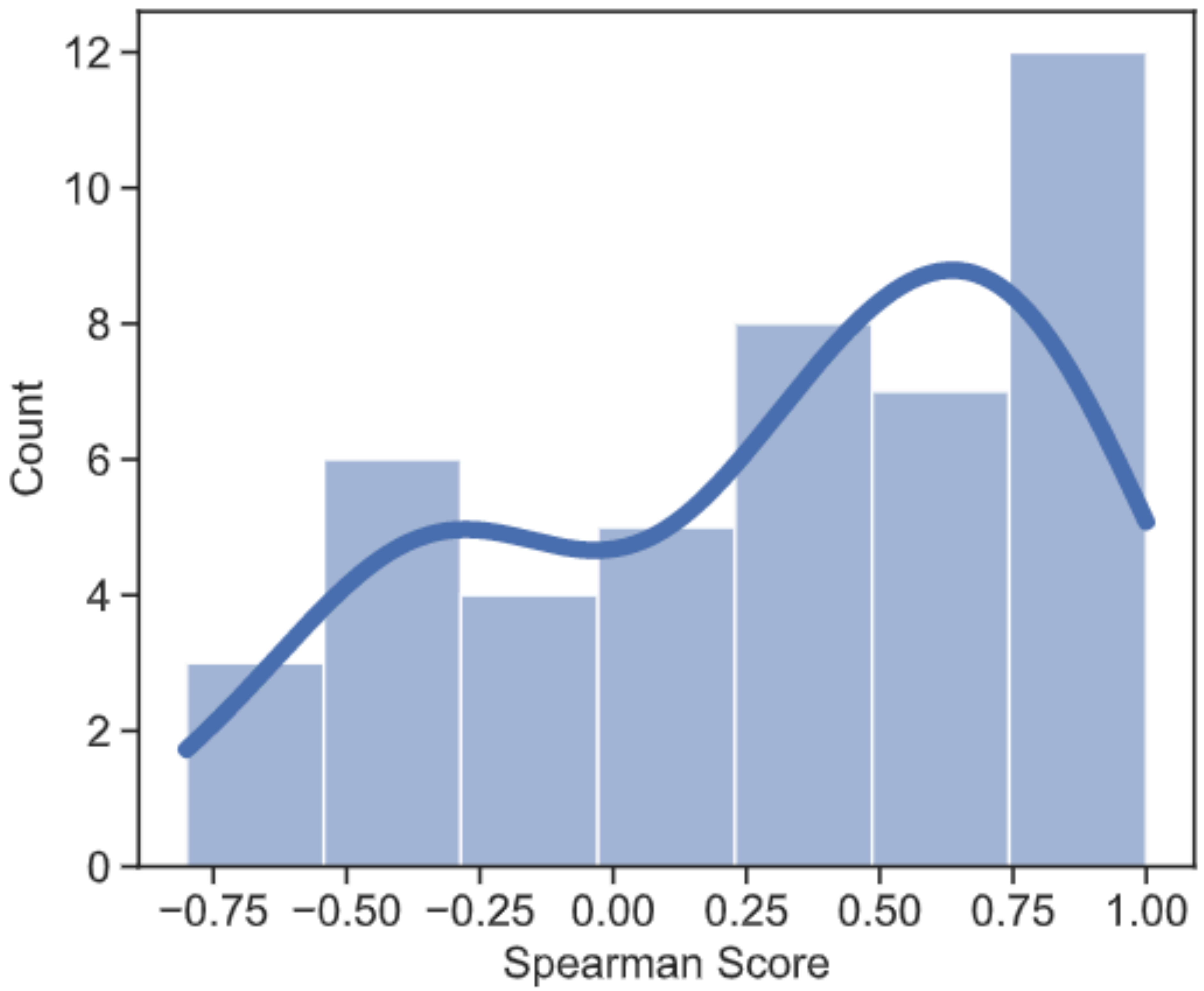
MUV is good indicator for MT performance

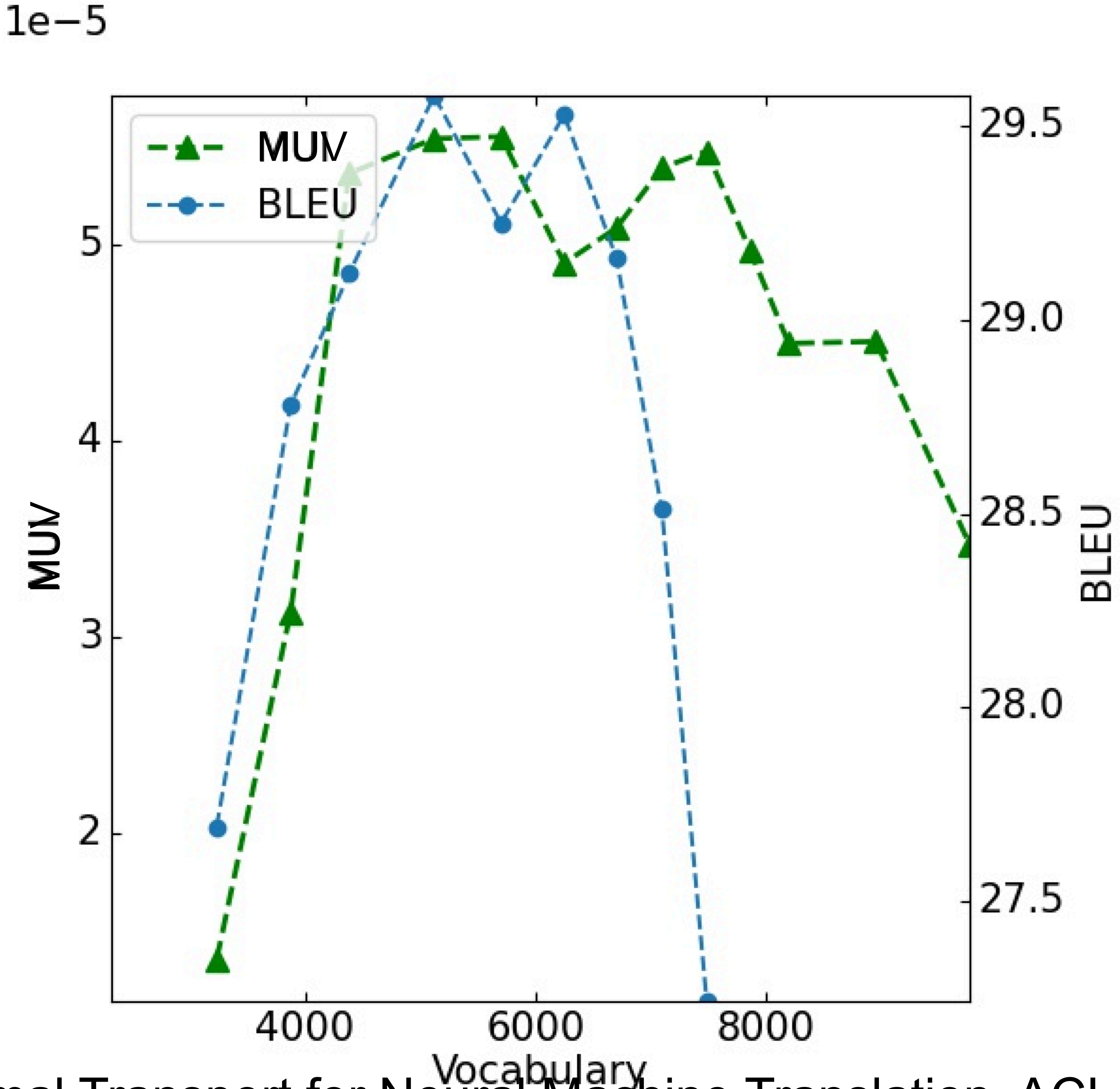


Xu, Zhou, Gan, Zheng, **Lei Li**. Vocabulary Learning via Optimal Transport for Neural Machine Translation. ACL 2021a

MUV Indicates MT Performance

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

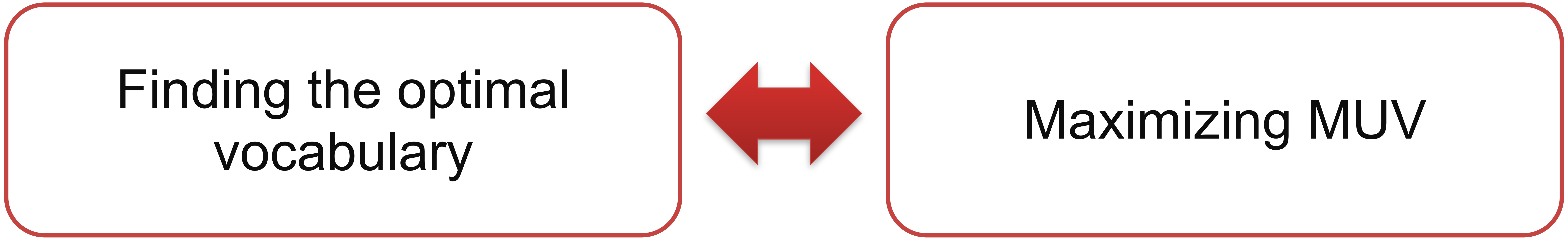




Xu, Zhou, Gan, Zheng, **Lei Li**. Vocabulary Learning via Optimal Transport for Neural Machine Translation. ACL 2021a

Proposed VOLT: Problem Reduction

- Goal: finding the optimal vocabulary



- MUV can be estimated efficiently.

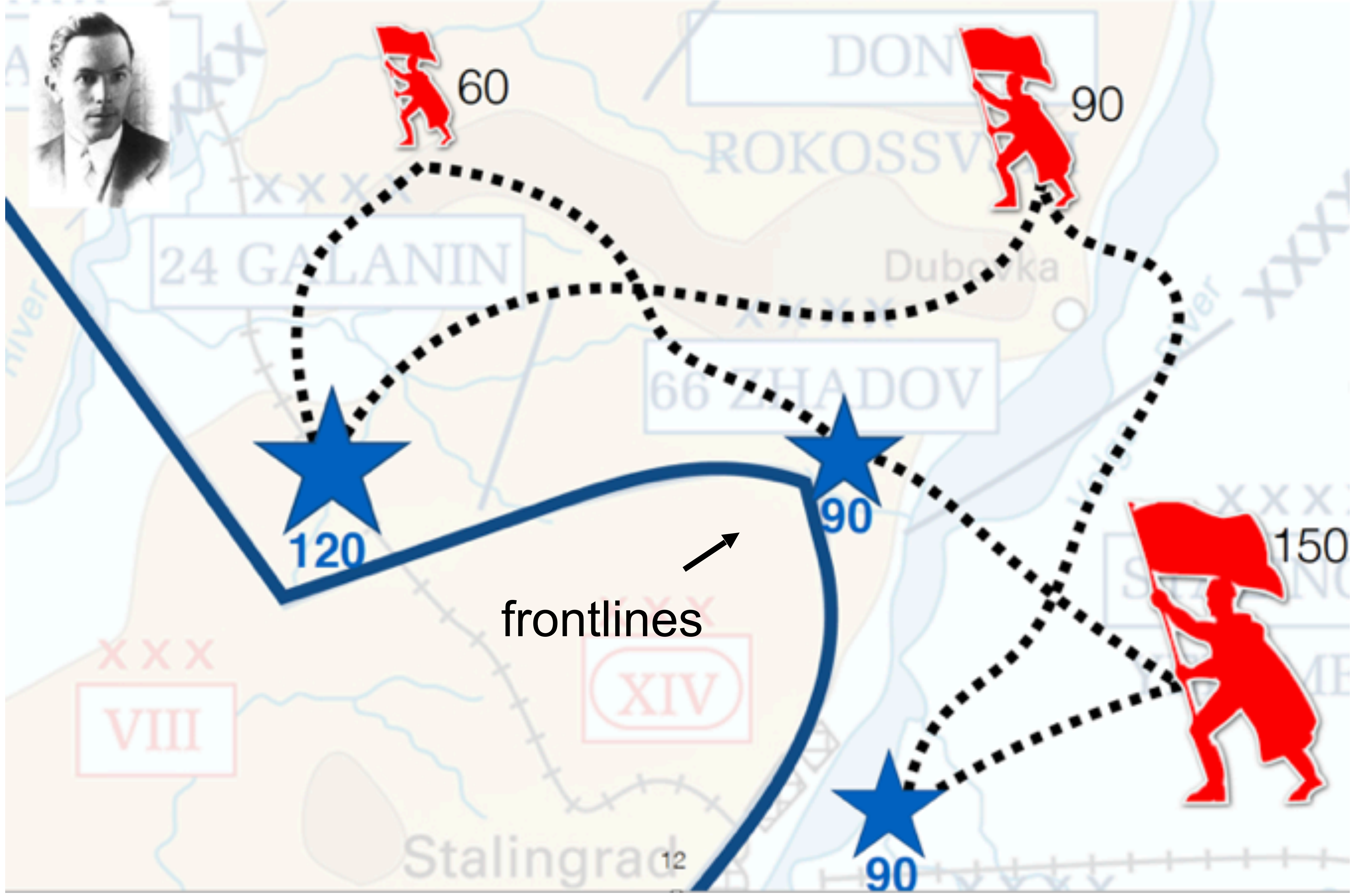
- How to find the vocabulary maximizing MUV?
    - Huge search space over possible vocabularies

Q2
How can we find the optimal vocabulary?

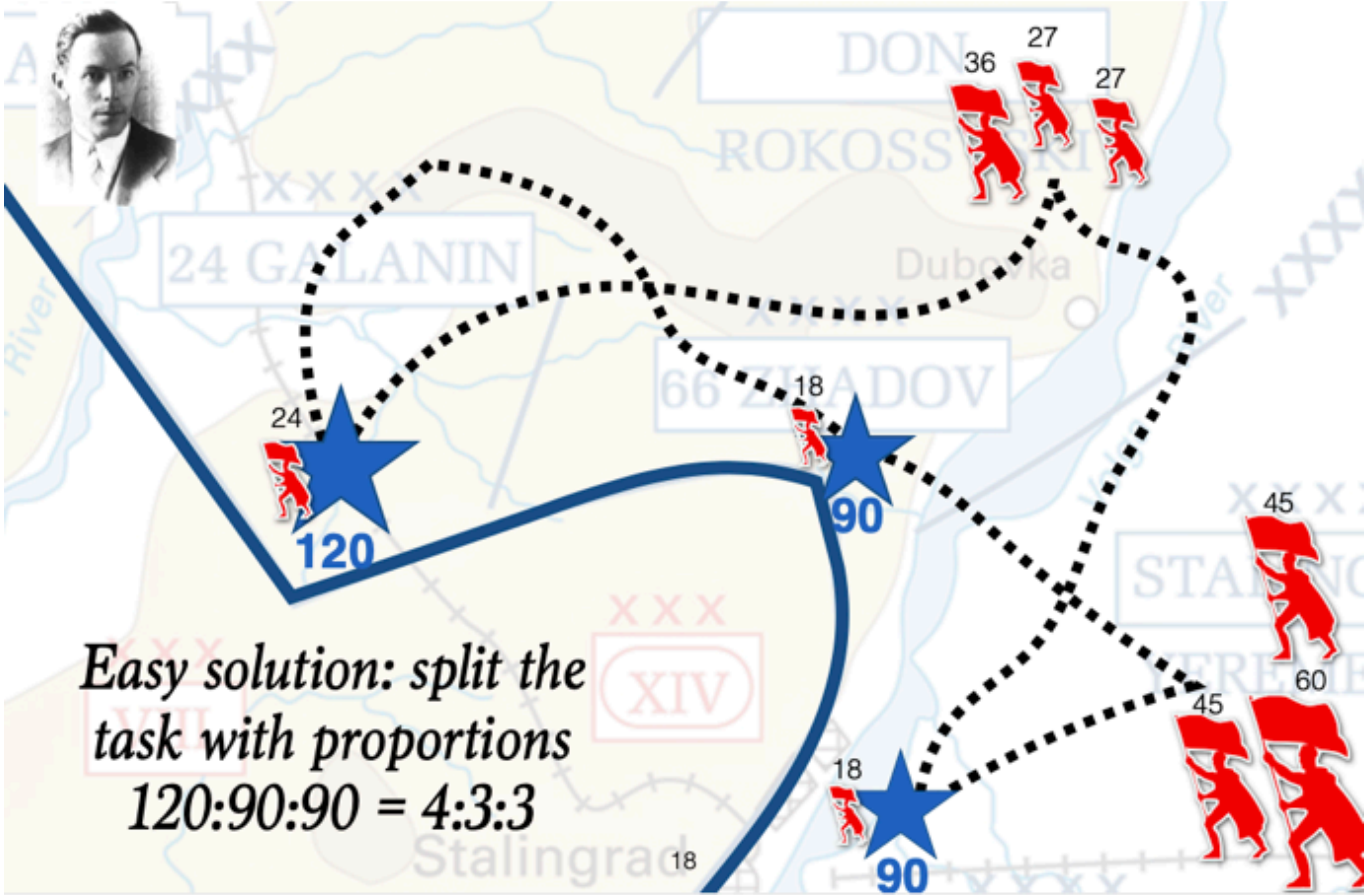
Proposed VOLT: Problem Reduction

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

Min cost to Transport soldiers from bases to frontlines

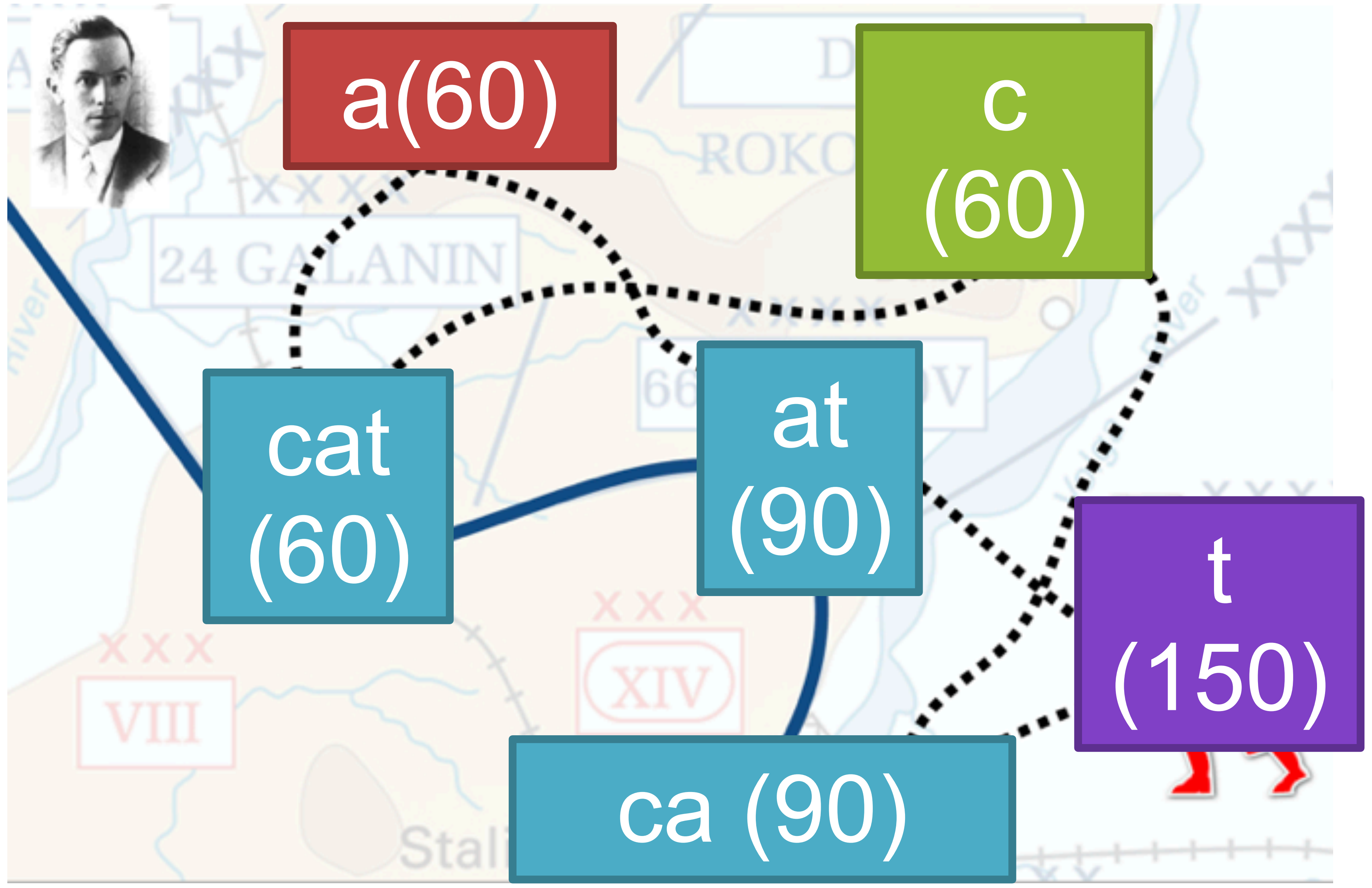


Optimal Transport



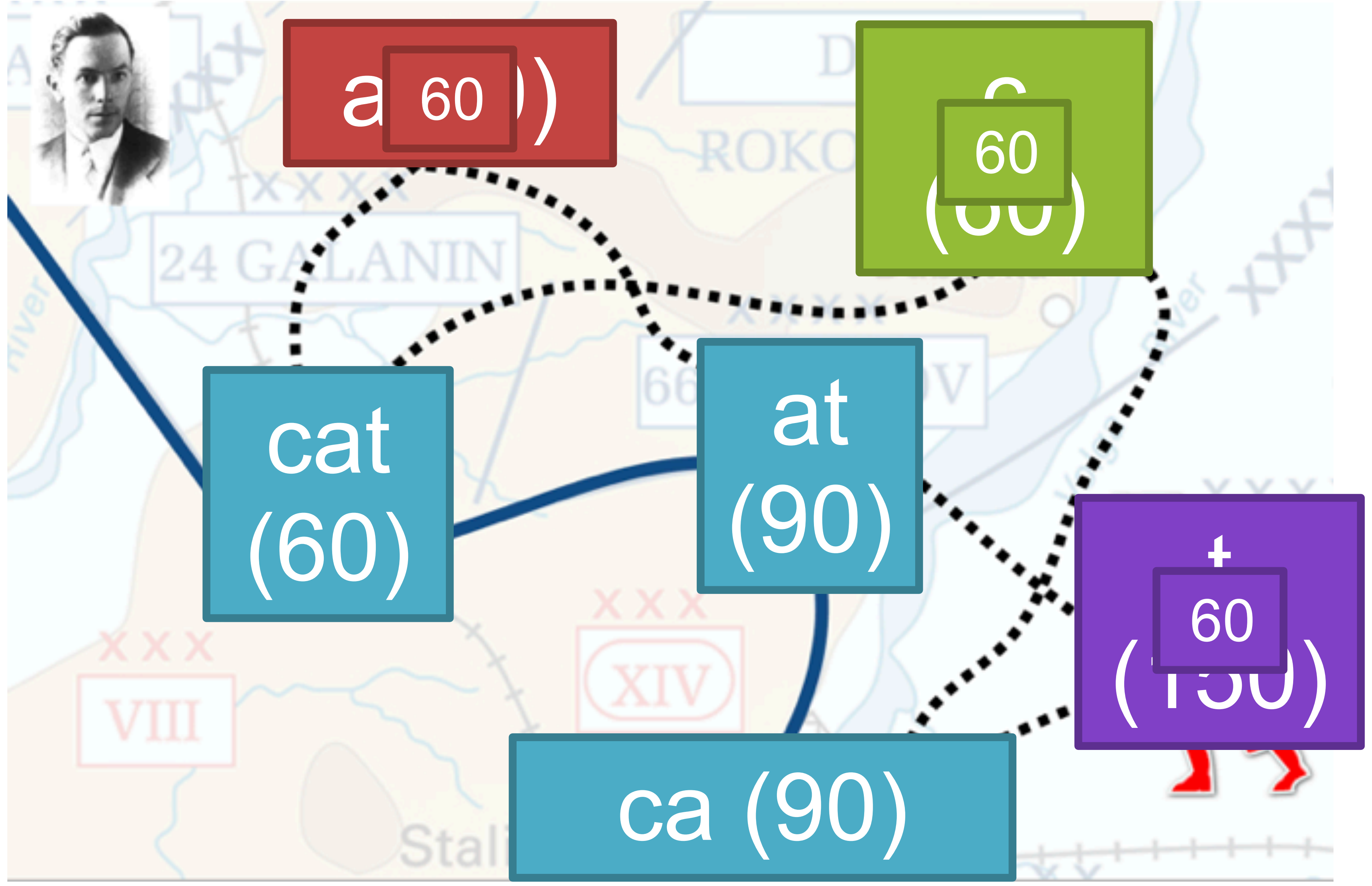
VOLT Formulation

Transport chars to tokens



VOLT Formulation

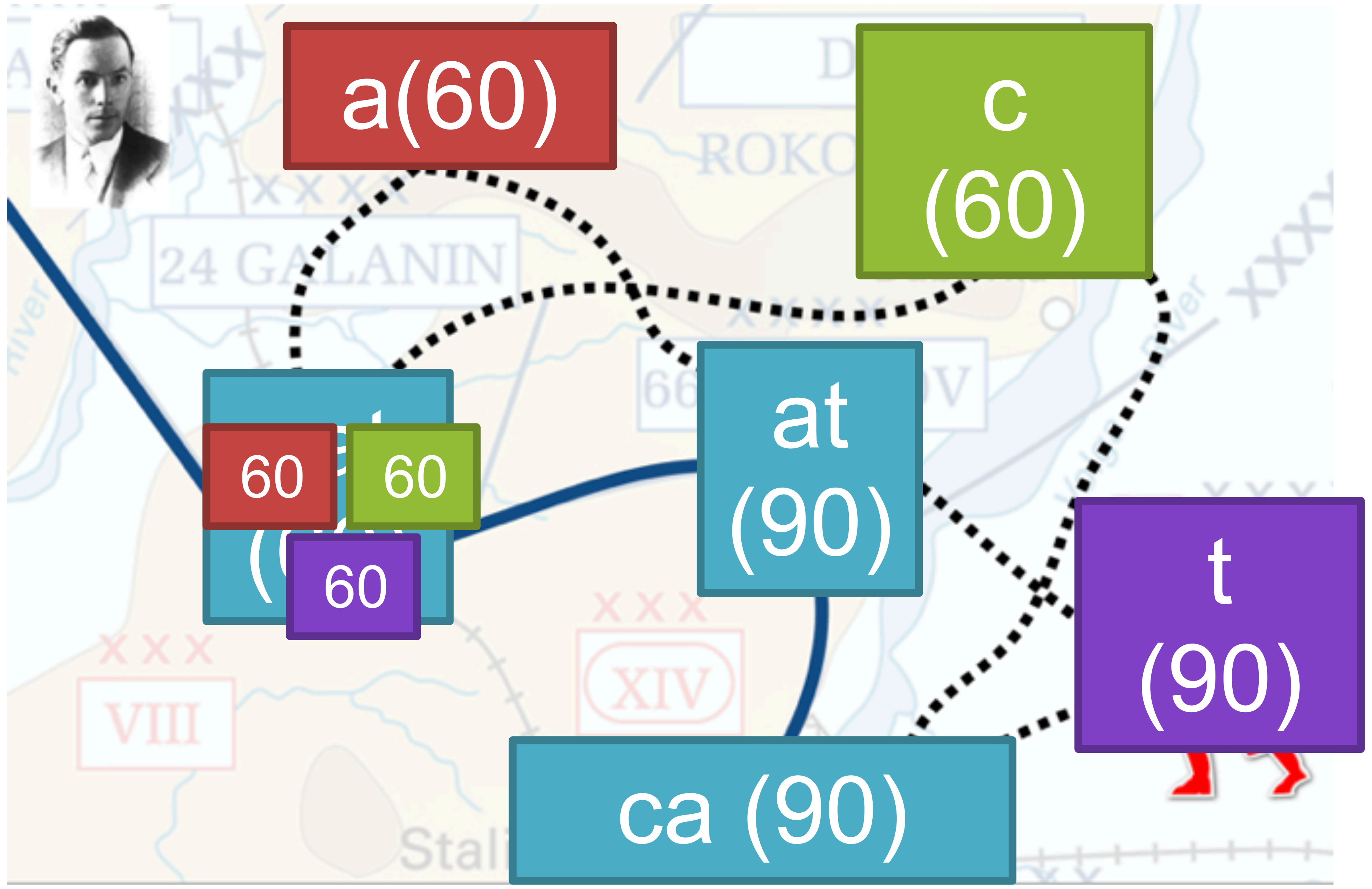
Not all tokens can get chars



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

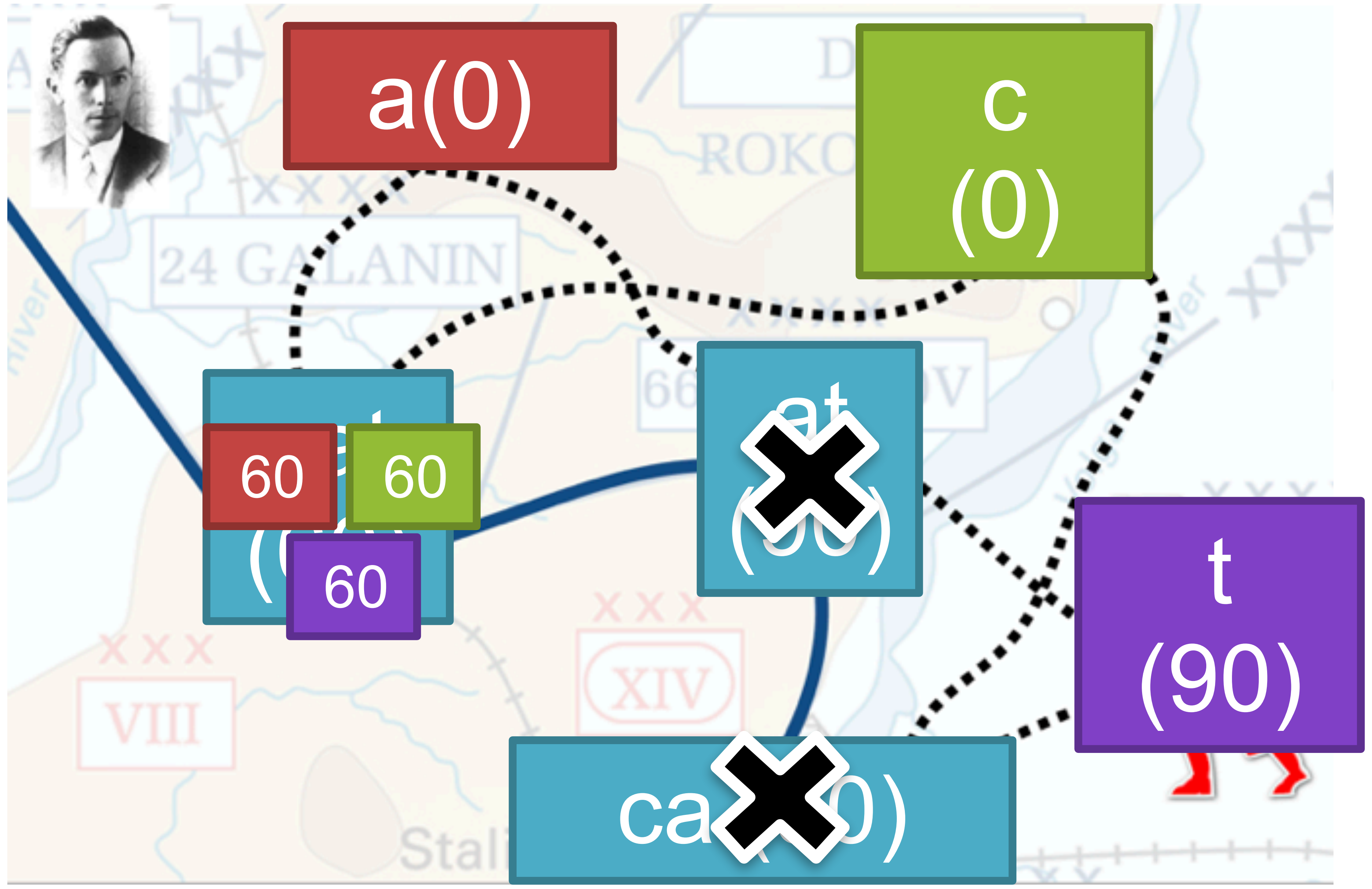
VOLT Formulation

Not all tokens can get chars

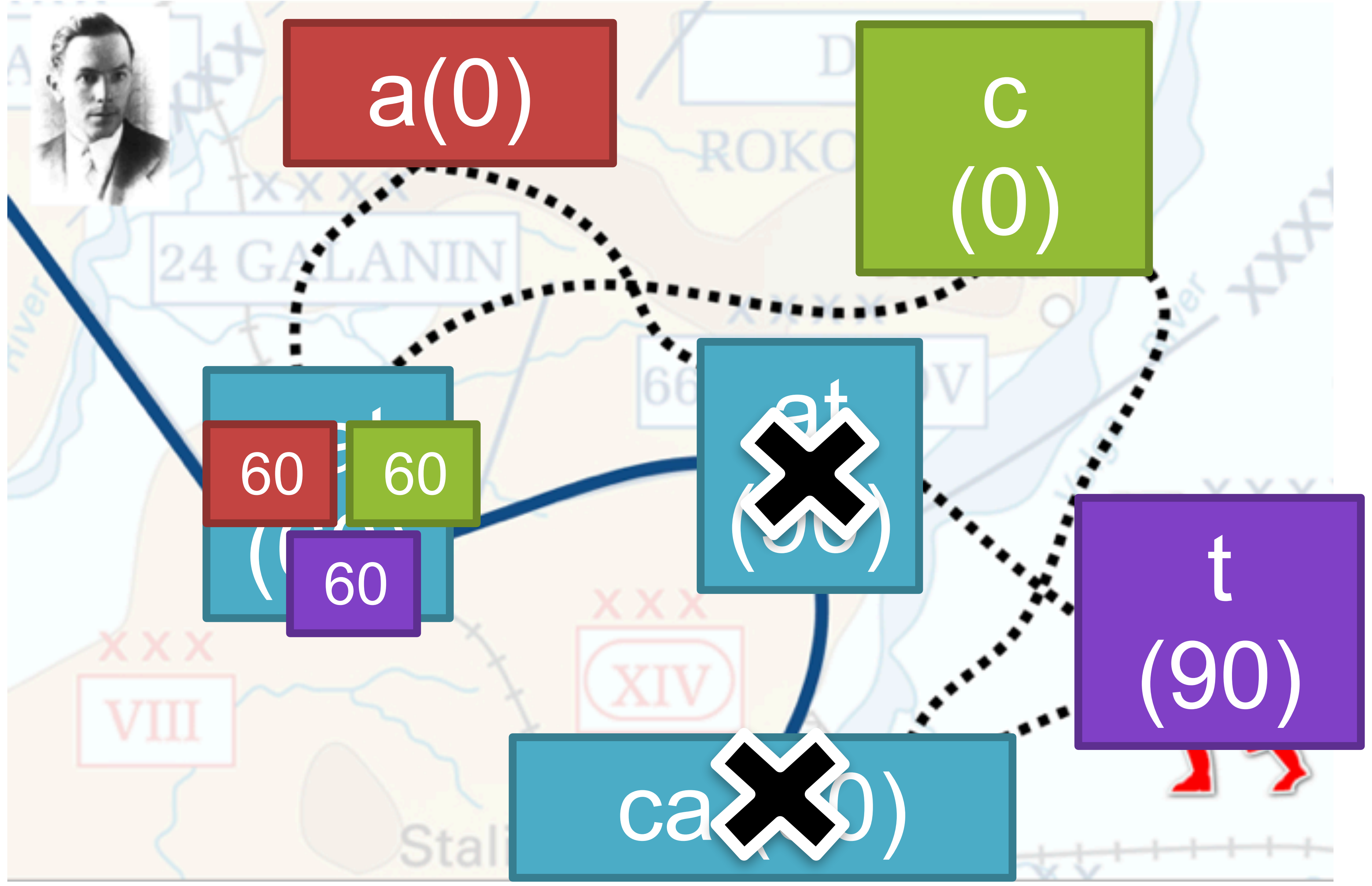


VOLT Formulation

Not all tokens can get chars



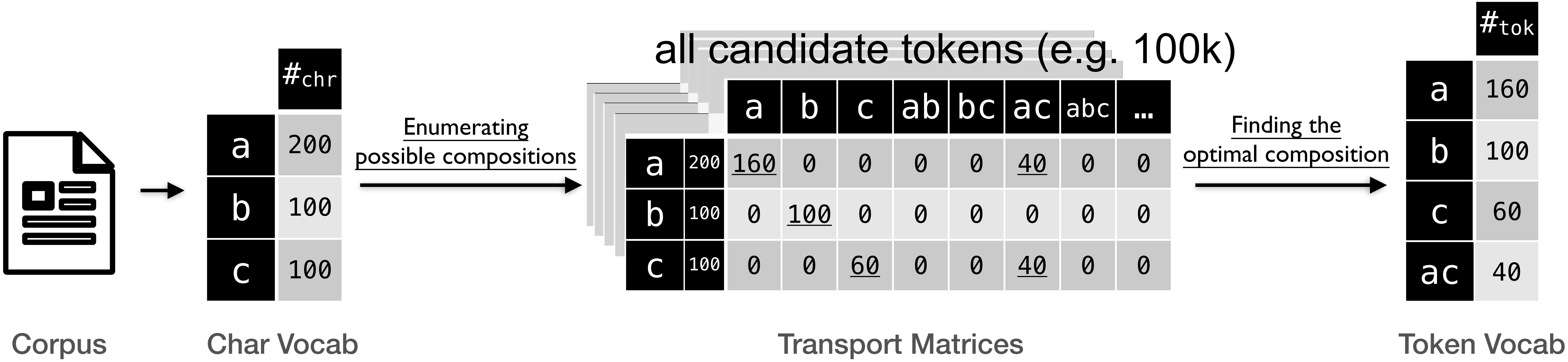
Each Transportation Defines a Vocabulary



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

Proposed VOLT: Vocabulary Building via Transportation

- Transport character occurrences to token occurrences



- Maximizing MUV for vocabulary

$$\max - (H(V_{t+1}) - H(V_t))$$

- Instead, maximizing the lower bound ==> Optimal Transport

$$\max_t(\max H(V_t) - \max H(V_{t+1}))$$

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.

Proposed VOLT: Finding the Optimal Vocabulary

- Entropy-regularized Optimal Transport

$$\min_{P \in \mathbb{R}^{m \times n}} \langle D, P \rangle - H(P)$$

subject to

$$\forall i \in \text{Char}, \sum_{j \in V_n} P_{i,j} = \hat{P}(i)$$

$$\forall j \in V_n, \left| \sum_{i \in \text{Char}} P_{i,j} - \hat{P}(j) \right| = \epsilon$$

Transportation matrix P

Char \ Tok	a	ab	bc
a	$P_{a,a}$	$P_{a,ab}$	$P_{a,bc}$
b	$P_{b,a}$	$P_{b,ab}$	$P_{b,bc}$
c	$P_{c,a}$	$P_{c,ab}$	$P_{c,bc}$

Cost matrix D

Char \ Tok	a	ab	bc
a	0	$\ln 2$	∞
b	∞	$\ln 2$	$\ln 2$
c	∞	∞	$\ln 2$

- Sinkhorn's algorithm (from [Sinkhorn 1967])

Encoding and Decoding with VOLT

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

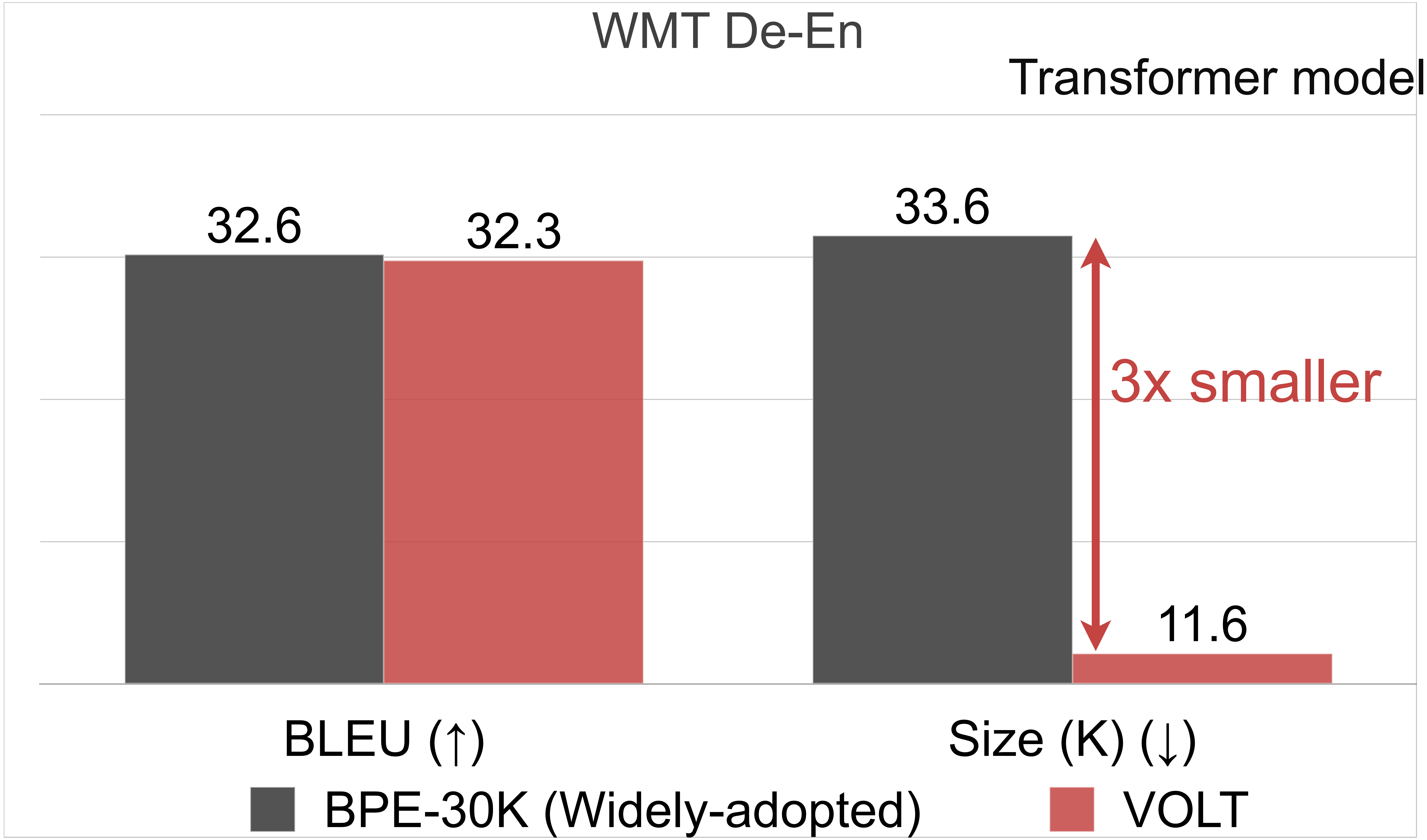
- The vocabulary includes all basic characters.
  - To encode text, it first splits sentences into character-level tokens.
  - Then, we merge two consecutive tokens into one token if the merged one is in the vocabulary solved by OT.
  - This process keeps running until no tokens can be merged.
  - Out-of-vocabulary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

Significance: VOLT is 700x Faster and Greener!

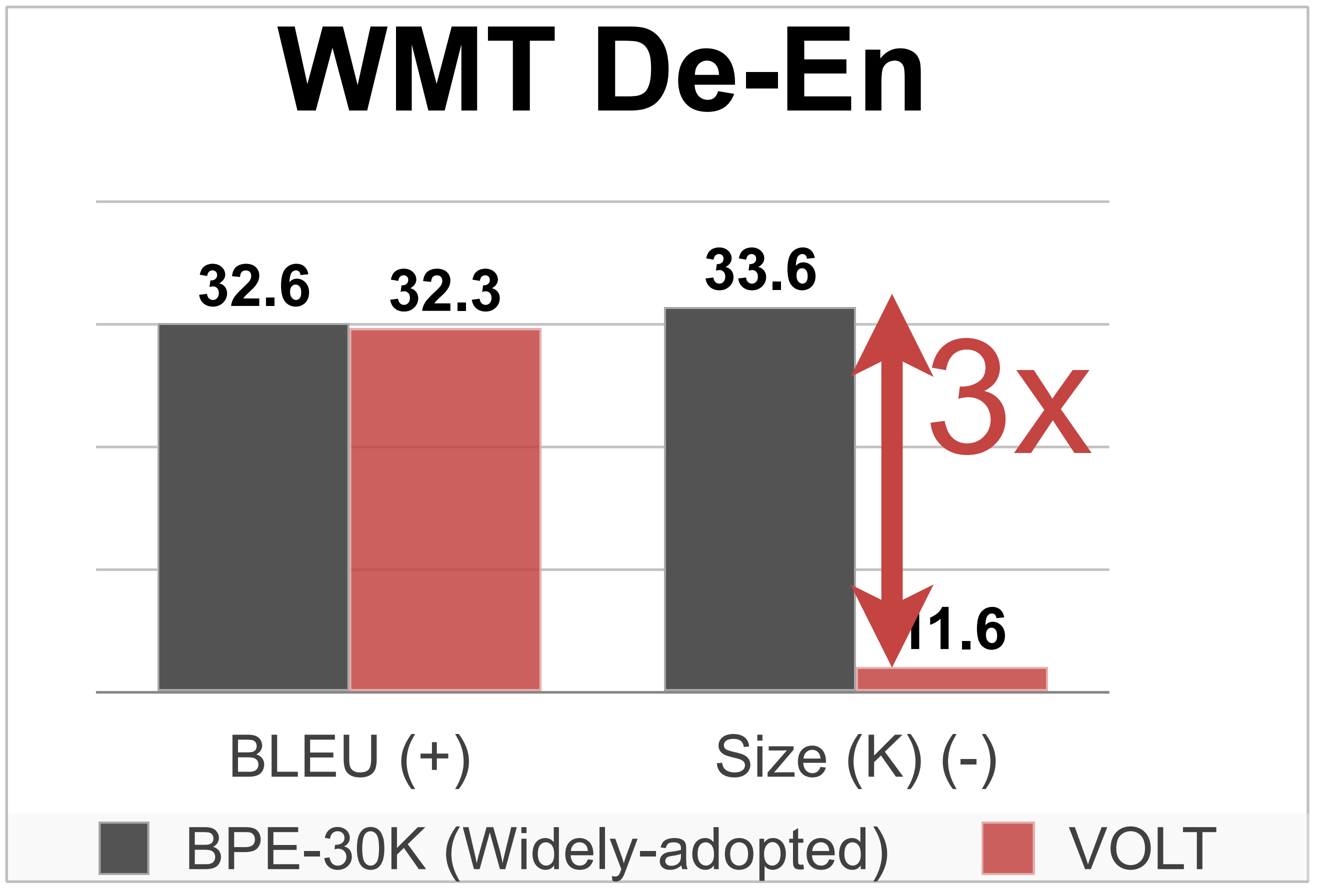
	Computation	Carbon Emission
BPE-Search	384 GPU hours	
VOLT	0.5 CPU hours	

700x faster!

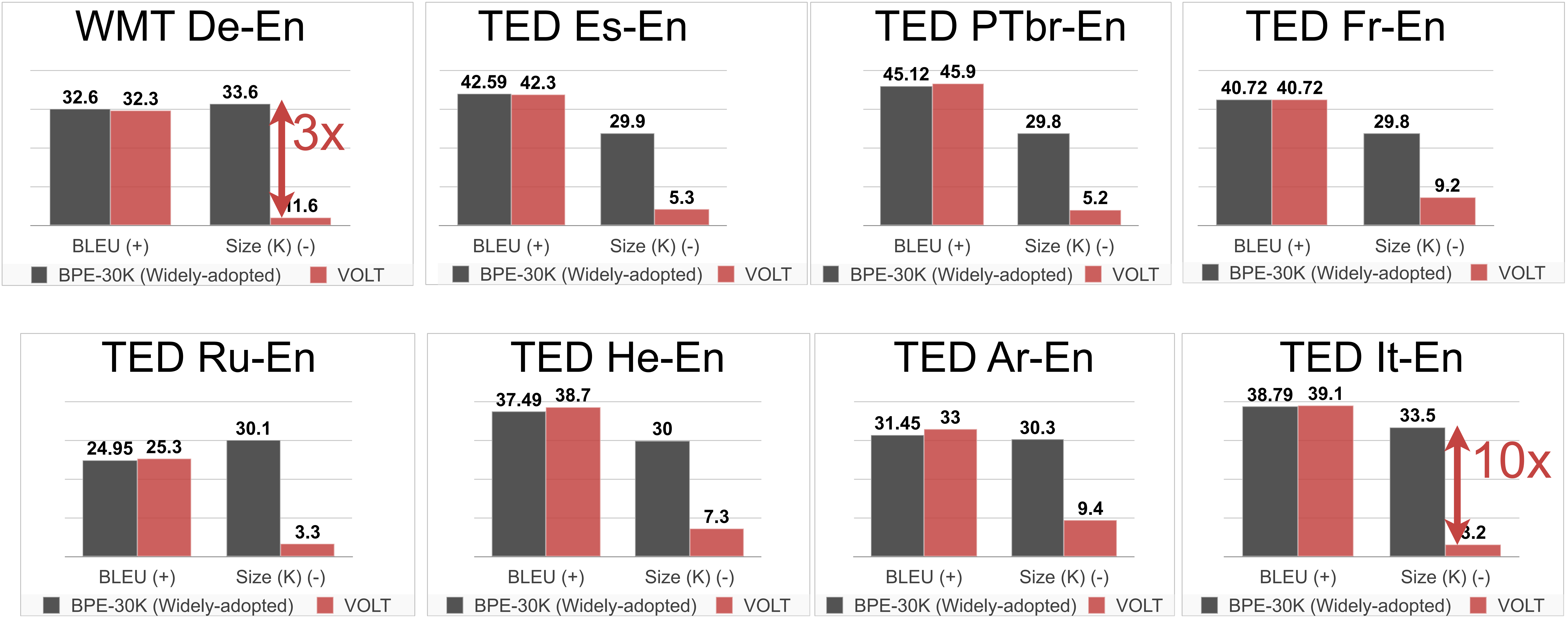
VOLT Finds Smaller Vocabulary on Bilingual MT



VOLT Finds Smaller Vocabulary on Bilingual MT



VOLT Finds Smaller Vocabulary on Bilingual MT



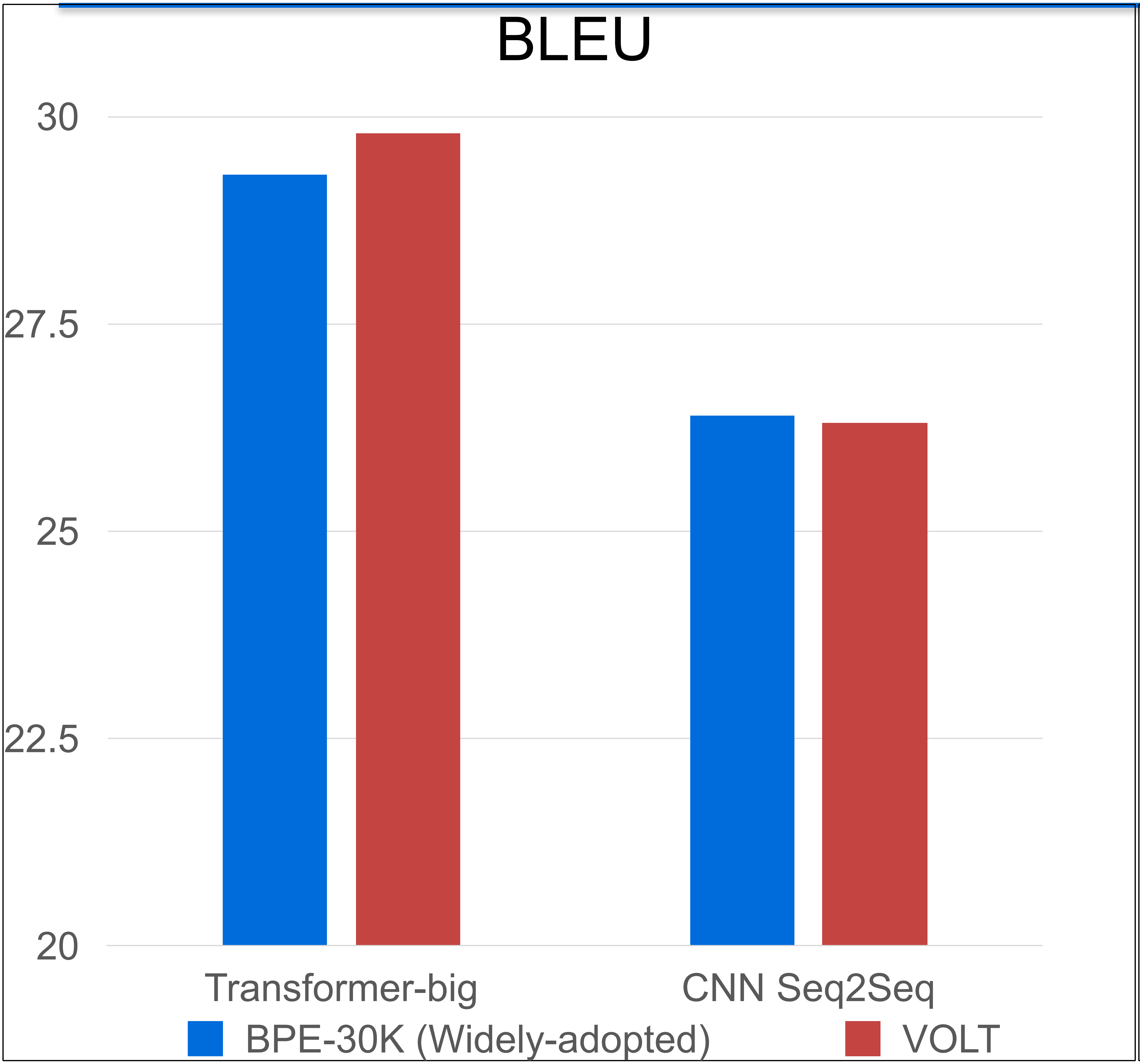
VOLT Finds Better Vocabulary on Multilingual MT

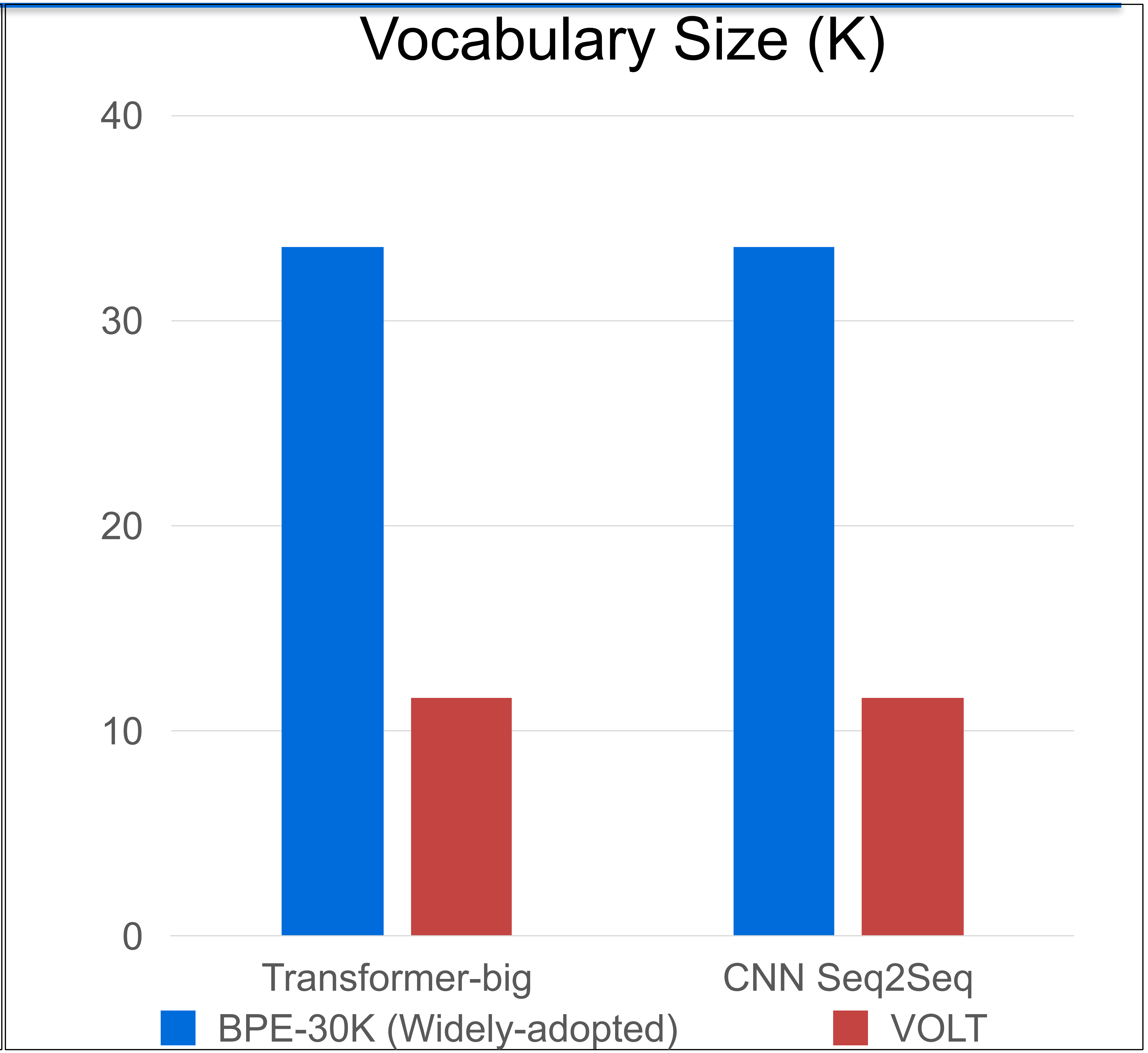
Xu, Zhou, Gan, Zheng, **Lei Li**. Vocabulary Learning via Optimal Transport for Neural Machine Translation. ACL 2021

VOLT Finds Better Vocabulary on Multilingual MT

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

VOLT Generalizes Well to Other Architectures



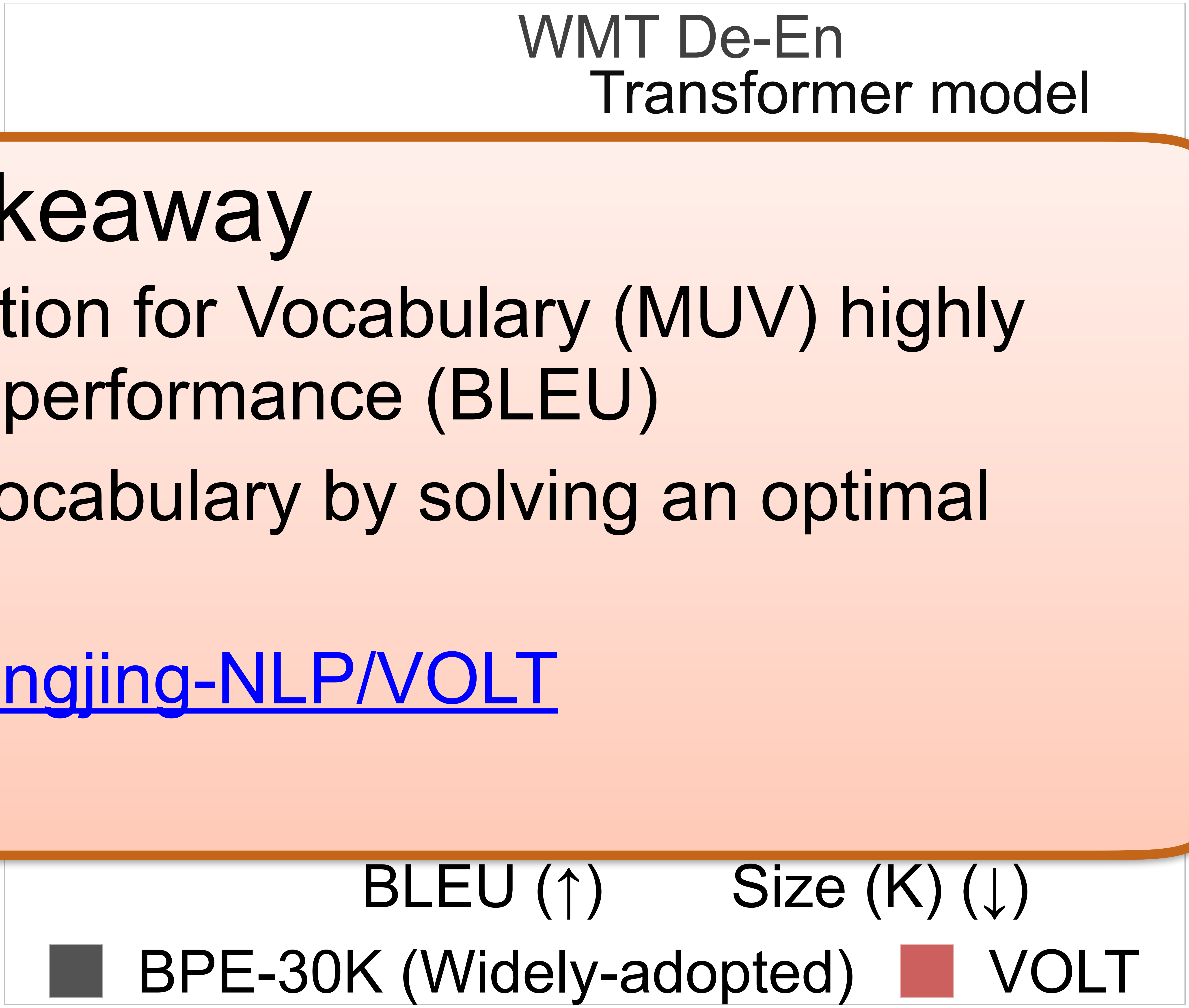


Xu, Zhou, Gan, Zheng, **Li**. Vocabulary Learning via Optimal Transport for Neural Machine Translation. ACL 2021

VOLT is Fast and Finds Smaller Vocabulary

Computation

BP



VOLT Takeaway

- Marginal Utility of information for Vocabulary (MUV) highly correlates with translation performance (BLEU)
- VOLT learns the optimal vocabulary by solving an optimal transport problem.
- code: https://github.com/Jingjing-NLP/VOLT

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

Language In 10