CS11-737 Multilingual NLP

Semi-supervised and Unsupervised Machine Translation

Lei Li

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

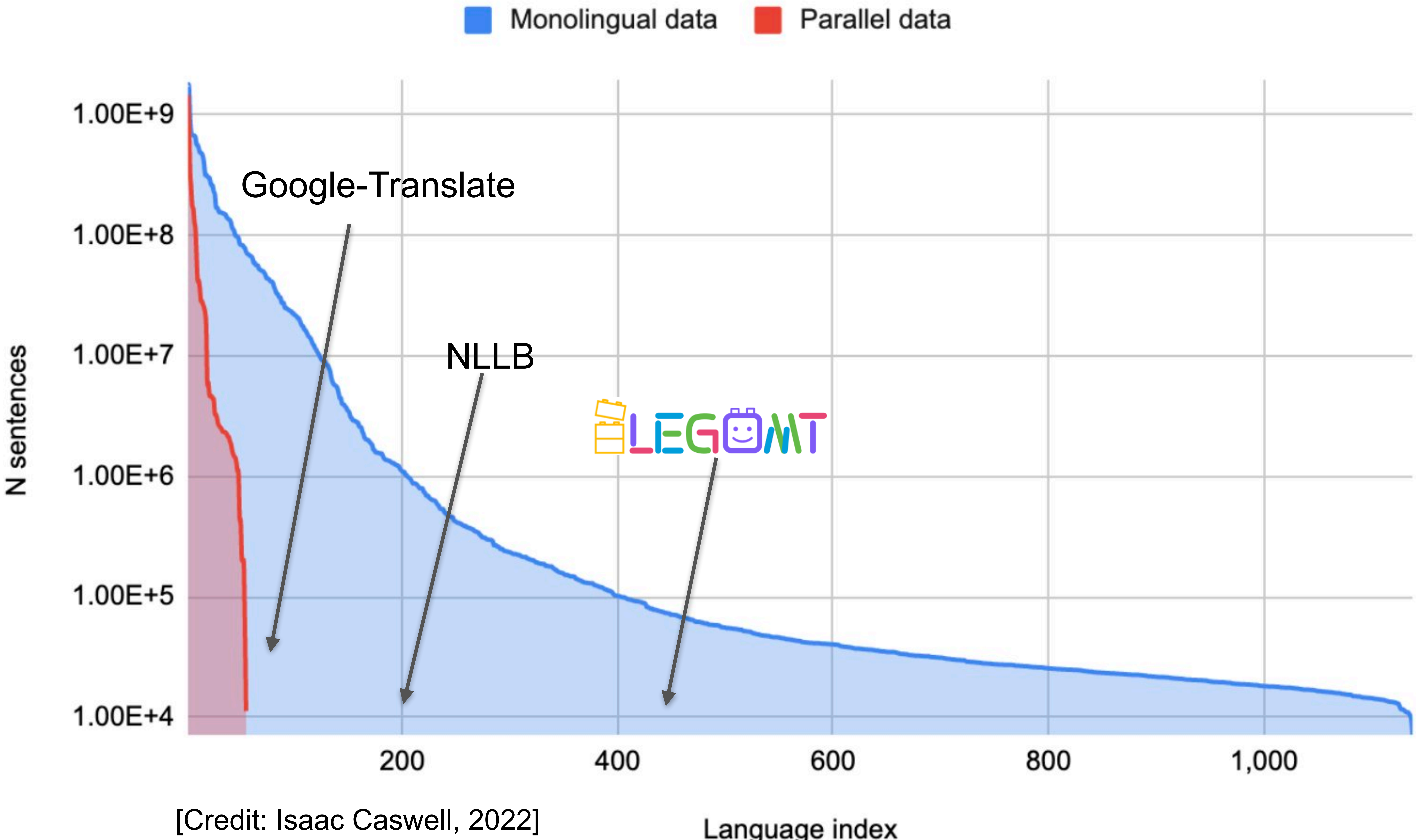
Carnegie Mellon University
Language Technologies Institute

Outline

- Semi-supervised NMT

- Unsupervised MT

Main Bottleneck for NMT: Data Scarcity

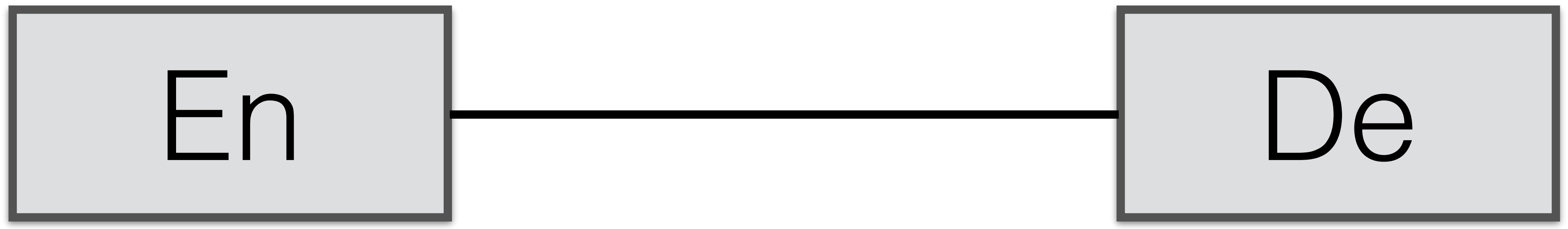


[Credit: Isaac Caswell, 2022]

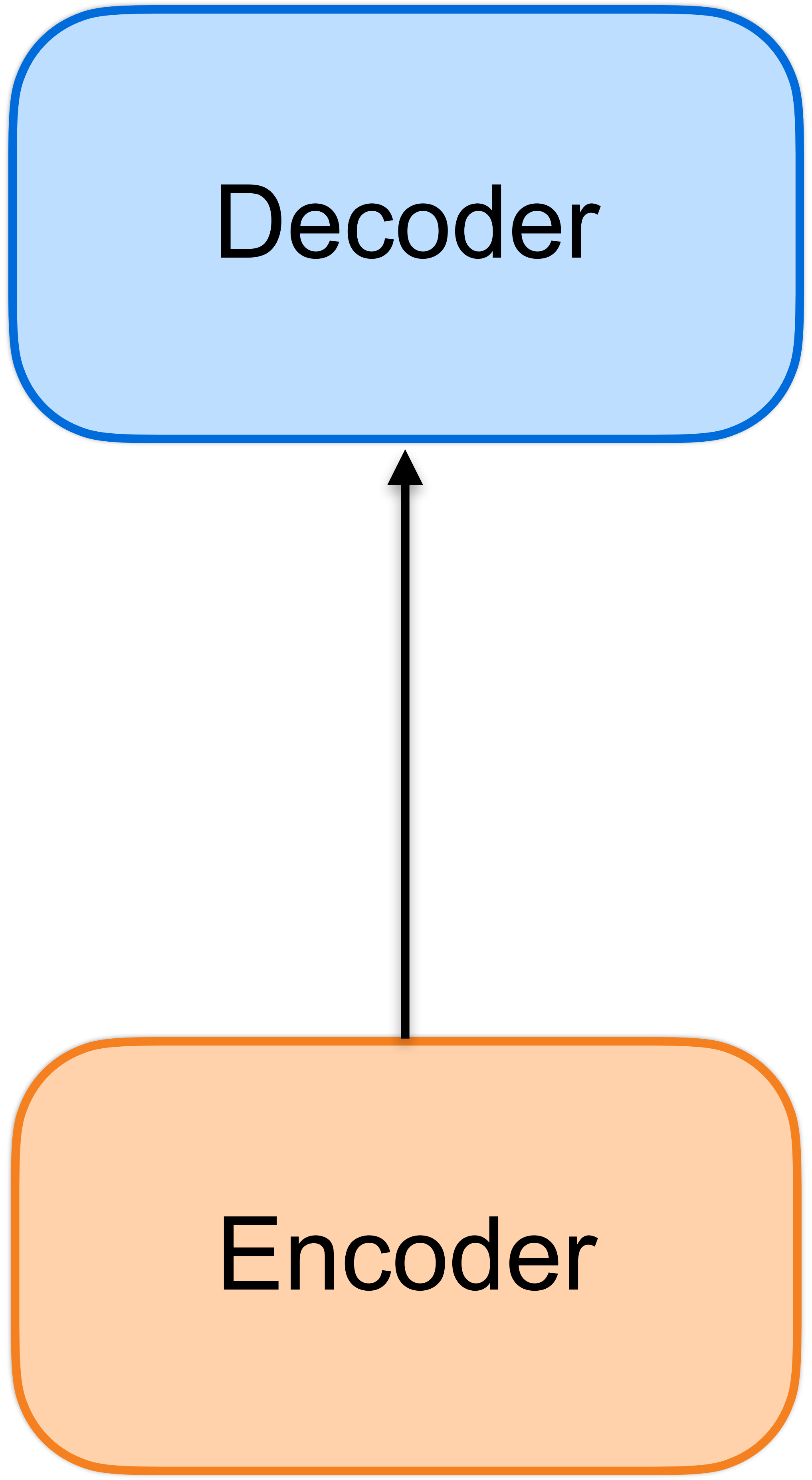
Semi-supervised Learning for MT

- Using both parallel corpus and monolingual data to train an MT system

Parallel corpus



Monolingual



WMT 23 General MT

- Testing MT's capability in general domain: news, conversation, social media
- https://www2.statmt.org/wmt23/translation-task.html
- Chinese to/from English
- German to/from English: document-level (testset won't be sentence breaked)
- Hebrew to/from English: low-resource
- Japanese to/from English
- Russian to/from English
- Ukrainian to/from English
- Czech to Ukrainian: non-English
- English to Czech

WMT 23 Data

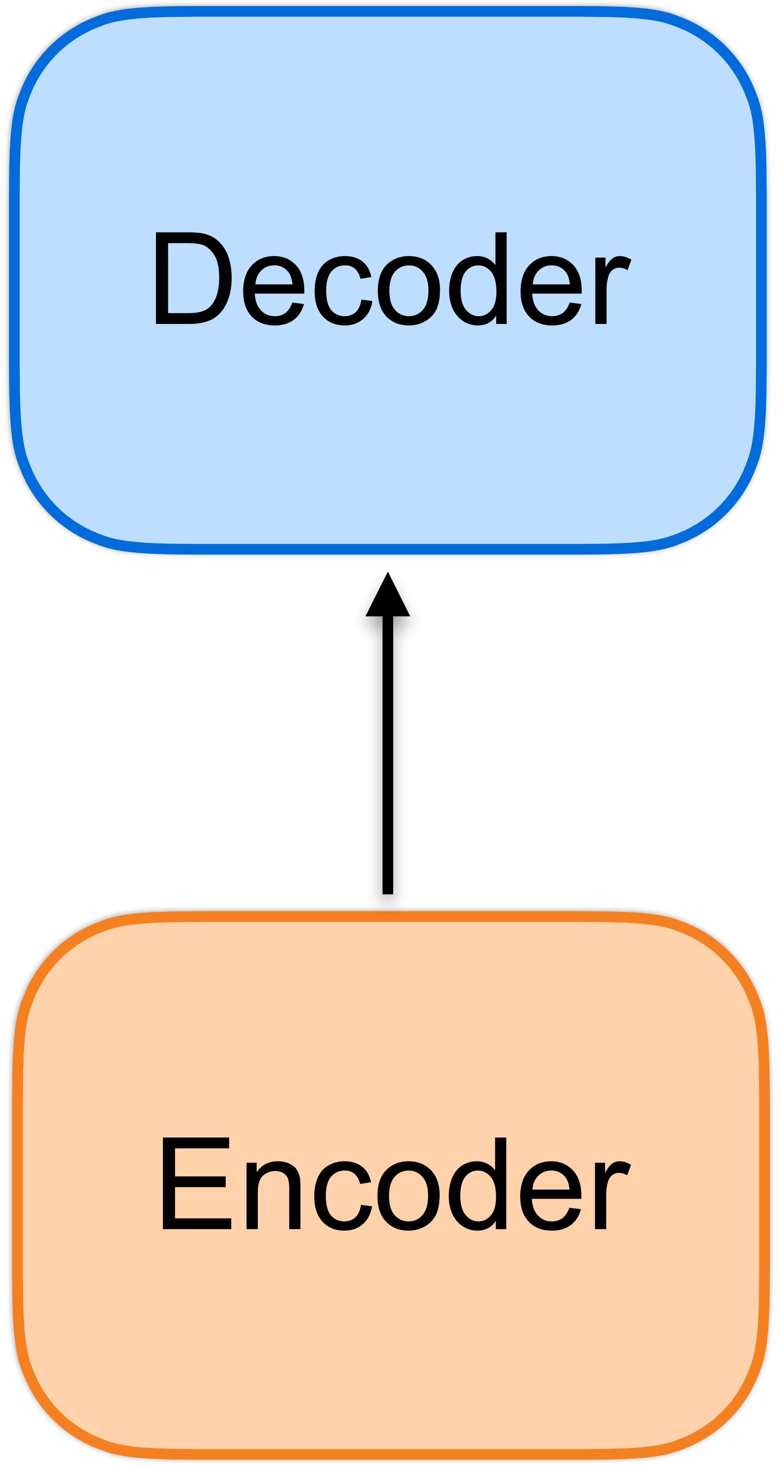
WMT23 Parallel Corpus

File	CS-EN	DE-EN	JA-EN	RU-EN	ZH-EN	HE-EN	UK-EN	UK-CS
Europarl v10	✓	✓						
ParaCrawl v9	✓	✓	✓	✓	✓		✓	
Common Crawl corpus	✓	✓		✓				
News Commentary v18.1	✓	✓	✓	✓	✓			
CzEng 2.0	✓							
Yandex Corpus				✓				
Wiki Titles v3	✓	✓	✓	✓	✓			
UN Parallel Corpus V1.0				✓	✓			
Tilde MODEL corpus	✓	✓		✓			✓	
CCMT Corpus					✓			
WikiMatrix	✓	✓	✓	✓	✓	✓	✓	✓
Back-translated news	✓			✓	✓			
Japanese-English Subtitle Corpus			✓					

WMT23 Monolingual Corpus

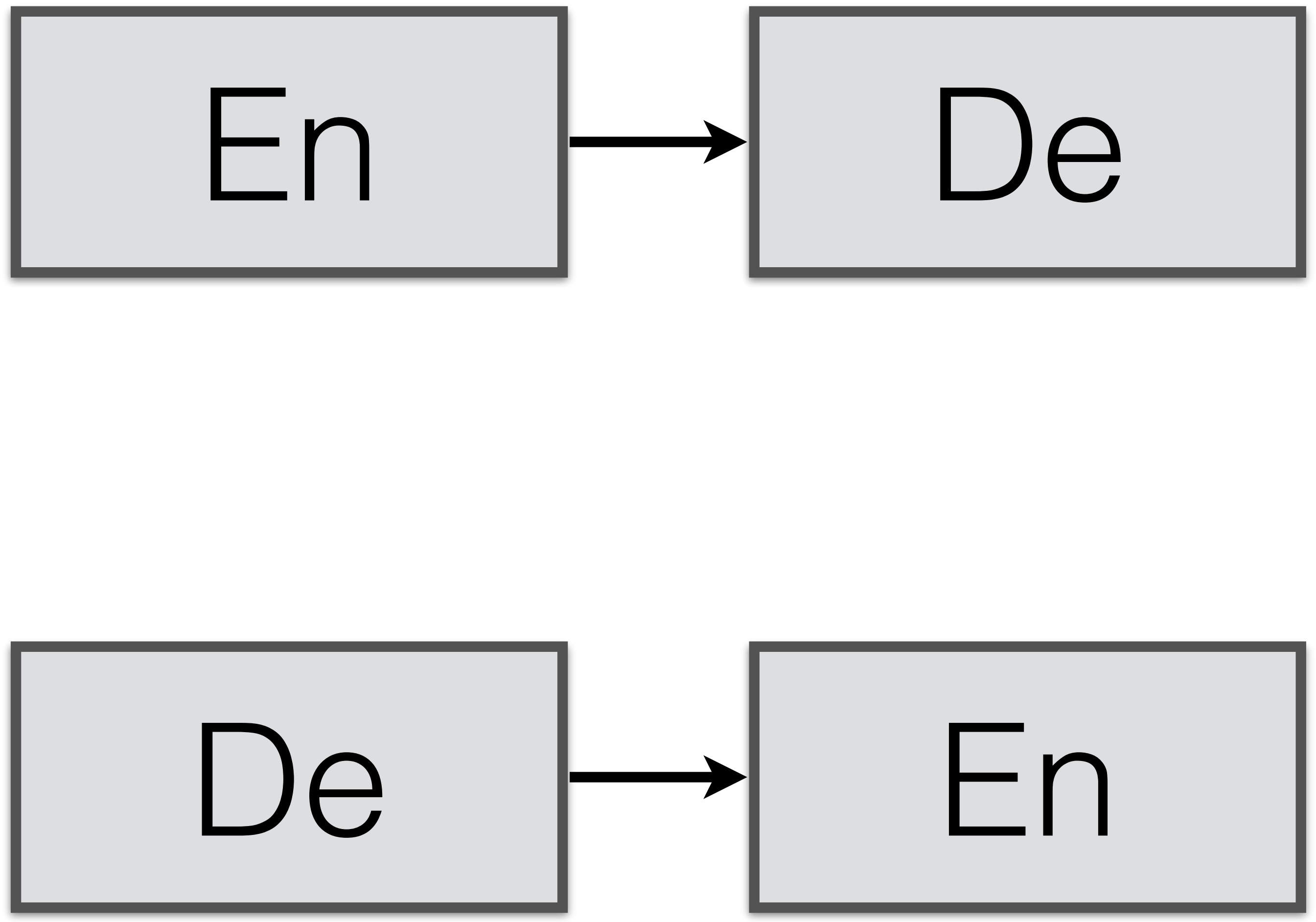
Corpus	CS	DE	EN	JA	RU	ZH	HE	UK
News crawl	✓	✓	✓	✓	✓	✓		✓
News discussions			✓					
Europarl v10	✓	✓	✓					
News Commentary	✓	✓	✓	✓	✓	✓		
Common Crawl	✓	✓	✓	✓	✓	✓		
Extended Common Crawl	✓	✓		✓	✓	✓		
UberText Corpus								✓
Leipzig Corpora	✓	✓	✓	✓	✓	✓	✓	✓
Legal Ukrainian								✓

Back Translation



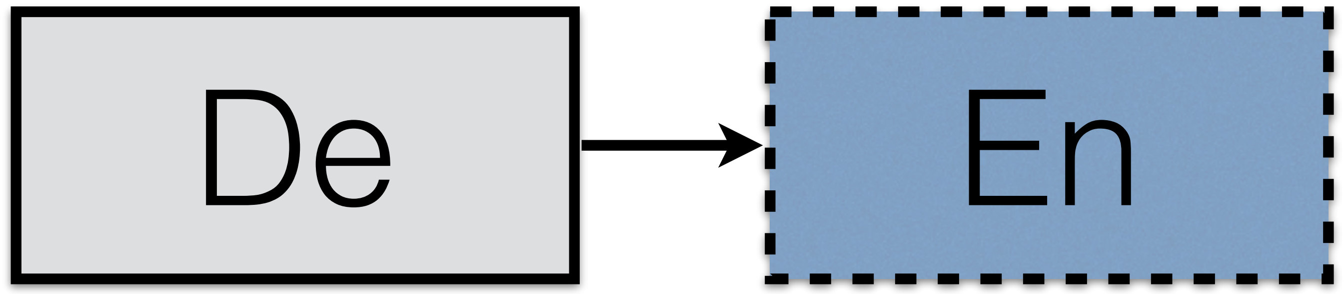
Step 1:

Training two MT models M1, M2 w/ Parallel corpus



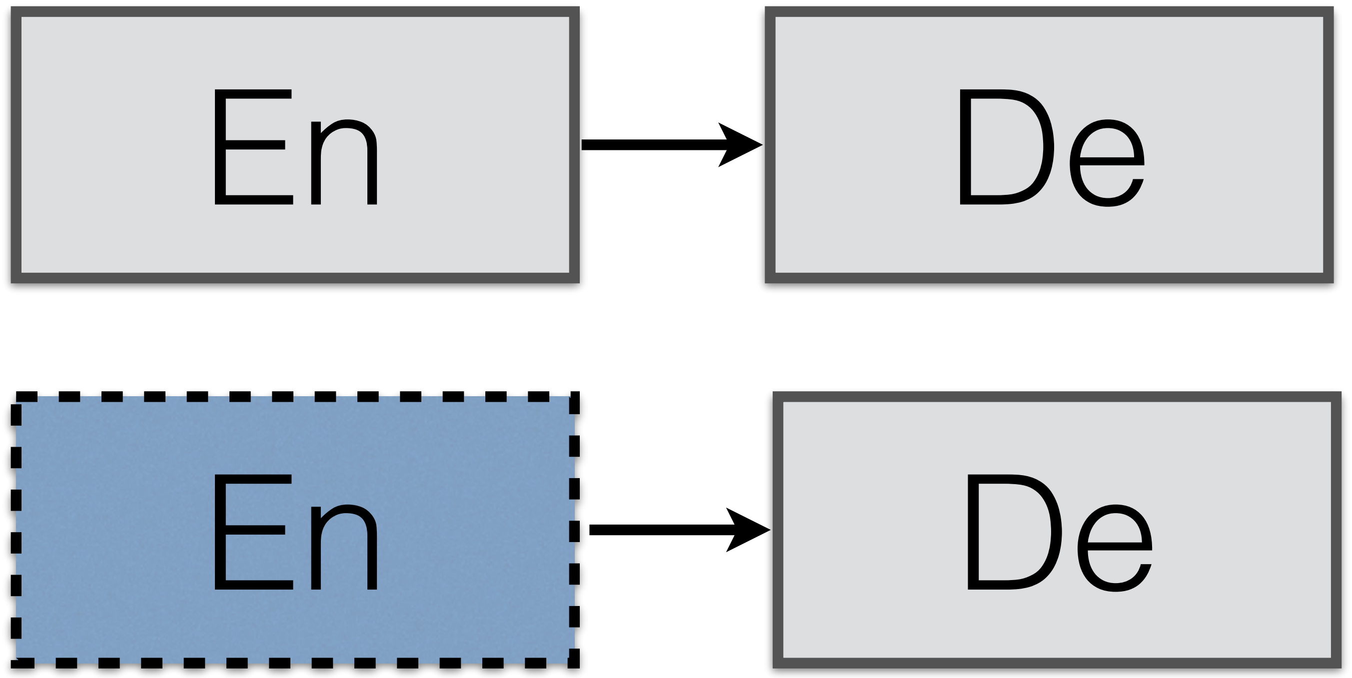
Step 2:

Use M2 to generate pseudo-parallel pairs



Step 3:

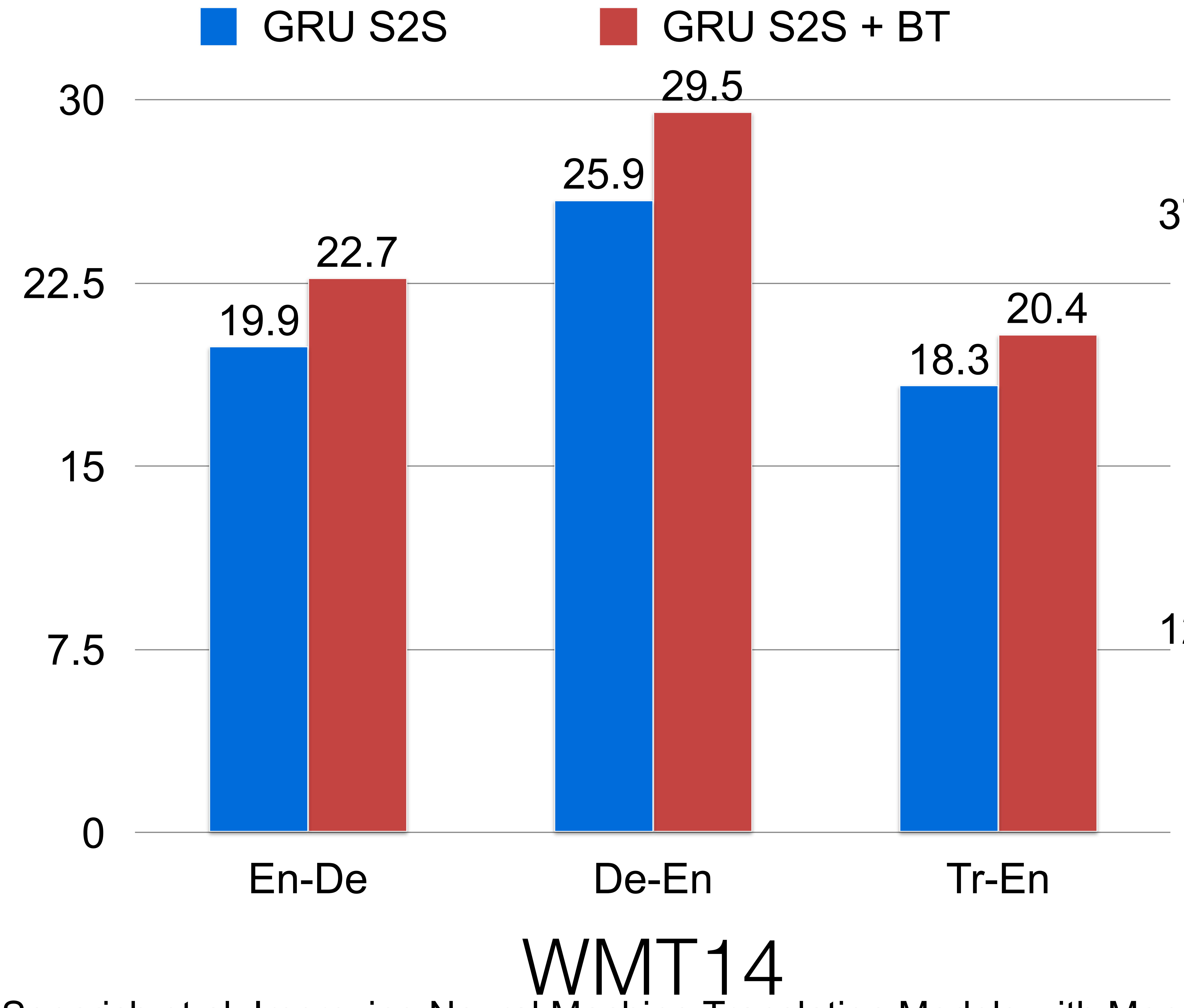
Training w/ both parallel and pseudo-parallel

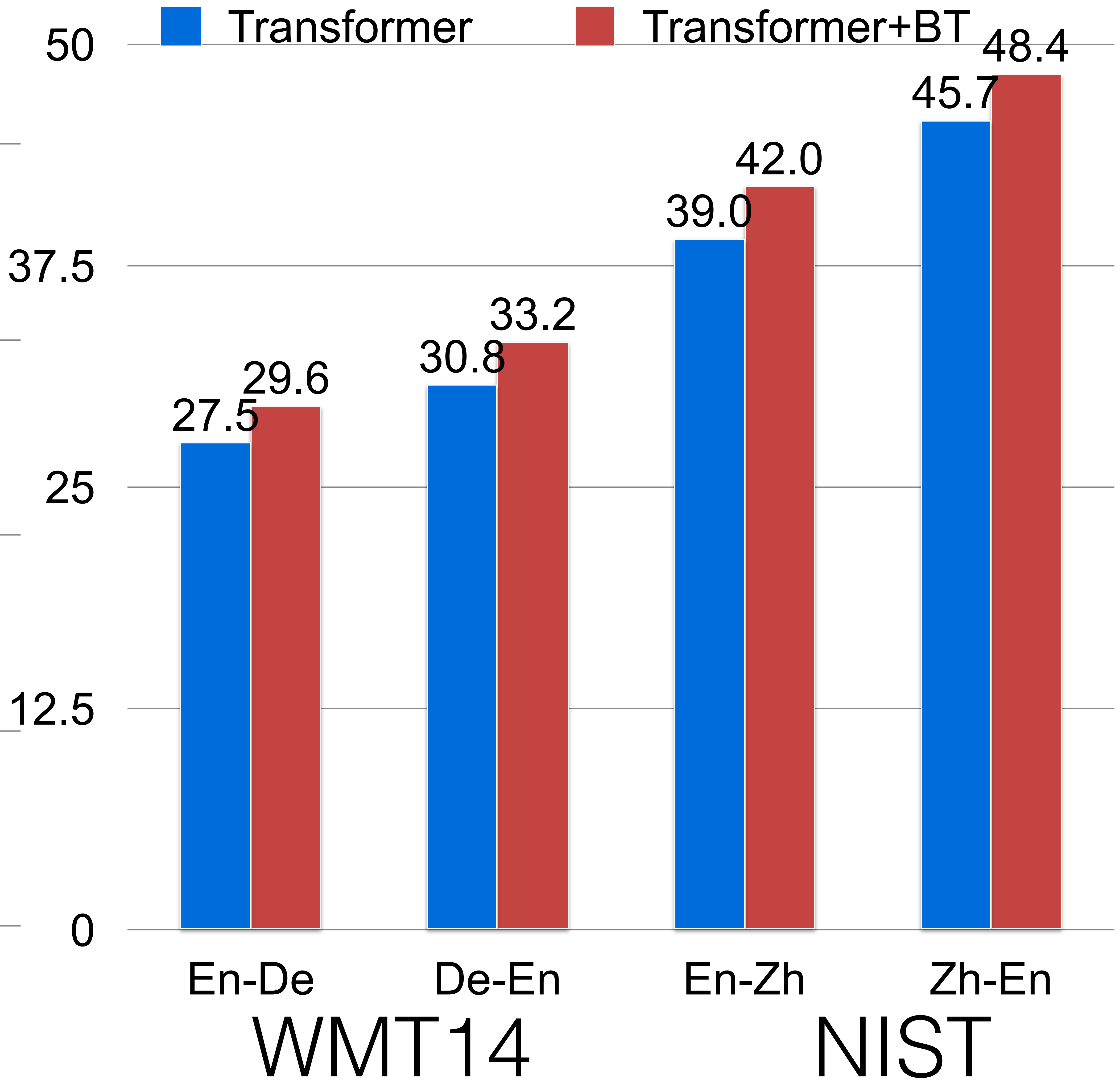


Back Translation Details

1. An initial parallel data D = <x, y> (e.g. De — En)
2. Target side monolingual data (En)
3. Train two separate NMT systems, M1 : x->y, and M2 : y->x
4. Now use M2 to generate translation for y —> x’ = M2(y), denote this synthetic pairs as D’ = {<x’, y>}
5. Combine both D and D’ —> D”=D U D’
6. Train a new model M from x -> y using D”

Does Back Translation work? Yes!



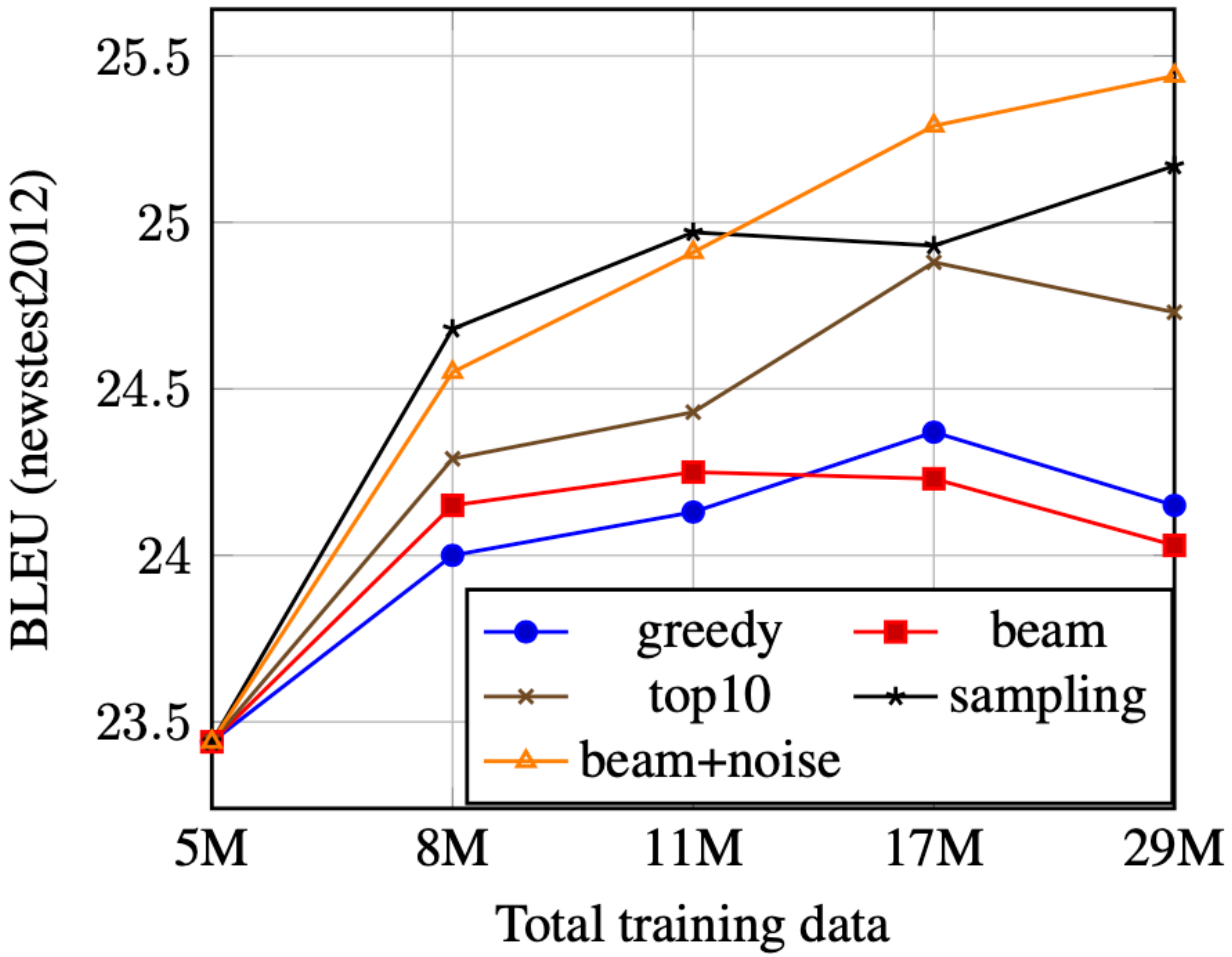


Sennrich et al. Improving Neural Machine Translation Models with Monolingual Data. ACL 2016.

Zheng et al. Mirror-Generative Neural Machine Translation. 2020

Decoding Strategy in Back Translation

- Two best practice (for high-resource):
  - Noisy beam search (adding noise to source side helps!)
    - Select the highest scoring output
    - Higher quality, but lower diversity, potential for data bias
  - Sampling (instead of beam search)
    - Randomly sample from back-translation model
    - Lower overall quality, but higher diversity



Edunov et al. Understanding Back-translation at Scale. 2018.

Some Consideration

- Why back-translation from target side to source?
  - why source is pseudo?
- Can we use source monolingual to generation synthetic pairs?
  - Forward-translation

Using Source Monolingual? Forward Translation

- Like back-translation
- Use the model x->y to create monolingual data
- Train x->y MT model again on

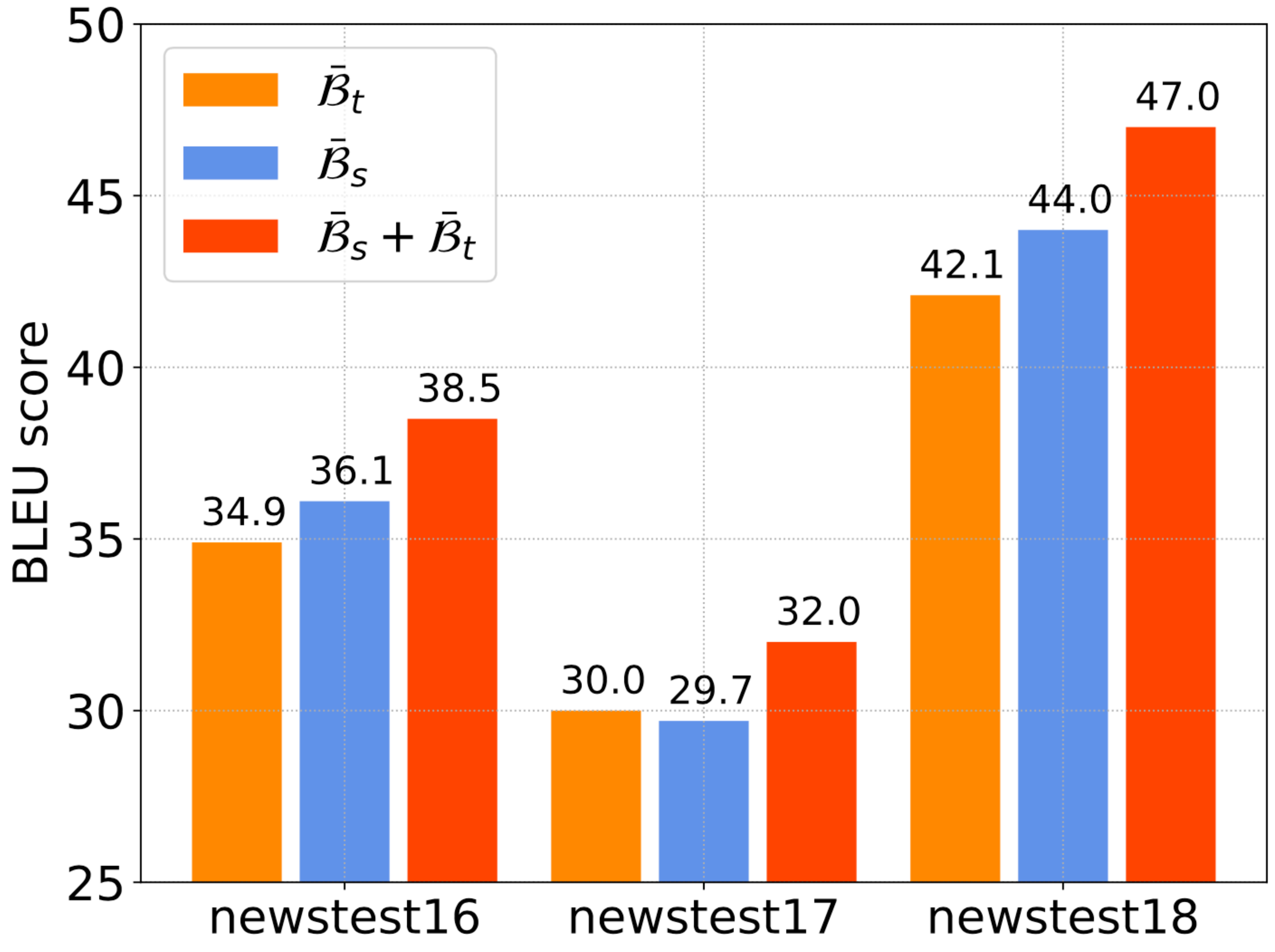


Figure 1: The de-tokenized SacreBLEU scores on En→De newstest2016, newstest2017 and newstest2018 of the models trained by different synthetic data: (1) $\bar{\mathcal{B}}_s$ from source-side monolingual data only, (2) $\bar{\mathcal{B}}_t$ from target-side monolingual data only and (3) the combination of $\bar{\mathcal{B}}_s$ and $\bar{\mathcal{B}}_t$.

Zhang & Zong. Exploiting Source-side Monolingual Data in Neural Machine Translation. 2016

Forward Translation + Back Translation + Noise

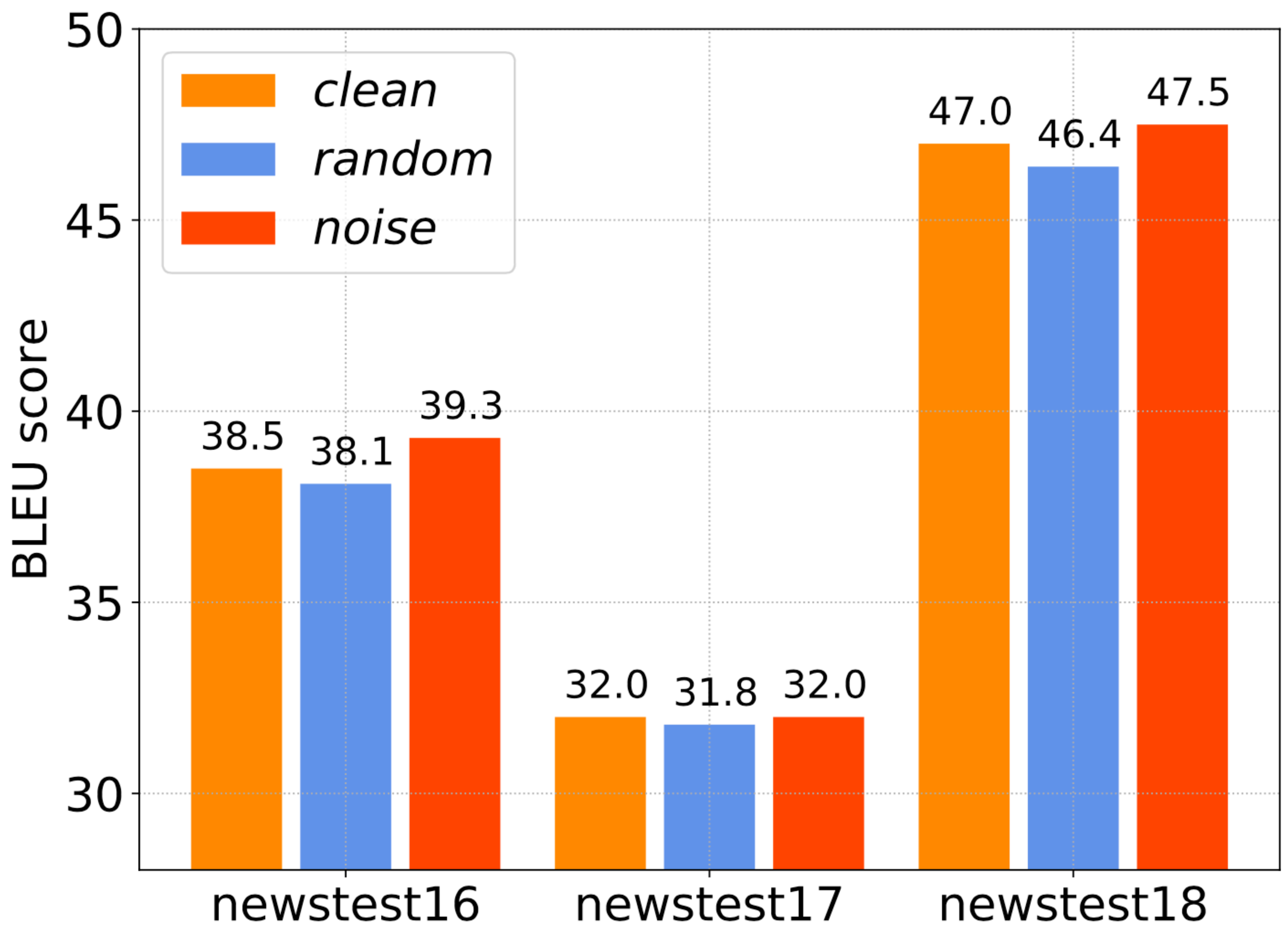


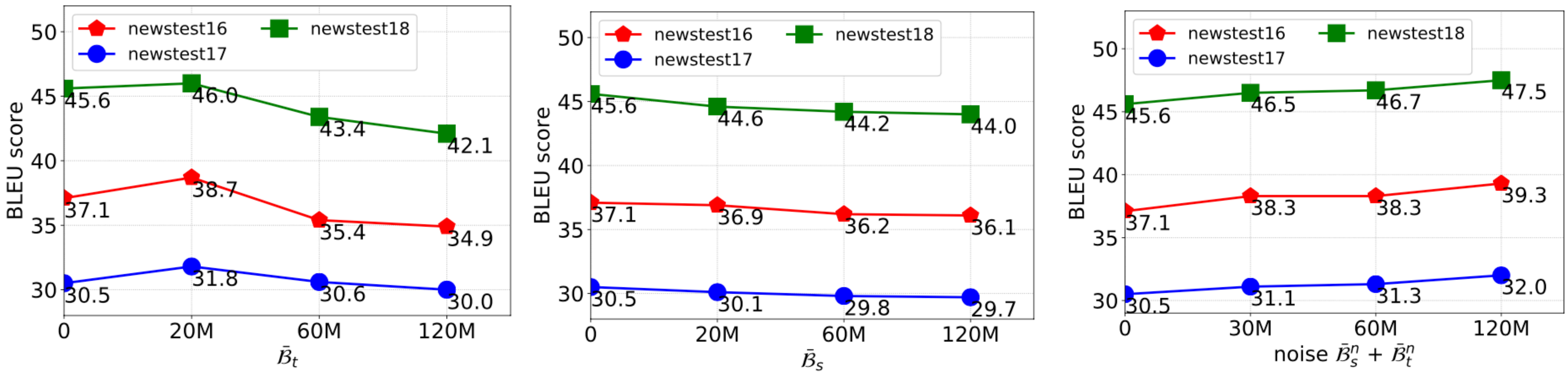
Figure 2: The de-tokenized SacreBLEU scores on En→De newstest2016, newstest2017 and newstest2018 of the models trained by synthetic data generated in different ways: (1) clean $\bar{\mathcal{B}}_s$ and $\bar{\mathcal{B}}_t$ data, (2) $\bar{\mathcal{B}}_s^r$ and randomly sampled $\bar{\mathcal{B}}_t^r$ data, and (3) noised $\bar{\mathcal{B}}_s^n$ and $\bar{\mathcal{B}}_t^n$ data.

Some Consideration

- What kind of monolingual data?
- How much monolingual data?
  - Ratio parallel vs. synthetic?
  - Usually 1:1

How much monolingual for BT?

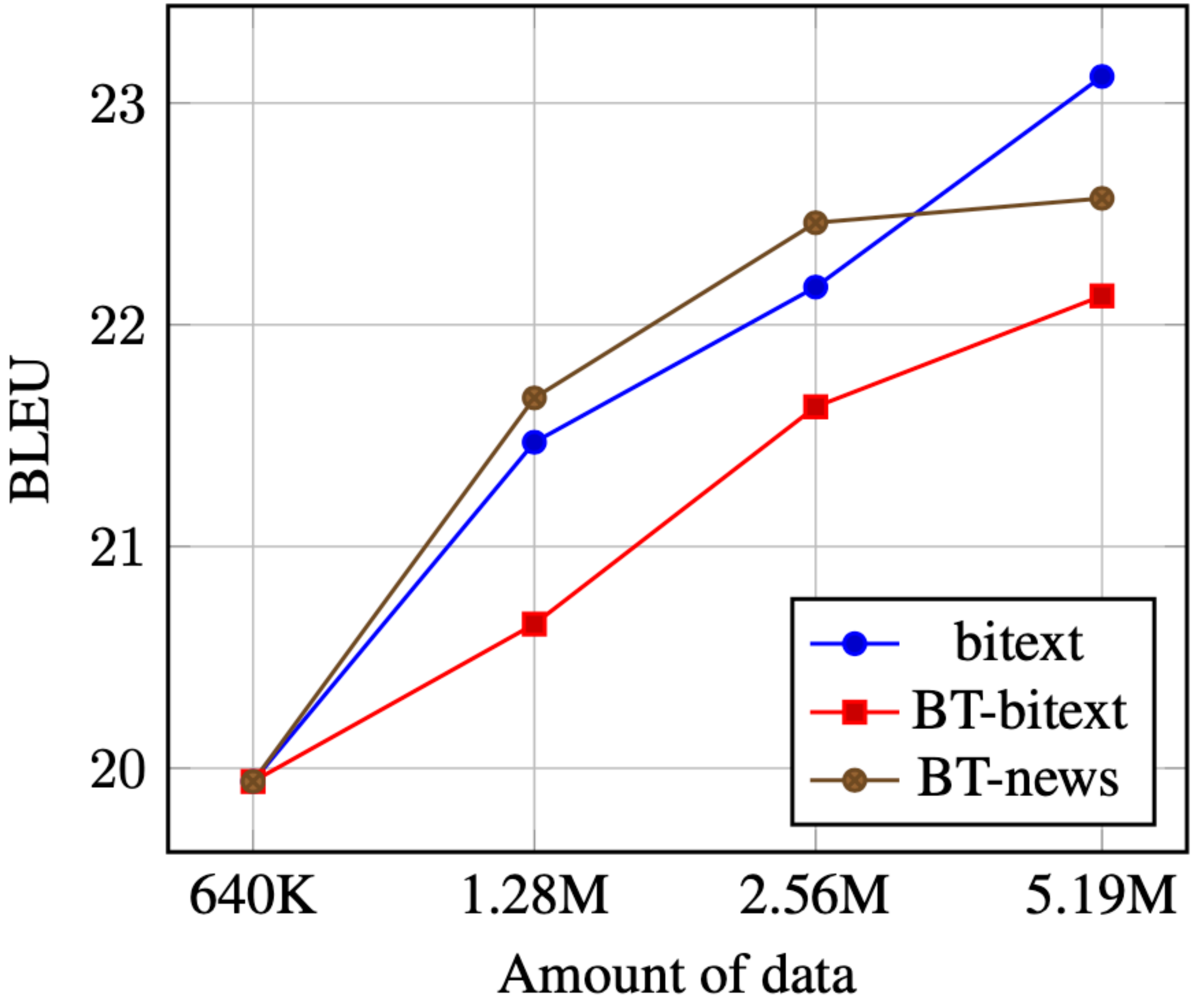
- More is better?
- Over BT hurts
- But noised-BT can sustain improvement!



(a) Different scales of $\bar{\mathcal{B}}_t$ data. (b) Different scales of $\bar{\mathcal{B}}_s$ data. c) Different scales of noised $\bar{\mathcal{B}}_s + \bar{\mathcal{B}}_t$ data.

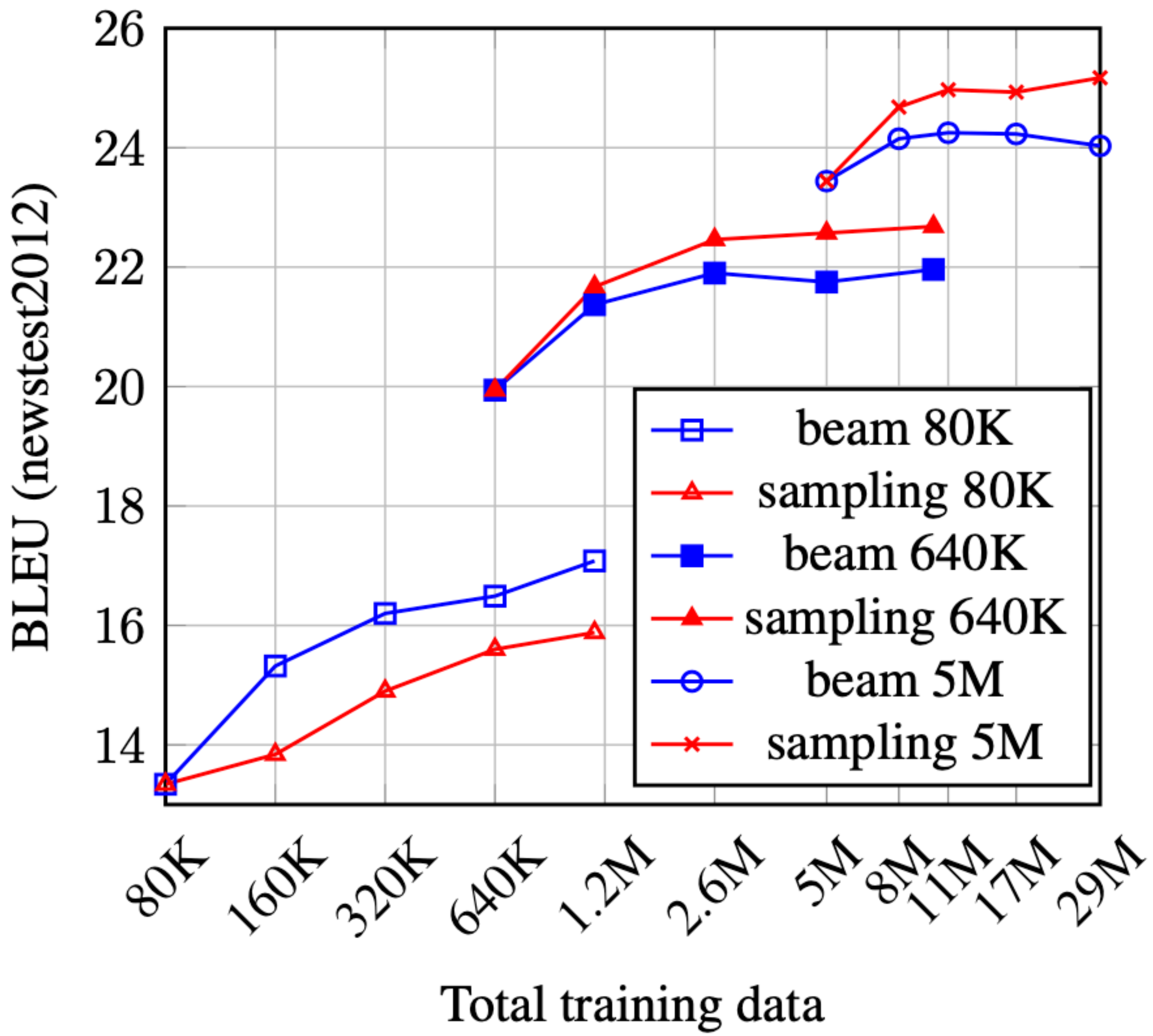
Target Domain for Back Translation

- Better to pick monolingual data the same as target domain



(a) newstest2012

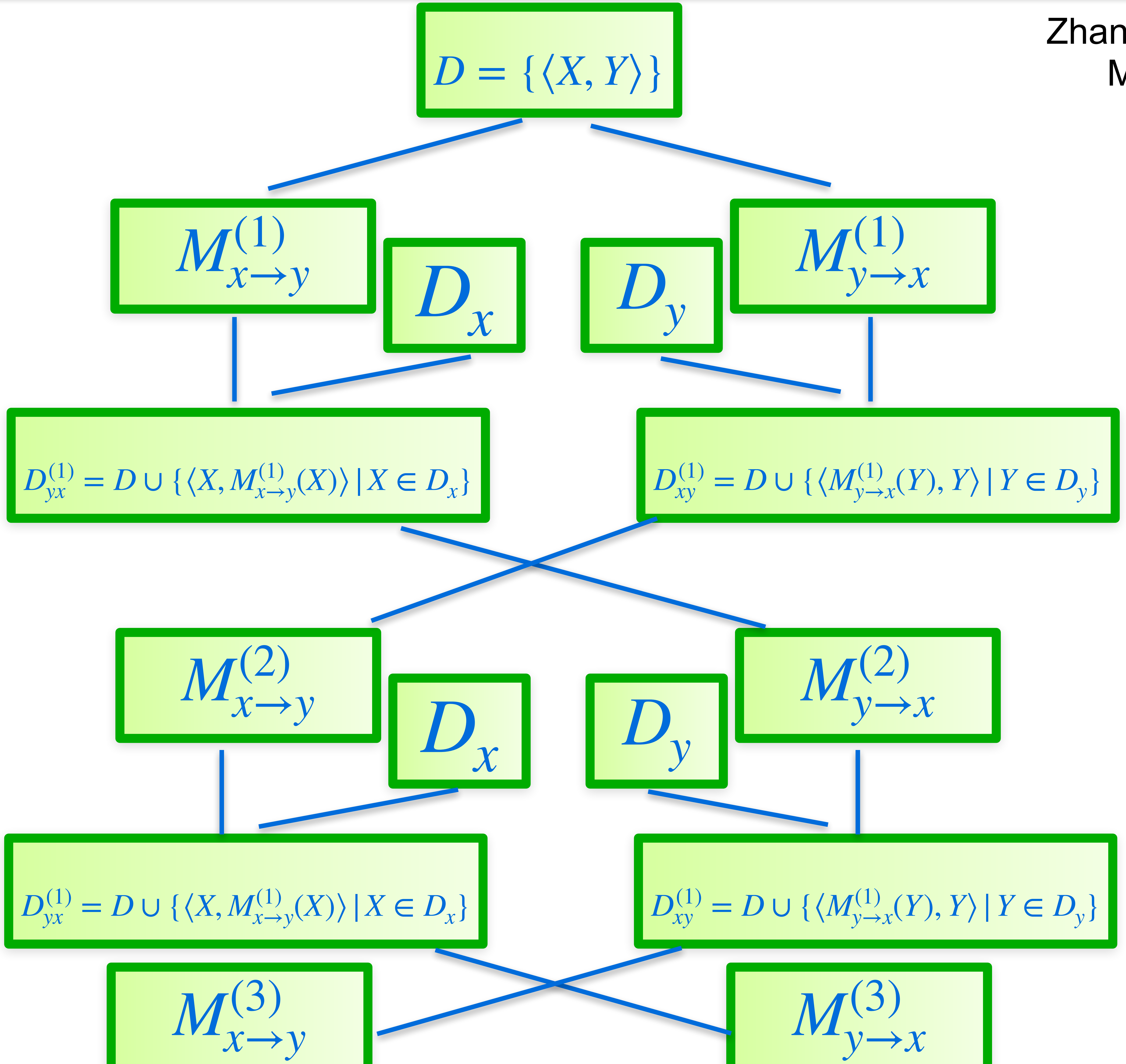
BT in Low-resource Setting



Edunov et al. Understanding Back-translation at Scale. 2018.

Iterative Joint Back Translation

Zhang et al. Joint Training for Neural Machine Translation Models with Monolingual Data. 2018



Probabilistic Model for Semi-Supervised MT

- For monolingual $Y_m \in D_y$, treat X as a random variable, $X \sim P(X \mid Y_m; \theta^{\leftarrow})$
- Training with parallel and monolingual corpus
  ℓ = CE + Expected reconstruction

$$= \sum_{\langle X_n, Y_n \rangle \in D} \log P(Y_n \mid X_n; \theta^{\rightarrow}) + \sum_{Y_m \in D_Y} \log \sum_{X \in V^*} P(Y_m \mid X; \theta^{\rightarrow}) P(X \mid Y_m; \theta^{\leftarrow})$$

$$\sum_{\langle X_n, Y_n \rangle \in D} \log P(X_n \mid Y_n; \theta^{\leftarrow}) + \sum_{X_m \in D_x} \log \sum_{Y \in V^*} P(Y \mid X_m; \theta^{\rightarrow}) P(X_m \mid Y; \theta^{\leftarrow})$$

Cheng et al. Semi-Supervised Learning for Neural Machine Translation. ACL 2016.

Training

- SGD
- An instance Monte-Carlo EM

$$\ell = \sum_{\langle X_n, Y_n \rangle \in D} \log P(Y_n | X_n; \theta^{\rightarrow}) + \sum_{Y_m \in D_Y} \log \sum_{X \in V^*} P(Y_m | X; \theta^{\rightarrow}) P(X | Y_m; \theta^{\leftarrow})$$

- $$\sum_{\langle X_n, Y_n \rangle \in D} \log P(X_n | Y_n; \theta^{\leftarrow}) + \sum_{X_m \in D_x} \log \sum_{Y \in V^*} P(Y | X_m; \theta^{\rightarrow}) P(X_m | Y; \theta^{\leftarrow})$$

- $$\frac{\partial \ell}{\partial \theta^{\rightarrow}} = \cdots + \sum_{Y_m \in D_Y} \sum_{X \in V^*} \frac{P(Y_m | X; \theta^{\rightarrow}) P(X | Y_m; \theta^{\leftarrow})}{\sum_{X' \in V^*} P(Y_m | X'; \theta^{\rightarrow}) P(X' | Y_m; \theta^{\leftarrow})} \frac{\partial \log P(Y_m | X; \theta^{\rightarrow})}{\partial \theta^{\rightarrow}} + \cdots$$

- Alg 1: generate top-k candidates, then compute the gradient.

Cheng et al. Semi-Supervised Learning for Neural Machine Translation. ACL 2016.

Back-translation as a Special Case

- $$\frac{\partial \ell}{\partial \theta^{\rightarrow}} = \cdots + \sum_{Y_m \in D_Y} \sum_{X \in V*} \frac{P(Y_m | X; \theta^{\rightarrow}) P(X | Y_m; \theta^{\leftarrow})}{\sum_{X' \in V*} P(Y_m | X'; \theta^{\rightarrow}) P(X' | Y_m; \theta^{\leftarrow})} \frac{\partial \log P(Y_m | X; \theta^{\rightarrow})}{\partial \theta^{\rightarrow}} + \cdots$$
- If instead of top-k, just pick the top-1 beam search result, ==> back-translation
- Back-translation is an instance of Semi-supervised MT
- Other ways to implement?

Also known as Dual Learning

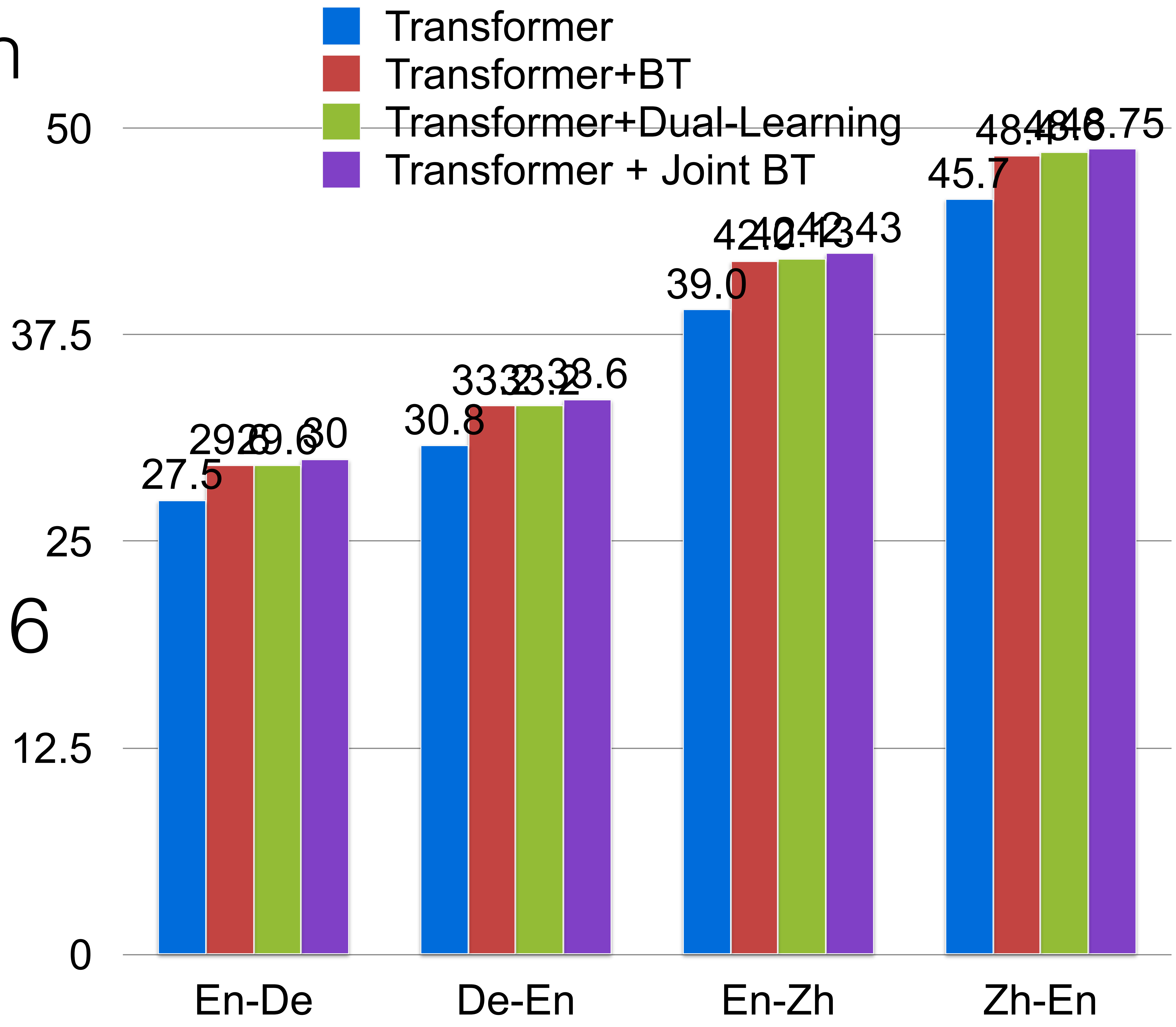
$$\ell = \sum_{Y_m \in D_Y} \sum_{X \in V^*} P(X \mid Y_m; \theta^{\leftarrow}) \left(\log P(Y_m \mid X; \theta^{\rightarrow}) + \log P(X; \theta_X)\right)$$

- essentially the lower bound of the complete log-likelihood (multiplies with language model probability)

He et al. Dual Learning for Machine Translation. 2016

Comparing Backtranslation and Dual Learning

- Back-translation [Sennrich 2016], Cheng 2016, Dual Learning [He 2016], joint back-translation [Zhang 2018], all have same performance.
- Formulation of Cheng 2016 and Zhang 2018 are the same.



Zheng et al. Mirror-Generative Neural Machine Translation. 2020.

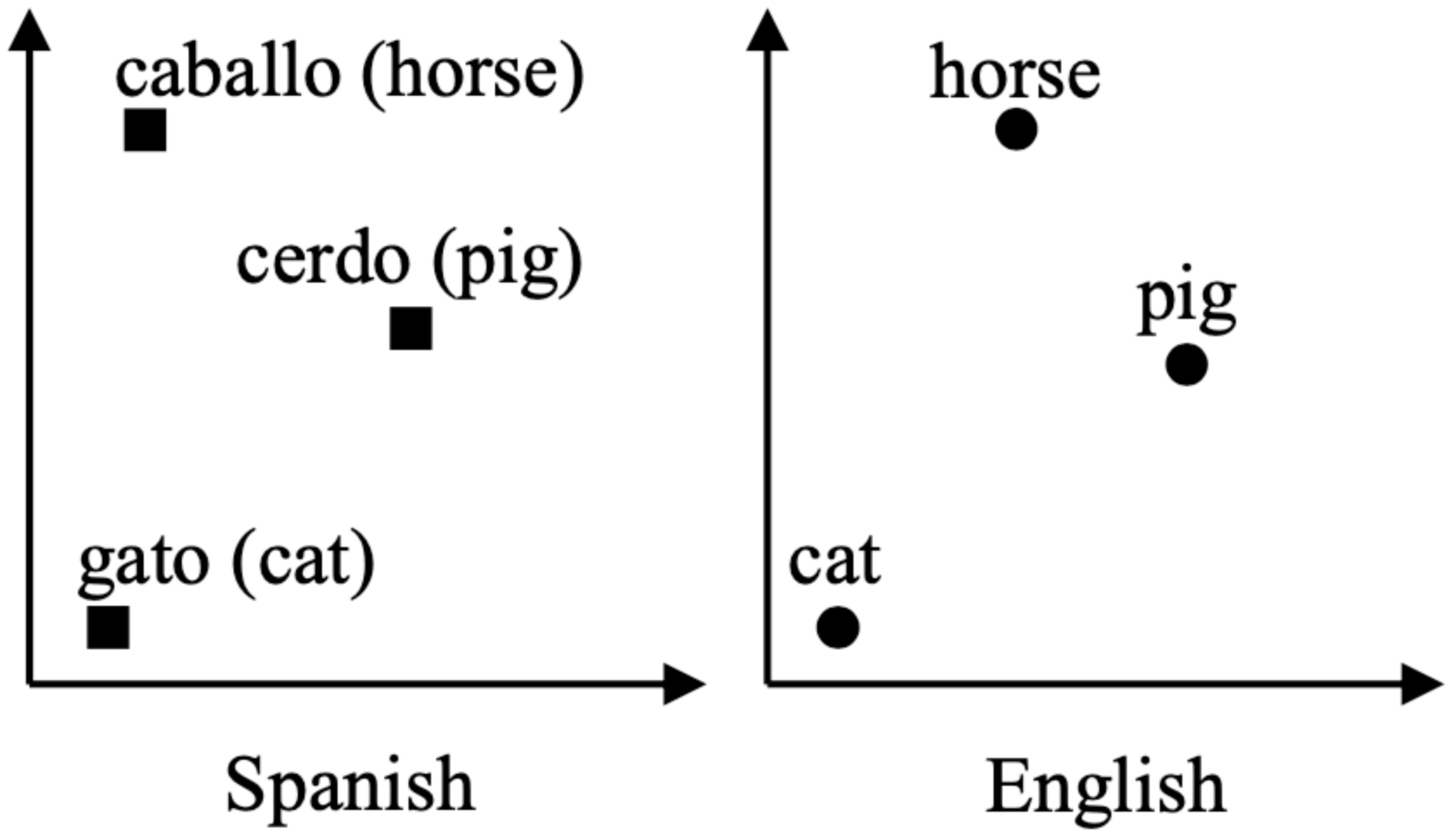
Unsupervised Neural Machine Translation

Unsupervised Machine Translation

- Learning without supervision
  - No parallel corpus, only monolingual data
- Why?
  - many language pairs do not have parallel sentences, or very expensive to create parallel sentences by human
  - but monolingual data are abundant
- How? Basic idea:
  - Cross-lingual pre-training
  - Weight sharing
  - Iterative Back Translation

Unsupervised Lexicon Induction

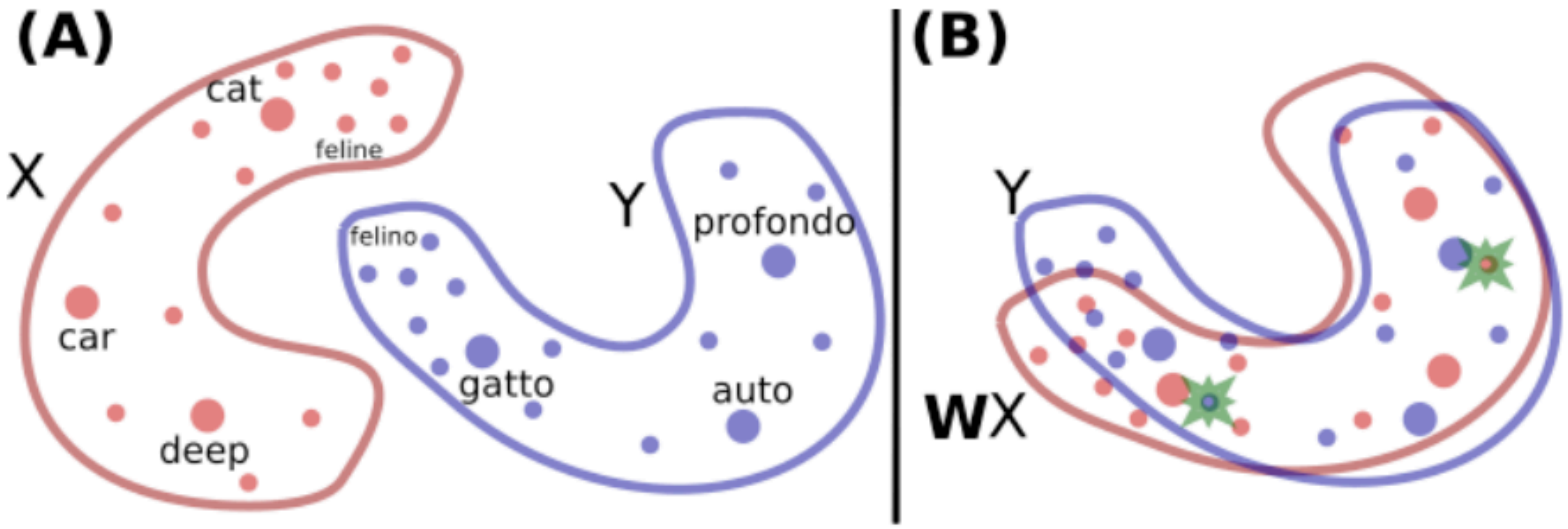
- Also called word translation
- Hypothesis: words with the same meaning in two languages share isomorphic embedding space



Zhang et al. Adversarial Training for Unsupervised Bilingual Lexicon Induction. 2017

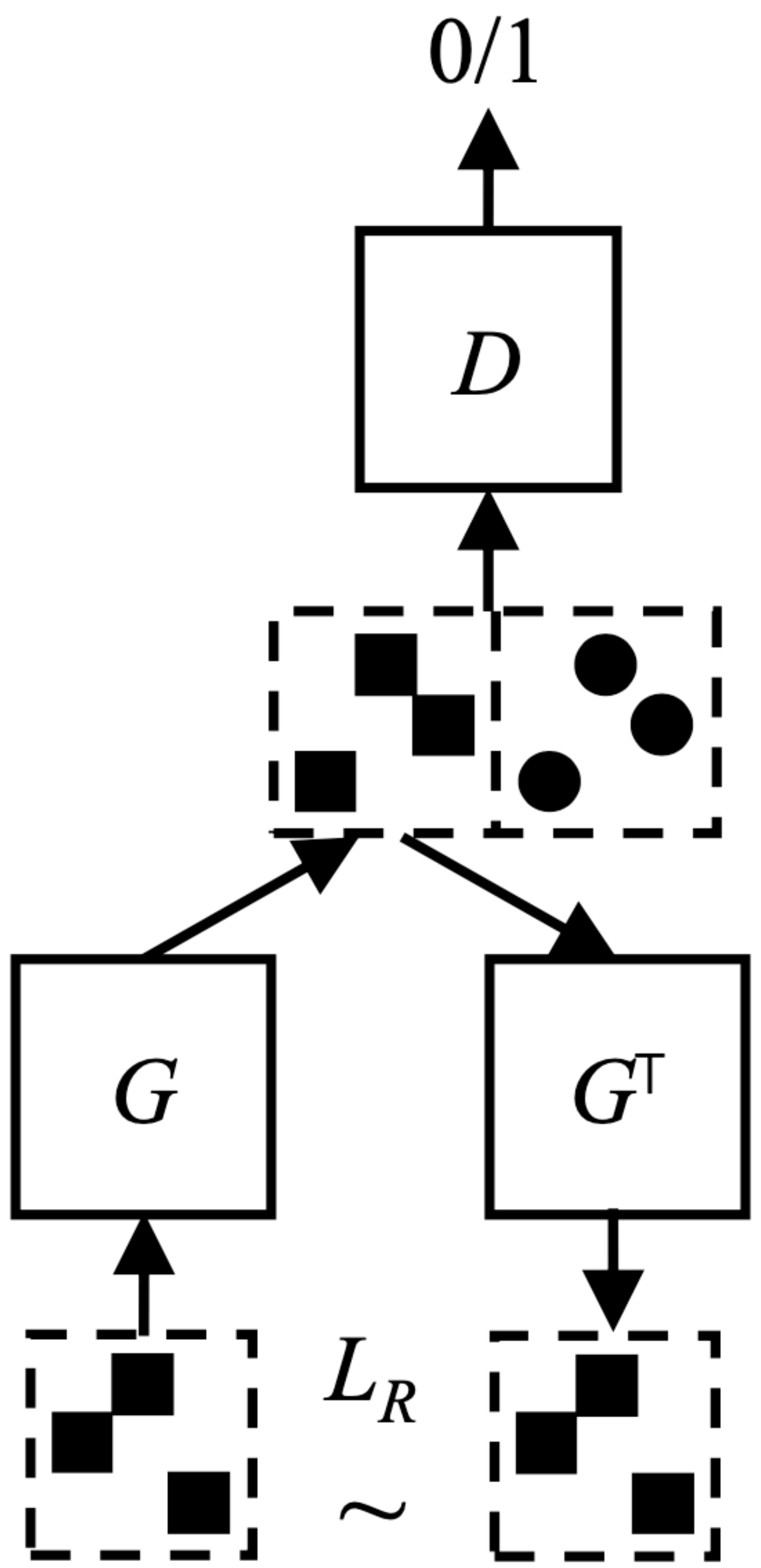
Lexicon Induction: Mapping of the Embedding Space

- To learn a matrix W
- Supervised setting (pairs of aligned words available)
  $\arg\min \|XW - Y\|_f$
  - closed form solution for this
- How to learn W without aligned word pairs?



Lexicon Induction via Adversarial Training

- x, y are pretrained word embeddings in two languages. But not aligned.
- Using a discriminator to distinguish between
  – Wx and y
  – A feedforward NN with 1 hidden layers.
- Alternating between
  – $\min_{D} L_D = -\log D(y) - \log(1 - D(Wx))$
  – $\min_{W} L_G = -\log D(Wx) - \cos(x, W^T Wx)$



Zhang et al. Adversarial Training for Unsupervised Bilingual Lexicon Induction. 2017

Find the closest words

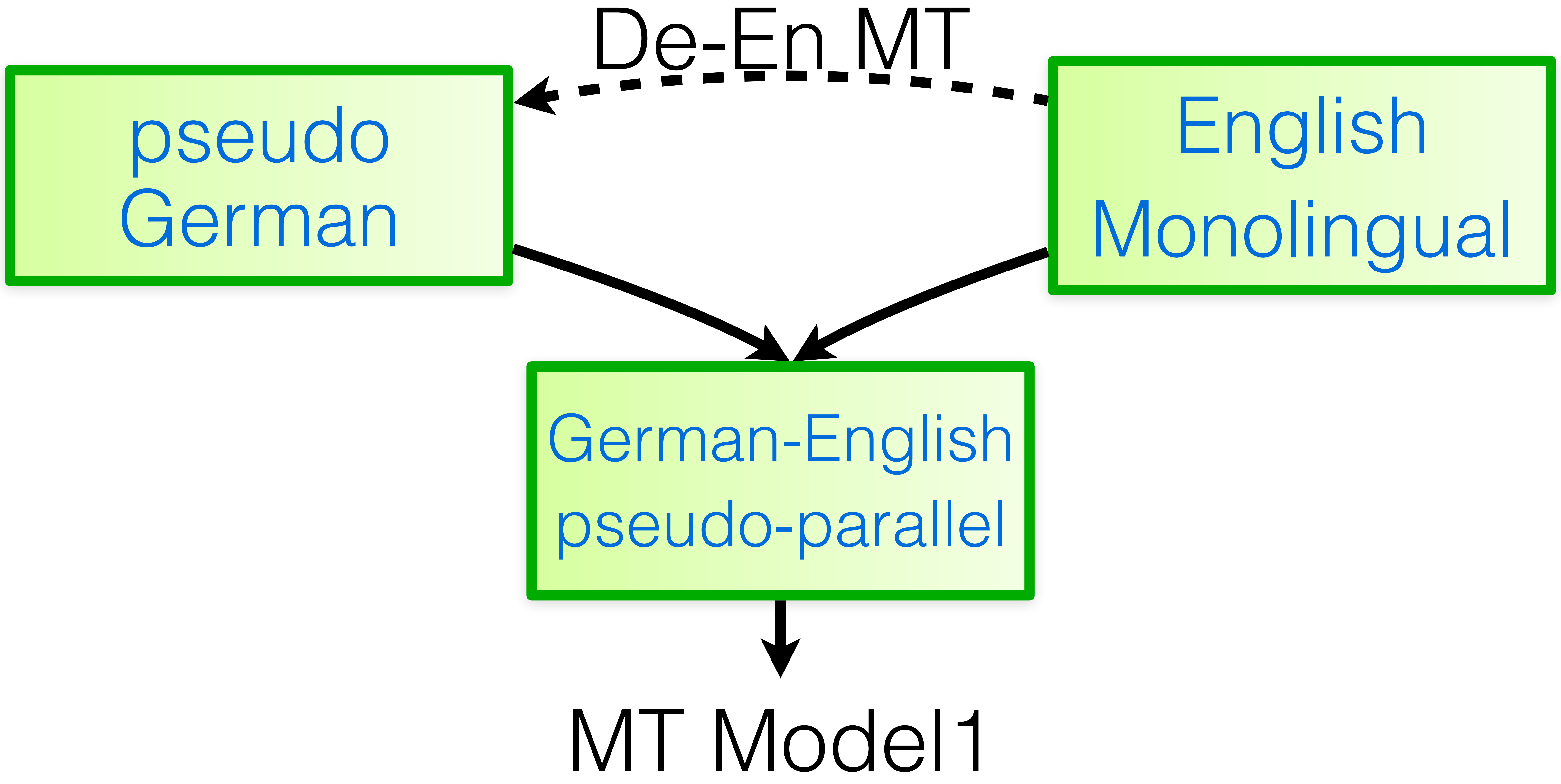
- Use this as the word-level translation

method	# seeds	es-en	it-en	ja-zh	tr-en
MonoGiza w/o embeddings	0	0.35	0.30	0.04	0.00
MonoGiza w/ embeddings	0	1.19	0.27	0.23	0.09
TM	50	1.24	0.76	0.35	0.09
	100	48.61	37.95	26.67	11.15
IA	50	39.89	27.03	19.04	7.58
	100	60.44	46.52	36.35	17.11
Ours	0	71.97	58.60	43.02	17.18

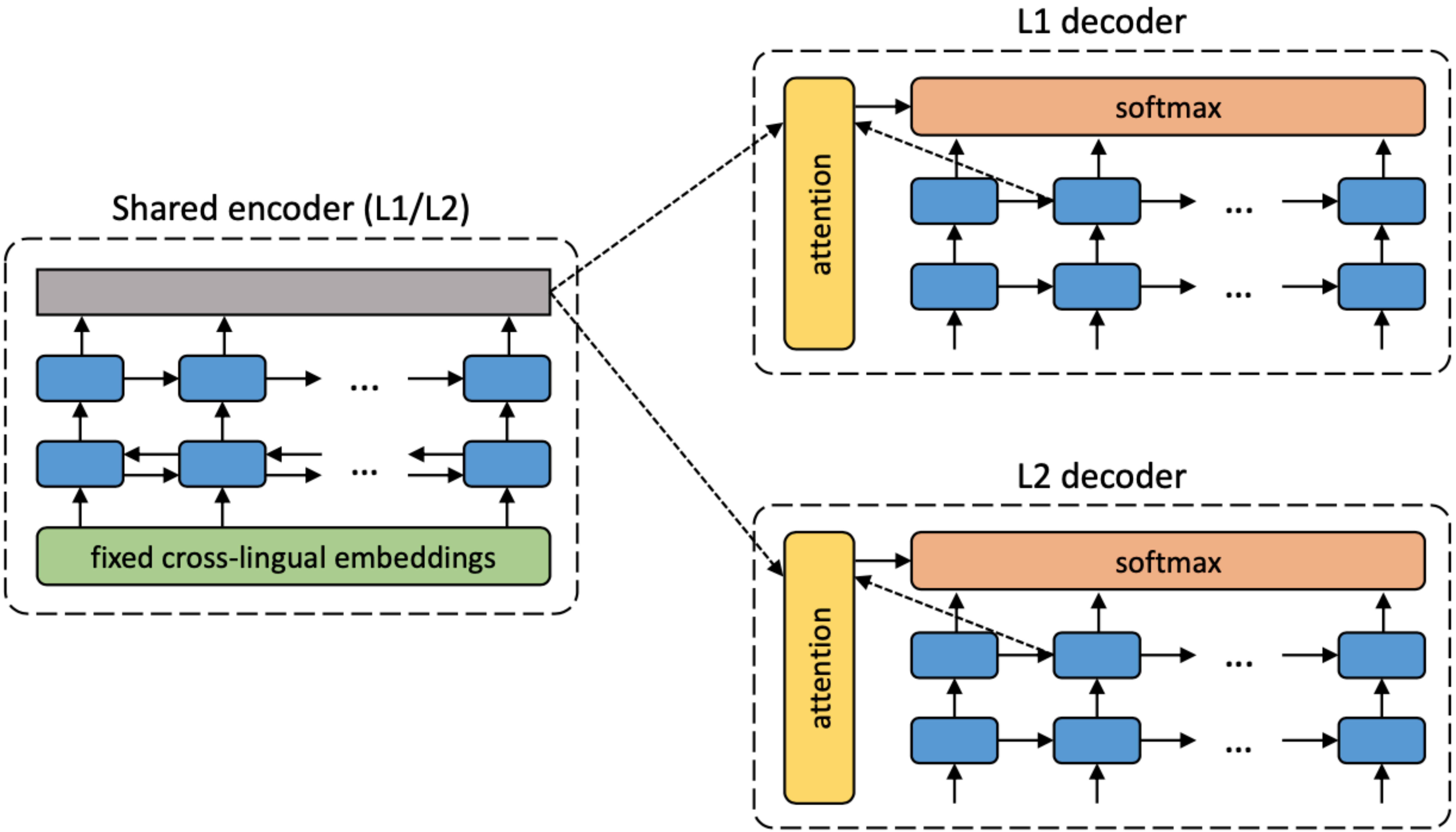
Zhang et al. Adversarial Training for Unsupervised Bilingual Lexicon Induction. 2017

Unsupervised Machine Translation

- Build an initial MT system to translate from English -> German, and German -> English using word-level translation
- Iterate

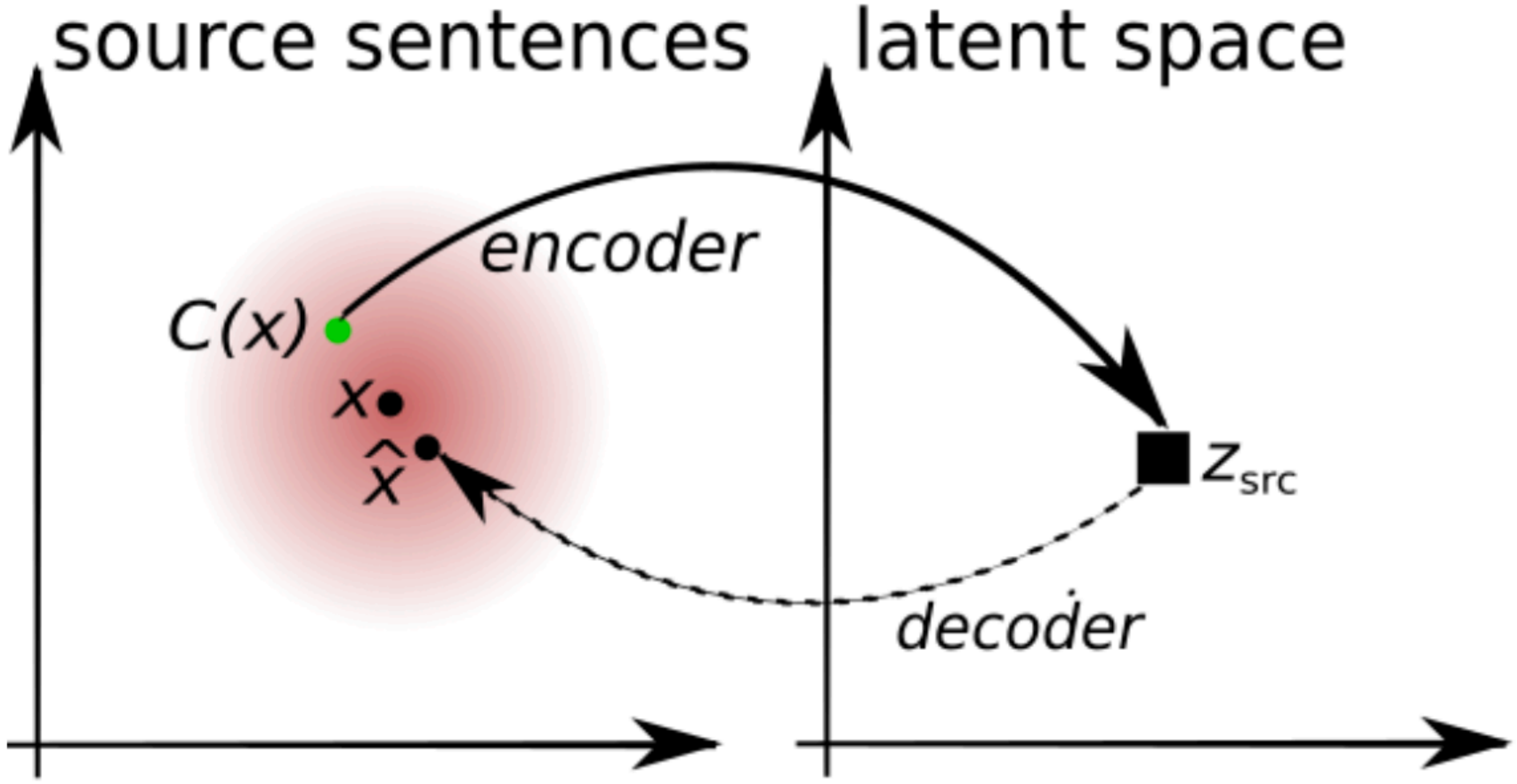


Shared Encoder with Dual Decoder



Training Objective 1: Denoising Autoencoder

- Create a noisy version of source sentence, and reconstruct using encoder-decoder
- Using cross-entropy loss on reconstructed sentence

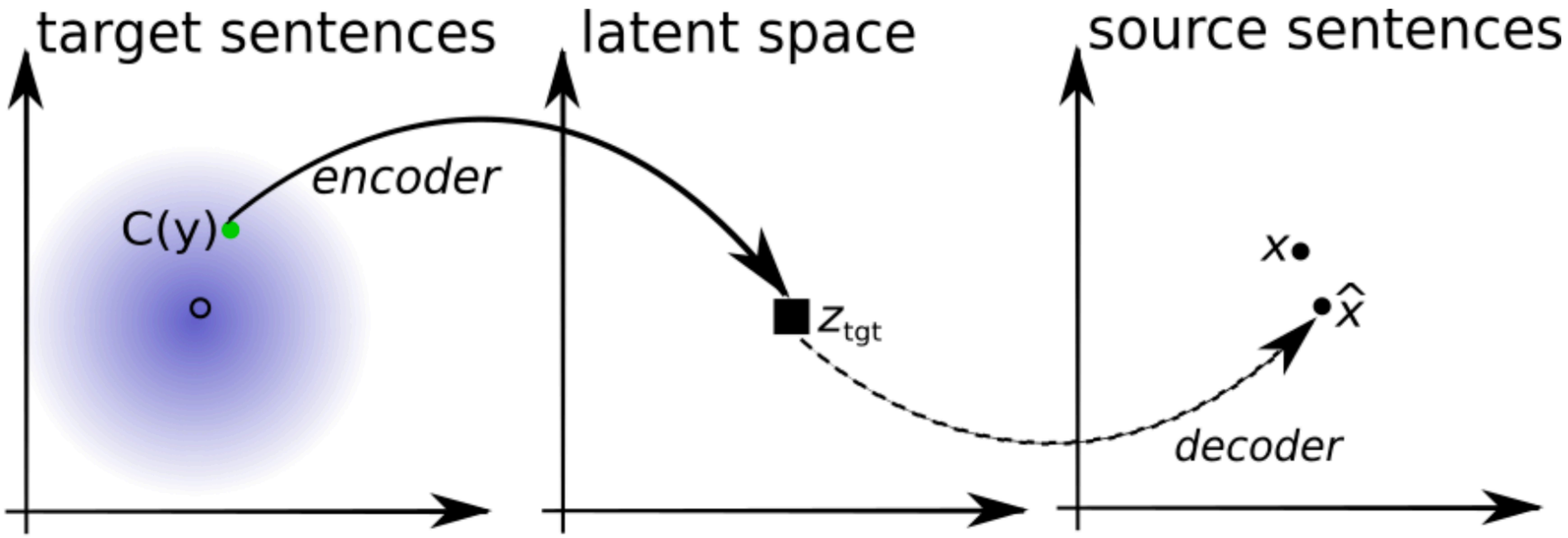


Artetxe et al. Unsupervised Neural Machine Translation. 2018

Lampele et al. Unsupervised Machine Translation Using Monolingual Corpora Only. 2018

Training Objective 2: Back-translation

- Back-translate: From target to generate pseudo-parallel source sentence



Artetxe et al. Unsupervised Neural Machine Translation. 2018

Lample et al. Unsupervised Machine Translation Using Monolingual Corpora Only. 2018

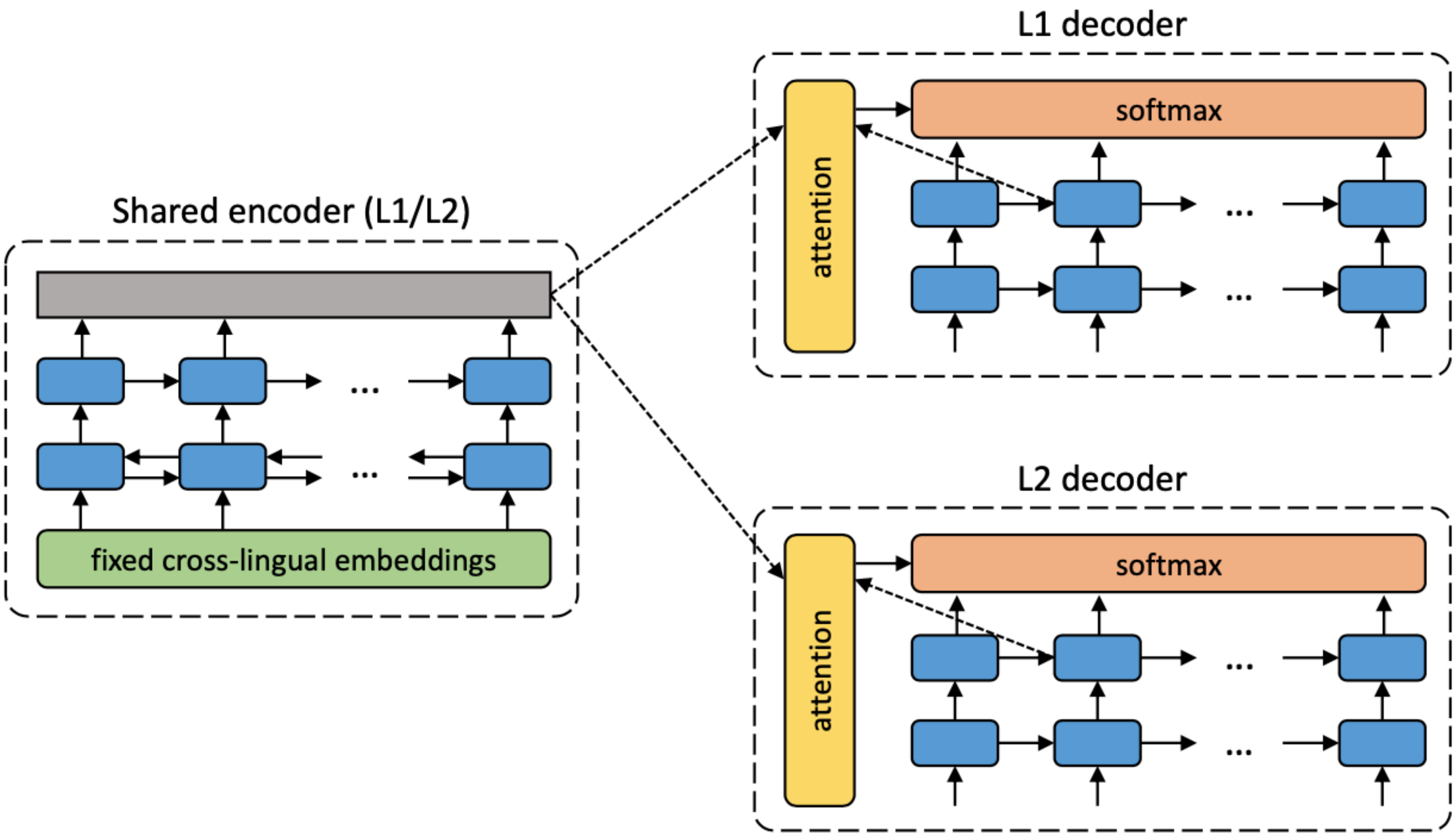
Training Objective 3: Adversarial Loss

- To distinguish between source and target sentence embeddings.

$$\min L_D = -\log P_D(0 \text{ or } 1 \mid \mathrm{emb}(\text{src or tgt}))$$

Lampe et al. Unsupervised Machine Translation Using Monolingual Corpora Only. 2018

Unsupervised Neural Machine Translation



Artetxe et al. Unsupervised Neural Machine Translation. 2018

Does it work?

	Multi30k-Task1				WMT			
	en-fr	fr-en	de-en	en-de	en-fr	fr-en	de-en	en-de
Supervised	56.83	50.77	38.38	35.16	27.97	26.13	25.61	21.33
word-by-word	8.54	16.77	15.72	5.39	6.28	10.09	10.77	7.06
word reordering	-	-	-	-	6.68	11.69	10.84	6.70
oracle word reordering	11.62	24.88	18.27	6.79	10.12	20.64	19.42	11.57
Our model: 1st iteration	27.48	28.07	23.69	19.32	12.10	11.79	11.10	8.86
Our model: 2nd iteration	31.72	30.49	24.73	21.16	14.42	13.49	13.25	9.75
Our model: 3rd iteration	32.76	32.07	26.26	22.74	15.05	14.31	13.33	9.64

Bidirectional LSTM encoder-decoder

Lample et al. Unsupervised Machine Translation Using Monolingual Corpora Only. 2018

When does Unsupervised NMT work?

- Similar languages with large monolingual data
- Distant languages are still difficult
- Eg. En-Tr 4.5 (unsupervised) vs. 20 (supervised)

Reading

- Sennrich et al. Improving Neural Machine Translation Models with Monolingual Data. ACL 2016.
- Cheng et al. Semi-Supervised Learning for Neural Machine Translation. ACL 2016.
- Artetxe et al. Unsupervised Neural Machine Translation. 2018
- Lample et al. Unsupervised Machine Translation Using Monolingual Corpora Only. 2018
- He et al. Dual Learning for Machine Translation. 2016.
- Gulcehre et al. On Using Monolingual Corpora in Neural Machine Translation. 2015
- Edunov et al. Understanding Back-translation at Scale. 2018.

Code Walk

- There will be no graded discussion, but we'll have a code walk through The Annotated Transformer https://nlp.seas.harvard.edu/2018/04/03/attention.html
- Organize into group to discuss some of the design decisions, their motivation, etc.