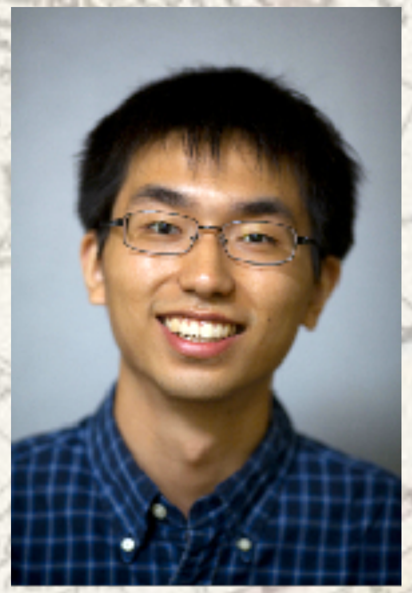
The Effect of Place in MT

Marc'Aurelio Ranzato
Facebook AI Research - NYC
ranzato@fb.com

joint work with:

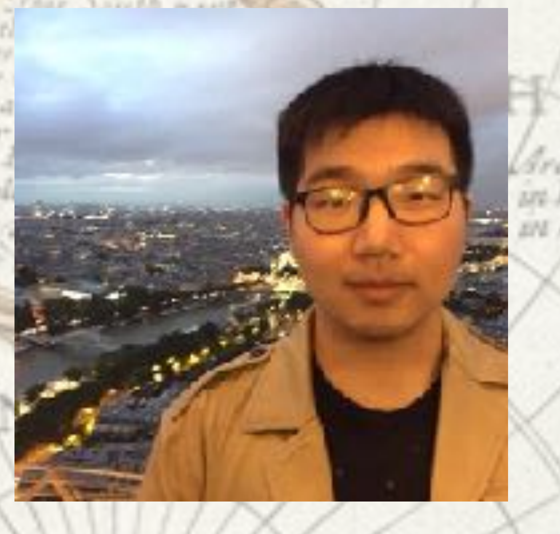

Peng-Jen Chen


Jiajun Shen


Matt Le


Junxian He


Myle Ott


Jiatao Gu


Michael Auli

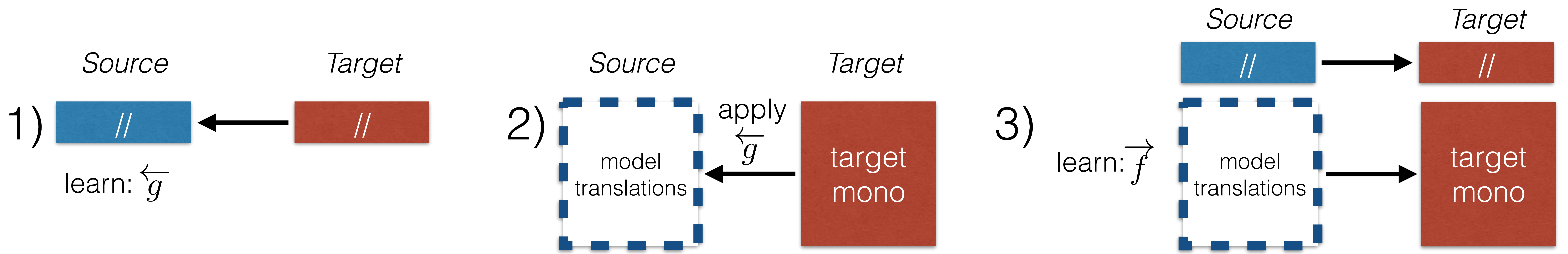
WeCNLP 6 September 2019

Machine Translation - AD 2019

- Human level performance on some languages in some domains under some conditions…
- Key factors of success:
    - big data (parallel & monolingual), big models & big compute
    - better architectures like transformer
    - **back-translation** (BT)

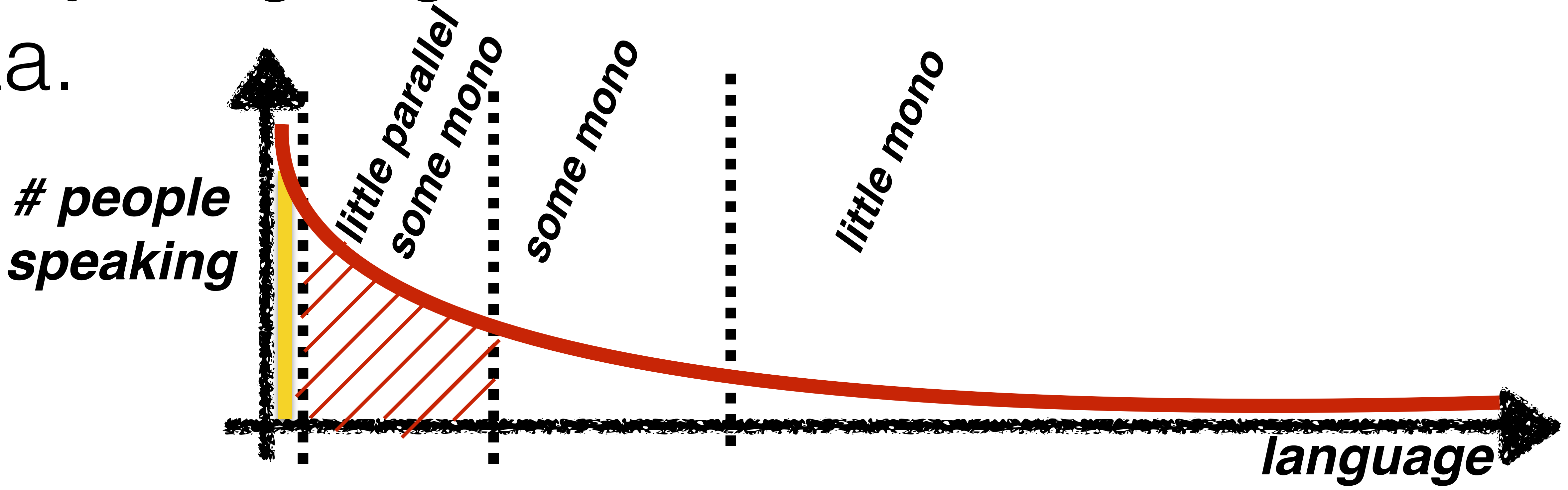
Attention is all you need. Vaswani et al. NIPS 2017
Improving NMT with monolingual data. Sennrich et al. ACL 2015

BT



Machine Translation - AD 2019

- There are 6000+ languages in the world. Very few enjoy large parallel resources.
- **MT for low resource languages**. In addition to back-translation, also:
  - initialization
  - multi-lingual training
  - noisy parallel data from ParaCrawl
- Problem: the tail is very long… for many languages these methods are not applicable because of lack of data.



Setting

- Data:
  - a small parallel dataset (~10K).
  - some monolingual data on both target and source language (~1M).

- Method:
  - Back-translation (BT).

The Power of BT

Simulating low-resource MT with a high resource language:
using *EuroParl* data with 20K parallel sentences and 100K monolingual target sentences.

EuroParl Fr—>En

only parallel data	30.4 BLEU
parallel data + BT	33.8 BLEU

+3.4 BLEU!

https://www.statmt.org/europarl/

A Worrisome Finding

BT sometimes yields very mild improvements.

Example #1

FB public posts En—>My

only parallel data	15.2 BLEU
parallel data + BT	15.3 BLEU

+0.1 BLEU!

Example #2

FLORES evaluation set En—>Ne

only parallel data	4.3 BLEU
parallel data + BT	6.8 BLEU

+2.5 BLEU!

The FLORES evaluation datasets… Guzmán, Chen et al. EMNLP 2019

Sports

- *football*
- *baseball*
- *basketball*

- *soccer*
- *cricket*
- *rowing*

News

- ***hurricane***
- ***trade war***
- ***elections***

- ***lake preservation***
- ***earthquake***
- ***heritage***

Food

- ***hamburger***
- ***clam chowder***
- ***apple pie***

- ***kimchi***
- ***bulgogi***
- ***bibimbap***

The Place Effect

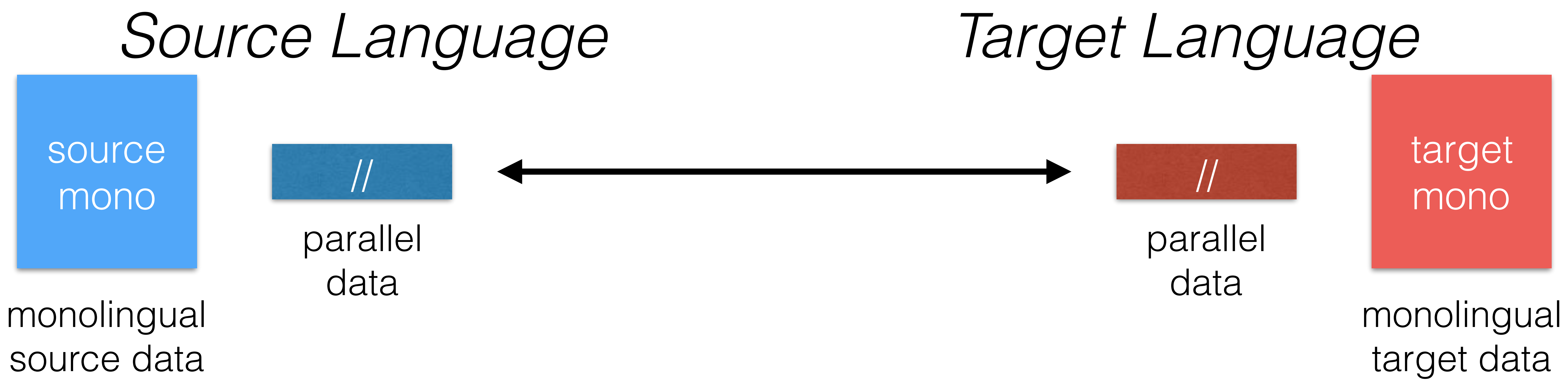
- Def.: Content produced in blogs, social networks, news outlets, etc. varies with the geographic location.

- The place effect is even more pronounced in low resource MT, where source & target geographic locations are typically farther apart and cultures have more distinct traits.

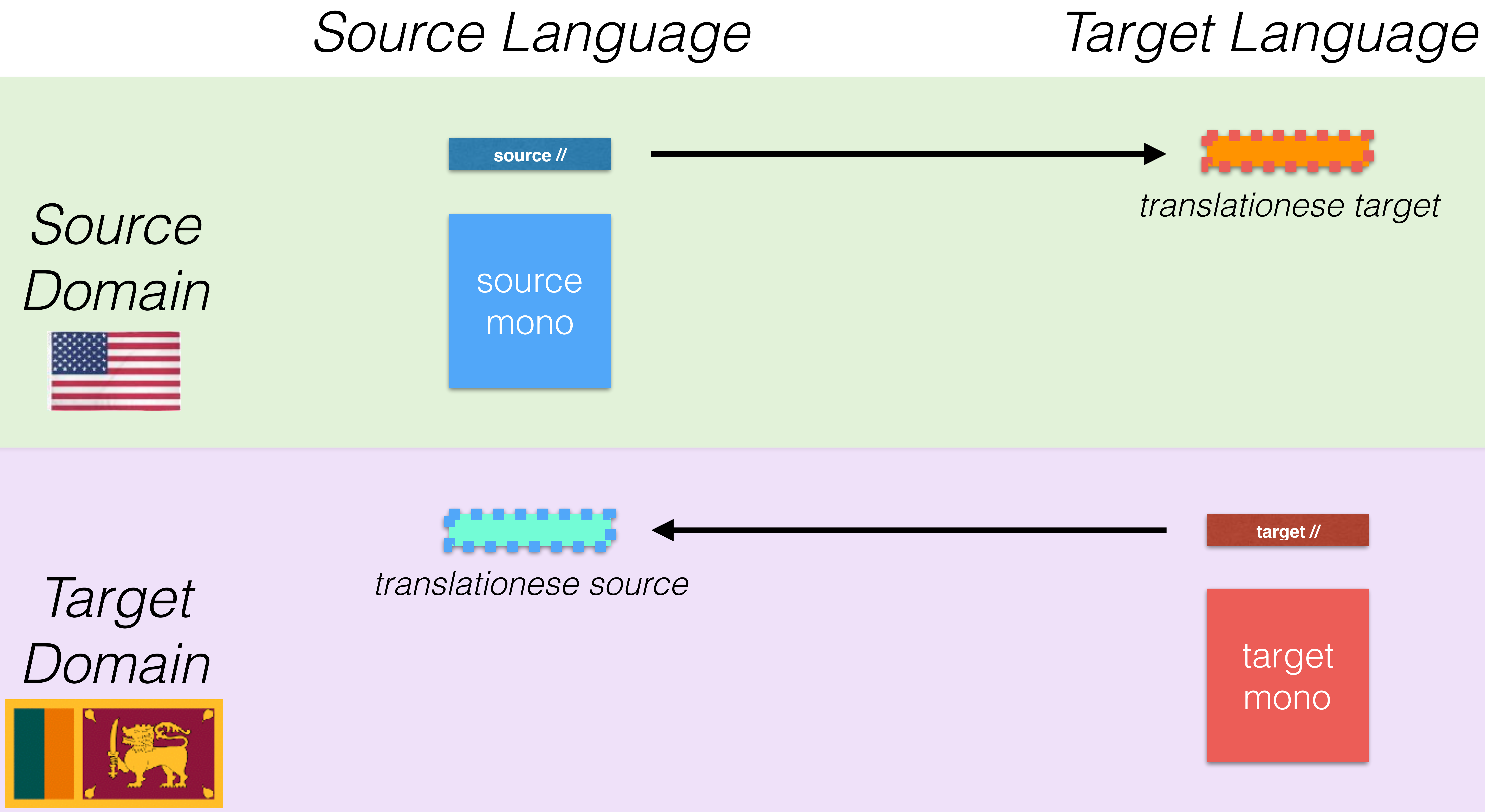
- The place effect makes the MT problem even harder, because of **source/target domain mismatch**.

Language and place. Johnstone Cambridge Univ. Press 2010
A latent variable model for geographic lexical variation. Eisenstein et al. ACL 2010

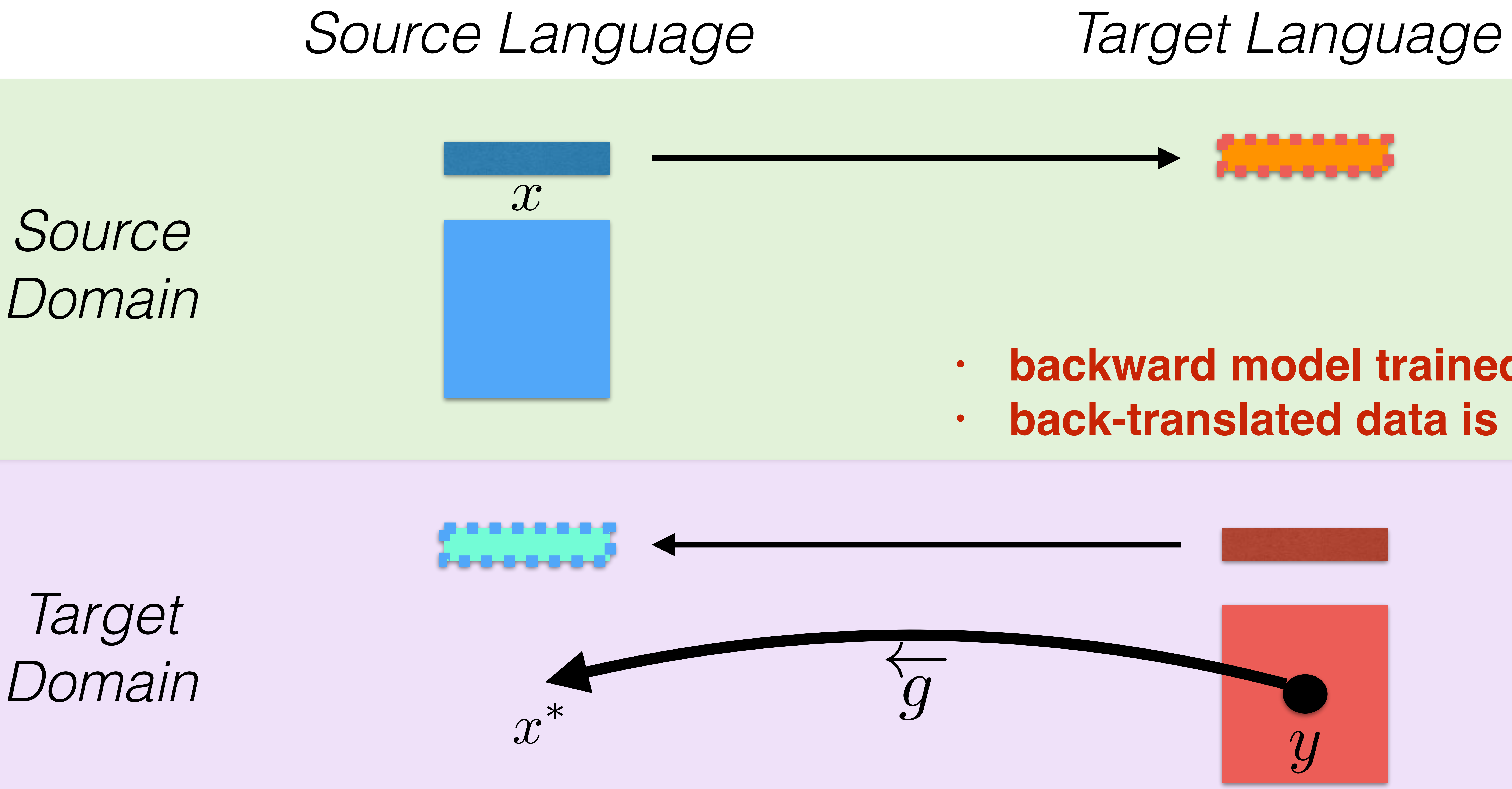
Source / Target Domain Mismatch



Source / Target Domain Mismatch



STDM Cripples BT



- **backward model trained with mixed-domain data**
- **back-translated data is out-of-domain**

Questions

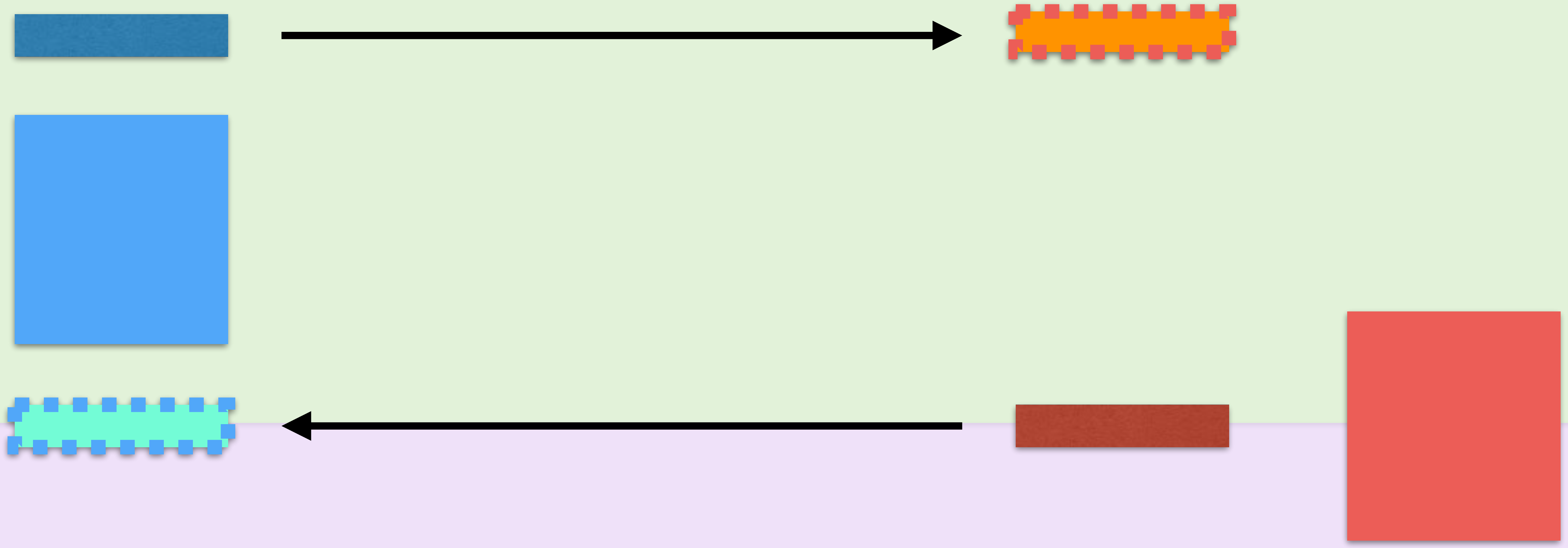
- Is it true that BT is less effective when there is STDM?
- What other baselines shall we consider when there is STDM?
- Is out-of-domain data worth using when there is STDM?
- What are general best practices when there is STDM?
- How to study STDM in a controlled setting?

Controlled Setting

Controlled Setting

Source Language: Fr *Target Language: En*

Source Domain
EuroParl

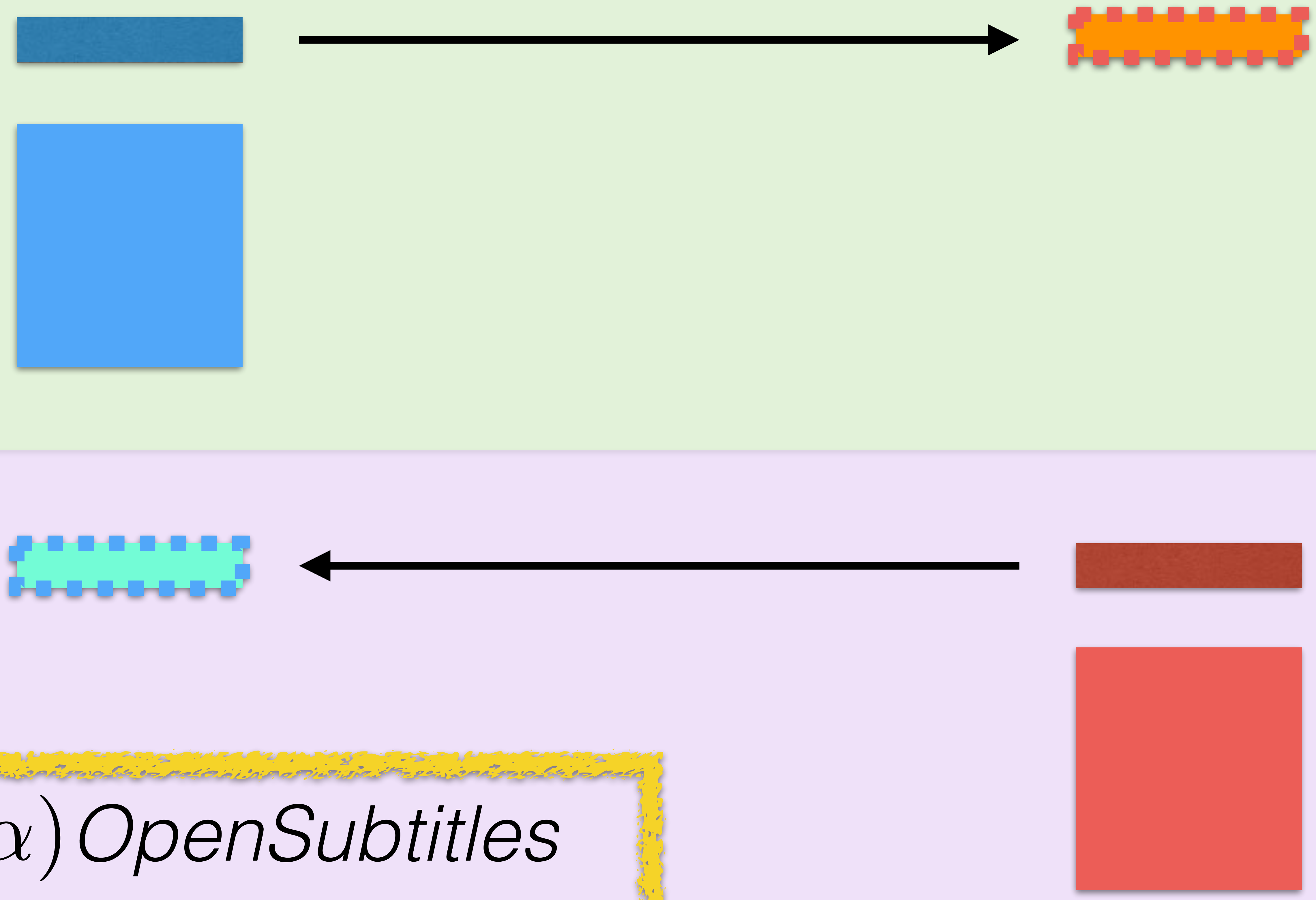


Target Domain
OpenSubtitles

Controlled Setting

Source Language: Fr *Target Language: En*

Source Domain
EuroParl



Target Domain

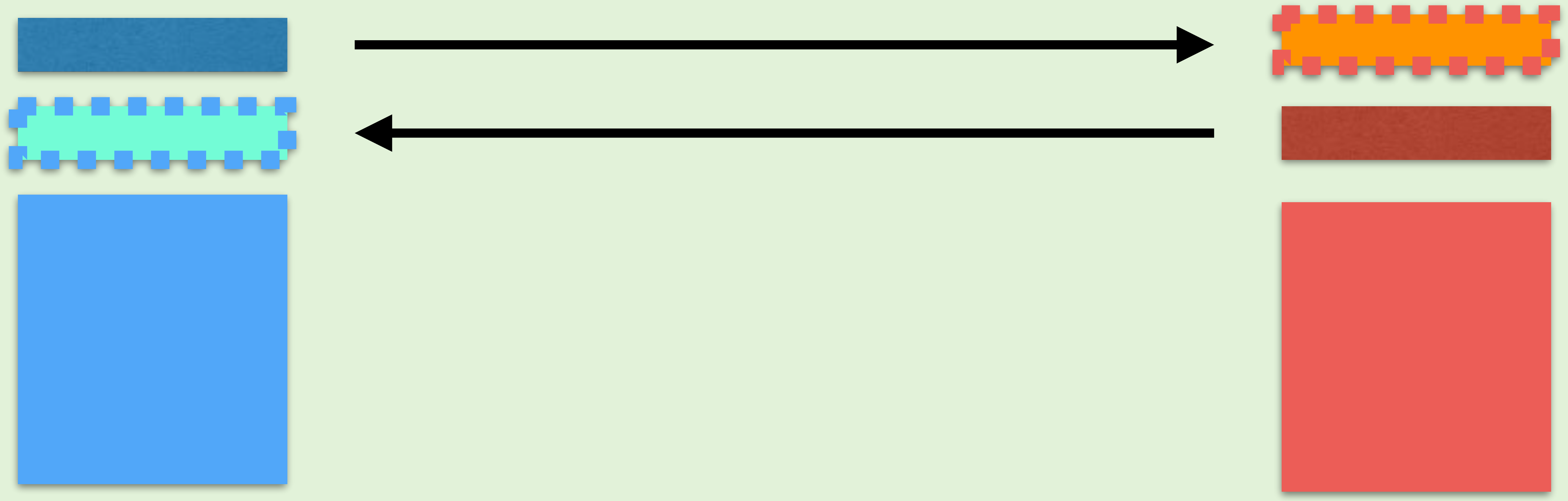
α *EuroParl* $+ (1 - \alpha)$ *OpenSubtitles*

$\alpha = 0$

Controlled Setting

Source Language: Fr *Target Language: En*

Source Domain
EuroParl



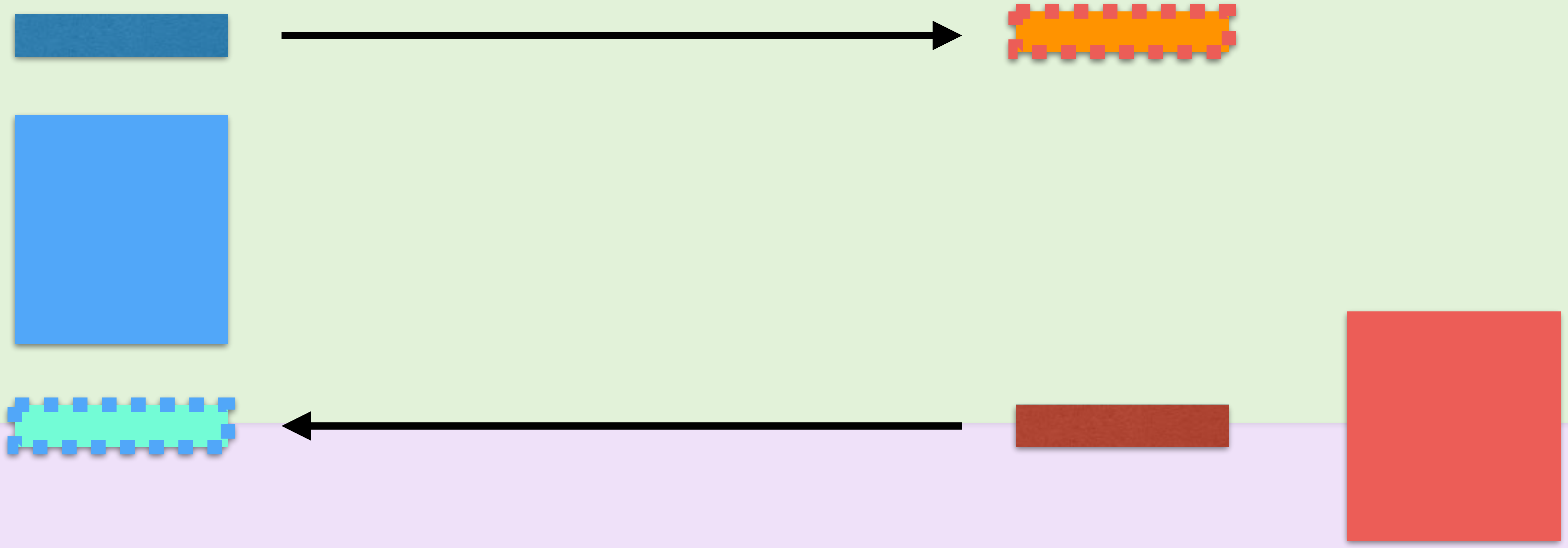
Target Domain

α *EuroParl* $+(1-\alpha)$ *OpenSubtitles*
$\alpha = 1$

Controlled Setting

Source Language: Fr *Target Language: En*

Source Domain
EuroParl

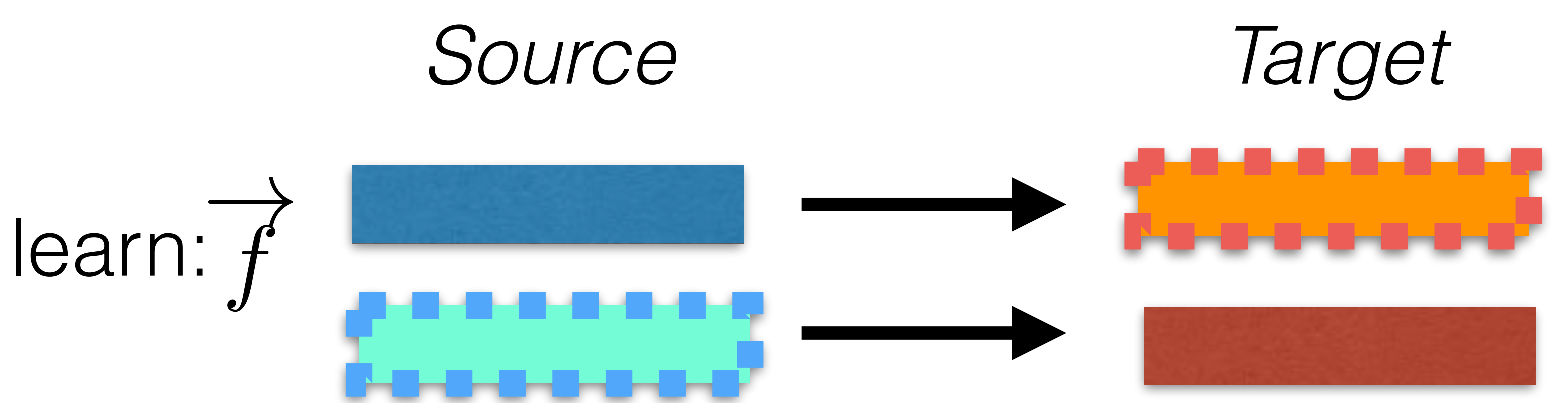


Target Domain

α *EuroParl* $+(1-\alpha)$ *OpenSubtitles*
intermediate value of α

Baseline Approaches

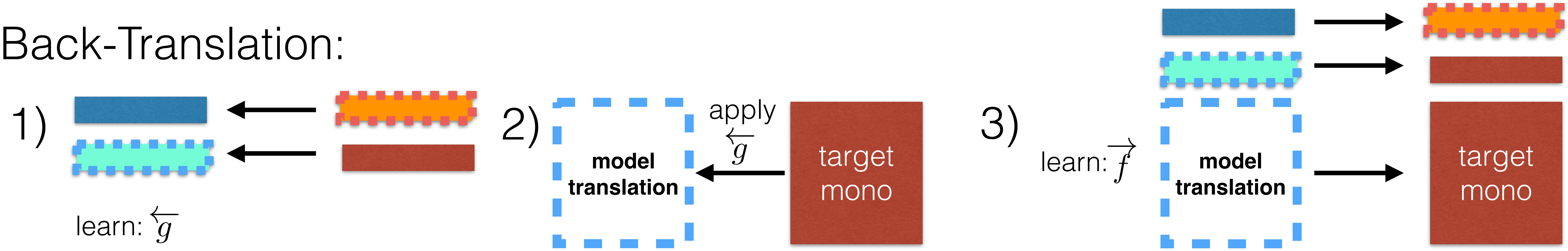
- Bitext only:



Baseline Approaches

- Bitext only:

- Back-Translation:



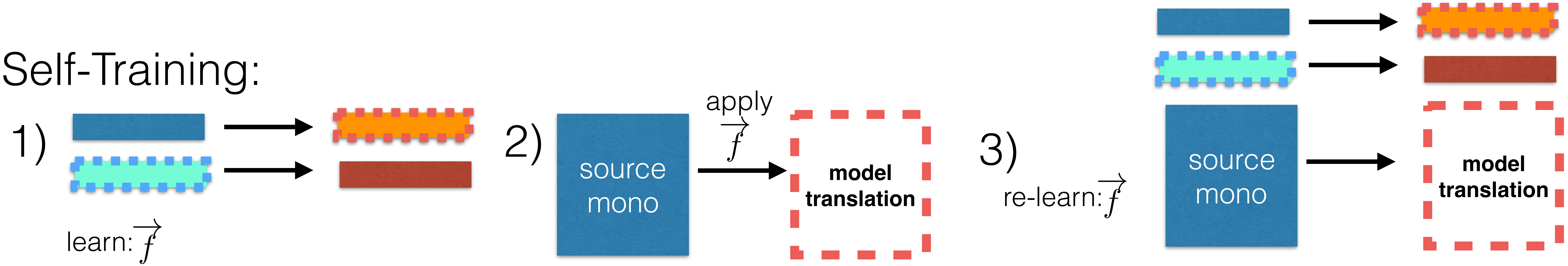
Baseline Approaches

- Bitext only:

Unsupervised word sense disambiguation … Yarowski ACL 1995
Using monolingual source-language data to improve MT performance. Ueffing IWSLT 2006
Exploiting source-side monolingual data in NMT. Zhang et al. EMNLP 2016

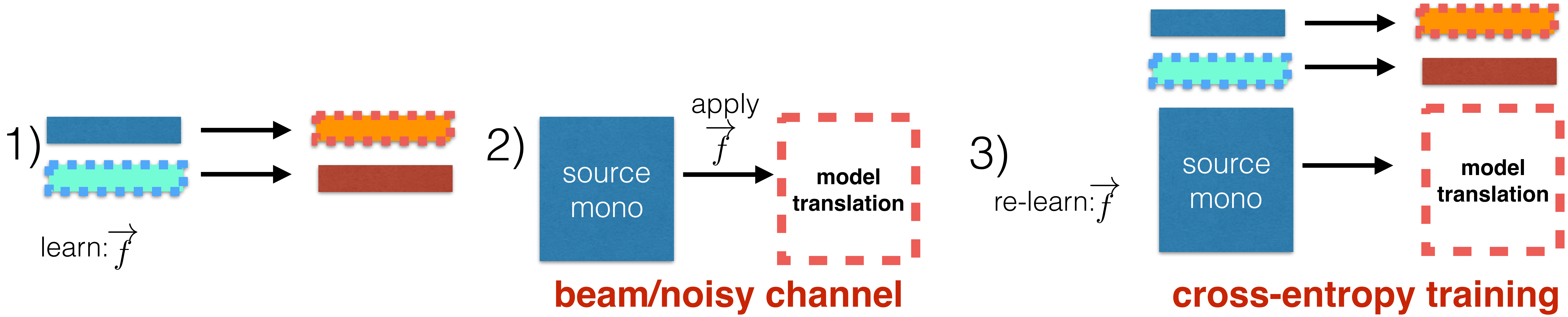
- Back-Translation:

- Self-Training:



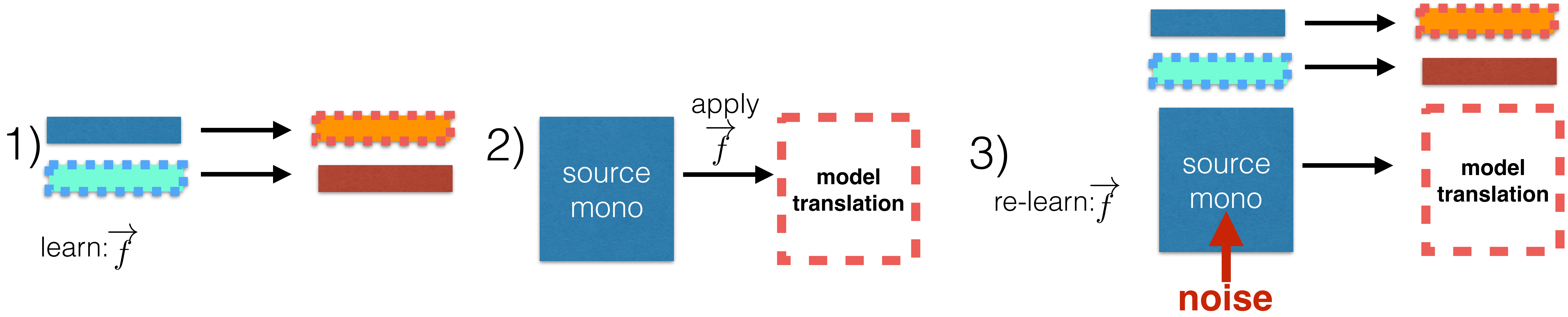
Why Self-Training May Work

- The model learns the decoding process.



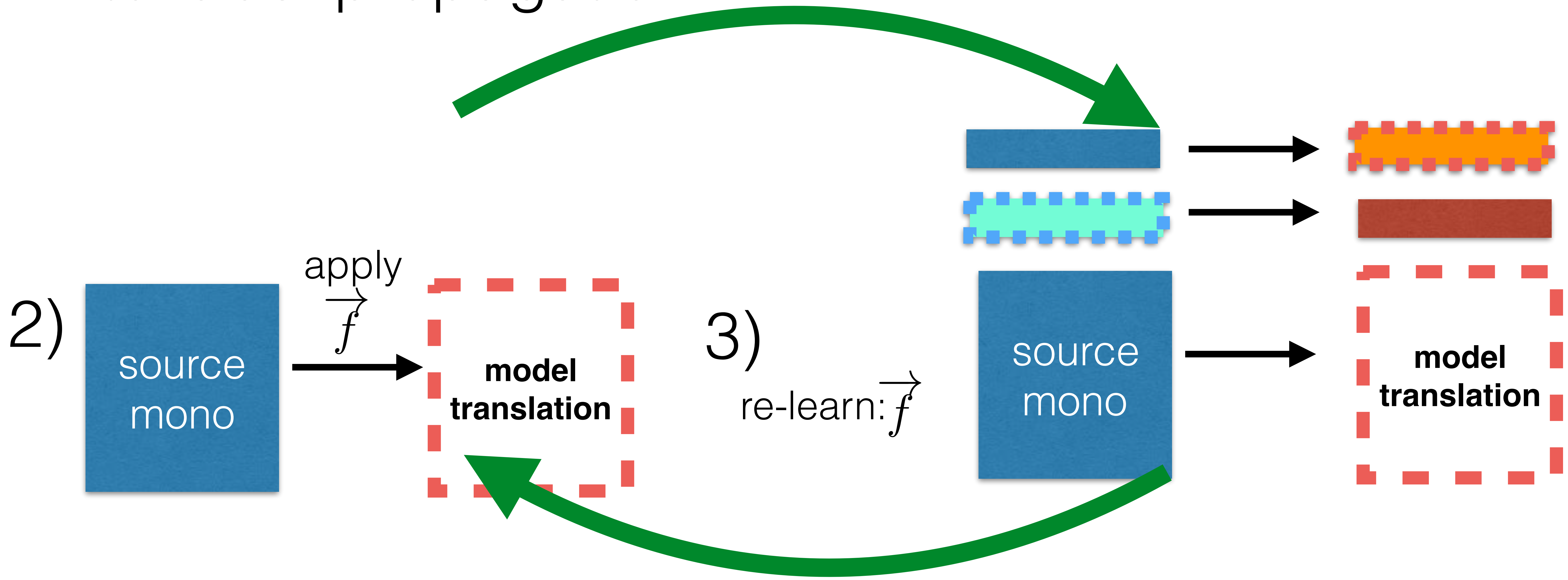
Why Self-Training May Work

- The model learns the decoding process.
- Noise helps mapping similar inputs to the same target.



Why Self-Training May Work

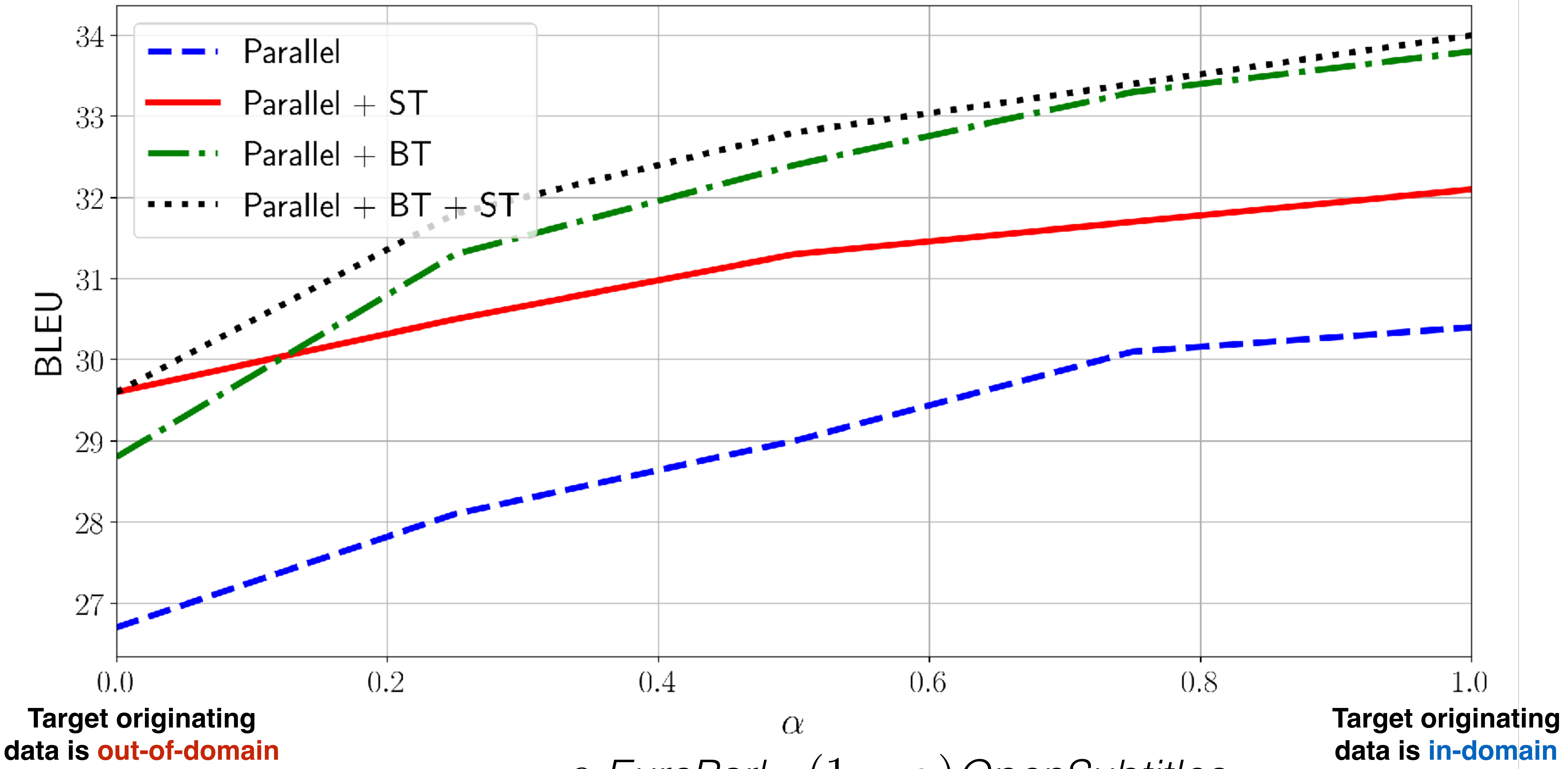
- The model learns the decoding process.
- Noise helps mapping similar inputs to the same target.
- Iterative ST is akin to label propagation.



	target mono data	*source mono data*	*in-domain mono data*
• Back-Translation:	✓	✗	✗
• Self-Training:	✗	✓	✓

Q.: Is it better to have clean targets but out-of-domain data, or noisy targets but in-domain data?
Q.: What's the effect of amount of parallel/monolingual data?
Q.: What's the effect of the quality of the model forward model when training with ST?

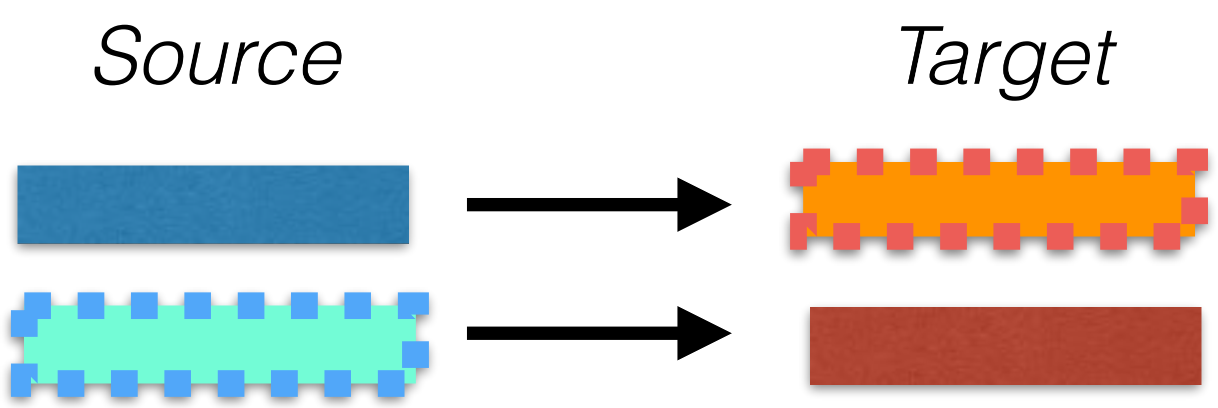
Varying Domain of Target Originating Data



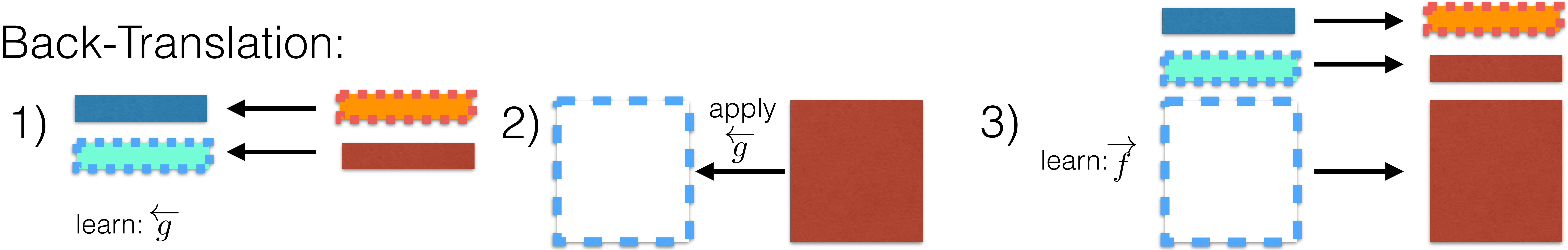
$\alpha\,EuroParl + (1-\alpha)\,OpenSubtitles$

Baseline Approaches

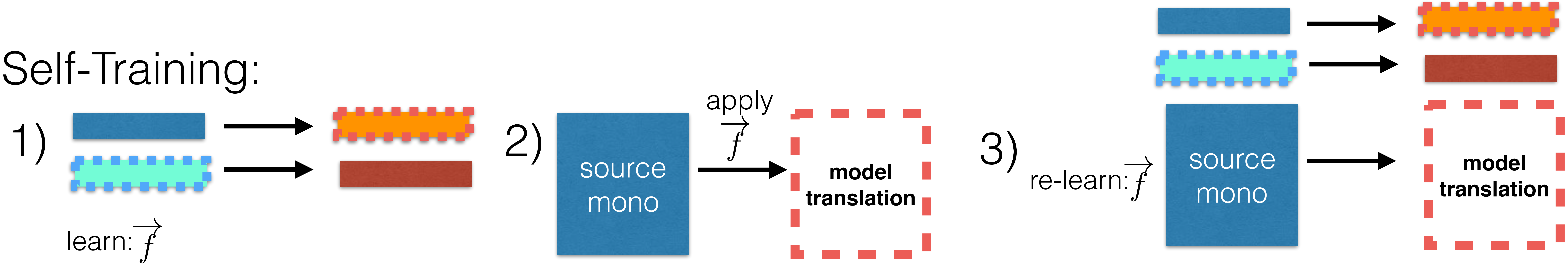
- Bitext only:



- Back-Translation:

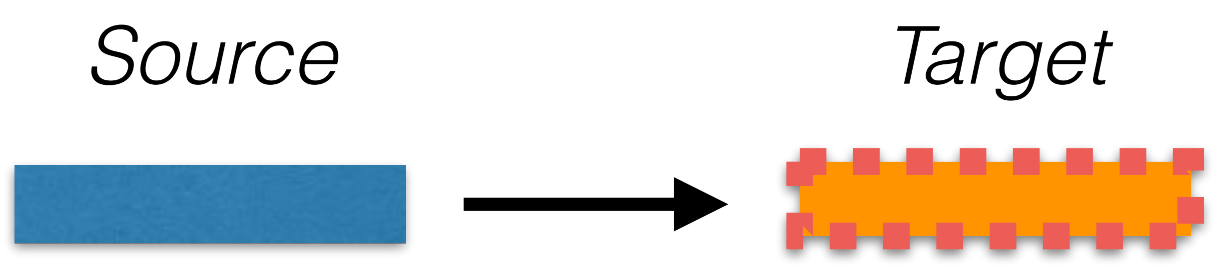


- Self-Training:

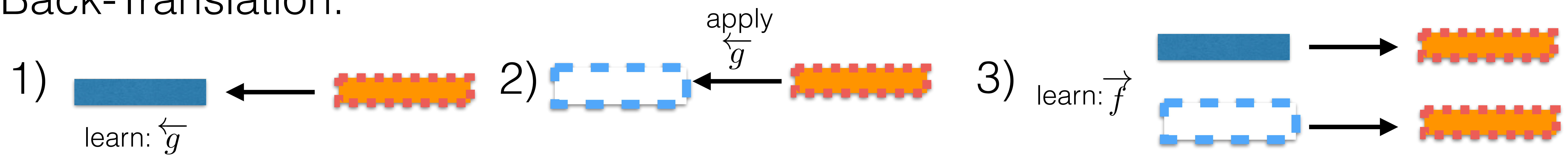


Baseline Approaches: Only In-Domain Data

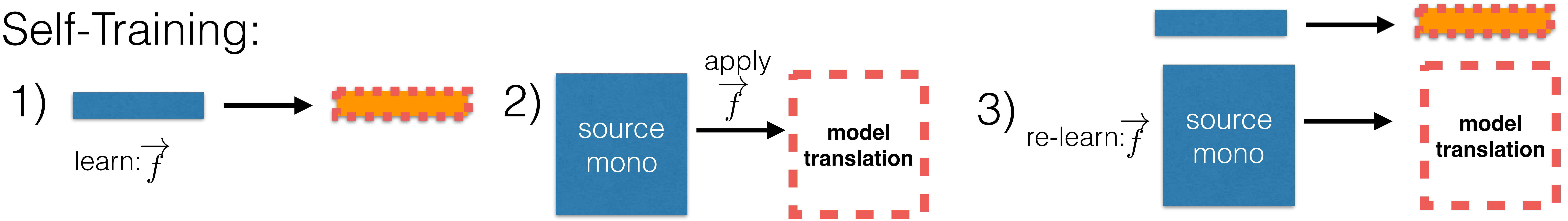
- Bitext only:



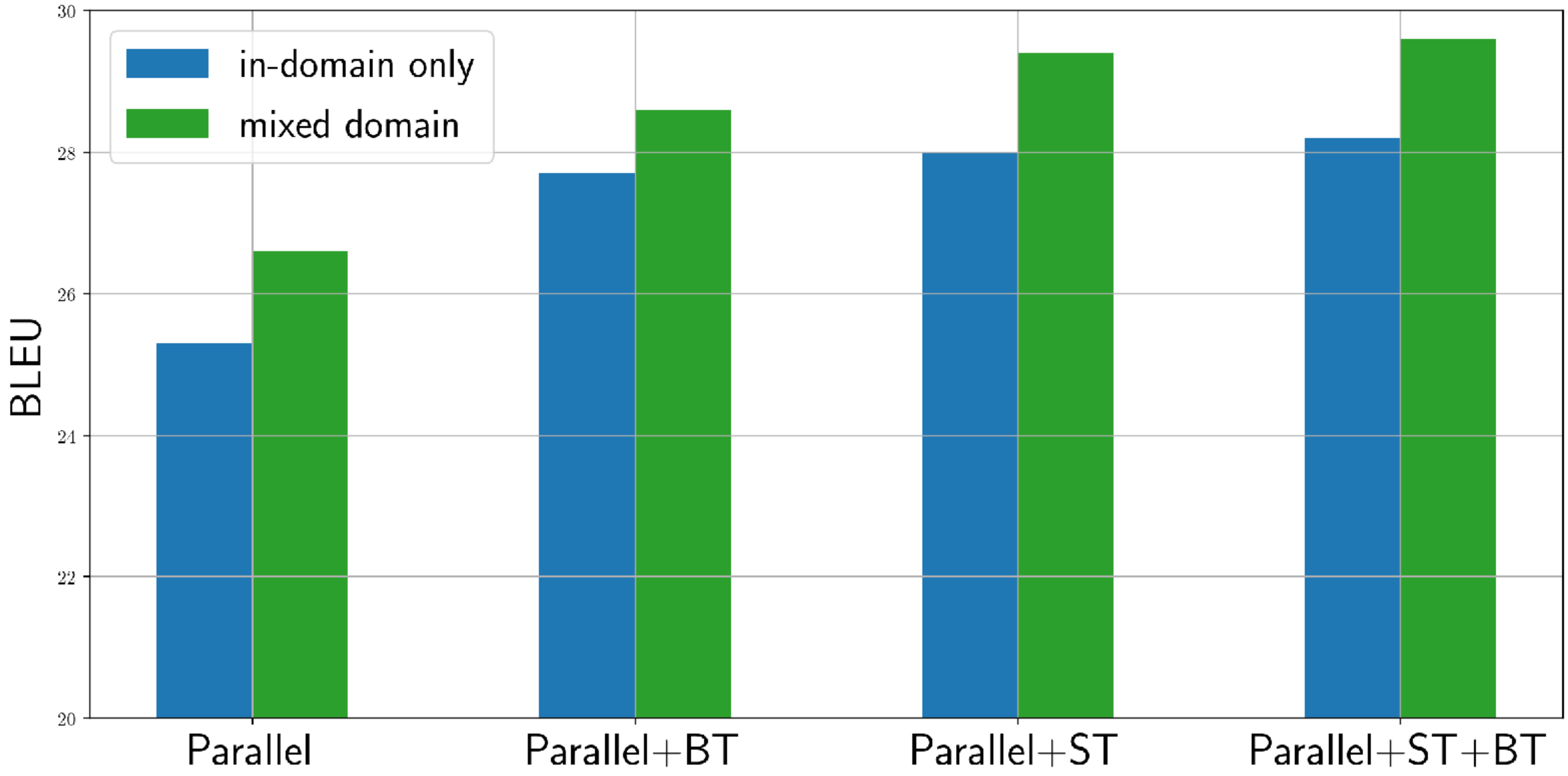
- Back-Translation:



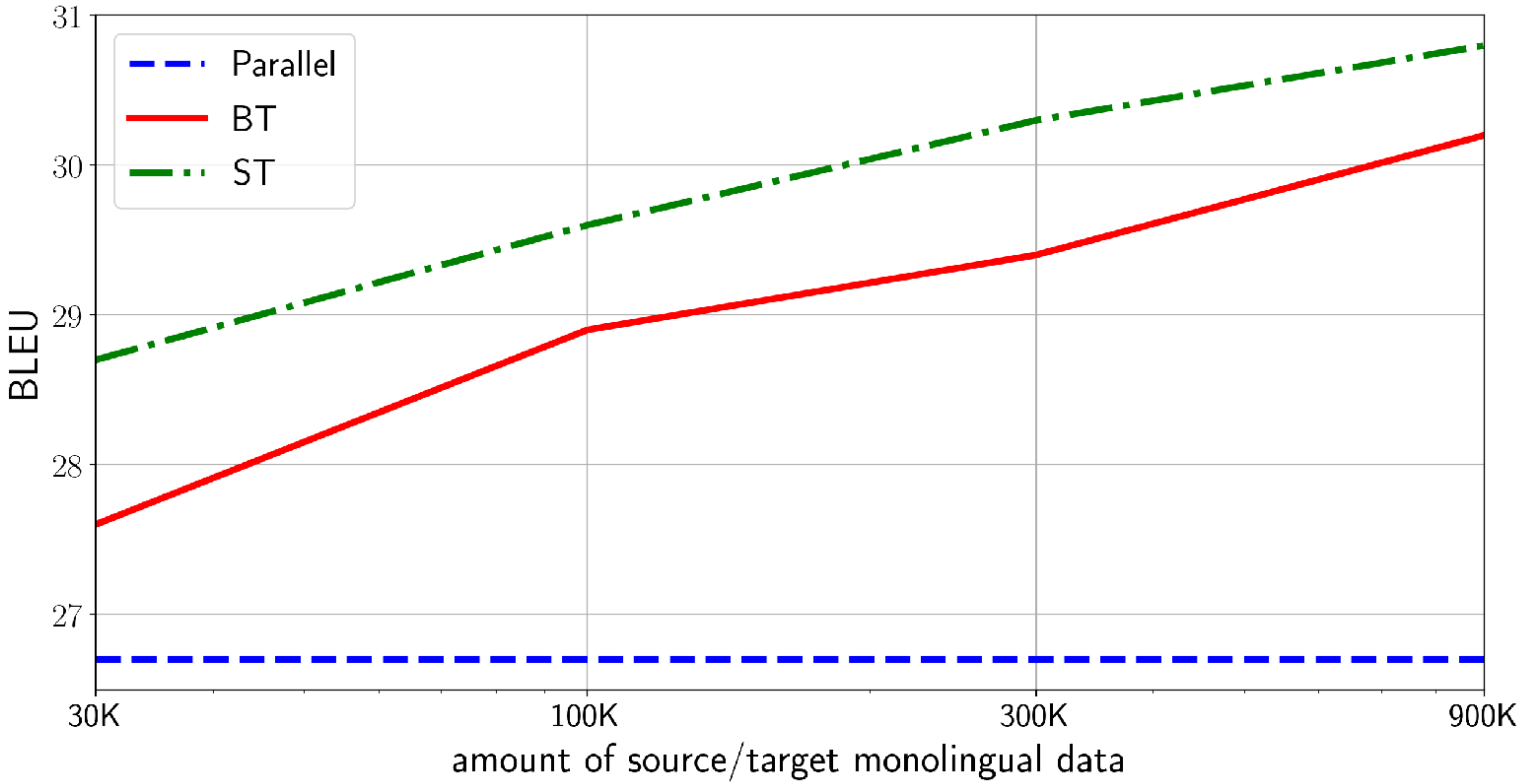
- Self-Training:



In-Domain Only VS. Mixed Domain



Varying Amount of Monolingual Data ($\alpha = 0$)

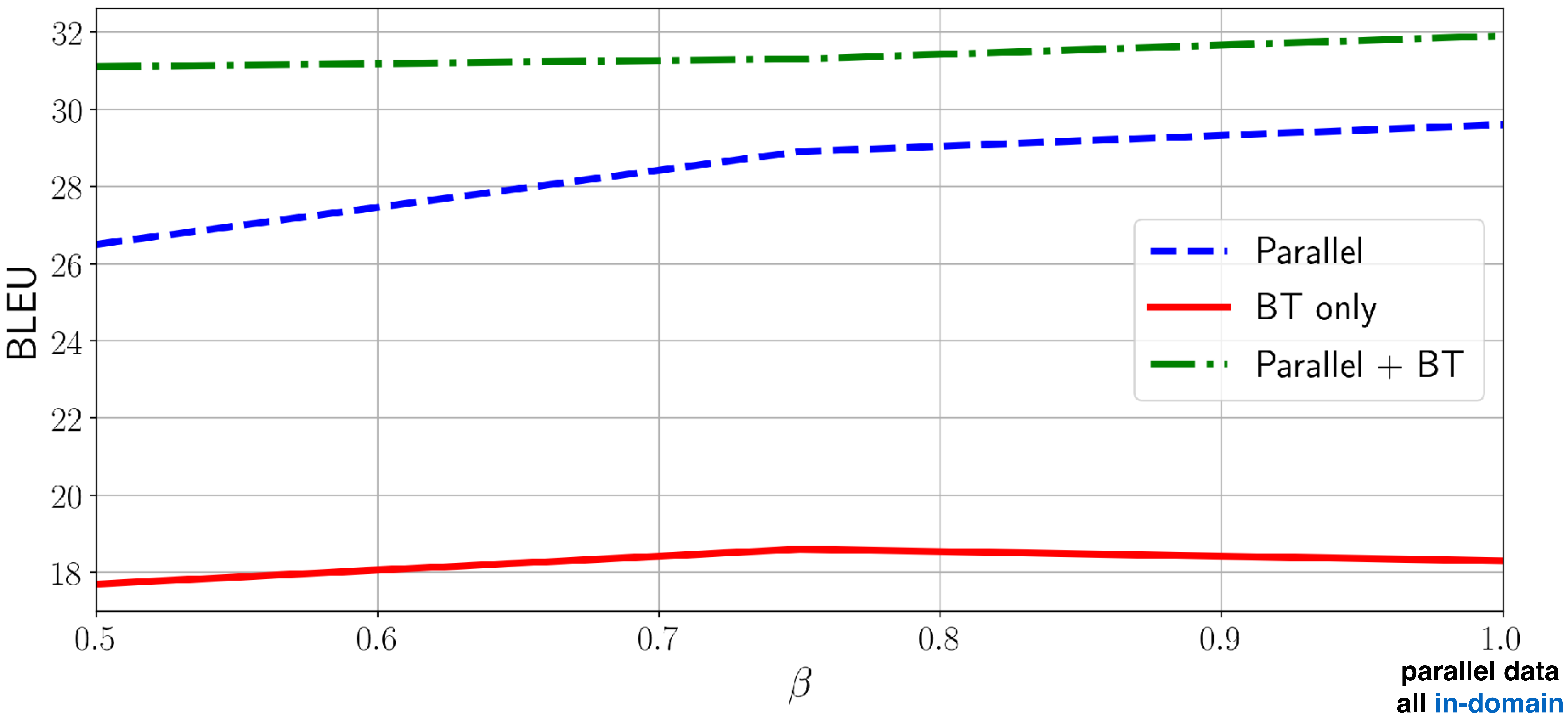


How To Construct Parallel Datasets

- Suppose that:
  - we are interested in forward translation only.
  - we can only translate 20K sentences in total.
  - target data is out-of-domain ($\alpha = 0$).
  - we have 100K target monolingual sentences for BT.

- Is it better to translate 20K sentences all originating from the source? or have some originating from the target as well?

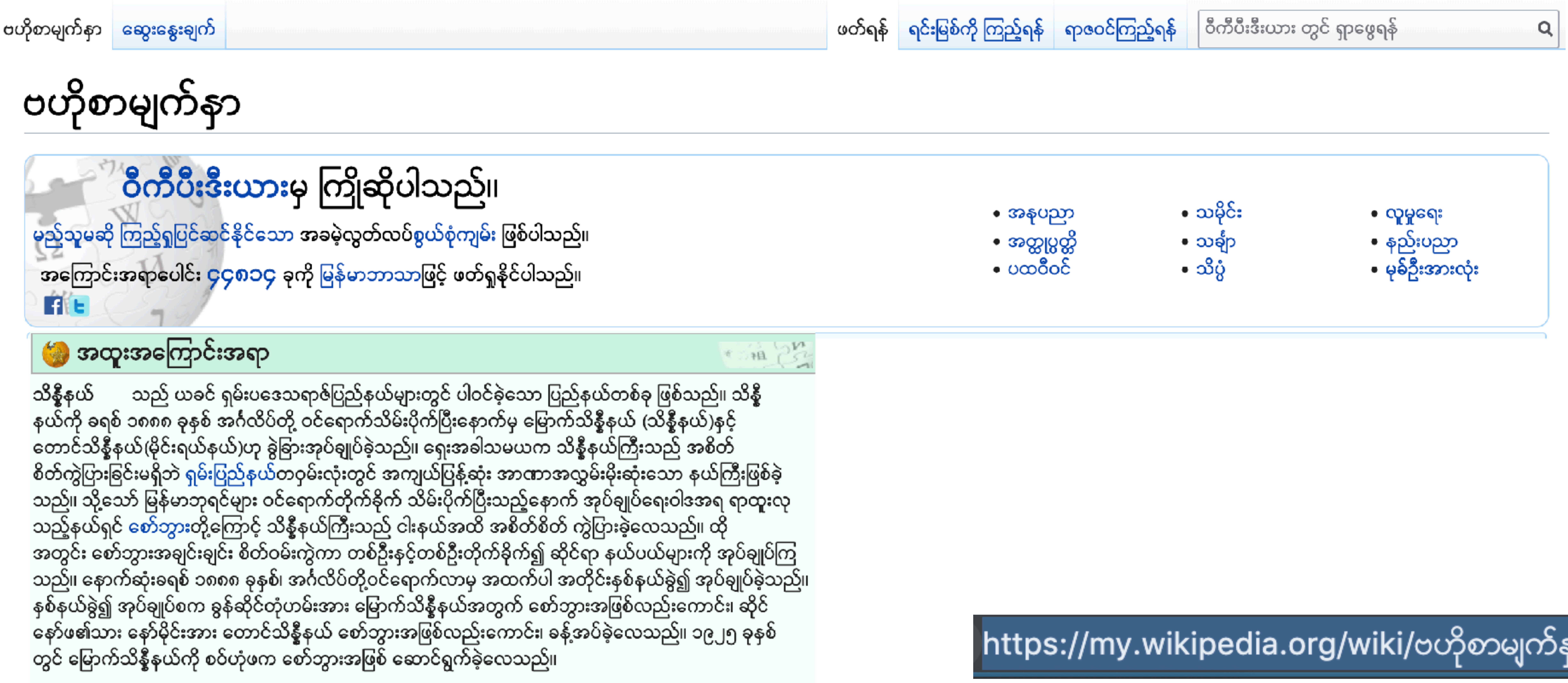
Parallel Dataset = $\underbrace{\beta \text{ EuroParl}}_{\textit{source originating data}} + \underbrace{(1-\beta) \text{ OpenSubtitles}}_{\textit{target originating data}}$



Parallel Dataset = β EuroParl + $(1-\beta)$ OpenSubtitles

Target Mono Dataset = OpenSubtitles

A Real Case-Study: English-Burmese

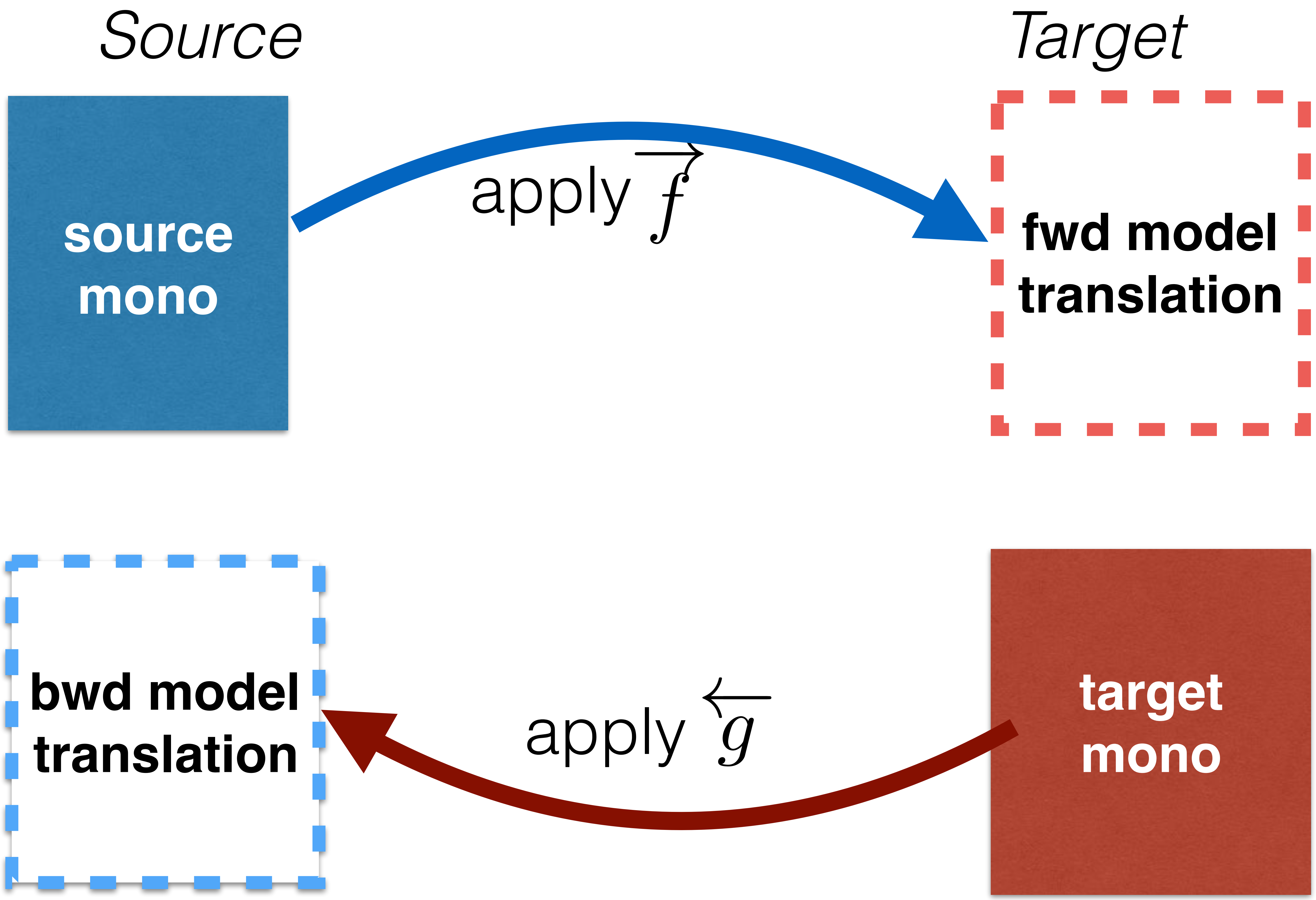


https://my.wikipedia.org/wiki/ဗဟိုစာမျက်နှာ

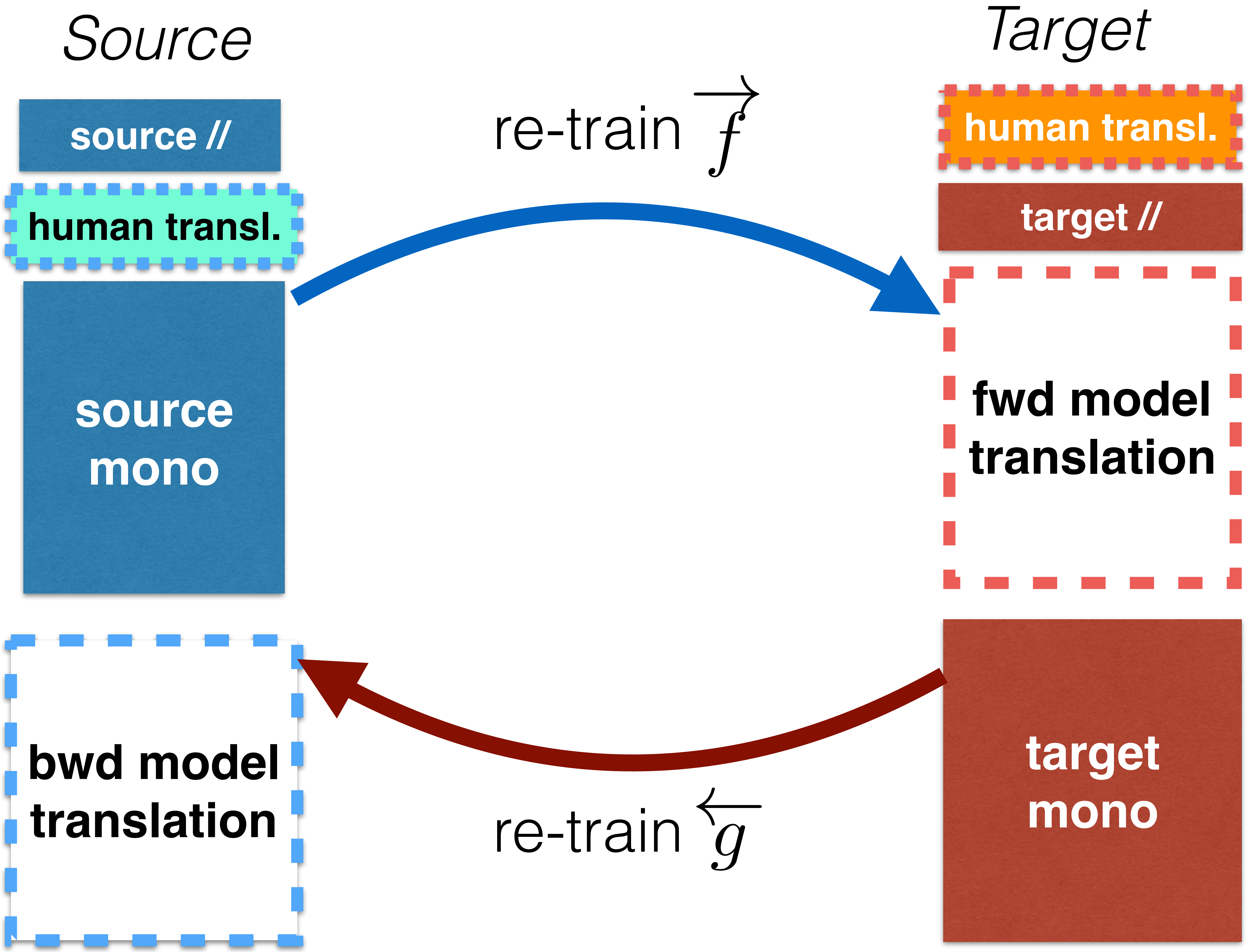
A Real Case-Study: English-Burmese

Workshop on Asian Translation @ EMNLP 2019: En-My

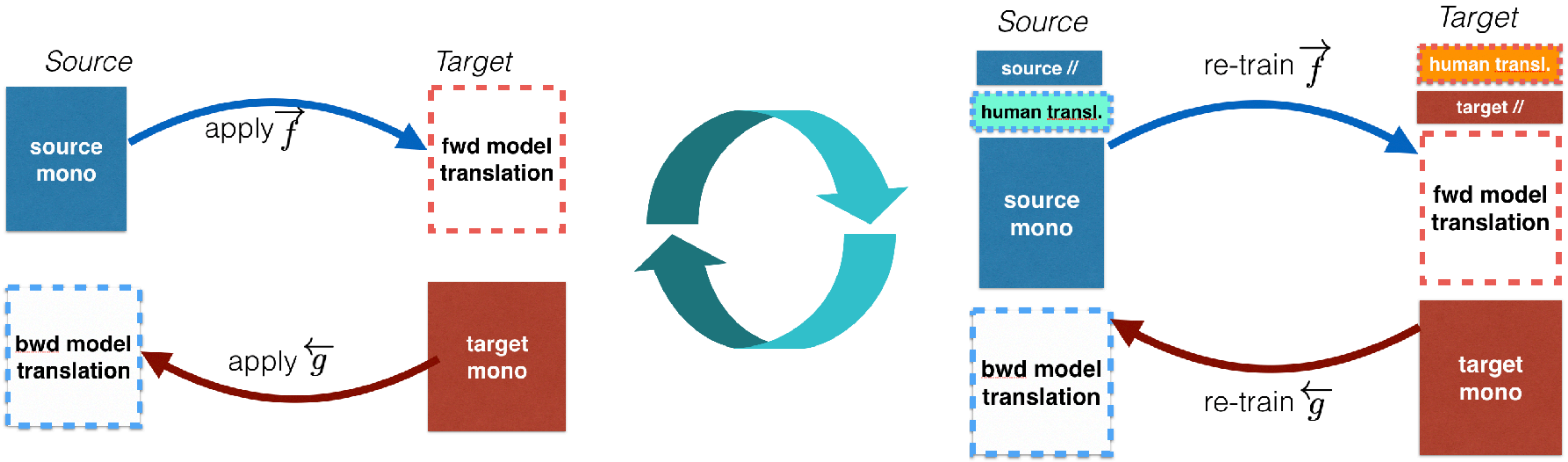
Iterative BT+ST



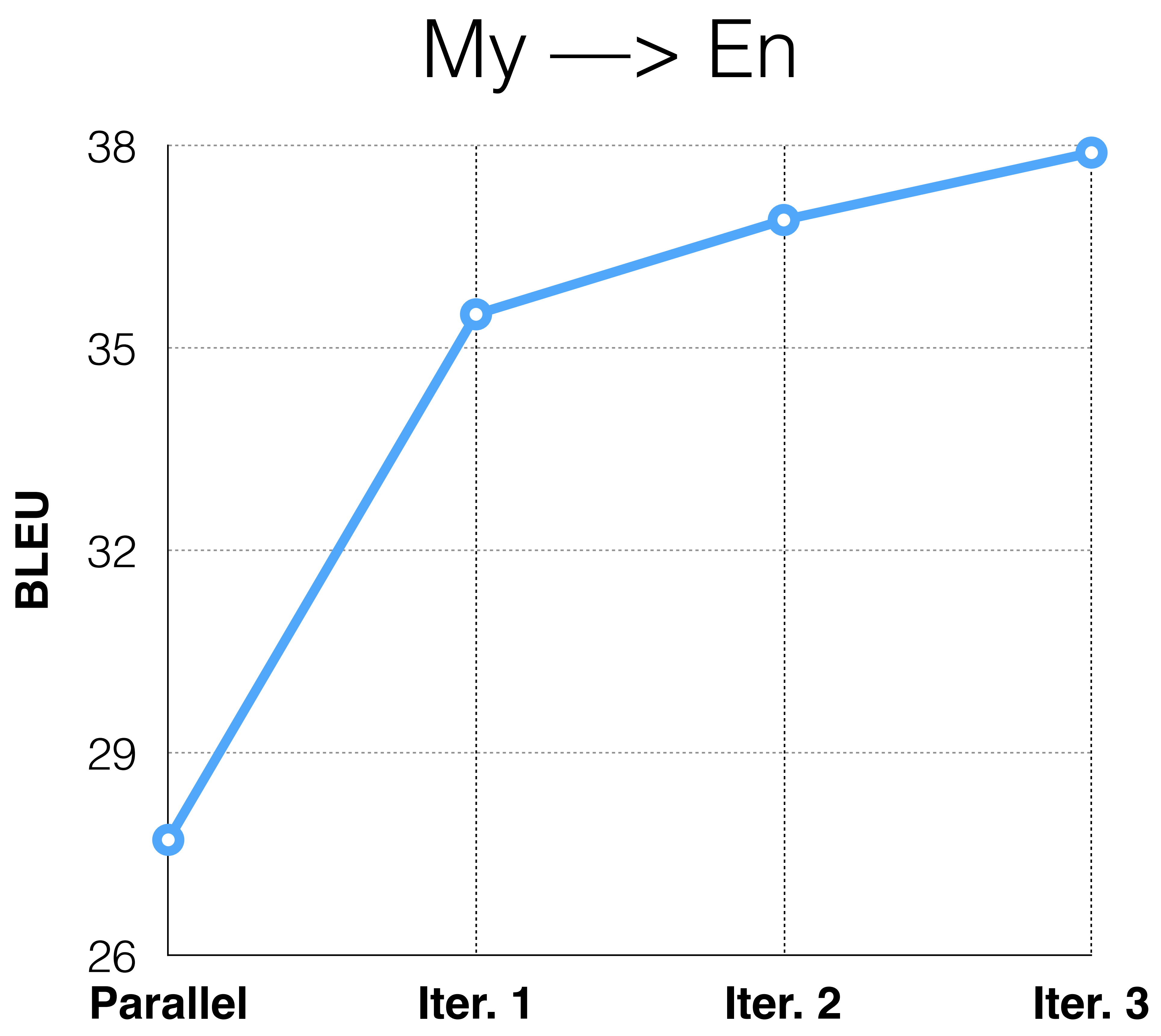
Iterative BT+ST

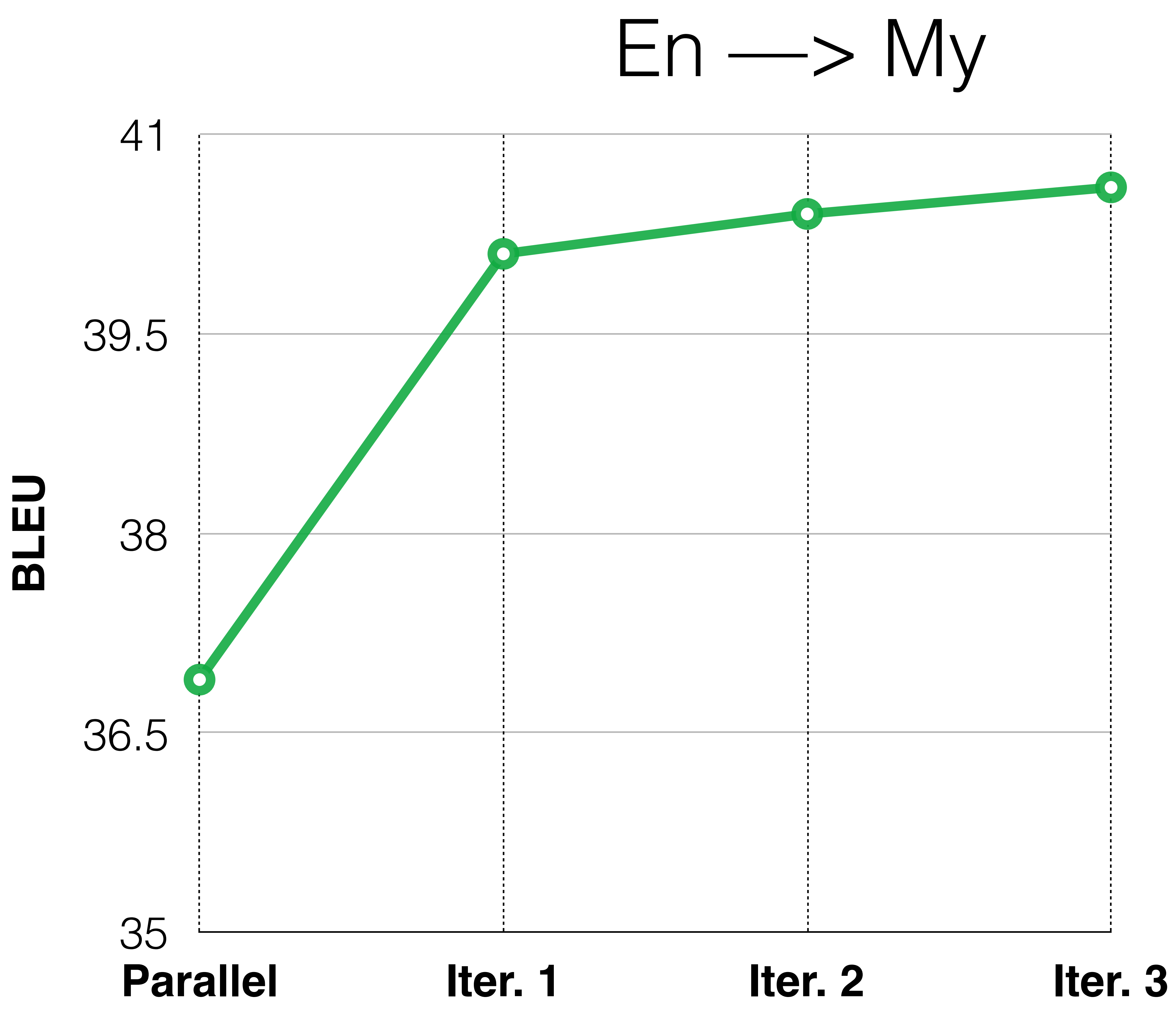


Iterative BT+ST

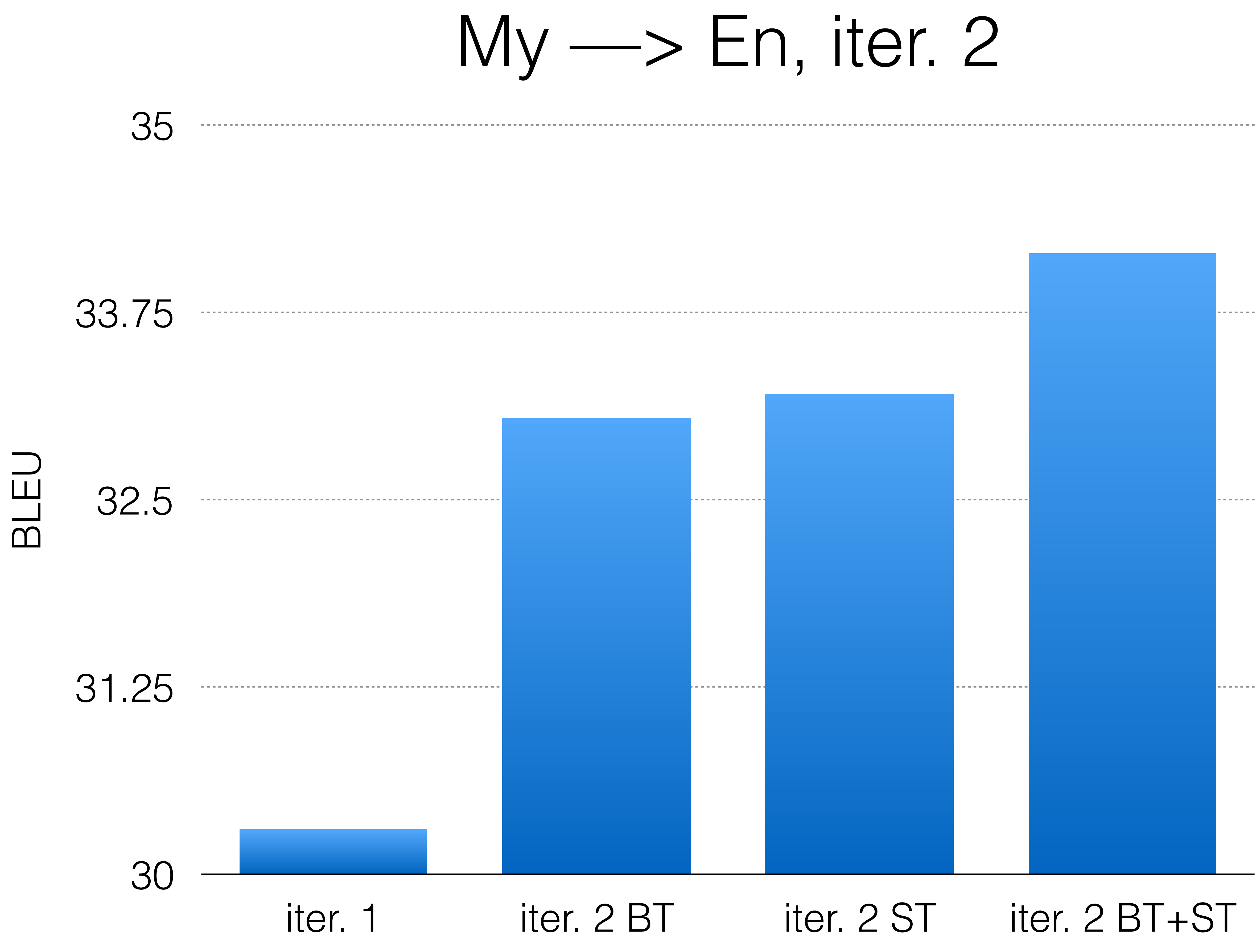


Results: Iterative ST+BT



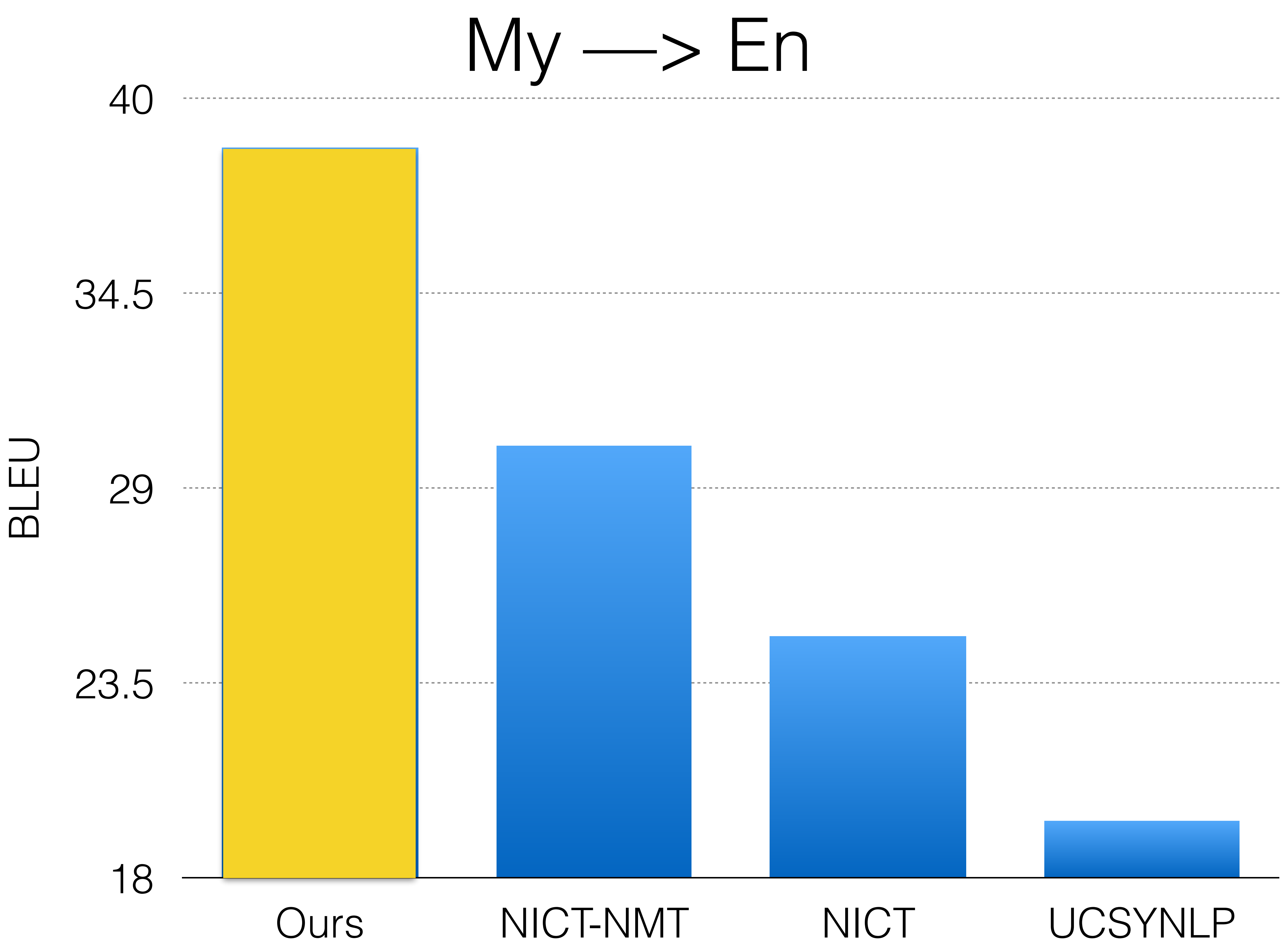


Results: BT vs ST vs BT+ST

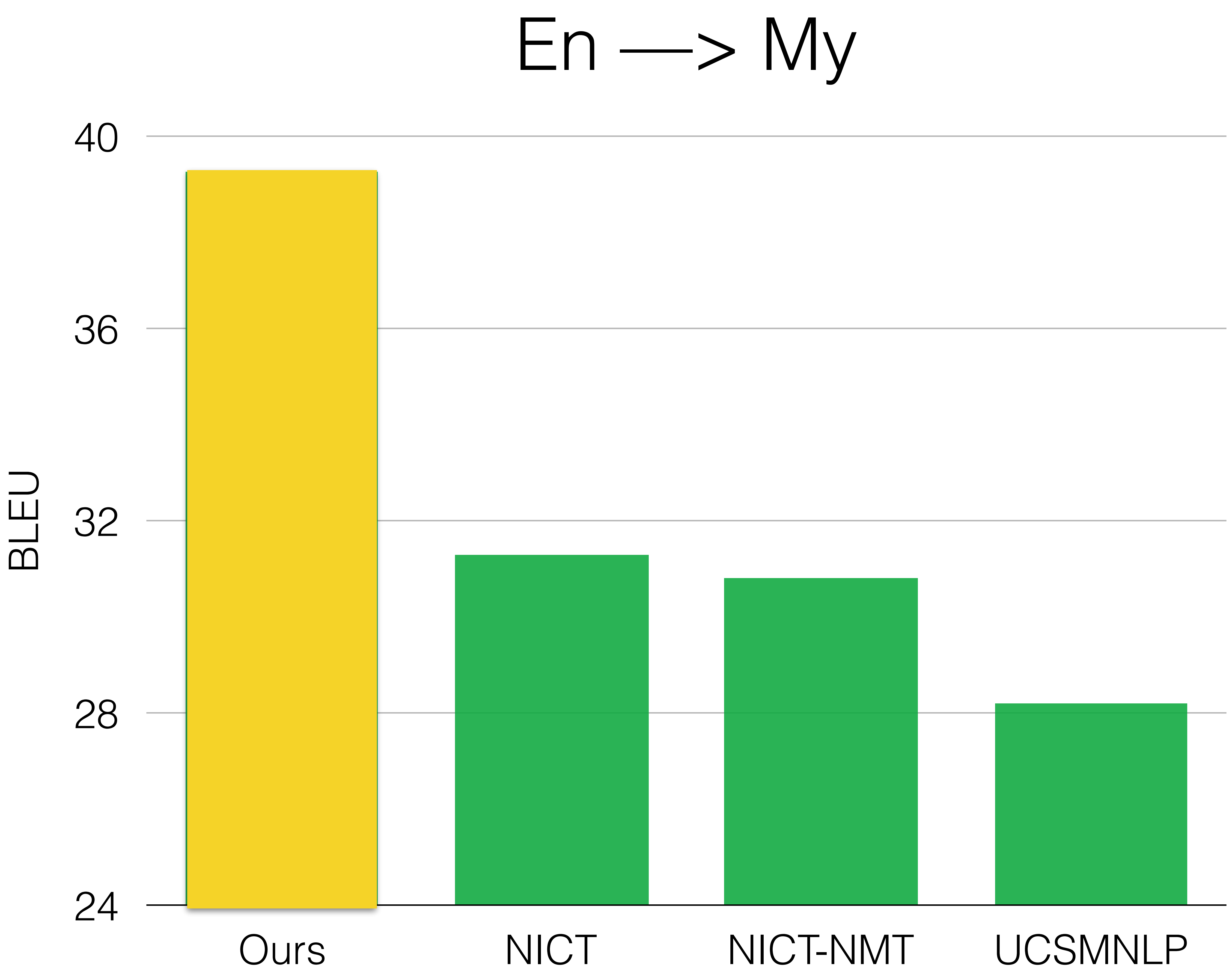


Final Results of 2019 Competition

+8 BLEU compared to second best



http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/list.php?t=70&o=4



http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/list.php?t=71&o=9

Conclusion

- The effect of **place** in MT is significant for low resource language pairs.
- Locality of topics is responsible for **source / target domain mismatch**. This decreases the effectiveness of back-translation.
- STDM can be easily simulated using public benchmarks.
- Self-training works well when there is little monolingual data on the target side and when there is extreme source / target domain mismatch.
- Self-training is complementary to back-translation. They can be combined in an iterative manner.

Open Research Questions

- Are there ways to measure STDM?
- What are good methods to cope with STDM?
- In general, how to adapt to the desired domain given little in-domain data (either parallel or just monolingual)?

THANK YOU