Low Resource Machine Translation

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

Machine Translation

English   **French**

Training data

Train NMT

NMT System

Ingredients:
- seq2seq with attention
- *SGD*

Test NMT

life is beautiful → **NMT System** → *la vie est belle*

*Ingredient:*
- beam
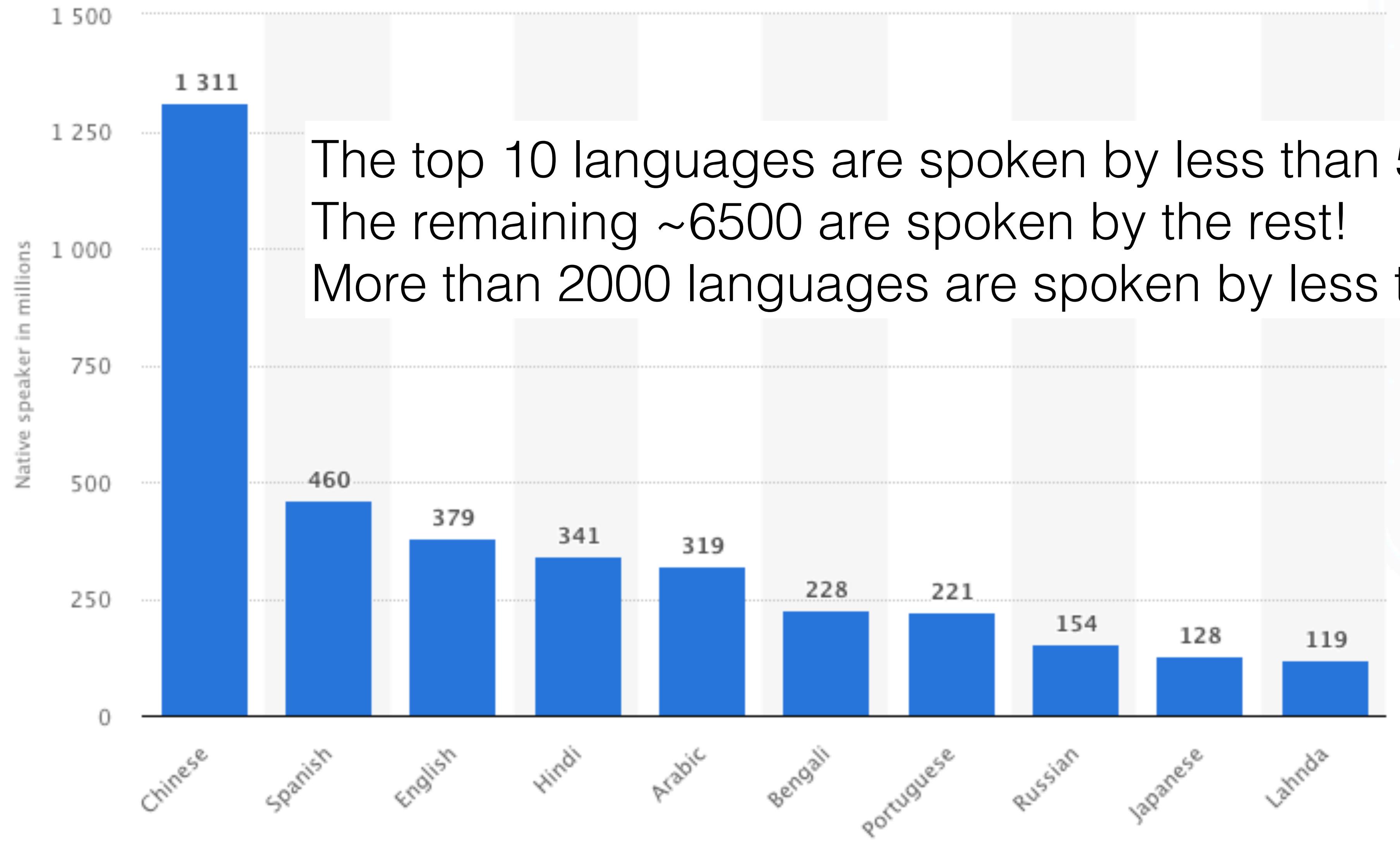
Some Stats

- 6000+ languages in the world
- 80% of the world population does not speak English
- Less than 5% of the people in the world are native English speakers.
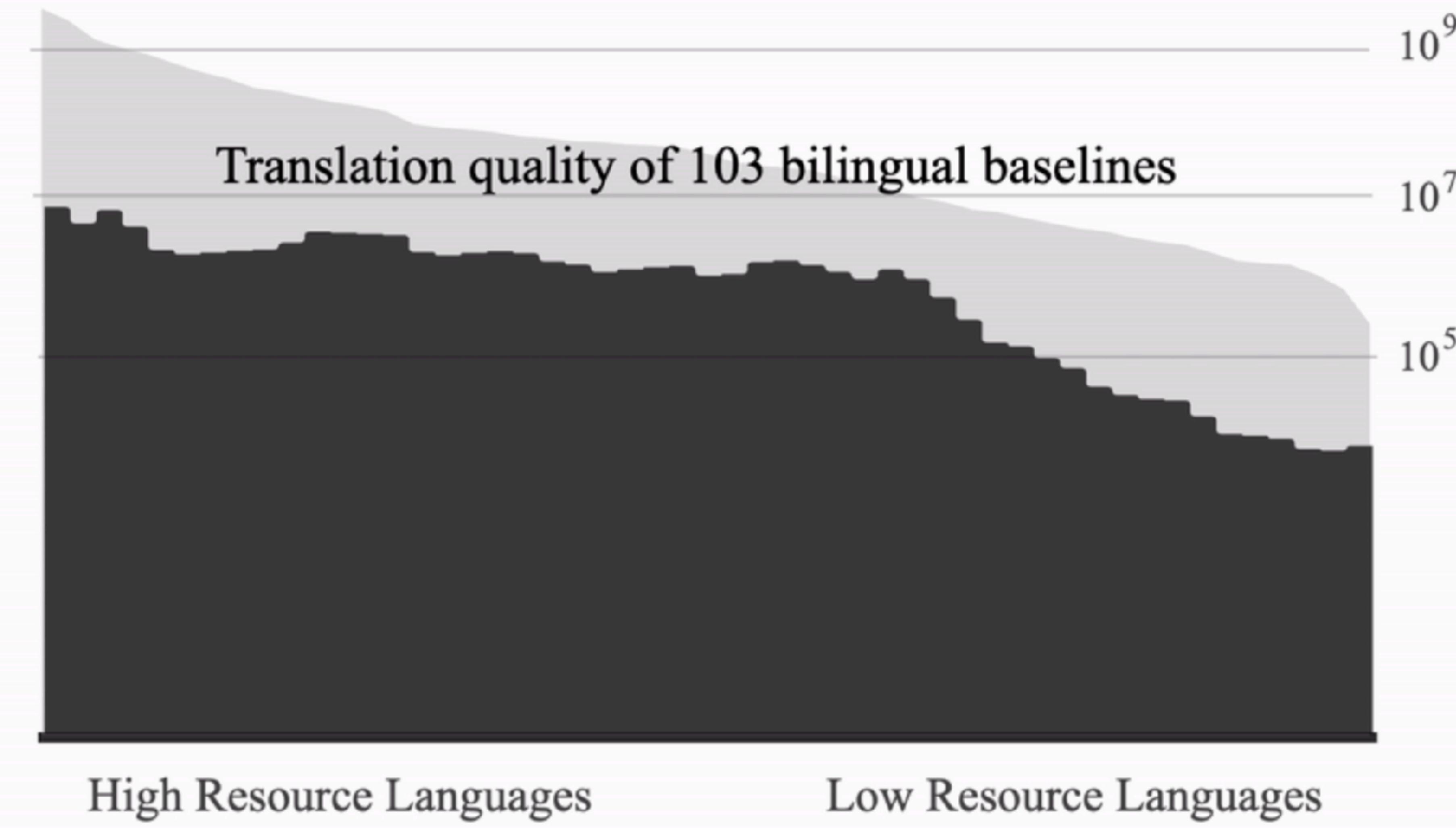
The Long Tail of Languages

The top 10 languages are spoken by less than 50% of the people.
The remaining ~6500 are spoken by the rest!
More than 2000 languages are spoken by less than 1000 people.

Native speaker in millions

Language	Speakers (millions)
Chinese	1 311
Spanish	460
English	379
Hindi	341
Arabic	319
Bengali	228
Portuguese	221
Russian	154
Japanese	128
Lahnda	119

source: https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/

Data distribution over language pairs (X to English)

Translation quality of 103 bilingual baselines

10^9

10^7

10^5

High Resource Languages

Low Resource Languages

source: https://ai.googleblog.com/2019/10/exploring-massively-multilingual.html

Machine Translation in Practice

English　　　　　　　　　　　**Nepali**
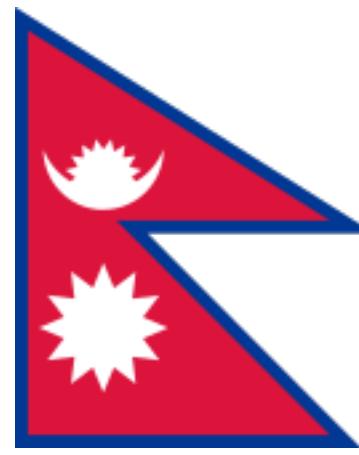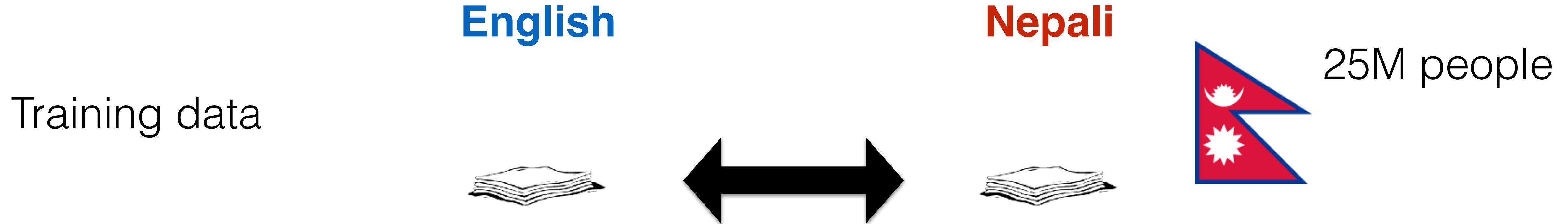
Training data

25M people

Machine Translation in Practice

English   **Nepali**

25M people

Training data

Parallel training data (collection of sentences with corresponding translation) is small!

Machine Translation in Practice

English   **Nepali**
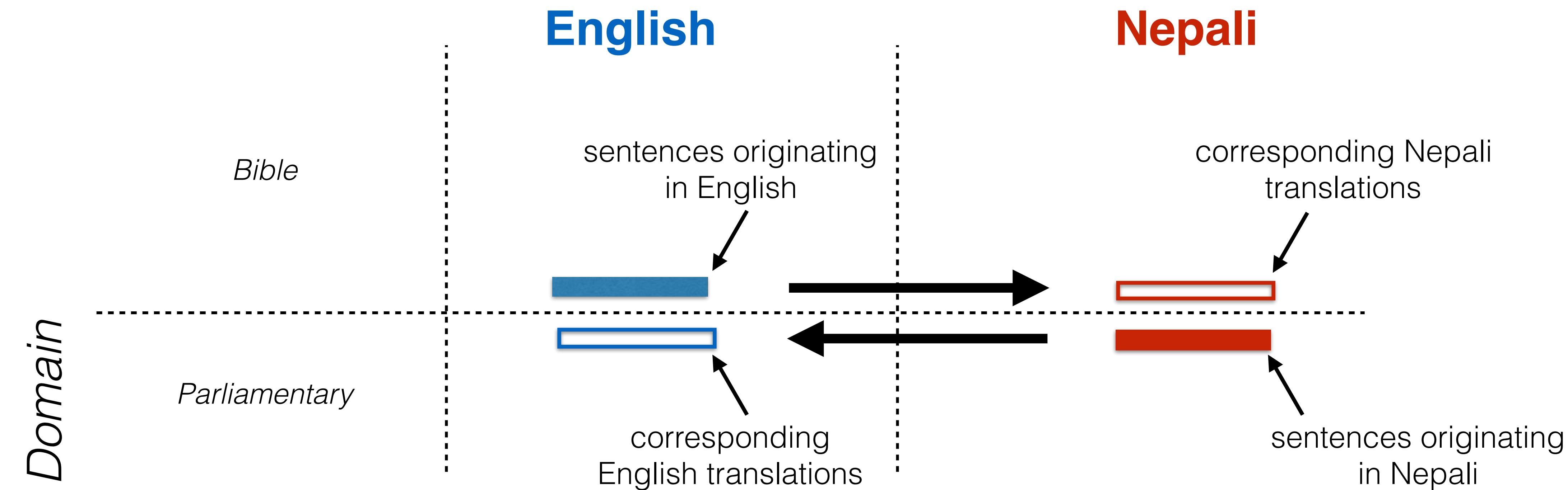
Training data

Let's represent data with rectangles. The color indicates the language.

Machine Translation in Practice

English   **Nepali**

Bible

sentences originating in English

corresponding Nepali translations

Domain

corresponding English translations

Parliamentary

sentences originating in Nepali
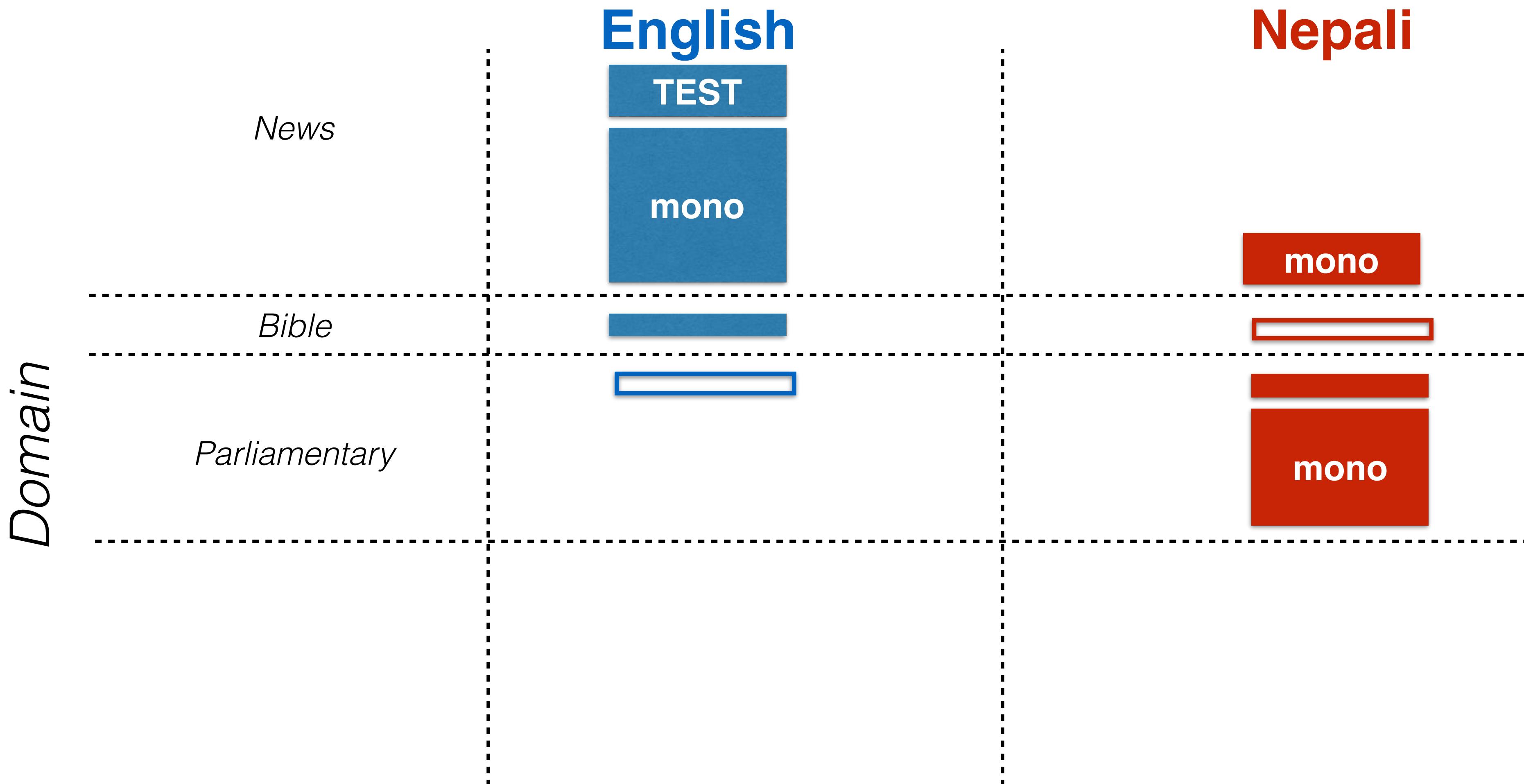
Let's represent (human) translations with empty rectangles.
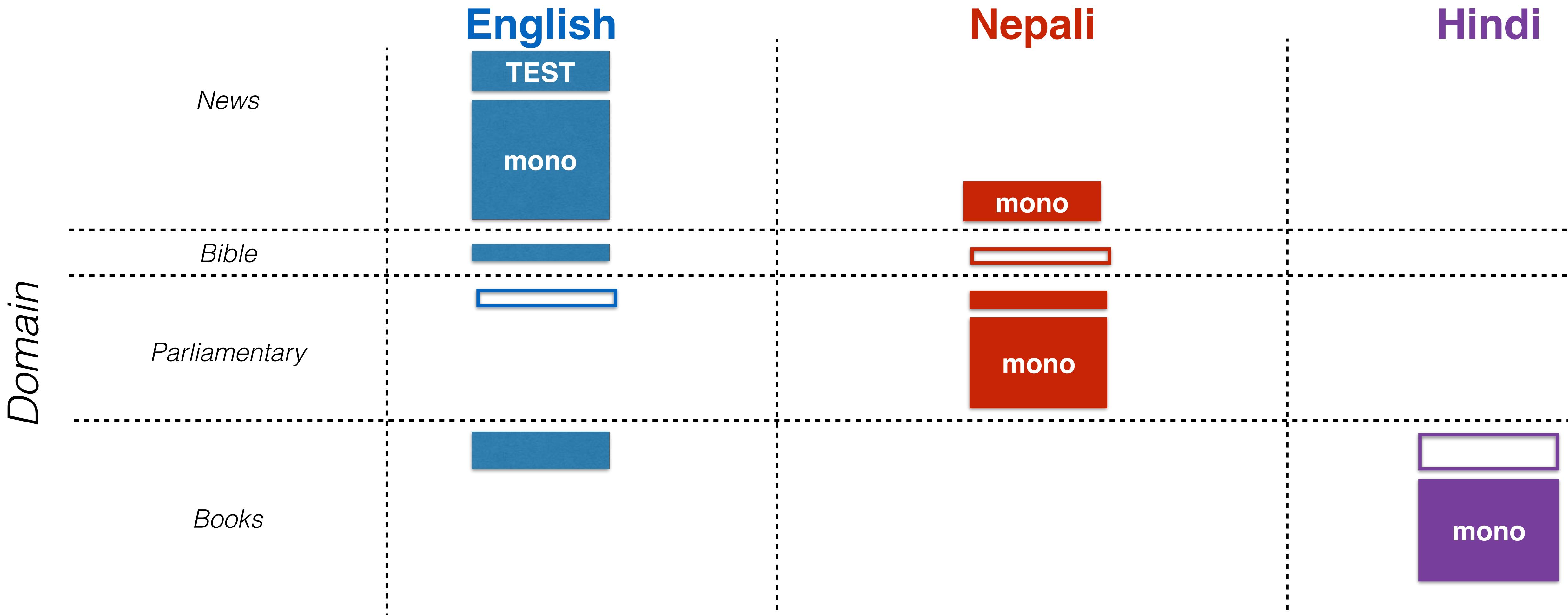
- Some parallel data originates in the source, some in the target language.
- Source and target domains may not match.

Machine Translation in Practice

English **Nepali**

Domain

News

Bible

Parliamentary

- Test data might be in another domain.
- There might exist source side in-domain monolingual data.

Machine Translation in Practice

English   **Nepali**   **Hindi**

News

TEST

mono

mono

Bible

Parliamentary

mono

mono

Books

mono

Domain

• There might be parallel and monolingual data with a high resource language close to the low resource language of interest. This data may belong to a different domain.

the *Mondrian* like learning setting! **...**

Low Resource Machine Translation

Loose definition: A language pair can be considered **low resource** when the number of parallel sentences is in the order of 10,000 or less.

Note: modern NMT systems have several hundred million parameters nowadays!
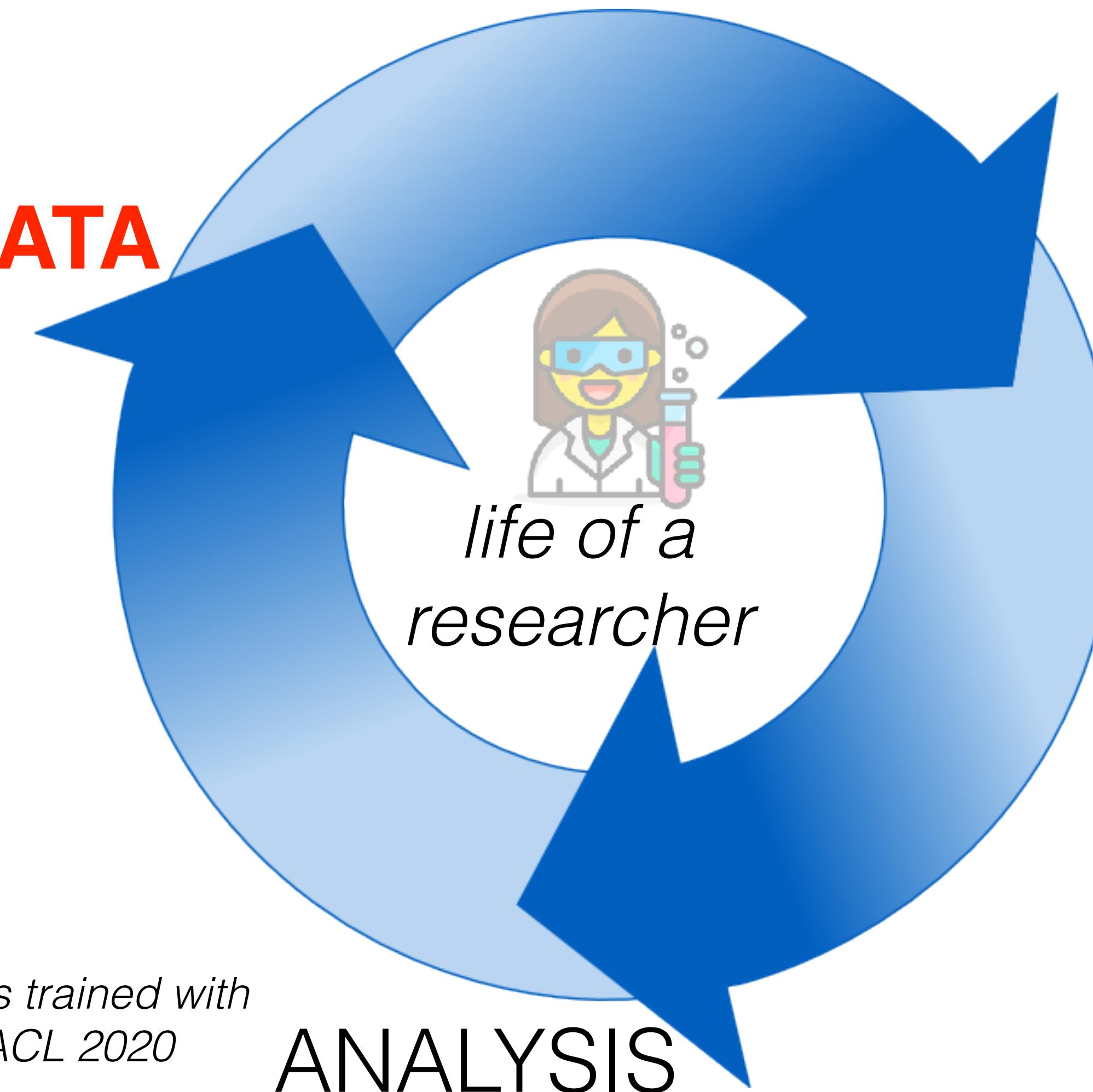
Challenges:
- data
    - sourcing data to train on
    - evaluation datasets
- modeling
    - unclear learning paradigm
    - domain adaptation
    - generalization

Why Low Resource MT Is Interesting?

- It is about learning with less labeled data.

- It is about modeling structured outputs and compositional learning.

- It is a real problem to solve.

Outline

DATA

*"The FLoRes evaluation for low resource MT:…" Guzmán, Chen et al. 'EMNLP 2019*

MODEL

"Phrase-based & Neural Unsup MT" Lample et al. EMNLP 2018

"FBAI WAT'19 My-En translation task submission" Chen et al., WAT@EMNLP 2019

"Investigating Multilingual NMT Representations at Scale" Kudugunta et al., EMNLP 2019

"Multilingual Denoising Pre-training for NMT" Liu et al., arXiv 2001:08210 2020

life of a researcher

"Analyzing uncertainty in NMT" Ott et al. ICML 2018

"On the evaluation of MT systems trained with back-translation" Edunov et al. ACL 2020

"The source-target domain mismatch problem in MT" Shen et al. arXiv 1909.13151 2019

ANALYSIS

15

A Big "Small-Data" Challenge

Oℝ**PUS** ... the open parallel corpus
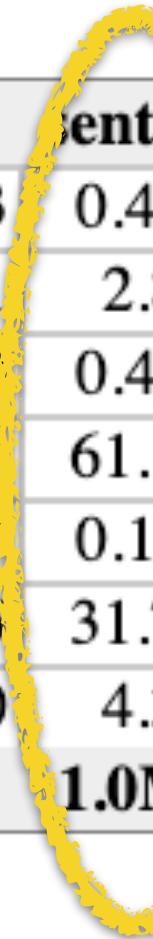
OPUS is a growing collection of translated texts from the web. In the OPUS project we try to convert and align free online data, to add linguistic annotation, a provide the community with a publicly available parallel corpus. OPUS is based on open source products and the corpus is also delivered as an open content package. We used several tools to compile the current collection. All pre-processing is done automatically. No manual corrections have been carried out.

The OPUS collection is growing! Check this page from time to time to see new data arriving ...
Contributions are very welcome! Please contact <jorg.tiedemann@helsinki.fi >
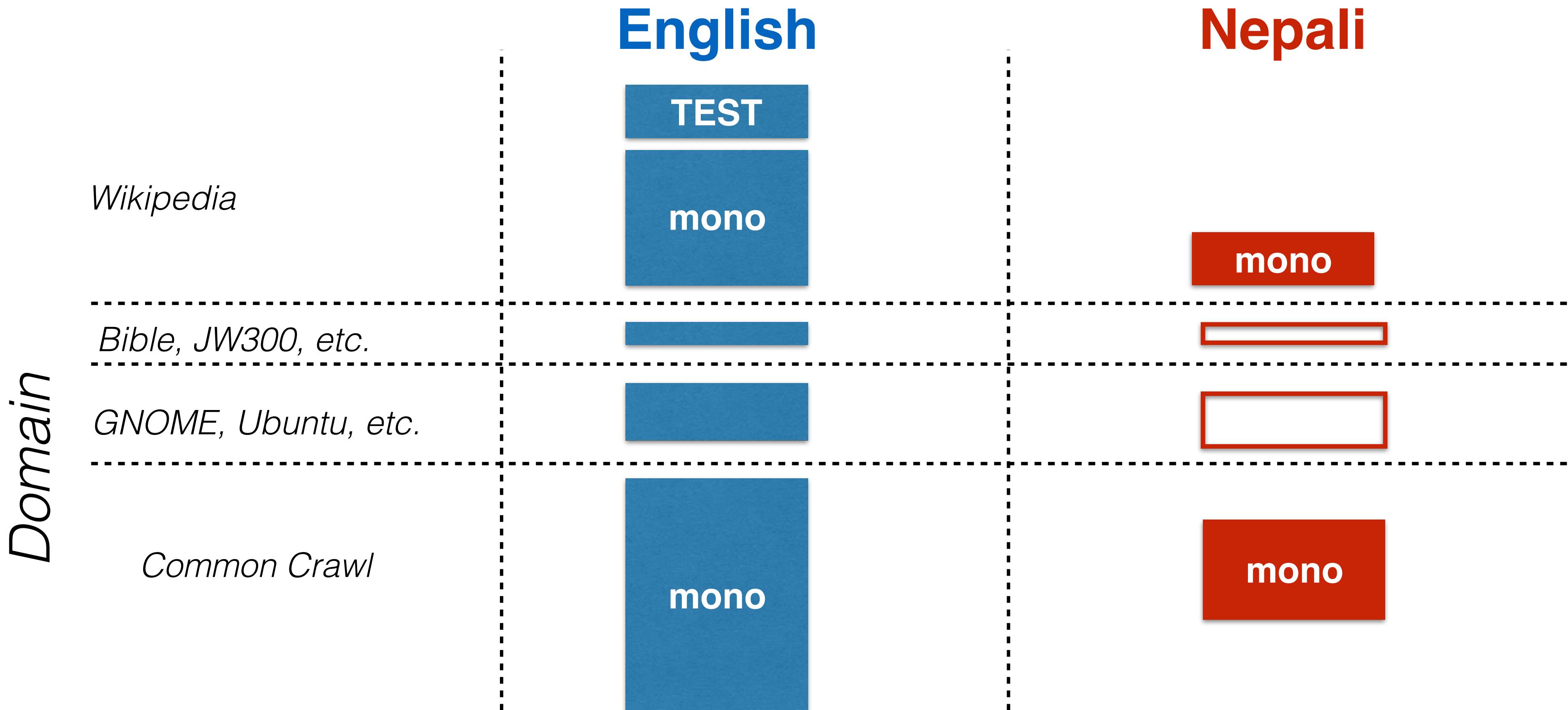
Search & download resources: [en (English) ▼] [ne (Nepali) ▼] [all ▼] ☐ show all versions

Language resources: click on [tmx | moses | xces | lang-id] to download the data! (raw = untokenized, ud = parsed with universal dependencies, alg = word alignments and phrase tables)

corpus	doc's	sent's	en tokens	ne tokens	XCES/XML	raw	TMX	Moses	mono	raw	ud	alg	dic	freq		other files
JW300 v1	4663	0.4M	6.5M	5.5M	xces en ne	en ne			en ne	en ne				en ne	sample	
wikimedia v20190628	1	2.8k	7.7M	1.2M	xces en ne	en ne	tmx	moses	en ne	en ne		alg smt	dic	en ne	sample	
GNOME v1	830	0.4M	1.8M	4.7M	xces en ne	en ne	tmx	moses	en ne	en ne		alg smt		en ne	sample	
bible-uedin v1	2	61.1k	1.8M	1.6M	xces en ne	en ne	tmx	moses	en ne	en ne		alg smt	dic	en ne	sample	
KDE4 v2	435	0.1M	0.6M	0.5M	xces en ne	en ne	tmx	moses	en ne	en ne		alg smt	dic	en ne	query	sample
Ubuntu v14.10	155	31.7k	0.3M	0.2M	xces en ne	en ne	tmx	moses	en ne	en ne		alg smt	dic	en ne	sample	
QED v2.0a	60	4.3k	69.8k	41.7k	xces en ne	en ne	tmx	moses	en ne	en ne		alg smt	dic	en ne	sample	
total	6146	1.0M	18.8M	13.7M	1.0M		0.6M	0.6M								

16

http://opus.nlpl.eu/

Case Study: En-Ne

English   **Nepali**

Wikipedia

Bible, JW300, etc.

GNOME, Ubuntu, etc.

Common Crawl

(English column, top to bottom: TEST, mono, thin bar, bar, mono)

(Nepali column: mono, thin outlined bar, outlined bar, mono)

In-domain data: no parallel, little monolingual.
Out-of-domain: little parallel, quite a bit monolingual
No translation originating from Nepali.

A Case Study: En-Ne

- Parallel Training data: versions of bible and ubuntu handbook (<1M sentences).

- Nepali Monolingual data: wikipedia (90K), common crawl (few millions).

- English Monolingual data: unlimited almost.

- Test data: ???

FLoRes Evaluation Benchmark

- Validation, test and hidden test set, each with 3000 sentences in English-Nepali and English-Sinhala.

- Sentences taken from *Wikipedia* documents.

Data Collection Process:

- Very expensive and slow.

- Very hard to produce high-quality translations:

    - automatic checks (language model filtering, transliteration filtering, length filtering, language id filtering, etc),

    - human assessment.

Guzmàn, Chen et al. "The FLoRes evaluation datasets for low resource MT…" EMNLP 2019

Examples

Si-En

original

අධ් යාපනයෙන් පසු හෝ පවුලේ යුතුකම් ඉටු කරන්නට හෝ රෝග තත්වයන් නිසා සංඝයා උපසම්පදාවෙන් නිතරම ඉවත් වෙති .

After education priests leave ordination in order to fulfill duties to the family or due to sickness.

translation

තර්ජන , ශාරීරික හිංසනය , දේපල හානිය , පහර දීම සහ මරාදැමීම මෙම දඬුවමිය .

Threatening, physical violence, property damage, assault and execution are these punishments.

En-Si

In Serious meets, the absolute score is somewhat meaningless.

🤔 **Wikipedia originating in Si has different topics than Wikipedia originating in En**

සැබෑ තරඟ වලදී ලකුණු සැසඳීම තේරුමක් තැති ක්‍රියාවකි .

Iphone users can and do access the internet frequently, and in a variety of places.

අයිෆෝන් භාවිත කරන්නට නිතරම සහ විවිධ ස්ථානවලදී අන්තර්ජාලයට පිවිසිය හැකිය .

Guzmàn, Chen et al. "The FLoRes evaluation datasets for low resource MT…" EMNLP 2019

Examples

Ne-En

पुरानो समयमा राजालाई सल्लाह दिने सभा ' संसद ' कहलाउँथ्यो ।

In the past, the assembly that advised the king were called 'parliament'.

कार्यकर्ताका रुपमा अफ्रिकन नेशनल कंग्रेसमा आबद्ध भए ।

As a worker African Mandela joined the Congress party.

En-Ne

The academic research tended toward the improvement of basic technologies, rather than their specific applications.

शैक्षिक अनुसन्धानले उनीहरूको विशिष्ट अनुप्रयोगहरूको सट्टा आधारभूत प्रविधिको सुधारको पक्षमा जोड दिए ।

It has automatic spell checking and correction, predictive word capabilities, and a dynamic dictionary that learns new words.

यसमा स्वचालित हिज्जे जाँच र सुधार छ , भविष्यवाणी शब्द क्षमताहरू , र गतिशील शब्दकोश हुन्छ जसले नयाँ शब्दहरू सिक्छ ।

Guzmàn, Chen et al. "The FLoRes evaluation datasets for low resource MT…" EMNLP 2019

FLoRes

- Useful to evaluate truly low resource language pairs.

  - WMT 2019 and WMT 2020 shared filtering task.

  - Several publications.

- Sustained effort, more to come…

FLoRes Low Resource MT Benchmark

This repository contains data and baselines from the paper:
The FLoRes Evaluation Datasets for Low-Resource Machine Translation: Nepali-English and Sinhala-English.

The data can be downloaded directly at:
https://github.com/facebookresearch/flores/raw/master/data/wikipedia_en_ne_si_test_sets.tgz

Baselines

The following instructions will can be used to reproduce the baseline results from the paper.

Requirements

The baseline uses the Indic NLP Library and sentencepiece for preprocessing; fairseq for model training; and sacrebleu for scoring.

Dependencies can be installed via pip:

```
$ pip install fairseq sacrebleu sentencepiece
```

The Indic NLP Library will be cloned automatically by the prepare-{ne,si}en.sh scripts.

Download and preprocess data

https://github.com/facebookresearch/flores
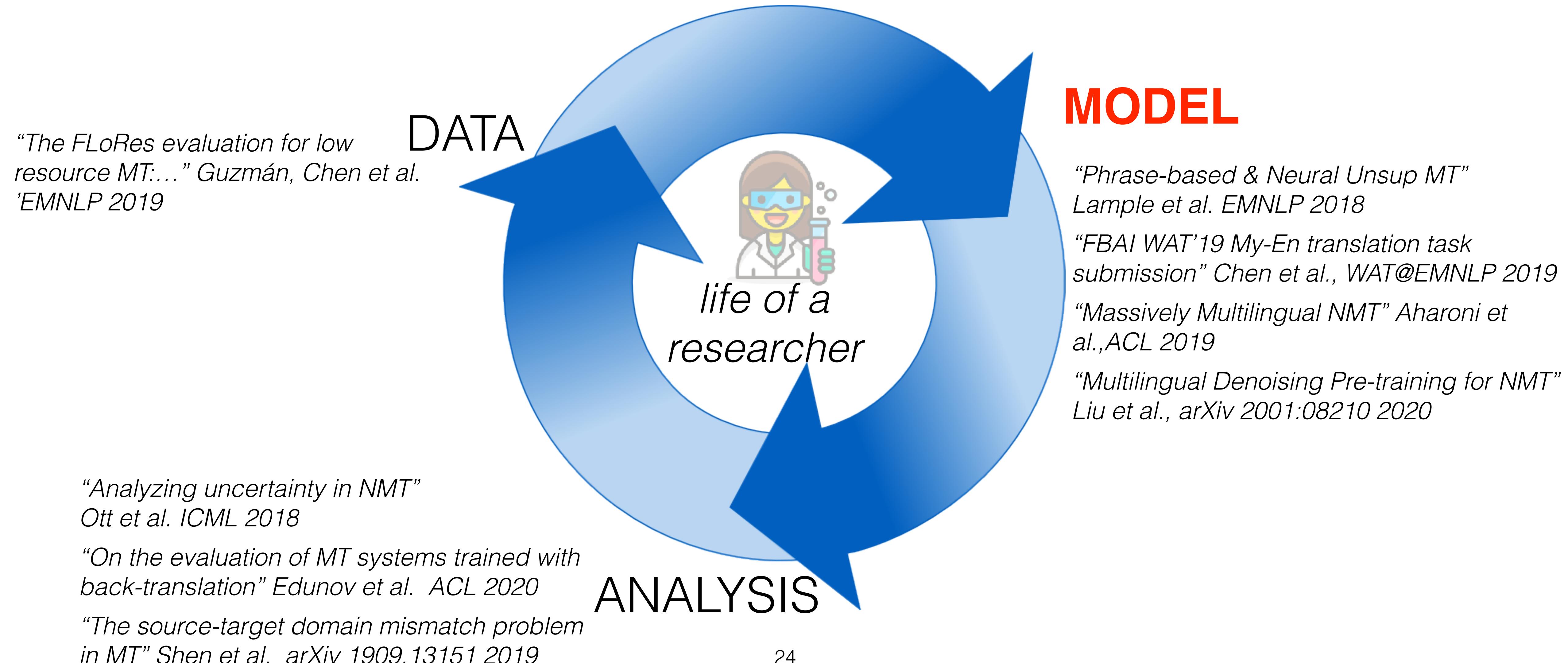
data & baseline models

What Did We Learn?

- Data is often as or more important than designing a model.

- Collecting data is not trivial.

- Look at the data!!

Outline

DATA

"The FLoRes evaluation for low resource MT:…" Guzmán, Chen et al. 'EMNLP 2019

MODEL

"Phrase-based & Neural Unsup MT" Lample et al. EMNLP 2018

"FBAI WAT'19 My-En translation task submission" Chen et al., WAT@EMNLP 2019

"Massively Multilingual NMT" Aharoni et al.,ACL 2019

"Multilingual Denoising Pre-training for NMT" Liu et al., arXiv 2001:08210 2020

life of a researcher

"Analyzing uncertainty in NMT" Ott et al. ICML 2018

"On the evaluation of MT systems trained with back-translation" Edunov et al. ACL 2020

"The source-target domain mismatch problem in MT" Shen et al. arXiv 1909.13151 2019

ANALYSIS

ML Perspective: Supervised Learning

DATA

English

TEST

TRAIN //

Nepali

en NMT system ne

x
input sentence

Encoder → **Decoder** → *prediction* → **Cross-Entropy Loss**

y
human reference

human translator

Training Dataset

$$\mathcal{D} = \{(x, y)_i\}_{i=1,..,N}$$

If N is small, how can we further regularize the model?
- dropout [1]
- label smoothing [2]

Learning Framework: Supervised Learning.

Per-sample loss:
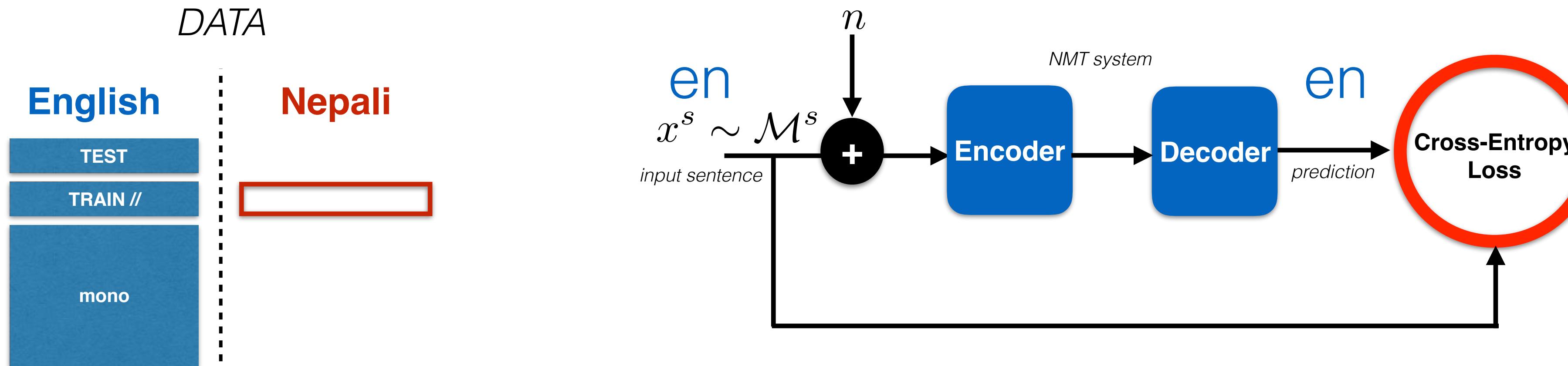
$$\mathcal{L}(\theta) = -\log p(y|x)$$

usual attention-based transformer

[1] Srivastava et al. "Dropout: a simple way to prevent neural networks from overfitting" JMLR 2014

[2] Szegedy et al. "Rethinking the inception architecture for computer vision" CVPR 2016

ML Perspective: Semi-Supervised Learning

DATA

English   **Nepali**

TEST

TRAIN //

mono

$$n$$

NMT system

en $x^s \sim \mathcal{M}^s$

input sentence   **+**   **Encoder** → **Decoder** → *prediction*   en   **Cross-Entropy Loss**

Training Dataset

$$\mathcal{D} = \{(x, y)_i\}_{i=1,..,N}$$
$$\mathcal{M}^s = \{x_j^s\}_{j=1,..,M_s}$$

Adding source-side monolingual data.
Idea: model p(x).

Noise: word drop, swap, etc.

E.g.: The cat the on sat mat.
The cat sat on the.
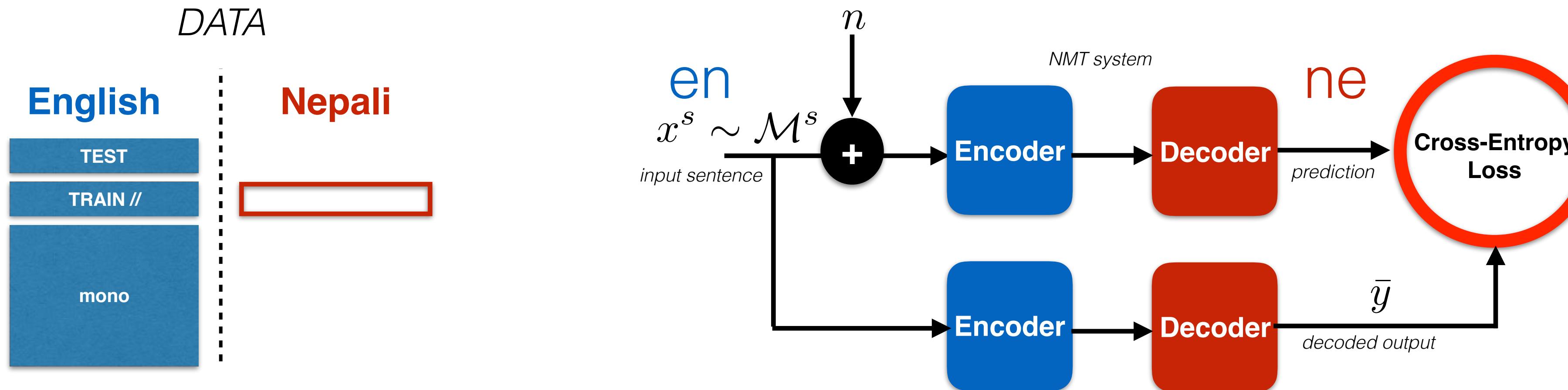
Learning Framework: DAE

Either pre-train or add a DAE loss to the supervised cross-entropy term.

$$\mathcal{L}^{DAE}(\theta) = -\log p(x|x+n)$$

good region

no noise   *noise only*

Vincent et al. "Stacked denoising auto-encoders:…" JMLR 2010
Liu et al. "Multilingual denoising pretraining for NMT" arXiv:2001.08210 2020

ML Perspective: Semi-Supervised Learning

DATA

English | **Nepali**

TEST

TRAIN //

mono

en $x^s \sim \mathcal{M}^s$ — NMT system — ne

n

input sentence — + — Encoder — Decoder — prediction — Cross-Entropy Loss

Encoder — Decoder — $\bar{y}$ — decoded output

Training Dataset

$$\mathcal{D} = \{(x,y)_i\}_{i=1,..,N}$$

$$\mathcal{M}^s = \{x^s_j\}_{j=1,..,M_s}$$

Adding source-side monolingual data. An alternative approach to DAE.

$$\mathcal{L}^{ST}(\theta) = -\log p(\bar{y}|x+n)$$

$$\mathcal{L}(\theta) = \mathcal{L}^{\sup}(\theta) + \lambda\mathcal{L}^{\mathrm{ST}}(\theta)$$
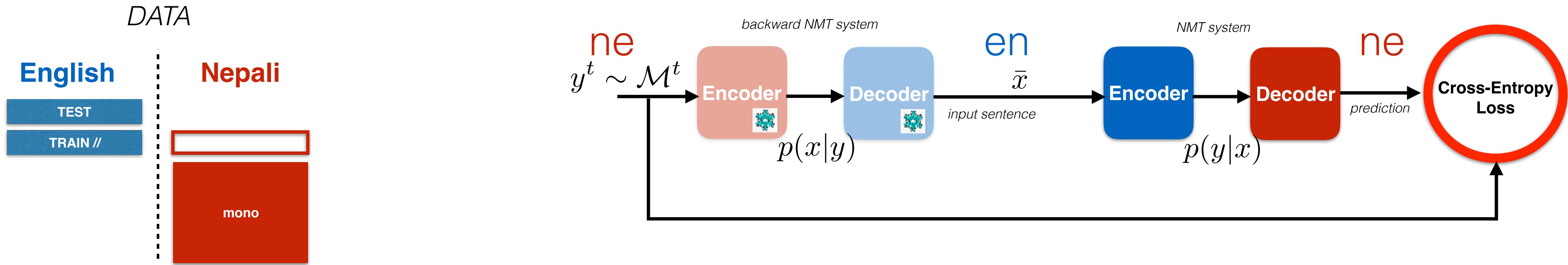
Learning Framework: Self-Training (ST).

ALGORITHM

- train model $p(y|x)$ on $\mathcal{D}$
- repeat
  - decode $x^s \sim \mathcal{M}^s$ to $\bar{y}$ and create additional dataset $\mathcal{A}^s = \{(x^s_j, \bar{y}_j)\}_{j=1,..,M_s}$
  - retrain model on: $\mathcal{D} \cup \mathcal{A}^s$

Key elements: decoding and training noise.

He et al. "Revisiting self-training for neural sequence generation" ICLR 2020

ML Perspective: Semi-Supervised Learning

DATA

Training Dataset

$$\mathcal{D} = \{(x, y)_i\}_{i=1,..,N}$$
$$\mathcal{M}^t = \{y_k^t\}_{k=1,..,M_t}$$

Adding target-side monolingual data.
Two benefits:
a) Decoder learns a good language model.
b) Better generalization via data augmentation.
c) Unlike ST, target is correct but input is not.
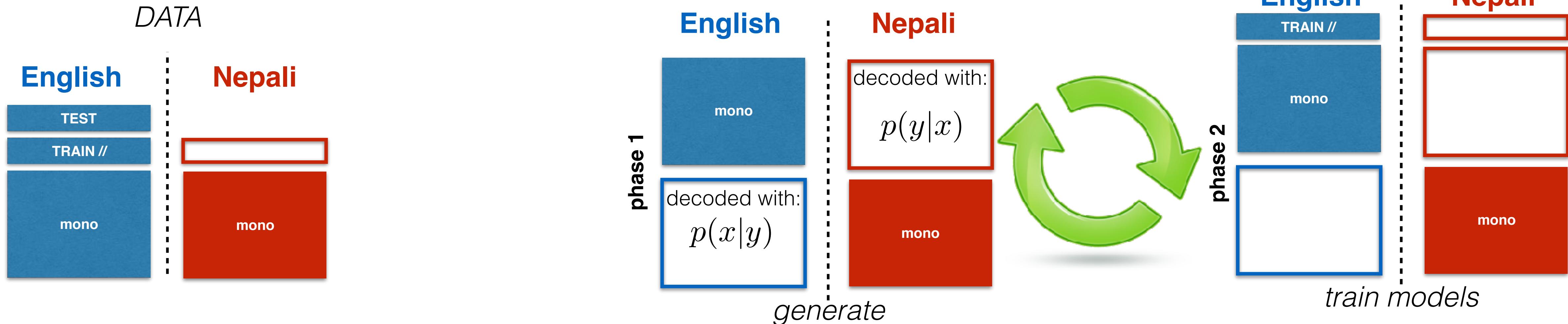
$$\mathcal{L}^{BT}(\theta) = -\log p(y|\bar{x})$$
$$\mathcal{L}(\theta) = \mathcal{L}^{\sup}(\theta) + \lambda \mathcal{L}^{\mathrm{BT}}(\theta)$$

Learning Framework: Back-Translation (BT).

ALGORITHM

- train model $p(x|y)$ and $p(y|x)$ on $\mathcal{D}$
- decode $y^t \sim \mathcal{M}^t$ to $\bar{x}$ with $p(x|y)$, create additional dataset $\mathcal{A}^t = \{(\bar{x}_k, y_k^t)\}_{k=1,..,M_t}$
- retrain model $p(y|x)$ on: $\mathcal{D} \cup \mathcal{A}^t$

Sennrich et al. "Improving NMT models with monolingual data" ACL 2016

ML Perspective: Semi-Supervised Learning

DATA

English   **Nepali**

| TEST |
| TRAIN // |
| mono |

| |
| mono |

Training Dataset

$$\mathcal{D} = \{(x, y)_i\}_{i=1,..,N}$$
$$\mathcal{M}^t = \{y_k^t\}_{k=1,..,M_t}$$
$$\mathcal{M}^s = \{x_j^s\}_{j=1,..,M_s}$$

Adding both source & target-side monolingual data.
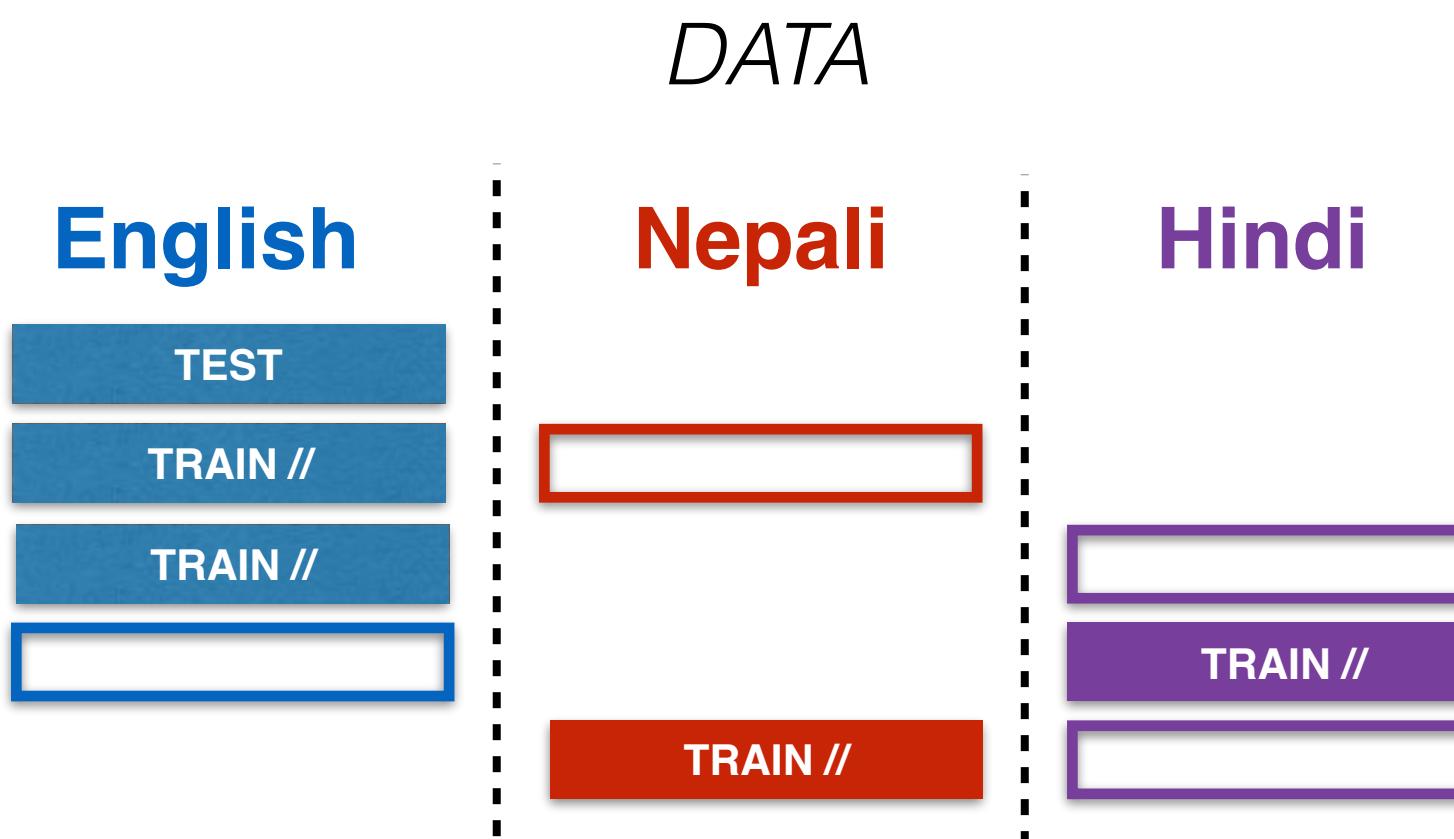
$$\mathcal{L}^{\text{total}}(\theta) = -\log p(y|x) - \lambda_1 \log p(y^t|\bar{x}^t) - \lambda_2 \log p(\bar{y}^s|x^s)$$

English   **Nepali**

phase 1

| mono |

| decoded with:
 $p(x|y)$ |

| decoded with:
 $p(y|x)$ |

| mono |

generate

English   **Nepali**

phase 2

| TRAIN // |
| mono |

| |

| |
| mono |

train models

Learning Framework: Iterative ST & BT.

ALGORITHM

- train model $p(x|y)$ and $p(y|x)$ on $\mathcal{D}$
- repeat
  - decode $y^t \sim \mathcal{M}^t$ to $\bar{x}$ with $p(x|y)$, create additional dataset $\mathcal{A}^t = \{(\bar{x}_k, y_k^t)\}_{k=1,..,M_t}$
  - decode $x^s \sim \mathcal{M}^s$ to $\bar{y}$ with $p(y|x)$, create additional dataset $\mathcal{A}^s = \{(x_j^s, \bar{y}_j)\}_{j=1,..,M_s}$
  - retrain both $p(y|x)$ and $p(x|y)$ on: $\mathcal{D} \cup \mathcal{A}^t \cup \mathcal{A}^s$

phase 1 / **phase 2**

Shen et al. "The source-target domain mismatch problem in MT" arXiv:1909.13151 2019
Chen et al. "FBAI WAT'19 Myanmar-English translation task submission" WAT@EMNLP 2019

ML Perspective: Multi-Task/Multi-Modal
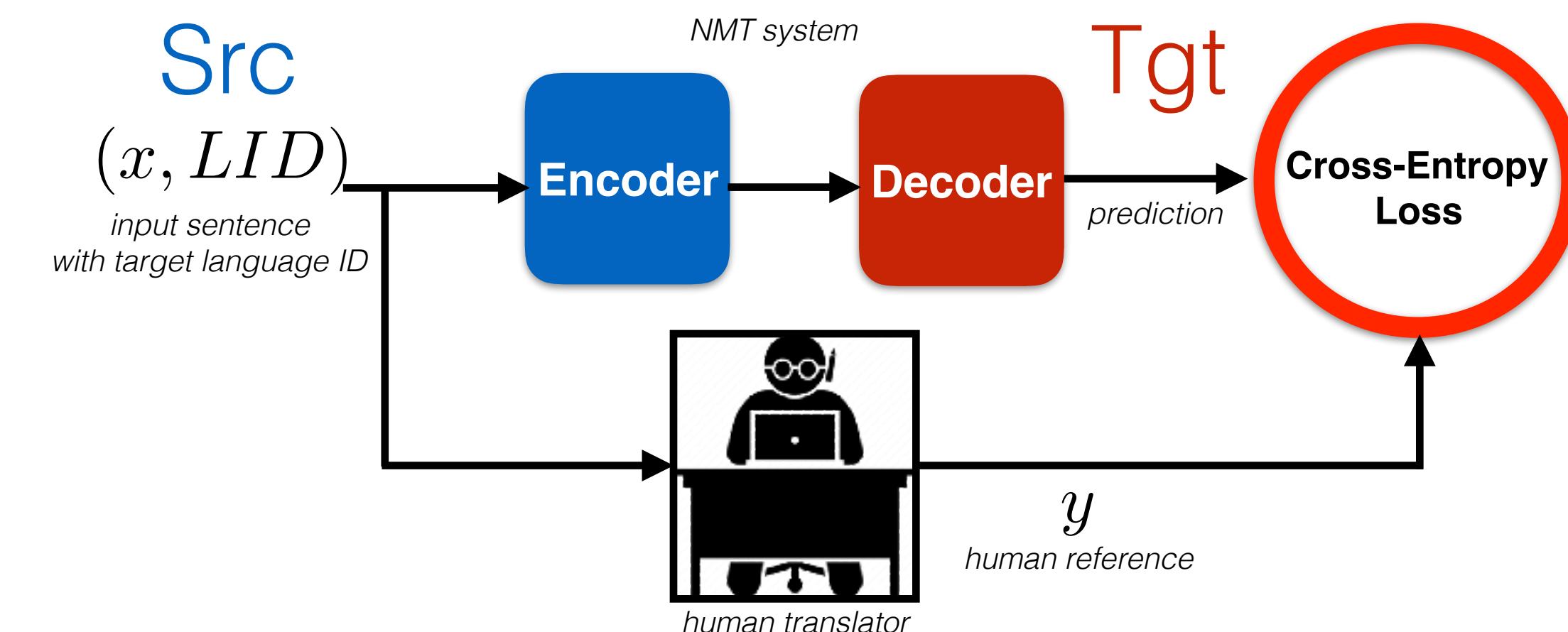
DATA

Training Dataset

$$\mathcal{D}_{\text{en,ne}} \cup \mathcal{D}_{\text{en,hi}} \cup \mathcal{D}_{\text{hi,en}} \cup \mathcal{D}_{\text{ne,hi}}$$

Adding parallel data in other languages.

$$\mathcal{L}(\theta) = -\sum_{s,t} \mathbb{E}_{(x,y)\sim\mathcal{D}_{s,t}}[\log p(y|x;t)]$$

Learning Framework: Multilingual Training

Share the same encoder and the same decoder with all the language pairs.
Prepend a target language identifier to the source sentence to inform decoder of desired language.
Concatenate all the datasets together.
Train using standard cross-entropy loss.
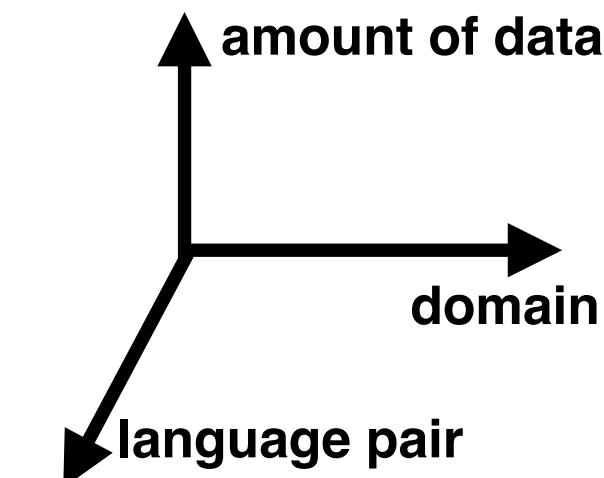
Johnson et al. "Google's multilingual NMT system…" ACL 2017
Aharoni et al. "Massively multilingual NMT" ACL 2019

Conclusion so far…
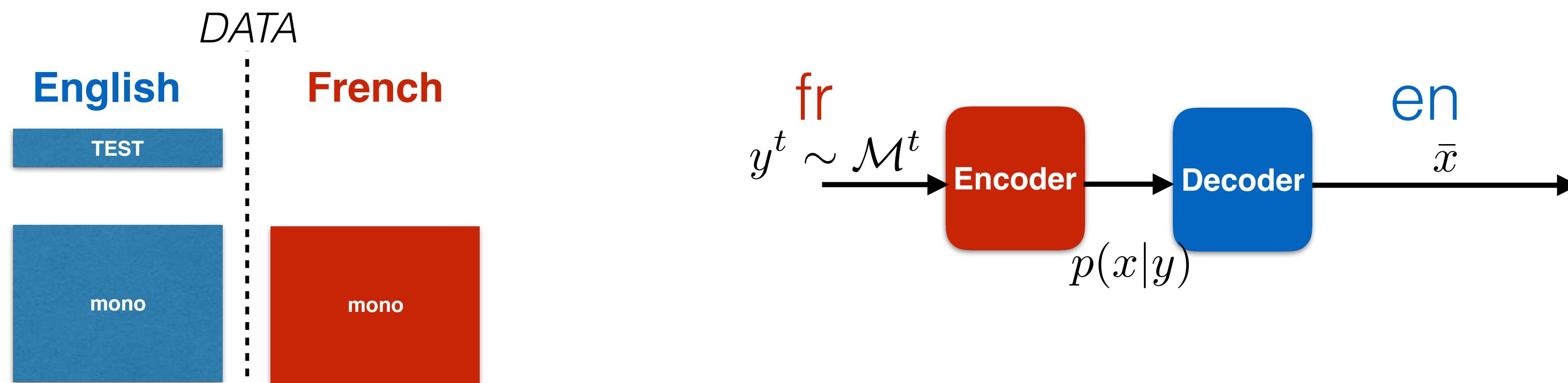
- Assuming no domain effect, there are lots of training paradigms depending on the available data.

- It is hard to predict what works best.

- In general, DAE pretraining, (iterative) BT and multi-lingual training perform strongly on low resource languages.

- All these methods can be combined together, but it requires some level of craftsmanship…

- Final touch: ensembling, fine-tuning, distillation, etc.

Open Challenges

- Diversity of domains and varying translation quality.

- Wildly varying dataset sizes.

- Diversity of language pairs.

- Training large models to handle large datasets, as soon as we put together lots of language pairs and monolingual data.
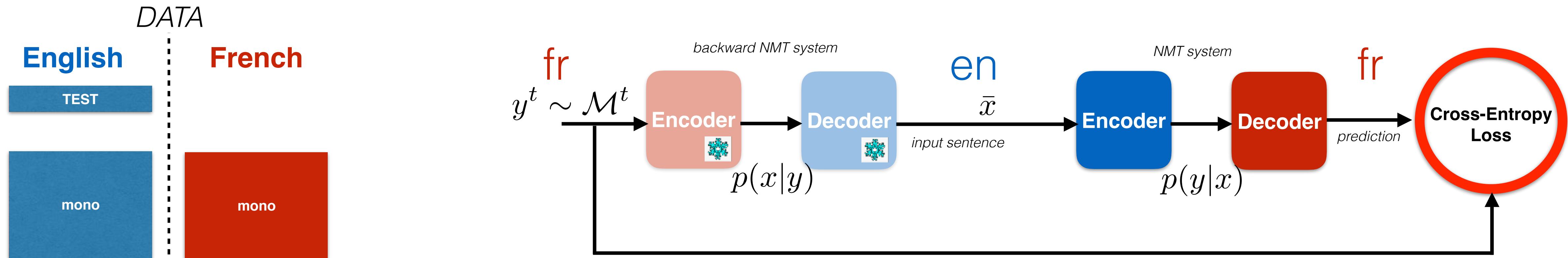
Case Study #1: Unsupervised MT

DATA

English | **French**

TEST

mono | mono

$$\mathcal{M}^t = \{y_k^t\}_{k=1,..,M_t}$$
$$\mathcal{M}^s = \{x_j^s\}_{j=1,..,M_s}$$

fr

en

$$y^t \sim \mathcal{M}^t \longrightarrow \boxed{\text{Encoder}} \longrightarrow \boxed{\text{Decoder}} \xrightarrow{\bar{x}}$$

$$p(x|y)$$

Lample et al. "Phrase-based and neural unsupervised MT" EMNLP 2018
*Artetxe et al. "An effective approach to unsupervised MT" ACL 2019*

Case Study #1: Unsupervised MT

DATA

English | **French**

TEST

mono | mono

$$\mathcal{M}^t = \{y_k^t\}_{k=1,..,M_t}$$
$$\mathcal{M}^s = \{x_j^s\}_{j=1,..,M_s}$$

fr

backward NMT system

en

NMT system

fr

$y^t \sim \mathcal{M}^t$

Encoder → **Decoder** → $\bar{x}$ → **Encoder** → **Decoder** → **Cross-Entropy Loss**

$p(x|y)$

input sentence

prediction

$p(y|x)$

…and vice versa starting from English.

This is an example of auto-encoding or cycle consistency.
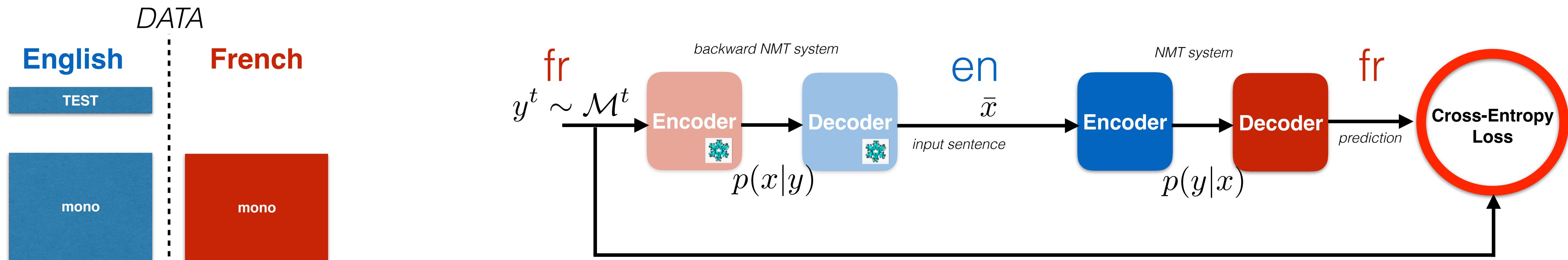
Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks

Jun-Yan Zhu*   Taesung Park*   Phillip Isola   Alexei A. Efros
Berkeley AI Research (BAIR) laboratory, UC Berkeley

Problem: lack of constrained on $\bar{x}$
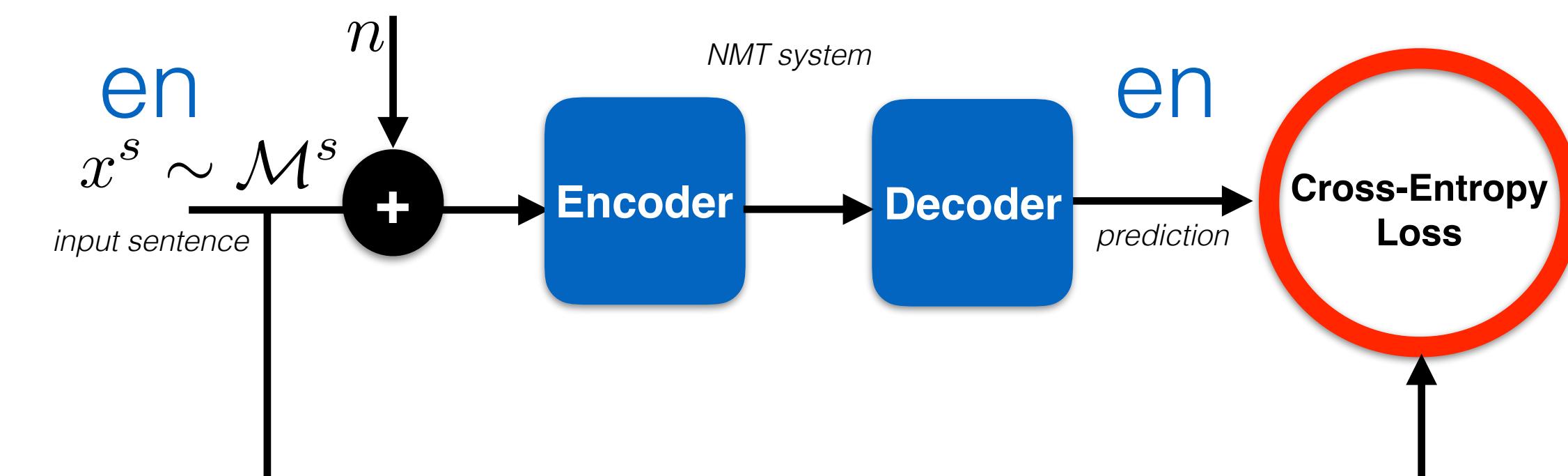
35

Case Study #1: Unsupervised MT

DATA

English | **French**

TEST

mono | mono

$$\mathcal{M}^t = \{y_k^t\}_{k=1,..,M_t}$$
$$\mathcal{M}^s = \{x_j^s\}_{j=1,..,M_s}$$

DAE makes sure decoder outputs fluently in the desired language.

fr — *backward NMT system* — en — *NMT system* — fr

$y^t \sim \mathcal{M}^t$ → **Encoder** → **Decoder** → $\bar{x}$ → **Encoder** → **Decoder** → $\xrightarrow{prediction}$ **Cross-Entropy Loss**

$p(x|y)$ *input sentence* $p(y|x)$

n

en   *NMT system*   en

$x^s \sim \mathcal{M}^s$ → **+** → **Encoder** → **Decoder** → $\xrightarrow{prediction}$ **Cross-Entropy Loss**

input sentence
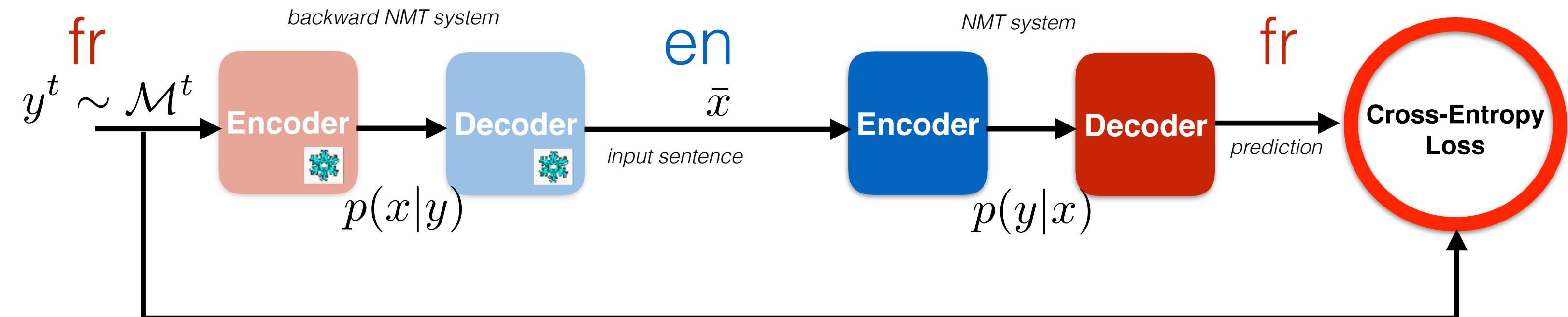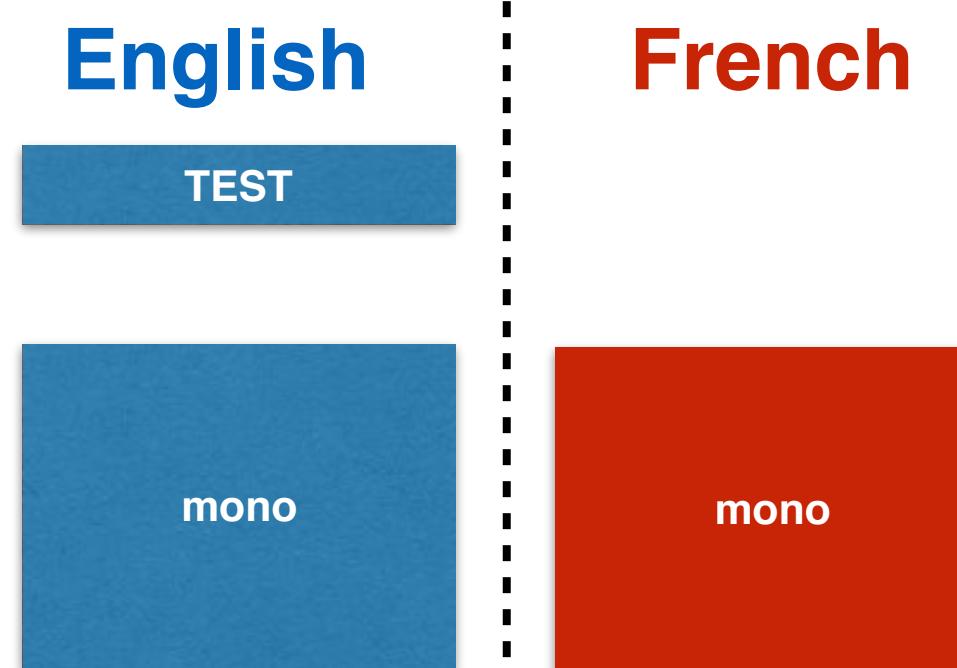
Problem: lack of modularity.

Decoder may behave differently when fed with representations from French encoder VS English encoder.

Lample et al. "Phrase-based and neural unsupervised MT" EMNLP 2018
Artetxe et al. "An effective approach to unsupervised MT" ACL 2019

Case Study #1: Unsupervised MT

DATA

English | **French**

TEST

mono | mono

$$\mathcal{M}^t = \{y_k^t\}_{k=1,..,M_t}$$
$$\mathcal{M}^s = \{x_j^s\}_{j=1,..,M_s}$$

DAE makes sure decoder outputs fluently in the desired language.

Like in multilingual NMT, share encoder and decoder parameters. Encoder is encouraged to produce shared representations (particularly if pre-trained).

fr   *backward NMT system*   en   *NMT system*   fr

$y^t \sim \mathcal{M}^t$

Encoder → **Decoder** $\bar{x}$ **Encoder** → **Decoder**   *prediction* → **Cross-Entropy Loss**

$p(x|y)$ *input sentence* $p(y|x)$

en n *NMT system*   en

$x^s \sim \mathcal{M}^s$

input sentence → **+** → **Encoder** → **Decoder**   *prediction* → **Cross-Entropy Loss**

Src   *NMT system*   Tgt

(x, LID)

Encoder → **Decoder**   *prediction*

input sentence with target language ID

37

Case Study #1: Unsupervised MT

DATA

English | **French**

TEST

mono | mono

$$\mathcal{M}^t = \{y_k^t\}_{k=1,..,M_t}$$
$$\mathcal{M}^s = \{x_j^s\}_{j=1,..,M_s}$$

DAE makes sure decoder outputs fluently in the desired language.

Like in multilingual NMT, share encoder and decoder parameters. Encoder is encouraged to produce shared representations (particularly if pre-trained).

fr
backward NMT system
NMT system
fr
$y^t \sim \mathcal{M}^t$
en
Encoder
$\bar{x}$
Cross-Entropy Loss
prediction

ITERATIVE BT

$p(x|y)$
$p(y|x)$

+

en
n
NMT system
en
$x^s \sim \mathcal{M}^s$
input sentence
Encoder
Encoder
prediction
Cross-Entropy Loss

DAE

+

Src
NMT system
Tgt
$(x, ...)$
input sentence with target language ID
prediction

Multi-Lingual

WMT'14 En-Fr

Same ideas can be applied to phrase-based statistical MT systems (PBSMT). NMT and PBSMT can be combined for even better results.

Since unsupMT was trained on about 10M sentences, each parallel sentence is worth 100 monolingual sentences (for this dataset and language pair).

Lample et al. "Phrase-based and neural unsupervised MT" EMNLP 2018
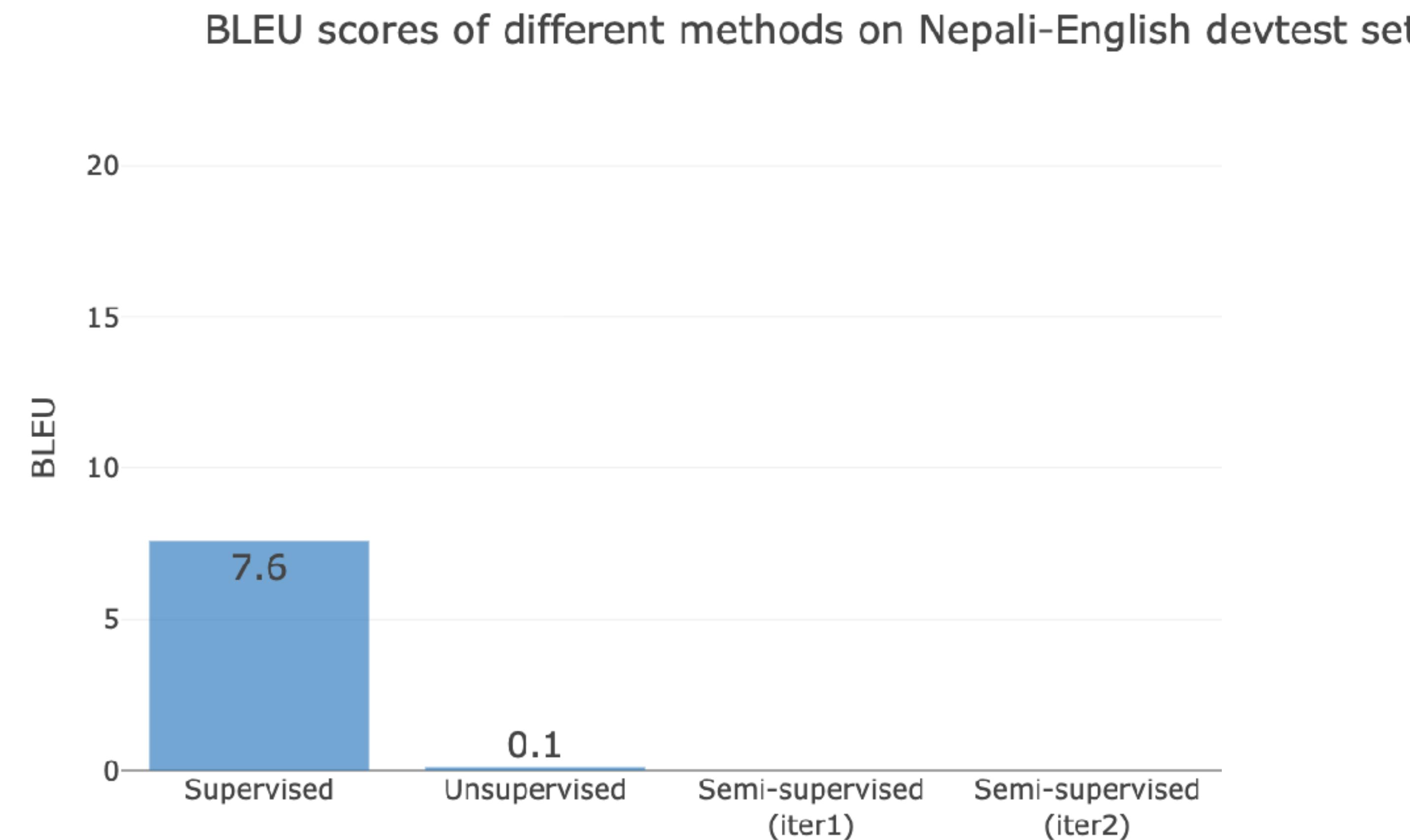
Case Study #2: FLoRes Ne-En

	In-domain (Wikipedia)	Out-of-domain
Parallel	None	500K sentences (Bible, GNOME/Ubuntu, OpenSubtitle, …) *Hindi: 1.5M
Monolingual	100K sentences	~5M sentences (CommonCrawl) *Hindi: 45M
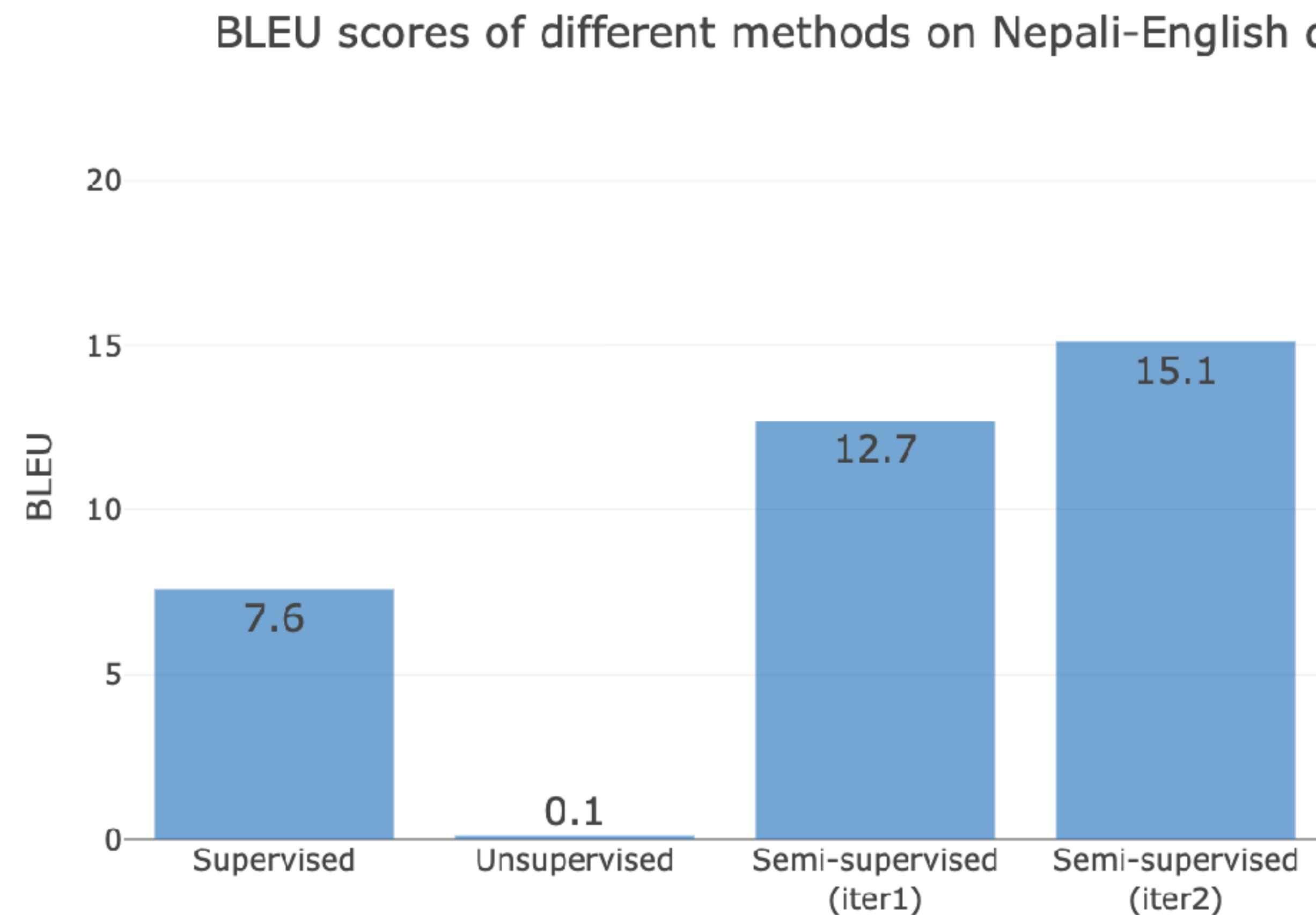
Results on FLoRes: Ne-En

BLEU scores of different methods on Nepali-English devtest set

Results on FLoRes: Ne-En

BLEU scores of different methods on Nepali-English devtest set

Results on FLoRes: Ne-En

BLEU scores of different methods on Nepali-English devtest set

Results on FLoRes: Ne-En

BLEU scores of different methods on Nepali-English devtest set
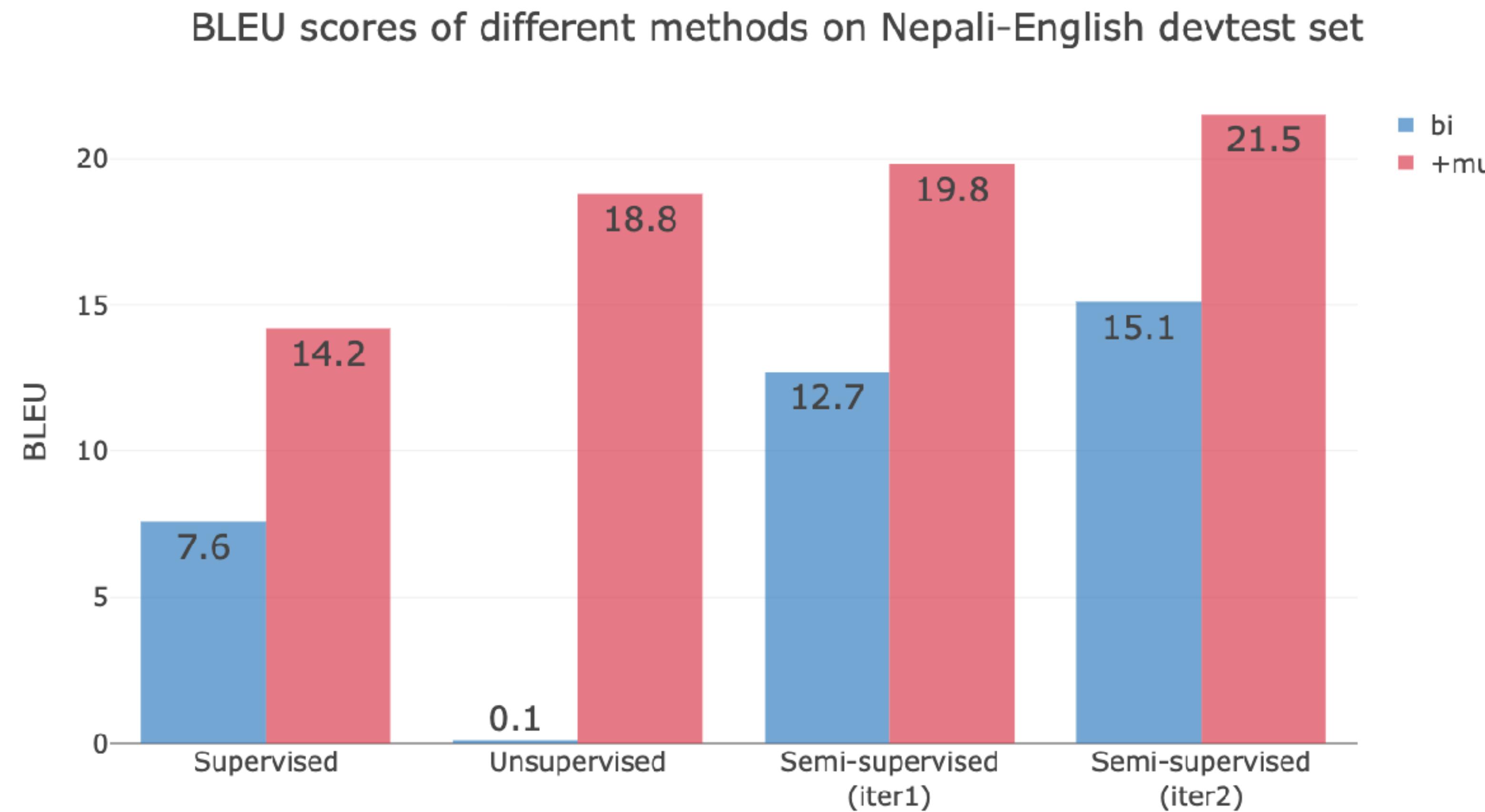
Case-Study #3: English-Burmese

ပင်မစာမျက်နှာ

ဝီကီပီးဒီးယားမှ ကြိုဆိုပါသည်။

မည်သူမဆို ကြည့်ရှုပြင်ဆင်နိုင်သော အခမဲ့လွတ်လပ်စွယ်စုံကျမ်း ဖြစ်ပါသည်။

အကြောင်းအရာပေါင်း ၄၄၈၁၄ ခုကို မြန်မာဘာသာဖြင့် ဖတ်ရှုနိုင်ပါသည်။

- အနုပညာ
- အတ္ထုပ္ပတ္တိ
- ပထဝီဝင်
- သမိုင်း
- သချာ
- သိပ္ပံ
- လူမှုရေး
- နည်းပညာ
- မုခ်ဦးအားလုံး

🏵 အထူးအကြောင်းအရာ

သိန္နီနယ်   သည် ယခင် ရှမ်းပဒေသရာဇ်ပြည်နယ်များတွင် ပါဝင်ခဲ့သော ပြည်နယ်တစ်ခု ဖြစ်သည်။ သိန္နီ နယ်ကို ခရစ် ၁၈၈၈ ခုနှစ် အင်္ဂလိပ်တို့ ဝင်ရောက်သိမ်းပိုက်ပြီးနောက်မှ မြောက်သိန္နီနယ် (သိန္နီနယ်)နှင့် တောင်သိန္နီနယ်(မိုင်းရယ်နယ်)ဟု ခွဲခြားအုပ်ချုပ်ခဲ့သည်။ ရှေးအခါသမယက သိန္နီနယ်ကြီးသည် အစိတ် စိတ်ကွဲပြားခြင်းမရှိဘဲ ရှမ်းပြည်နယ်တစ်ဝှမ်းလုံးတွင် အကျယ်ပြန့်ဆုံး အာဏာအလွှမ်းမိုးဆုံးသော နယ်ကြီးဖြစ်ခဲ့ သည်။ သို့သော် မြန်မာဘုရင်များ ဝင်ရောက်တိုက်ခိုက် သိမ်းပိုက်ပြီးသည်နောက် အုပ်ချုပ်ရေးဝါဒအရ ရာထူးလု သည့်နယ်ရှင် စော်ဘွားတို့ကြောင့် သိန္နီနယ်ကြီးသည် ငါးနယ်အထိ အစိတ်စိတ် ကွဲပြားခဲ့လေသည်။ ထို အတွင်း စော်ဘွားအချင်းချင်း စိတ်ဝမ်းကွဲကာ တစ်ဦးနှင့်တစ်ဦးတိုက်ခိုက်၍ ဆိုင်ရာ နယ်ပယ်များကို အုပ်ချုပ်ကြ သည်။ နောက်ဆုံးခရစ် ၁၈၈၈ ခုနှစ်၊ အင်္ဂလိပ်တို့ဝင်ရောက်လာမှ အထက်ပါ အတိုင်းနှစ်နယ်ခွဲ၍ အုပ်ချုပ်ခဲ့သည်။ နှစ်နယ်ခွဲ၍ အုပ်ချုပ်စက ခွန်ဆိုင်တုံဟန်းအား မြောက်သိန္နီနယ်အတွက် စော်ဘွားအဖြစ်လည်းကောင်း၊ ဆိုင် နော်ဖၝ၏သား နော်မိုင်းအား တောင်သိန္နီနယ် စော်ဘွားအဖြစ်လည်းကောင်း၊ ခန့်အပ်ခဲ့လေသည်။ ၁၉၂၅ ခုနှစ် တွင် မြောက်သိန္နီနယ်ကို စပ်ဟုံဖက စော်ဘွားအဖြစ် ဆောင်ရွက်ခဲ့လေသည်။
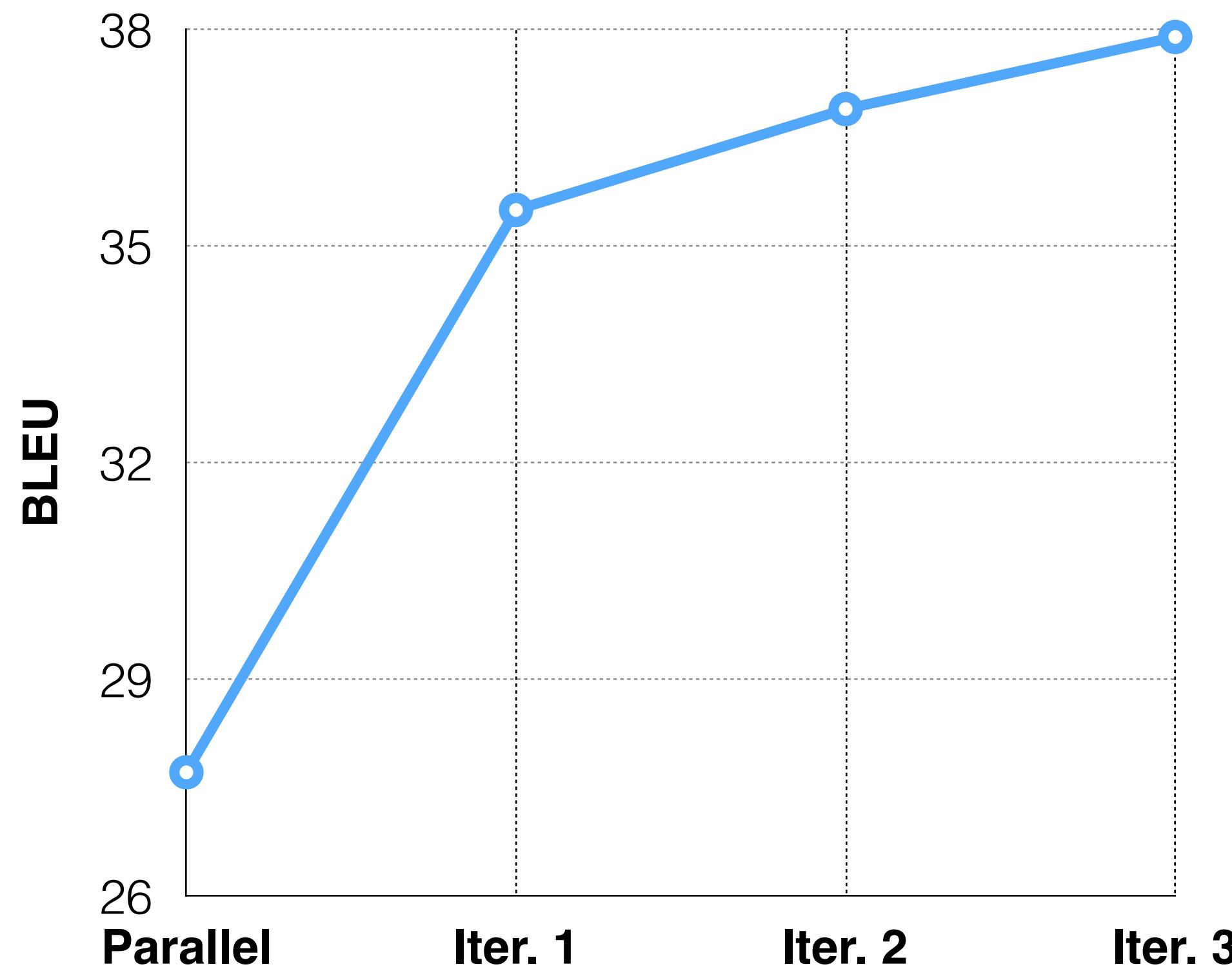
Workshop on Asian Translation 2019: English-Burmese
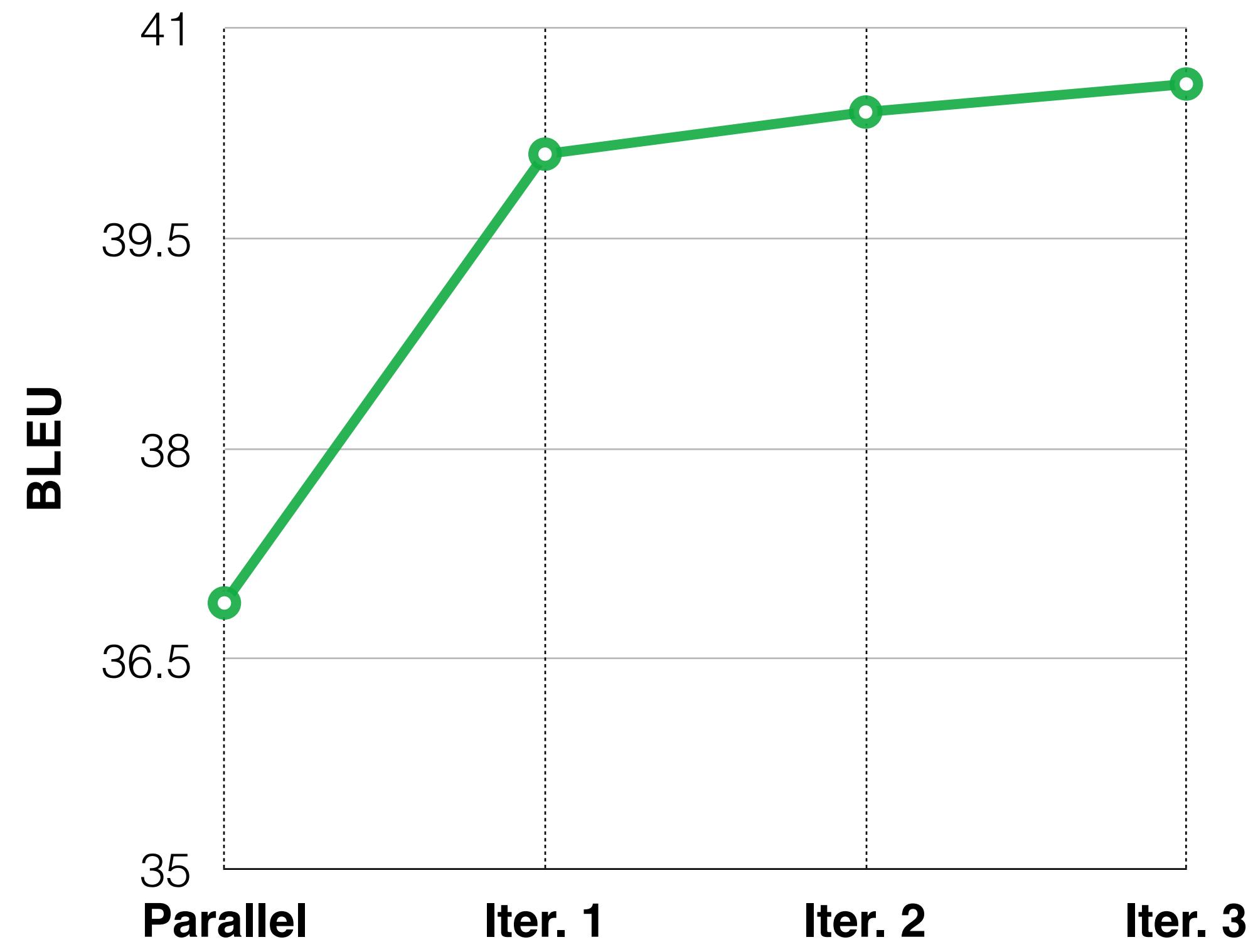
	In-domain (News)	**Out-of-domain**
Parallel	20K sentences	200K sentences
Monolingual	~79M sentences (En only)	~23M sentences (My only)

"FBAI WAT'19 My-En translation task submission" Chen et al., WAT@EMNLP 2019

Results: Iterative ST+BT
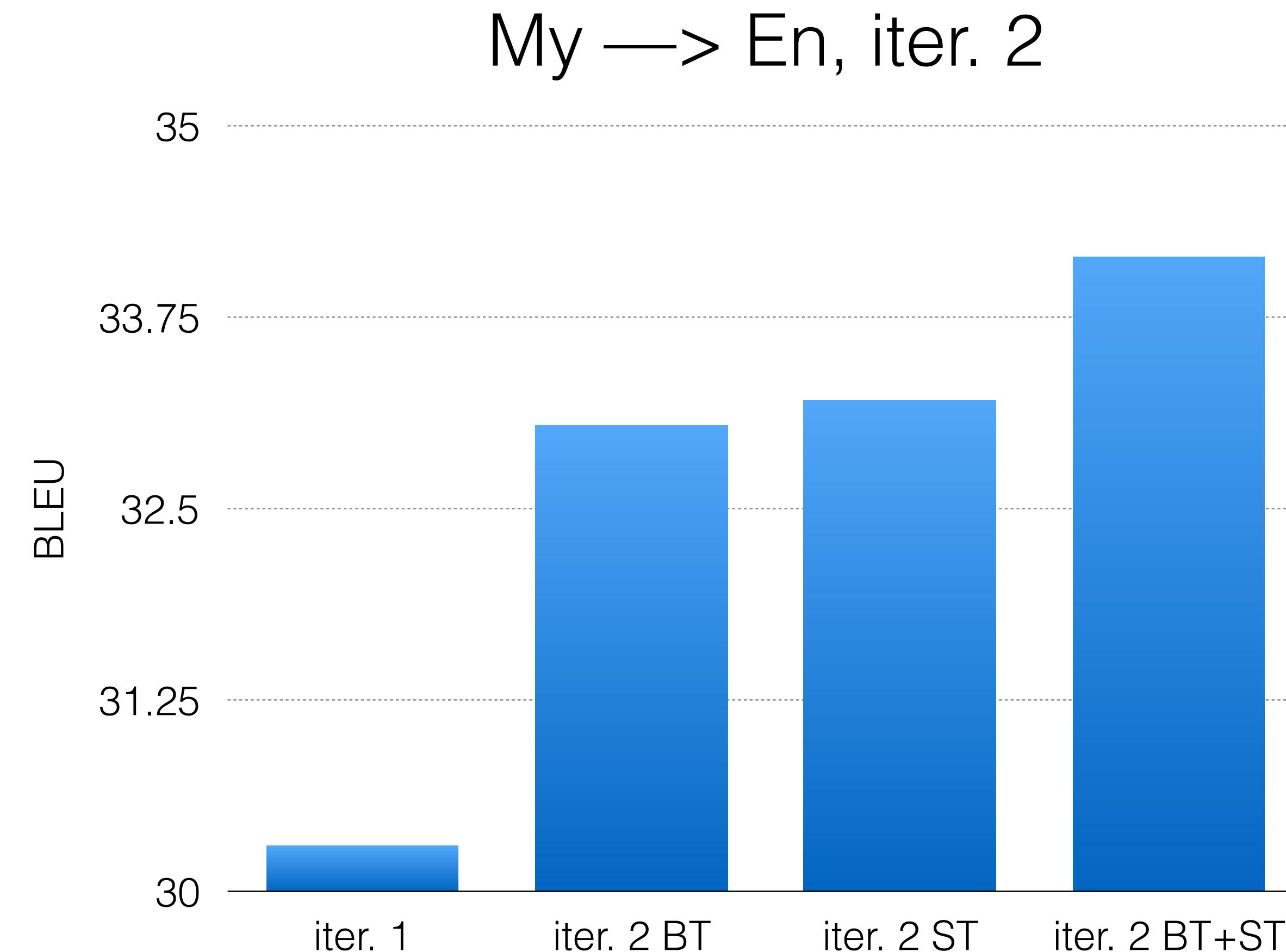
My —> En

En —> My
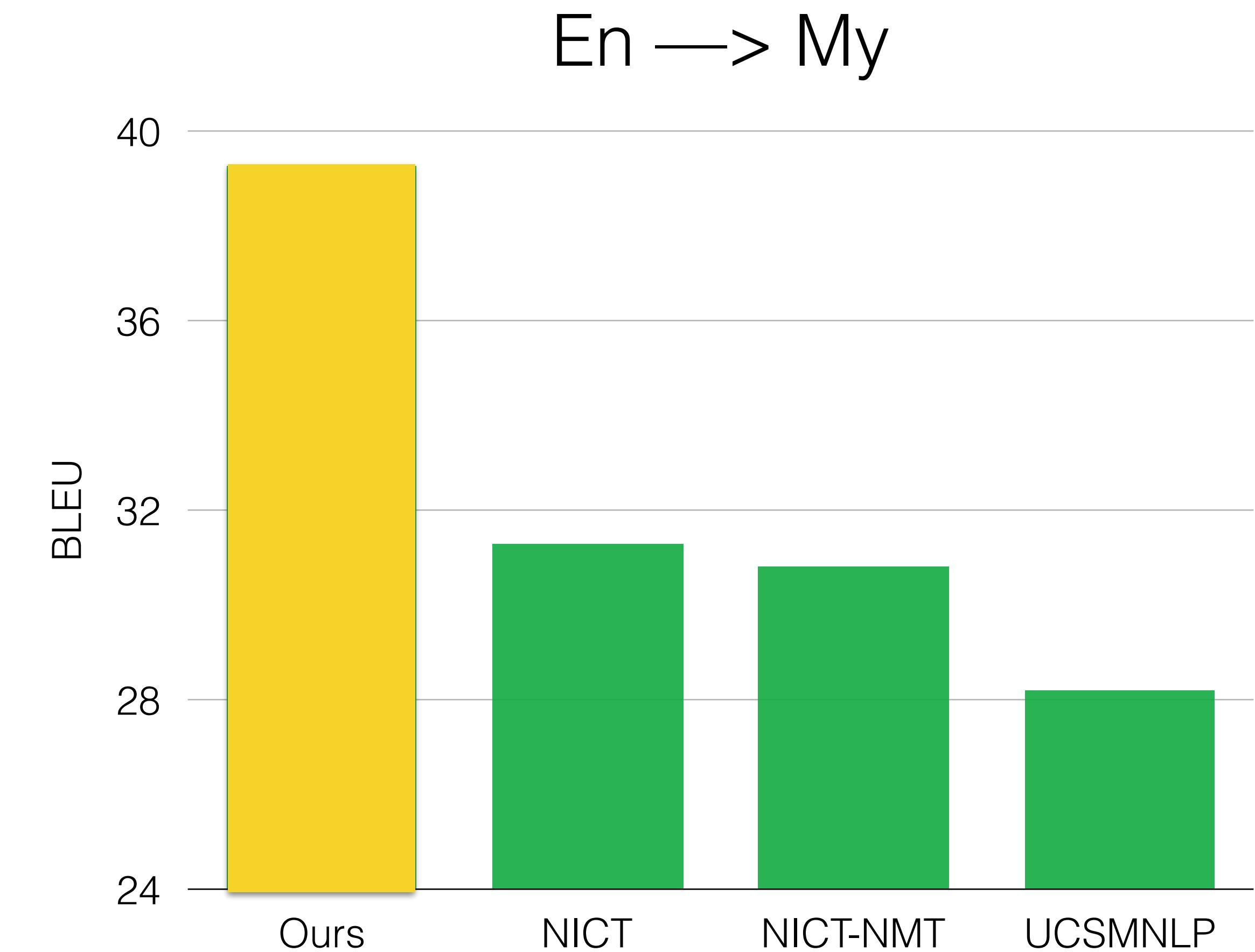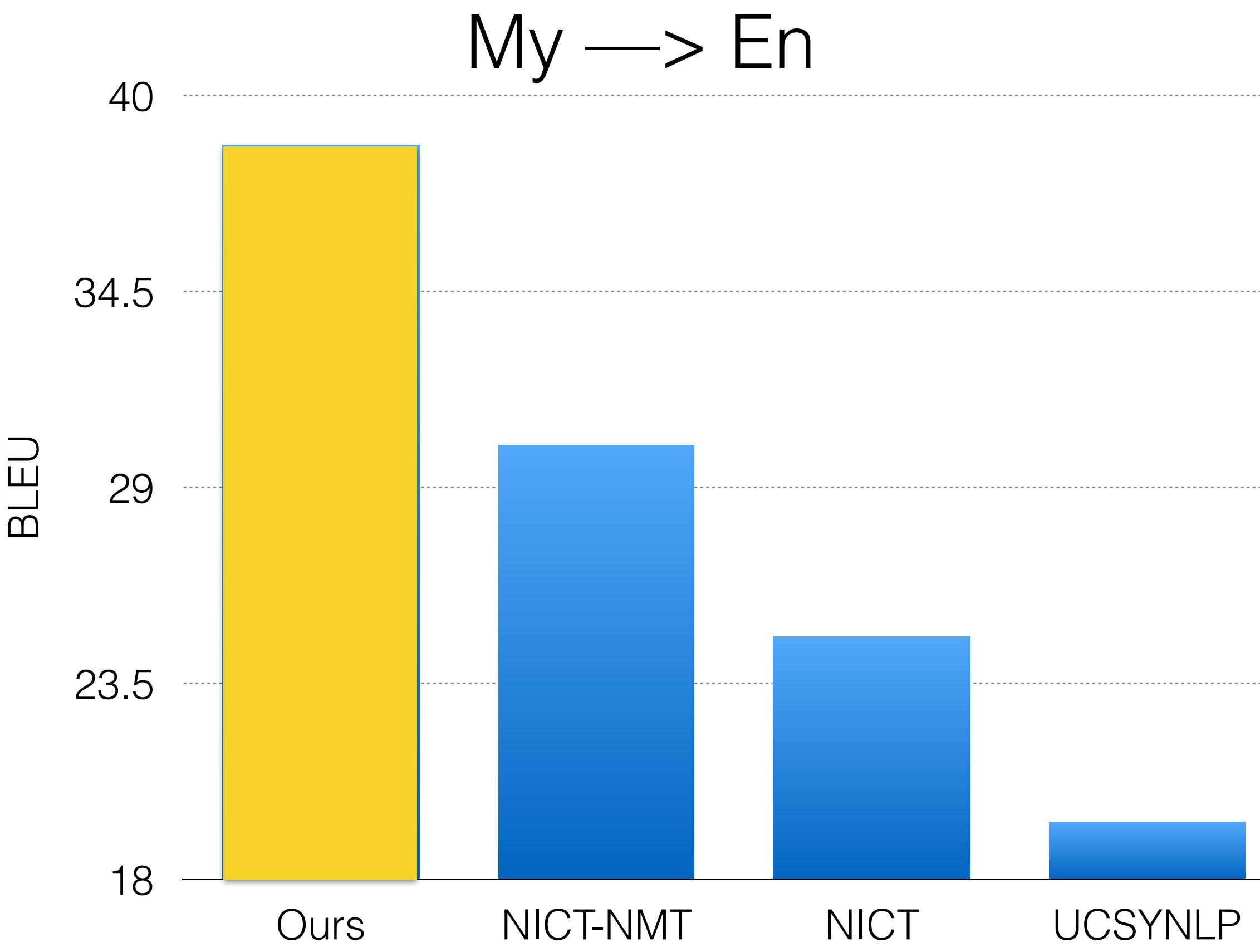
Results: BT vs ST vs BT+ST

My —> En, iter. 2

Final Results of 2019 Competition

+8 BLEU compared to second best

My —> En

En —> My

"FBAI WAT'19 My-En translation task submission" Chen et al., WAT@EMNLP 2019

Demo (Burmese —> English)

slide credit to Peng-Jen Chen
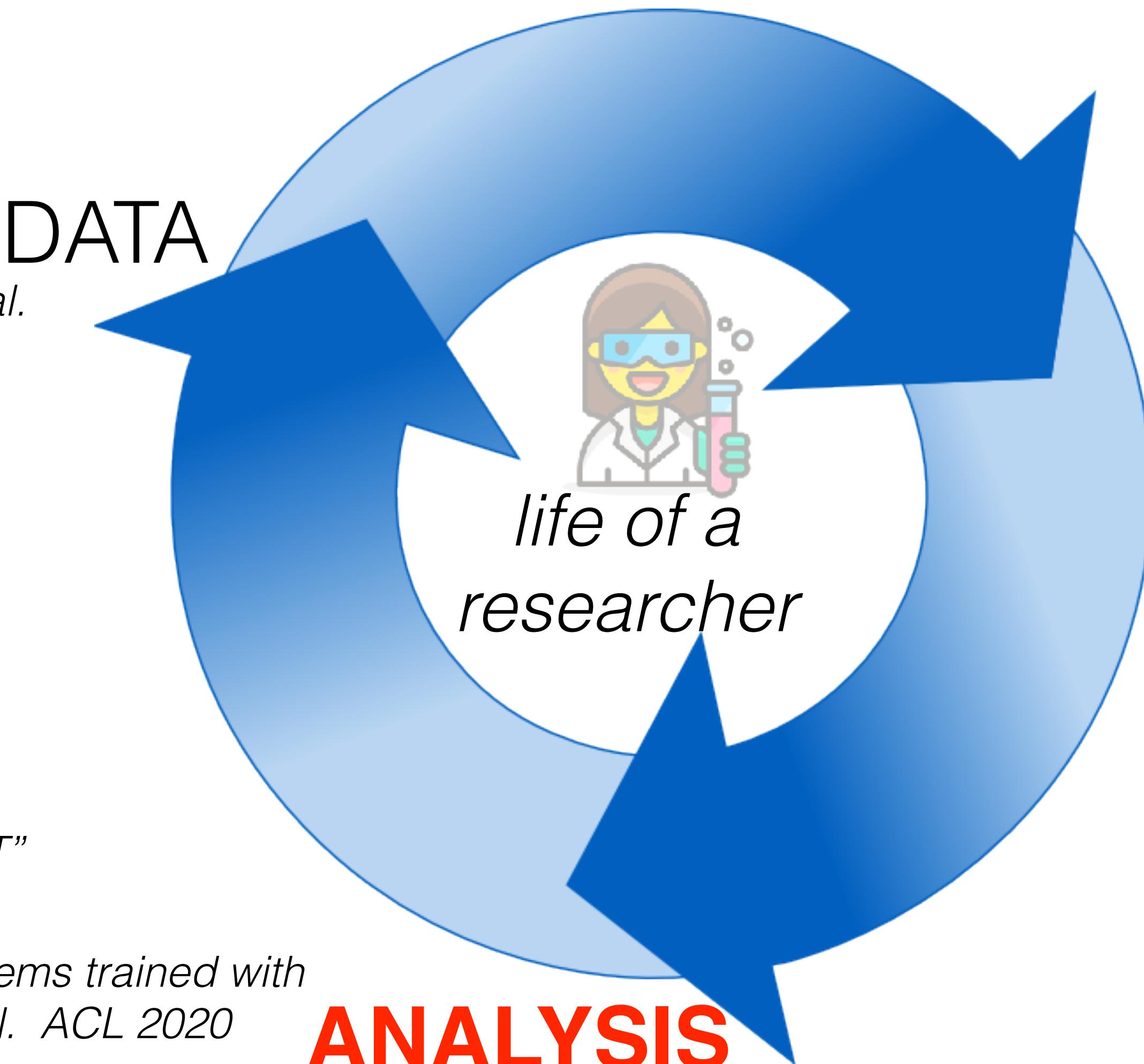
Conclusion so far…

- Iterative back-translation, multi-lingual training work remarkably well.

- By feeding more data (BT, ST, pre-training, multi-lingual training) we can afford training bigger models. Bigger models train on more data generalize better.

- Low-resource MT requires big compute! Remember that about 100 monolingual sentences give the same training signal as a single pair of parallel sentence.

Outline

DATA

MODEL

life of a researcher

ANALYSIS

"The FLoRes evaluation for low resource MT:…" Guzmán, Chen et al. 'EMNLP 2019

"Phrase-based & Neural Unsup MT" Lample et al. EMNLP 2018

"FBAI WAT'19 My-En translation task submission" Chen et al., WAT@EMNLP 2019

"Massively Multilingual NMT" Aharoni et al.,ACL 2019

"Multilingual Denoising Pre-training for NMT" Liu et al., arXiv 2001:08210 2020

"Analyzing uncertainty in NMT" Ott et al. ICML 2018

"On the evaluation of MT systems trained with back-translation" Edunov et al. ACL 2020

"The source-target domain mismatch problem in MT" Shen et al. arXiv 1909.13151 2019

52
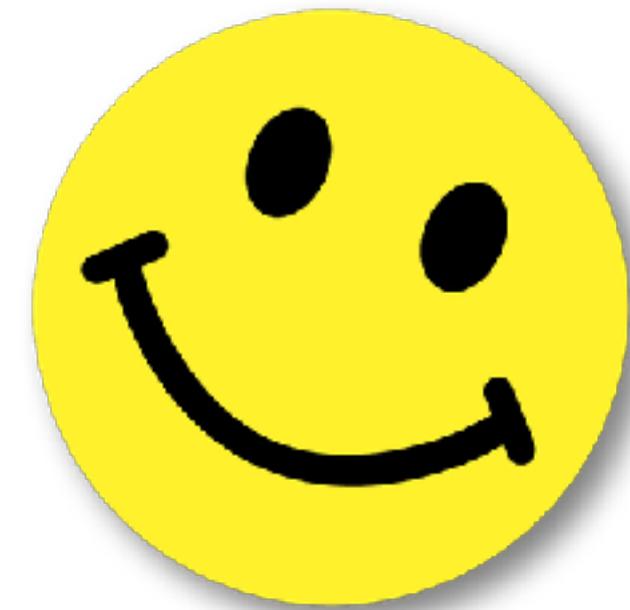
Simulating Low-Resource MT

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!

A Worrisome Finding

BT sometimes yields very mild improvements.
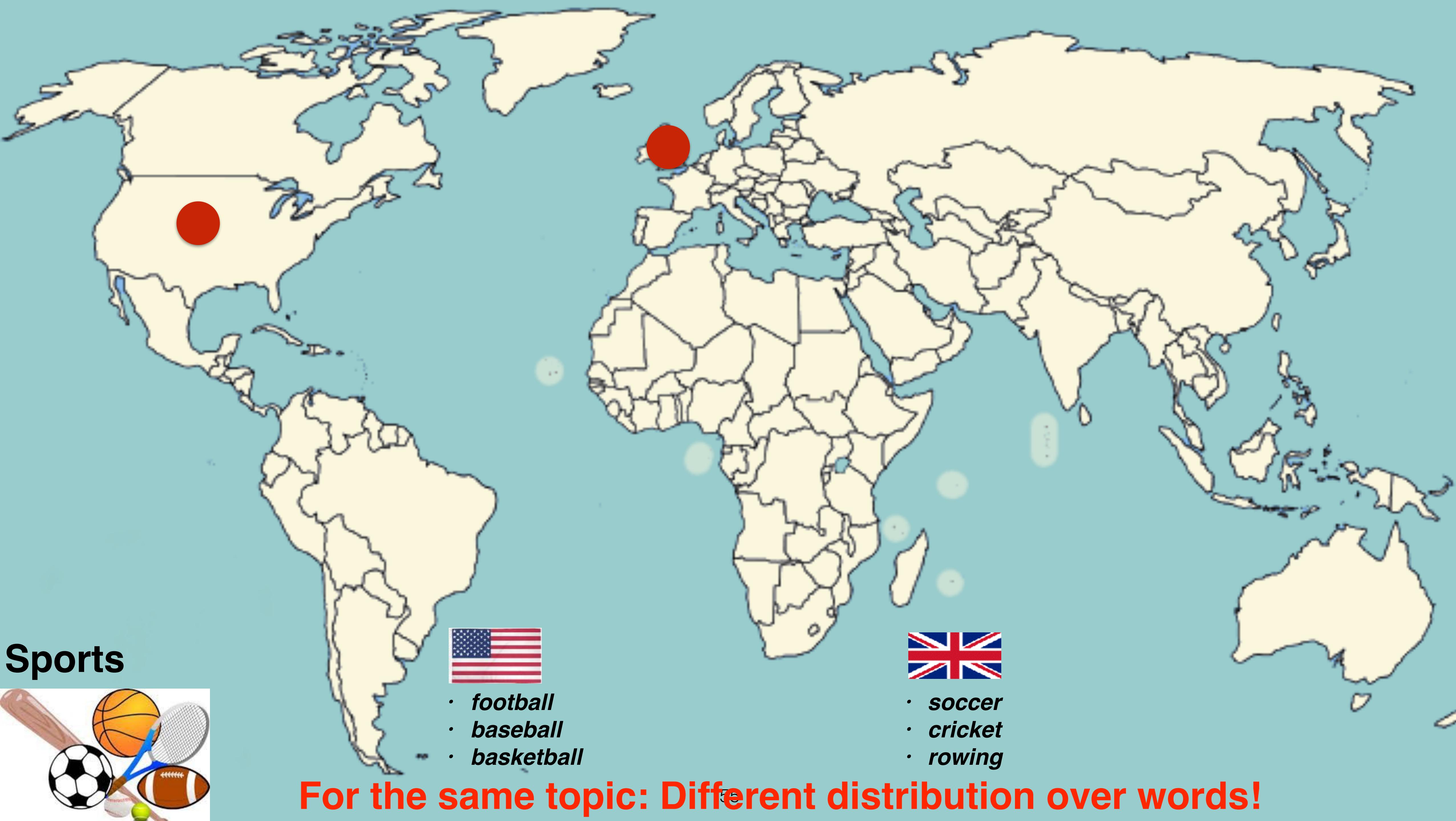
Example
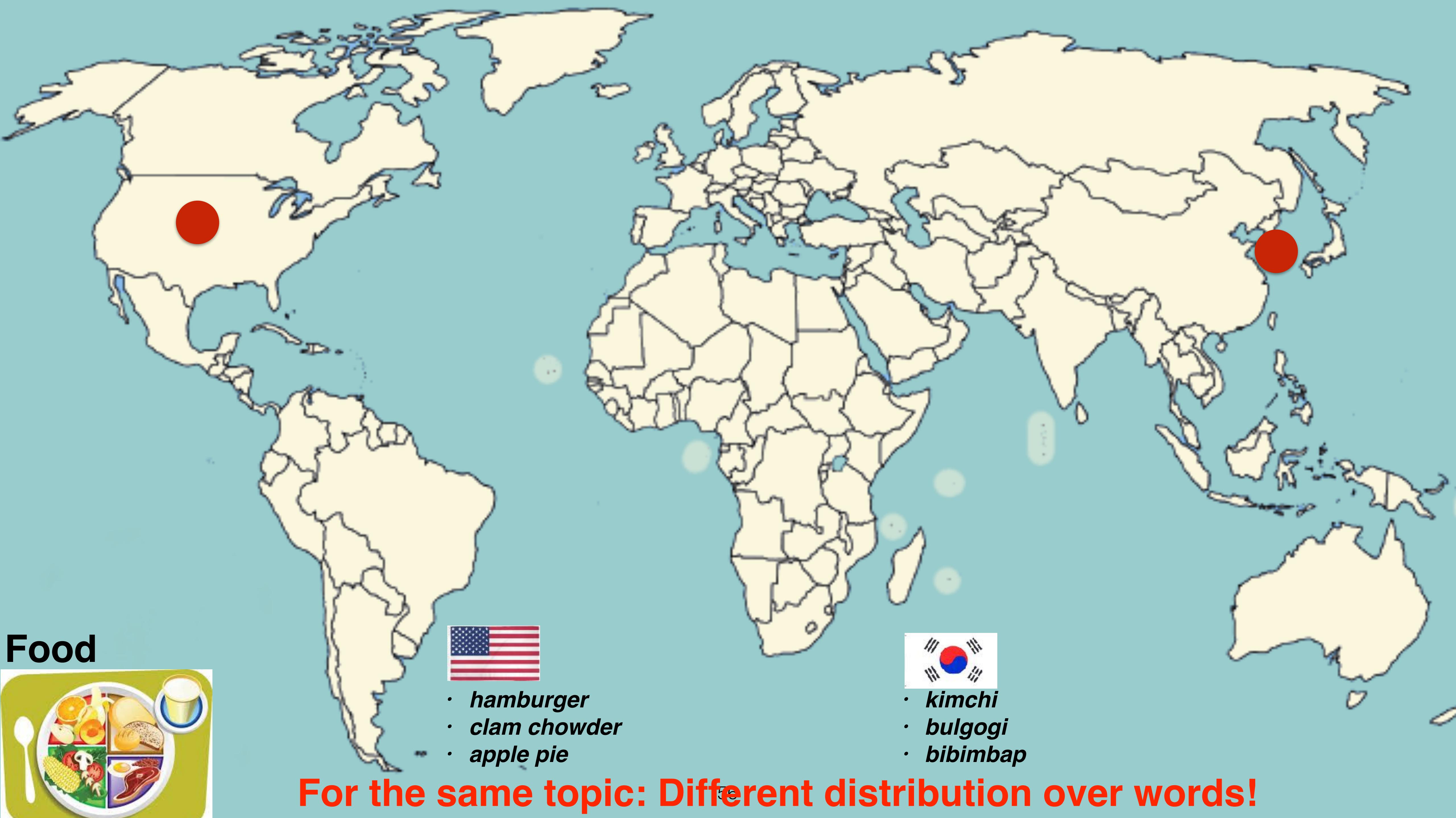
FB public posts En—>My

only parallel data	15.2 BLEU
parallel data + BT	15.3 BLEU
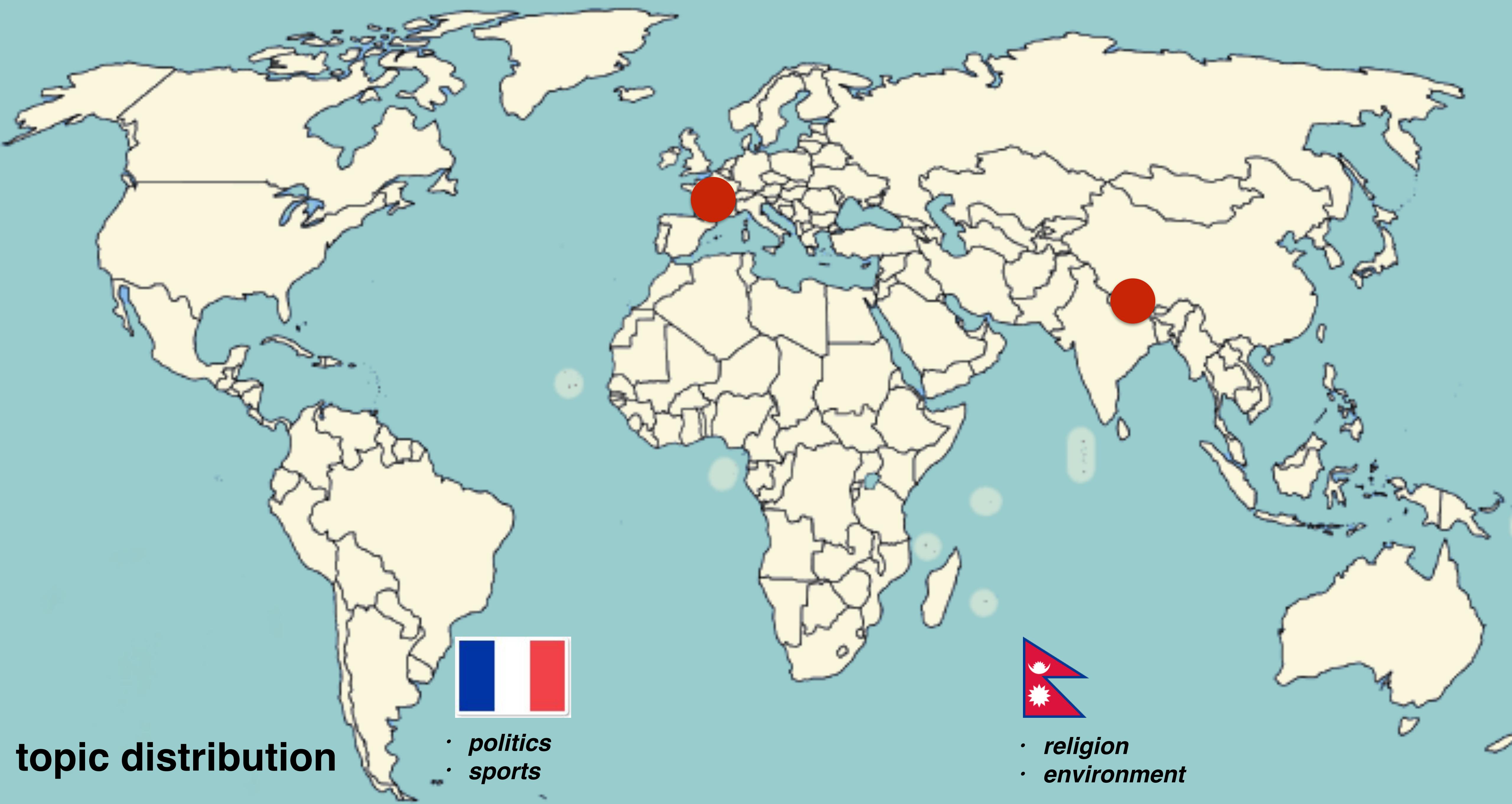
+0.1 BLEU!

Why is BT not working as well?

Sports

🇺🇸
- *football*
- *baseball*
- *basketball*

🇬🇧
- *soccer*
- *cricket*
- *rowing*

For the same topic: Different distribution over words!

Food

🇺🇸

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

🇰🇷

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

For the same topic: Different distribution over words!

topic distribution

politics
sports

religion
environment

Domains differ in the topic distribution

Examples from FLoRes

Si-En

original

අධ් යාපනයෙන් පසු හෝ පවුලේ යුතුකම් ඉටු කරන්නට හෝ රෝග තත්වයන් නිසා සංසයා උපසම්පදාවෙන් නිතරම ඉවත් වෙති .

After education priests leave ordination in order to fulfill duties to the family or due to sickness.

translation

තර්ජන , ශාරීරික හිංසනය , දේපල හානිය , පහර දීම සහ මරාදැමීම මෙම දඬුවමිය .

Threatening, physical violence, property damage, assault and execution are these punishments.

Wikipedia in Sinhala has different topic distribution.

En-Si

In Serious meets, the absolute score is somewhat meaningless.

සැබෑ තරහ වලදී ලකුණු සැසදීම තේරුමක් තැති ක්‍රියාවකි .

Iphone users can and do access the internet frequently, and in a variety of places.

අයිෆෝන් භාවිත කරන්නට නිතරම සහ විවිධ ස්ථානවලදී අන්තර්ජාලයට පිවිසිය හැකිය .

Guzmàn, Chen et al. "The FLoRes evaluation datasets for low resource MT…" EMNLP 2019
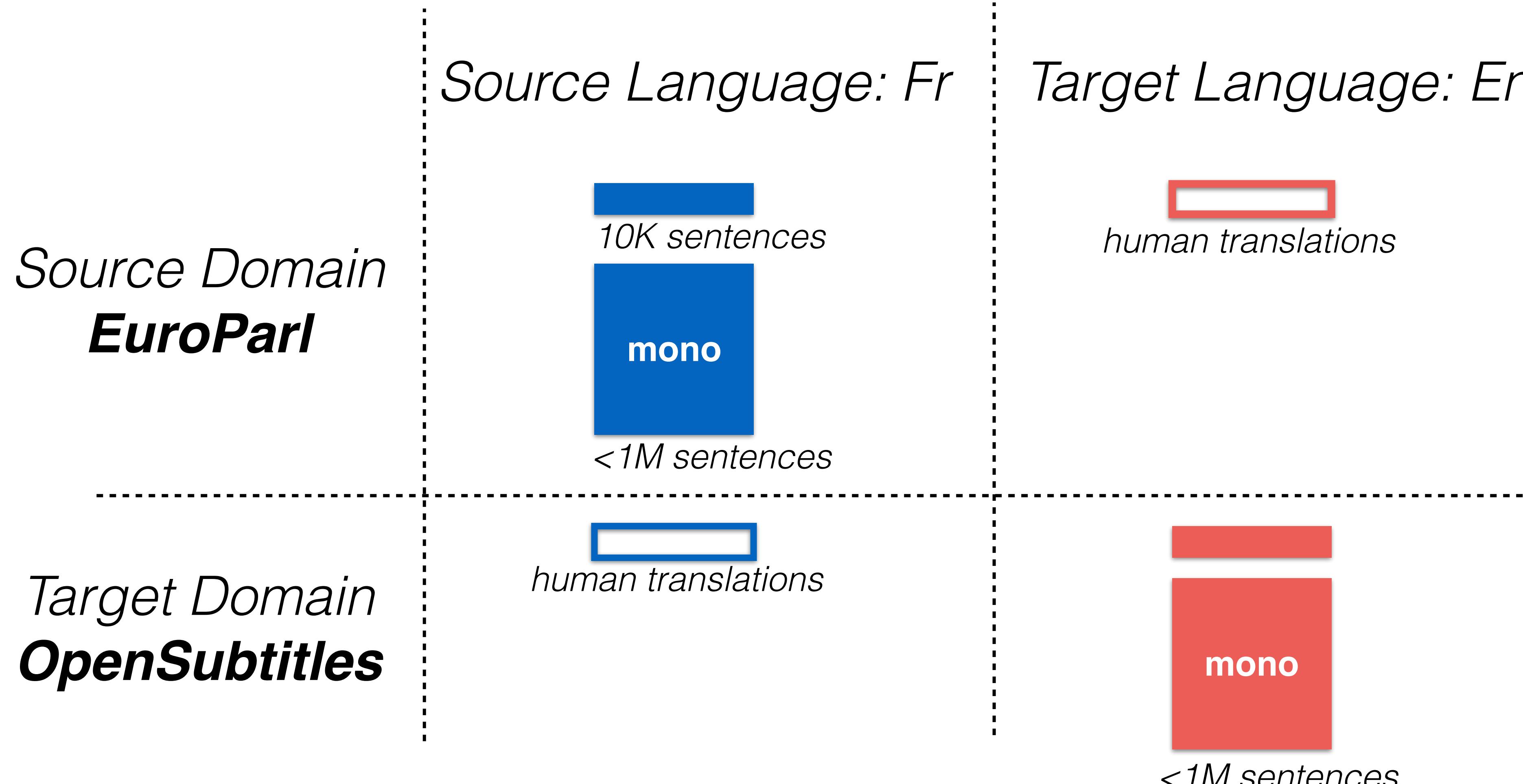
Source-Target Domain Mismatch (STDM)

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

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

- STDM manifests itself in terms of: Different distribution of topics, and for the same topic different distribution over words.

- STDM is hard to measure because of the unknown effect of translationese.

- STDM makes the MT problem even harder: source originating data and target originating data are not comparable in general. BT won't work as well. UnsupMT won't work as well.

Language and place. Johnstone Cambridge Univ. Press 2010
Leech et al. "Computer corpora: What do they tell us about culture?"Journal Computers in English Linguistics 1992
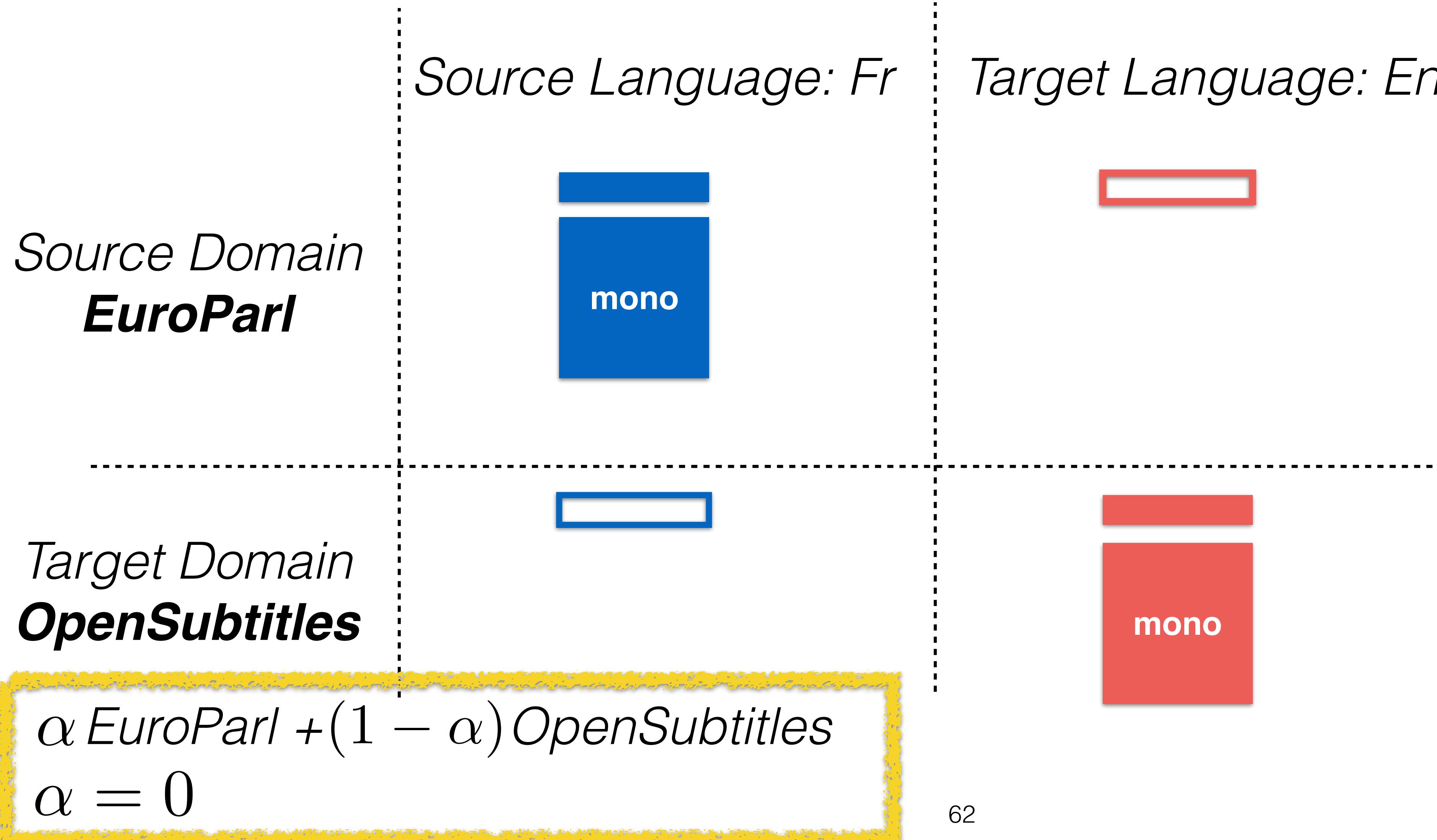
Questions

- Is it true that BT is less effective when there is STDM?

- What baselines shall we consider when there is STDM?

- What are general best practices when there is STDM?

- How to study STDM in a controlled setting?
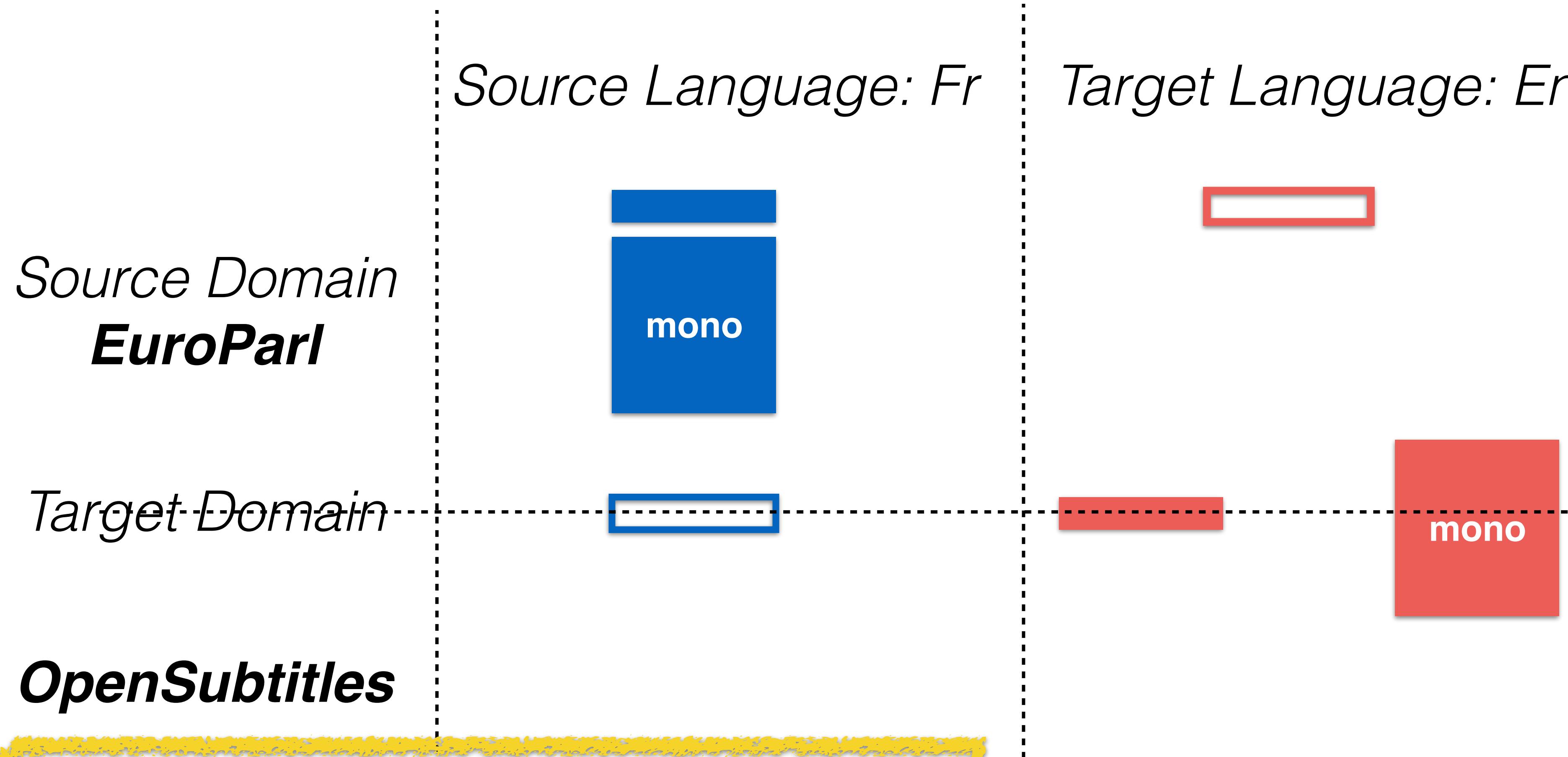
Controlled Setting

Source Language: Fr   *Target Language: En*

Source Domain
EuroParl

10K sentences

mono

<1M sentences

human translations

Target Domain
OpenSubtitles

human translations

mono

<1M sentences

Controlled Setting

Source Language: Fr | *Target Language: En*

Source Domain
EuroParl

mono

Target Domain
OpenSubtitles

mono

$$\alpha \, EuroParl + (1 - \alpha) \, OpenSubtitles$$
$$\alpha = 0$$

Controlled Setting

Source Language: Fr *Target Language: En*

Source Domain
EuroParl

mono

Target Domain

mono

OpenSubtitles

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

Controlled Setting

Source Language: Fr | *Target Language: En*

Source Domain &
Target Domain
EuroParl

mono

mono

$\alpha \, EuroParl + (1 - \alpha) \, OpenSubtitles$
$\alpha = 1$
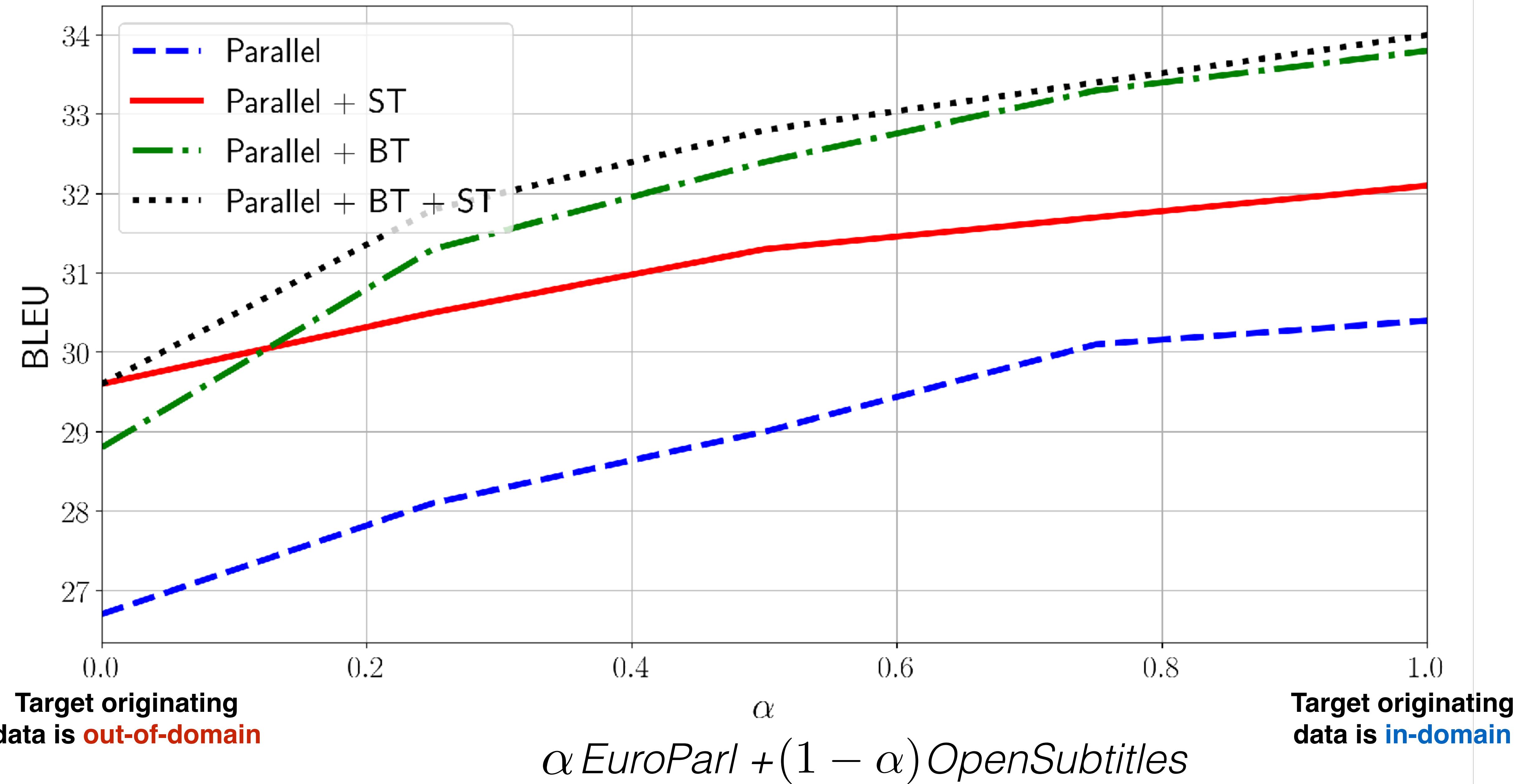
	target mono data	*source mono data*	*in-domain mono data*
• Back-Translation:	✓	✗	✗
• Self-Training:	✗	✓	✓

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?
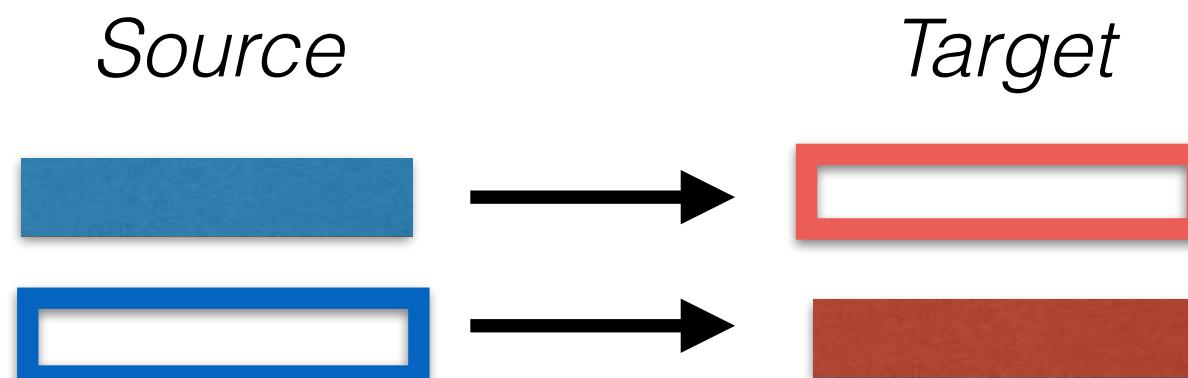
Varying Domain of Target Originating Data

Legend:
- Parallel (blue dashed)
- Parallel + ST (red solid)
- Parallel + BT (green dash-dot)
- Parallel + BT + ST (black dotted)

BLEU (y-axis)

α (x-axis)

Target originating data is out-of-domain

Target originating data is in-domain

$$\alpha \, EuroParl + (1 - \alpha) \, OpenSubtitles$$
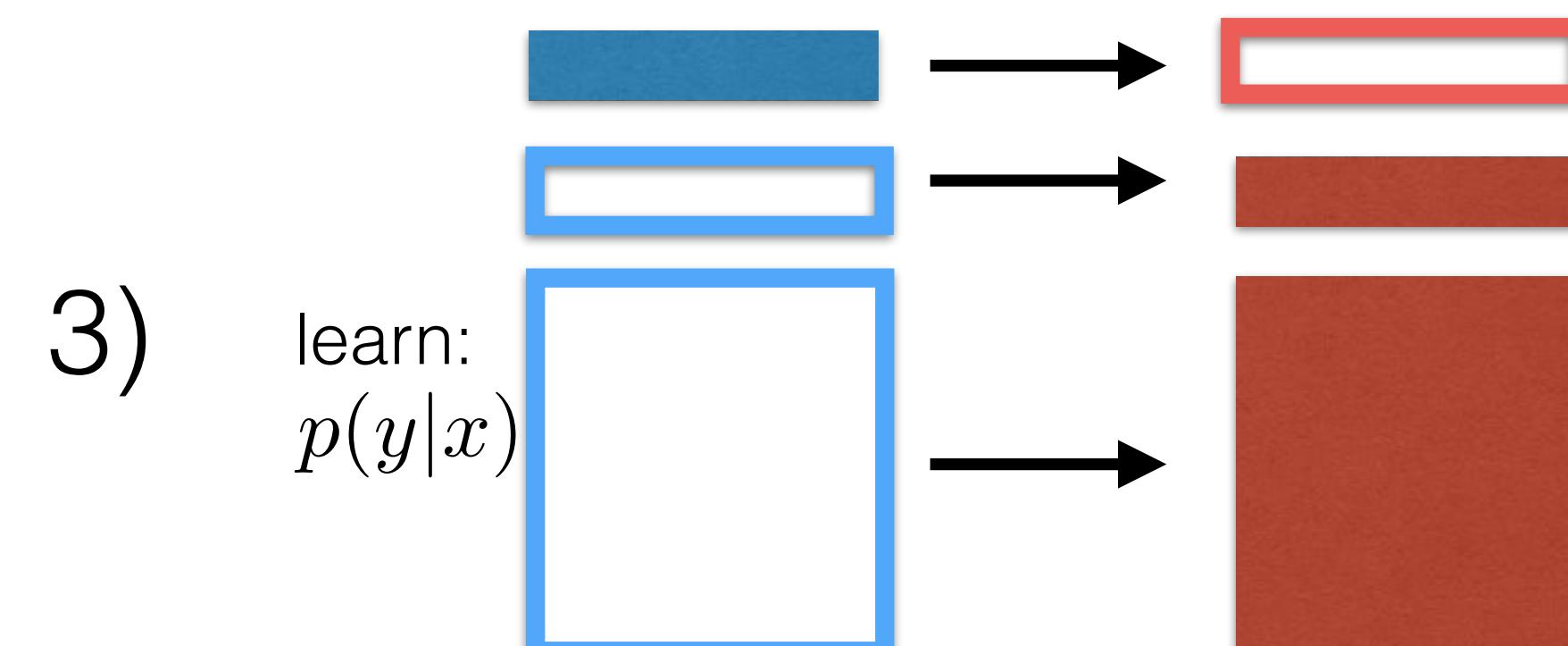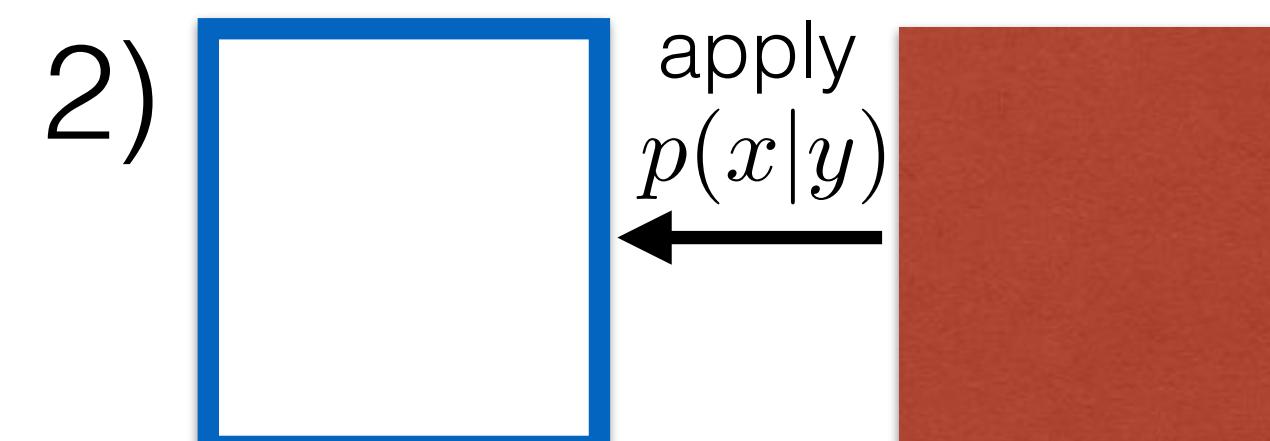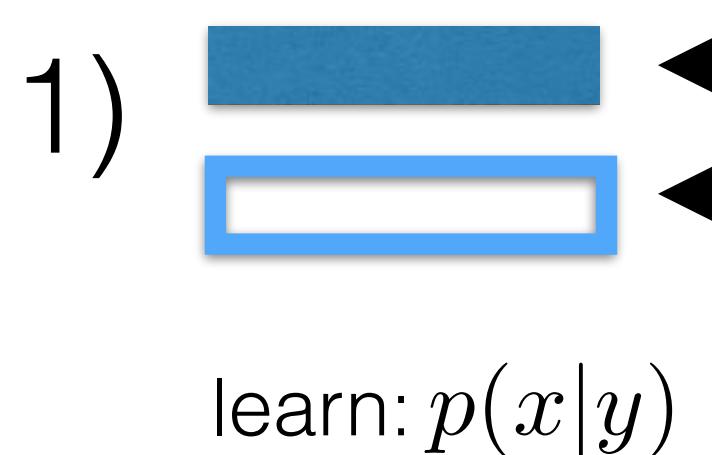
Baseline Approaches
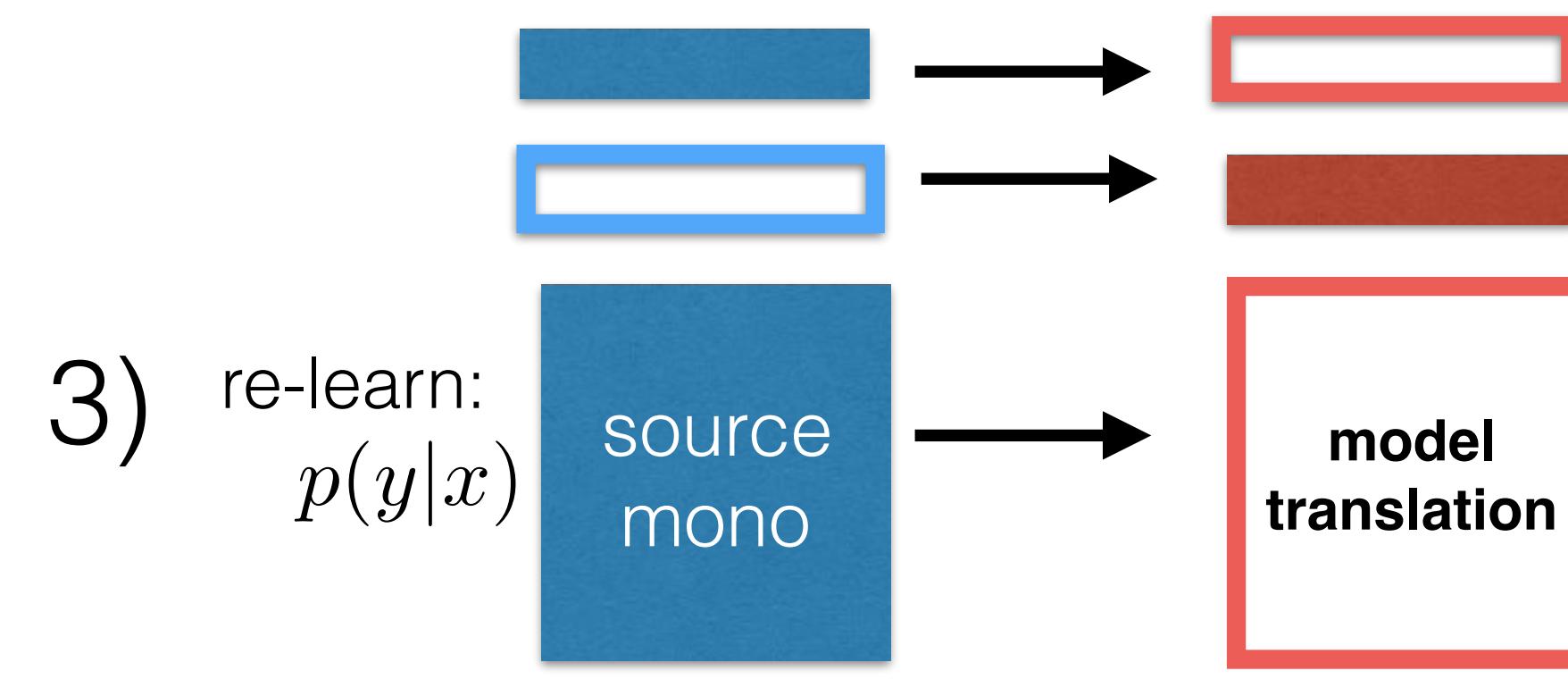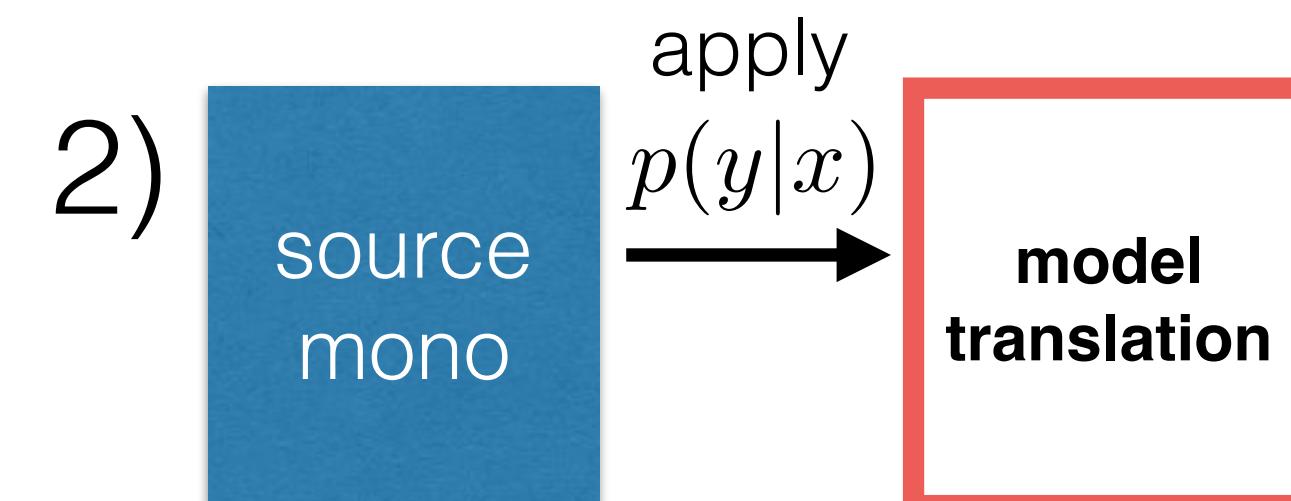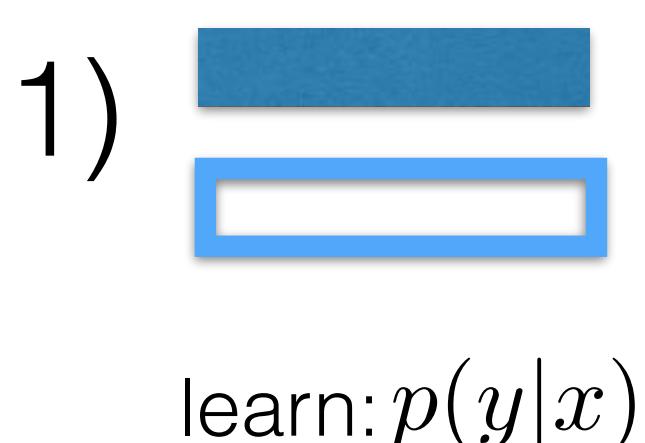
- Bitext only:

- Back-Translation:

1) learn: $p(x|y)$

2) apply $p(x|y)$

3) learn: $p(y|x)$

- Self-Training:

1) learn: $p(y|x)$

2) source mono → apply $p(y|x)$ → **model translation**

3) re-learn: $p(y|x)$ source mono → **model translation**

Baseline Approaches: Only In-Domain Data

- Bitext only:

Source   *Target*

- Back-Translation:

1)   learn: $p(x|y)$

2)   apply $p(x|y)$

3)   learn: $p(y|x)$

- Self-Training:

1)   learn: $p(y|x)$

2)   apply $p(y|x)$

source mono   **model translation**

3)   re-learn: $p(y|x)$

source mono   **model translation**
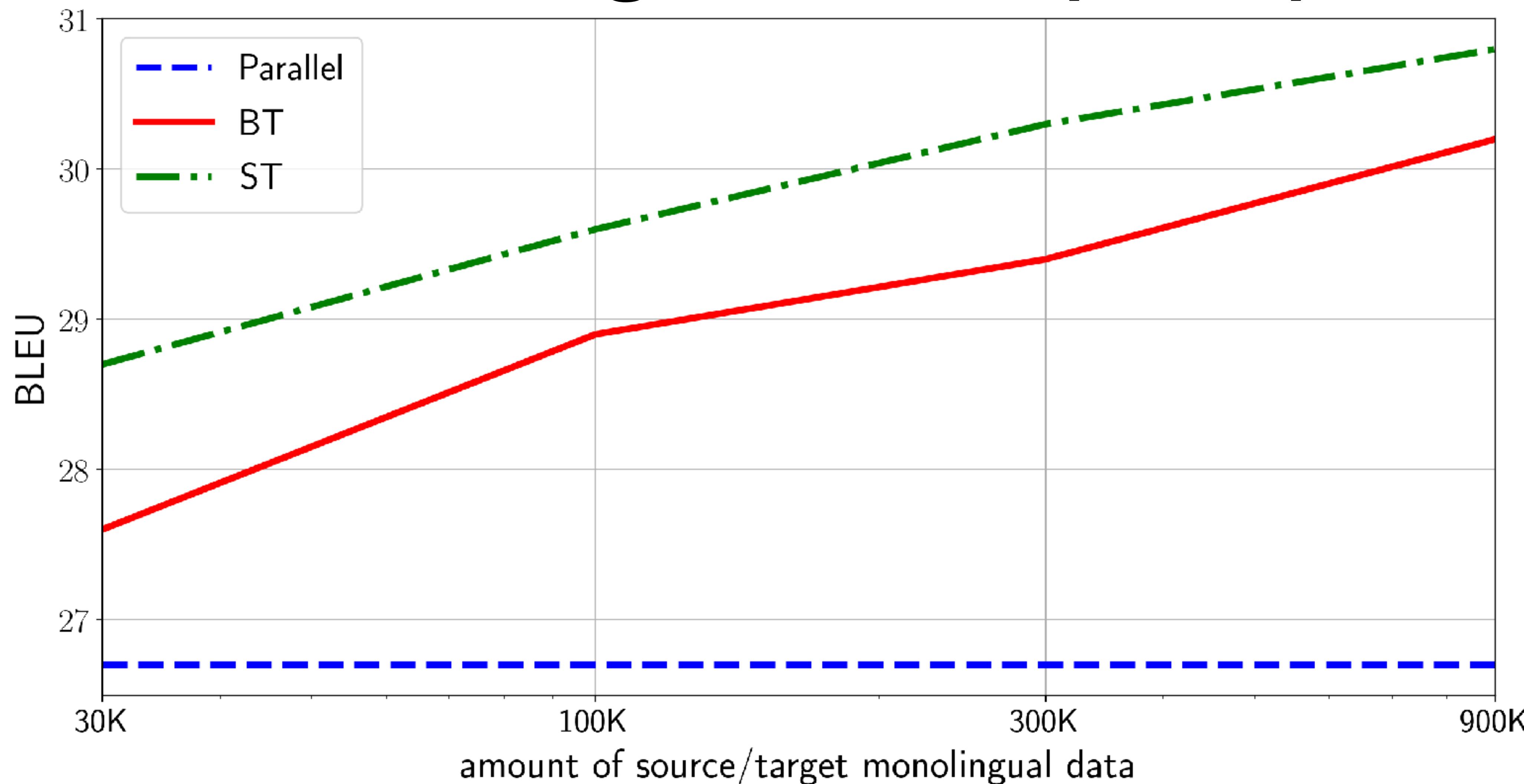
In-Domain Only VS. Mixed Domain

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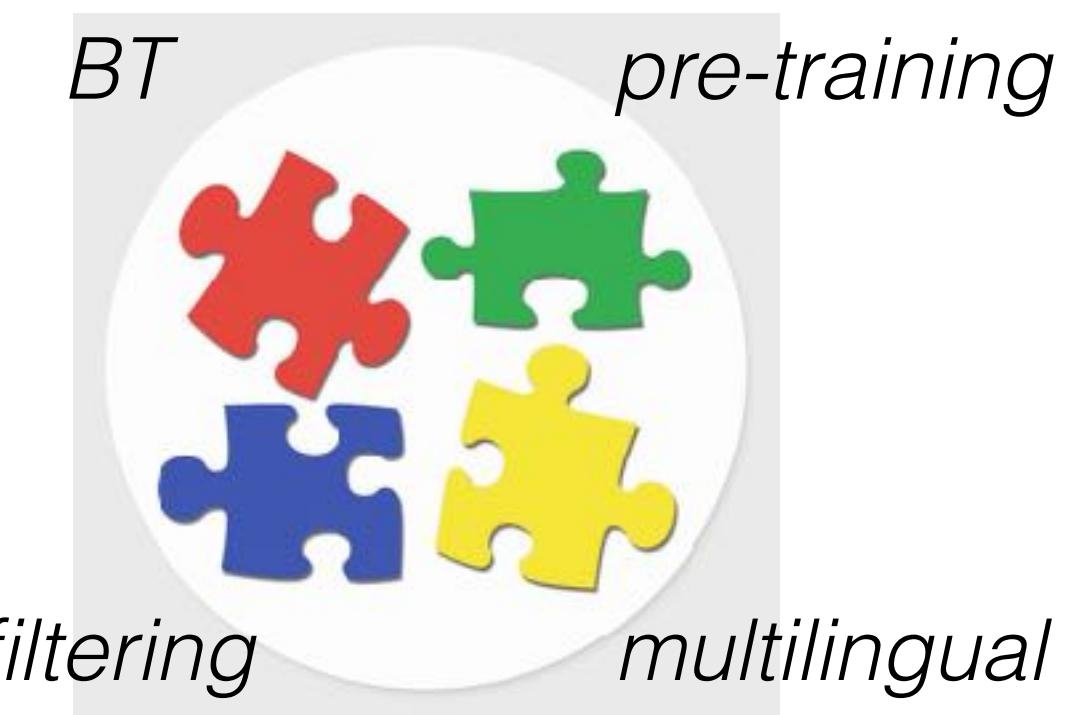
Conclusion

- STDM is particularly significant in low resource language pairs.

- Controlled setting helps studying STDM in isolation.

- ST is more robust than BT to STDM. We have already seen that combining ST & BT worked best in En-My.

- In practice, the influence of STDM depends on several factors, such as the amount of parallel and monolingual data, the domains, etc. In particular, if domains are not too distinct, STDM may even help regularizing!
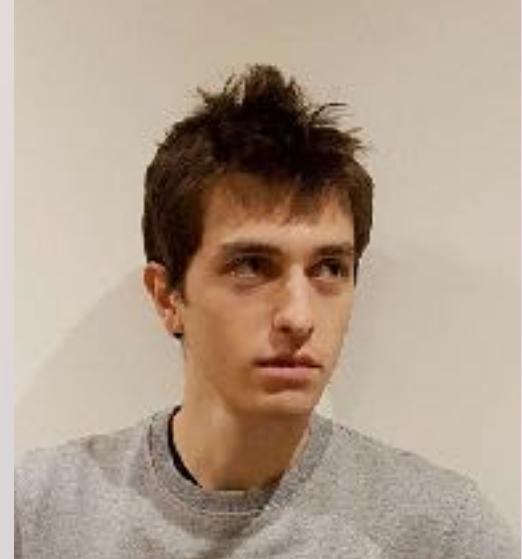
What I did not talk about: Filtering

- Data: extract clean version of common crawl.

- Learn a joint embedding space.

- Use approximate nearest neighbor methods to find closest matching sentence in embedding space.

- Translation quality on several languages is even higher than using actual existing bilingual + mono data!

BT *pre-training*

filtering *multilingual*

Schwenk et al. "CCMatrix: mining billions of high-quality parallel sentences on the WEB" arXiv:1911.04944 2019

Final Remarks

- Low resource MT is a good use case applications of several long standing ML problems: aligning domains, learning with less supervision, leveraging compositionality, etc.

- The importance and difficulty of data collection should not be under-estimated.

- A healthy cycle of research: data, modeling, analysis.

- Low resource MT key idea: use as many auxiliary tasks and data.

- Low resource MT requires lots of data and compute.

- Lots of open challenges.
  - Specific to low resource: dealing with all sorts of domain mismatch, learning from little data, quality of evaluation sets…
  - General of text generation: better use of context, common sense, striking a good trade-off between accuracy and speed, controllability, safety, biases…

Questions?
Вопросы?
¿Preguntas?
Domande?

Guillaume Lample

Ludovic Denoyer

Myle Ott
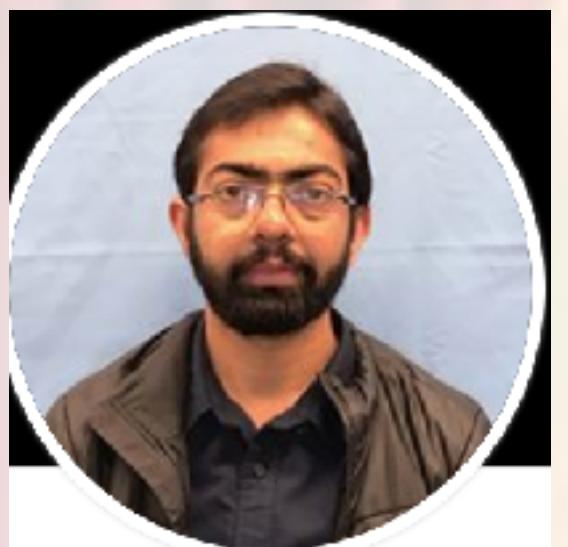
Peng-Jen Chen

Paco Guzmán
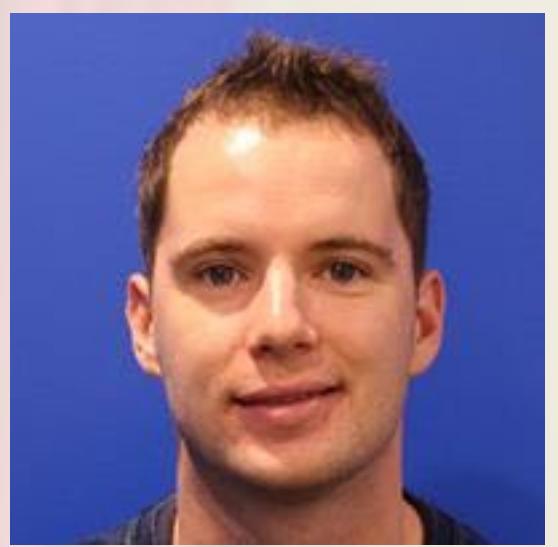
Jiajun Shen

Naman Goyal

Jiatao Gu

Xian Li

Matt Le

Alexis Conneau

Philipp Kohen
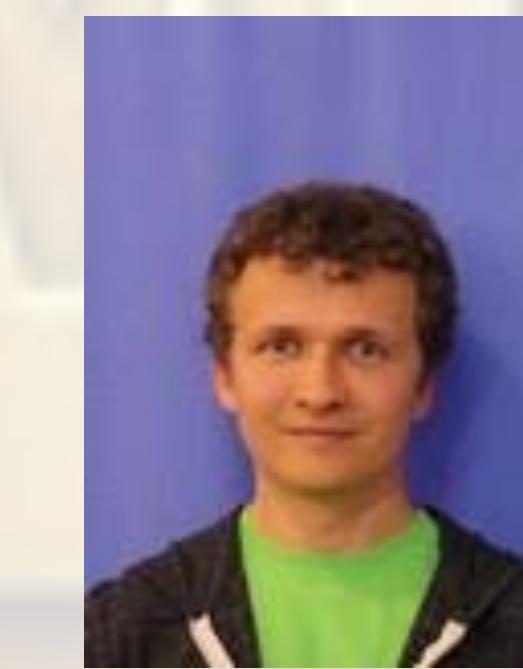
Juan-Miguel Pino

Hervé Jegou

Michael Auli

Junxian He

Sergey Edunov

Vishrav Chaudhary