

Low Resource Machine Translation

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

\$whoami

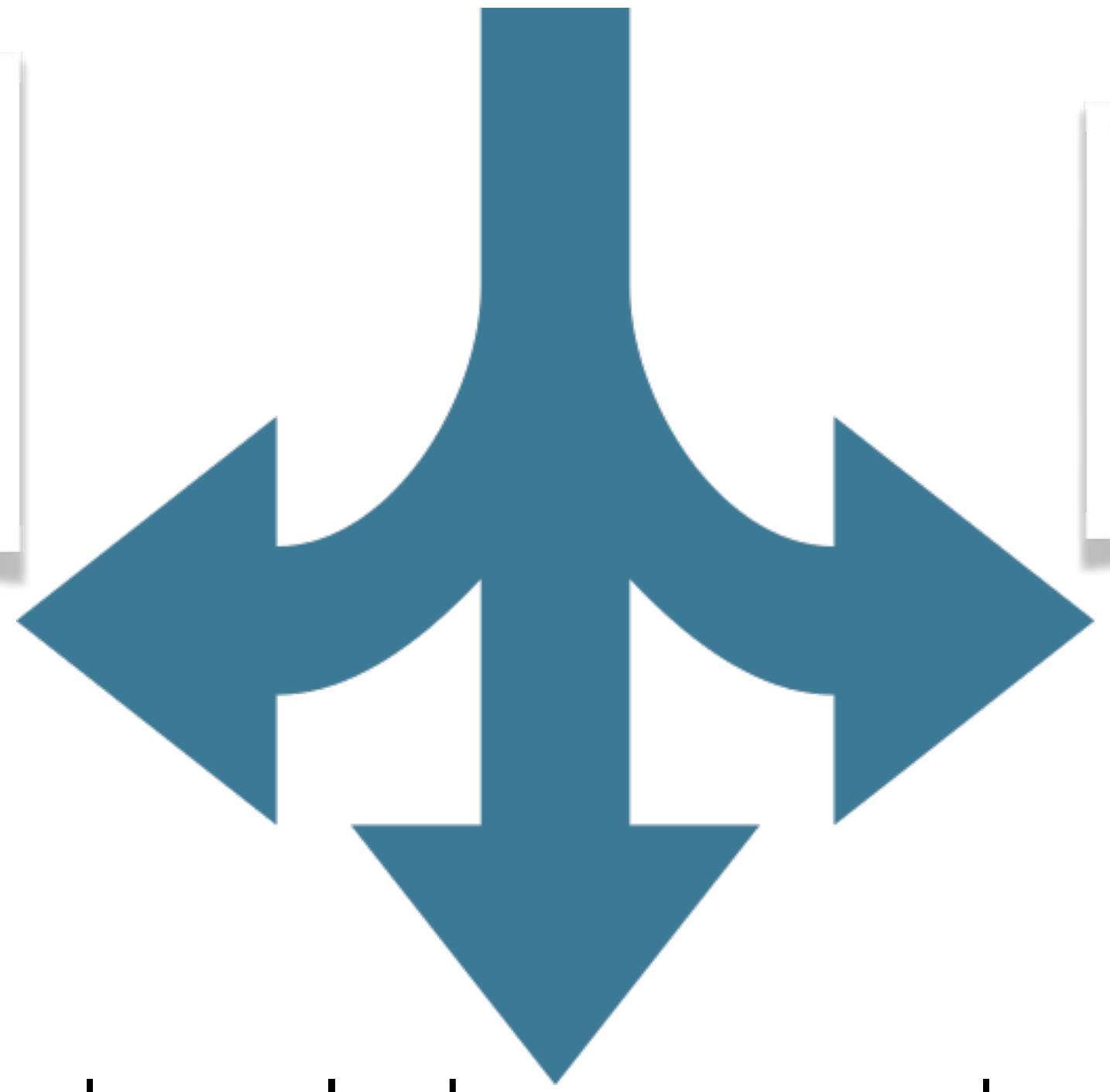
Goal: effective and efficient learning with limited supervision.

Phrase-Based & Neural Unsupervised Machine Translation

Guillaume Lample[†] Facebook AI Research Sorbonne Universités glample@fb.com	Myle Ott Facebook AI Research myleott@fb.com	Alexis Conneau Facebook AI Research Université Le Mans aconneau@fb.com
Ludovic Denoyer[†] Sorbonne Universités ludovic.denoyer@lip6.fr	Marc'Aurelio Ranzato Facebook AI Research ranzato@fb.com	

Leverage unlabeled data

- unsupervised learning
- semi-supervised learning

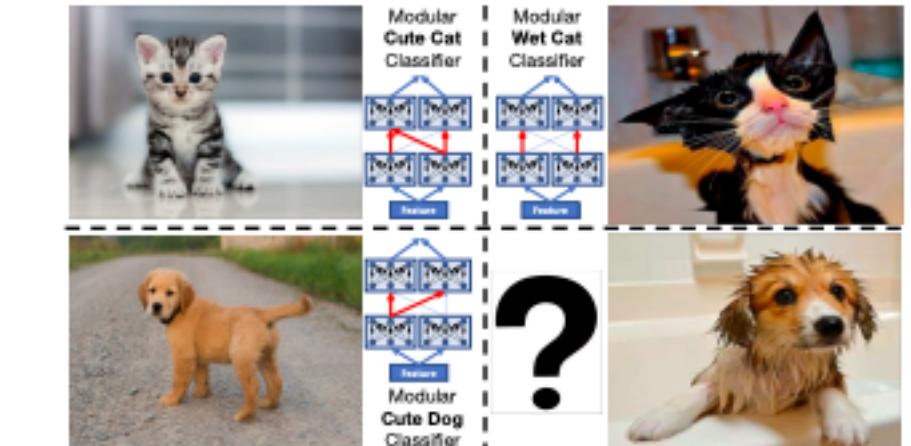


Task-Driven Modular Networks for Zero-Shot Compositional Learning

Senthil Purushwalkam* Carnegie Mellon University spurushw@andrew.cmu.edu	Maximilian Nickel Facebook AI Research maxn@fb.com	Abhinav Gupta Facebook AI Research gabhinav@fb.com
Marc'Aurelio Ranzato Facebook AI Research ranzato@fb.com		

Leverage inductive biases

- modularity
- compositionality



Leverage knowledge accrued on other tasks

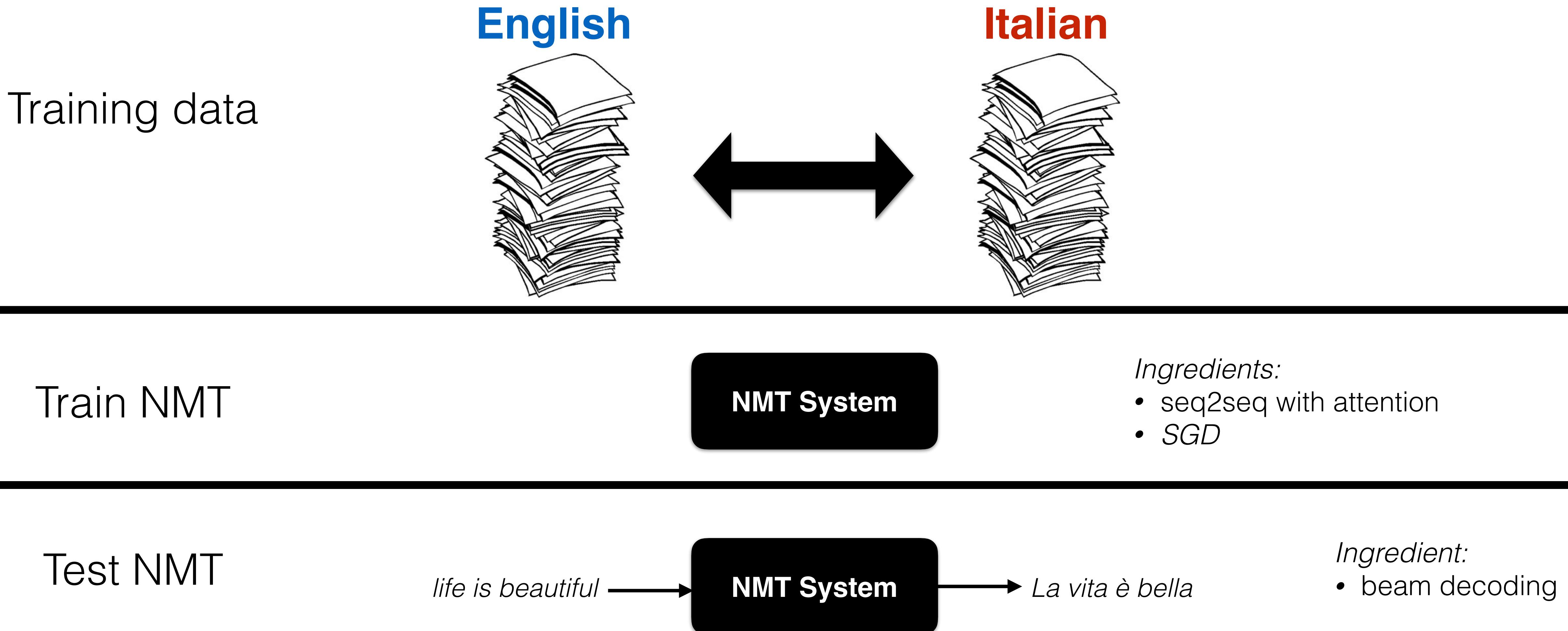
- continual learning
- few-shot learning
- meta-learning

EFFICIENT CONTINUAL LEARNING WITH MODULAR NETWORKS AND TASK-DRIVEN PRIORS

Tom Veniat
LIP6, Sorbonne Université, France
tom.veniat@lip6.fr

Ludovic Denoyer & Marc'Aurelio Ranzato
Facebook Artificial Intelligence Research
{denoyer, ranzato}@fb.com

Machine Translation

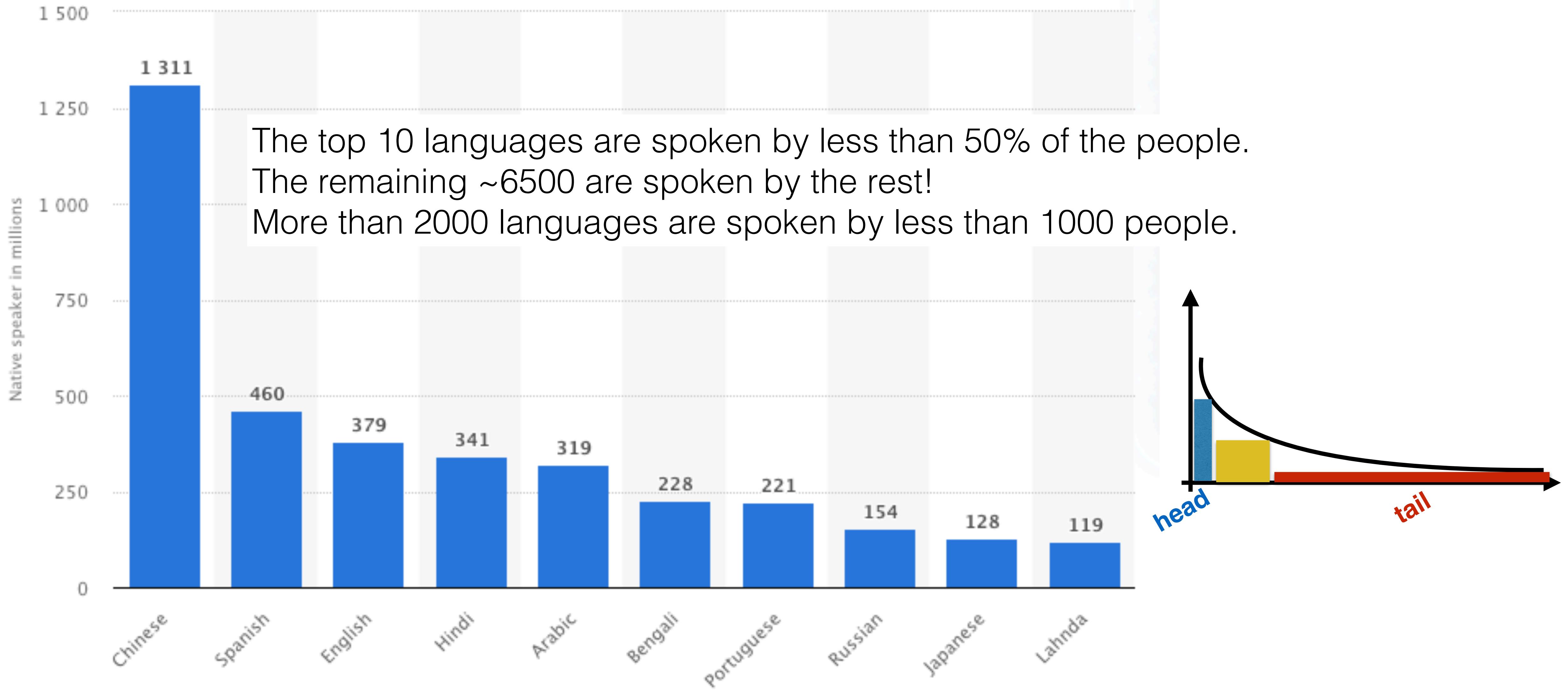


Some Stats

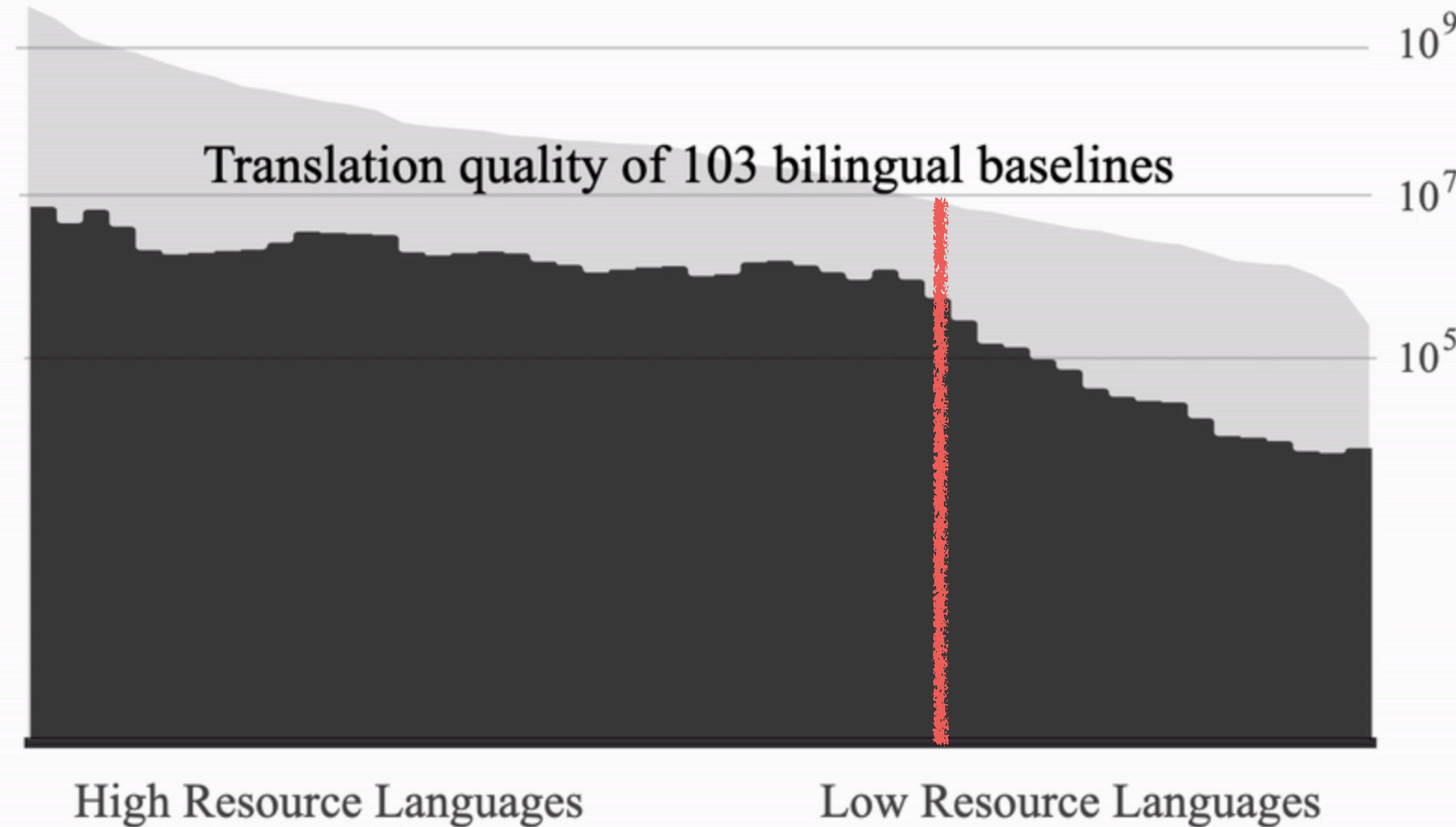
- 6000+ languages in the world
- 80% of the world population does not speak English
- Less than 5% of the people in the world are native English speakers.



The Long Tail of Languages



Data distribution over language pairs (X to English)



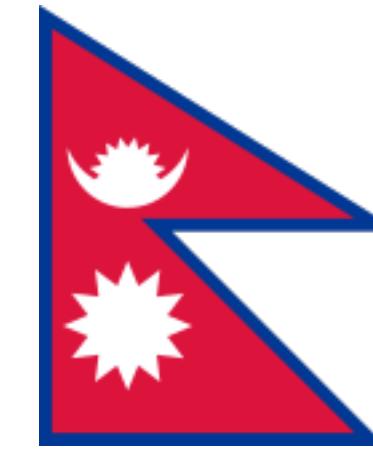
Machine Translation in Practice

Training data

English



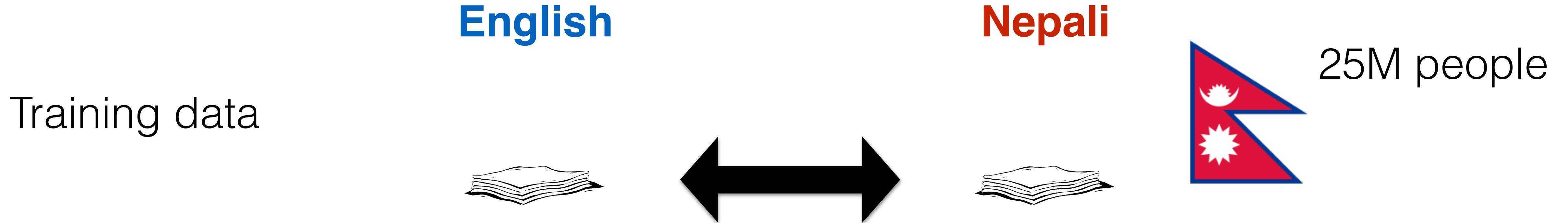
Nepali



25M people

Goal: Build MT system that can translate English news in Nepali.

Machine Translation in Practice



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

OPUS

... 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 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 some 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: 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
WikiMatrix v1	1	40.5k	1.0G	4.3M	xces en ne	en ne	tmx	moses	en ne	en ne			en ne		sample
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
ParaCrawl v7.1	2	92.1k	3.5M	4.4M	xces en ne	en ne	tmx	moses	en ne	en ne			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
GlobalVoices v2018q4	158	2.8k	0.1M	0.2M	xces en ne	en ne	tmx	moses	en ne	en ne	alg smt	dic	en ne		sample
tico-19 v2020-10-28	1	3.1k	80.5k	0.1M	xces en ne	en ne	tmx	moses	en ne	en ne	alg smt		en ne		sample
TED2020 v1	44	4.1k	73.9k	91.0k	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	6352	1.1M	1.1G	22.9M	1.1M		0.7M	0.7M							

OPUS

... 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 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 some 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: 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
WikiMatrix v1	1	40.5k	1.0G	4.3M	xces en ne	en ne	tmx	moses	en ne	en ne			en ne		sample
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
ParaCrawl v7.1	2	92.1k	3.5M	4.4M	xces en ne	en ne	tmx	moses	en ne	en ne			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
GlobalVoices v2018q4	158	2.8k	0.1M	0.2M	xces en ne	en ne	tmx	moses	en ne	en ne	alg smt	dic	en ne		sample
tico-19 v2020-10-28	1	3.1k	80.5k	0.1M	xces en ne	en ne	tmx	moses	en ne	en ne	alg smt		en ne		sample
TED2020 v1	44	4.1k	73.9k	91.0k	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	6352	1.1M	1.1G	22.9M	1.1M		0.7M	0.7M							



... the open parallel corpus

OPUS
with a
collection

The C
Conten

Search

linguistic annotat
ackage. We used se

**Relative to the number of parameters O(100M+),
there are very few parallel sentences to learn from.**

There are multiple domains and varying quality of translation.

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)

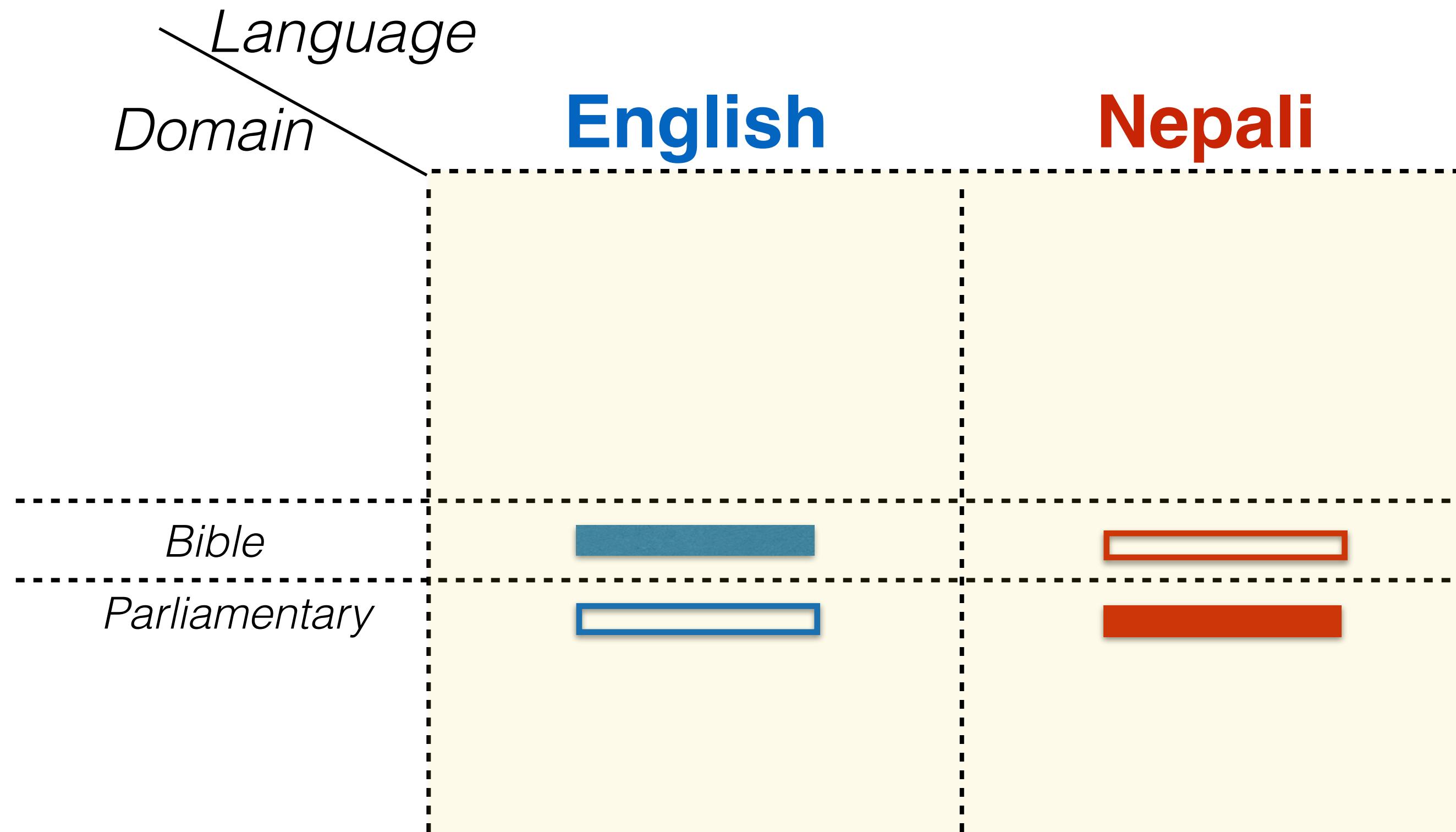
corpus	doc's	sent's	en tokens	ne tokens	XCES/XML	raw	TMX	Moses	mono	raw	ud	alg	dic	freq	other files
WikiMatrix v1	1	40.5k	1.0G	4.3M	xces en ne	en ne	tmx	moses	en ne	en ne			en ne		sample
JW300 v1	4663	0.4M	6.5M	5.5M	xces en ne	en ne			en ne	en ne			en ne		sample
wikimedia v20190628	1	2.8k	7.7M	1.2M	xces en ne	en ne	tmx	moses	en ne	en ne	alg smt	dic	en ne		sample
ParaCrawl v7.1	2	92.1k	3.5M	4.4M	xces en ne	en ne	tmx	moses	en ne	en ne			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
GlobalVoices v2018q4	158	2.8k	0.1M	0.2M	xces en ne	en ne	tmx	moses	en ne	en ne	alg smt	dic	en ne		sample
tico-19 v2020-10-28	1	3.1k	80.5k	0.1M	xces en ne	en ne	tmx	moses	en ne	en ne	alg smt		en ne		sample
TED2020 v1	44	4.1k	73.9k	91.0k	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	6352	1.1M	1.1G	22.9M	1.1M		0.7M	0.7M							

Machine Translation in Practice



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

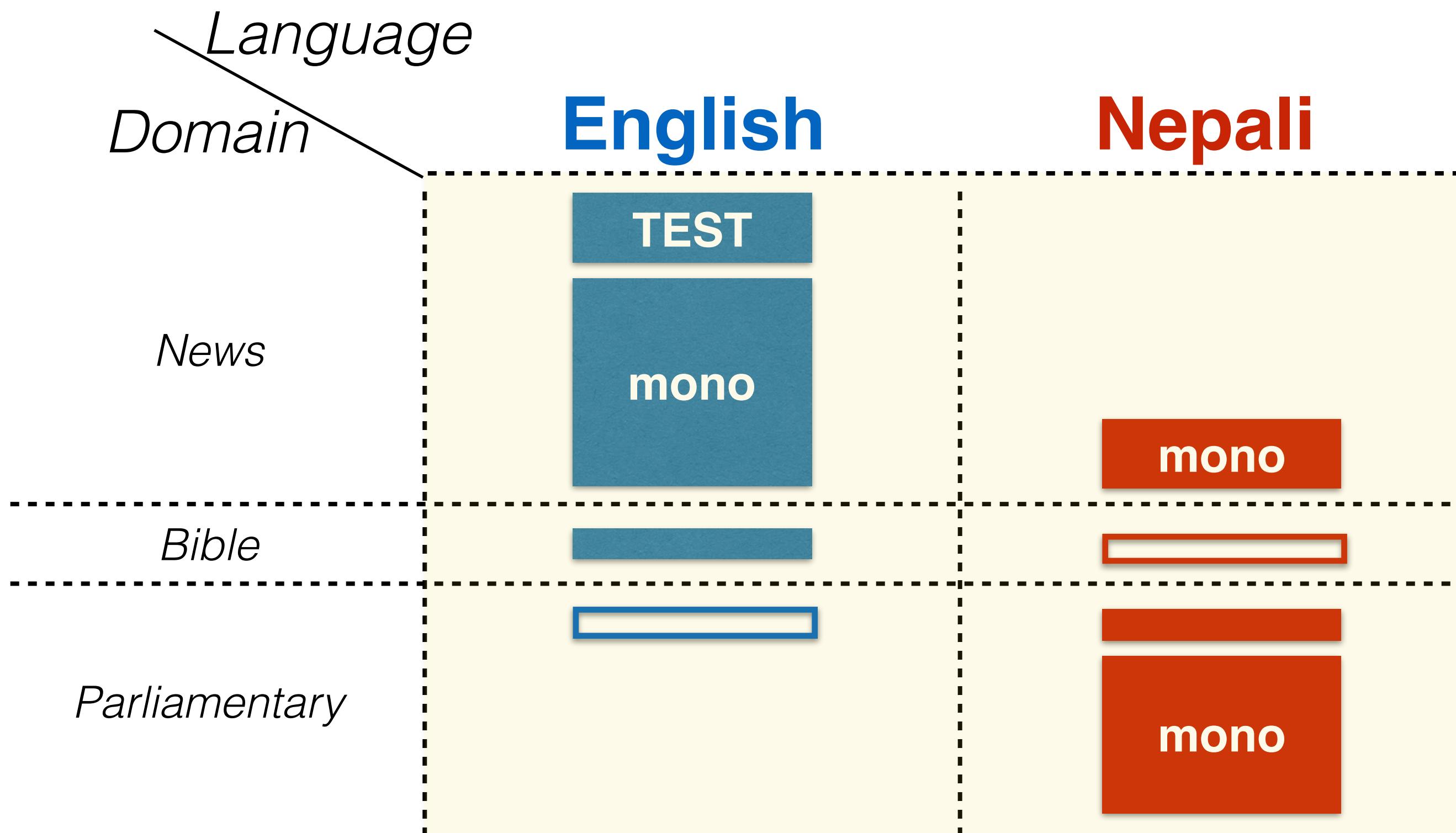
Machine Translation in Practice



Let's represent original text with filled boxes and (human) translations with empty rectangles.

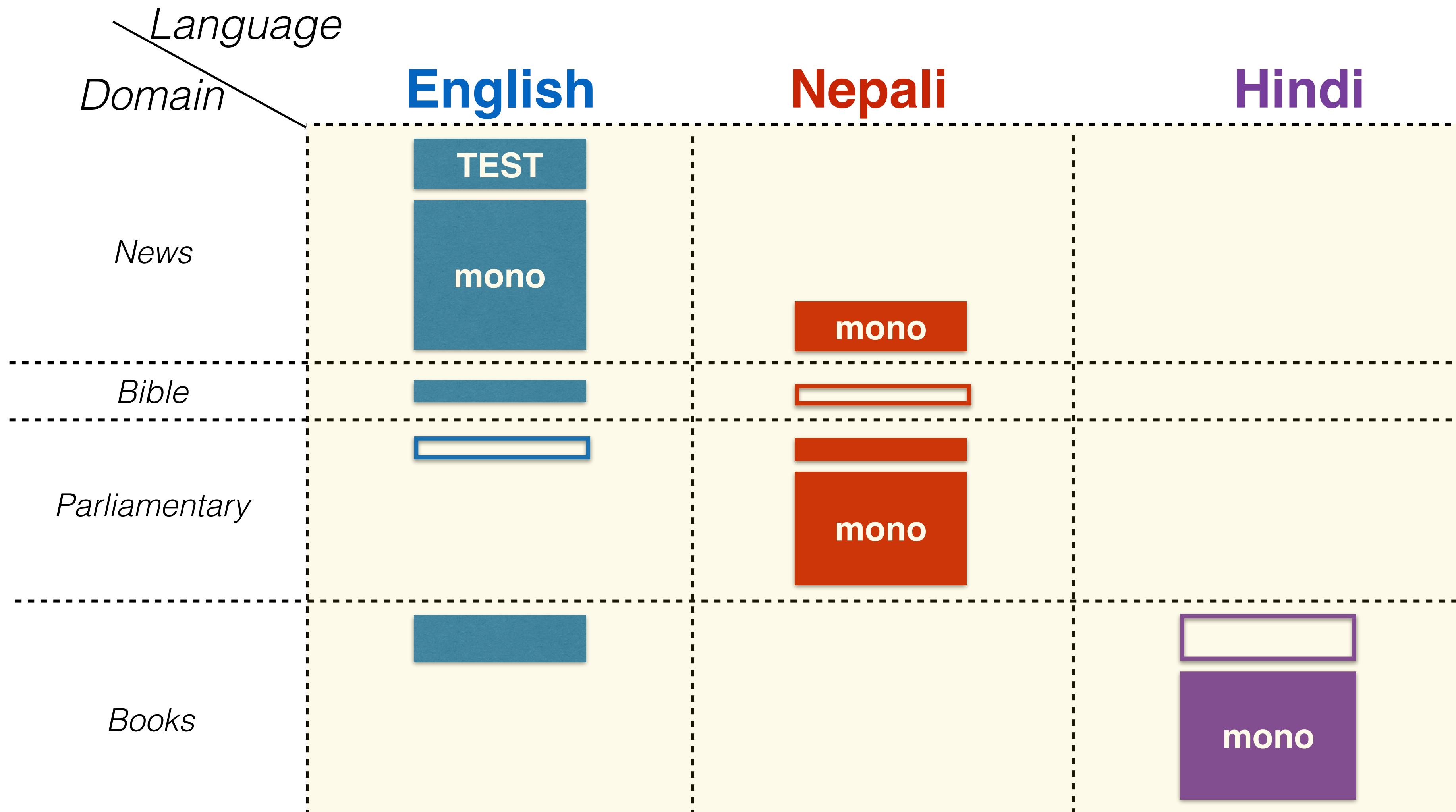
- Some parallel data originates in the source, some in the target language.
- Source and target domains may not match.

Machine Translation in Practice



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

Machine Translation in Practice



- 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.

English

Nepali

Hindi

Sinhala

Bengali

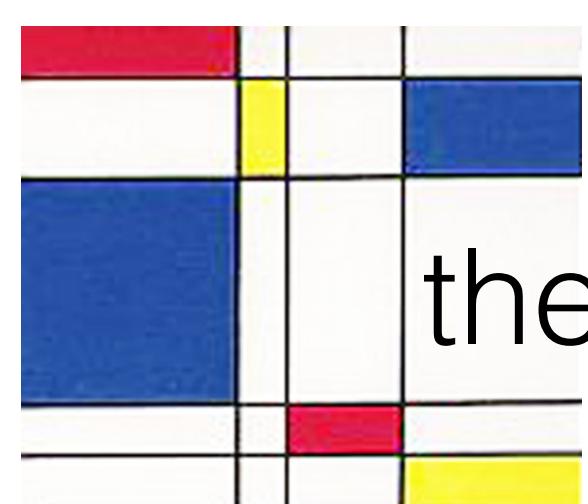
Spanish

Tamil

Gujarati

TEST

Domain



the *Mondrian* like learning setting! ...



Low Resource Machine Translation

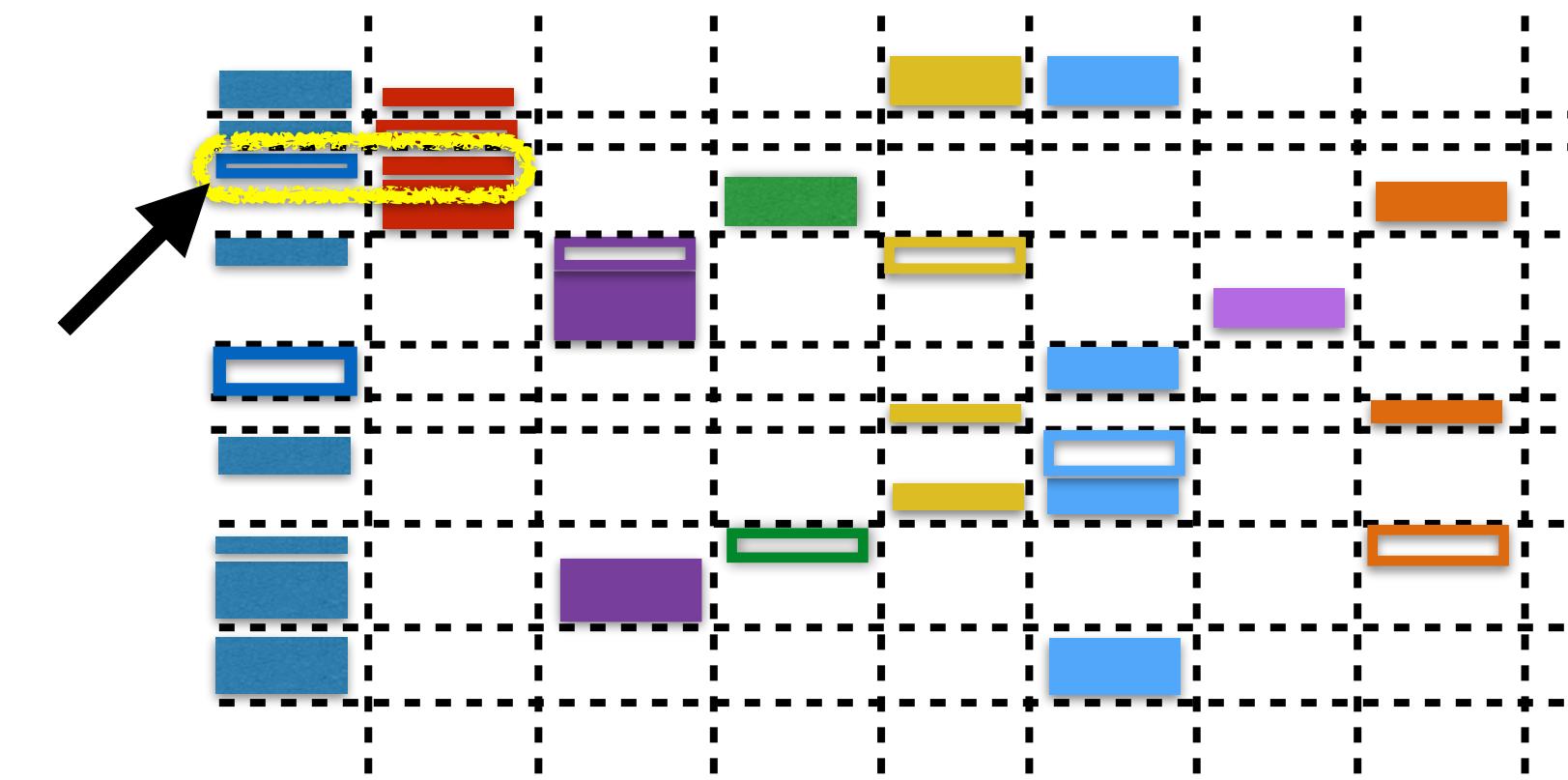
Loose definition: A language pair can be considered **low resource** when the number of parallel sentences is in the order of 10,000 or less.

Challenges:

- Datasets
 - Sourcing data to train on
 - High quality evaluation datasets
- Metrics
 - Human evaluation
 - Automatic evaluation
- Modeling
 - Learning paradigm
 - Domain adaptation
 - Generalization
- Scaling



Low-resource MT is about **large**-scale learning!

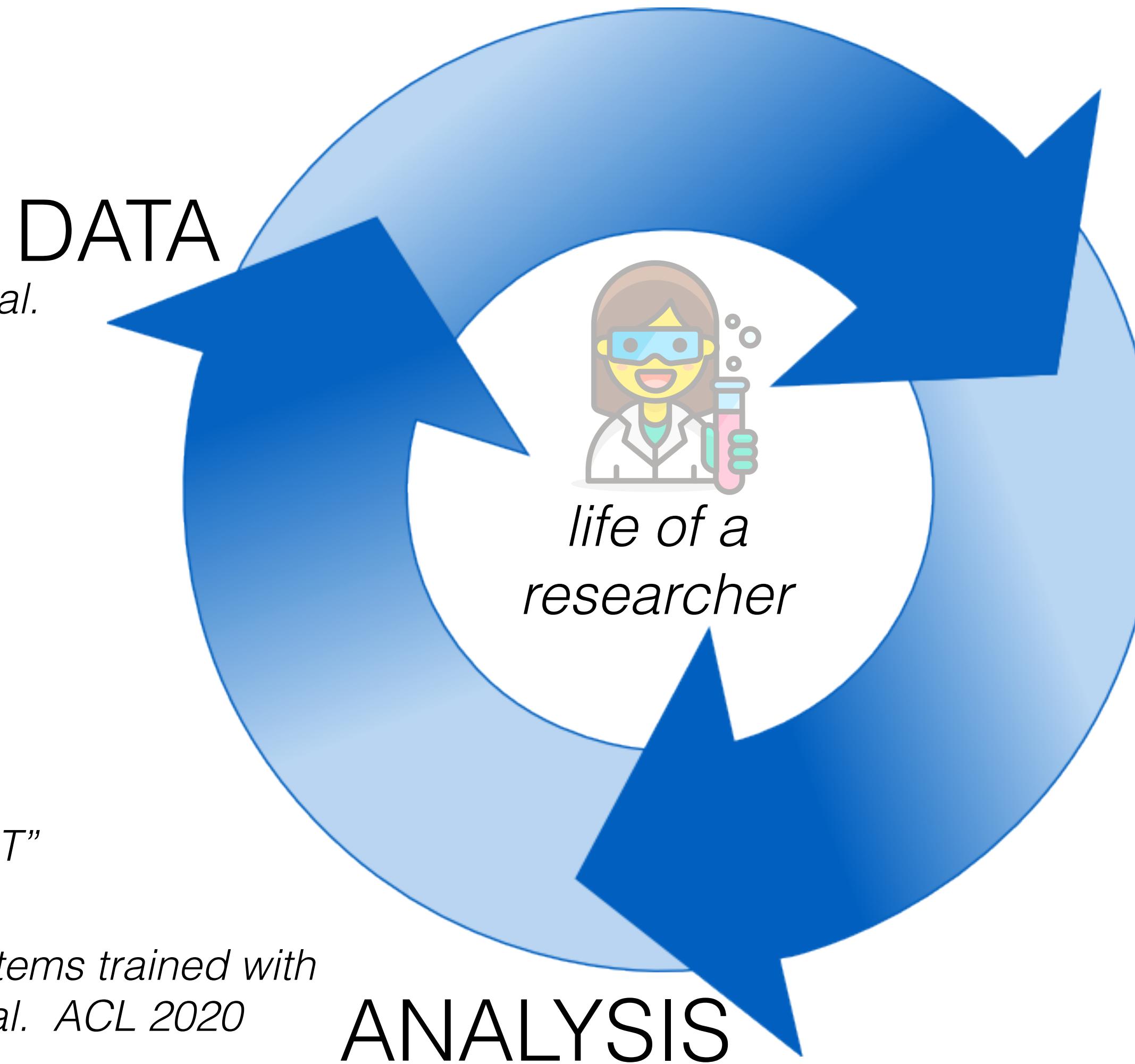


General ML Tip: whenever you lack supervised data, come up with auxiliary tasks or even fantasize it.

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.

The Cycle of Research



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

*“Analyzing uncertainty in NMT”
Ott et al. IJCAI 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. EACL 2021

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

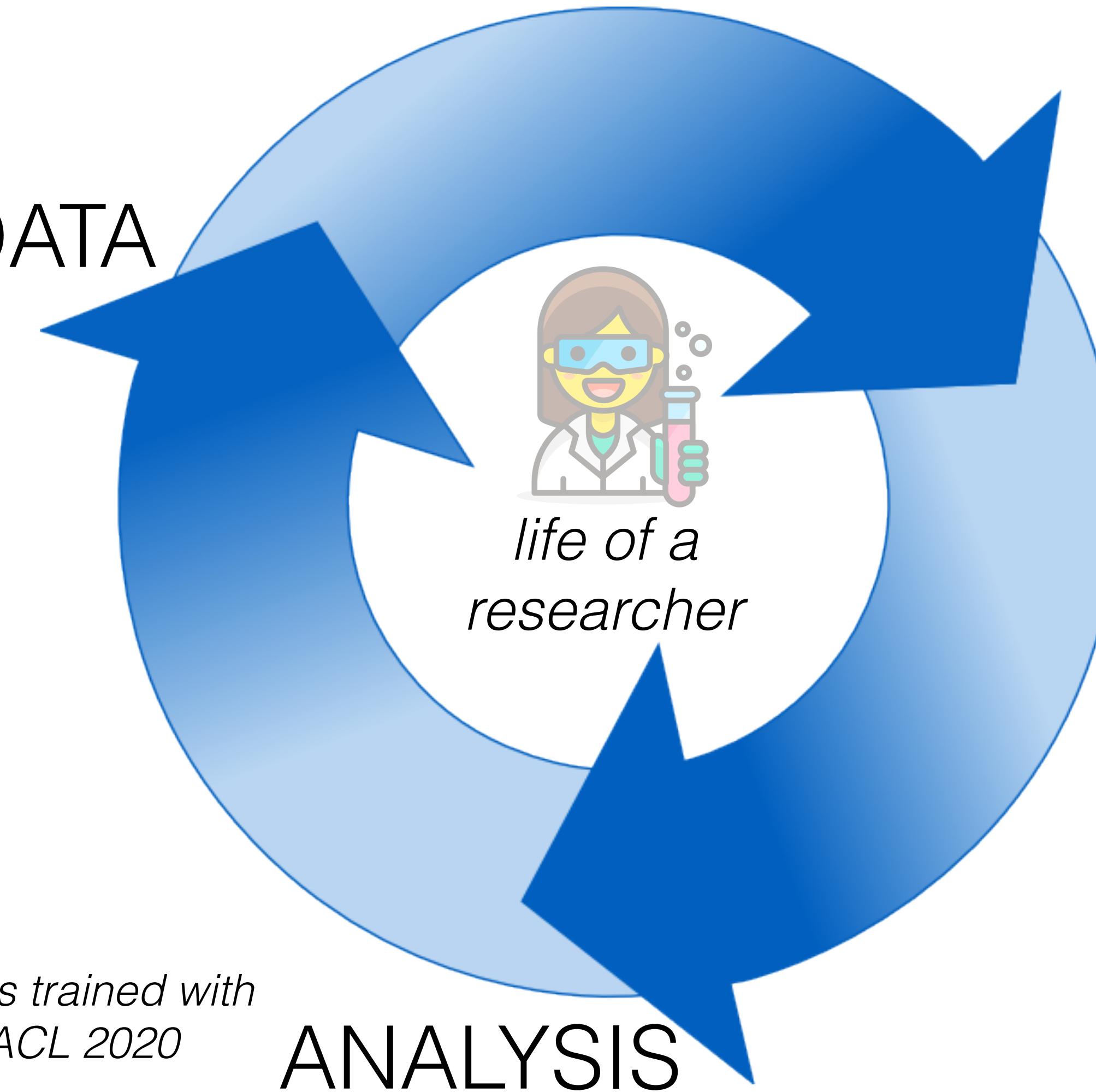
“Investigating Multilingual NMT Representations at Scale” Kudugunta et al., EMNLP 2019

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

“Beyond English-Centric MMT” Fan et al., arXiv 2010:11125 2020

“Complete Neural MMT” Freitag et al., arXiv 2010:10239 2020

The Cycle of Research



“The FLoRes evaluation for low resource MT:...” Guzmán, Chen et al.
‘EMNLP 2019

“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. EACL 2021

“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

“Beyond English-Centric MMT” Fan et al., arXiv 2010:11125 2020

“Complete Neural MMT” Freitag et al., arXiv 2010:10239 2020

Outline

- What is low-resource MT and why is it important?
- **ML perspective on low resource MT**
- Case studies:
 - Unsupervised MT
 - En-Ne
 - En-My
- Perspectives

English

Nepali

Hindi

Sinhala

Bengali

Spanish

Tamil

Gujarati

Domain

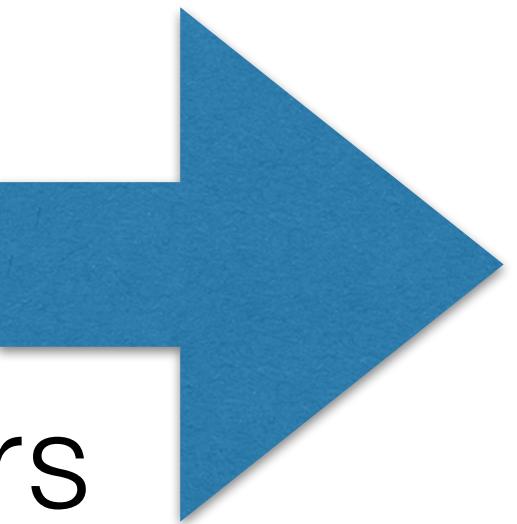
TEST



ML Perspective

NLP/MT

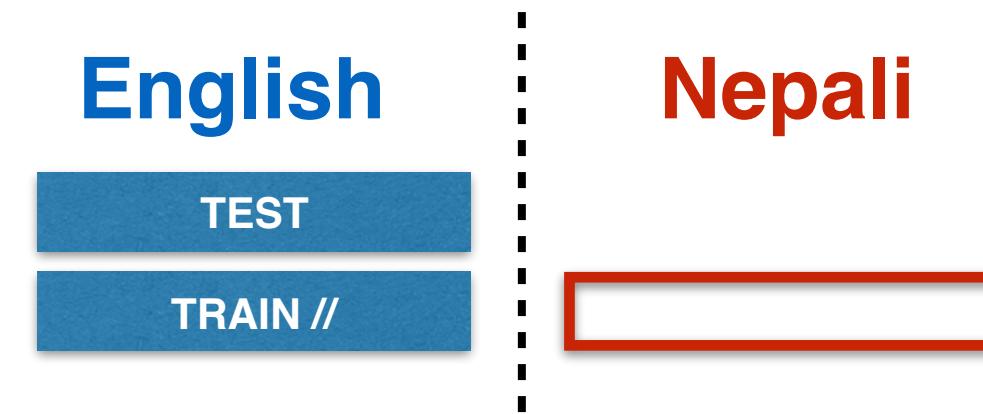
- Parallel dataset
- Monolingual data
- Multiple language pairs
- Multiple domains



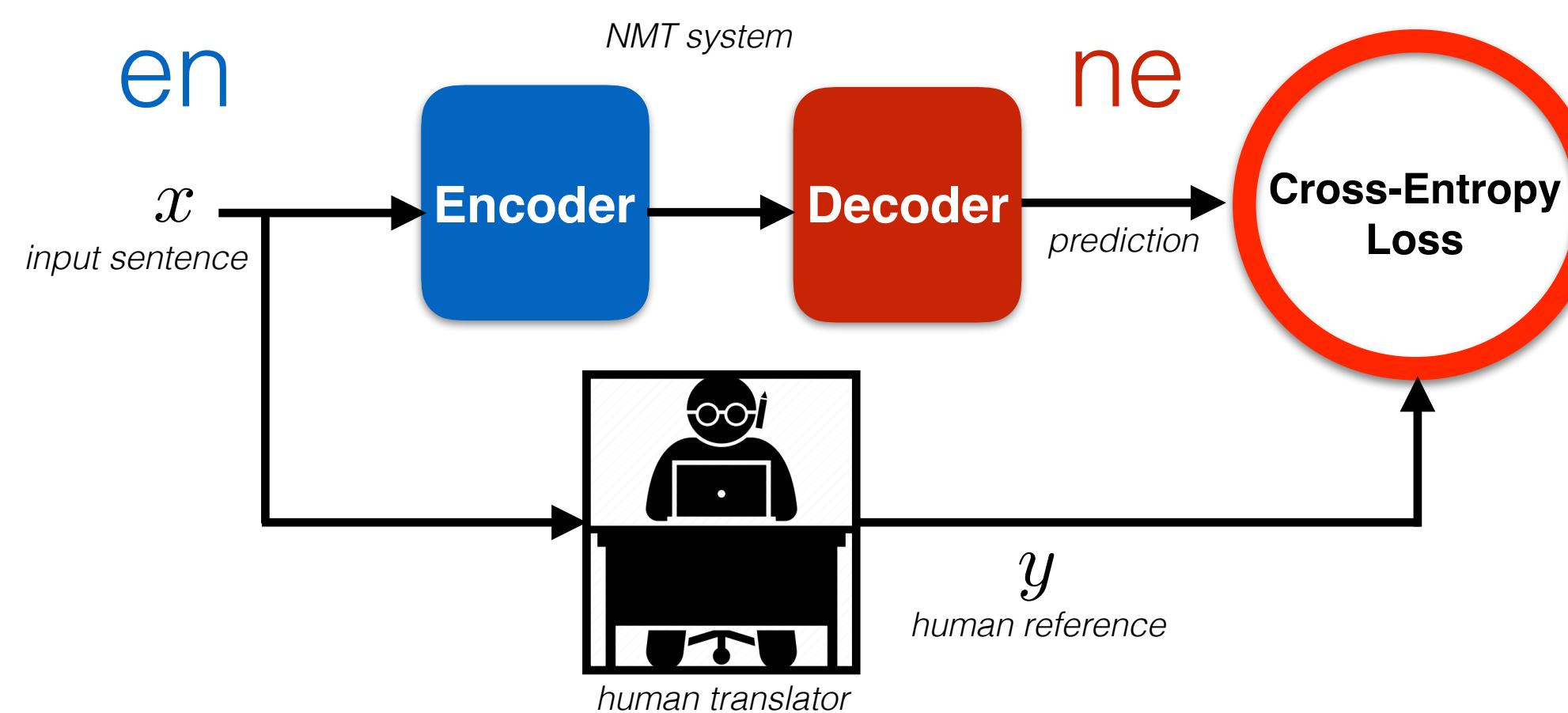
ML

- Supervised learning
- Semi-supervised learning
- Multi-task/multi-modal learning
- Domain adaptation

Supervised Learning



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



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

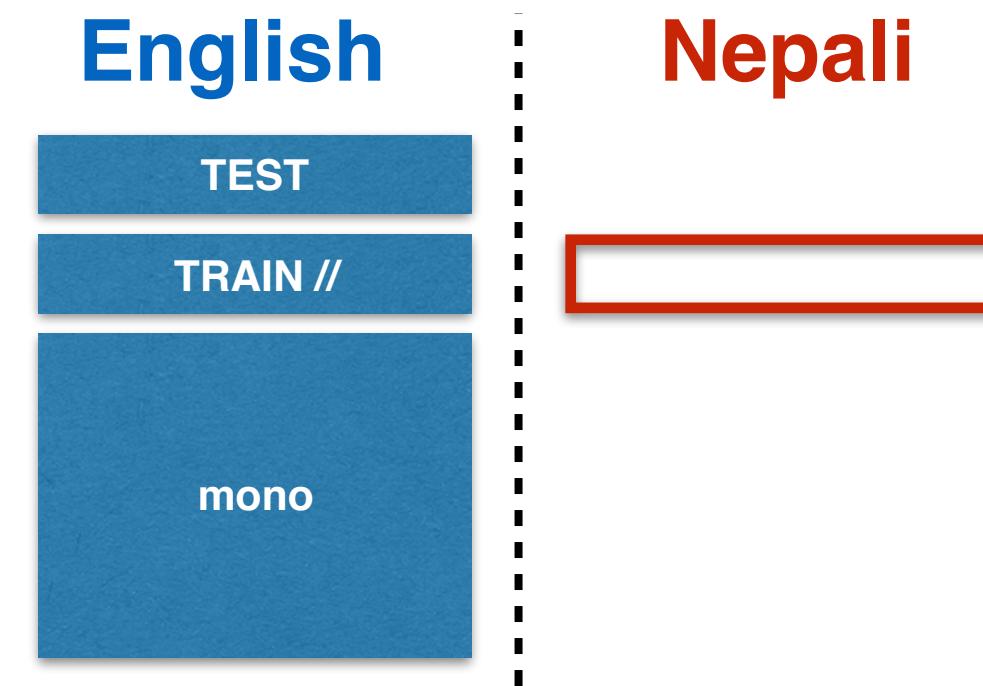
usual attention-based transformer

Regularize the model using:
- dropout [1]
- label smoothing [2]

[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

Semi-Supervised Learning (DAE)

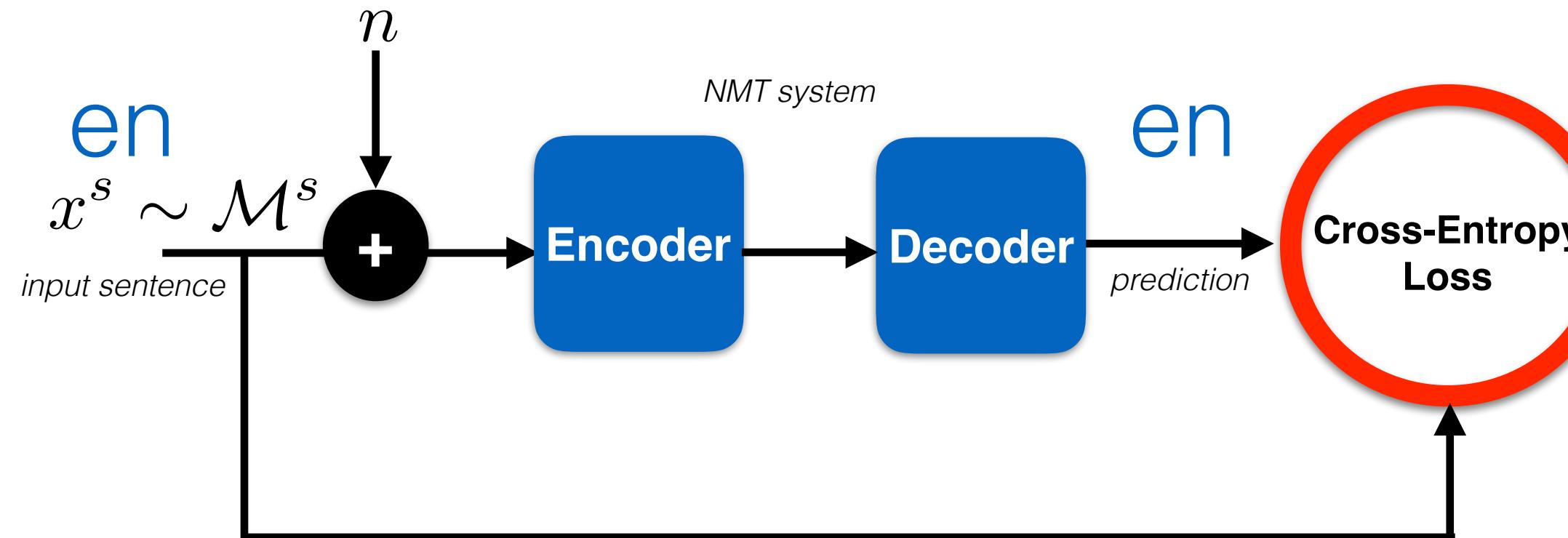


$$\mathcal{D} = \{(x, y)_i\}_{i=1,\dots,N}$$

$$\mathcal{M}^s = \{x_j^s\}_{j=1,\dots,M_s}$$

Additional source side monolingual data.

Idea: model $p(x)$ with a denoising auto-encoder.



Noise: word drop, swap, etc.

E.g.: *The cat the on sat mat.*
The sat cat on the.

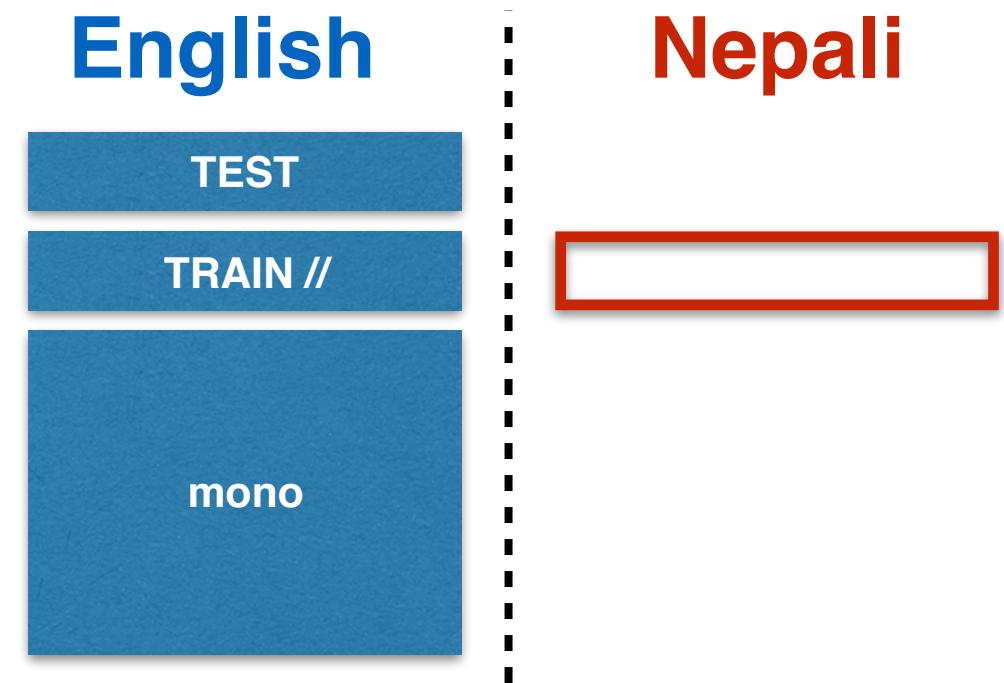
Learning Framework: DAE

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

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

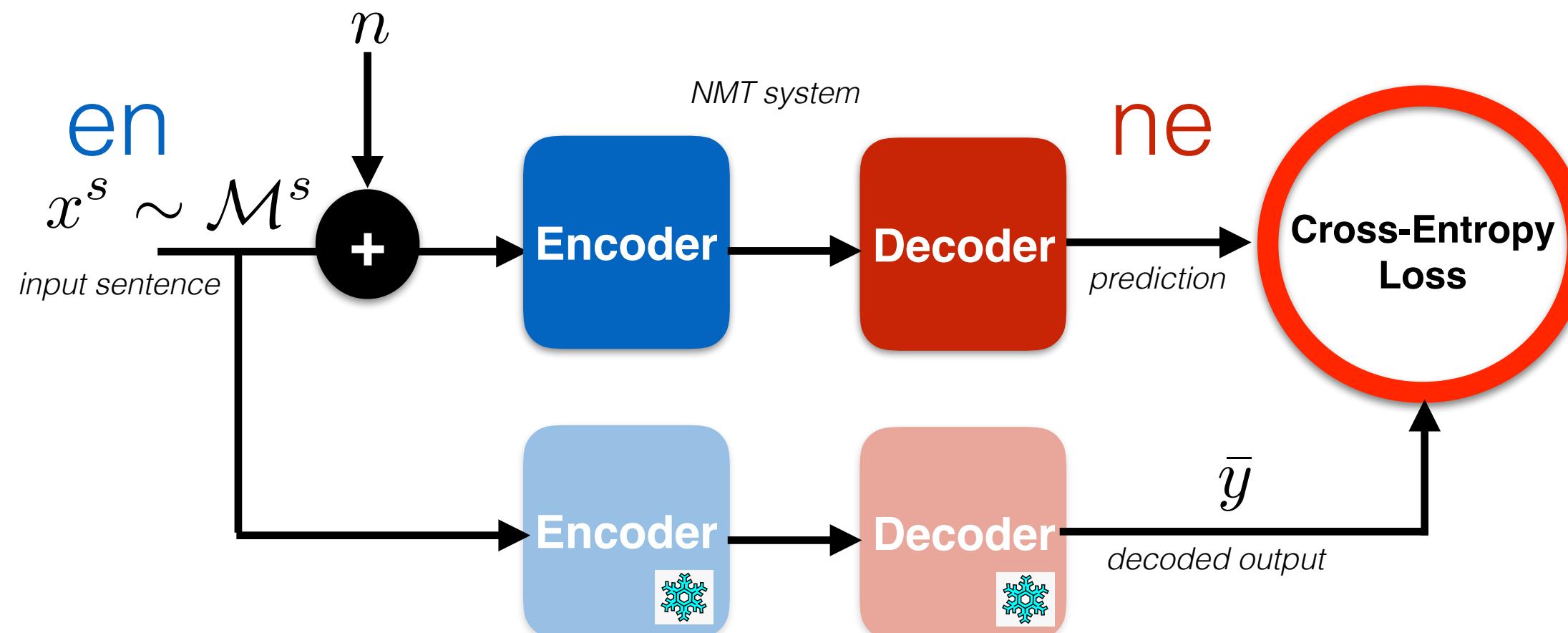
Vincent et al. "Stacked denoising auto-encoders:..." JMLR 2010
Liu et al. "Multilingual denoising pretraining for NMT" arXiv:2001.08210 2020

Semi-Supervised Learning (ST)



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

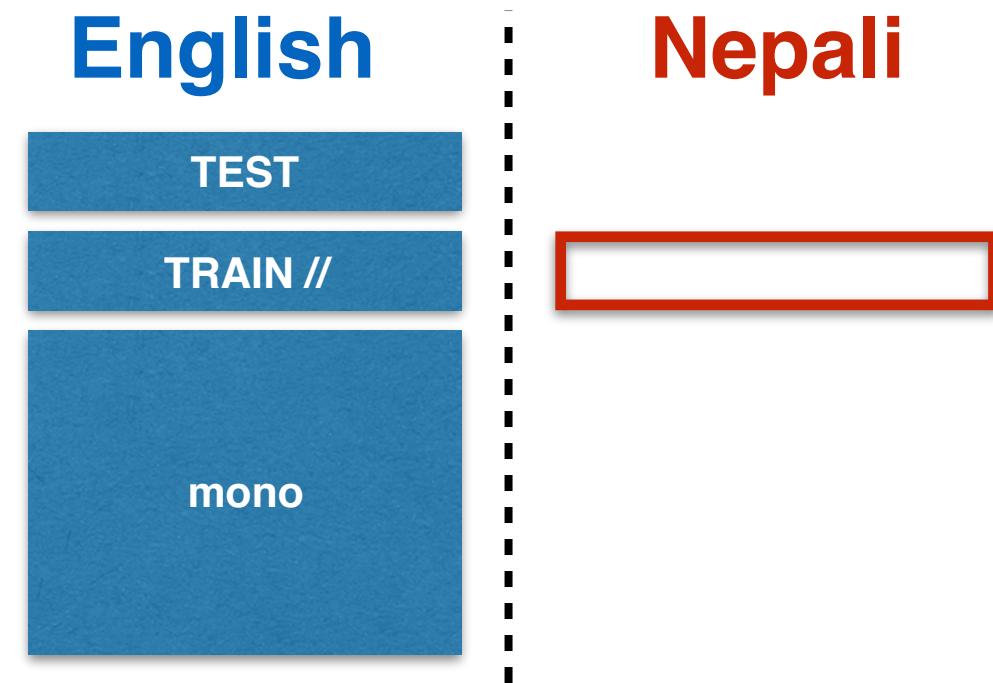
Idea: predict missing labels.



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

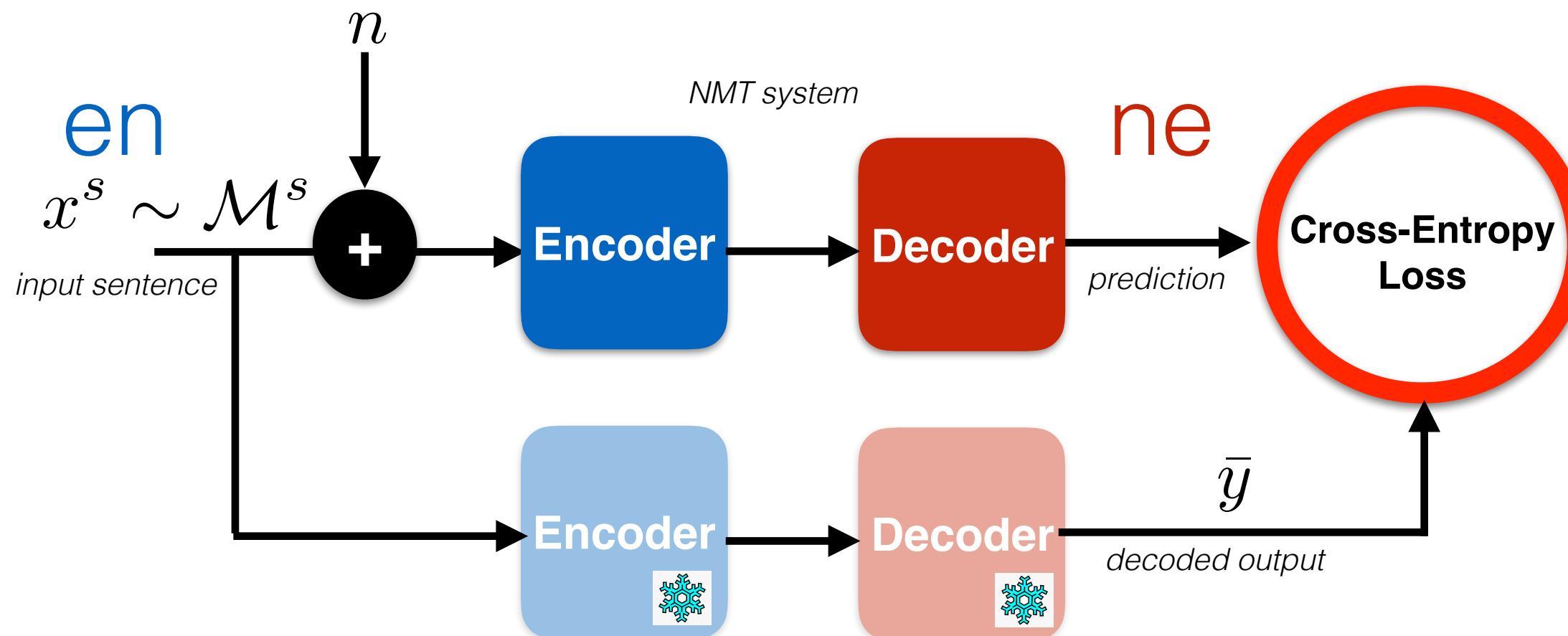
Key elements: decoding and training noise.

Semi-Supervised Learning (ST)



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

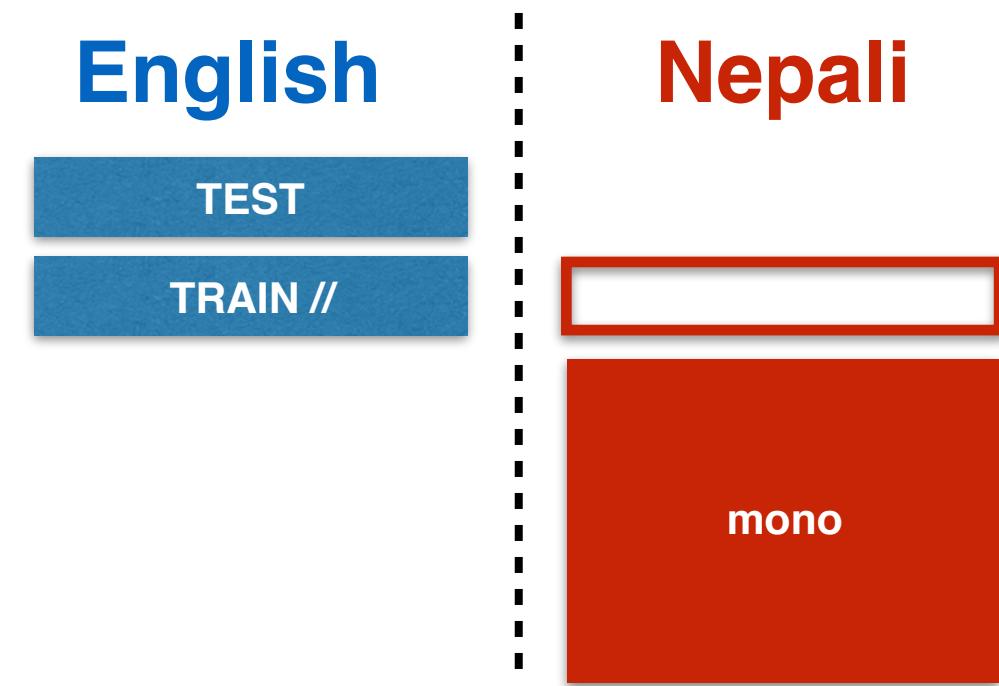
Idea: predict missing labels.



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

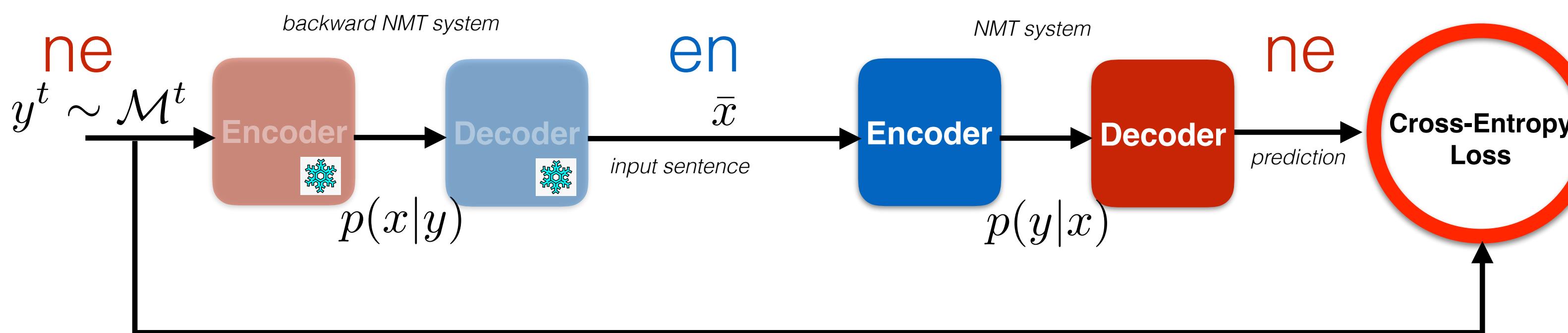
Key elements: decoding and training noise.

Semi-Supervised Learning (BT)



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

Additional target side monolingual data.



Adding target-side monolingual data.

Two benefits:

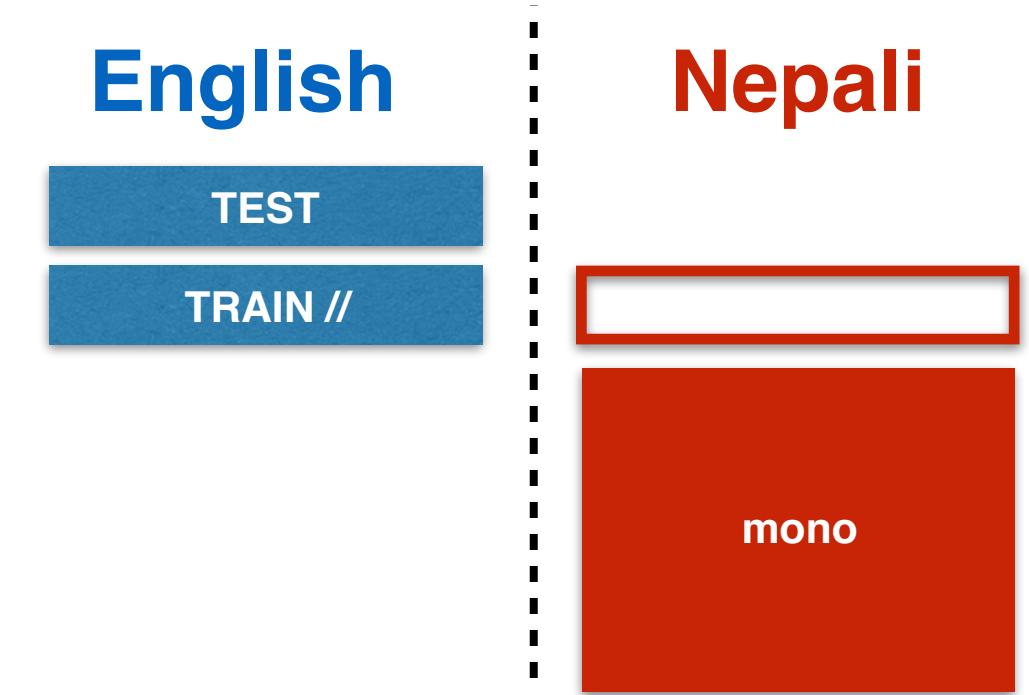
- Decoder learns a good language model.
- Better generalization via data augmentation.
- Unlike ST, target is correct but input is not.

Learning Framework:
Back-Translation (BT).

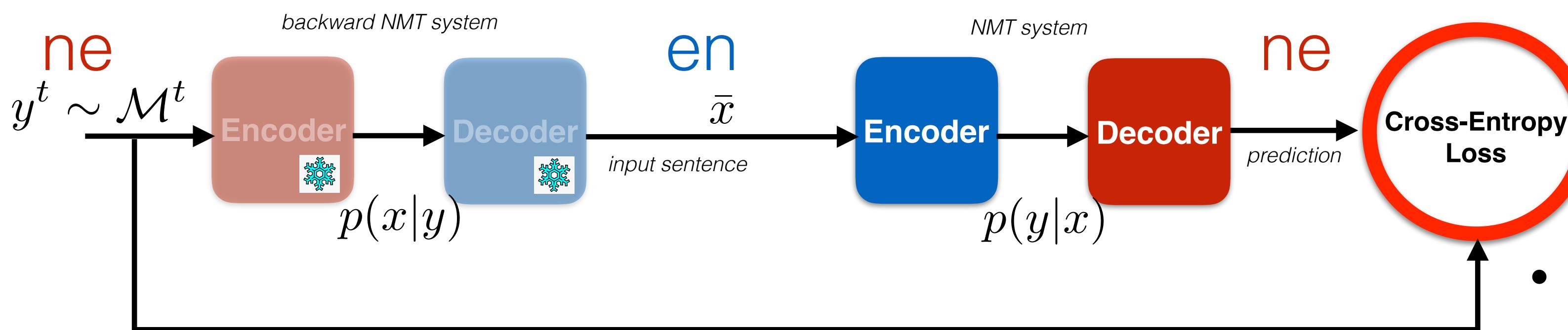
$$\mathcal{L}^{BT}(\theta) = -\log p(y|\bar{x})$$

$$\mathcal{L}(\theta) = \mathcal{L}^{\text{sup}}(\theta) + \lambda \mathcal{L}^{BT}(\theta)$$

Semi-Supervised Learning (BT)



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



Adding target-side monolingual data.

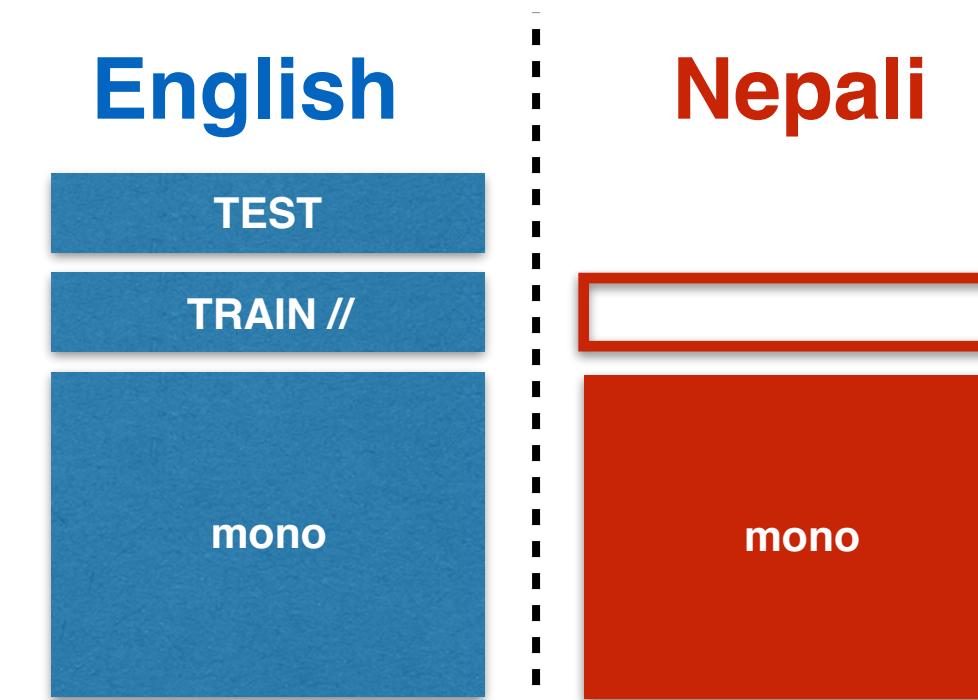
Two benefits:

- Decoder learns a good language model.
- Better generalization via data augmentation.
- Unlike ST, target is correct but input is not.

ALGORITHM

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

Semi-Supervised Learning (ST+BT)

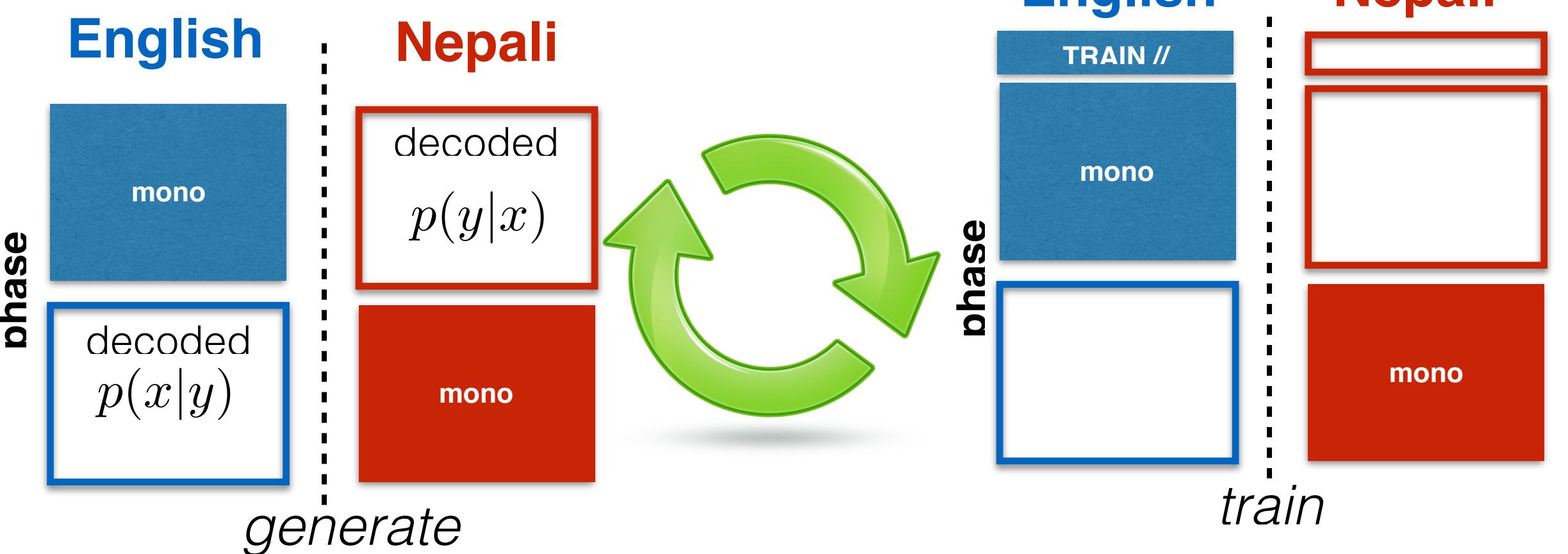


$$\mathcal{D} = \{(x, y)_i\}_{i=1,\dots,N}$$

$$\mathcal{M}^t = \{y_k^t\}_{k=1,\dots,M_t}$$

$$\mathcal{M}^s = \{x_j^s\}_{j=1,\dots,M_s}$$

Additional source & target side monolingual data.

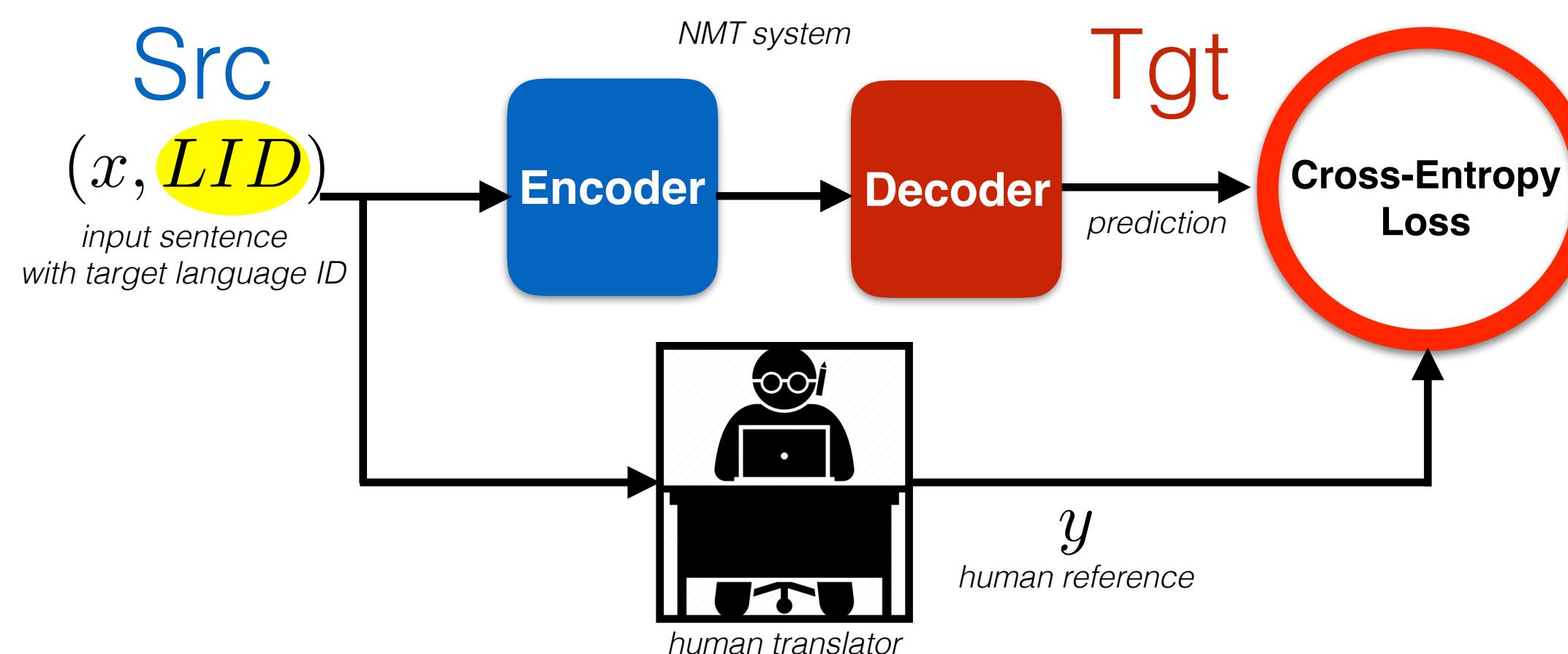


ALGORITHM

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

$$\mathcal{L}^{\text{total}}(\theta) = -\log p(y|x) - \lambda_1 \log p(y^t|\bar{x}^t) - \lambda_2 \log p(\bar{y}^s|x^s)$$

Multi-Task/Multi-Modal Learning



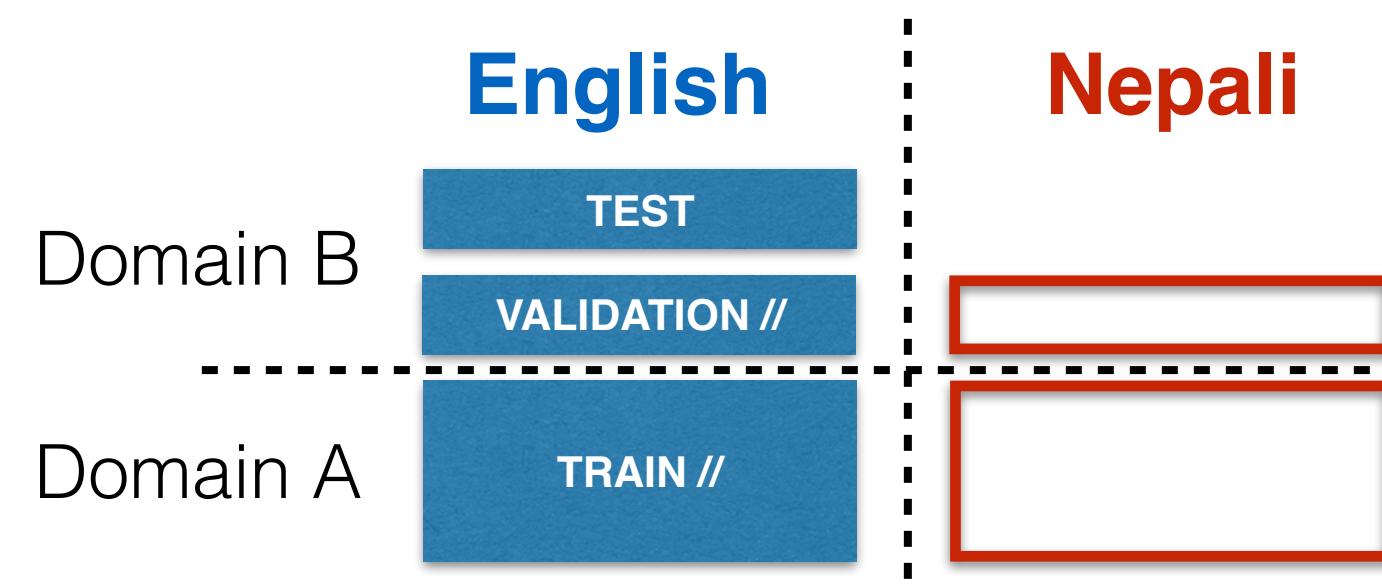
Learning Framework: Multilingual Training

Share encoder and decoder across 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.

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

Johnson et al. “Google’s multilingual NMT system...” ACL 2017
Aharoni et al. “Massively multilingual NMT” ACL 2019
Fan et al. “Beyond English centric MMT” arXiv 2020

Domain Adaptation



Learning Framework: Fine-tuning.

Train on domain A.

Finetune on domain B by continuing training for a little bit on the validation set.



Several basic learning approaches can be used and combined to tackle low-resource MT, thanks to the end-to-end learning based approach (there is nothing too specific to the task and language pair).

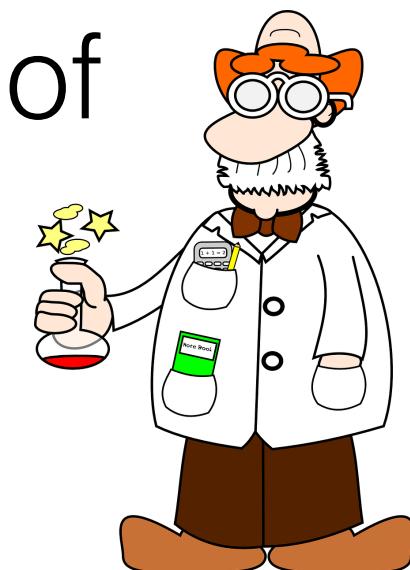
General ML Tip #1: keep architecture and learning algorithm as general as possible, let the model learn from data what the task is about.

General ML Tip #2: data augmentation is often a very powerful way to improve generalization.

What's so special about MT? The symmetry of the prediction task. This is what is exploited to fantasize data in BT.

Conclusion so far...

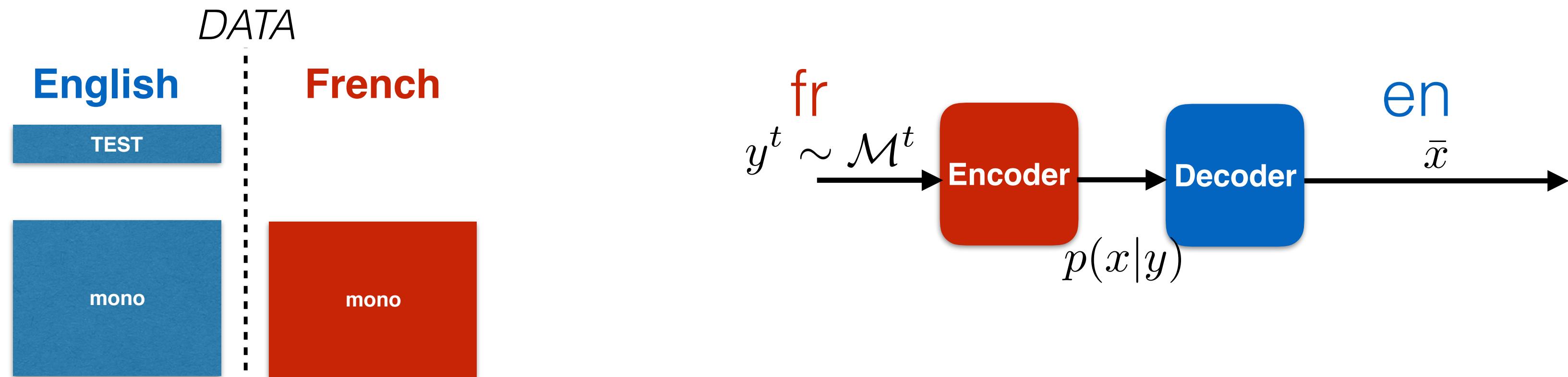
- Even assuming no domain effect, there are lots of training paradigms depending on the available data.
- Complex interaction: generalization, domain, language pair, capacity, amount of parallel and monolingual data, etc.
- In general, DAE pretraining, (iterative) BT and multi-lingual training perform 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.



Outline

- What is low-resource MT and why is it important?
- ML perspective on low resource MT
- **Case studies:**
 - Unsupervised MT
 - En-Ne
 - En-My
- Perspectives

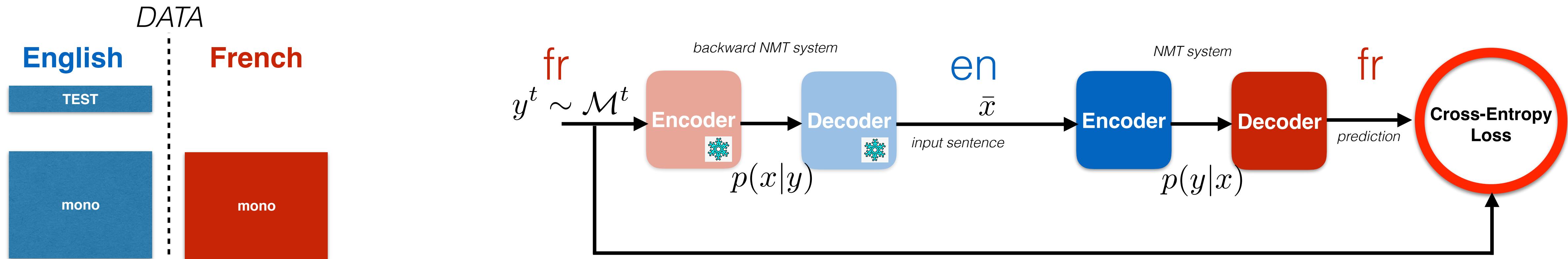
Case Study #1: Unsupervised MT



$$\mathcal{M}^t = \{y_k^t\}_{k=1,\dots,M_t}$$

$$\mathcal{M}^s = \{x_j^s\}_{j=1,\dots,M_s}$$

Case Study #1: Unsupervised MT



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

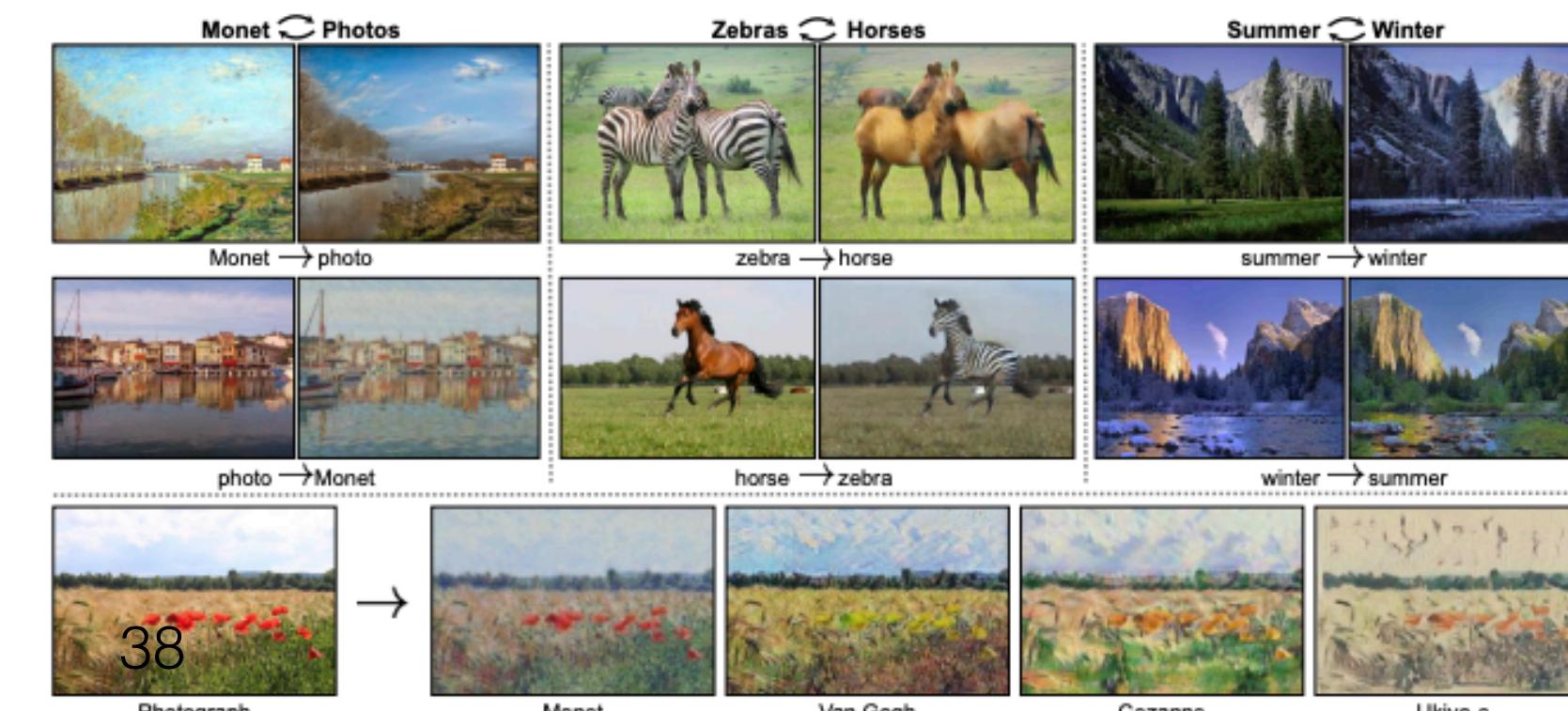
$$\mathcal{M}^s = \{x_j^s\}_{j=1,..,M_s}$$

...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

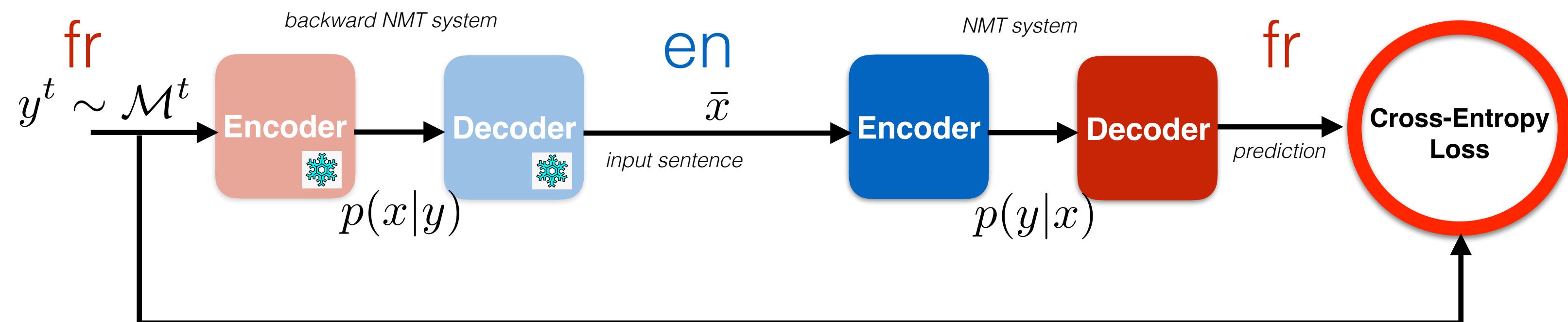
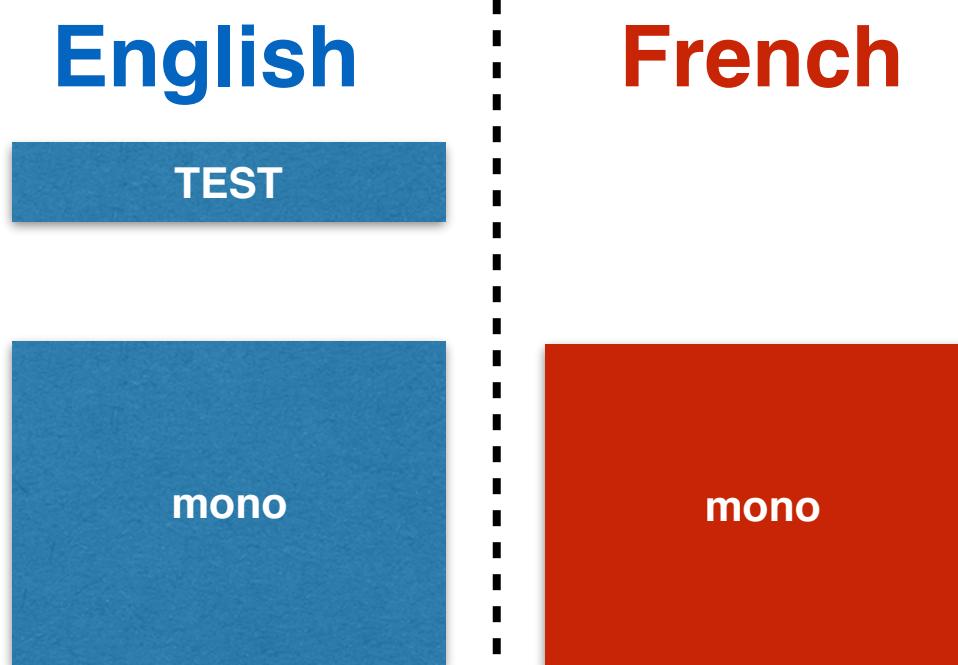
Jun-Yan Zhu* Taesung Park* Phillip Isola Alexei A. Efros
Berkeley AI Research (BAIR) laboratory, UC Berkeley



Problem: lack of constraints on \bar{x}

Case Study #1: Unsupervised MT

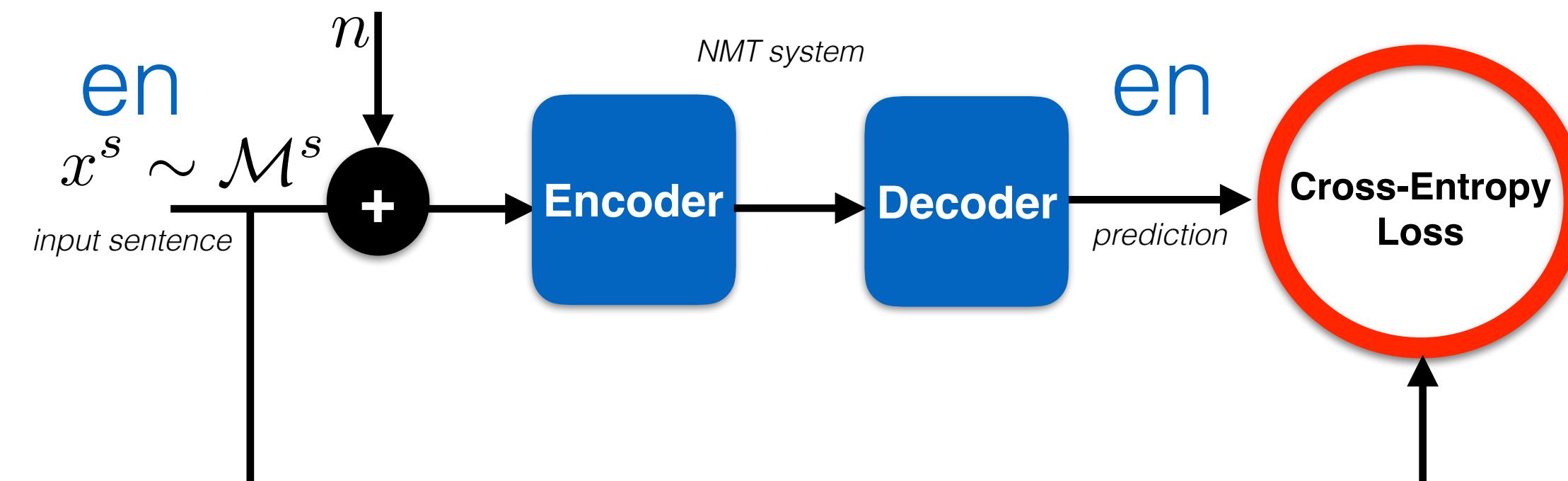
DATA



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

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

DAE makes sure decoder outputs fluently in the desired language.

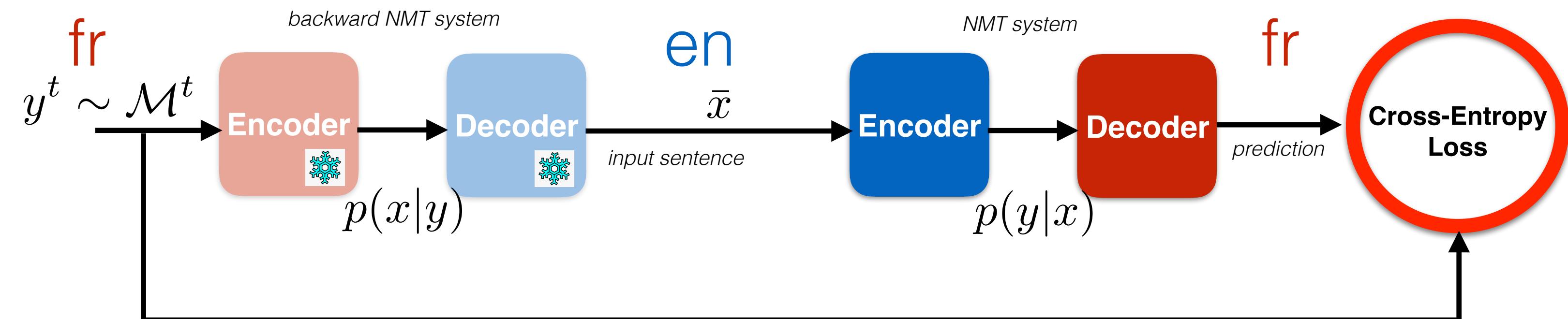
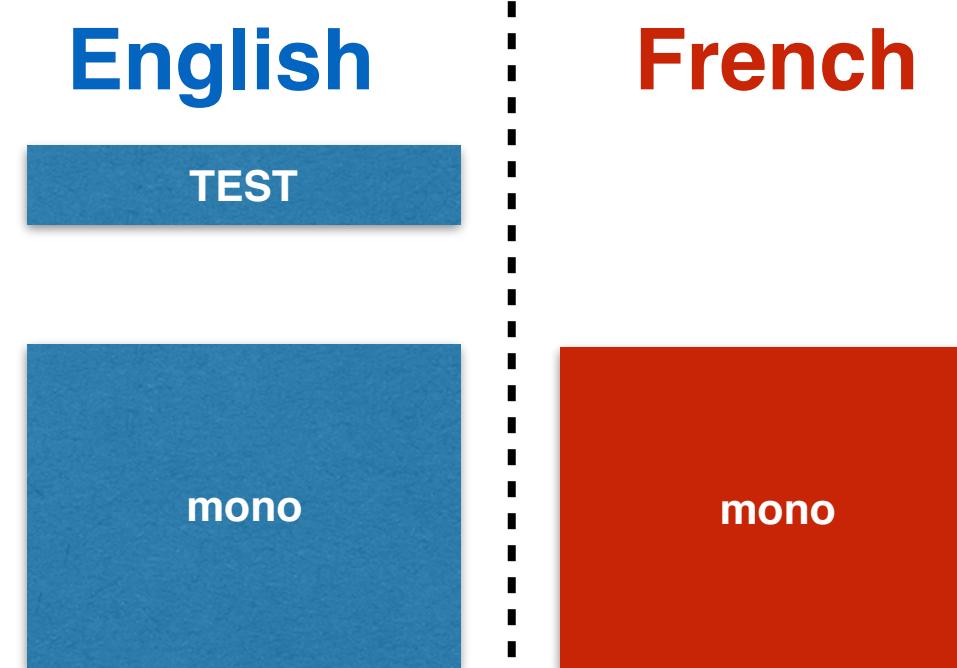


Problem: lack of modularity.

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

Case Study #1: Unsupervised MT

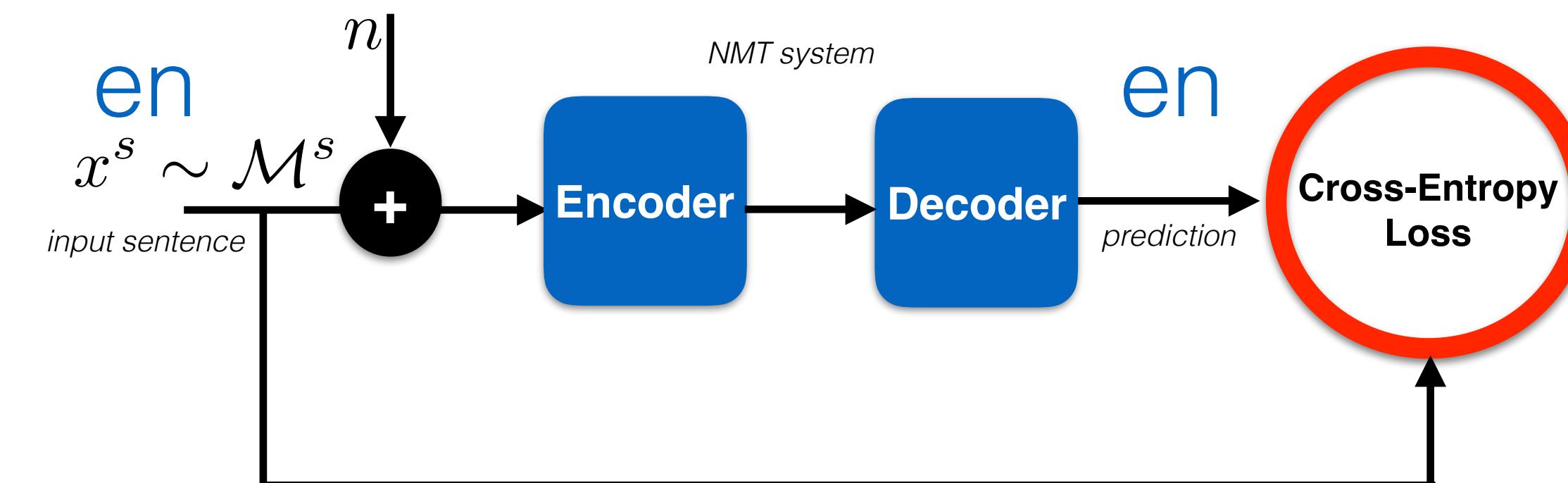
DATA



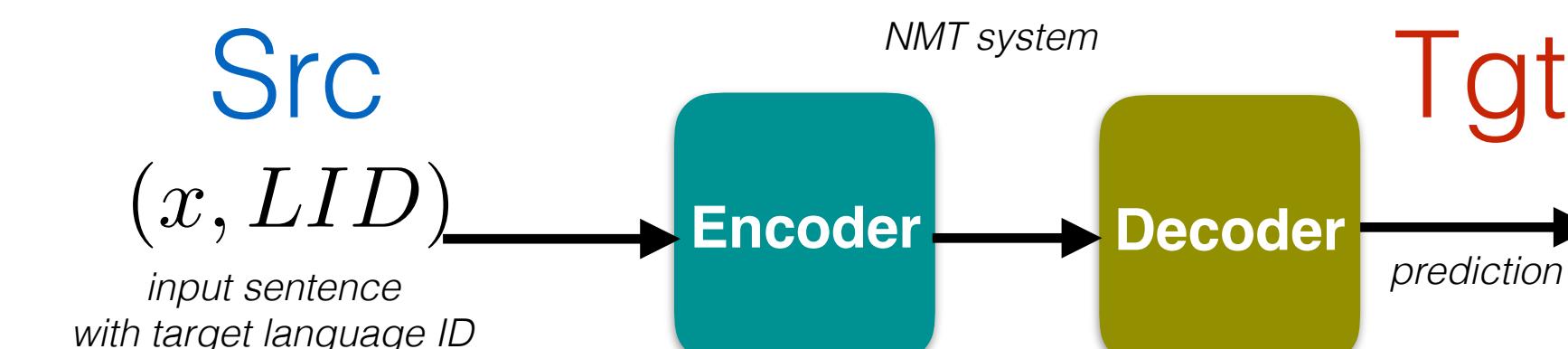
$$\mathcal{M}^t = \{y_k^t\}_{k=1,..,M_t}$$

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

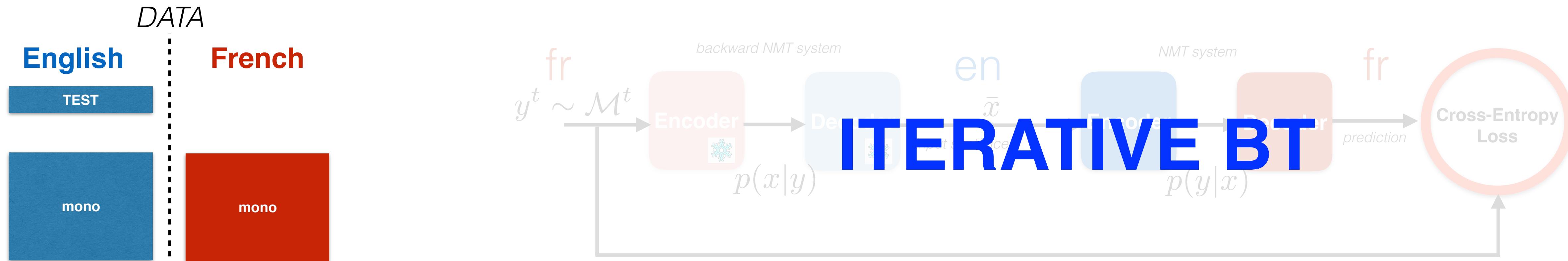
DAE makes sure decoder outputs fluently in the desired language.



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



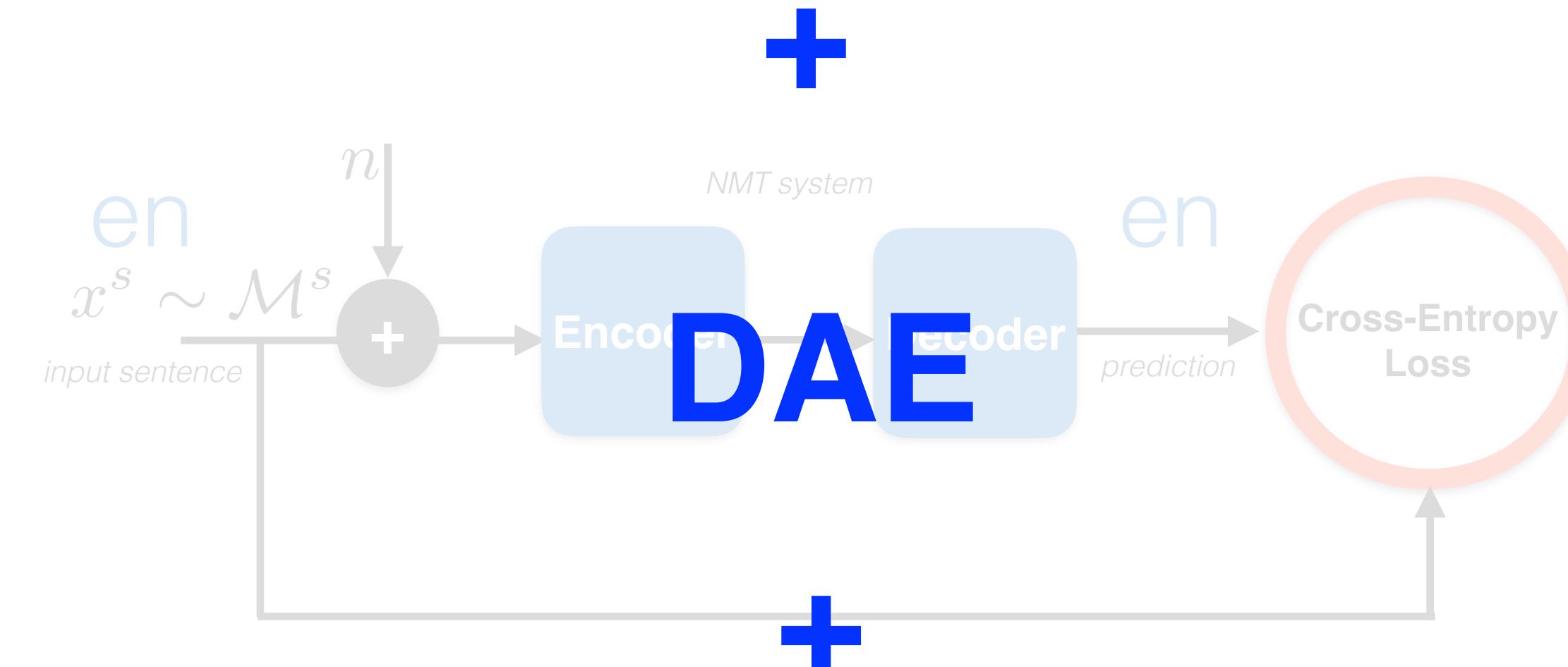
Case Study #1: Unsupervised MT



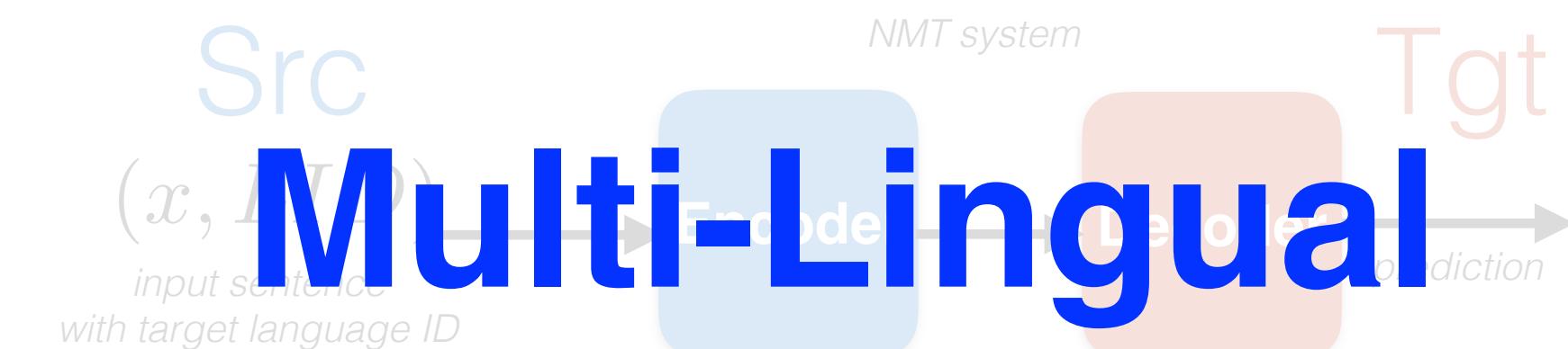
$$\mathcal{M}^t = \{y_k^t\}_{k=1,..,M_t}$$

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

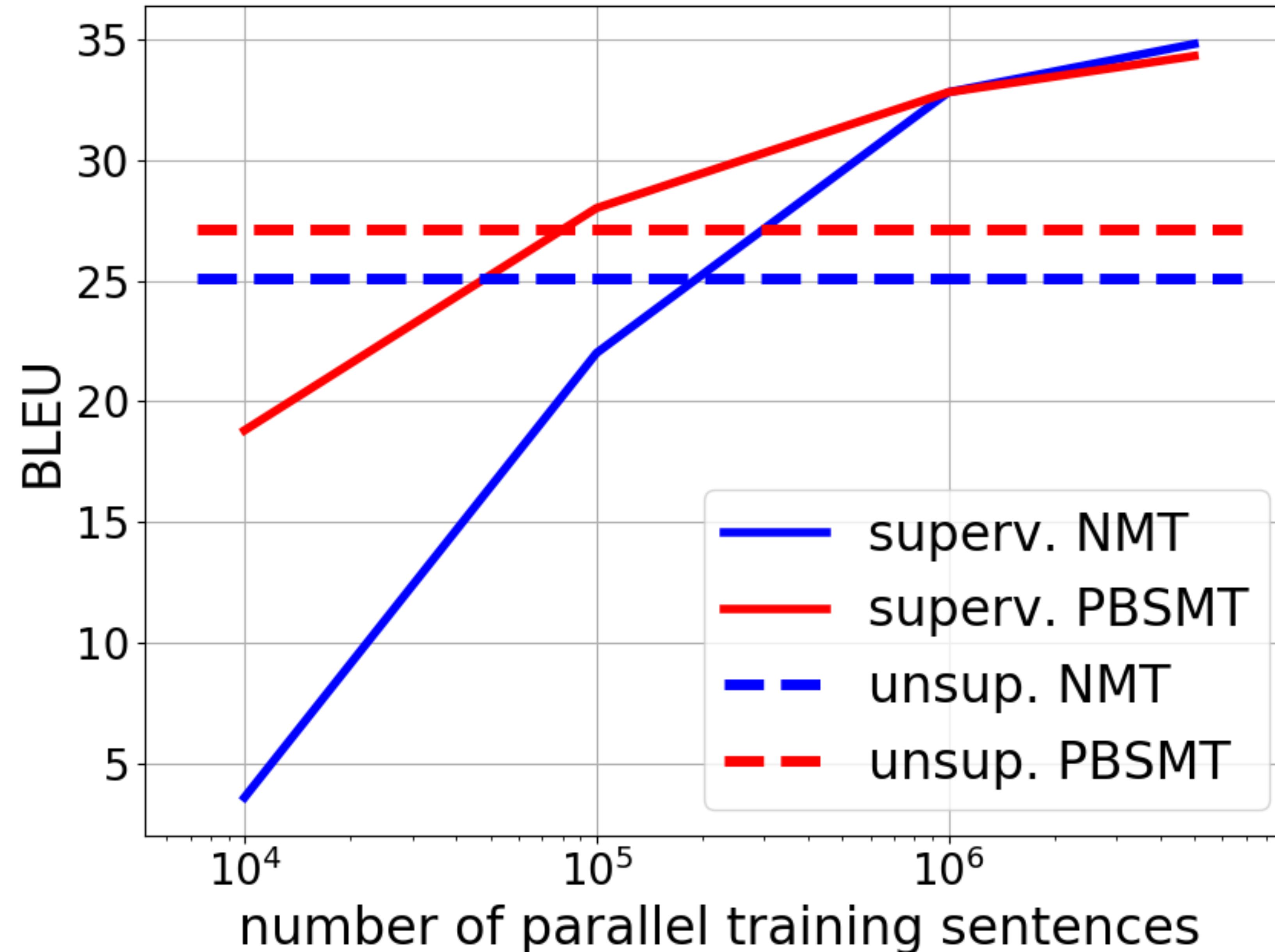
DAE makes sure decoder outputs fluently in the desired language.



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



WMT'14 En-Fr



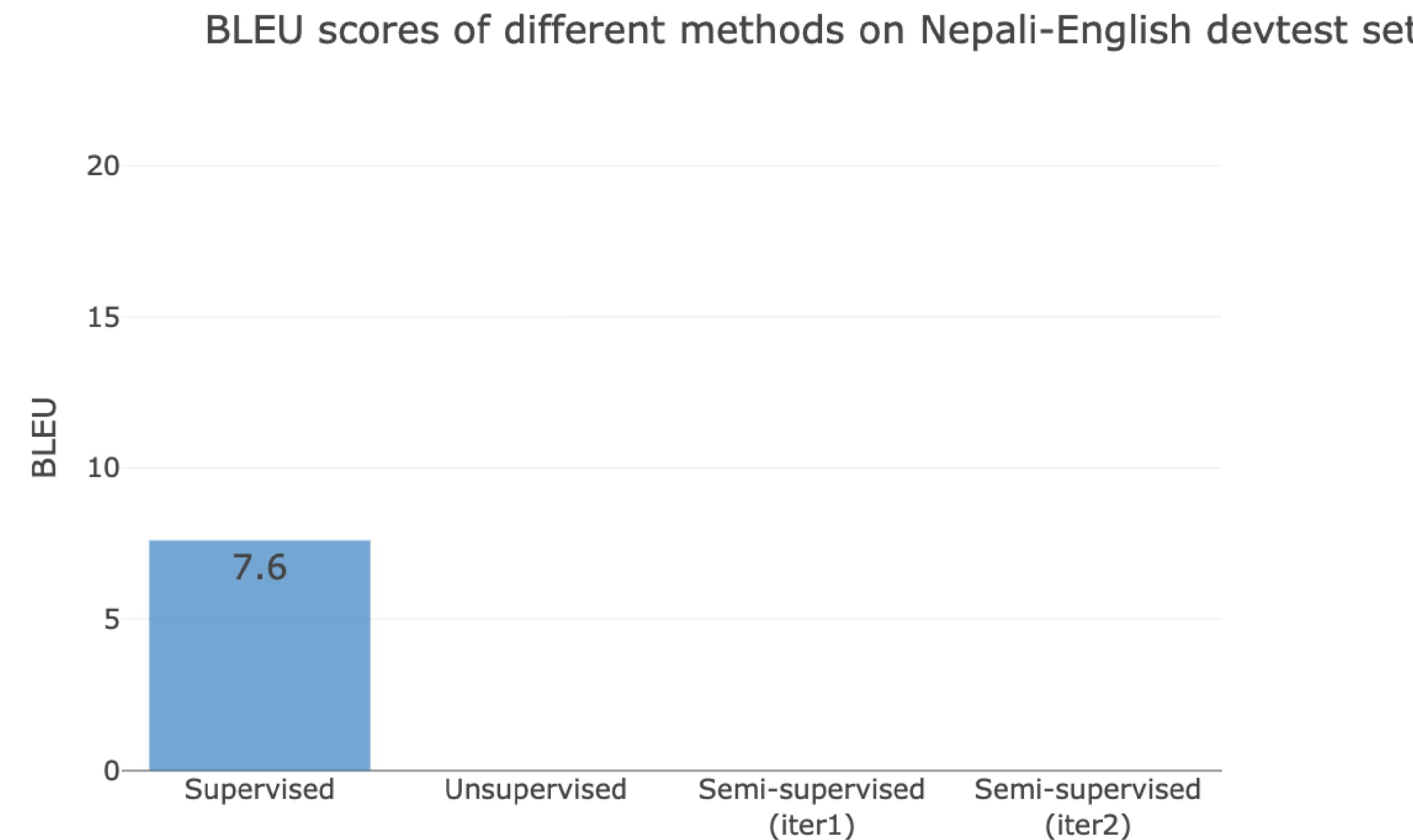
Same ideas can be applied to phrase-based statistical MT systems (PBSMT). NMT and PBSMT can be combined for even better results.

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

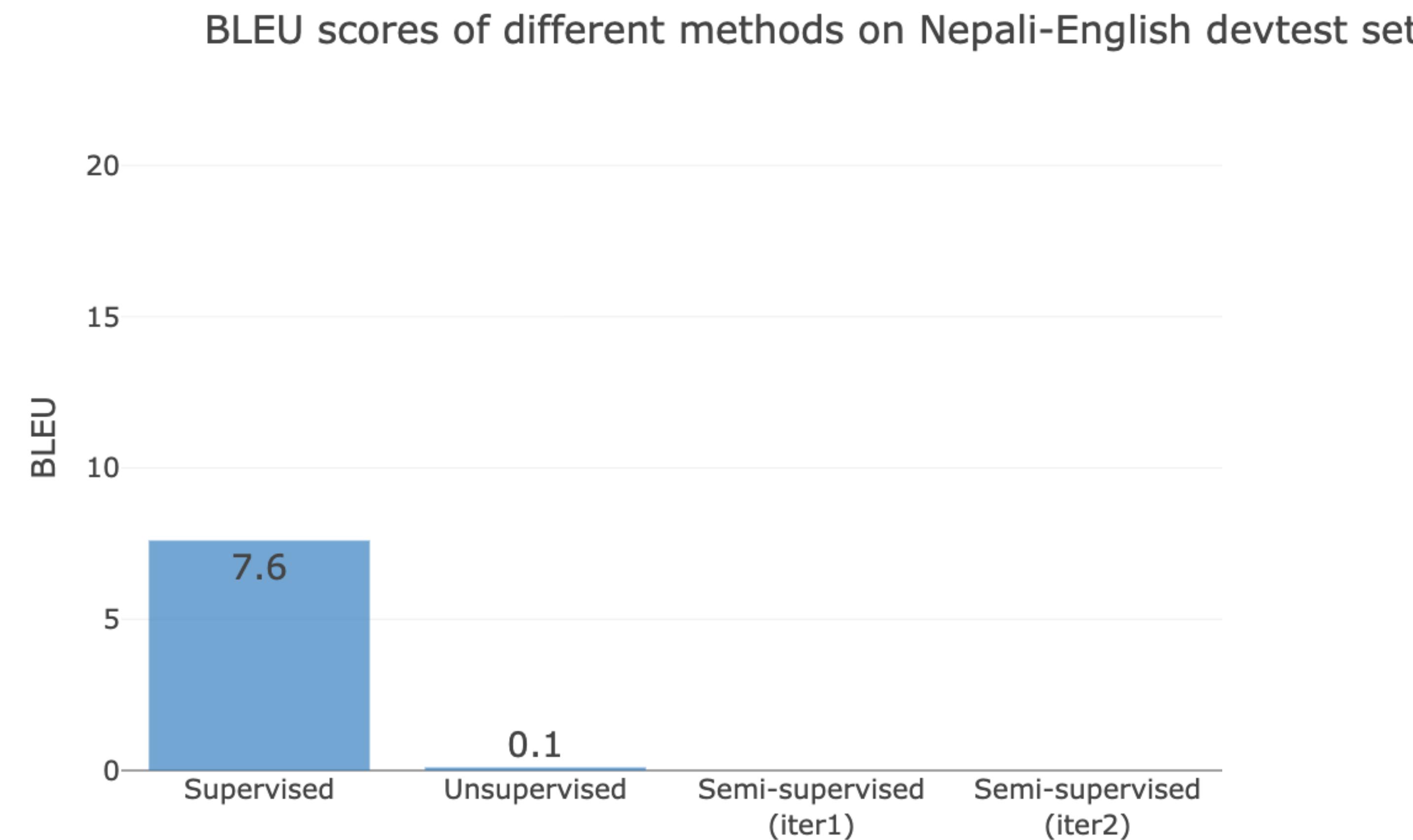
Case Study #2: FLoRes Ne-En

	In-domain (Wikipedia)	Out-of-domain
Parallel	None	500K sentences (Bible, GNOME/Ubuntu, OpenSubtitle, ...) *Hindi: 1.5M
Monolingual	Ne: 100K sentences En: 70M	~5M sentences (CommonCrawl) *Hindi: 45M

Results on FLoRes: Ne-En

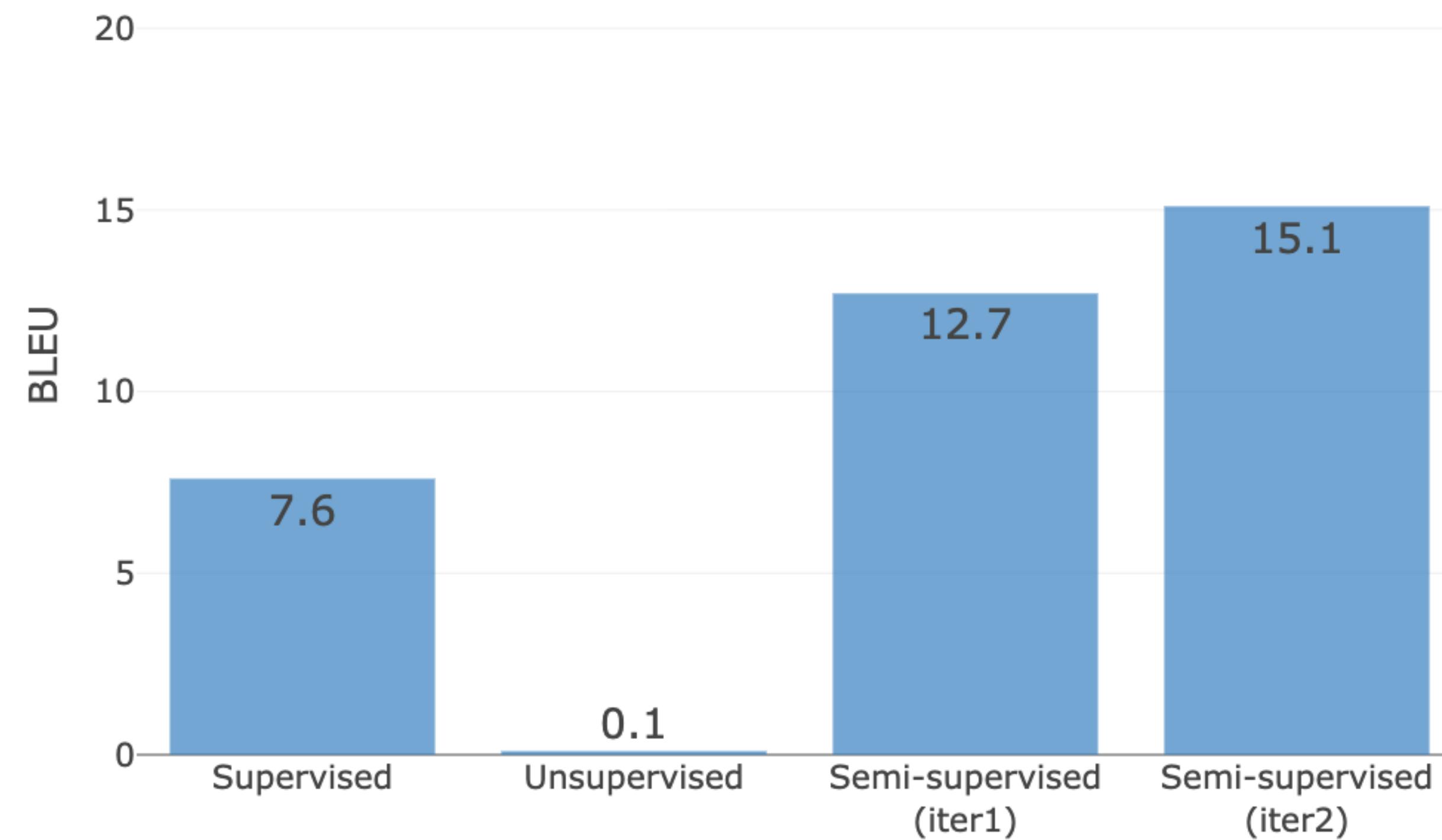


Results on FLoRes: Ne-En

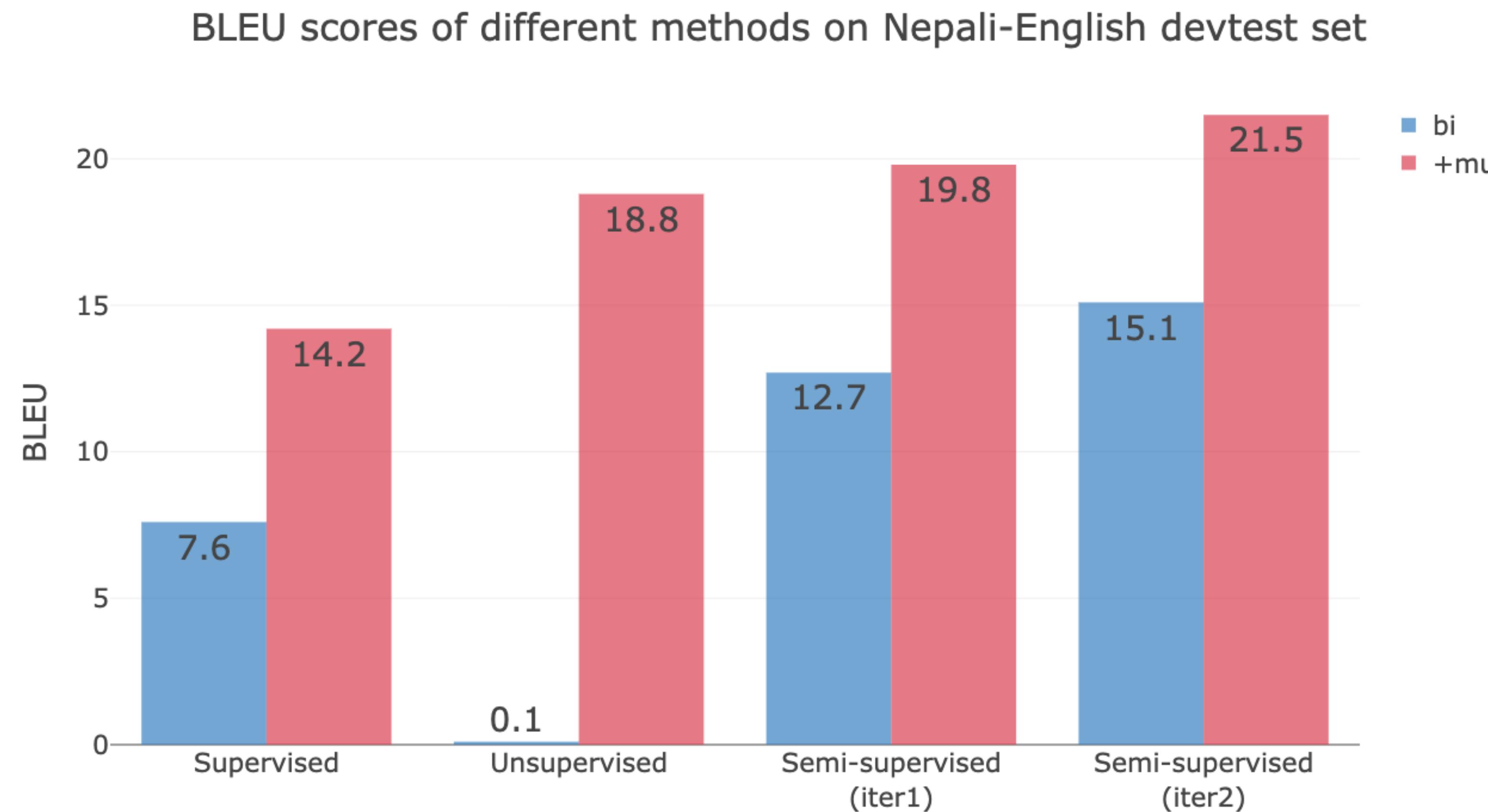


Results on FLoRes: Ne-En

BLEU scores of different methods on Nepali-English devtest set



Results on FLoRes: Ne-En



Case-Study #3: English-Burmese

ပဟိုတမျက်နှာ

ဆွေးနွေးချက်

ဖတ်ရန်

ရင်းမြစ်ကို ကြည့်ရန်

ရာဇ်ဝင်ကြည့်ရန်

ဝိကိုပီးဒီးယား တွင် ရှာဖွေရန်



ပဟိုတမျက်နှာ

ဝိကိုပီးဒီးယားမှ ကြိုဆိုပါသည်။

မည်သူမဆုံး ကြည့်ရပ်ငါးဆင်ရှင်သော အခမဲ့လွှတ်လပ်စွဲယုံကြုံများ ဖြစ်ပါသည်။
အကြောင်းအရာပေါင်း ၄၄၈၁၄ ခုကို မြန်မာဘာသာဖြင့် ဖတ်ရှုနိုင်ပါသည်။



- အနုပညာ
- အစ္စပွဲတို့
- ပထဝိဝင်
- သမိုင်း
- သချာ
- မှုပိုင်းအားလုံး

အထူးအကြောင်းအရာ

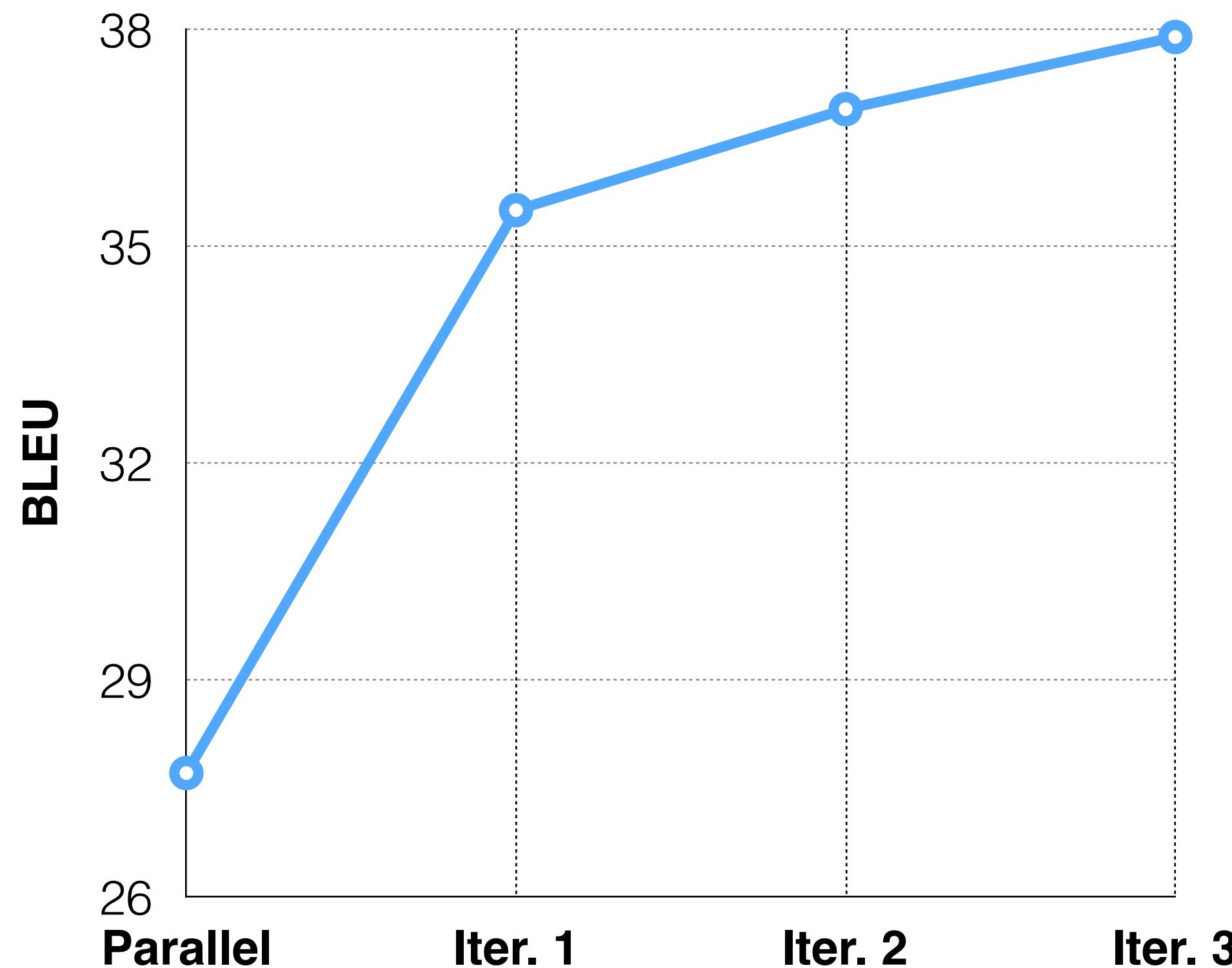
သိန့်နယ် သည် ယခင် ရှမ်းပဒေသရာဇ်ပြည်နယ်များတွင် ပါဝင်ခဲ့သော ပြည်နယ်တစ်ခု ဖြစ်ပါသည်။ သိန့်နယ်ကို ခရစ် ၁၈၈၈ ခုနှစ် အက်လိပ်တို့ ဝင်ရောက်သိမ်းပိုက်ပြီးနောက်မှ မြောက်သိန့်နယ် (သိန့်နယ်)နှင့် တောင်သိန့်နယ်(မိုင်းရယ်နယ်)ဟု ခွဲခြားအုပ်ချုပ်ခဲ့သည်။ ရေးအခါသမယက သိန့်နယ်ကြီးသည် အစိတ်စိတ်ကွဲပြားခြင်းမရှိဘဲ ရှမ်းပြည်နယ်တရာ့မ်းလုံးတွင် အကျယ်ပြန့်ဆုံး အာဏာအလွမ်းမိုးဆုံးသော နယ်ကြီးဖြစ်ခဲ့သည်။ သို့သော် မြန်မာဘုရင်များ ဝင်ရောက်တိုက်ခိုက် သိမ်းပိုက်ပြီးသည့်နောက် အုပ်ချုပ်ရေးဝါဒအရ ရာထူးလူသည့်နယ်ရှင် တော်ဘွားတို့ကြောင့် သိန့်နယ်ကြီးသည် ငါးနယ်အထိ အစိတ်စိတ် ကွဲပြားခဲ့လေသည်။ ထိုအတွင်း တော်ဘွားအချင်းချင်း စိတ်ဝမ်းကွဲကာ တစ်ဦးနှင့်တစ်ဦးတိုက်ခိုက်၏ ဆိုင်ရာ နယ်ပယ်များကို အုပ်ချုပ်ကြသည်။ နောက်ဆုံးခရစ် ၁၈၈၈ ခုနှစ် အက်လိပ်တို့ဝင်ရောက်လာမှ အထက်ပါ အတိုင်းနှစ်နယ်ခွဲ၏ အုပ်ချုပ်ခဲ့သည်။ နှစ်နယ်ခွဲ၏ အုပ်ချုပ်စက ခွန်ဆိုင်တံ့ဟမ်းအား မြောက်သိန့်နယ်အတွက် တော်ဘွားအဖြစ်လည်းကောင်း၊ ဆိုင်နော်ယော်သား နော်မိုင်းအား တောင်သိန့်နယ် တော်ဘွားအဖြစ်လည်းကောင်း၊ ခန့်အပ်ခဲ့လေသည်။ ၁၉၂၅ ခုနှစ်တွင် မြောက်သိန့်နယ်ကို စင်ဟုံဖက တော်ဘွားအဖြစ် ဆောင်ရွက်ခဲ့လေသည်။

Workshop on Asian Translation 2019: English-Myanmar

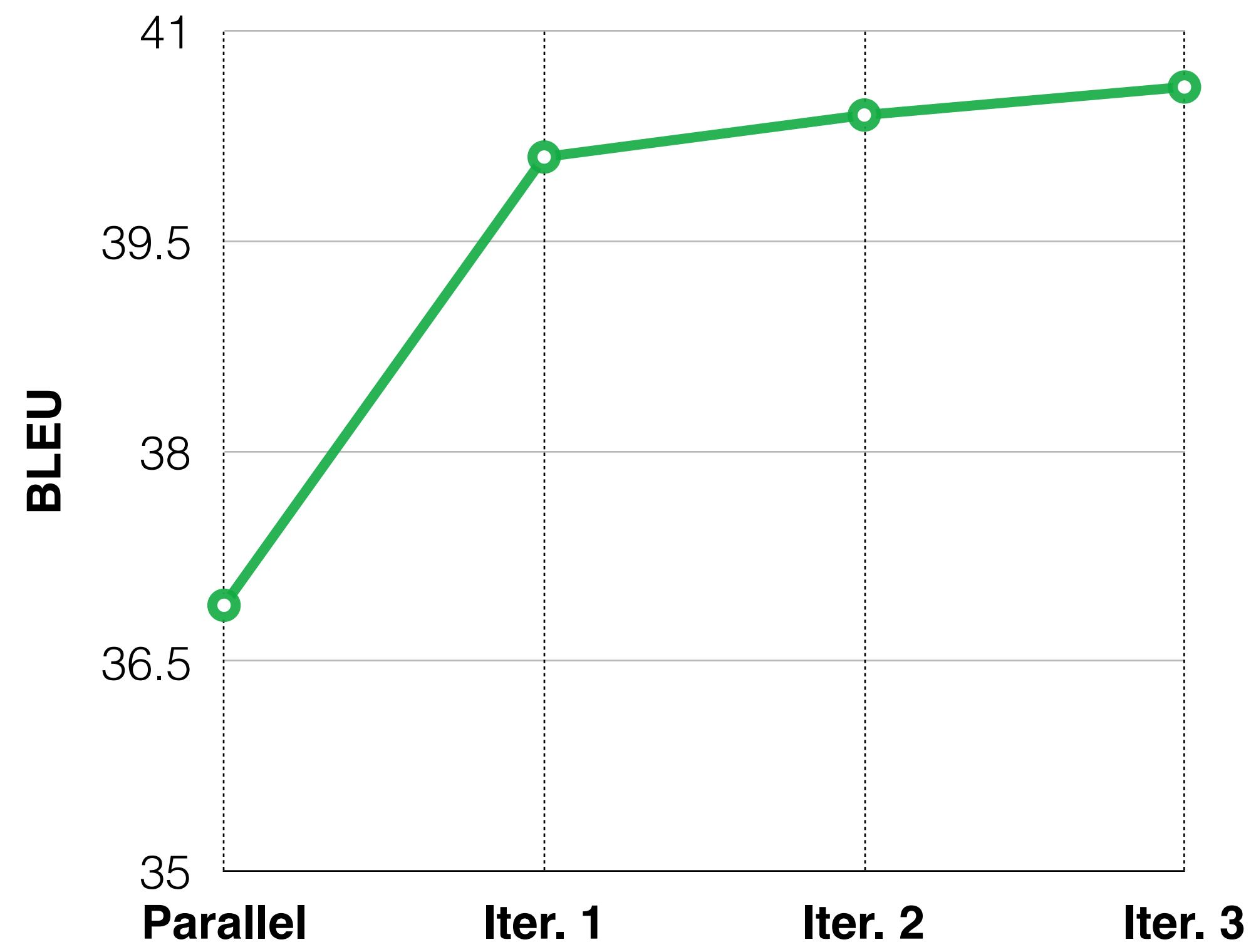
	In-domain (News)	Out-of-domain
Parallel	20K sentences	200K sentences
Monolingual	~79M sentences (En only)	~23M sentences (My only)

Results: Iterative ST+BT

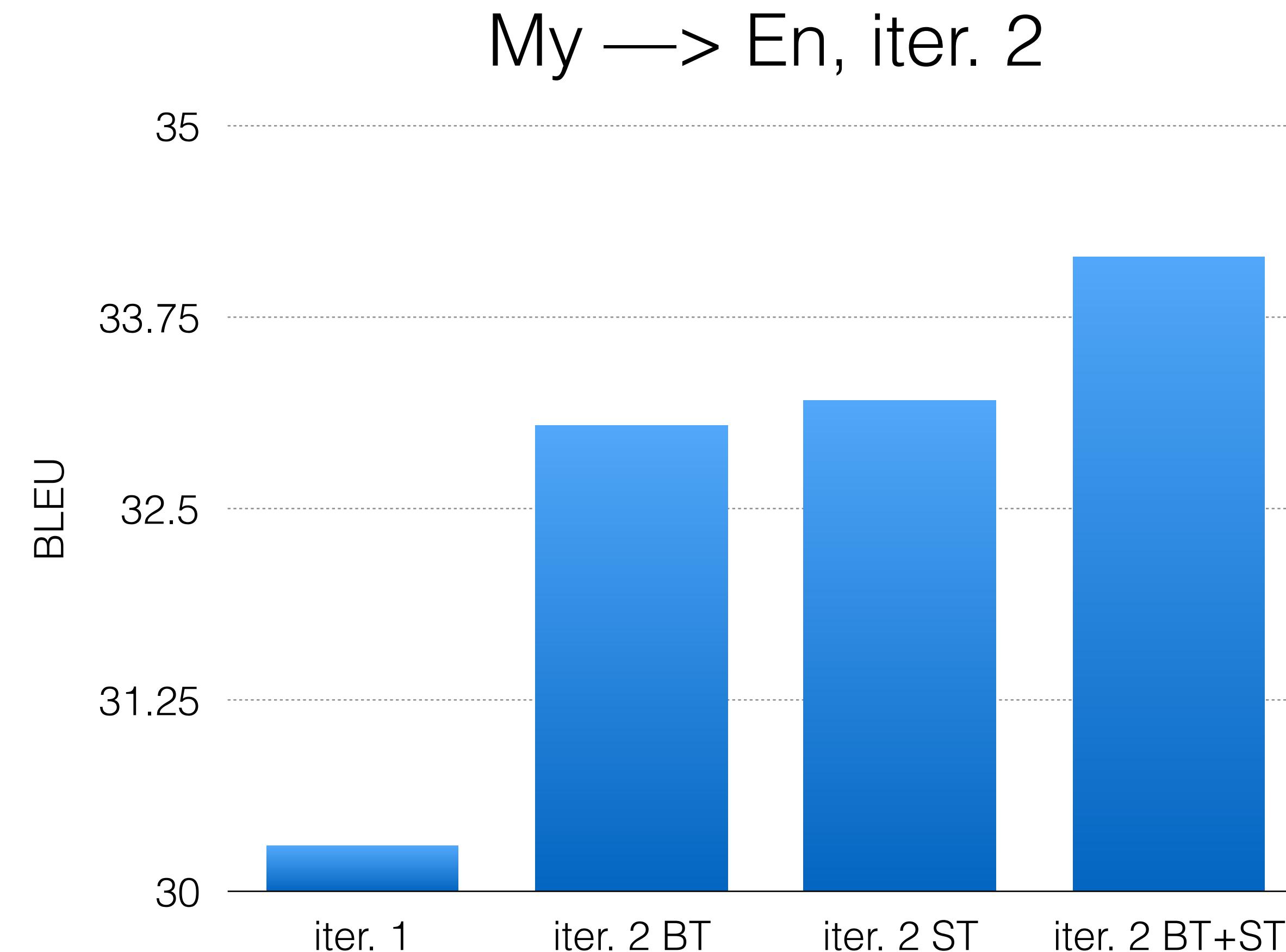
My → En



En → My

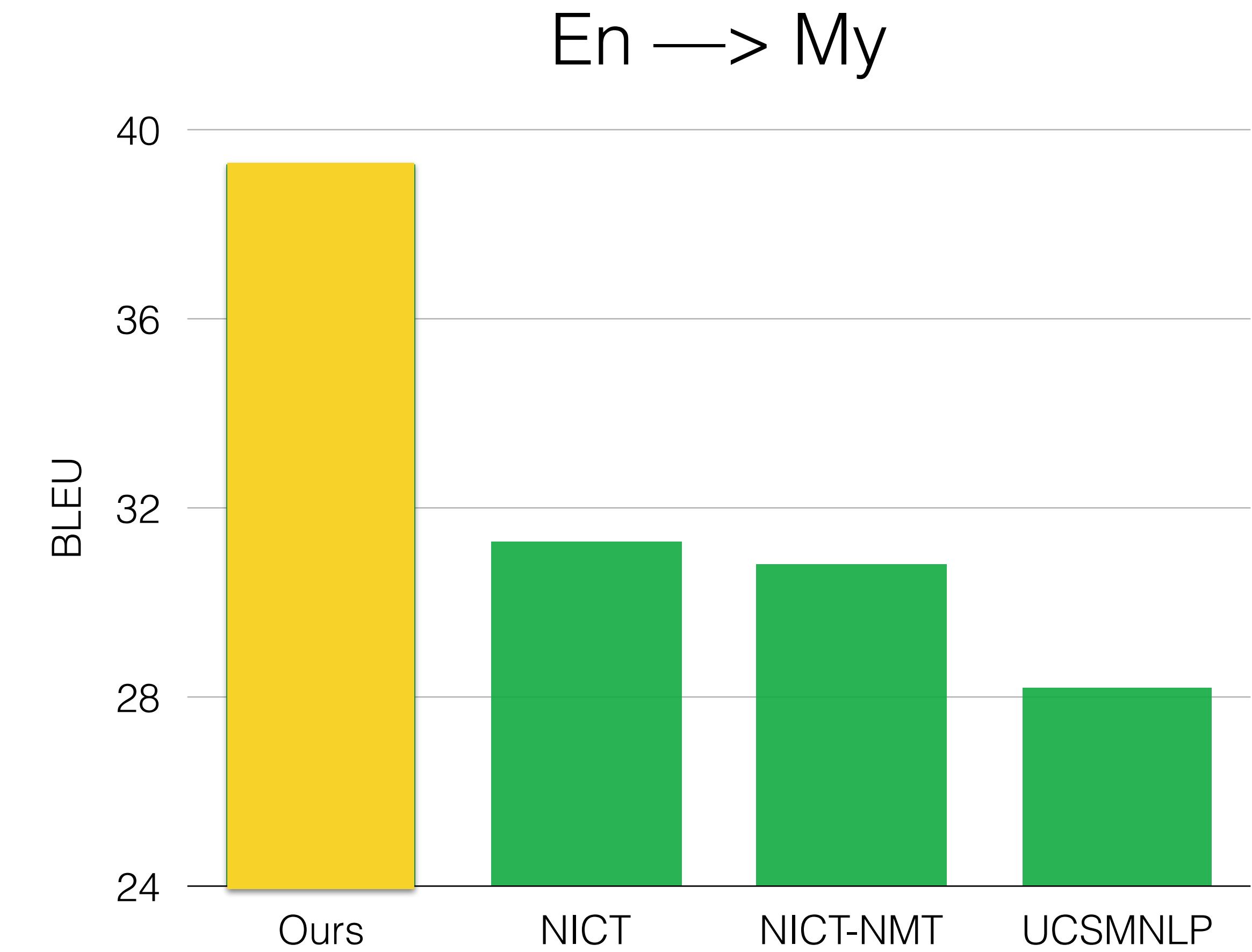
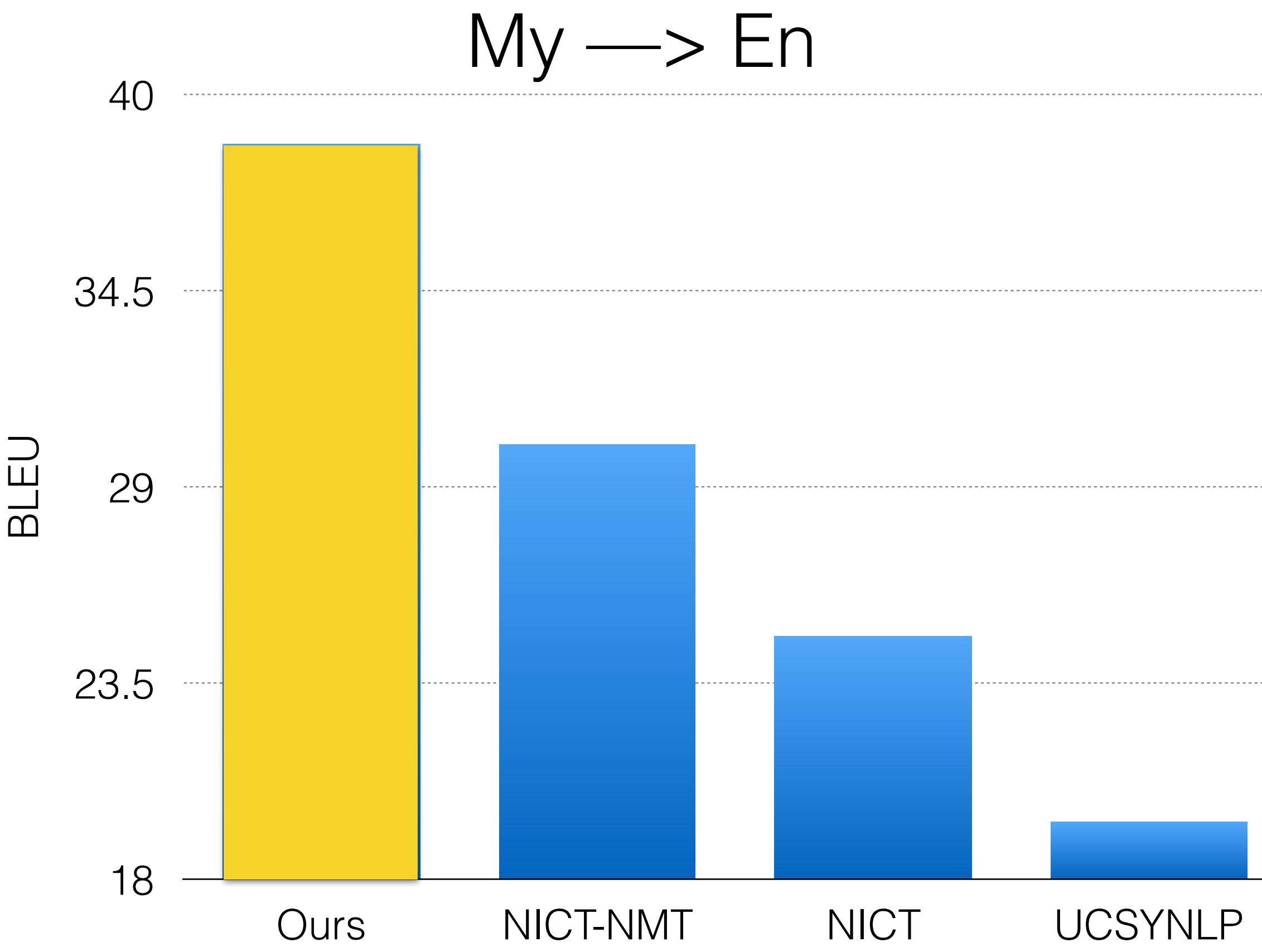


Results: BT vs ST vs BT+ST



Final Results of 2019 Competition

+8 BLEU compared to second best



Demo (Myanmar → English)



ဗြိတ်သူ နိုင်ငံရေး အကျပ်အတည်း

① 4 စက်တင်ဘာ 2019

f m t e ပေါ်ပေါ်

ဗြိတ်သူ ဝန်ကြီးချုပ် ဘောဂစ်ဂျွဲန်ဆင်ဟာ ရွှေးကောက်ပွဲ ကျင်းပတိ အတွက် ပါလီမန်ကို ဒီဇန်၊
တောင်းဆီ ဖို့ များနေတယ်လိုအနုံမှန်းထားကြပါတယ်။

Conclusion so far...

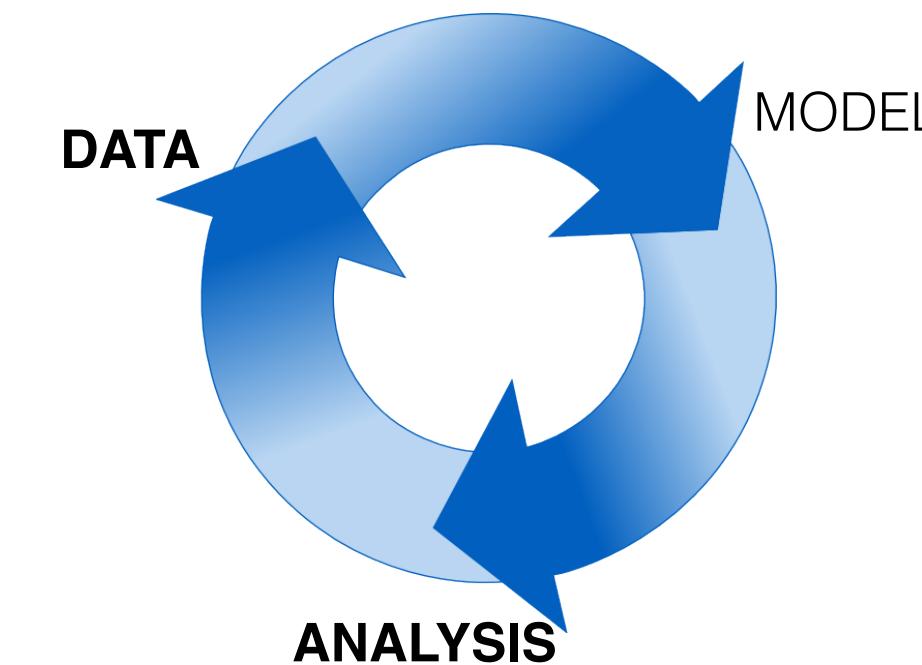
- Iterative back-translation, multi-lingual training work remarkably well.
- By feeding more data (BT, ST, pre-training, multi-lingual training) we can afford training bigger models. Bigger models train on more data generalize better.
- Low-resource MT requires big data and big compute!

Outline

- What is low-resource MT and why is it important?
- ML perspective on low resource MT
- Case studies:
 - Unsupervised MT
 - En-My
- **Perspectives**

What I did not talk about...

- The other two pillars



- Filtering (*)



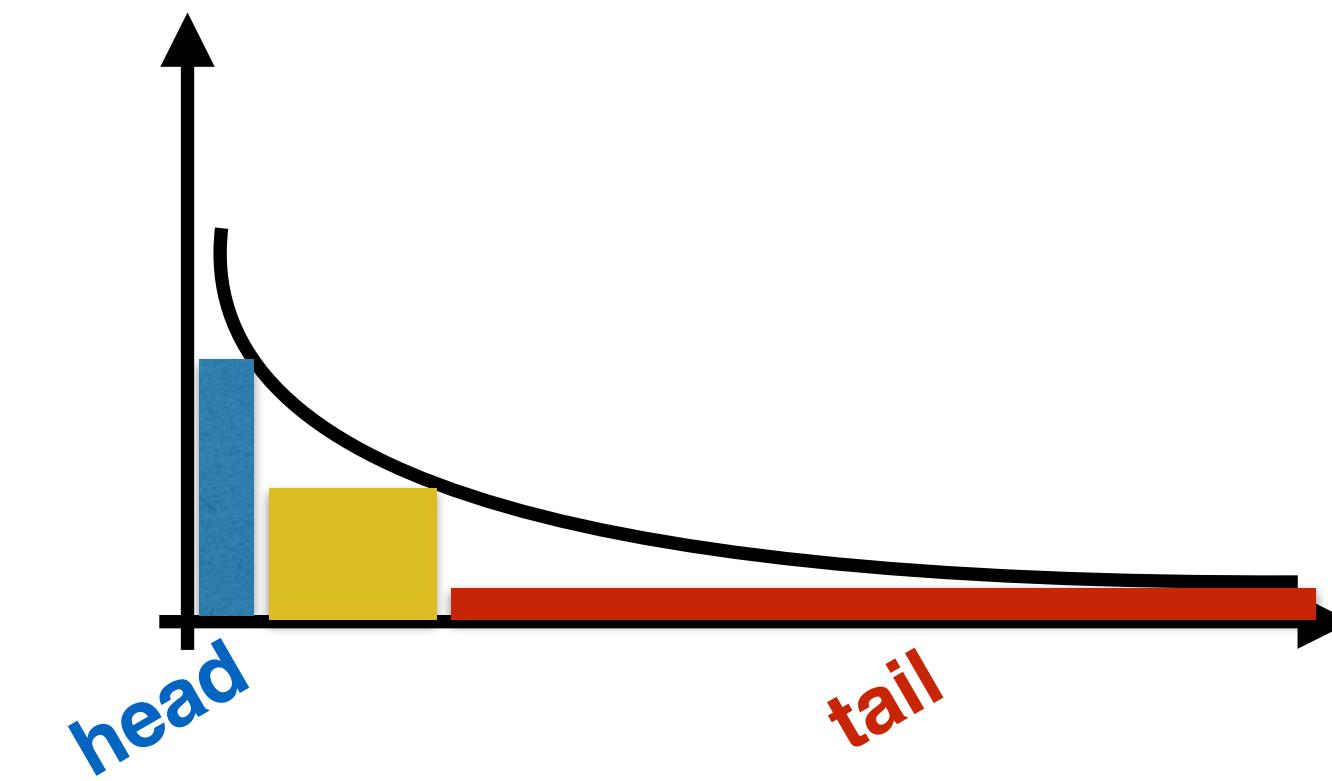
- Context: document level, multimodal, ... MT
- Scaling & Efficient prediction

(*) Schwenk et al. “CCMatrix: mining billions of high-quality parallel sentences on the WEB” arXiv:1911.04944 2019

Fan et al. “Beyond English centric MMT” arXiv 2020

What I did not talk about...

- The long tail of languages



- Bias

Example: the director decided to proceed with the acquisition -> ITA “director” is translated with male gender.

- Metrics and long tail of mistakes
- Idiomatic and fluid use of language



- Limited supervision is very common in practical applications.
- It usually does not pay off to scale down the model when there is little supervision.
- It is better to use more data and to scale-up the model instead. Model will learn lots of things, hopefully some of these will be relevant to the task of interest.
- The less direct supervision the more data (from auxiliary tasks) is needed.
- When dealing with lots of similar tasks (translation of various language pairs), it is better to be as language-pair agnostic as possible.
- Key techniques:
 - Data augmentation
 - Sharing a big model across several tasks
 - Iterative refinement
 - Domain adaptation

Debugging NMT



- **Data:** plot basic statistics (sentence length, token frequency).
 - Deduplicate sets, make sure there is no overlap between training/validation/test sets. Do not use test set ever.
- **Reproducibility:** Start simple and build on top of what is known to work. First reproduce then create something new. Be systematic and force yourself to come up with reproducible approaches. Releasing code does not suffice, if code is full of dataset-dependent hacks.
- **Analysis:** what do generation look like?
 - Often, your method encompasses some previous method as special case. Check things work as expected in that case. Adopt bottom-up approach to research.
 - Sort generations by sentence level BLEU and observe if there is any pattern (repetitions, excessively long/short sentences, etc.).

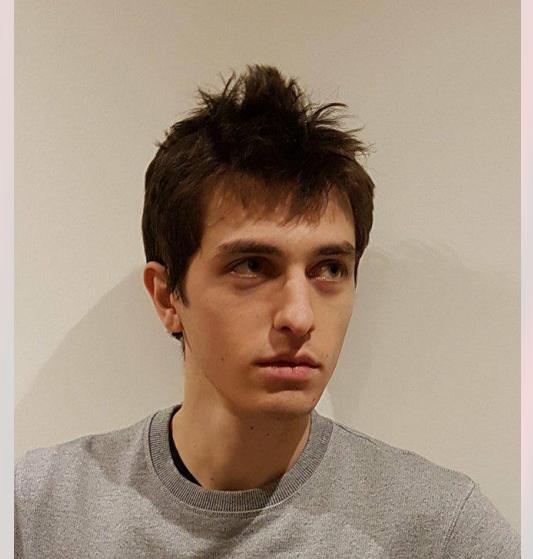
Debugging NMT



- **Optimization:** does the training loss decrease on the training set?
 - If training on a few mini-batches, can the training loss go to zero?
 - If not, check initialization, normalization layers and optimizer hyper-parameters.
- **Overfitting:** plot training and validation loss over time.
 - If overfitting is an issue, check dropout rate, label smoothing, input noise, add back-translation, do multilingual training, etc.
- **Domain adaptation:** does performance on held out portion of training set differ from validation set? Is the training set composed of several datasets? What's the statistics of sentence lengths and token frequency in each dataset?
 - Tagging, finetuning, example/dataset weighting methods.
- **Fooling:** Be rigorous and optimize the baselines as well you tune your method!

Opportunities @fb

- Internships
- Postdocs
- Full time positions
- Visiting scholar positions
- Scholarships



Guillaume Lample



Ludovic Denoyer



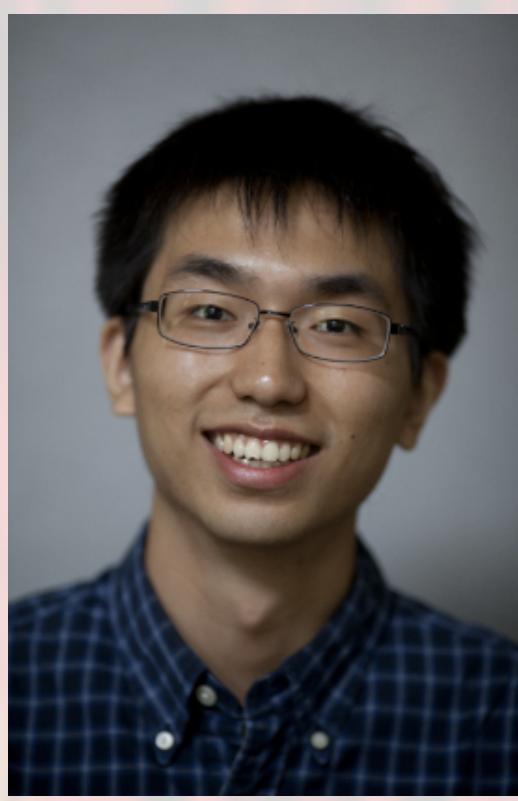
Myle Ott



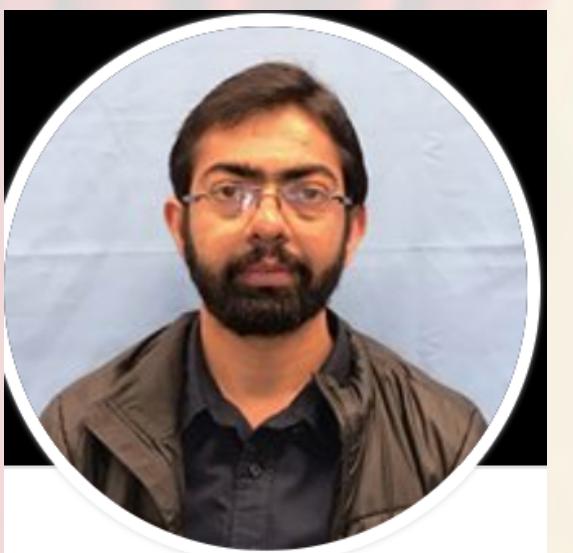
Peng-Jen Chen



Paco Guzmán



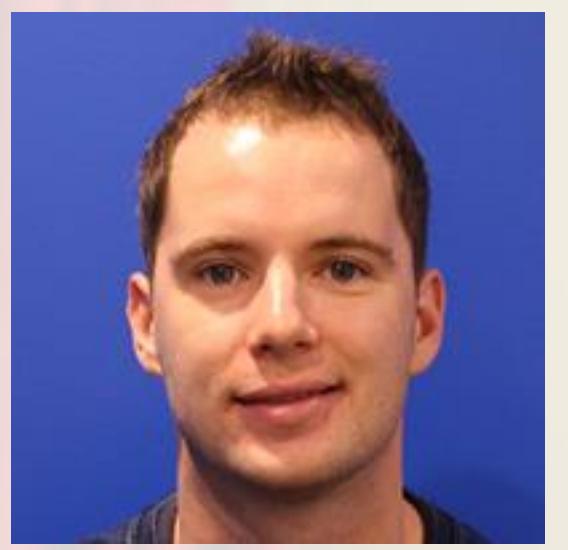
Jiajun Shen



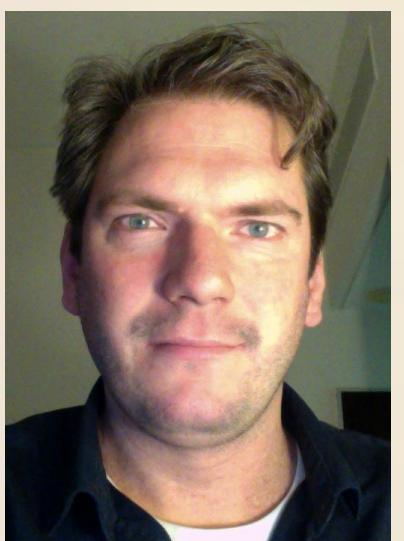
Naman Goyal



Jiatao Gu



Alexis Conneau



Philipp Kohen



Michael Auli



Junxian He



Sergey Edunov



Xian Li



Juan-Miguel Pino



Vishrav Chaudhary



Hervé Jegou

Questions?

Вопросы?

¿Preguntas?

Domande?