Low Resource Machine Translation

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

Course on Deep Learning - NYU CDS, 20 April 2021

Machine Translation

Training data

Italian

English

Collection of *parallel* sentences.

E.g.: It: Il gatto si è seduto sul tappetino.
En: The cat sat on the mat.

Machine Translation

Training data

Italian

English

Collection of *parallel* sentences.

E.g.: It: Il gatto si è seduto sul tappetino.
En: The cat sat on the mat.

Train NMT

NMT System

Ingredients:

- seq2seq with attention
- *SGD*

$$-\log p(y|x;\theta) = -\sum_{j=1}^{n} \log p(y_j|y_{j-1},\ldots,y_1,x;\theta)$$

Auto-regressive modeling reduces the structure prediction problem to a independent multi class logistic regression problems. E.g.:

$-\log p(y_2 = \text{"cat"}|y_1 = \text{"the"}, x = \text{"il gatto si è seduto sul tappetino"};\theta)$ $-\log p(y_4 = \text{"on"}|y_{1:3} = \text{"the cat sat"}, x = \text{"il gatto si è seduto sul tappetino"};\theta)$ ...

Neural Machine Translation

(in 3 slides)

Example:

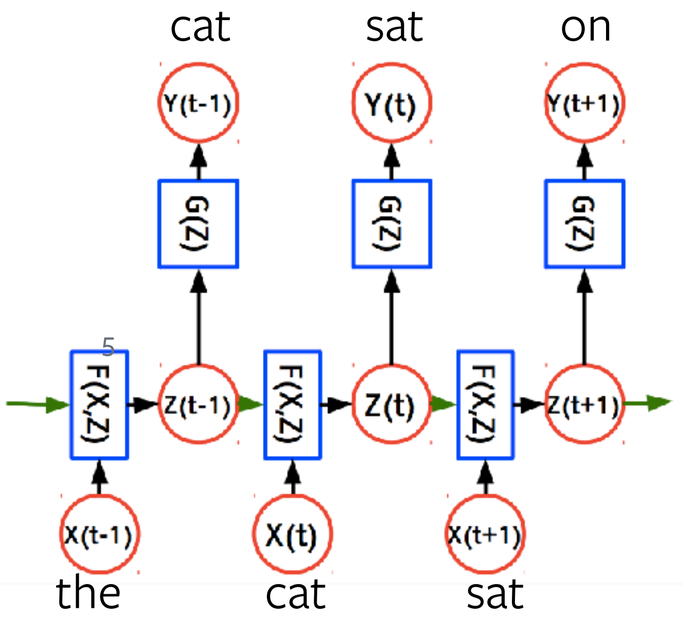
ITA (source) : Il gatto si e' seduto sul tappetino.

EN (target) : The cat sat on the mat.

Approach:

Have one RNN/CNN/Transformer (encoder) represent the source sentence, and another RNN/CNN/Transformer (decoder) predict the target sentence.
The decoder learns to (soft) align via attention.

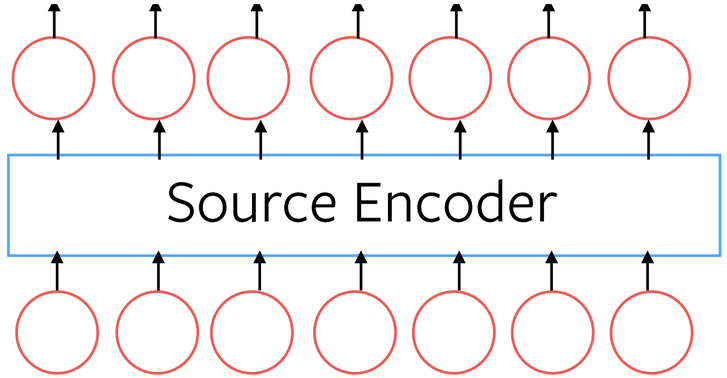
Neural machine translation by jointly learning to align and translate, Bahdanau et al. ICLR 2015



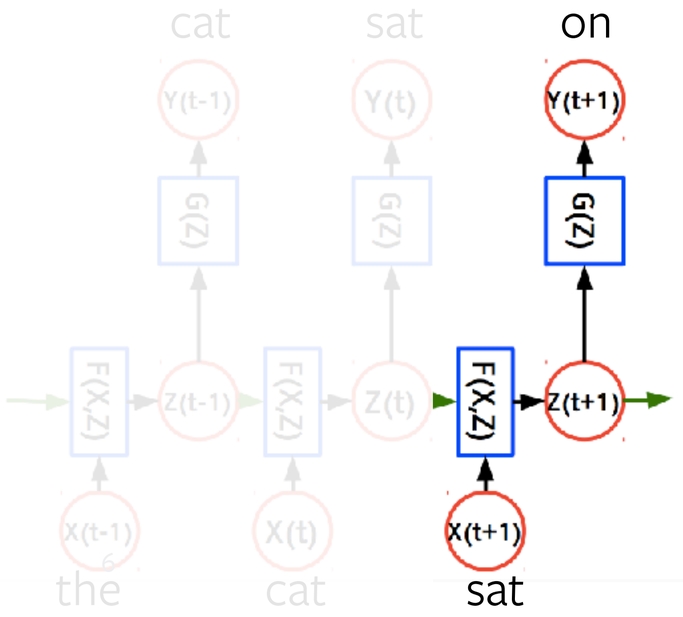
Y. LeCun's diagram

Source

1) Represent source



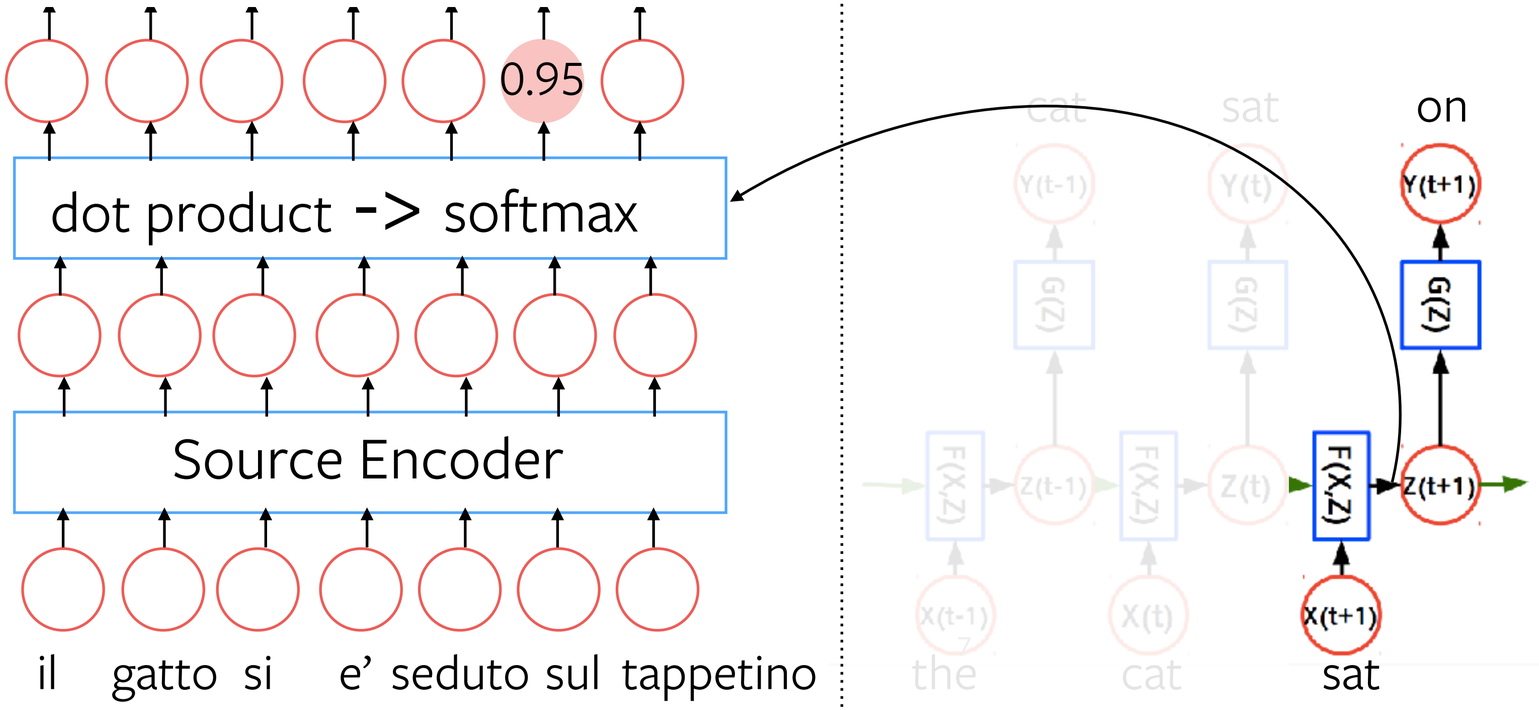
Target

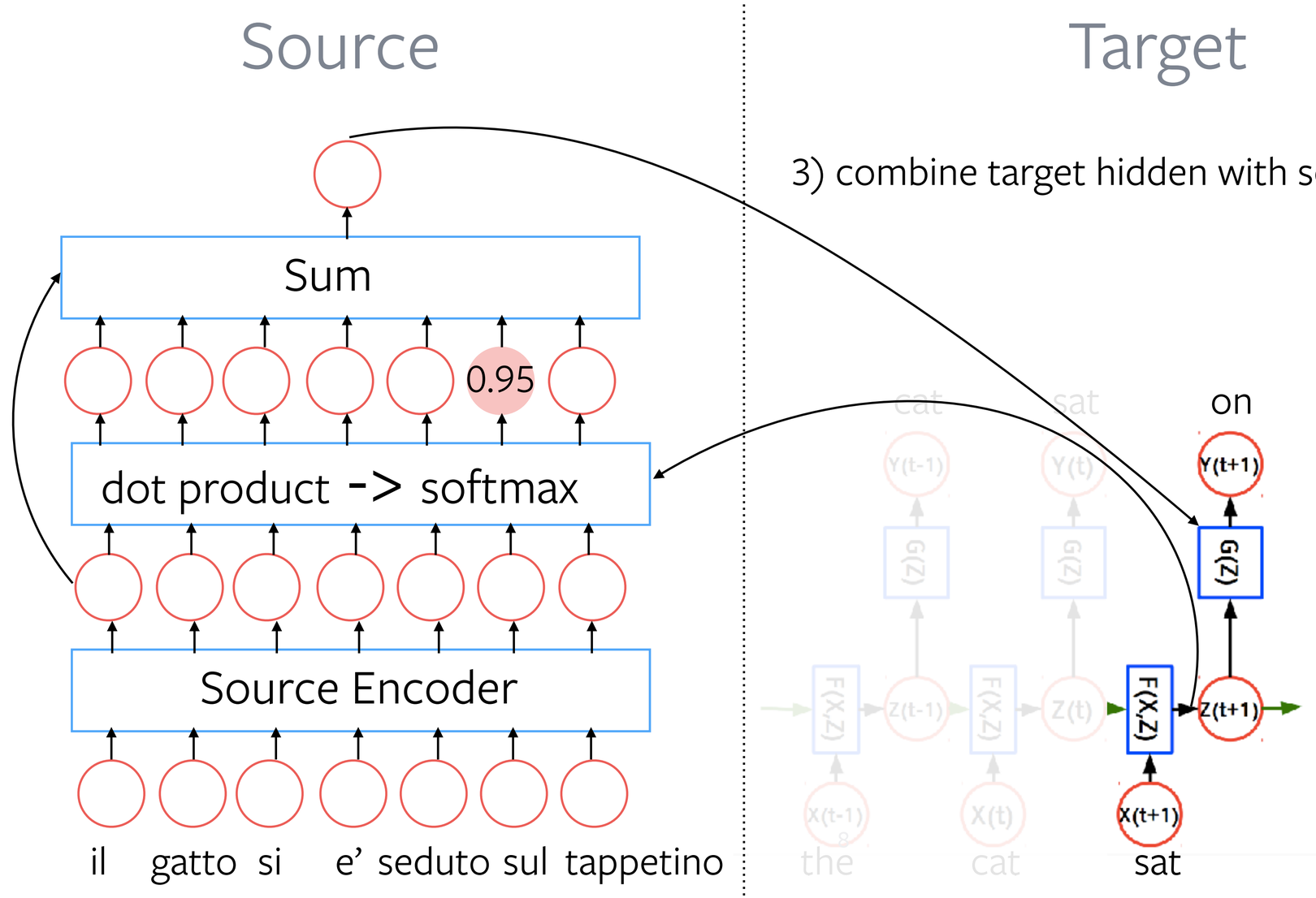


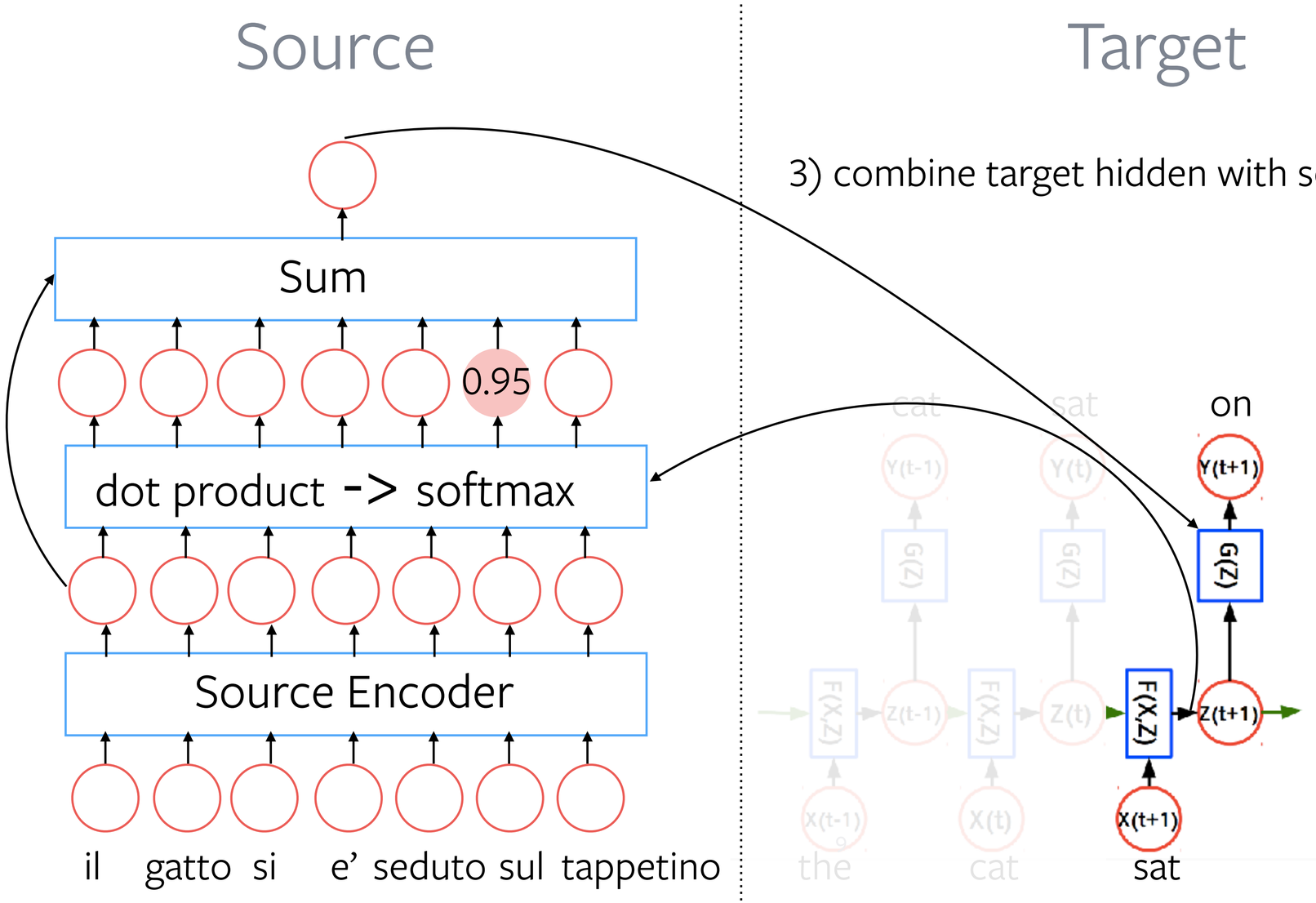
Source

Target

2) score each source word (attention)




Source
Target
3) combine target hidden with source vector
Sum
0.95
dot product -> softmax
Source Encoder
il gatto si e' seduto sul tappetino
cat
sat
on
Y(t-1)
Y(t)
Y(t+1)
G(Z)
G(Z)
G(Z)
F(X,Z)
Z(t-1)
F(X,Z)
Z(t)
F(X,Z)
Z(t+1)
X(t-1)
X(t)
X(t+1)
the
cat
sat

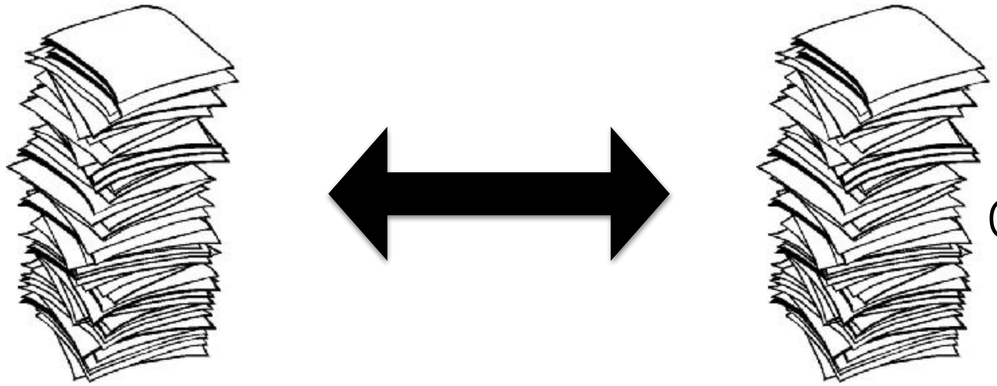


Alignment is learnt implicitly. **With CNNs & Transformers: All tokens processed in parallel.**

Machine Translation

Training data

Italian **English**



Collection of *parallel* sentences.

E.g.: It: Il gatto si è seduto sul tappetino.
En: The cat sat on the mat.

Train NMT

NMT System

Ingredients:
- seq2seq with attention
- *SGD*

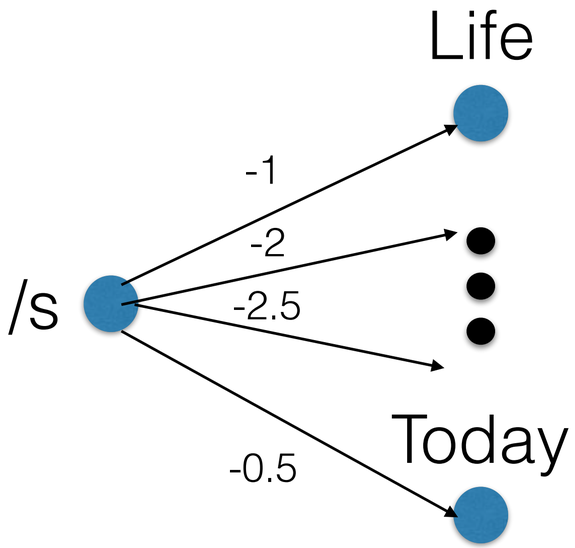
Test NMT

La vita è bella → **NMT System** → *Life is beautiful*

$$\arg\max \log p(y|x;\theta)$$

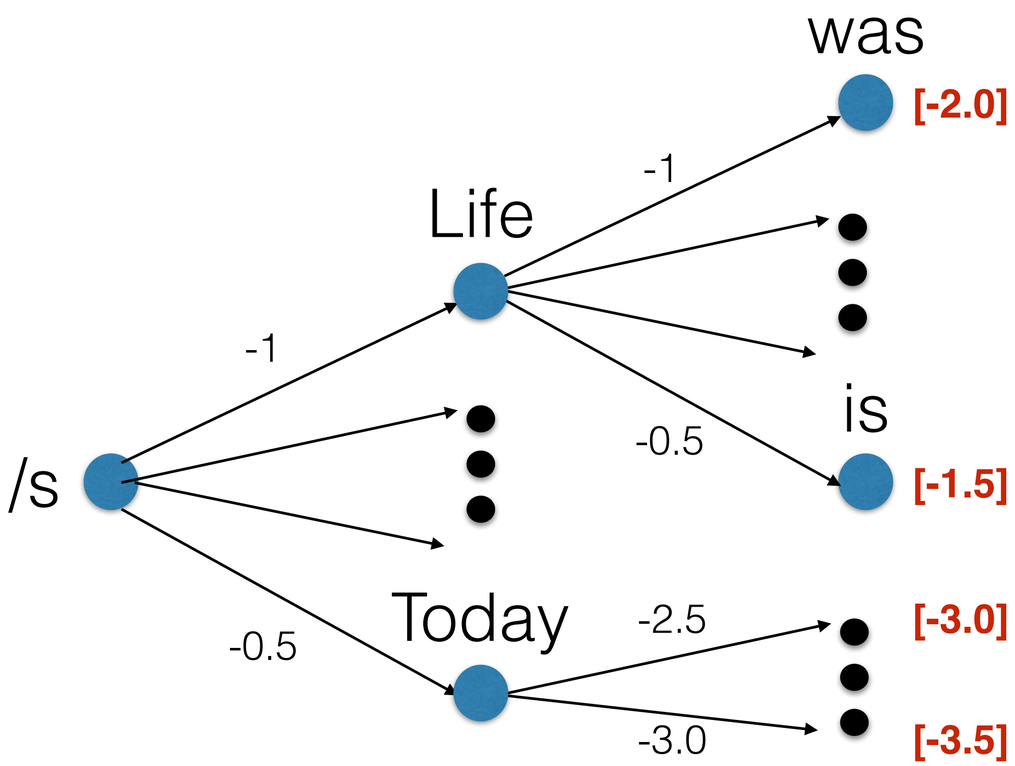
Ingredient:
- beam decoding

Beam Search



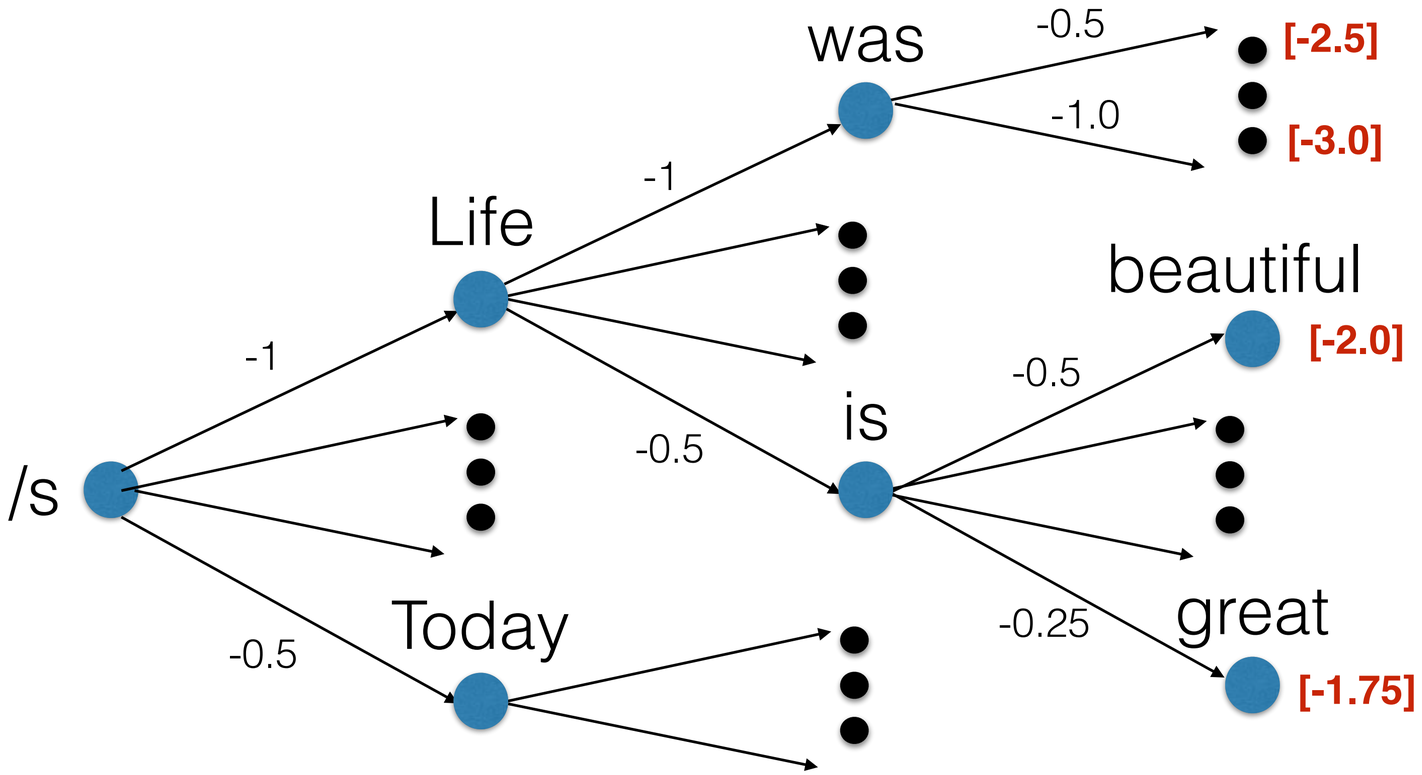
At every step maintain queue with k top scoring paths,
expand each of them and retain the top k scoring paths.

Beam Search



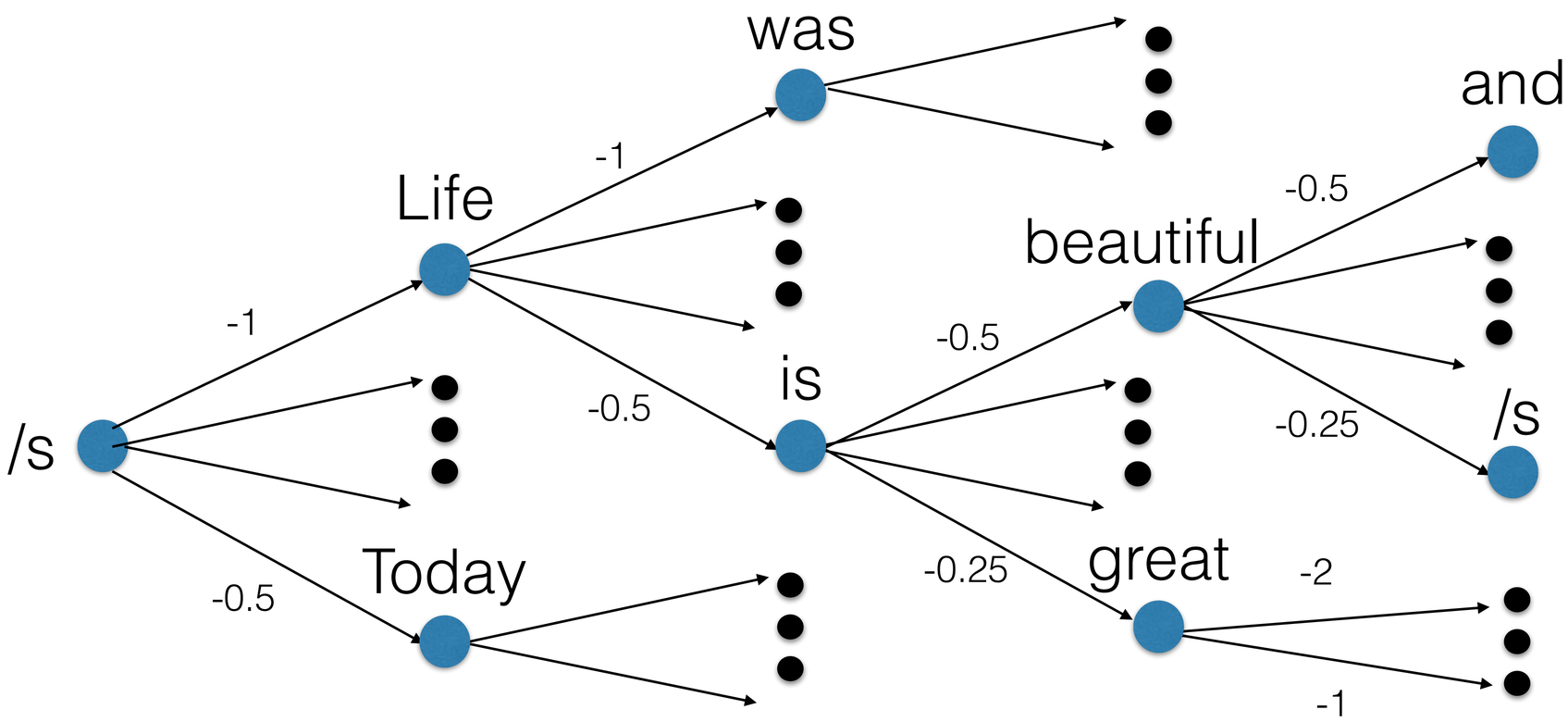
At every step maintain queue with k top scoring paths,
expand each of them and retain the top k scoring paths.

Beam Search



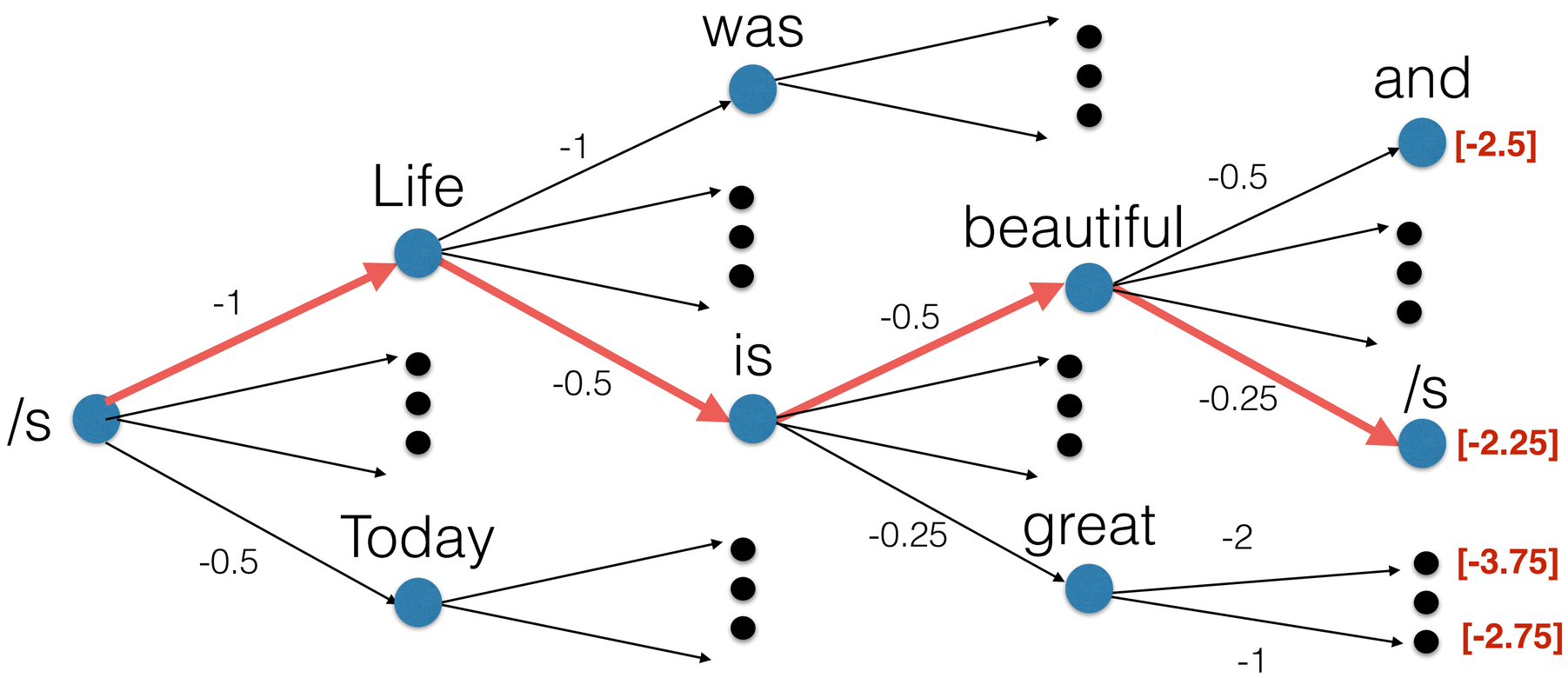
At every step maintain queue with k top scoring paths,
expand each of them and retain the top k scoring paths.

Beam Search



At every step maintain queue with k top scoring paths,
expand each of them and retain the top k scoring paths.

Beam Search



At the very last step, select the highest scoring path.

Trade-off between computational cost and approximation error.

Other decoding methods: sampling, top-k sampling, generative and discriminative reranking.

NMT Training & Inference

Training: predict one target token at the time and minimize cross-entropy loss.

Inference: find the most likely target sentence (approximately) using beam search.

Evaluation: compute BLEU on hypothesis returned by the inference procedure

$$p_n = \frac{\sum_{\text{generated sentences}} \sum_{\text{ngrams}} Clip(Count(\text{ngram matches}))}{\sum_{\text{generated sentences}} \sum_{\text{ngrams}} Count(\text{ngram})} \qquad \text{BLEU} = \text{BP}\ e^{\sum_{n=1}^{N} \frac{1}{N} \log p_n}$$

BLEU: a method for automatic evaluation of machine translation, Papineni et al. ACL 2002

Machine Translation

Training data

Italian

English

Collection of *parallel* sentences.

E.g.: It: Il gatto si è seduto sul tappetino.
En: The cat sat on the mat.

Train NMT

NMT System

Ingredients:
- seq2seq with attention
- *SGD*

Test NMT

La vita è bella. → **NMT System** → *Life is beautiful.*

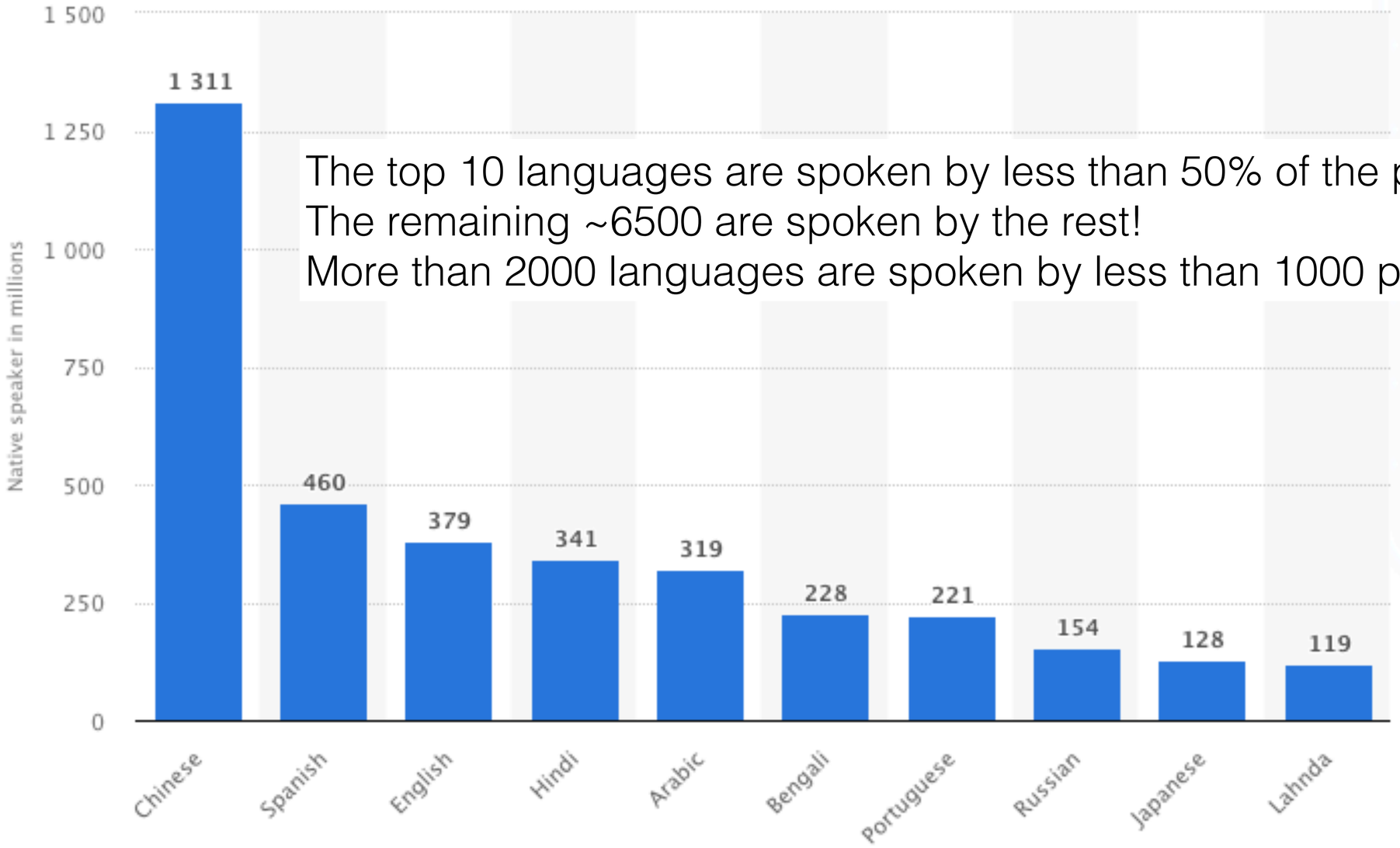
$$\arg\max \log p(y|x;\theta)$$

Ingredient:
- beam decoding

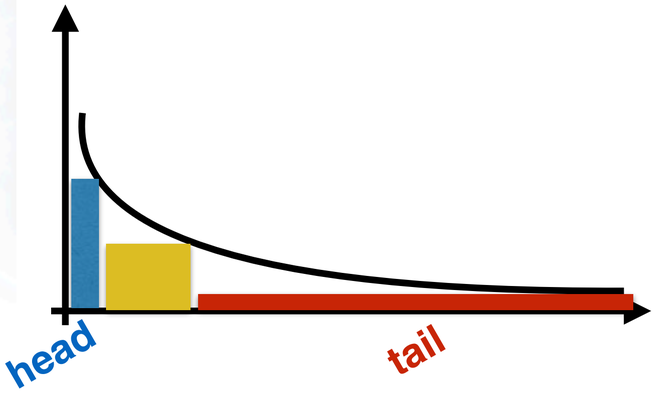
Some Stats

- 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.

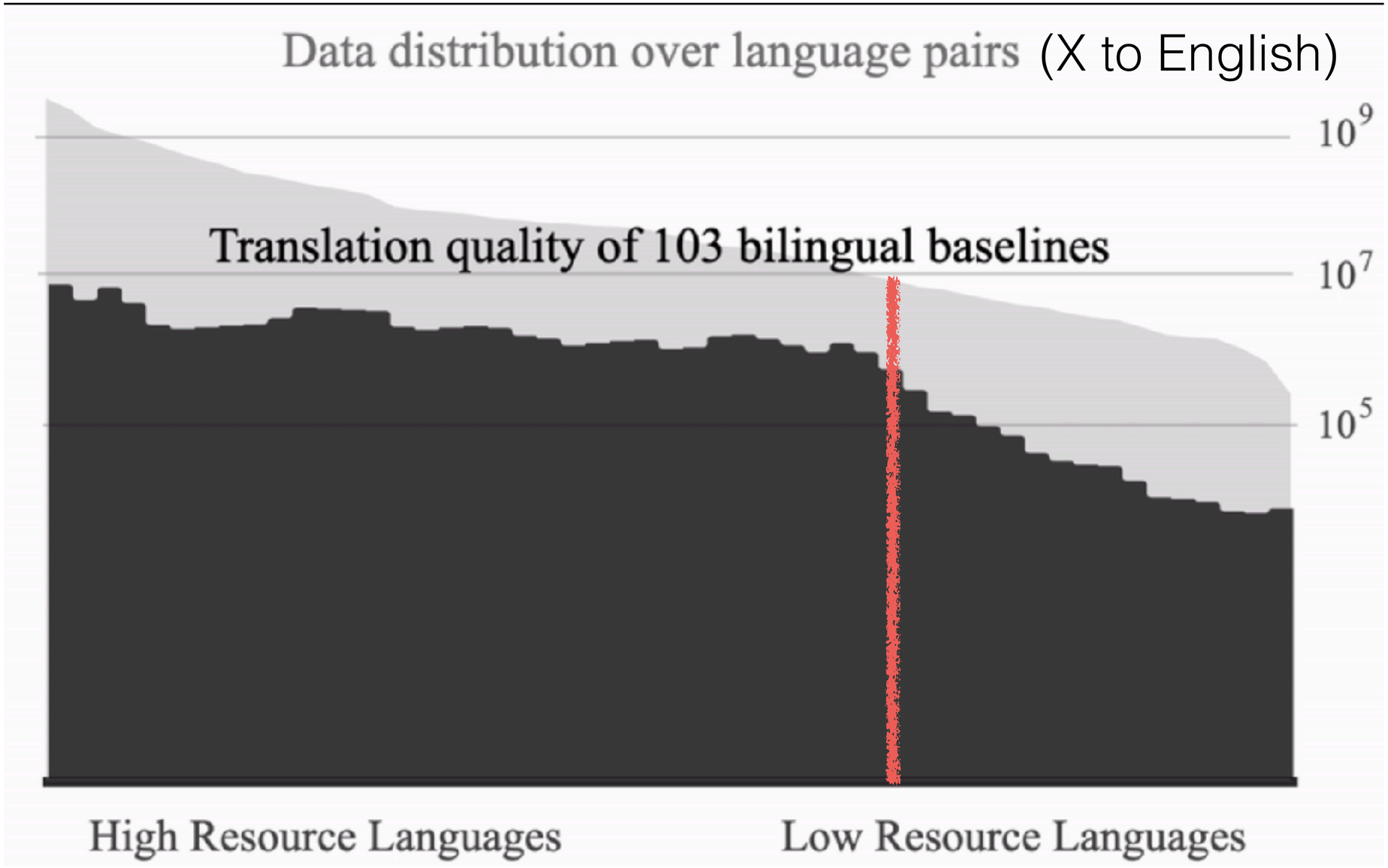
The Long Tail of Languages



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.



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



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

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

English **Nepali**

Training data

25M people

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 annotat
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 s
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
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**									

http://opus.nlpl.eu/

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 annotat
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 s
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
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**									

http://opus.nlpl.eu/

OPUS ... the open parallel corpus

OPU… linguistic annotat…
with … …ckage. We used s…
colle…

The …
Contr…

Searc…

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**									

http://opus.nlpl.eu/

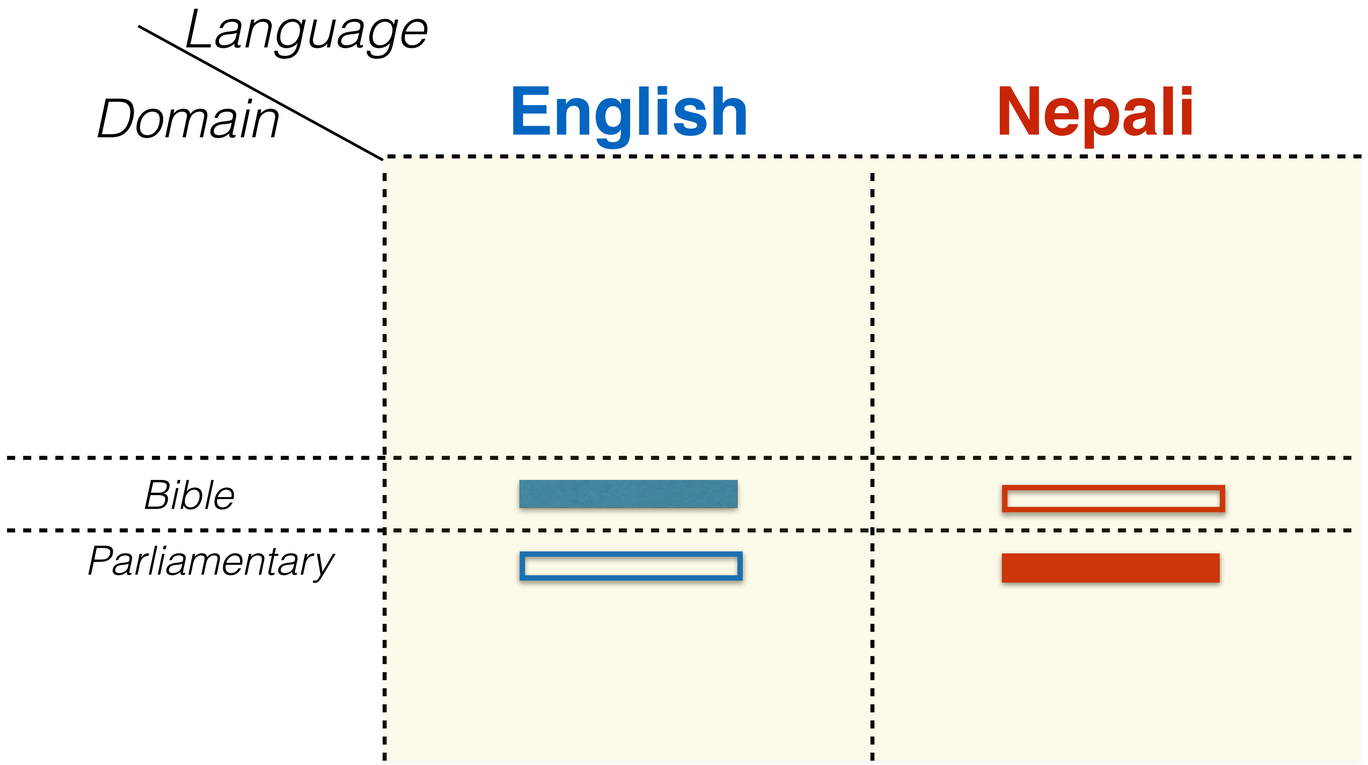
Machine Translation in Practice

English **Nepali**

Training data
(parallel sentences)

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

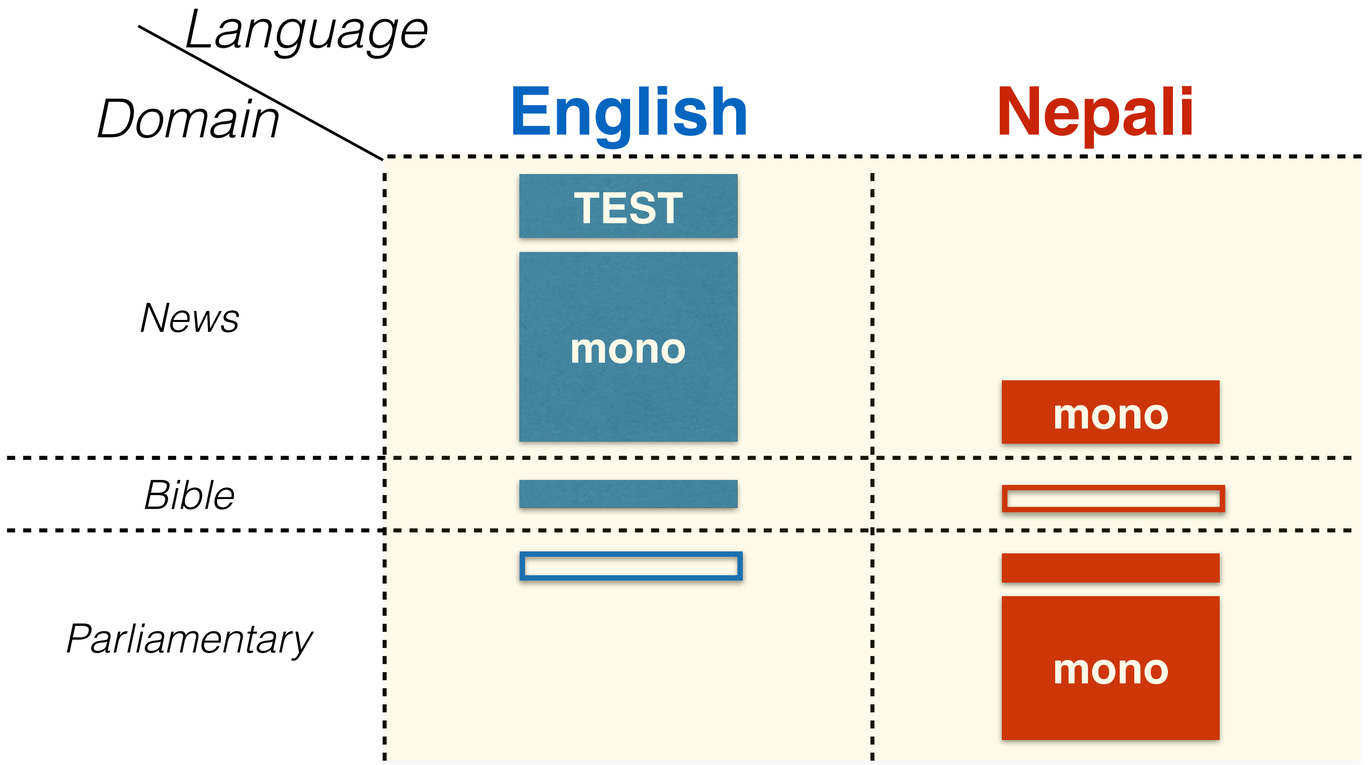
Machine Translation in Practice



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

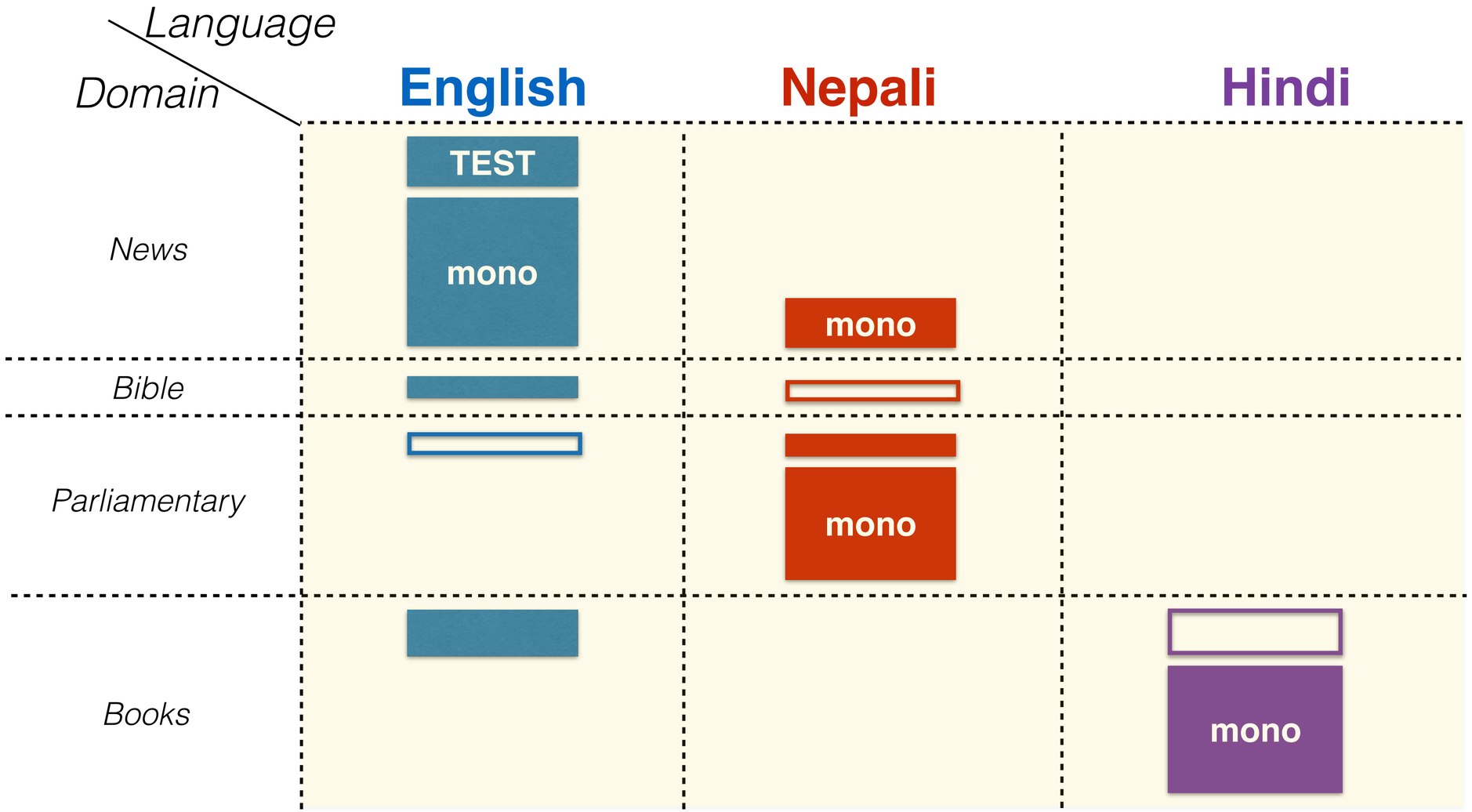
- Some parallel data originates in the source, some in the target language.
- Source and target domains may not match.

Machine Translation in Practice



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

Machine Translation in Practice



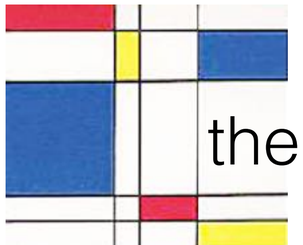
- There might be parallel and monolingual data with a high resource language close to the low resource language of interest. This data may belong to a different domain.

English Nepali Hindi Sinhala Bengali Spanish Tamil Gujarati

TEST

Domain

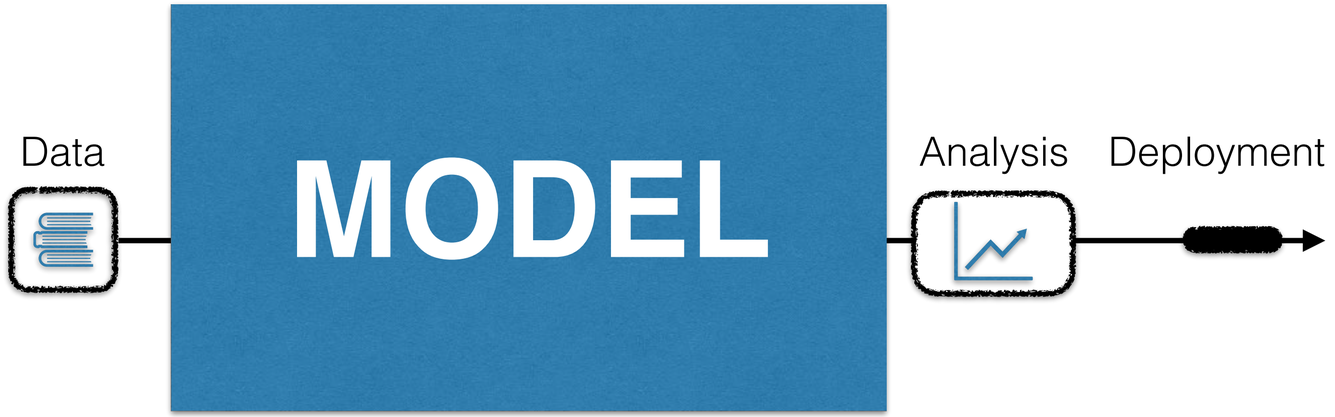
...



the *Mondrian* like learning setting! ...

Lesson #1

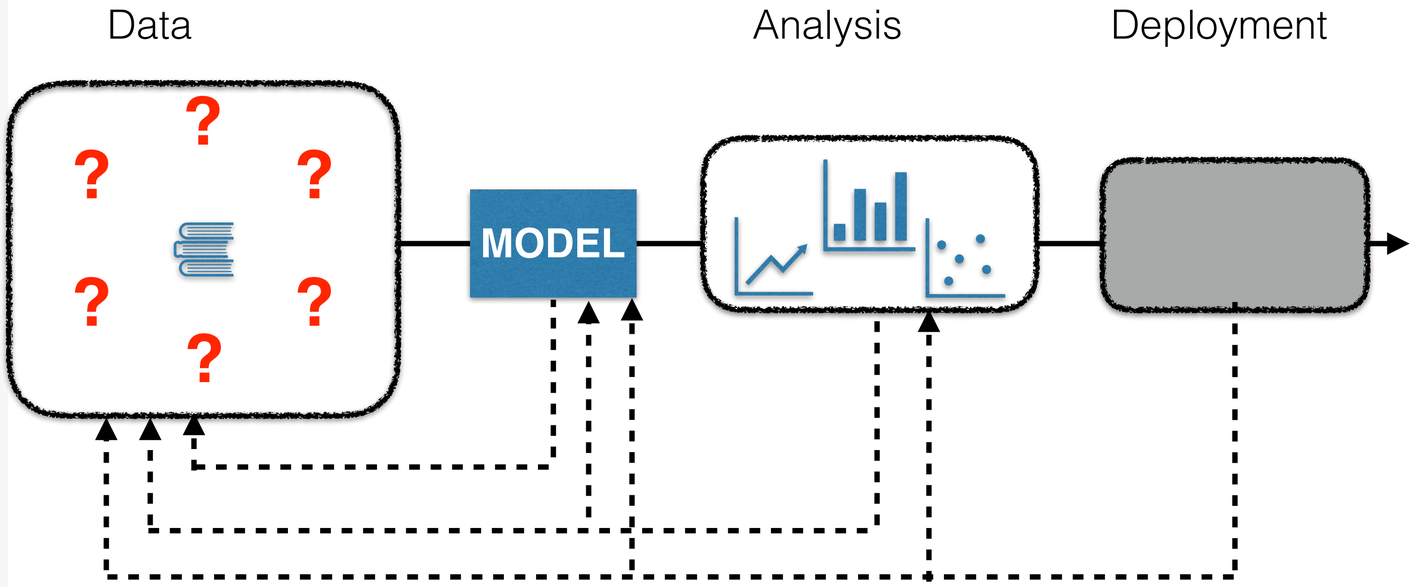
Dream Machine Learning



Model model model…

Lesson #1

Real Machine Learning

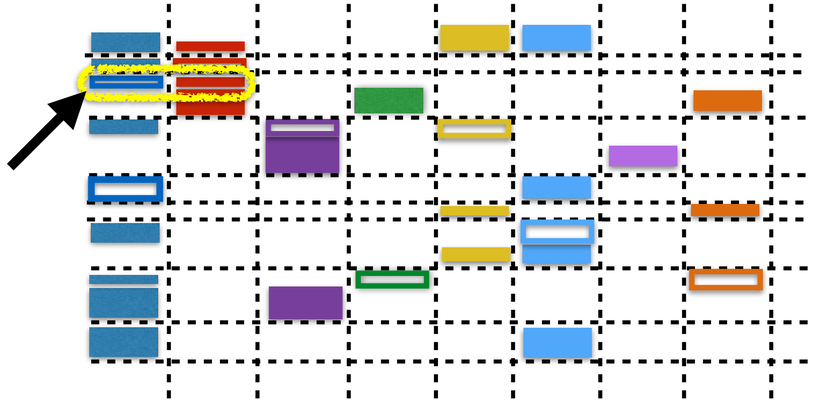


Data data data.. and iterate!

- Where is data coming from?
- Does the data surface the problem of interest?
- Does the data have biases?
- What properties does the data have?
- How to get more data..?

Lesson #2

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



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

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 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

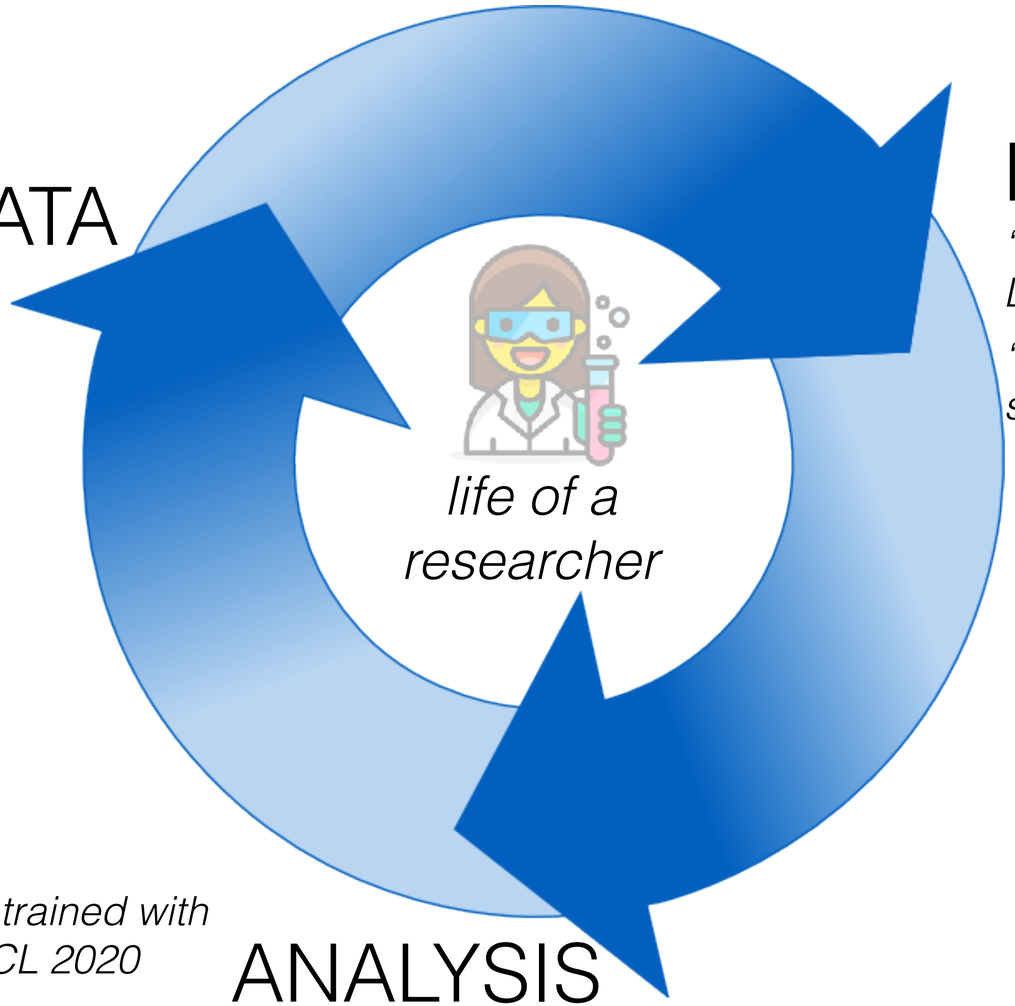
General MT challenges:

- Exposure bias (training for generation)
- Modeling uncertainty
- Automatic evaluation
- Budgeted computation
- Modeling the tails
- Efficiency
- …

Why Low Resource MT Is Interesting?

- It is about learning with little labeled data.
- It is about modeling structured outputs and compositional learning.
- It is a real problem to solve.

The MAD Cycle of Research



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

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

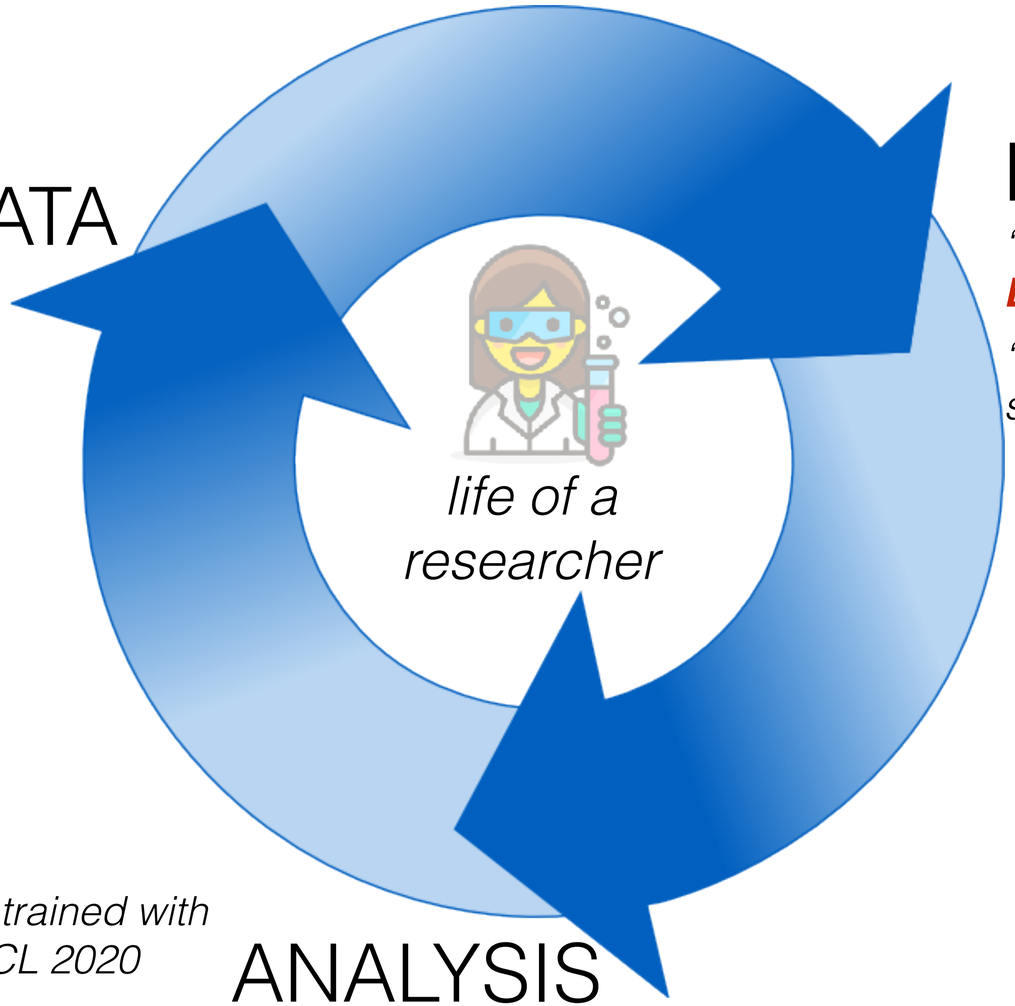
"FBAI WAT'19 My-En translation task submission" Chen et al., WAT@EMNLP 2019

"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

The Cycle of Research



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

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

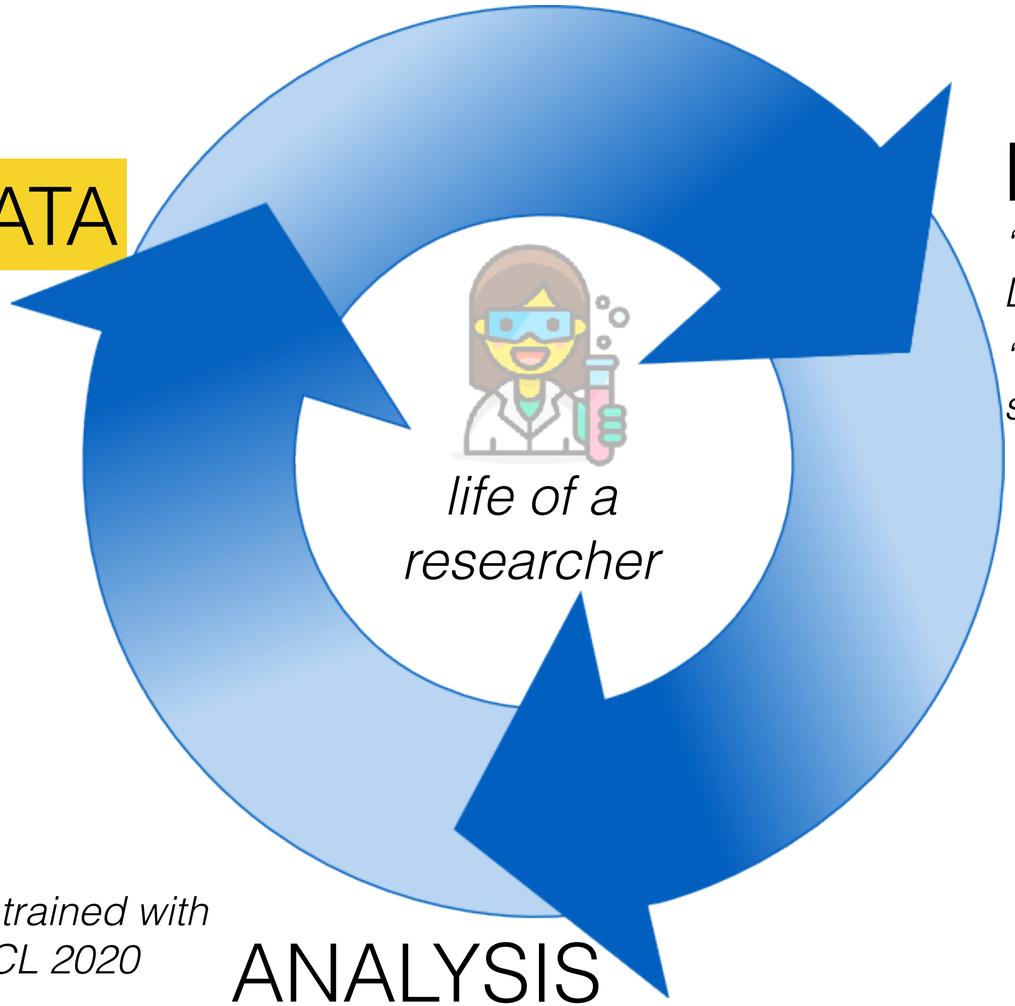
"FBAI WAT'19 My-En translation task submission" Chen et al., WAT@EMNLP 2019

"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

The Cycle of Research



"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

"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

A Big “Small-Data” Challenge

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, 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 >

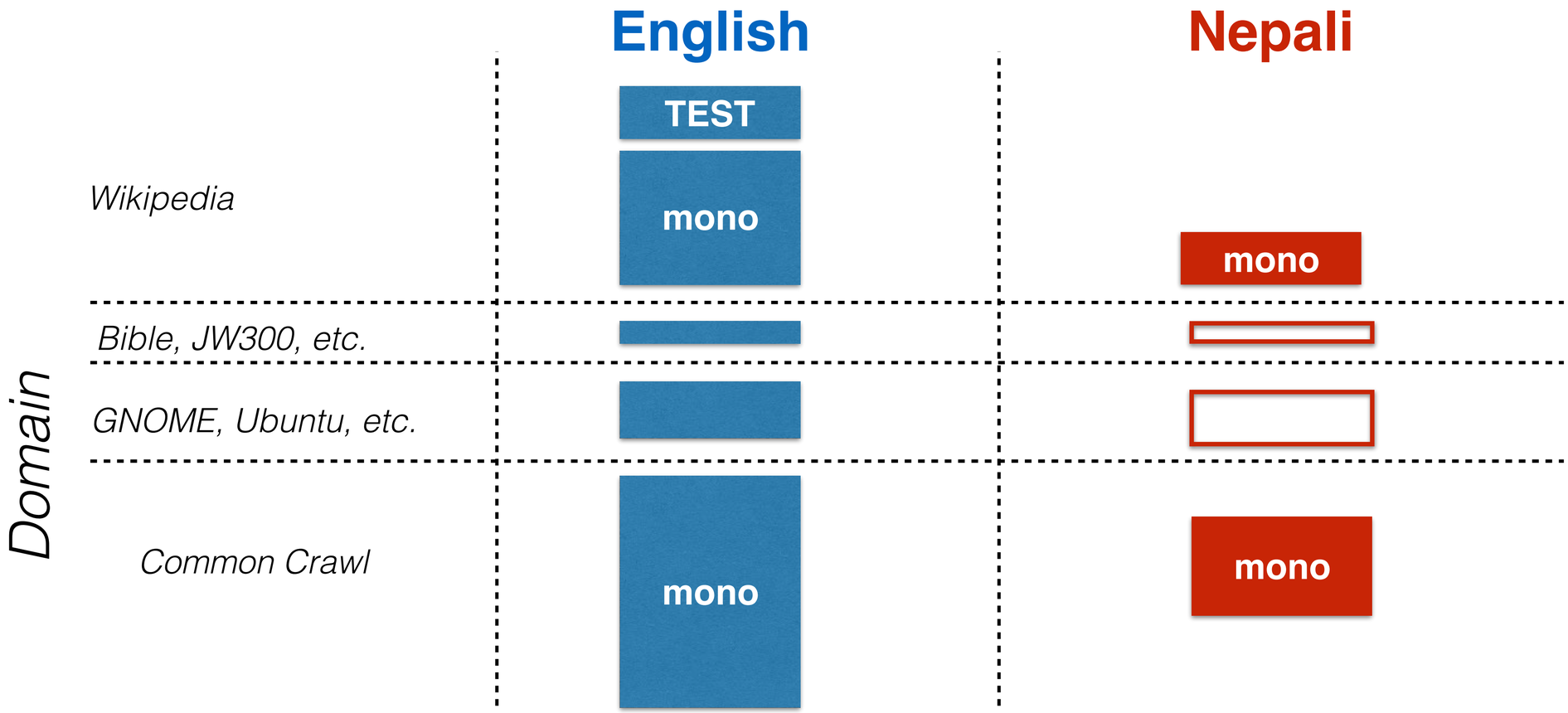
Search & download resources: en (English) | ne (Nepali) | all | show all versions

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**									

http://opus.nlpl.eu/

Case Study: En-Ne



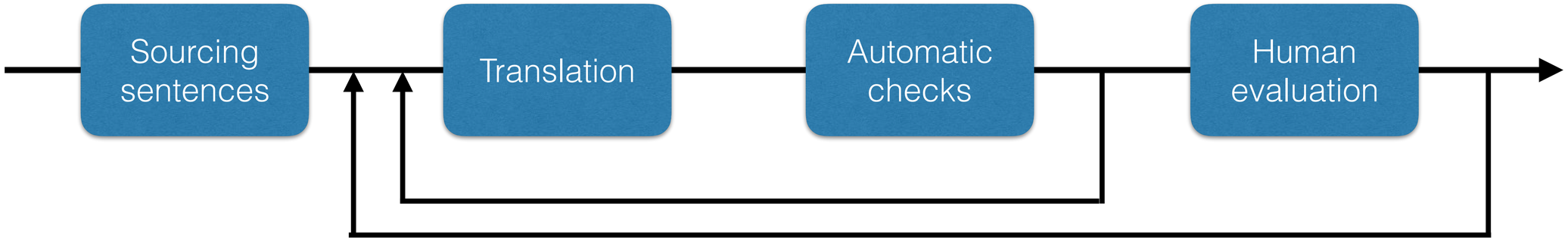
In-domain data: no parallel, little monolingual.
Out-of-domain: little parallel, quite a bit monolingual
No translation originating from Nepali.

FLoRes Evaluation Benchmark

- Validation, test and hidden test set, each with 3000 sentences: 1500 from En, 1500 from X, with X = {Nepali, Sinhala, Khmer, Pashto}.
- Sentences taken from *Wikipedia* documents.

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

Flores: Data Collection Process



- Very expensive and slow.
- Very hard to produce high-quality translations:
  - create guidelines for translators and evaluators,
  - define automatic checks, such as language model filtering, transliteration filtering, length filtering, language id filtering, etc.
  - define suitable thresholds for human assessment.

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

Examples

Si-En

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

original

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

නර්ජන , ශාරීරික හිංසනය , දේපල හානිය , පහර දීම සහ මරාදැමීම මෙම දඩුවම්ය .

translation

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

En-Si

In Serious meets, the absolute score is somewhat meaningless.

සැබෑ නරඟ වලදී ලකුණු සැසඳීම තේරුමක් නැති ක්‍රියාවකි .

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

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

අයිෆෝන් භාවිත කරන්නන්ට 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 ।

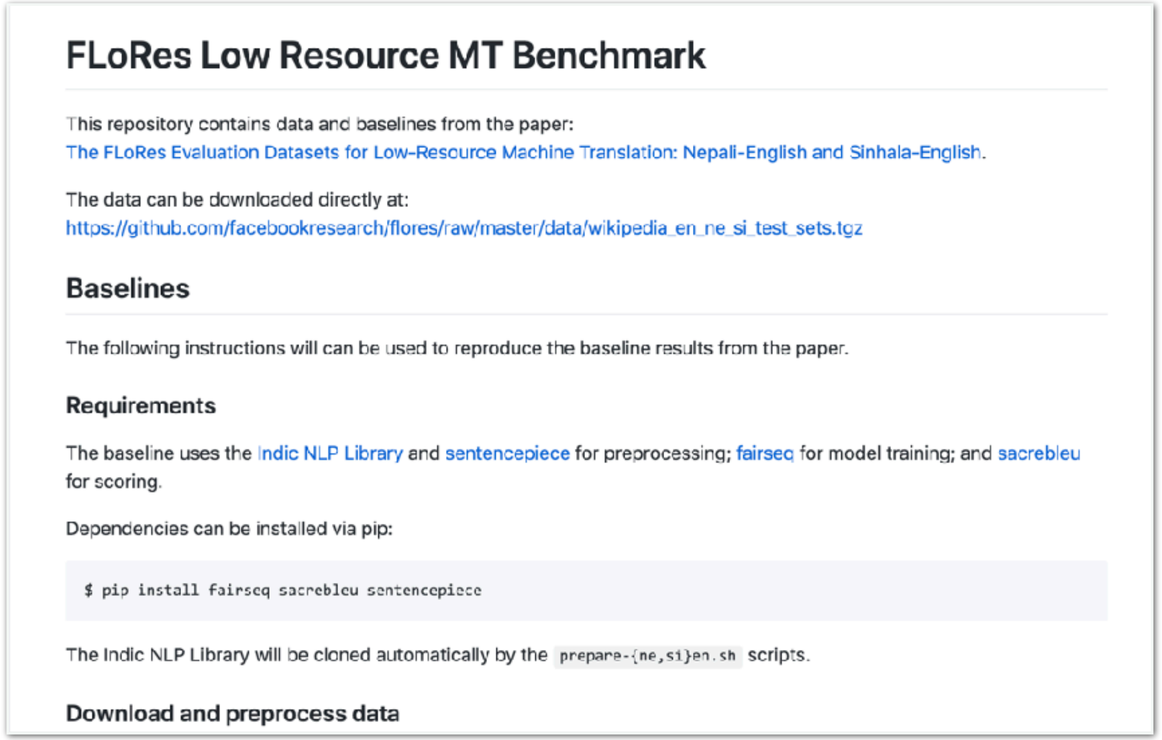
It has automatic spell checking and correction, predictive word capabilities, and a dynamic dictionary that learns new words.

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
```

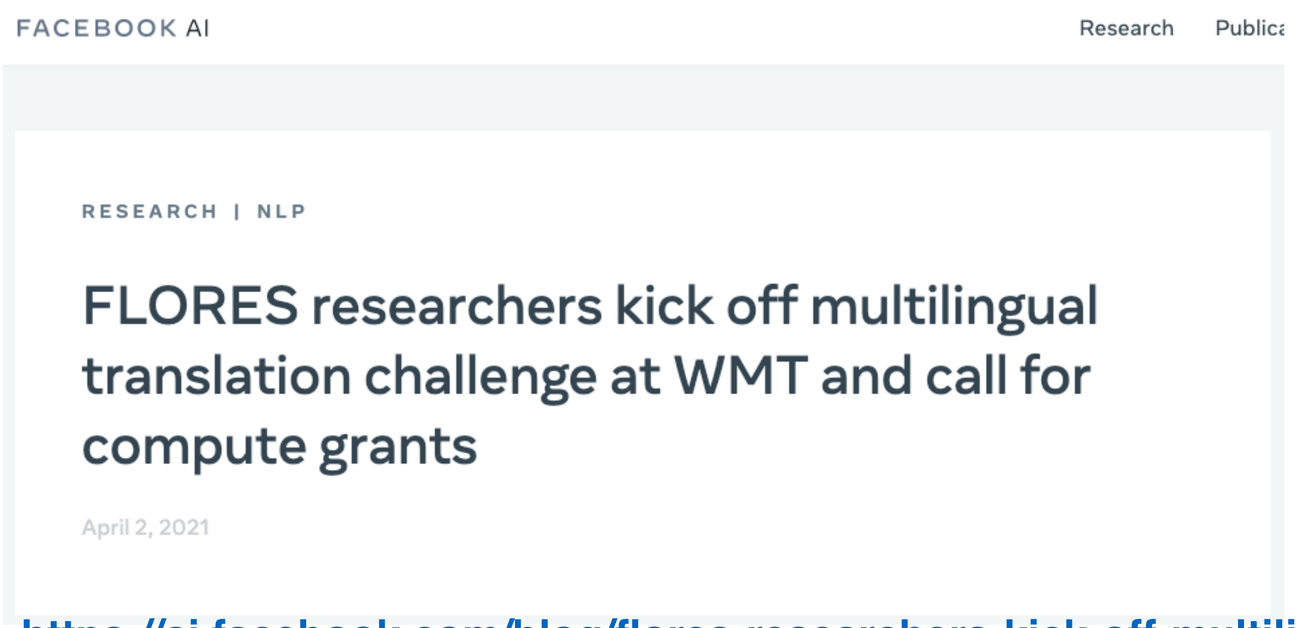
The Indic NLP Library will be cloned automatically by the `prepare-{ne,si}en.sh` scripts.

Download and preprocess data

https://github.com/facebookresearch/flores

data & baseline models

Upcoming Flores 101
(June 2021)



https://ai.facebook.com/blog/flores-researchers-kick-off-multilingual-translation-challenge-at-wmt-and-call-for-compute-grants/

EMNLP 2021
SIXTH CONFERENCE ON
MACHINE TRANSLATION (WMT21)

November 10-11 , 2021
Punta Cana, Dominican Republic

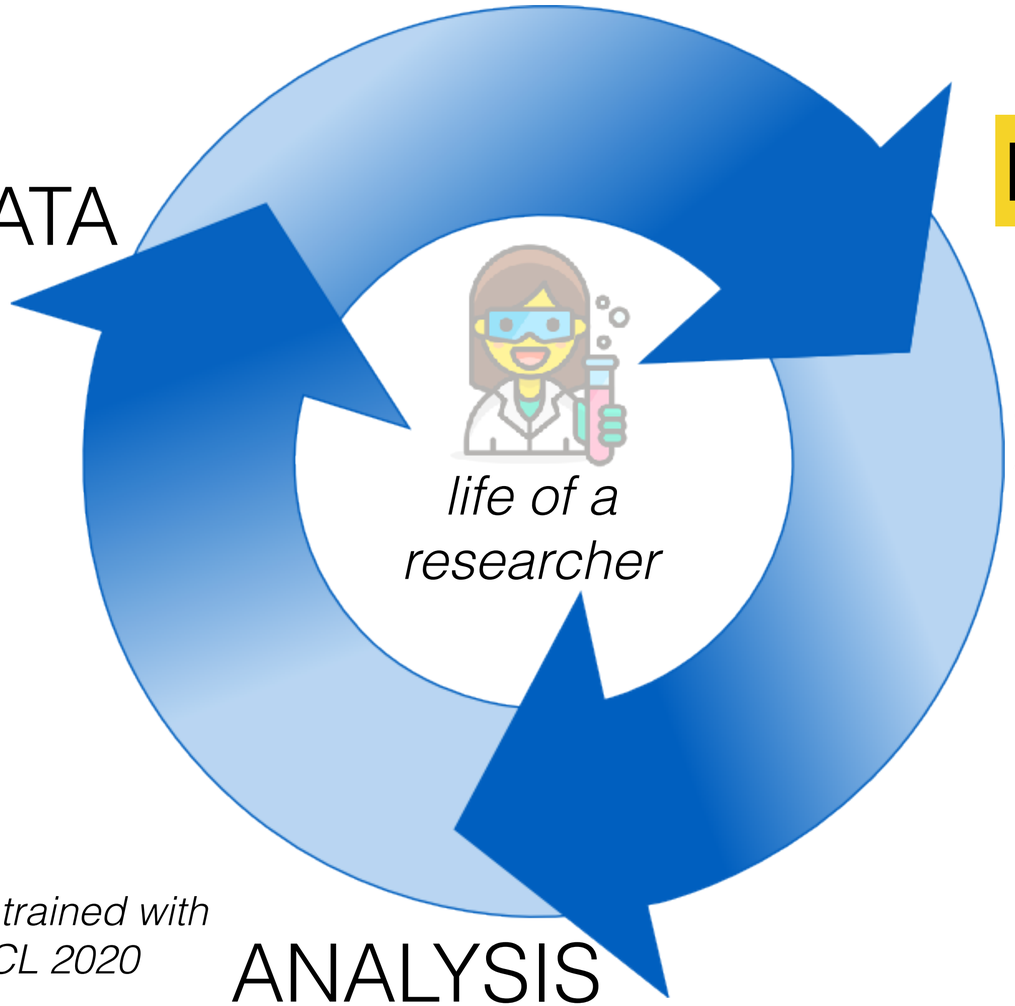
Shared Task: Large-Scale Multilingual Machine Translation

http://statmt.org/wmt21/large-scale-multilingual-translation-task.html

Lesson #3

- Data is often as or more important than designing a model.
- Data collection is not trivial.
- Look at the data!!

The Cycle of Research



"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

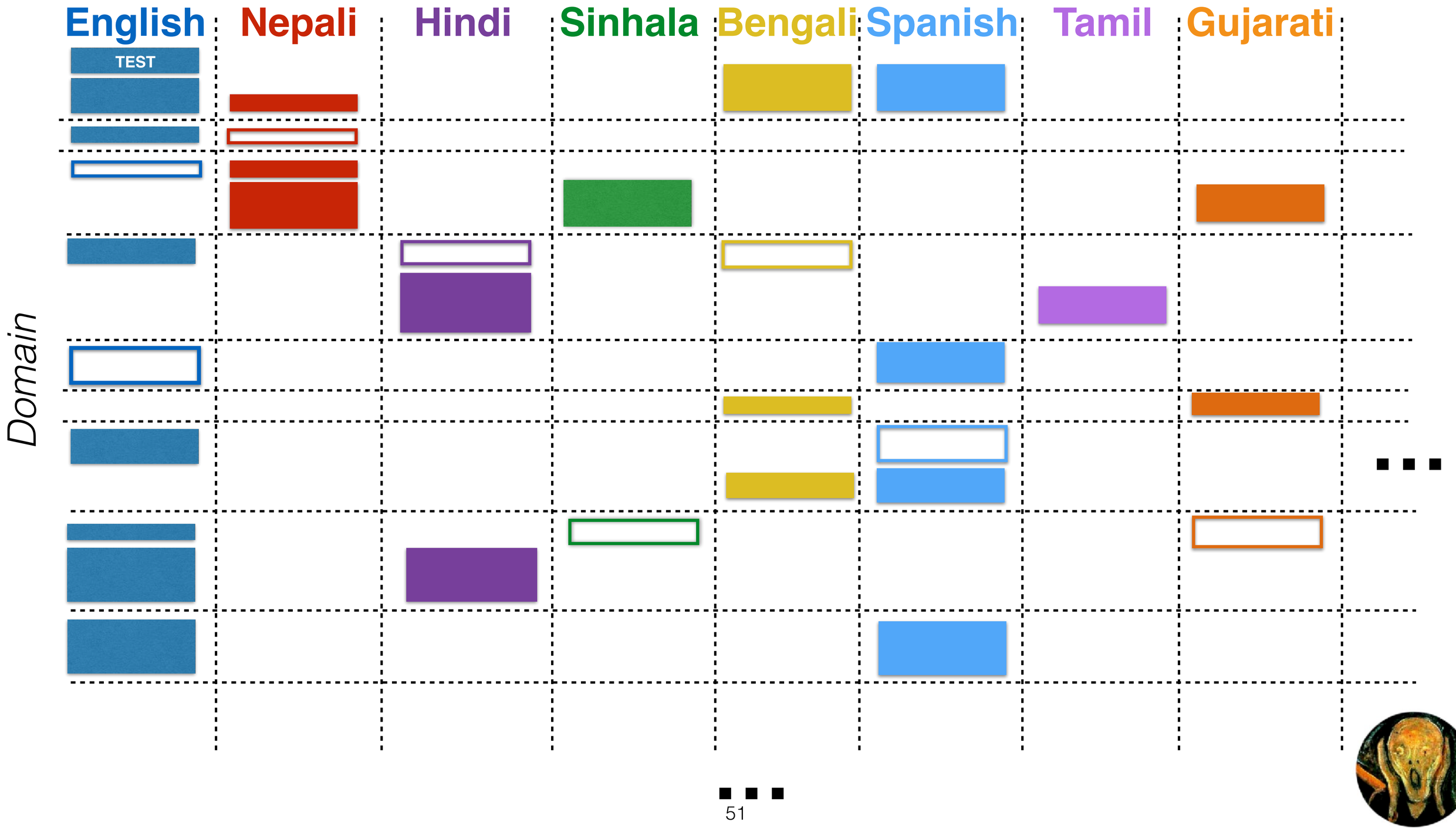
"Analyzing uncertainty in NMT" Ott et al. ICML 2018

"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

Outline

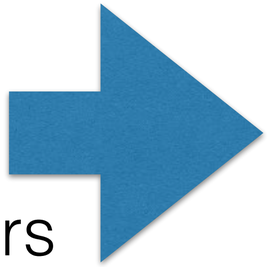
- **ML perspective on low resource MT**
- Case studies:
  - Unsupervised MT
  - En-Ne
  - En-My
- Perspectives


English
Nepali
Hindi
Sinhala
Bengali
Spanish
Tamil
Gujarati
TEST
Domain

ML Perspective

NLP/MT Data

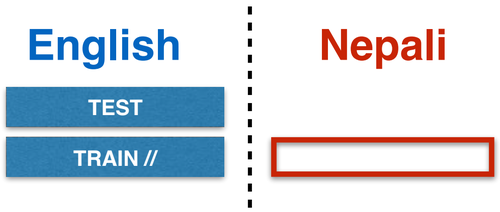
- Parallel dataset
- Monolingual data
- Multiple language pairs
- Multiple domains



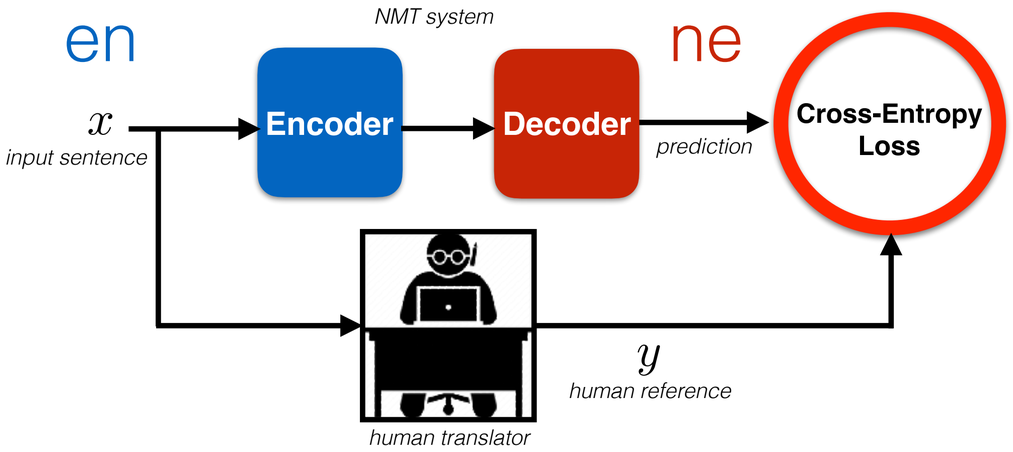
ML Techniques

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

Supervised Learning



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



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

usual attention-based transformer

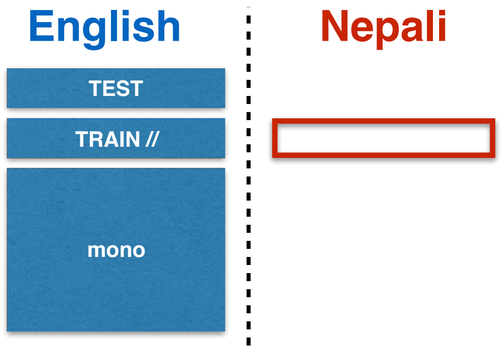
Regularize the model using:

- 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)

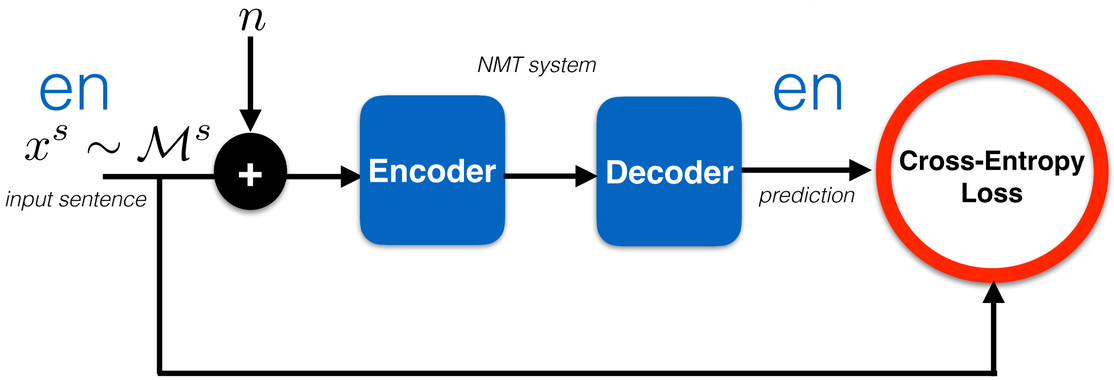


$$\mathcal{D} = \{(x, y)_i\}_{i=1,..,N}$$

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

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.

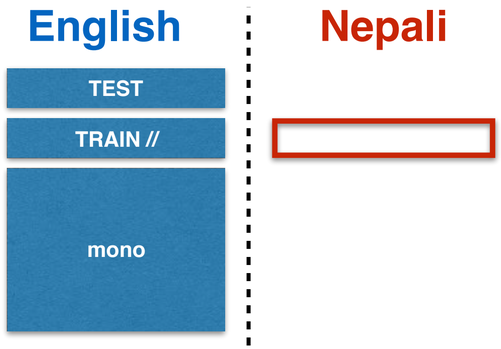
Learning Framework: DAE

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

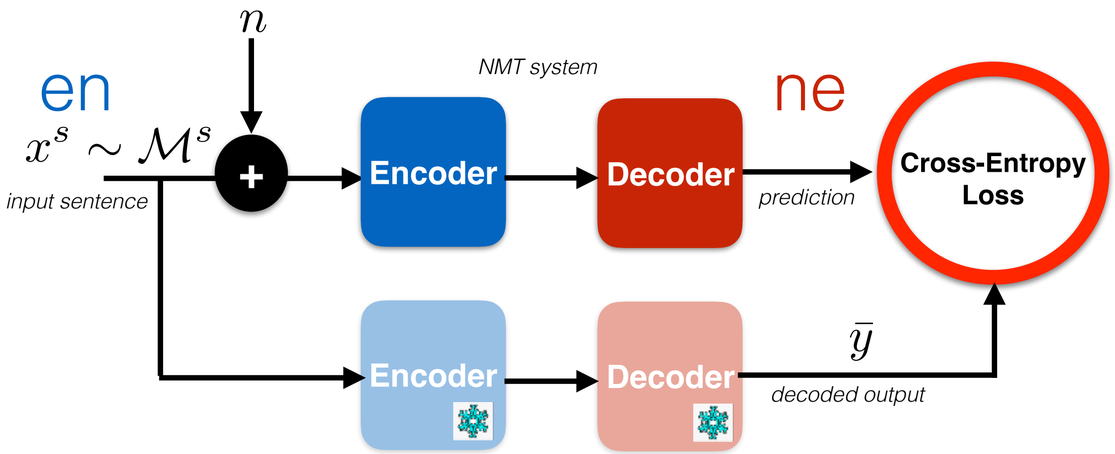
Semi-Supervised Learning (ST)



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

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

Idea: predict missing labels.



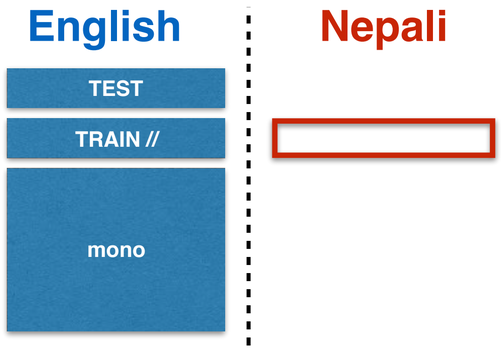
$$\mathcal{L}^{ST}(\theta) = -\log p(\bar{y}|x+n)$$

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

Key elements: decoding and training noise.

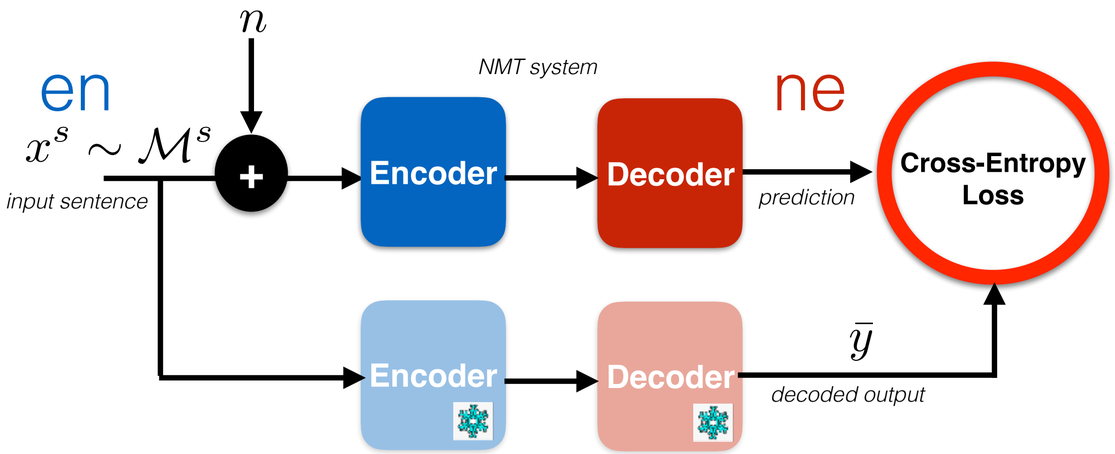
He et al. "Revisiting self-training for neural sequence generation" ICLR 2020

Semi-Supervised Learning (ST)



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

Idea: predict missing labels.



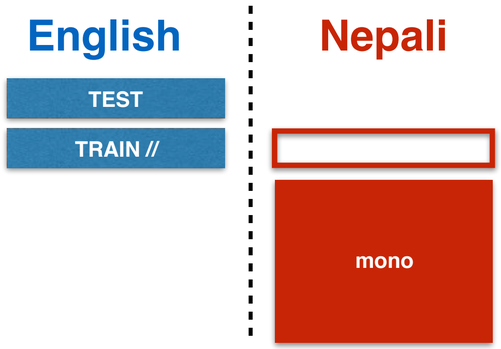
Key elements: decoding and training noise.

ALGORITHM

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

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

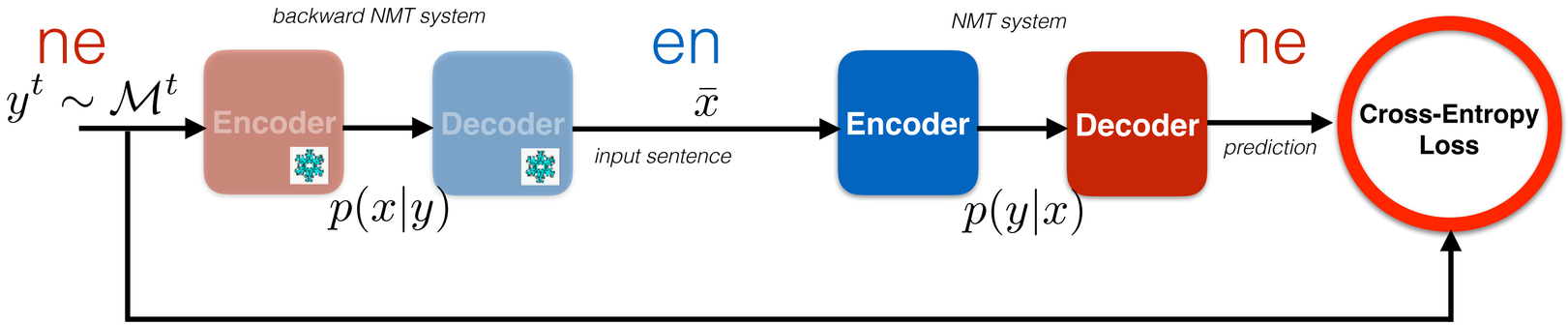
Semi-Supervised Learning (BT)



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

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

Additional target side monolingual data.



Learning Framework:

Back-Translation (BT).

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

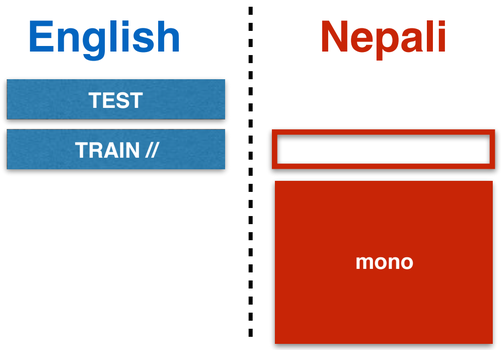
$$\mathcal{L}(\theta) = \mathcal{L}^{\text{sup}}(\theta) + \lambda \mathcal{L}^{\text{BT}}(\theta)$$

Adding target-side monolingual data.
Two benefits:

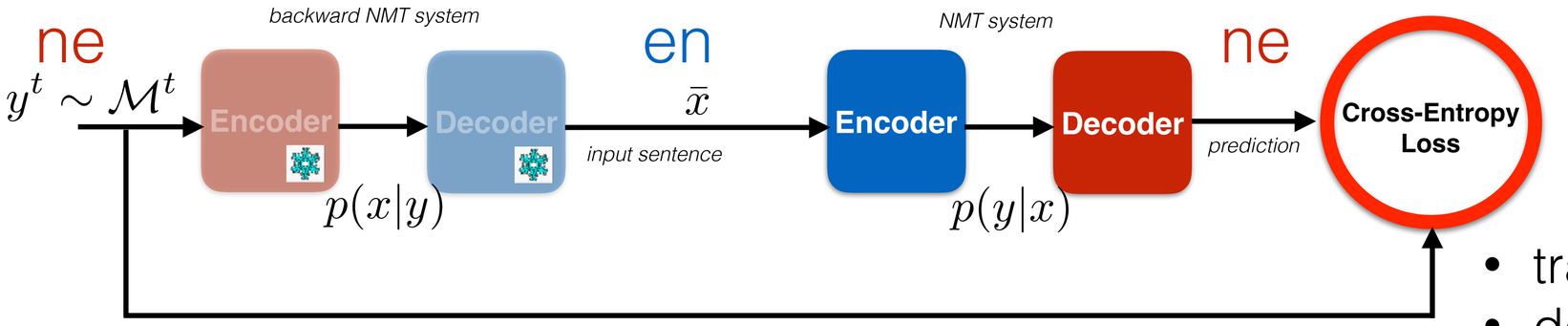
a) Decoder learns a good language model.
b) Better generalization via data augmentation.
c) Unlike ST, target is correct but input is not.

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

Semi-Supervised Learning (BT)



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



Adding target-side monolingual data.
Two benefits:

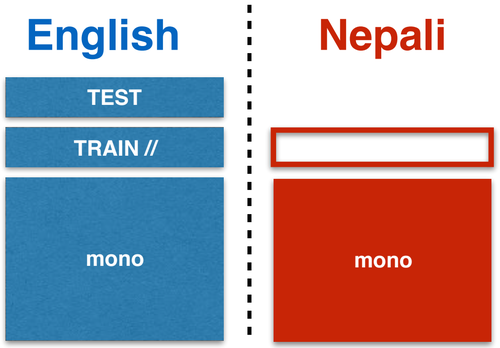
a) Decoder learns a good language model.
b) Better generalization via data augmentation.
c) Unlike ST, target is correct but input is not.

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 $\mathcal{A}^t = \{(\bar{x}_k, y_k^t)\}_{k=1,..,M_t}$
- train model $p(y|x)$ on: $\mathcal{D} \cup \mathcal{A}^t$

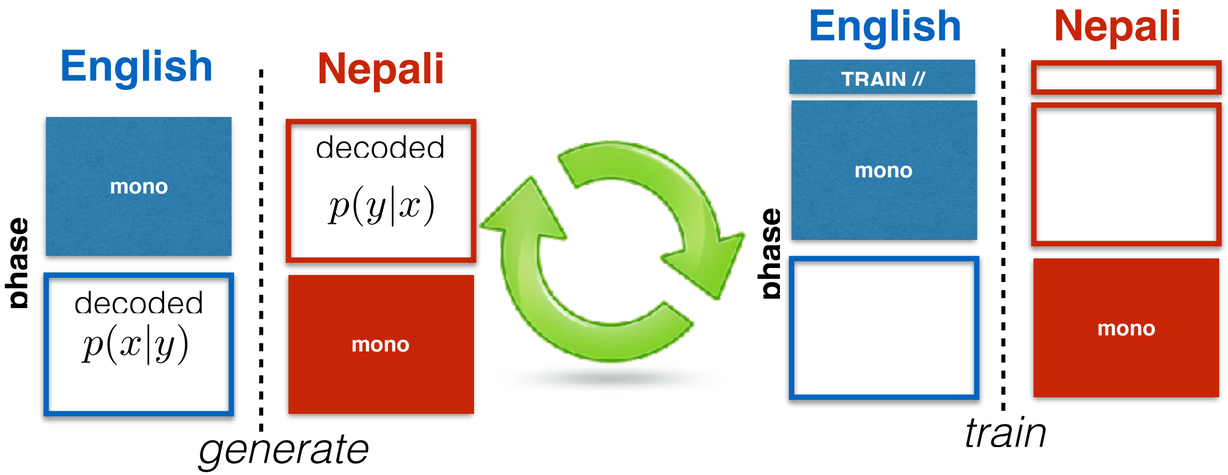
Sennrich et al. "Improving NMT models with monolingual data" ACL 2016

Semi-Supervised Learning (ST+BT)



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

Additional source & target side monolingual data.



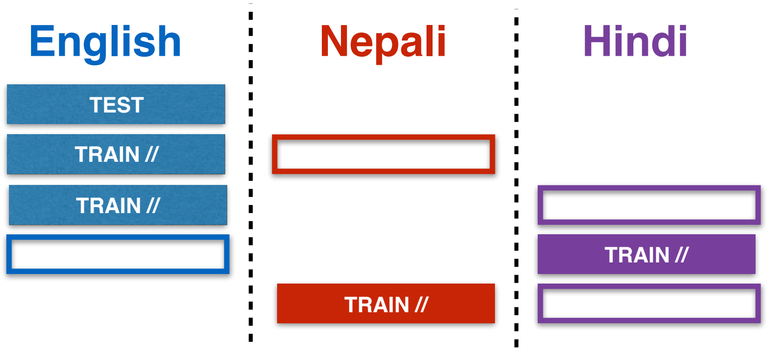
ALGORITHM

- train model $p(x|y)$ and $p(y|x)$ on $\mathcal{D}$
- repeat
  - **phase 1**
    - 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 $\mathcal{A}^s = \{(x_j^s, \bar{y}_j)\}_{j=1,..,M_s}$
  - **phase 2**
    - retrain both $p(y|x)$ and $p(x|y)$ on: $\mathcal{D} \cup \mathcal{A}^t \cup \mathcal{A}^s$

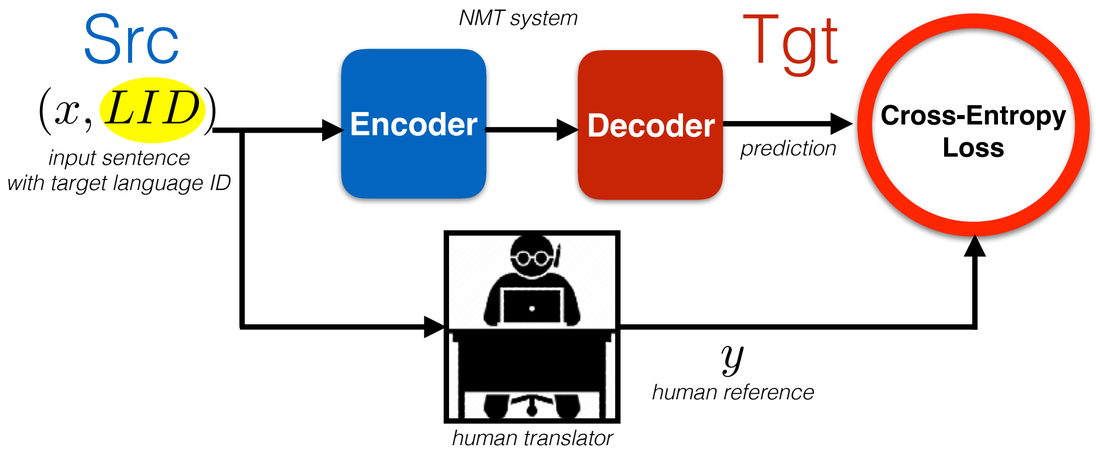
$$\mathcal{L}^{\text{total}}(\theta) = -\log p(y|x) - \lambda_1 \log p(y^t|\bar{x}^t) - \lambda_2 \log p(\bar{y}^s|x^s)$$

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

Multi-Task/Multi-Modal Learning



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



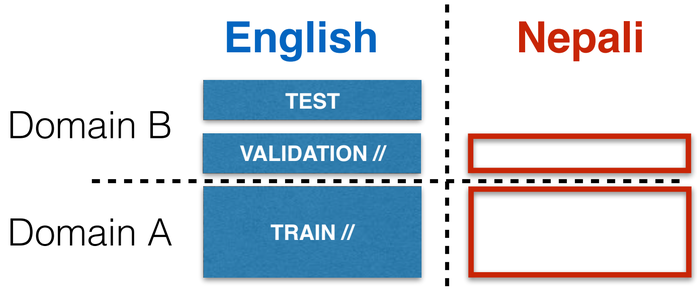
Learning Framework: Multilingual Training

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.

Lesson #4

Several basic learning approaches can be used and combined to tackle low-resource MT.

General ML Tip: 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.

Lesson #5

End-to-end learning: there is nothing too specific to the task and language pair.

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

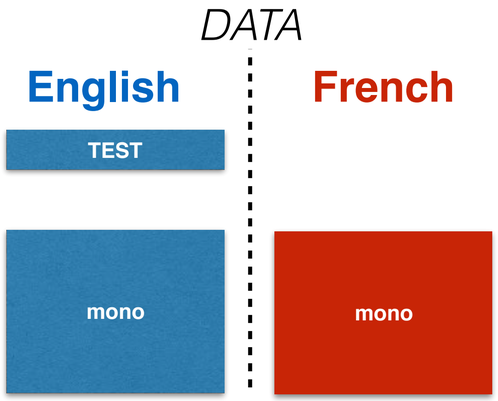
Conclusion so far…

- 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.

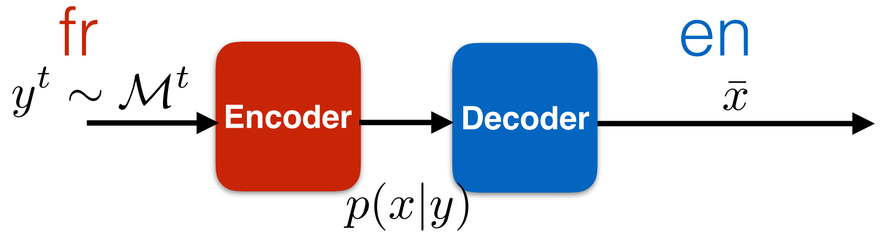
Outline

- ML perspective on low resource MT
- **Case studies:**
  - Unsupervised MT
  - En-Ne
  - En-My
- Perspectives

Case Study #1: Unsupervised MT

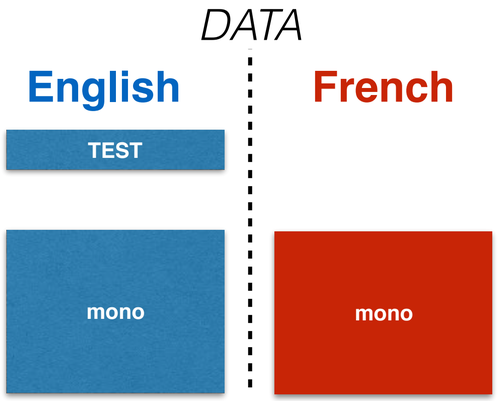


$$\mathcal{M}^t = \{y_k^t\}_{k=1,..,M_t}$$
$$\mathcal{M}^s = \{x_j^s\}_{j=1,..,M_s}$$

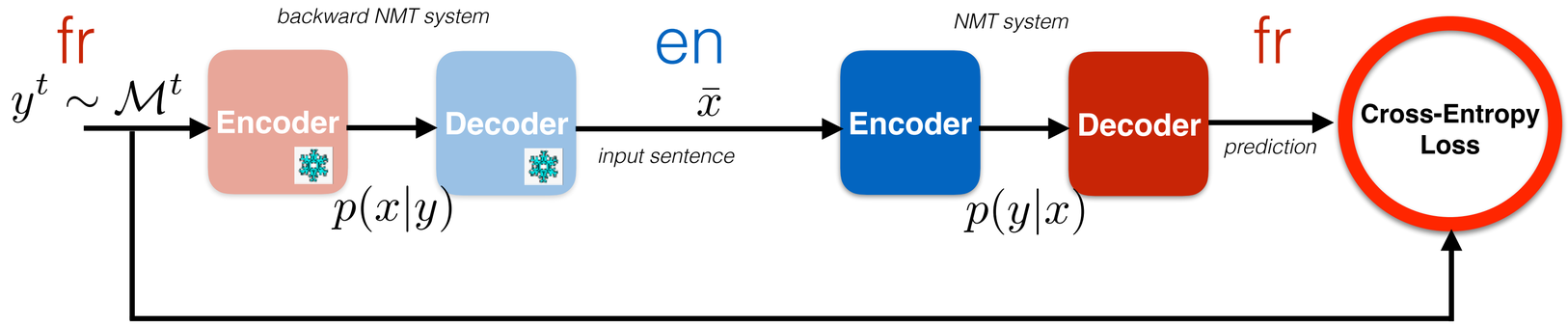


Lample et al. "Phrase-based and neural unsupervised MT" EMNLP 2018
Artetxe et al. "An effective approach to unsupervised MT" ACL 2019

Case Study #1: Unsupervised MT



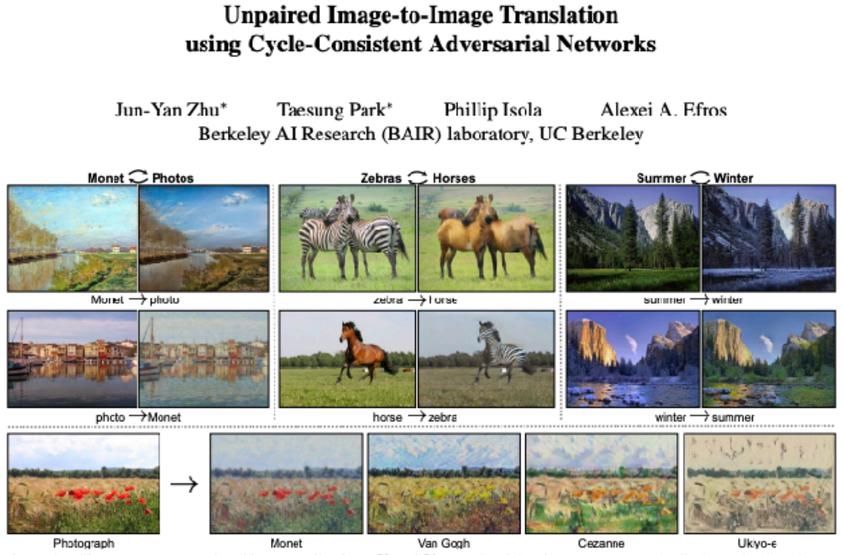
$$\mathcal{M}^t = \{y_k^t\}_{k=1,..,M_t}$$
$$\mathcal{M}^s = \{x_j^s\}_{j=1,..,M_s}$$



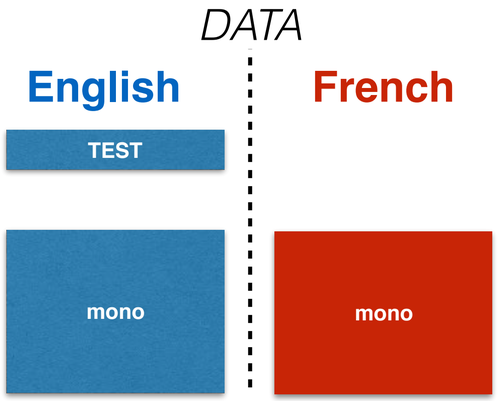
…and vice versa starting from English.

This is an example of auto-encoding or cycle consistency.

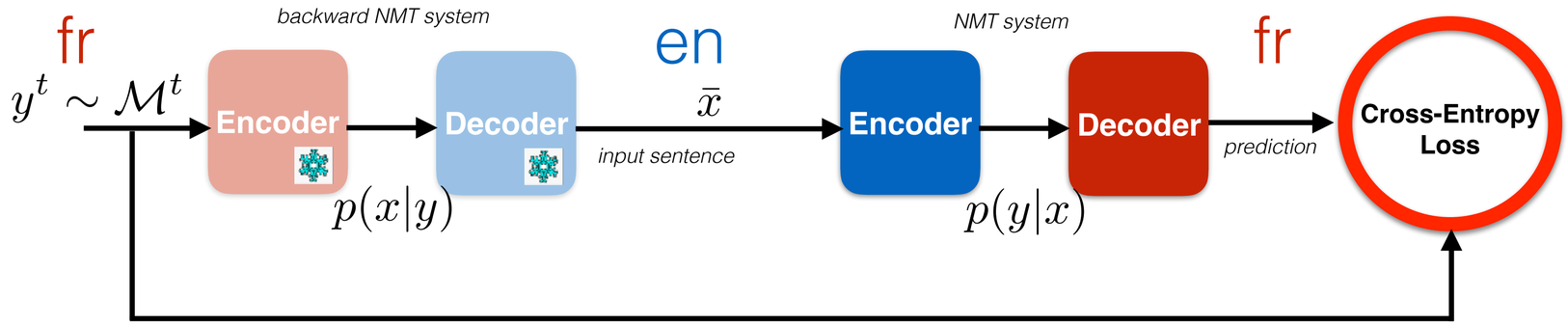
Problem: lack of constrains on $\bar{x}$

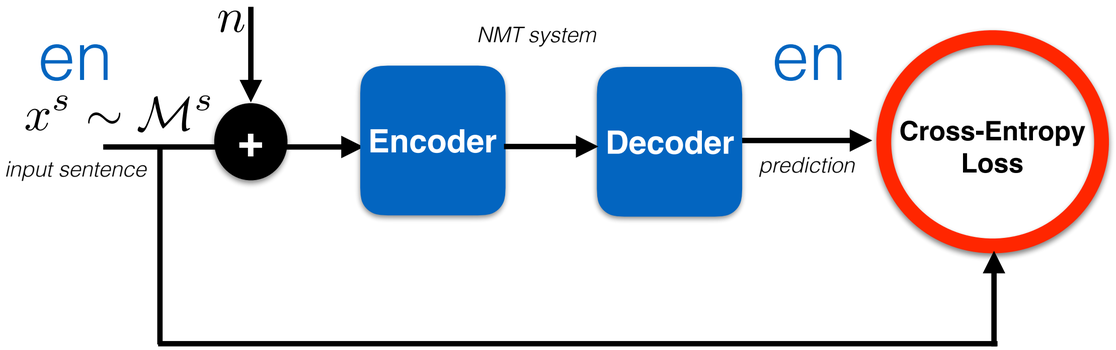


Case Study #1: Unsupervised MT



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





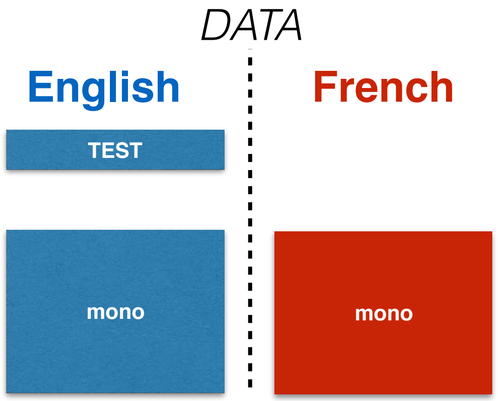
DAE makes sure decoder outputs fluently in the desired language.

Problem: lack of modularity.

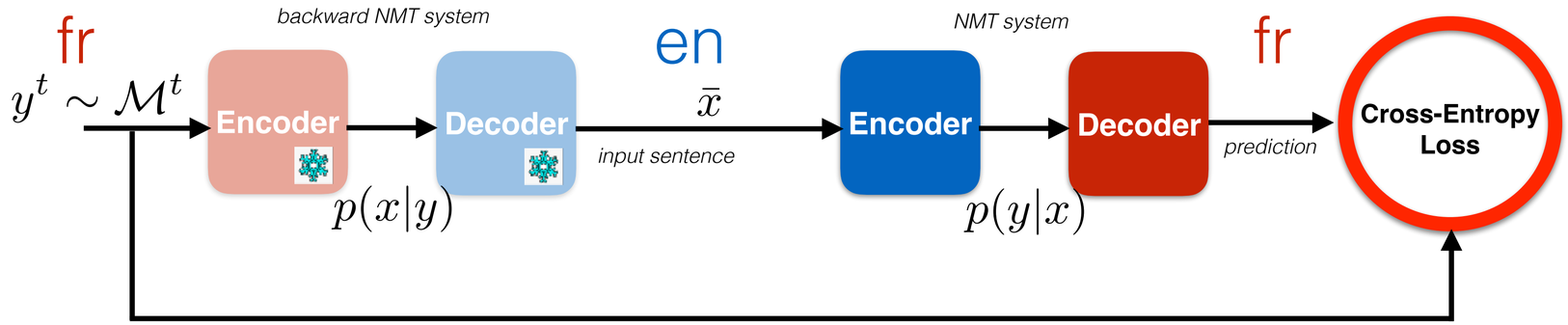
Decoder may behave differently when fed with representations from French encoder VS English encoder.

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

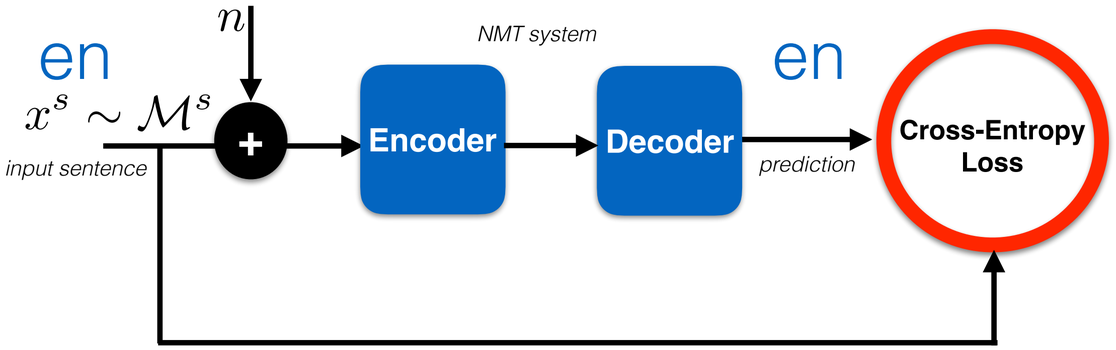
Case Study #1: Unsupervised MT



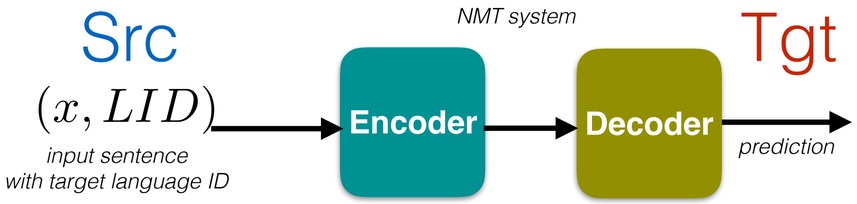
$$\mathcal{M}^t = \{y_k^t\}_{k=1,..,M_t}$$
$$\mathcal{M}^s = \{x_j^s\}_{j=1,..,M_s}$$



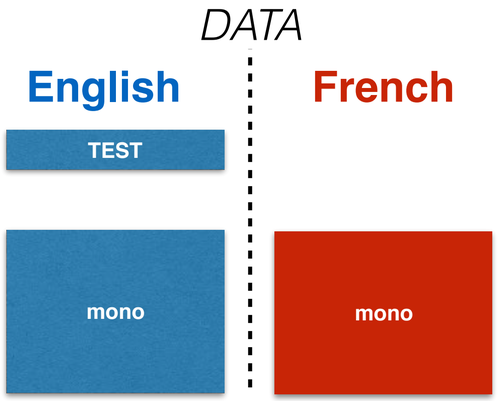
DAE makes sure decoder outputs fluently in the desired language.



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



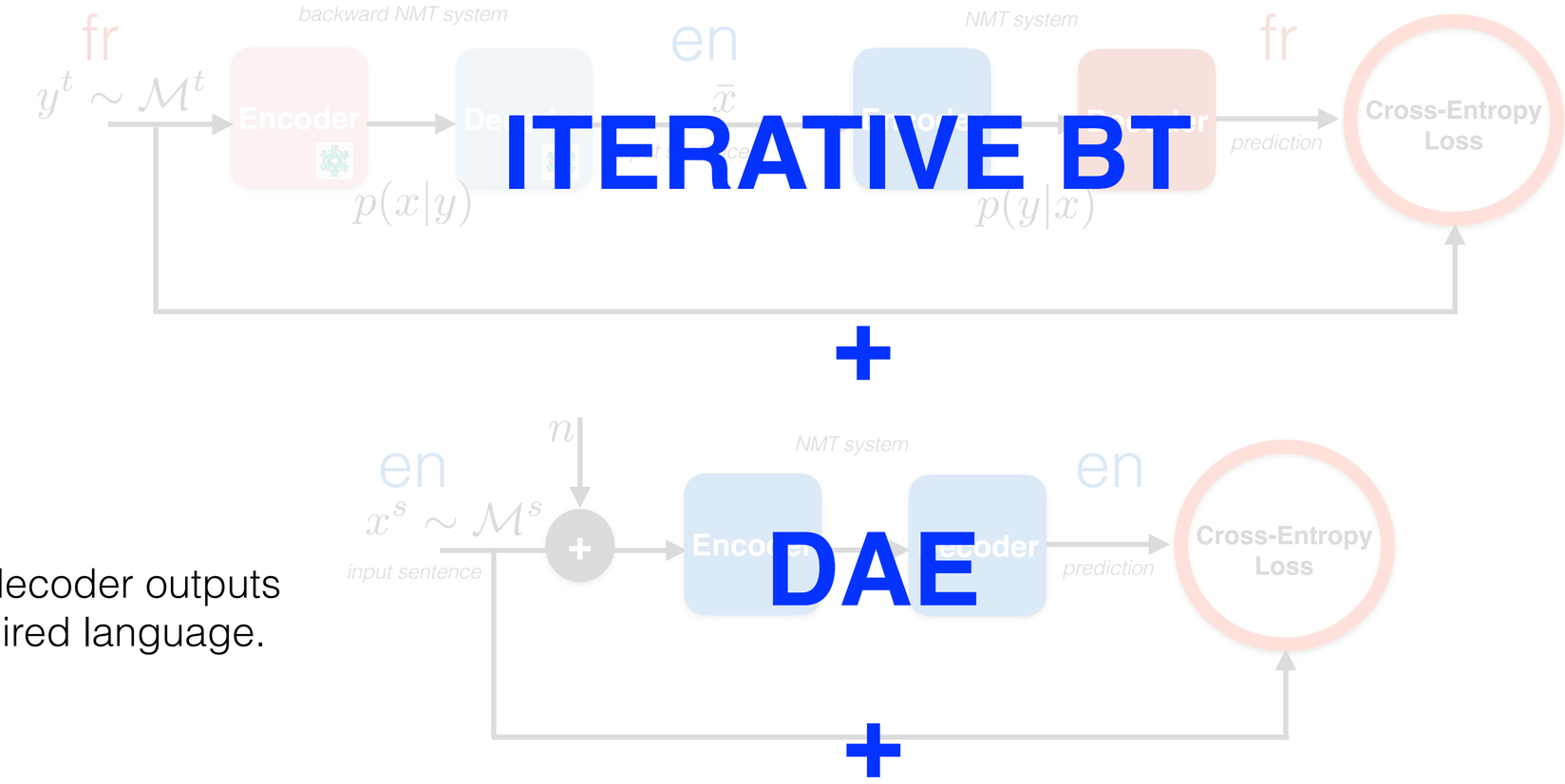
Case Study #1: Unsupervised MT

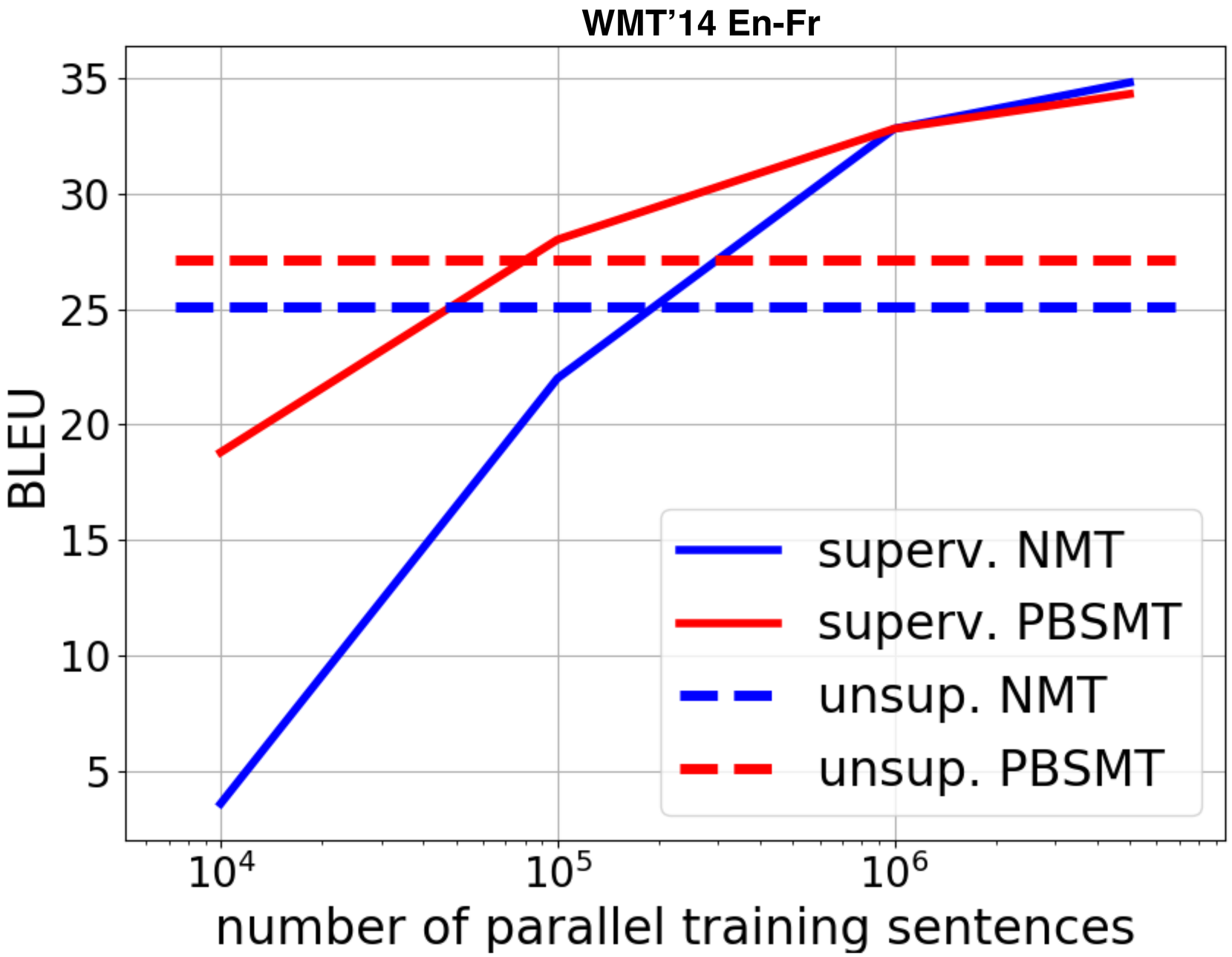


$$\mathcal{M}^t = \{y_k^t\}_{k=1,..,M_t}$$
$$\mathcal{M}^s = \{x_j^s\}_{j=1,..,M_s}$$

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).





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

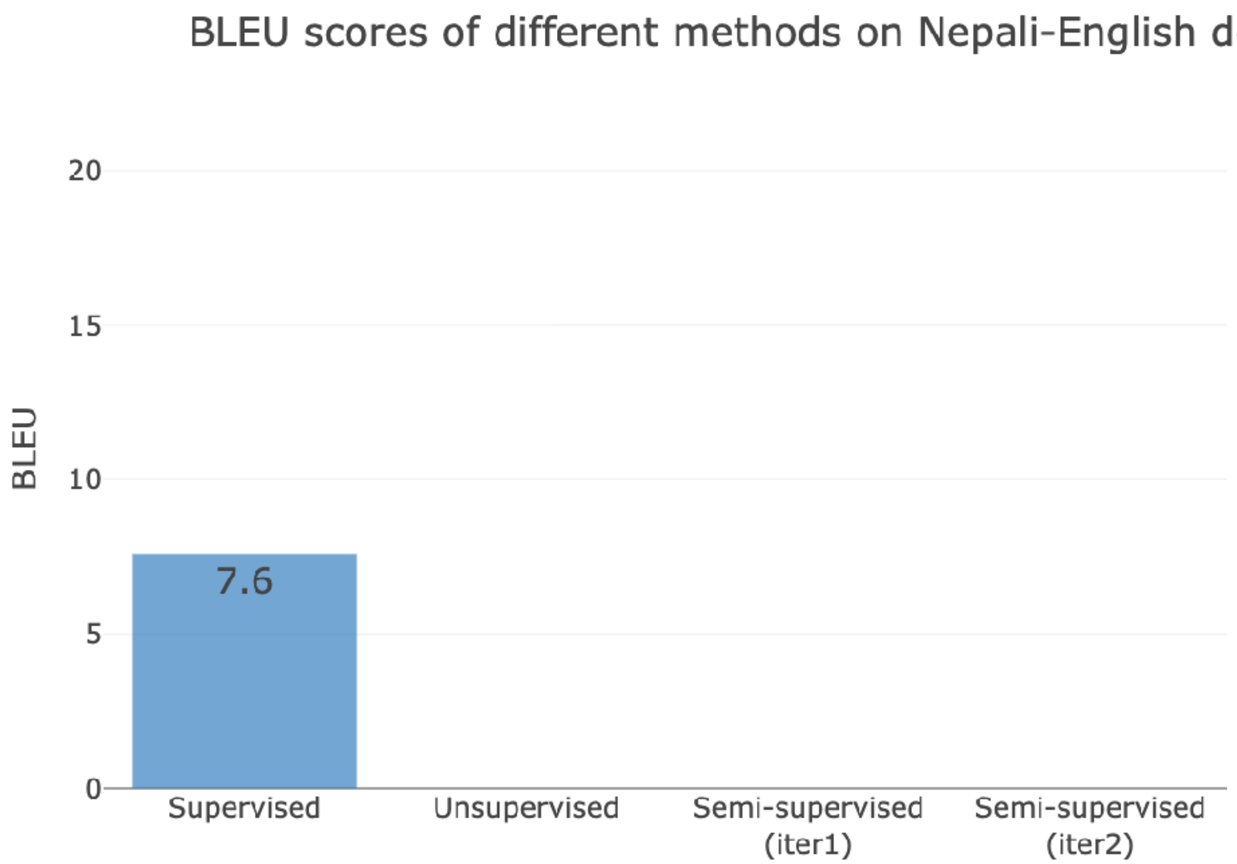
Since unsupMT was trained on about 10M sentences, each parallel sentence is worth 100 monolingual sentences (for this dataset and language pair).

Lample et al. "Phrase-based and neural unsupervised MT" EMNLP 2018

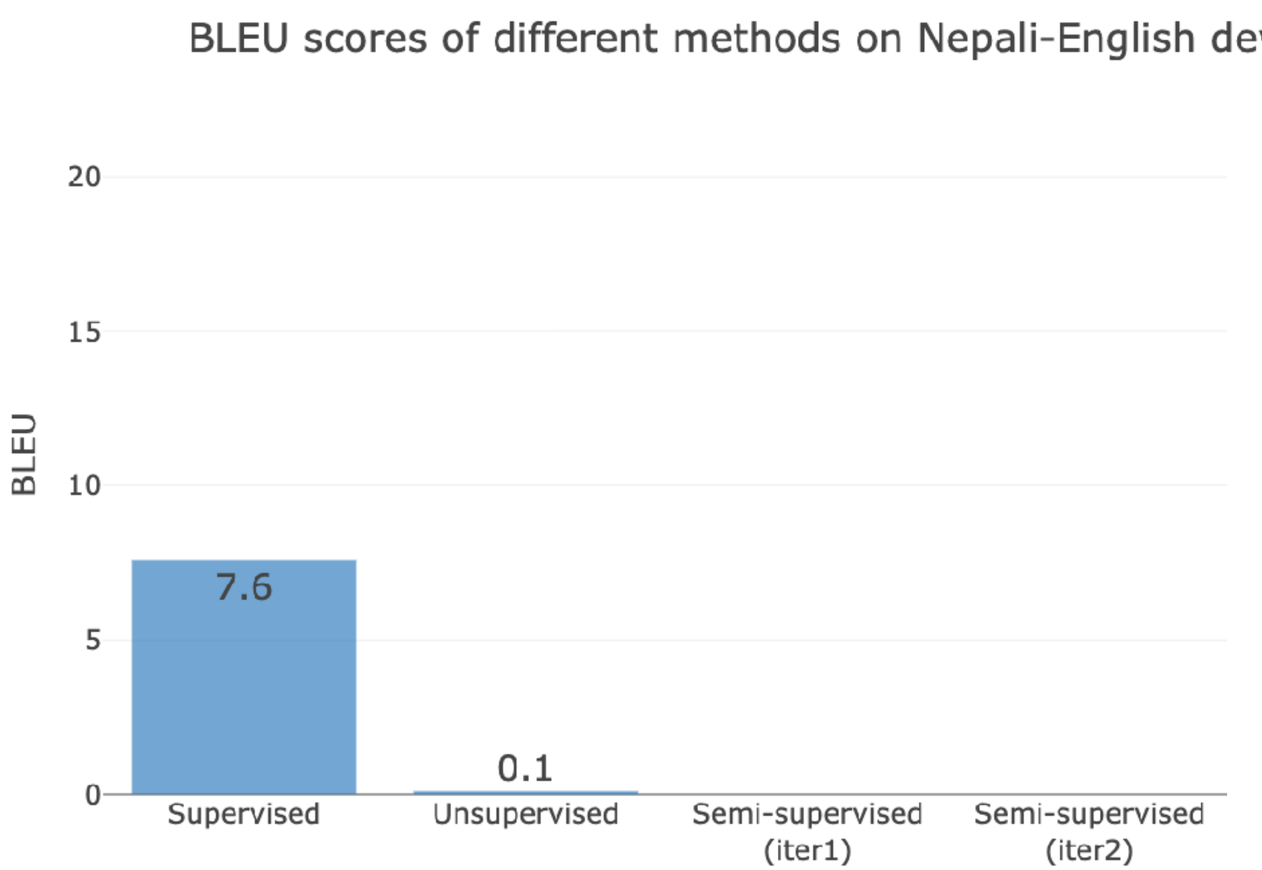
Case Study #2: FLoRes Ne-En

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

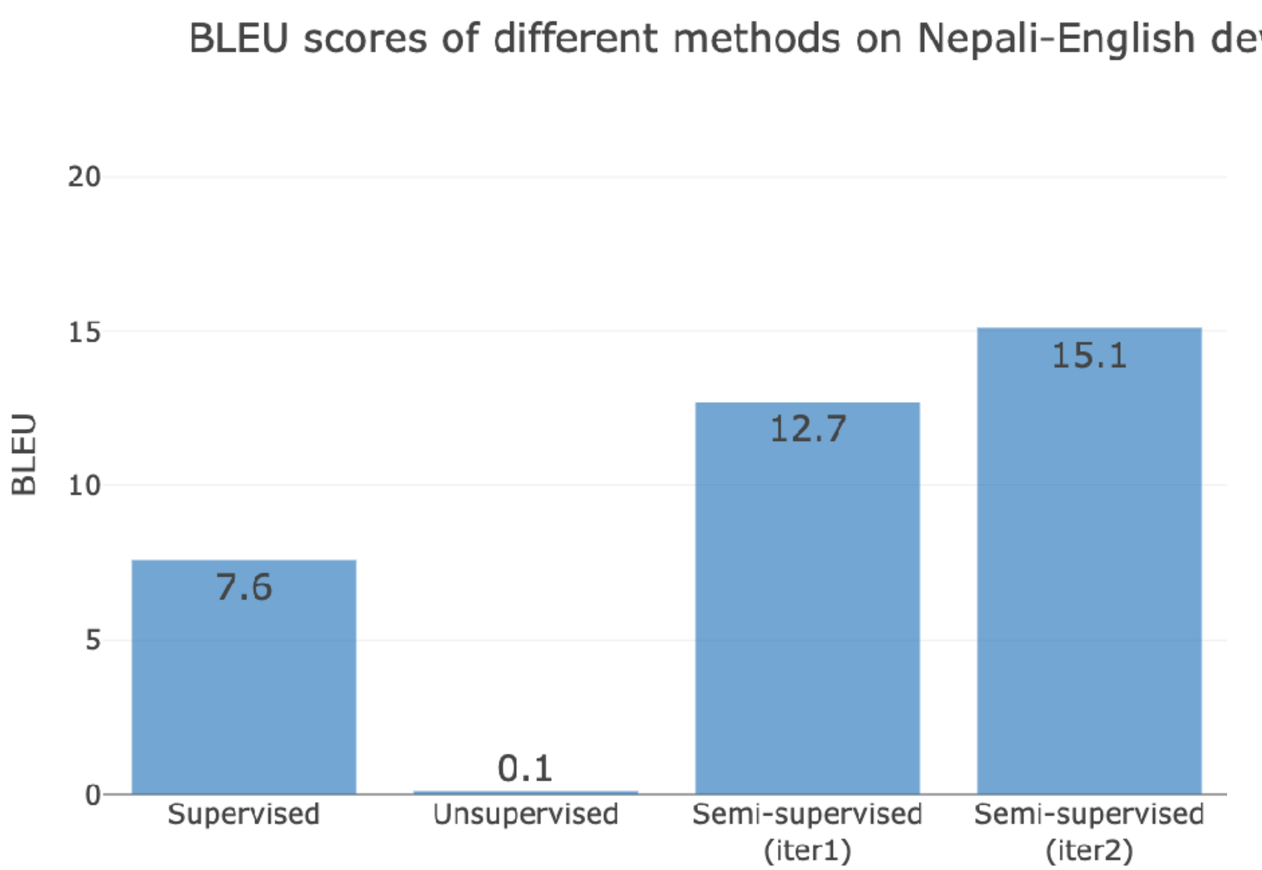
Results on FLoRes: Ne-En



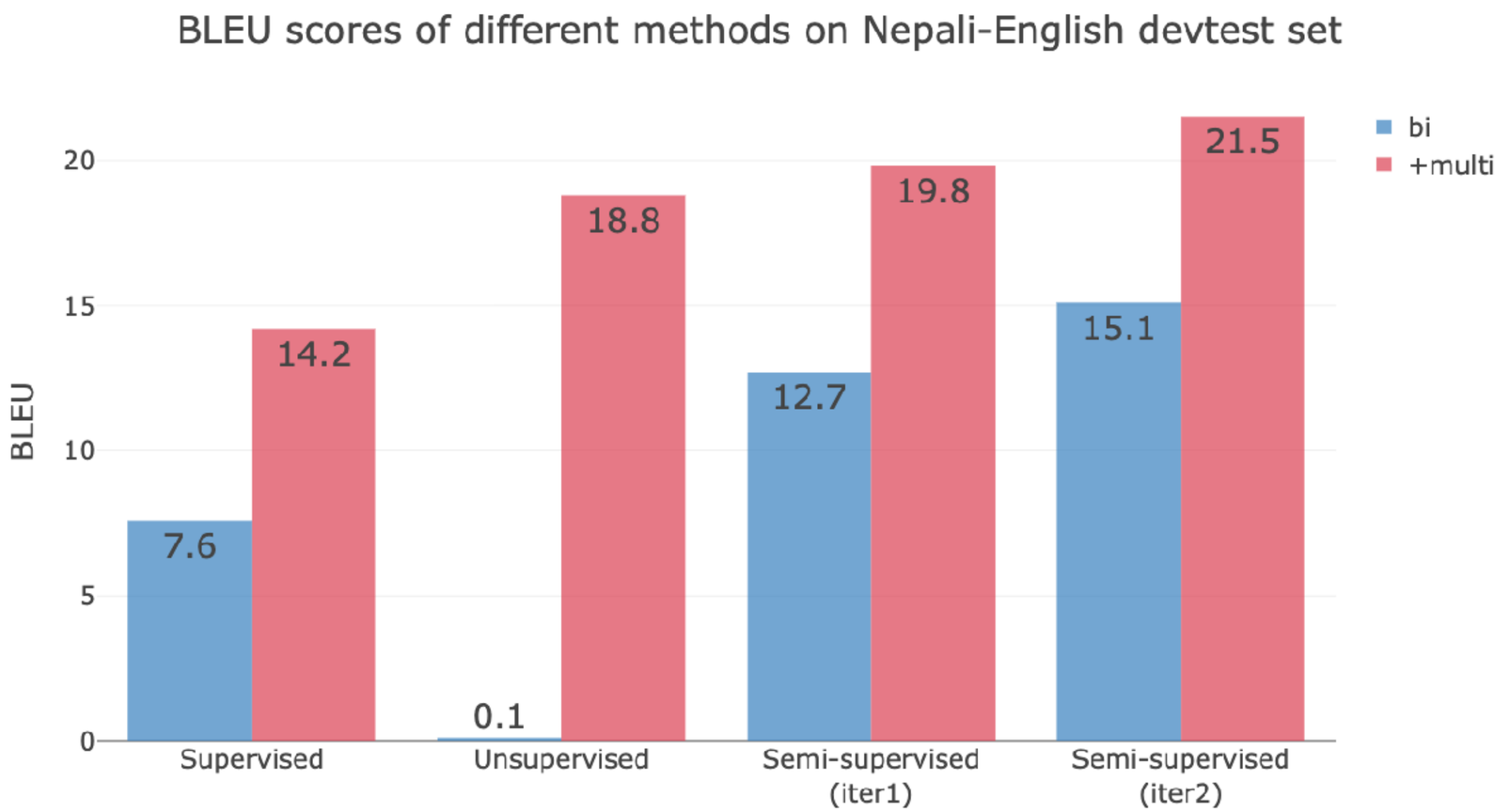
Results on FLoRes: Ne-En



Results on FLoRes: Ne-En



Results on FLoRes: Ne-En



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 trained on more data generalize better.
- Low-resource MT requires big data and big compute!

Lesson #6

The less labeled data you have, the more data you need to use…

Supervised learning:

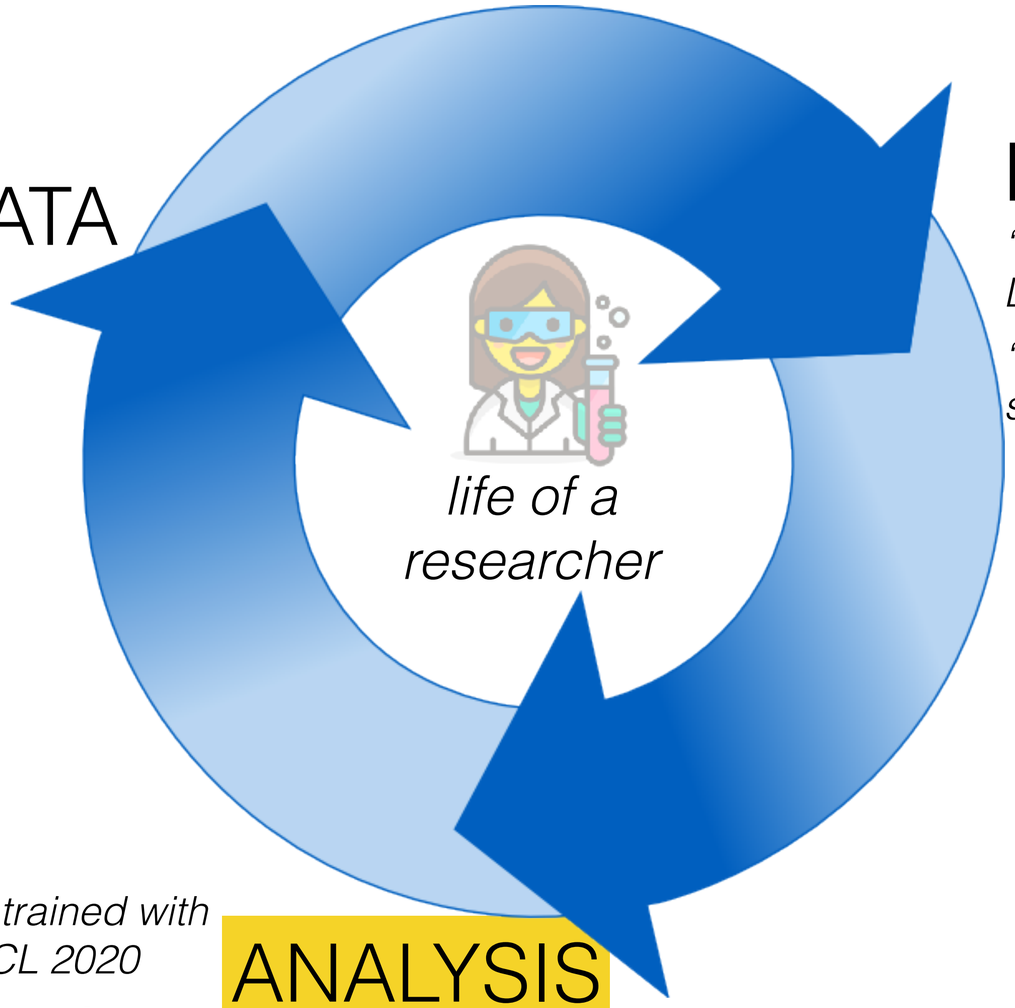
- Each datum yields X bits of information useful to solve the task.
- Need N samples.
- Need model of size Y MB.

Unsupervised learning:

- Each datum yields X/1000 bits (for instance).
- Need N*1000 samples.
- Need model of size Y*f(1000) MB

Low-resource MT <=> Large-Scale Learning!

The Cycle of Research



"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

"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

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!

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

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?

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

Examples from FLoRes

Si-En

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

original

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

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

translation

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

Wikipedia in Sinhala has a different topic distribution than Wikipedia in English.

En-Si

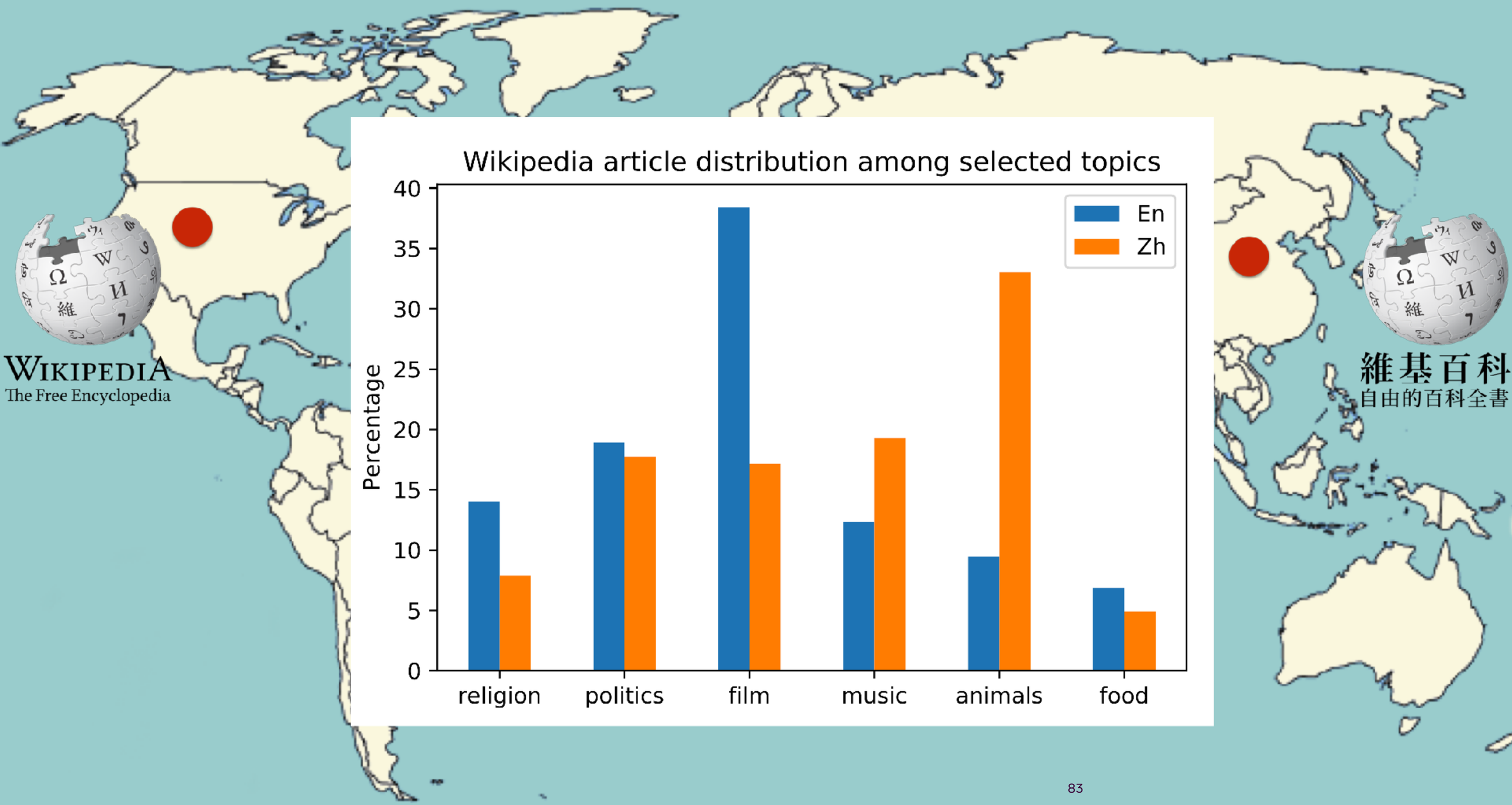
In Serious meets, the absolute score is somewhat meaningless.

සැබෑ නරඟ වලදී ලකුණු සැසඳීම තේරුමක් නැති ක්‍රියාවකි .

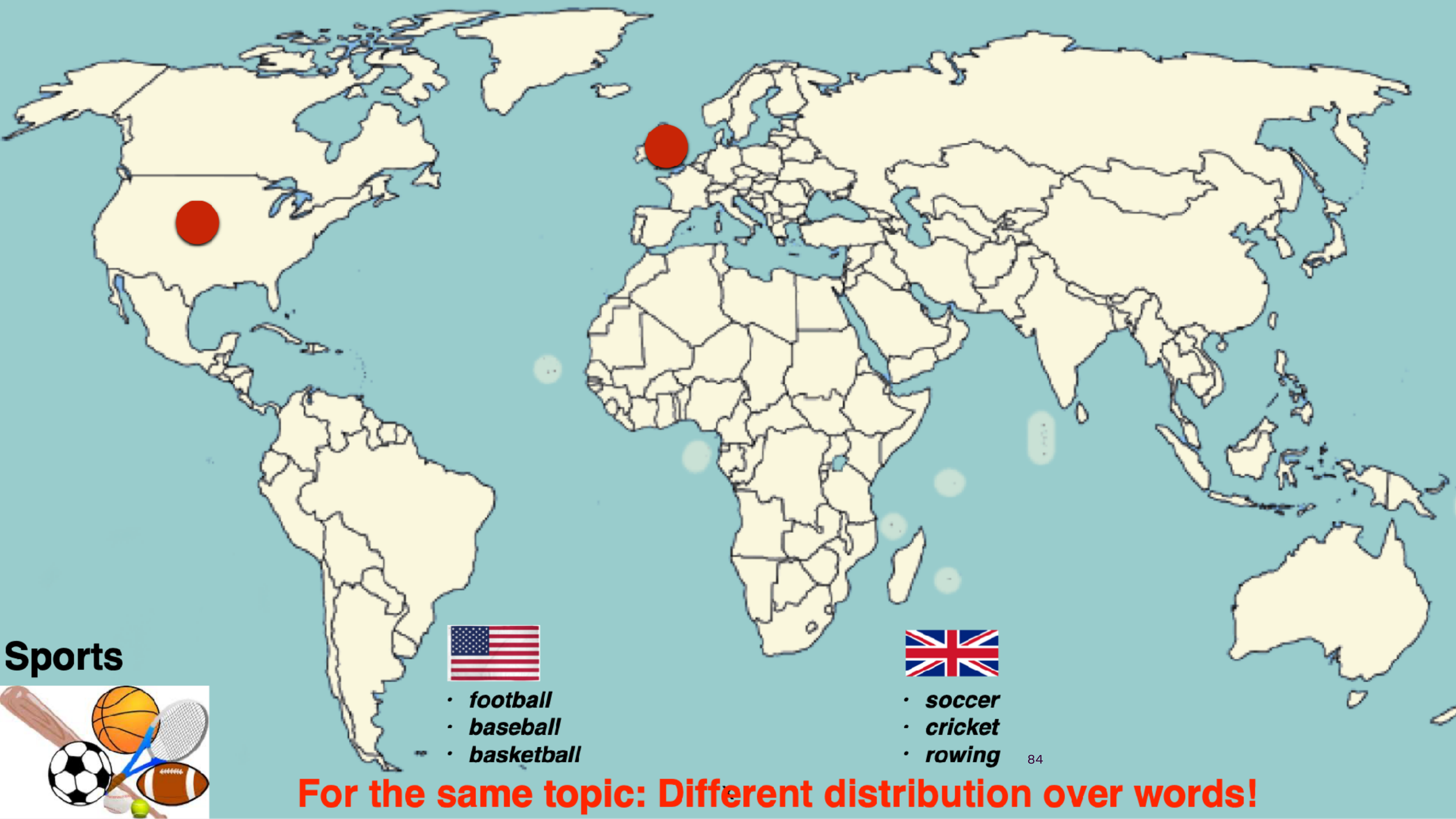
Iphone users can and do access the internet frequently, and in a variety of places.

අයිෆෝන් භාවිත කරන්නන්ට නිතරම සහ විවිධ ස්ථානවලදී අන්තර්ජාලයට පිවිසිය හැකිය .

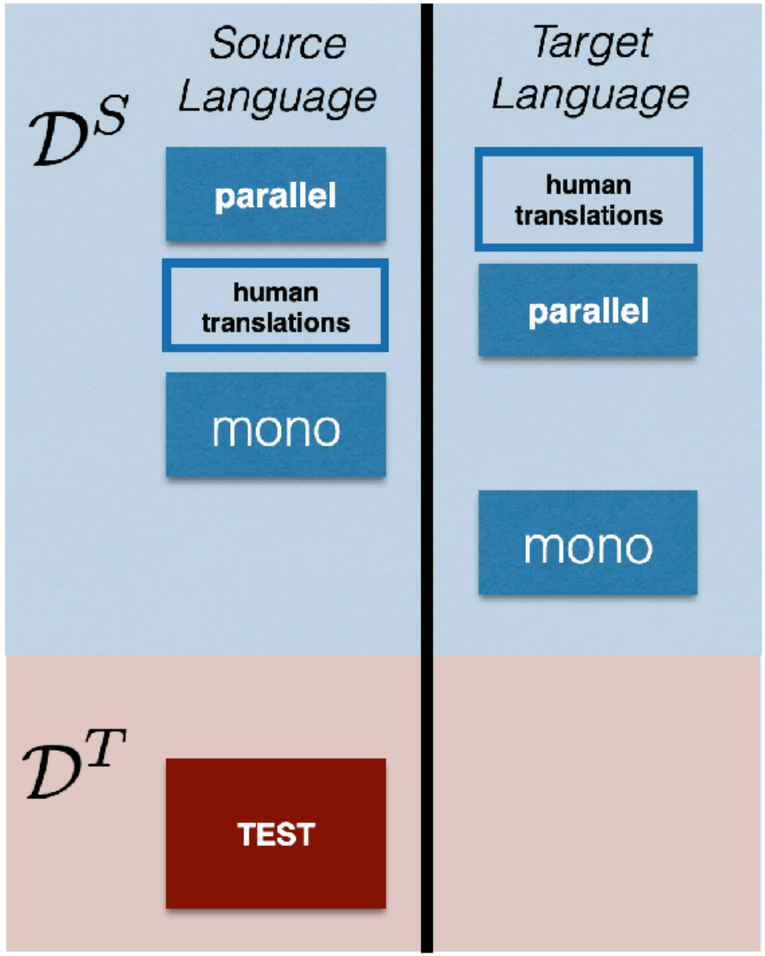
Guzmàn, Chen et al. “The FLoRes evaluation datasets for low resource MT…” EMNLP 2019



Domains differ in the topic distribution

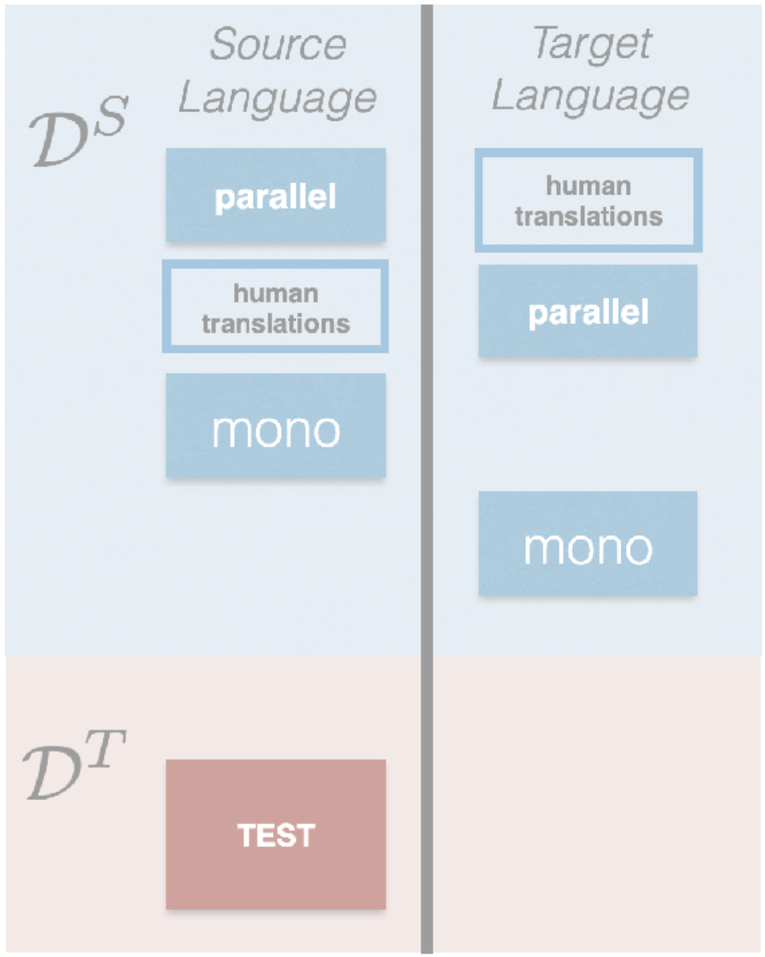

Sports
· football
· baseball
· basketball
· soccer
· cricket
· rowing
For the same topic: Different distribution over words!

Training-Test Domain Mismatch

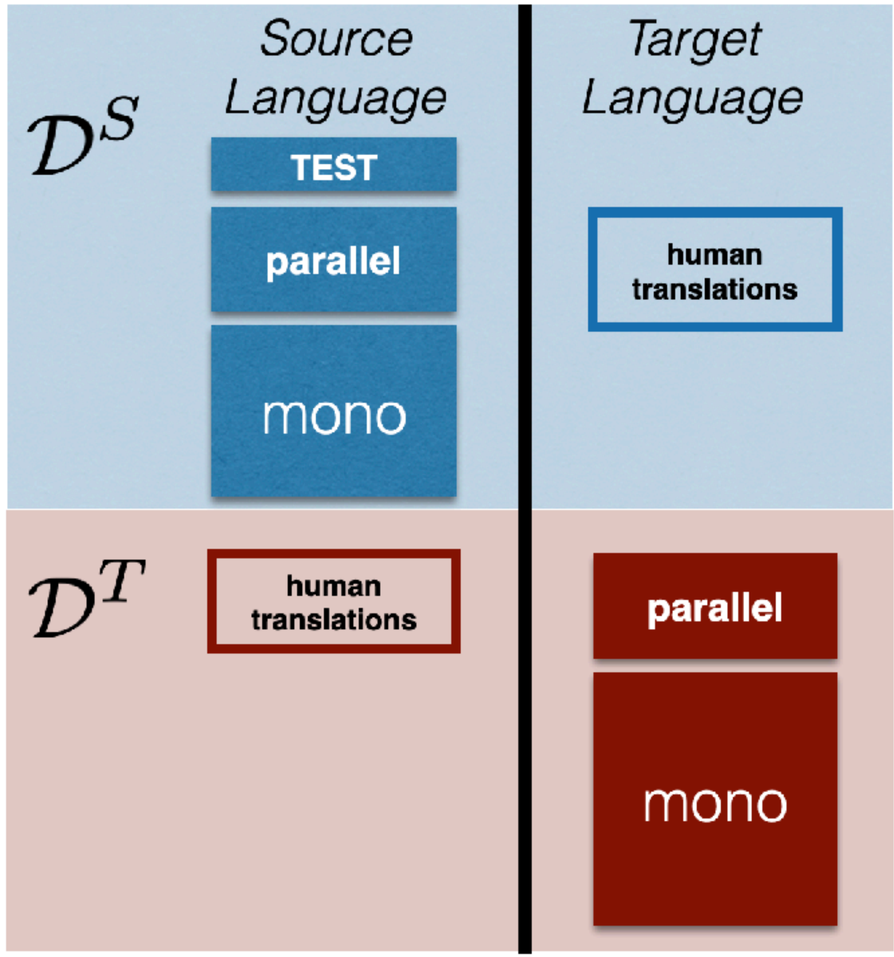


Train-Test Domain Mismatch in MT

Source-Target Domain Mismatch (STDM)



Train-Test Domain Mismatch in MT



Source-Target Domain Mismatch

Research Questions

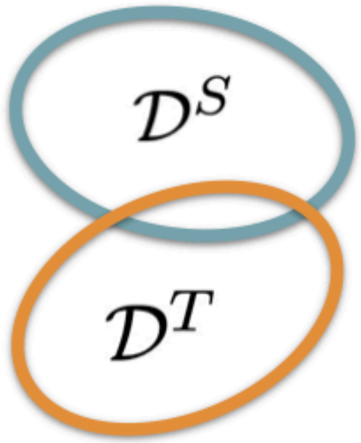
1. How to quantify STDM?

2. Would STDM affect training of MT systems and how?

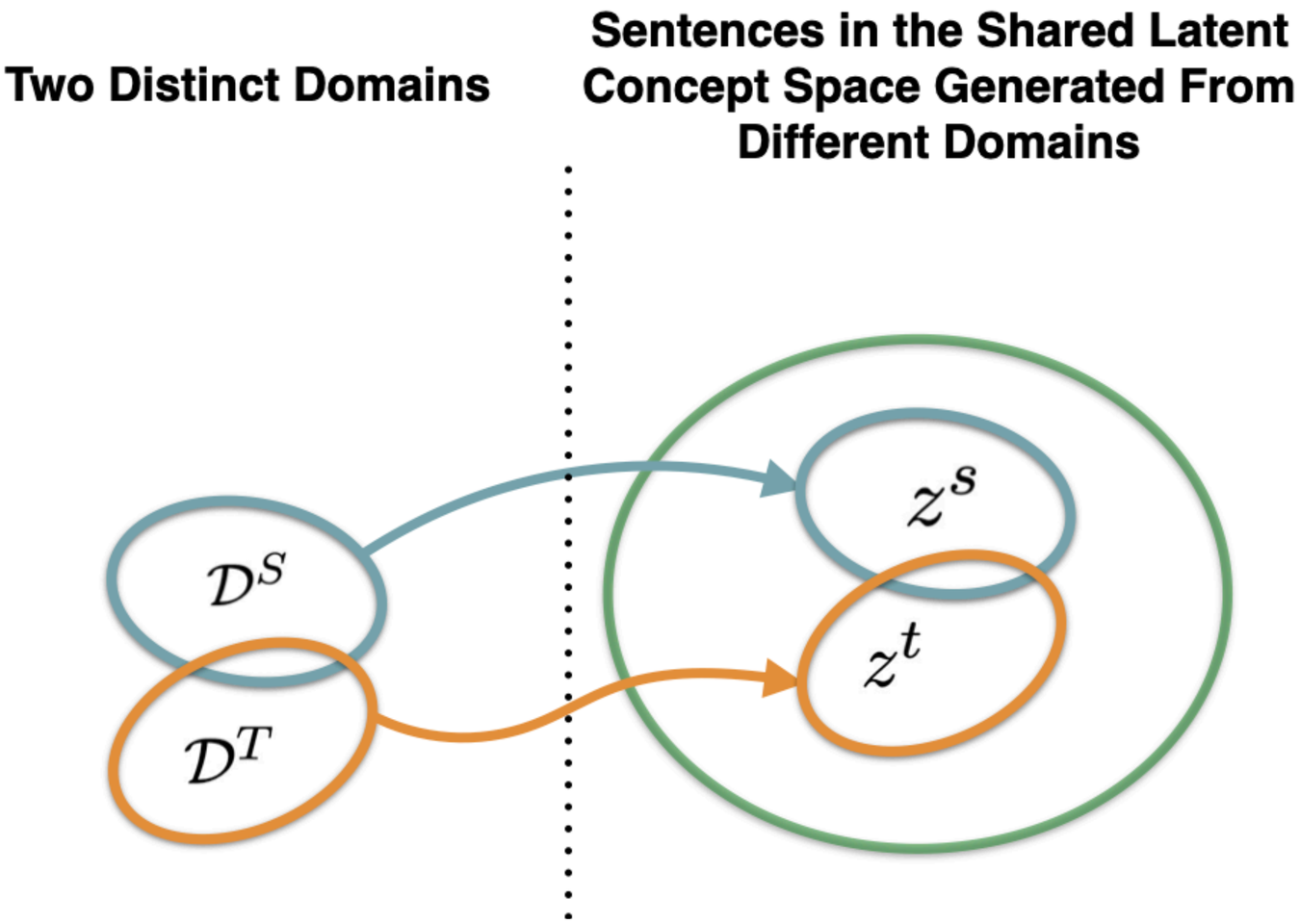
3. Any approaches to combat STDM?

How to quantify STDM?

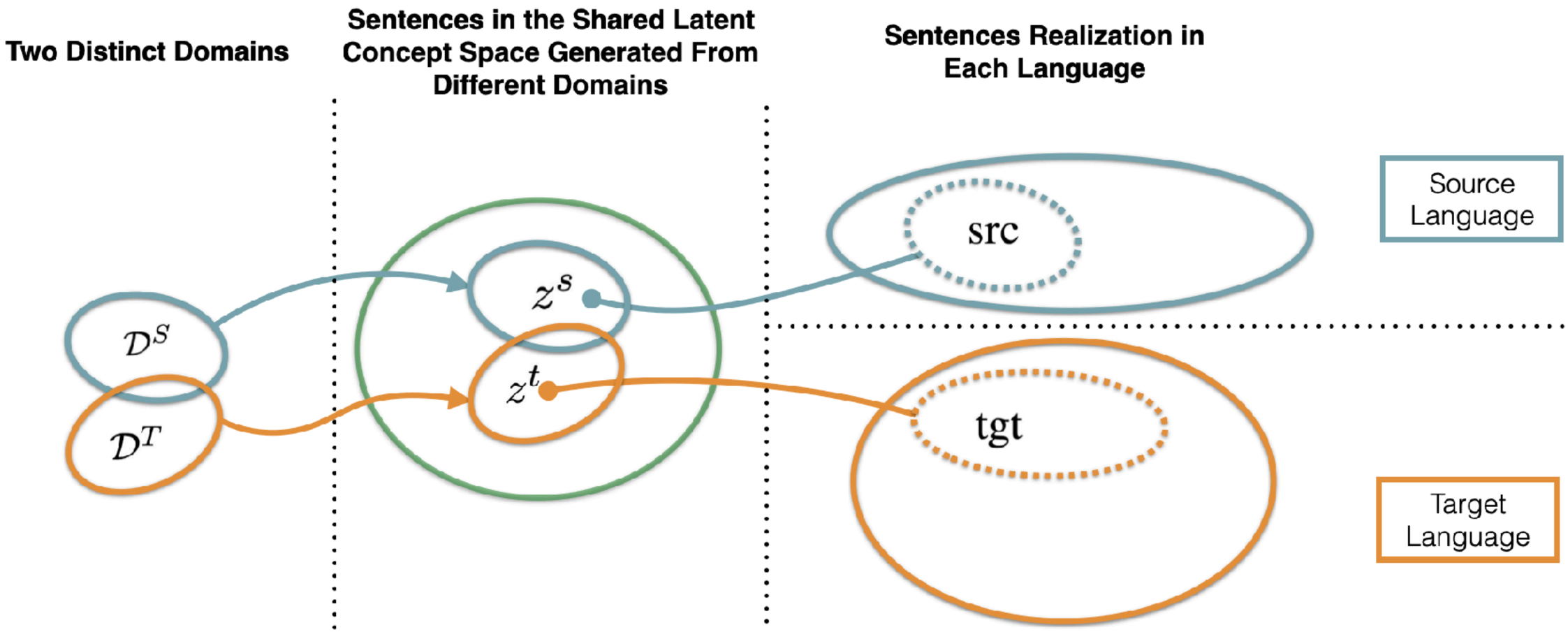
Two Distinct Domains



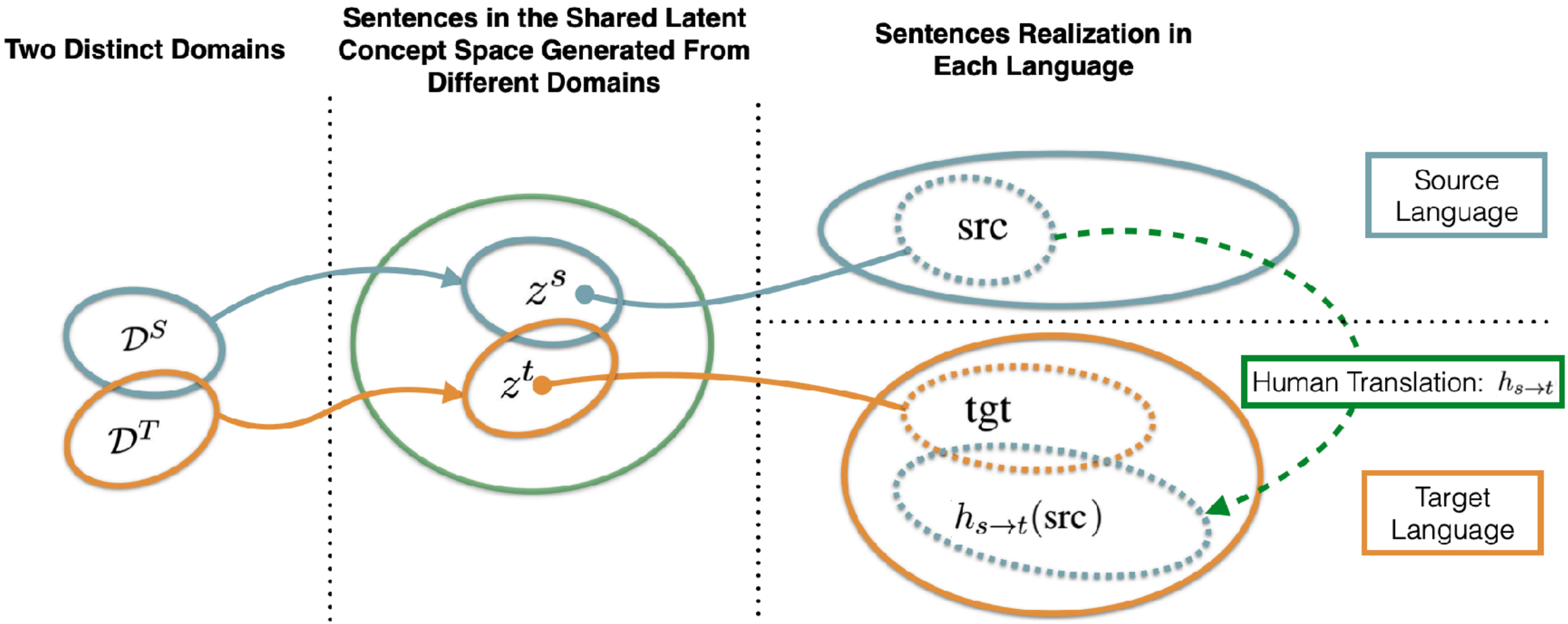
How to quantify STDM?



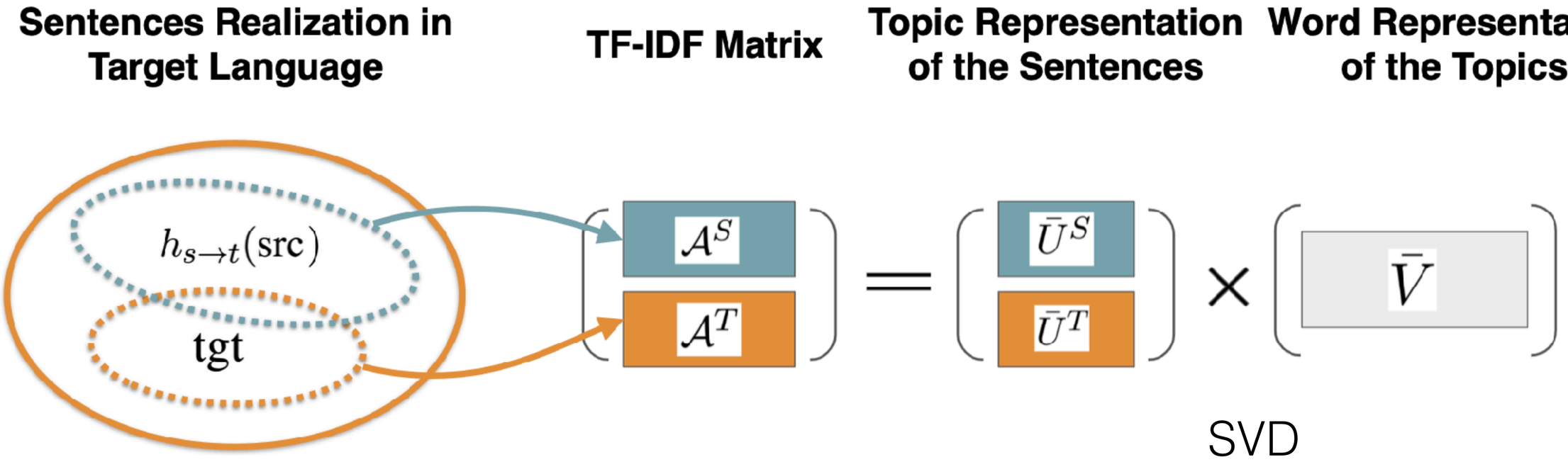
How to quantify STDM?



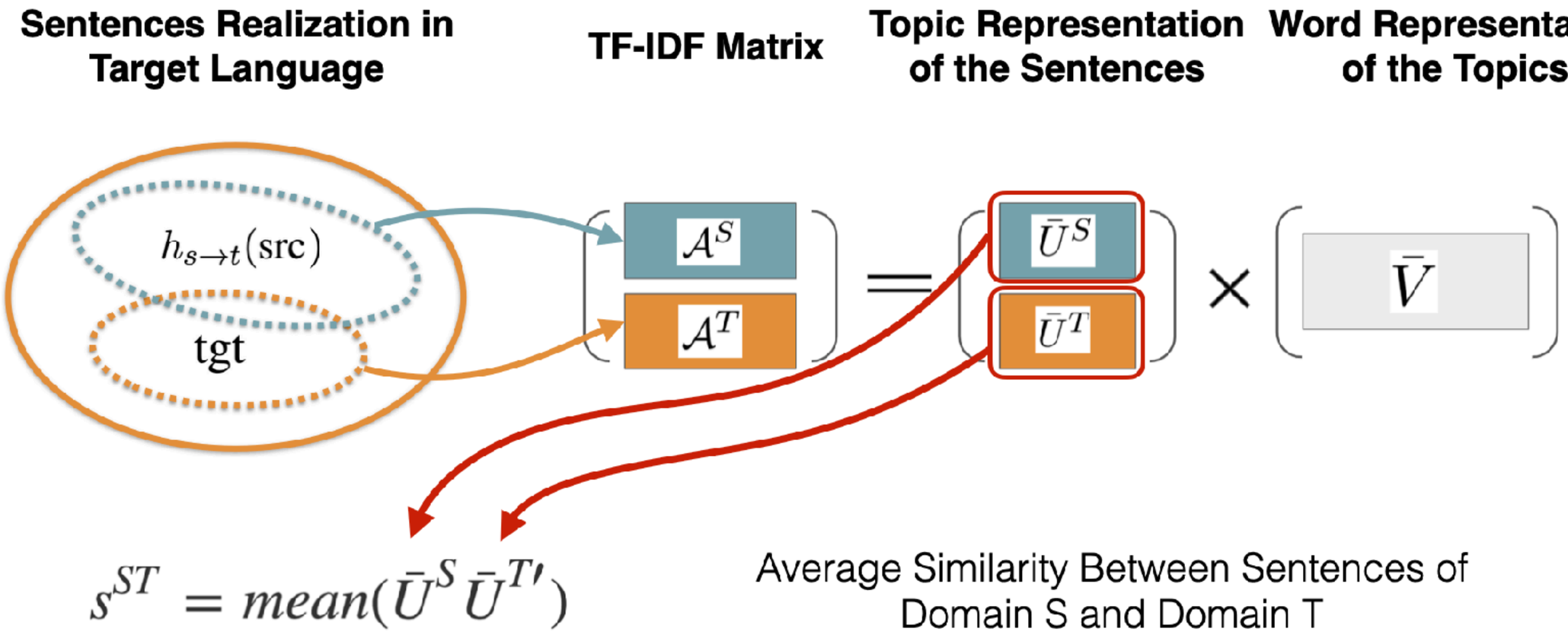
How to quantify STDM?



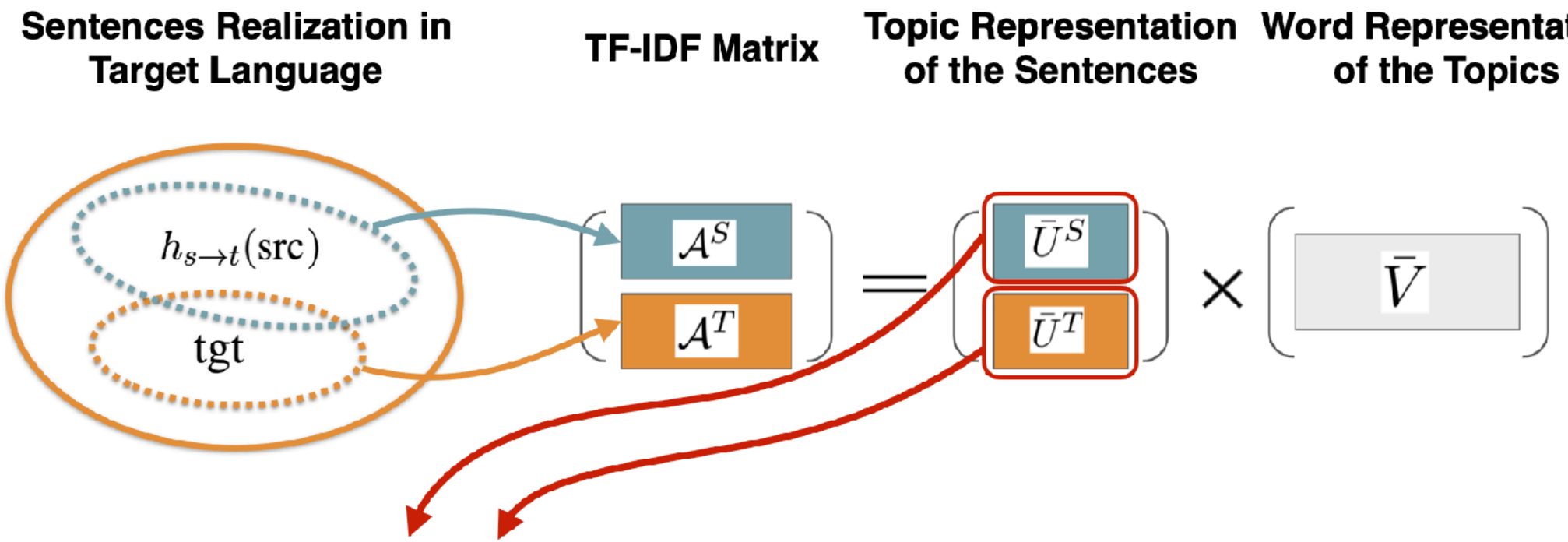
How to quantify STDM?



How to quantify STDM?



How to quantify STDM?

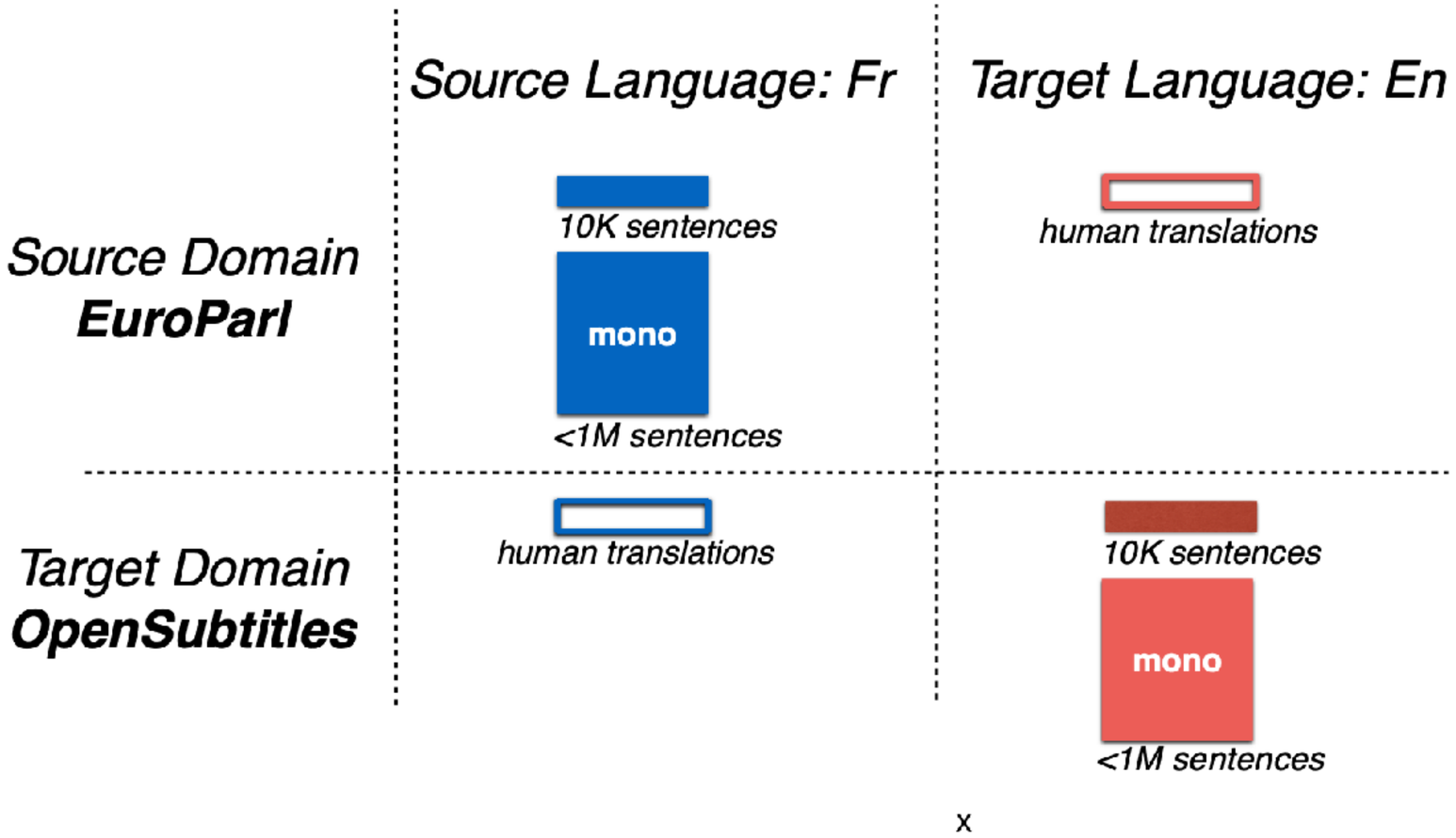


$$s^{ST} = mean(\bar{U}^S \bar{U}^{T\prime})$$

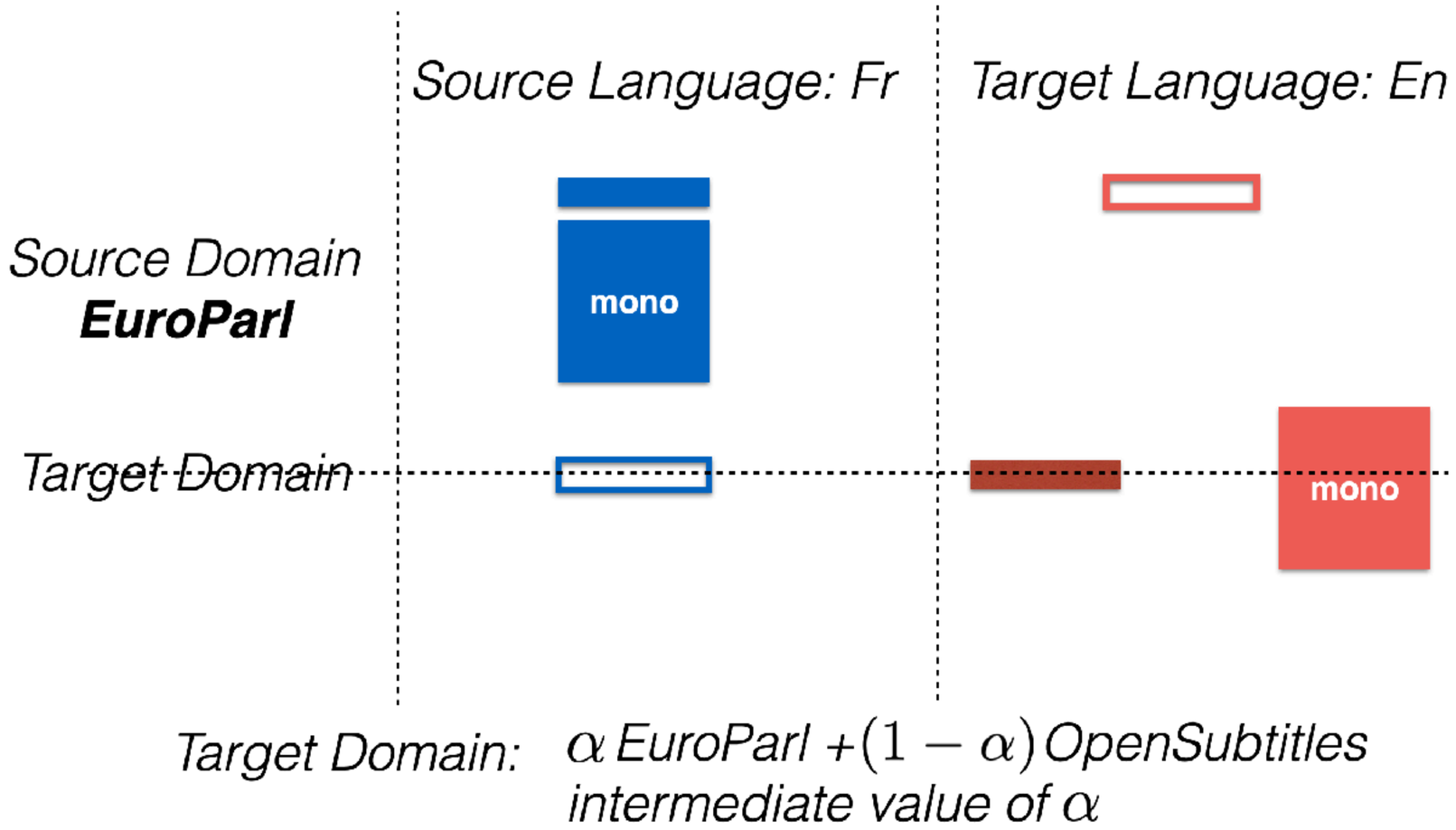
Average Similarity Between Sentences of Domain S and Domain T

$$\text{score} = \frac{s^{\text{ST}} + s^{\text{TS}}}{s^{\text{SS}} + s^{\text{TT}}}$$

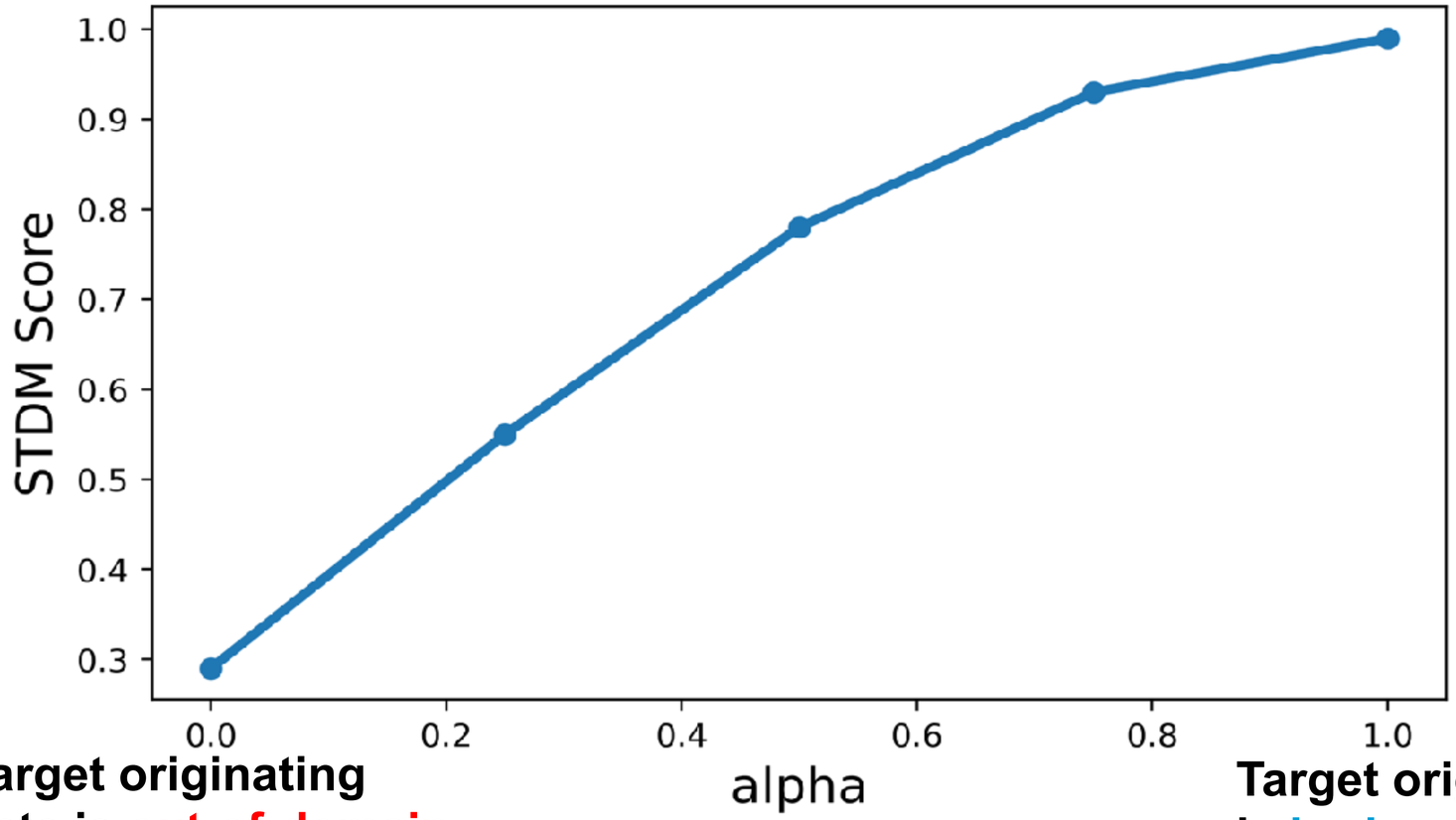
Studying STDM in a Controlled Setting



Studying STDM in a Controlled Setting



Sanity Check: STDM Scores for Datasets Under Control Setting



$\mathcal{D}^S$: *EuroParl*

$\mathcal{D}^T$: α *EuroParl* +$(1-\alpha)$*OpenSubtitles*

Target originating data is out-of-domain

Target originating data is in-domain

STDM Scores on Real Datasets

	De-En	Fi-En	Ru-En	Ne-En	Zh-En	Ja-En
WMT	0.79	0.79	0.76	-	0.65	-
MTNT	-	-	-	-	-	0.69
SMD	0.81	0.71	0.71	0.64	0.71	0.61

Mild signs of STDM and negligible difference across language pairs as they contain mostly international news.

More severe sign of STDM as it contains much more local news.

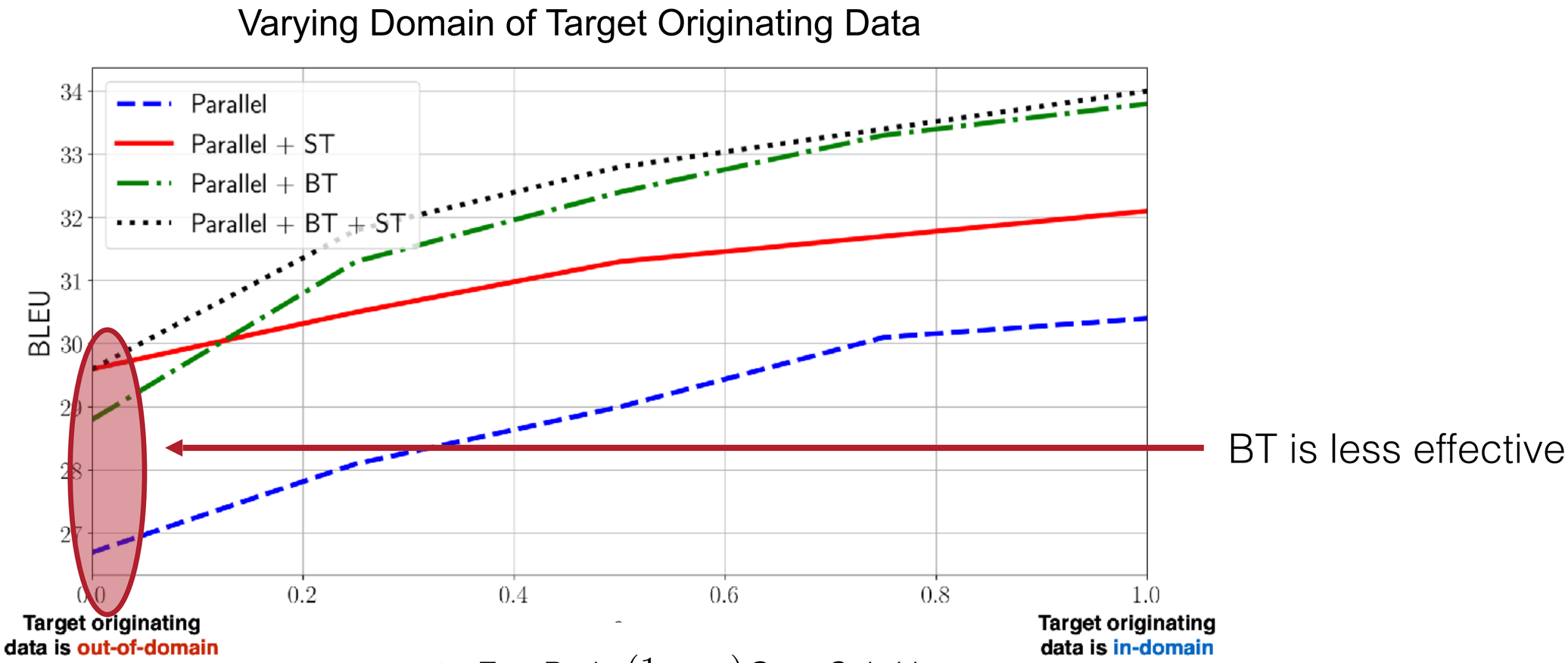
STDM Scores on Real Datasets

	De-En	Fi-En	Ru-En	Ne-En	Zh-En	Ja-En
WMT	0.79	0.79	0.76	-	0.65	-
MTNT	-	-	-	-	-	0.69
SMD	0.81	0.71	0.71	0.64	0.71	0.61

More severe signs of STDM for distant languages !

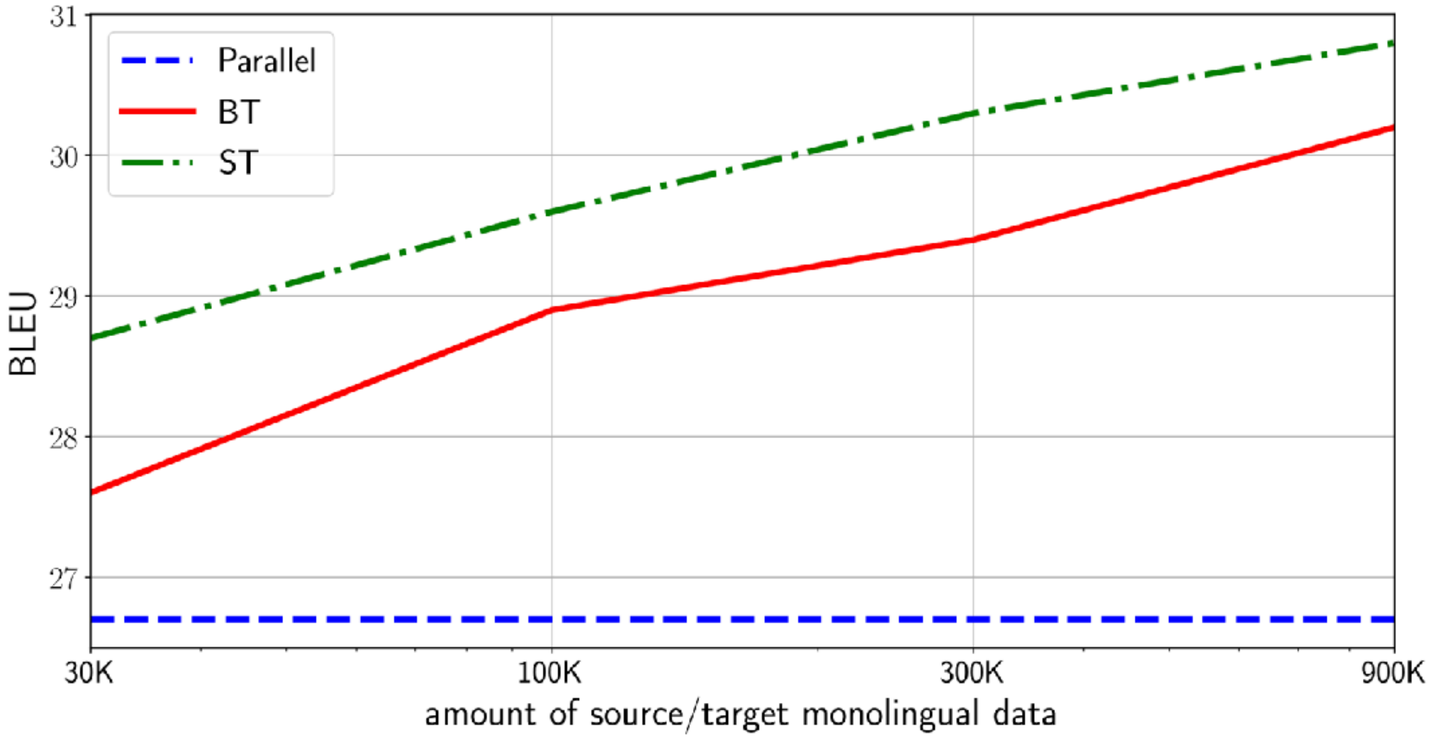
How does STDM affect MT training?

How does STDM affect MT training?



How does STDM affect MT training?

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

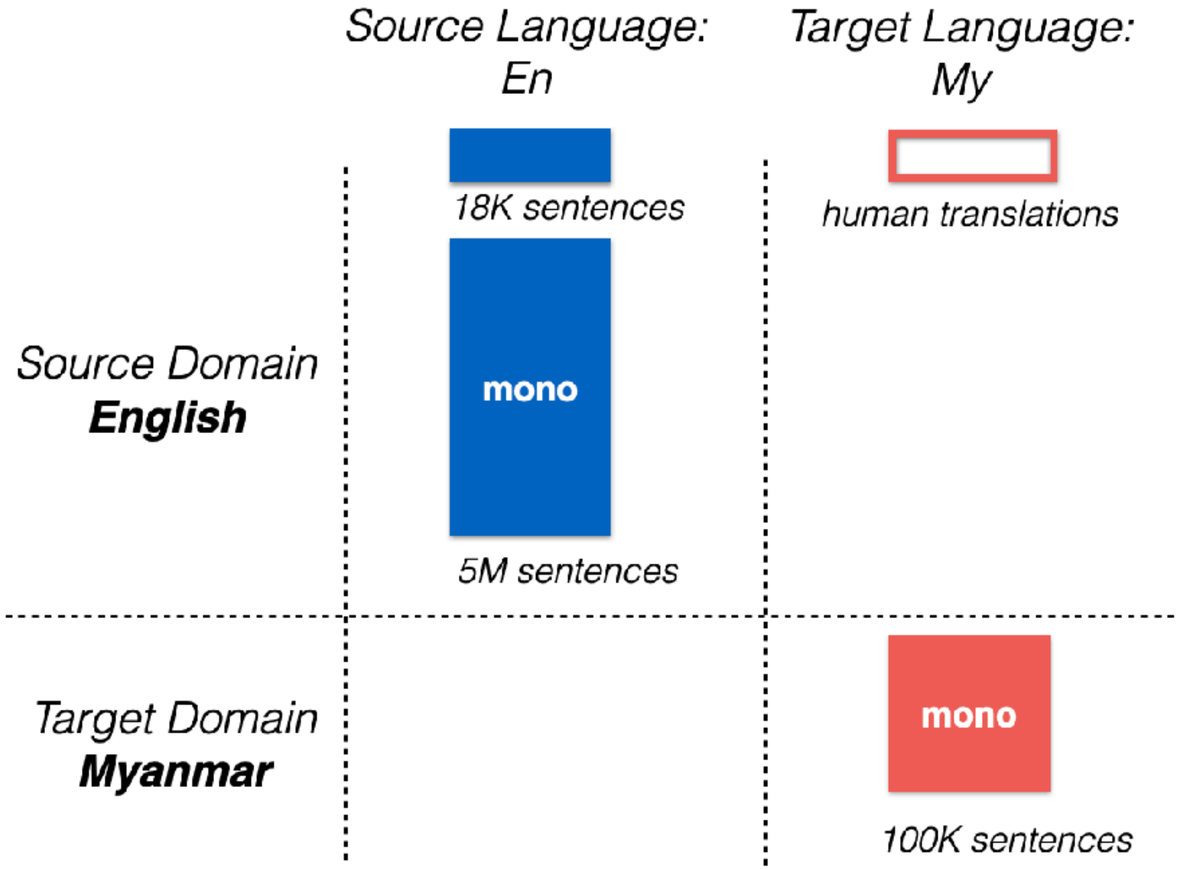


Increasing the amount of monolingual data can compensate for domain mismatch.

How does STDM affect MT training?

Model	En → My STDM score=0.27
baseline	28.1
BT	30.0
ST	31.9
ST + BT	**32.4**

BLEU scores for the English to Myanmar translation task



Conclusion

- STDM: a new kind of domain mismatch, intrinsic to the MT task.
- STDM is particularly significant in low resource language pairs.
- Metric & controlled setting enable study and better understanding of STDM.
- Methods that leverage source side monolingual data are more robust to STDM.
- In practice, the influence of STDM depends on several factors, such as the amount of parallel and monolingual data, the domains, language pair, etc. In particular, if domains are not too distinct, STDM may even help regularizing!

Summary

- 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
- Data collection is non trivial.
- There are several kinds of domain mismatch, which affect generalization.

Debugging NMT

- **Literature review:** do due diligence and understand what people have done for that particular language pair. Check entries at yearly WMT competition.
- **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.
    - Check originating datasets, quality of data, domain, etc. Will you need domain adaptation methods?
    - Is there data or an existing model that can be used for pretraining? Are there good auxiliary tasks to retrain or multitask?
- **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.). Are generations matching what's in the data? Are there bugs in how strings are pre or post-processed?

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.
    - If overfitting is not an issue, than scale up the model or try to better optimize.
- **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!

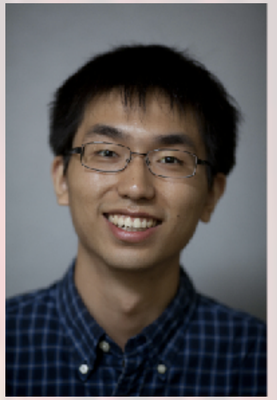
Guillaume Lample

Ludovic Denoyer

Myle Ott

Peng-Jen Chen

Paco Guzmán



Jiajun Shen

Questions?
Вопросы?
¿Preguntas?
Domande?

Jiatao Gu

Xian Li

Naman Goyal

Matt Le

Philipp Kohen

Juan-Miguel Pino

Alexis Conneau

Hervé Jegou

Michael Auli

Junxian He

Sergey Edunov

Vishrav Chaudhary