Multilinguality

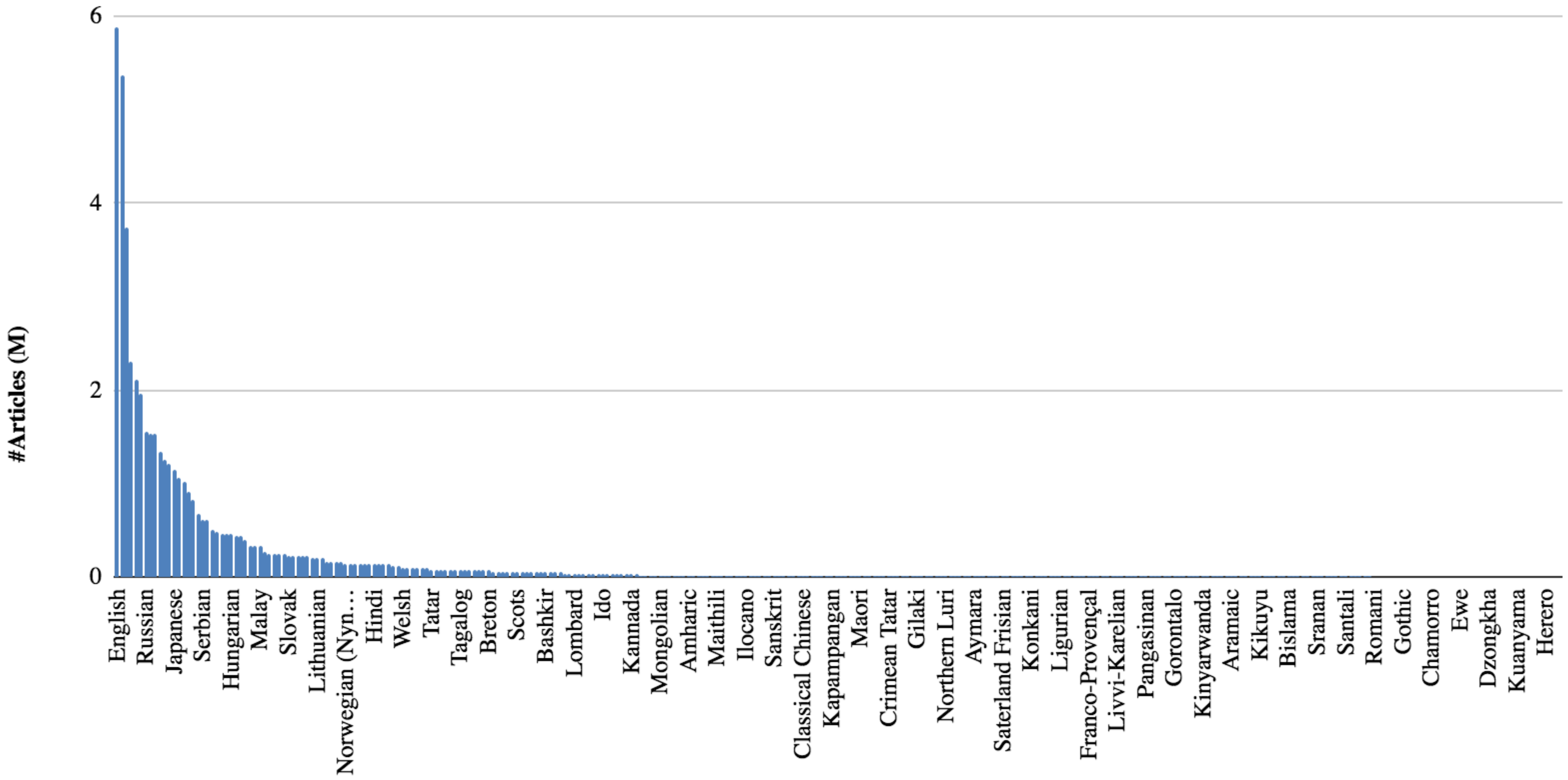
CS6120: Natural Language Processing
Northeastern University

David Smith
with slides from Graham Neubig

Two varieties of multilingual NLP

- Monolingual NLP in Multiple Languages
  - QA, sentiment analysis, chatbots, code generation
  - in English, Chinese, Hindi, Japanese, Spanish, …
- Cross-lingual NLP
  - Machine translation
  - Cross-language information retrieval
  - Cross-lingual QA

Data are mostly sparse



Data Source: Wikipedia articles from different languages

- Big disparity in monolingual data available for training
- Even less annotated data for MT, sequence labeling, dialogue, etc., in many languages

Linguistic peculiarities

- Most methods are tested first on English, but many languages differ from English in, e.g.,
  - Rich morphology (case, gender, mood, etc.)
  - Accents/diacritics
  - Different scripts
  - Variety and status of dialects
  - Lack of formal writing systems

Multilingual learning

- We would like to learn models that process multiple languages
- Why?
  - **Transfer Learning:** Improve accuracy on *lower-resource* languages by transferring knowledge from higher-resource languages
  - **Memory Savings:** Use one model for all languages, instead of one for each
  - **Time Savings:** We don’t need to decide which language we’re processing

Code switching

```
Ulikuwa ukiongea a lot of nonsense.    (Code-switching, English in bold)

"You were talking a lot of nonsense."   (Translation)
```

Embedded language	**Code-switched text**
Latin	... die *fortitudo animi, magnitudo* (c. 5.) seinem Sohne schildert, ...
English	... die man das Westende nennt, *the west end of the town*, und wo die vornehmere und minder beschäftigte Welt lebt.
French	... eine frivole Laune, ein "*car tel est notre plaisir*" des Geistes ...
Greek	Diess ist das Spinnrad des βάϑος; diess ist das Spinnrad, welches die Gedanken spinnet, ...

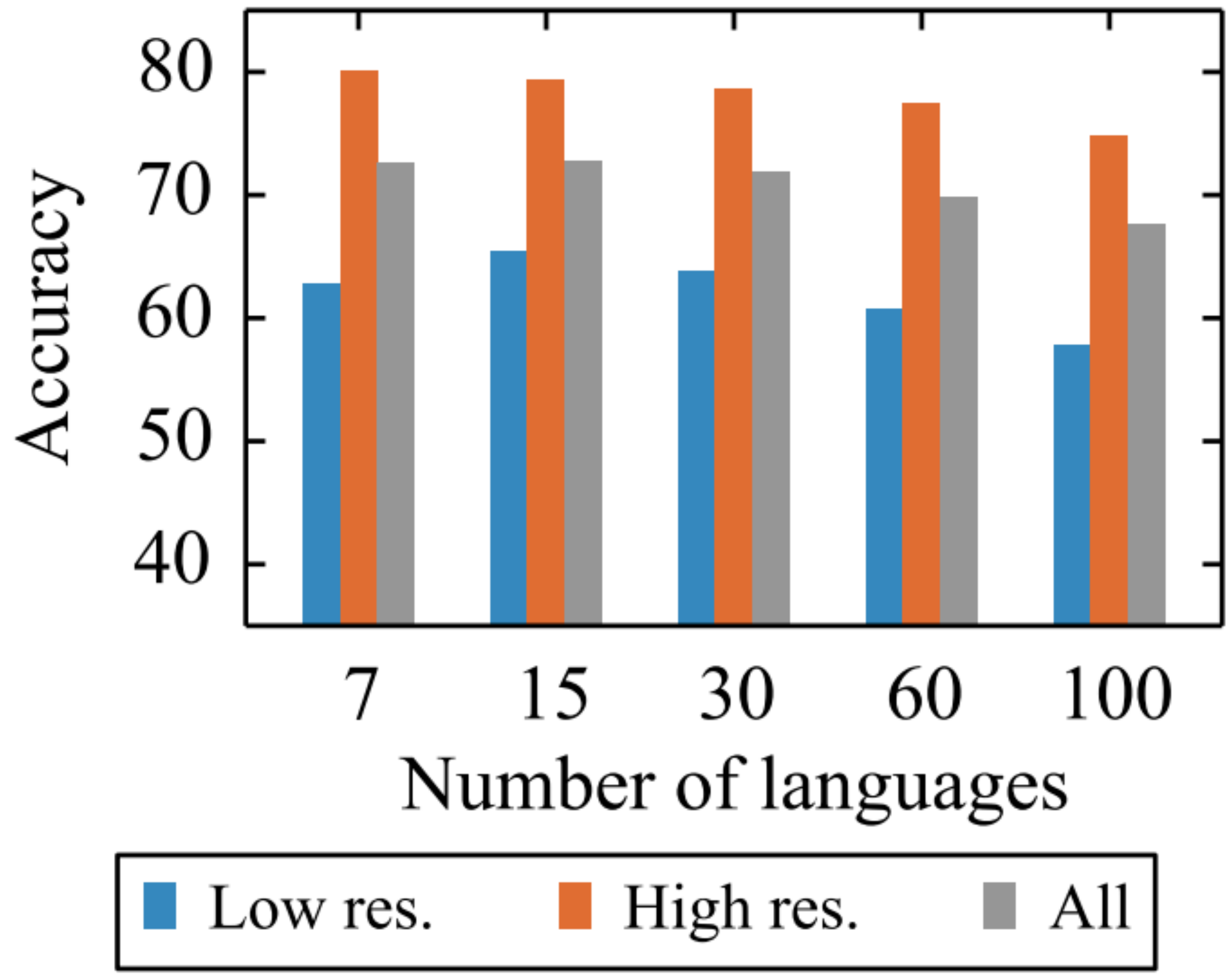
Multilingual Language Modeling

Simple multilingual modeling

- It is possible to learn a single model that handles several languages
- **Multilingual Input:** Can just process different input languages using the same network (Wu and Dredze 2019)
  - ceci est un exemple → this is an example
  - これは例です → this is an example
- **Multilingual Output:** Add a tag or prompt about the target language for generation (Johnson et al. 2016)
  - <fr> this is an example → ceci est un exemple
  - <ja> this is an example → これは例です

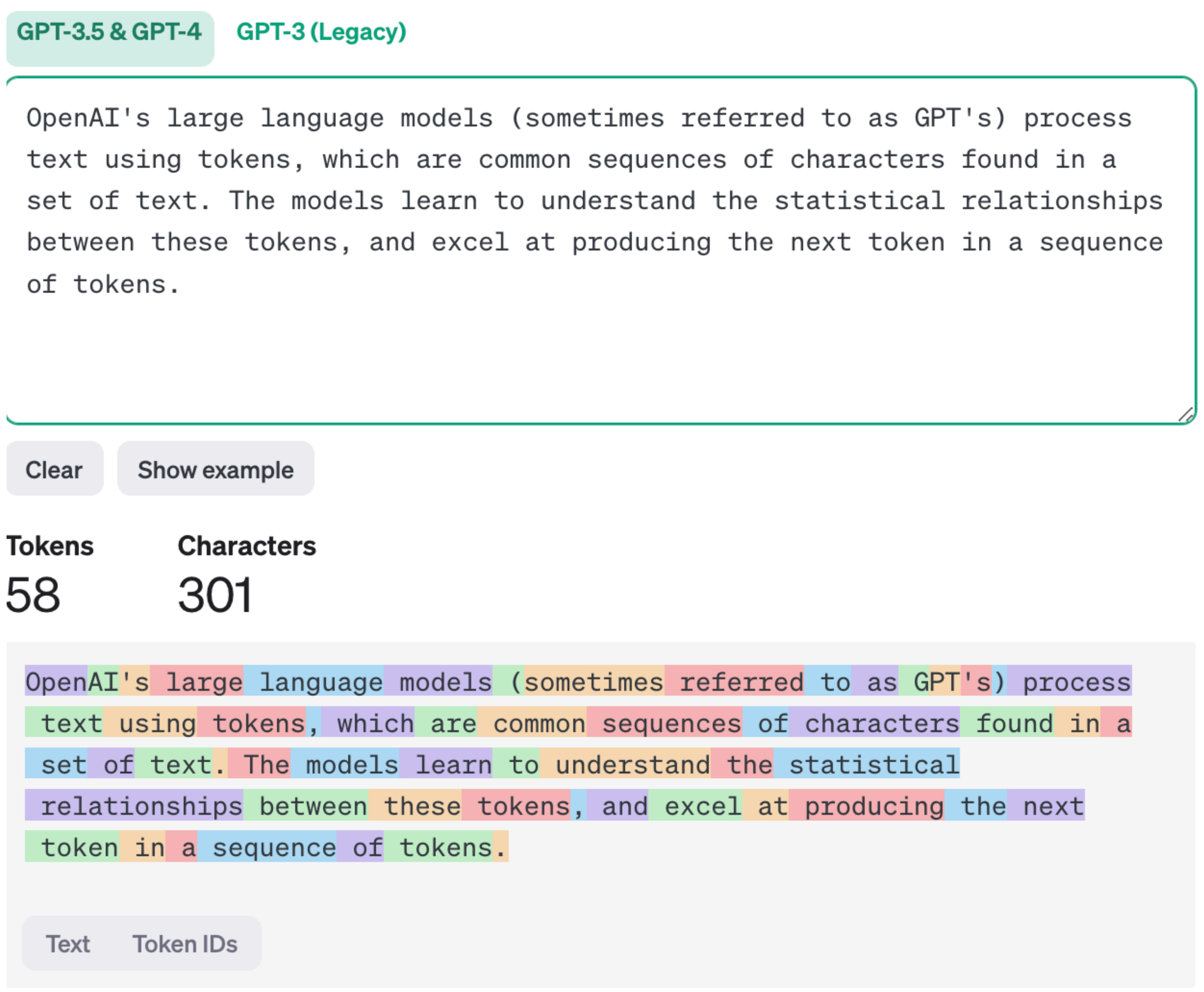
Difficulties in fully multilingual learning

- "**Curse of Multilinguality**" For a fixed sized model, the per-language capacity decreases as we increase the number of languages (Conneau et al., 2019)
- Increasing the number of low-resource languages→decrease in the quality of high-resource language translations (Aharoni et al., 2019)
- How to mitigate? **Better data balancing, better parameter sharing**

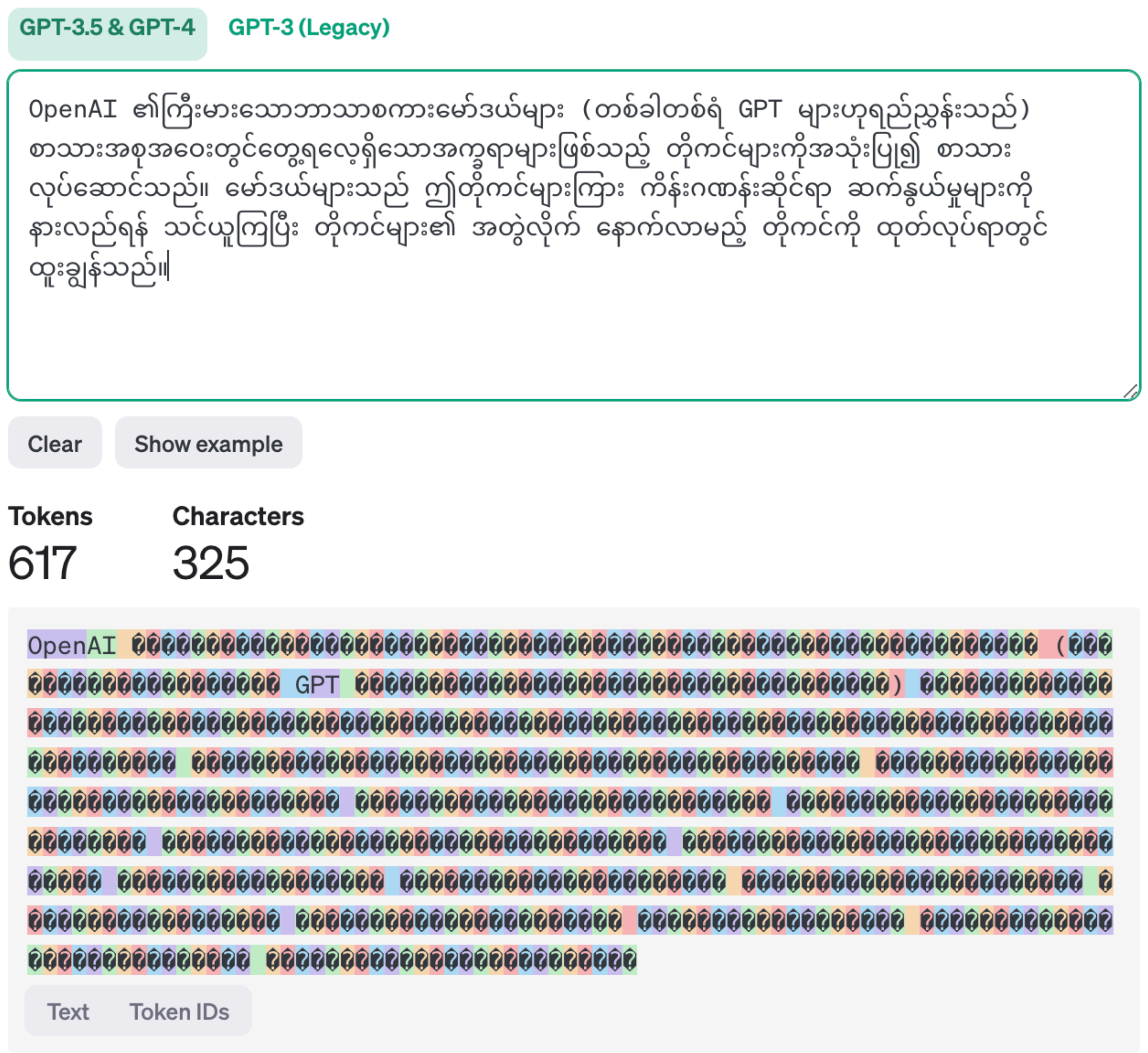


Tokenization disparities

English

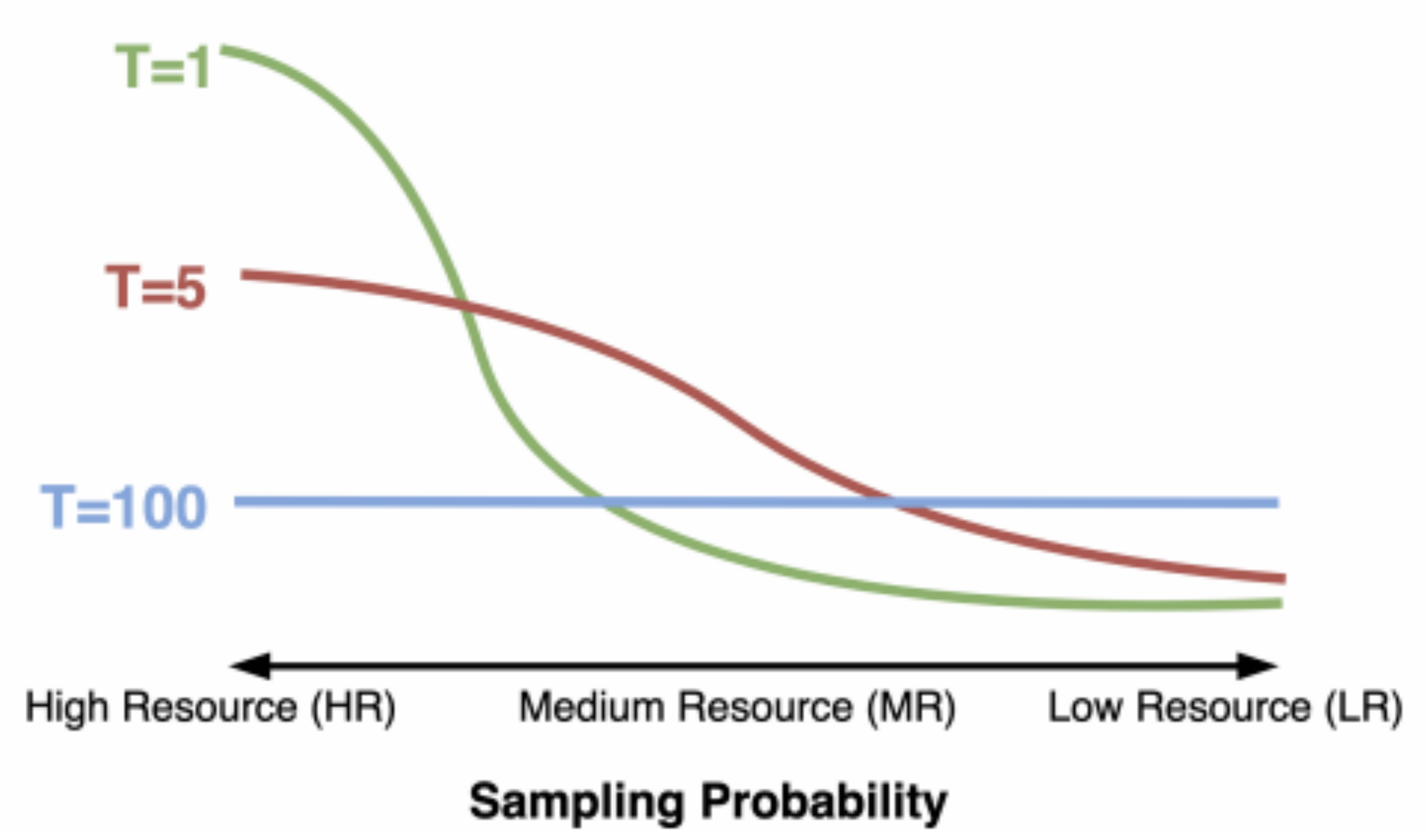


Burmese/Myanmar (Google Translated)



Similar content, 10.6x the tokens!

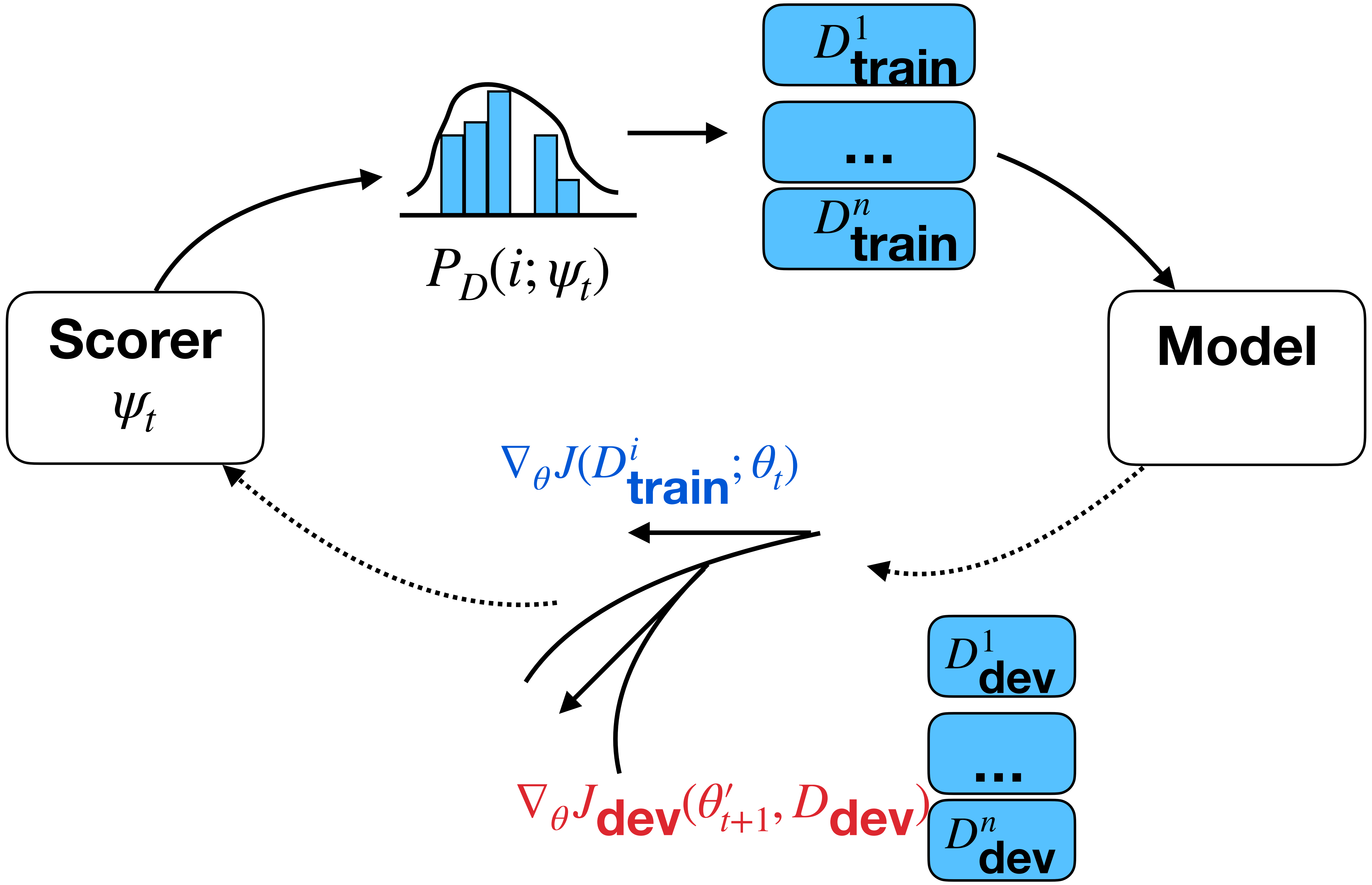
Heuristic sampling



Massively Multilingual Neural Machine Translation in the Wild. Arivazhagan et. al. 2019

- Sample data based on dataset size scaled by a temperature term
- Sample at model training time, or vocabulary construction time

Learning to balance data

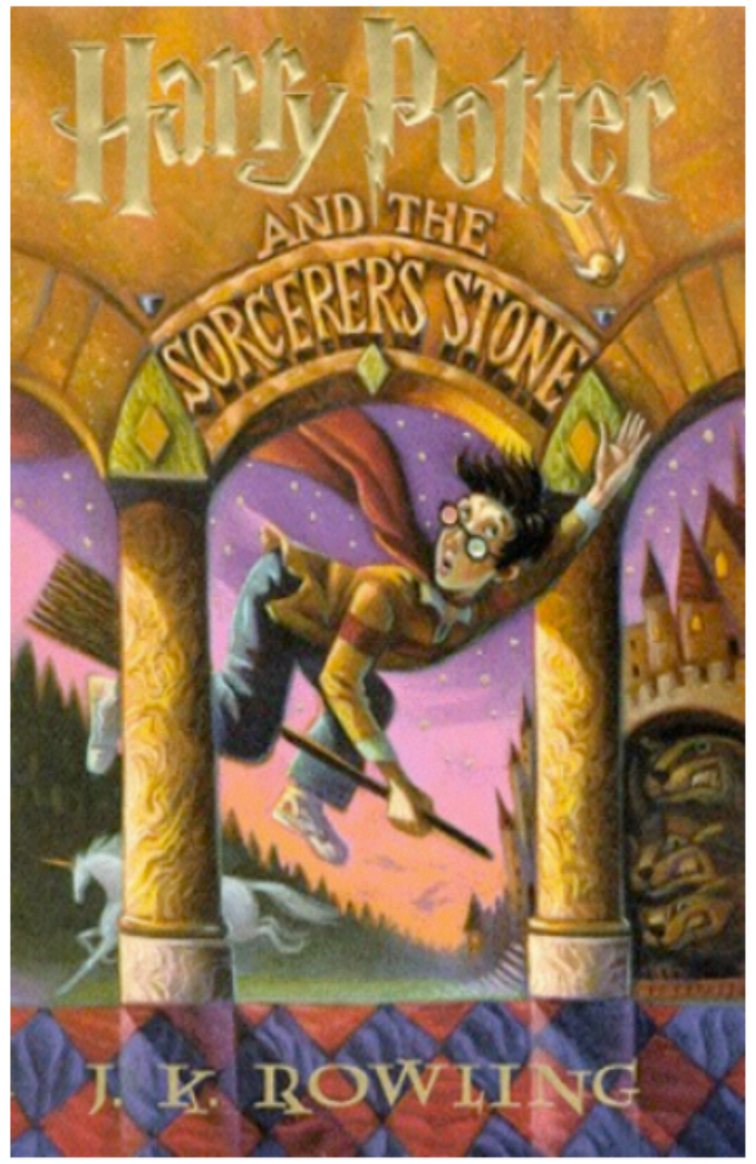


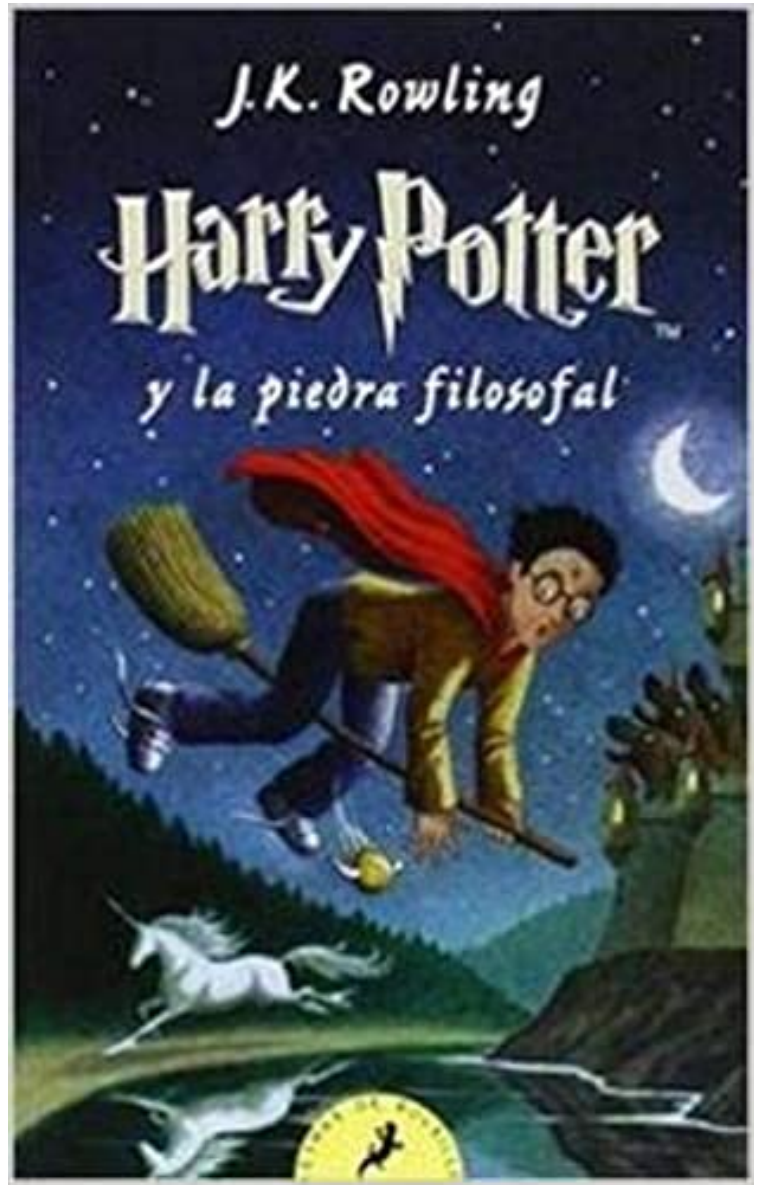
Balancing Training for multilingual neural machine translation. Wang et. al. 2020

- Optimize the data sampling distribution during training
- Upweight languages that have similar gradient with the multilingual dev set

Machine Translation

Translation





Mr. and Mrs. Dursley, who lived at number 4 on Privet Drive, were proud to say they were very normal, fortunately.

El señor y la señora Dursley, que vivían en el número 4 de Privet Drive, estaban orgullosos de decir que eran muy normales, afortunadamente.

Even if you don't know Spanish, can you find the correspondences between them?

Difficulties of translation: Syntactic divergences

The development of artificial intelligence is a really big deal.

El desarrollo de la inteligencia artificial es un asunto realmente importante.

The development of artificial intelligence is a really big deal.

人工知能の発展は本当にすごいことです。

Difficulties of translation: Syntactic divergences

(1) **Thematic divergence:**
E: I like Mary ⇔ S: María me gusta a mí
'Mary pleases me'

(2) **Promotional divergence:**
E: John usually goes home ⇔ S: Juan suele ir a casa
'John tends to go home'

(3) **Demotional divergence:**
E: I like eating ⇔ G: Ich esse gern
'I eat likingly'

(4) **Structural divergence:**
E: John entered the house ⇔ S: Juan entró en la casa
'John entered in the house'

(5) **Conflational divergence:**
E: I stabbed John ⇔ S: Yo le di puñaladas a Juan
'I gave knife-wounds to John'

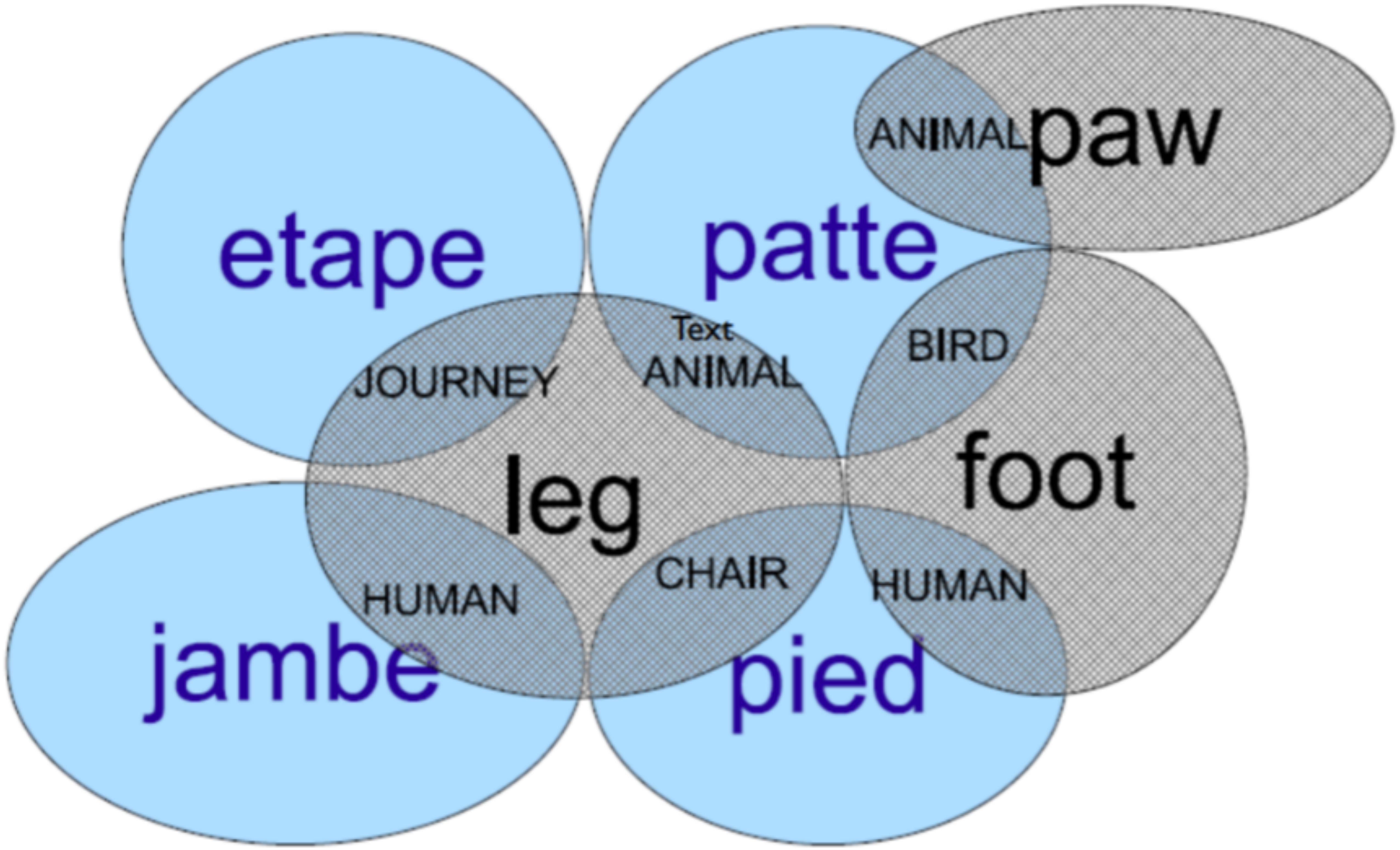
(6) **Categorial divergence:**
E: I am hungry ⇔ G: Ich habe Hunger
'I have hunger'

(7) **Lexical divergence:**
E: John broke into the room ⇔ S: Juan forzó la entrada al cuarto
'John forced (the) entry to the room'

Dorr 1994

Difficulties of translation: Lexical divergences

- Lexical ambiguities and divergences across languages

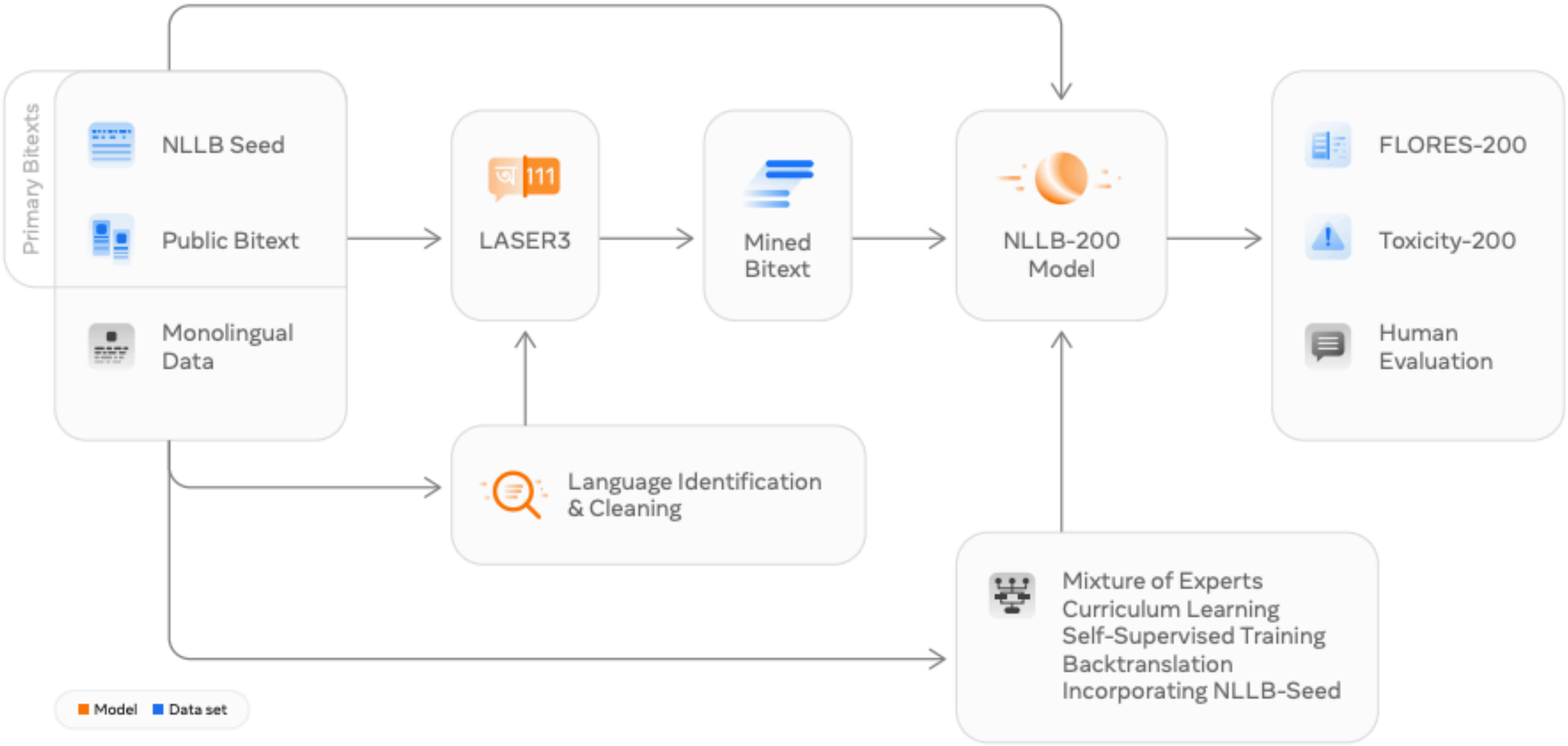


[Example from Jurafsky & Martin Speech and Language Processing 2nd ed.]

Translation tasks

- **WMT (the Conference on Machine Translation)** shared tasks—run every year for translation, evaluation, etc.
- **FLORES**: a dataset in 200 languages translated from English Wikipedia
- **IWSLT**: tasks on speech translation

No Language Left Behind (2022)



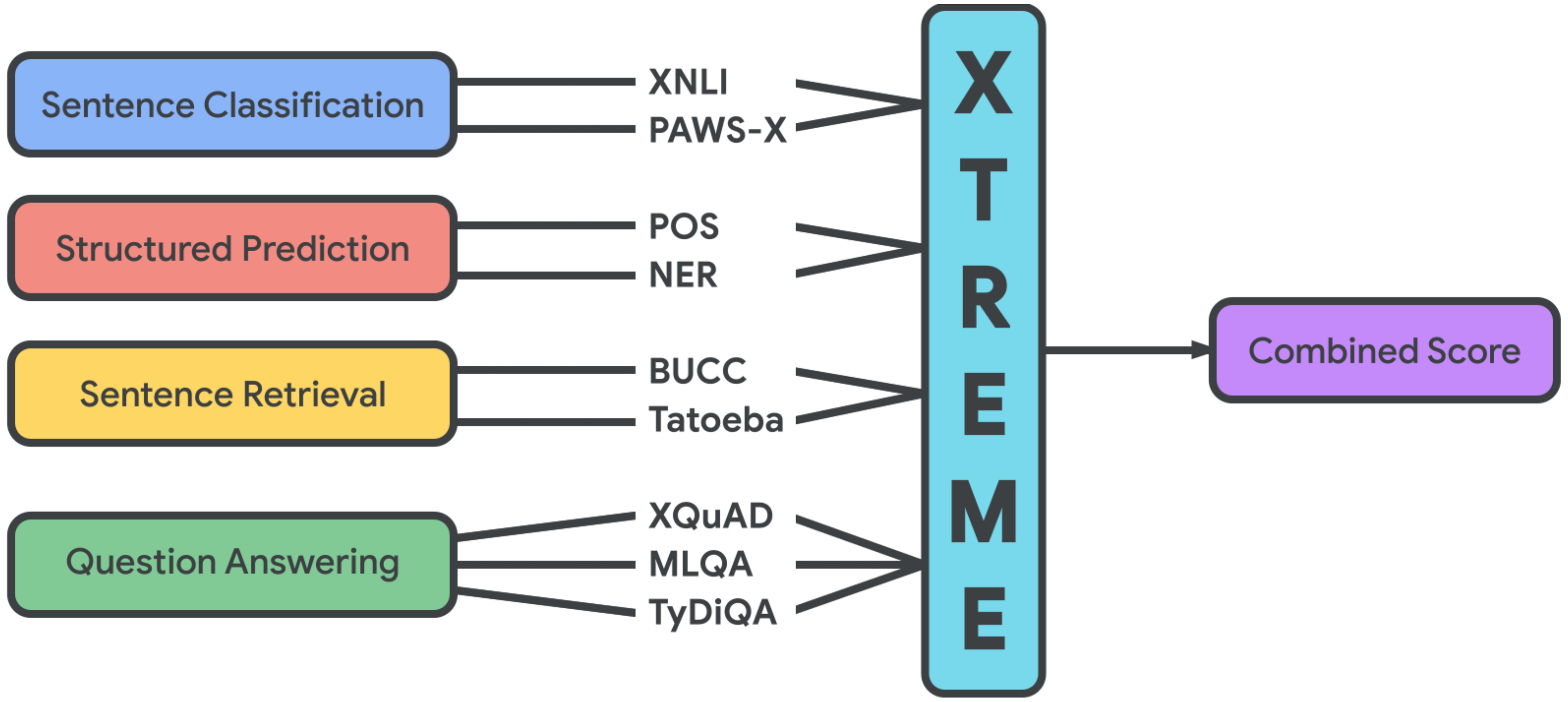
Multilingual Pre-trained Models

Multilinguality of general LLMs

- Closed LLMs such as GPT-4 are typically incidentally multilingual due to large training data
- Open LLMs (e.g., OLMo) often do data filtering to allow for good performance on English, and can be less multilingual
- Models such as mBERT, XLM, XLM-R extend BERT for multi-lingual pre-training

Multilingual representation evaluation

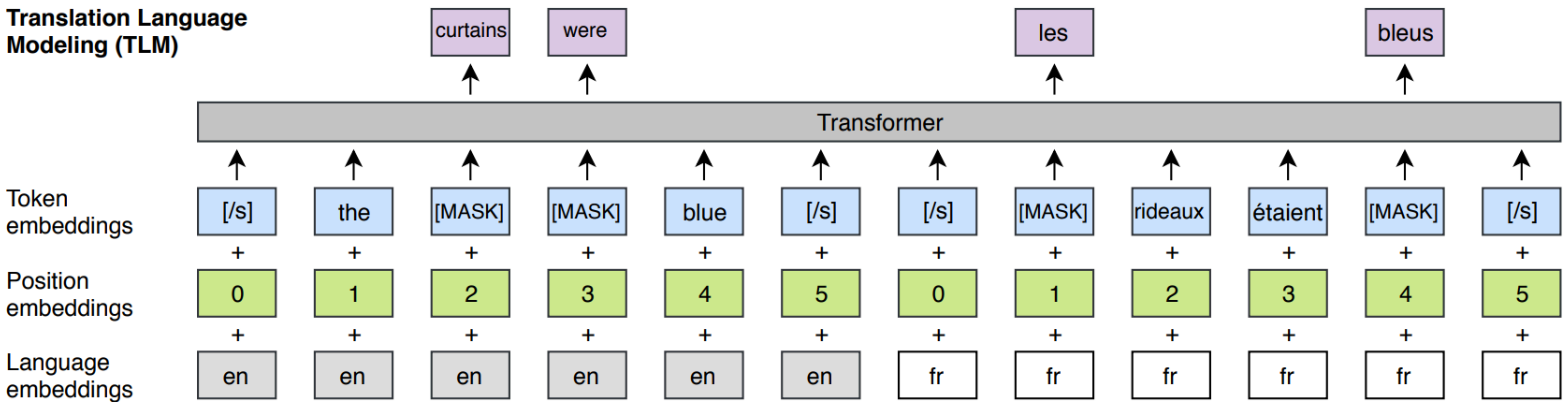
- Large-scale benchmarks that cover many tasks
  - **XTREME**: 40 languages, 9 tasks (Hu et al. 2020)



  - **XGLUE**: less typologically diverse but contains generation (Liang et al. 2020)
  - **XTREME-R** harder version based on XTREME (Ruder et al. 2021)

Multilingual masked language modeling

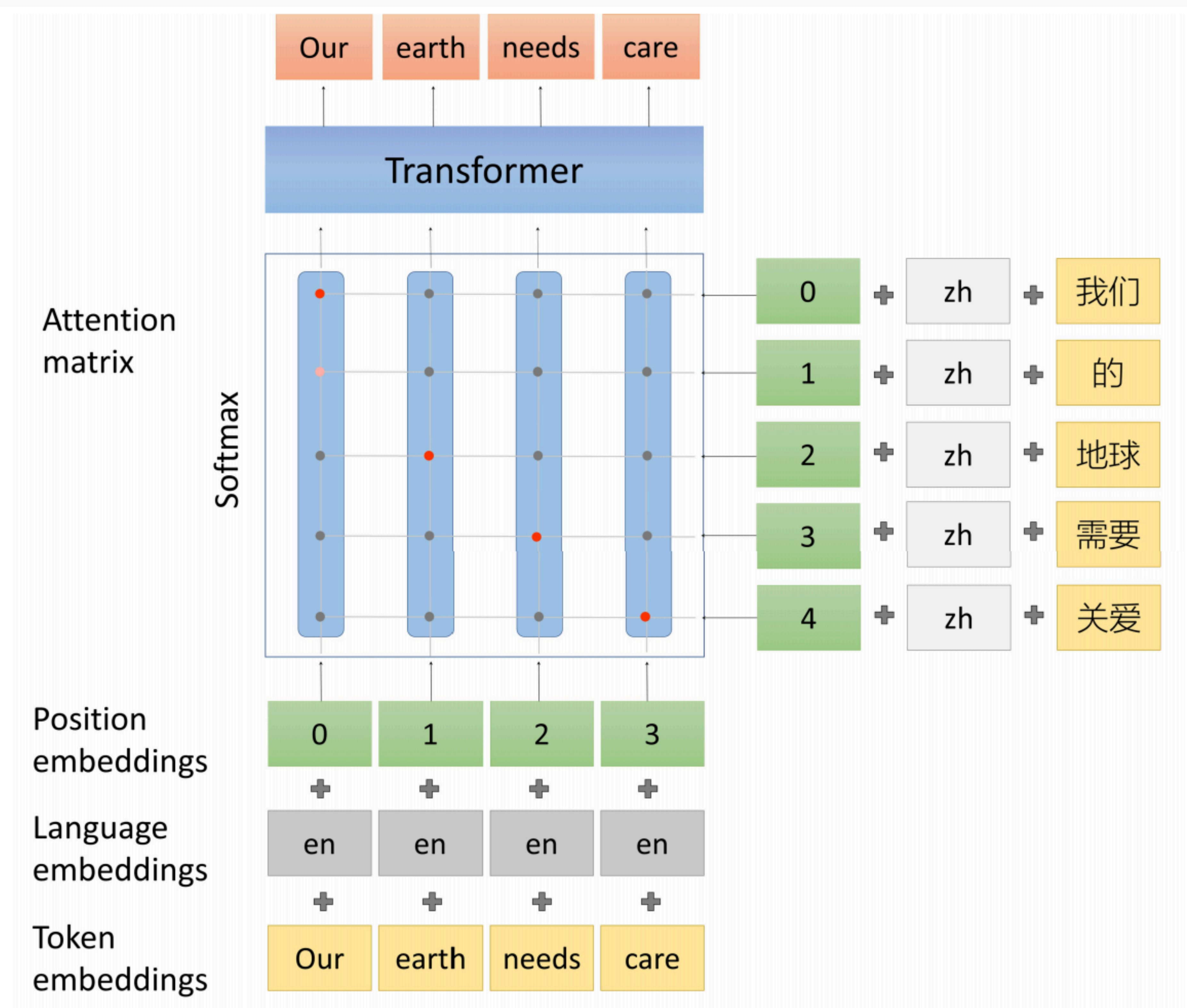
Also called translation language modeling (Lample and Conneau , 2019)



More explicit alignment objectives

Unicoder (Huang et al. 2019)

"cross-lingual word recovery"



AMBER (Hu et al. 2020)

bidirectional explicit alignment objective

$$\ell_{\mathrm{WA}}(x, y) = 1 - \frac{1}{H} \sum_{h=1}^{H} \frac{\mathrm{tr}\left(\mathbf{A}_{y \to x}^{h}{}^{T} \mathbf{A}_{x \to y}^{h}\right)}{\min(|x|, |y|)}$$

Multilingual encoder-decoder

- mT5 (Xue et al., 2020) is a multilingual encoder-decoder
- Trained on many languages, high performance

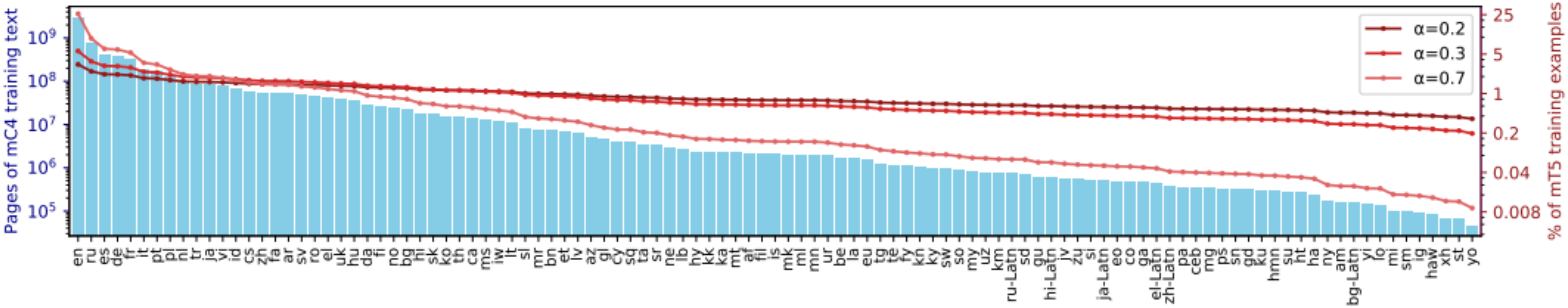
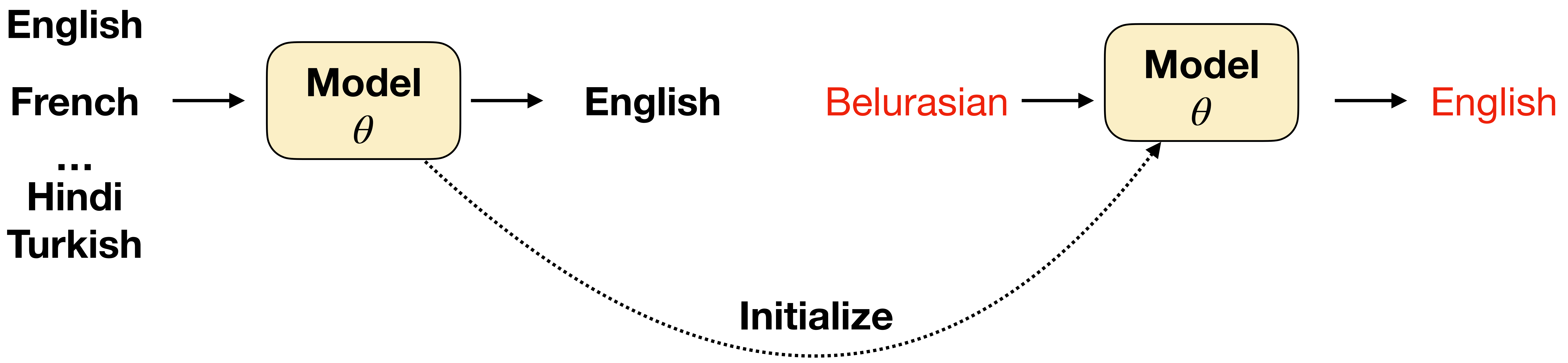


Figure 1: Page counts per language in mC4 (left axis), and percentage of mT5 training examples coming from each language, for different language sampling exponents α (right axis). Our final model uses α=0.3.

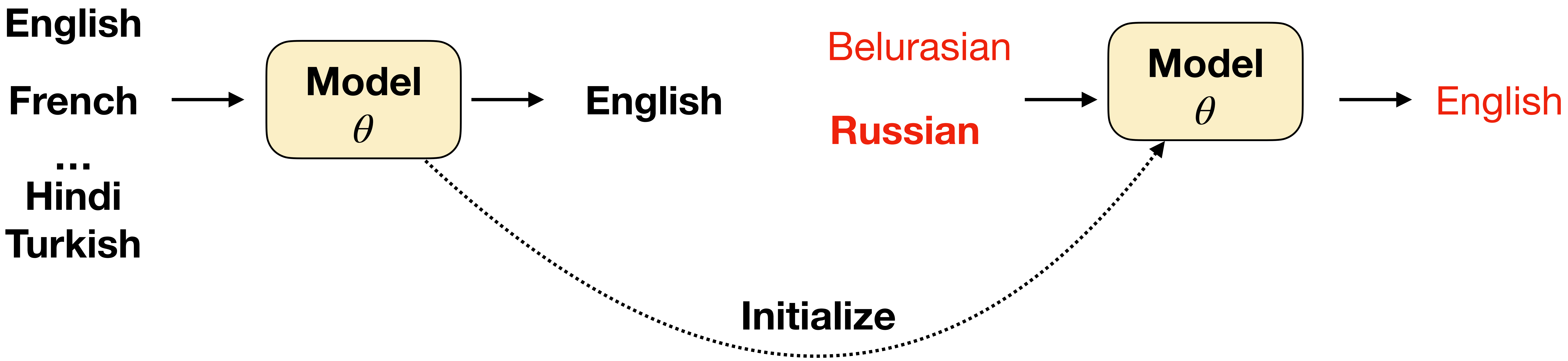
Pre-train and fine-tune



- First, do multilingual training on many languages (eg. 58 languages in the paper)
- Next fine-tune the model on a new low-resource language

Rapid adaptation of Neural Machine Translation to New Languages. Neubig et. al. 2018

Similar language regularization



- Regularized fine-tuning: fine-tune on low-resource language and its related high-resource language to avoid overfitting

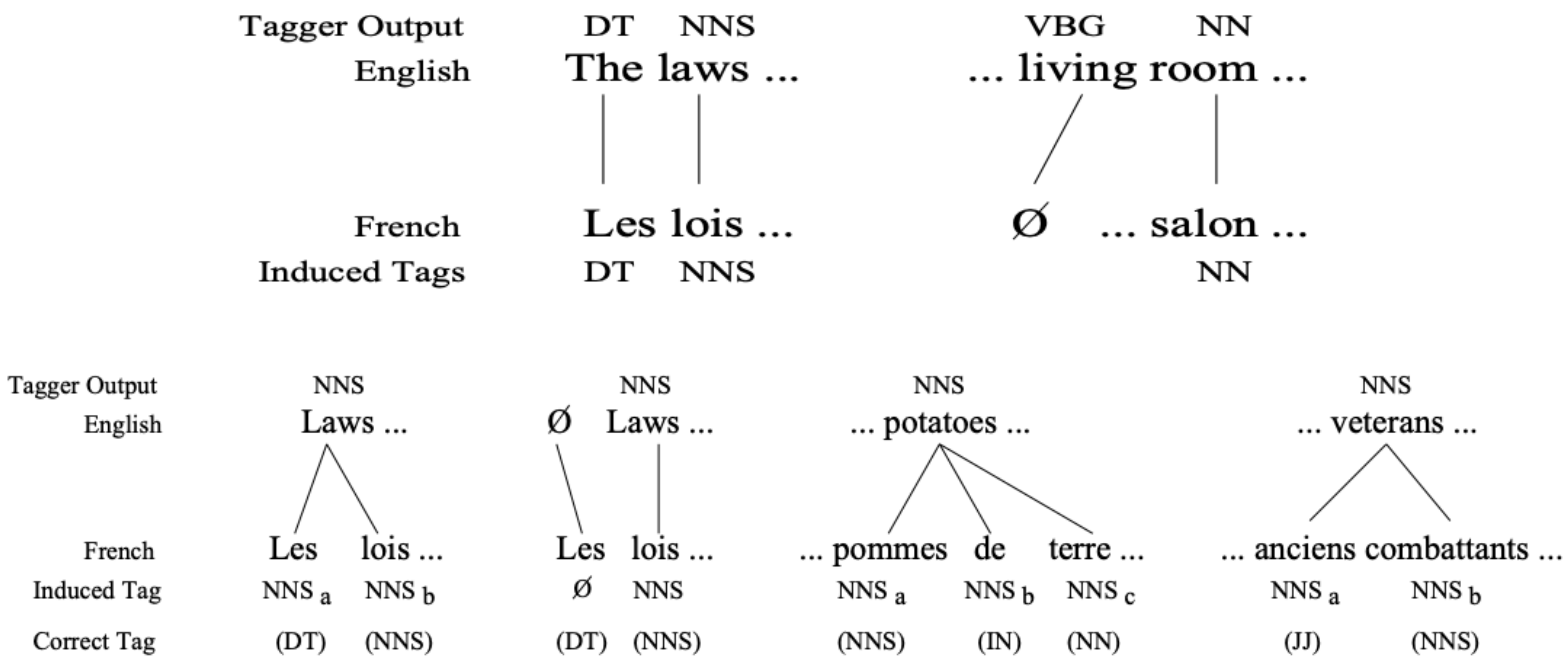
Rapid adaptation of Neural Machine Translation to New Languages. Neubig et. al. 2018

Zero-shot transfer for pretrained representations

- Pretrain: large language model using **monolingual data** from many different langauges
- Fine-tune: using **annotated data** in a given language (e.g., English)
- Test: test the fine-tuned model on a **different language** from the fine-tuning language (e.g., French)
- Multilingual pretraining learns a language-universal representation!
  - *How multilingual is multilingual BERT?* (Pires et al., 2019)

Annotation projection

Induce annotations in the target language using parallel data or bilingual dictionary (Yarowsky et al, 2001).

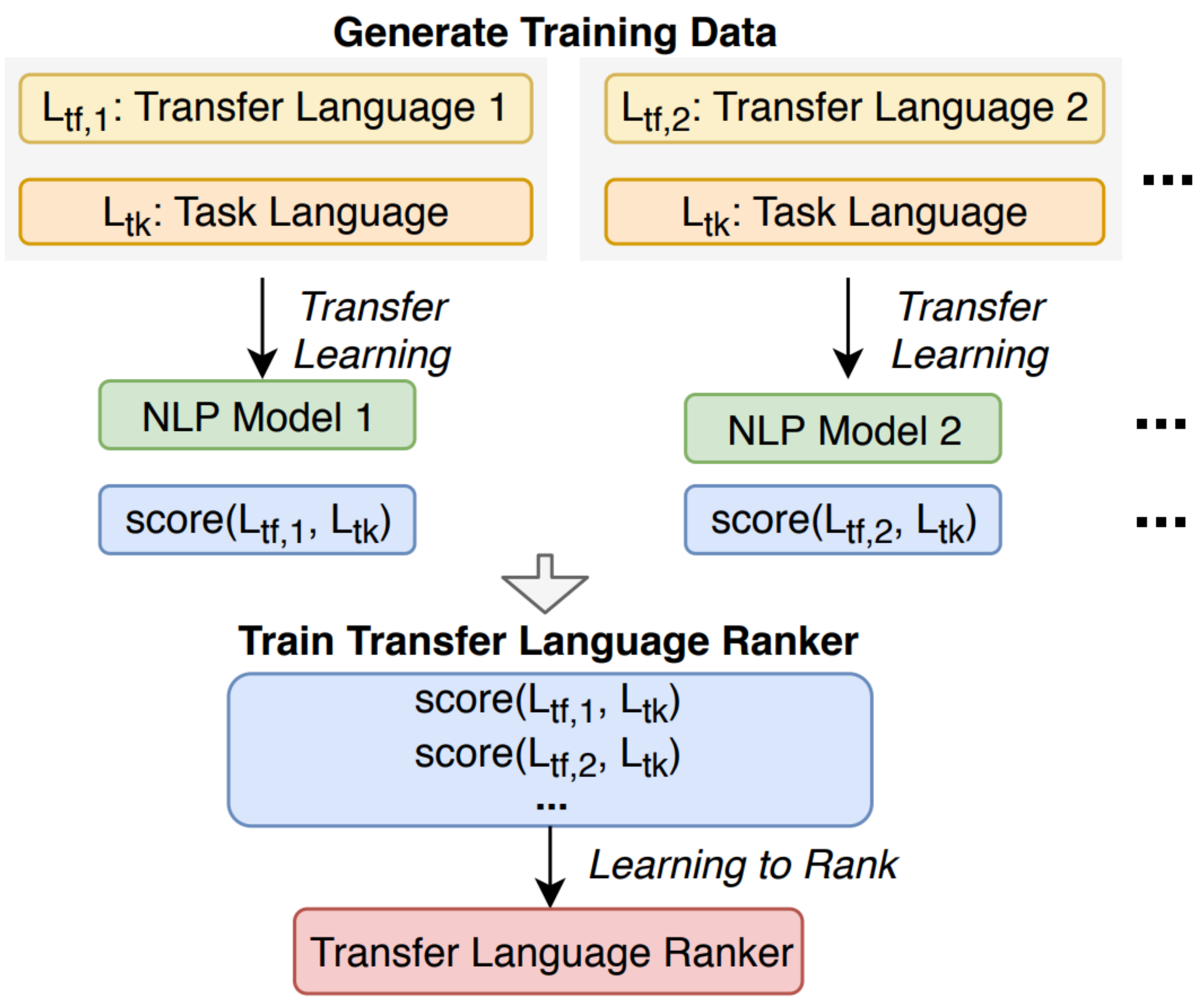


Which language to use for transfer?

When transferring from another language, it is ideal that it is

- **Similar** to the target language
- **Data-rich**

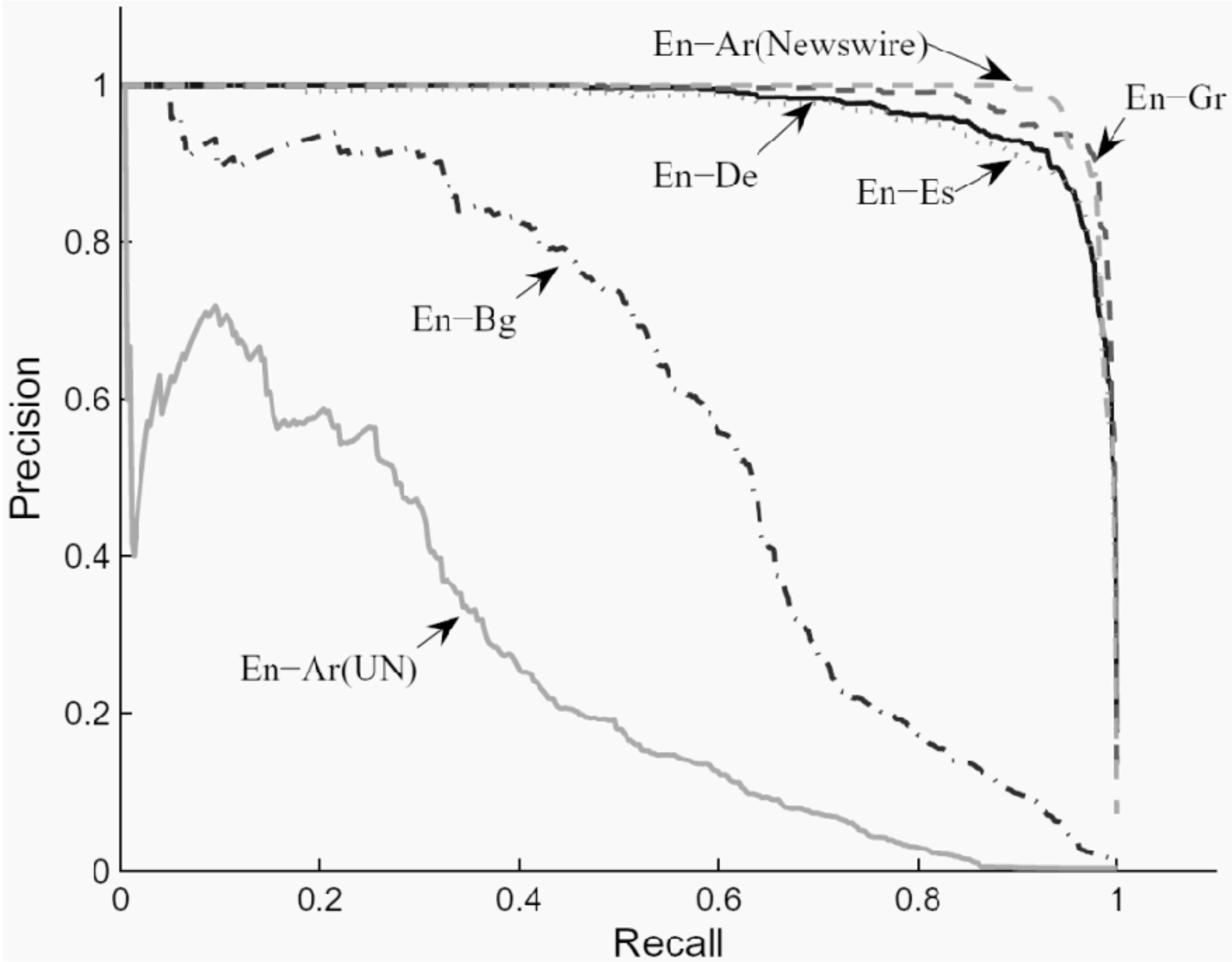
Lin et al. (2019) examine how to identify better transfer languages



	Method	MT	EL	POS	DEP
dataset	word overlap o_w	28.6	30.7	13.4	52.3
	subword overlap o_{sw}	29.2	–	–	–
	size ratio s_{tf}/s_{tk}	3.7	0.3	9.5	24.8
	type-token ratio d_{ttr}	2.5	–	7.4	6.4
ling. distance	genetic d_{gen}	24.2	50.9	14.8	32.0
	syntactic d_{syn}	14.8	46.4	4.1	22.9
	featural d_{fea}	10.1	47.5	5.7	13.9
	phonological d_{pho}	3.0	4.0	9.8	43.4
	inventory d_{inv}	8.5	41.3	2.4	23.5
	geographic d_{geo}	15.1	49.5	15.7	46.4
	LANGRANK (all)	51.1	**63.0**	**28.9**	**65.0**
	LANGRANK (dataset)	**53.7**	17.0	26.5	**65.0**
	LANGRANK (URIEL)	32.6	58.1	16.6	59.6

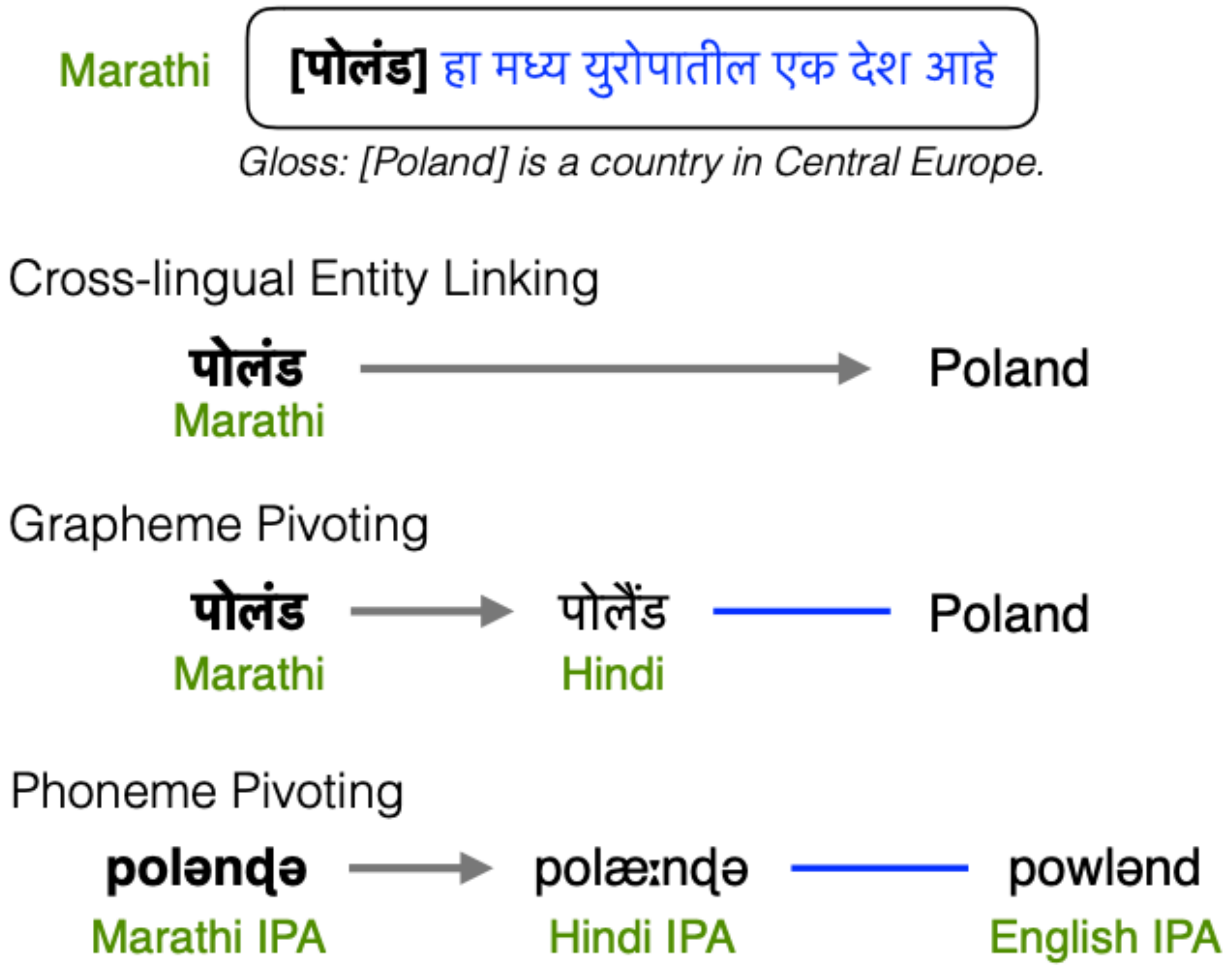
What if languages don't share a script?

- Some tokens (e.g., numerals, dates) can still be shared
- Can get high accuracy at the document level
- High variance depending on orthography (Krstovski et al., 2011)



What if languages don’t share a script?

- Use phonological representations to make the similarity between languages apparent.
- e.g.: Rijhwani et al (2019) use a pivot-based entity linking system for low-resource languages.



Multilingual summary

- LLMs can work with more languages than any human!
- But the "curse of multilinguality" imposes tradeoffs
- How to balance depends on our goals
  - Perform tasks independently in multiple languages?
    - Eventually specialize for important languages
  - Perform cross-language tasks?
    - Source-target asymmetry
      - Cf. speech→text and text→speech
- Much more with Terra Blevins on Tuesday!