CS11-711 Advanced NLP

Multilingual NLP

Graham Neubig

Carnegie Mellon University
Language Technologies Institute

https://phontron.com/class/anlp-fall2024/

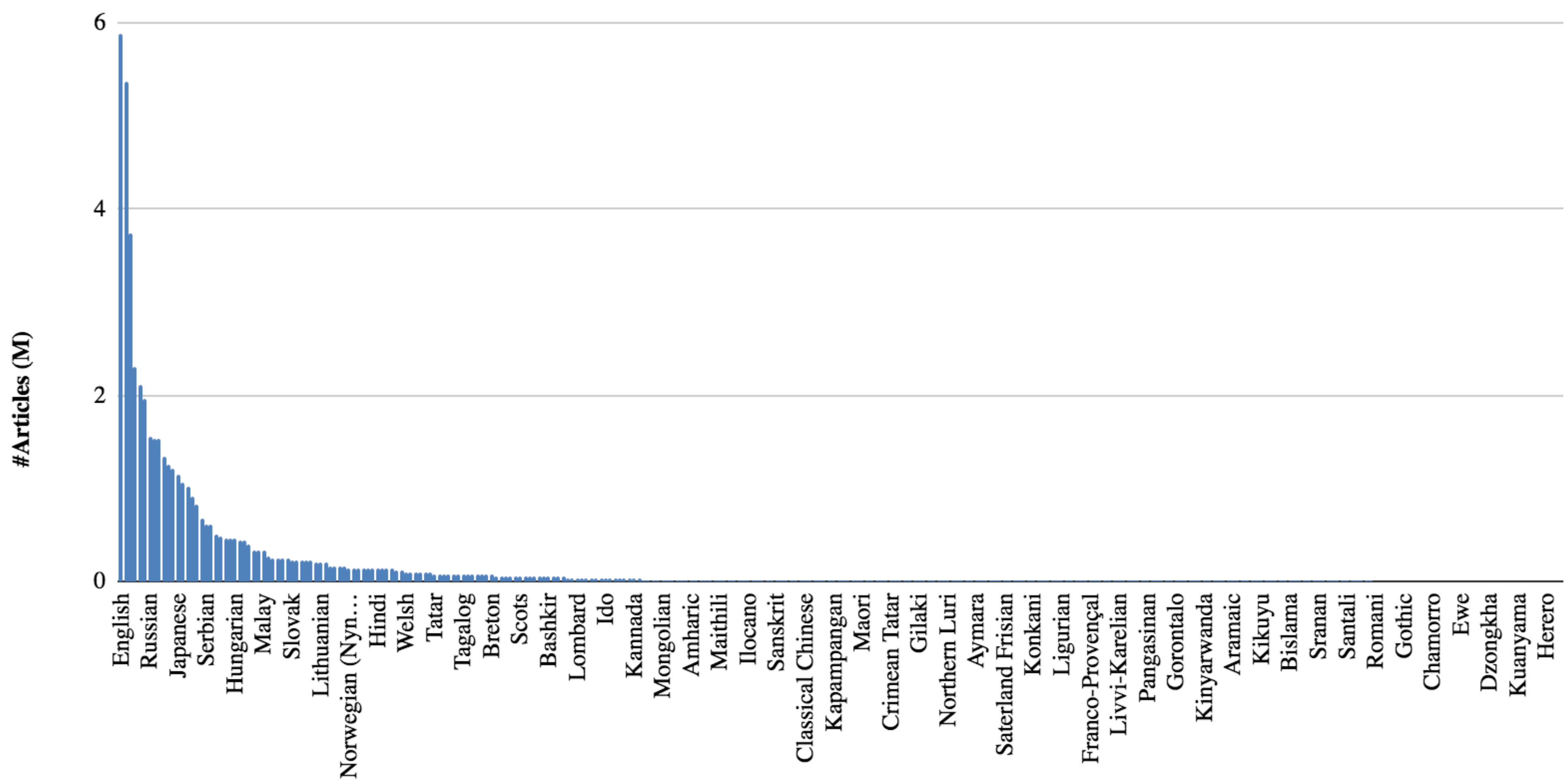
w/ Slides by Aditi Chaudhary, Xinyi Wang

Multilingual NLP and its Difficulties

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-lingual QA
  - …

Paucity of data



- Big disparity in monolingual data available for training
- Even less annotated data for NMT, sequence label, dialogue…

Data Source: Wikipedia articles from different languages

Linguistic Peculiarities

- Most methods are tested first on English, but not all languages are the same as English
- e.g.
  - Rich morphology (case, gender, etc.)
  - Accents/diacritics
  - Different scripts such as CJK
  - Dialectal language
  - Lack of formal writing systems

Data Source: Wikipedia articles from different languages

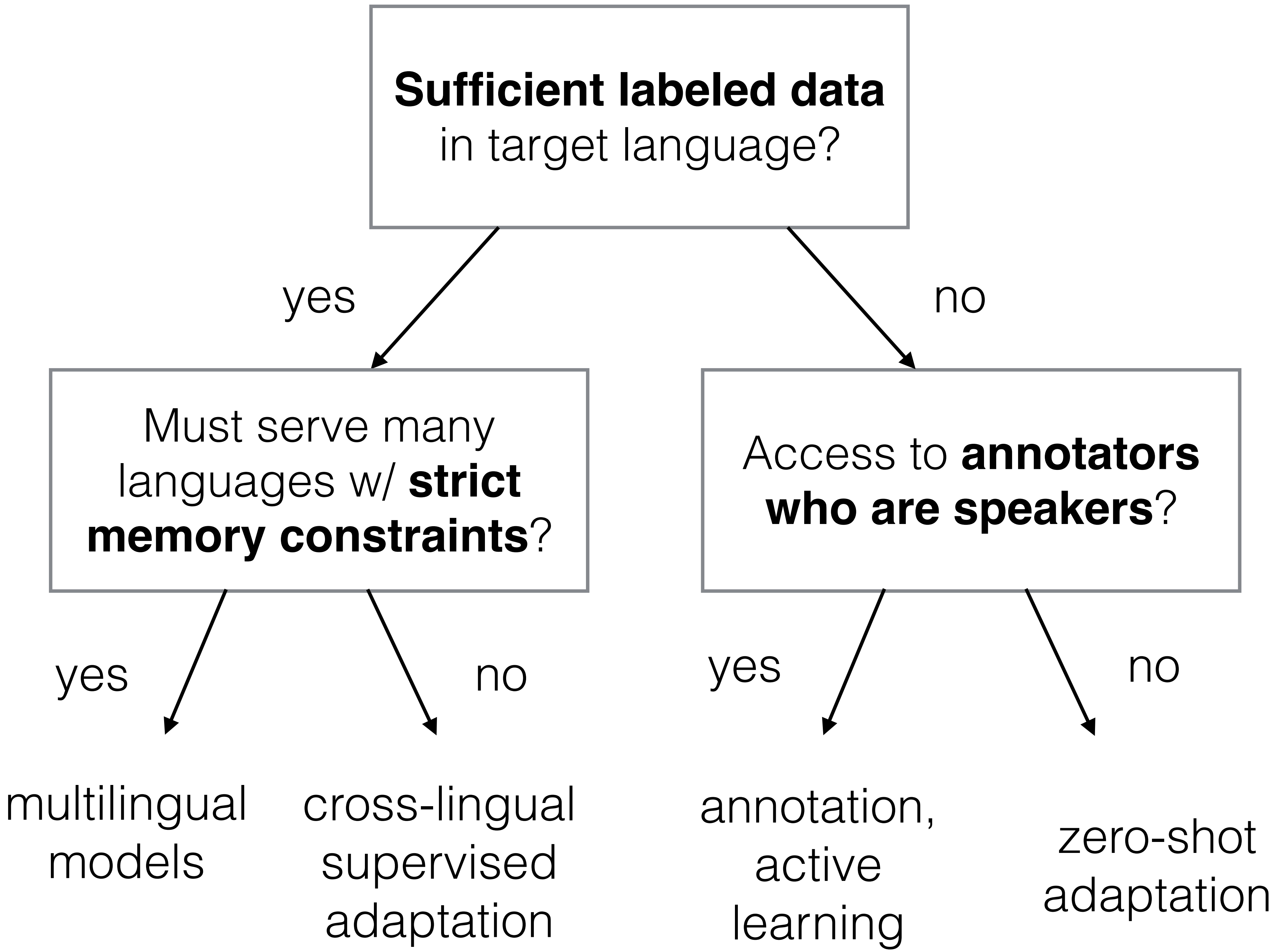
An Aside: Language or Dialect?

- There is also an interest in *dialect processing*
- *Idealized definition:* languages are mutually unintelligible
- *Reality:* “a language is a dialect with an army”
- Processing dialects is similar to processing similar languages but
  - Data scarcity is worse
  - Arguments are more nuanced

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

High-level Multilingual Learning Flowchart



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

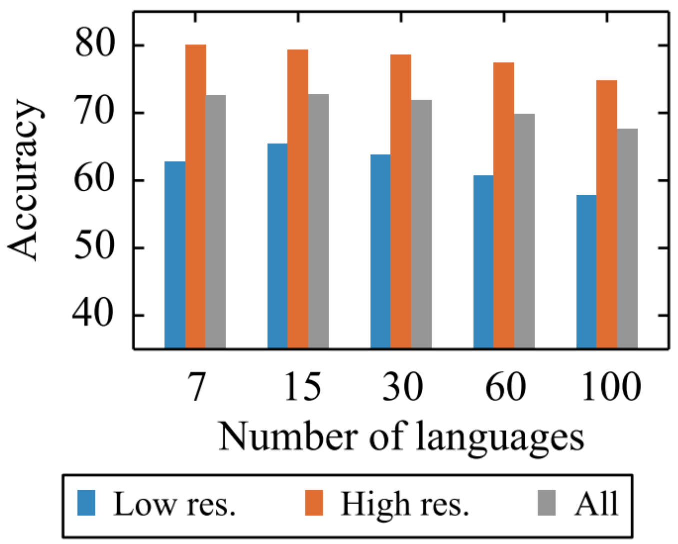
- **Multilingual Output:** Add a tag or prompt about the target language for generation (Johnson et al. 2016)

<fr> this is an example → ceci est un exemple

<ja> this is an example → これは例です

Difficulties in Fully Multi-lingual 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)

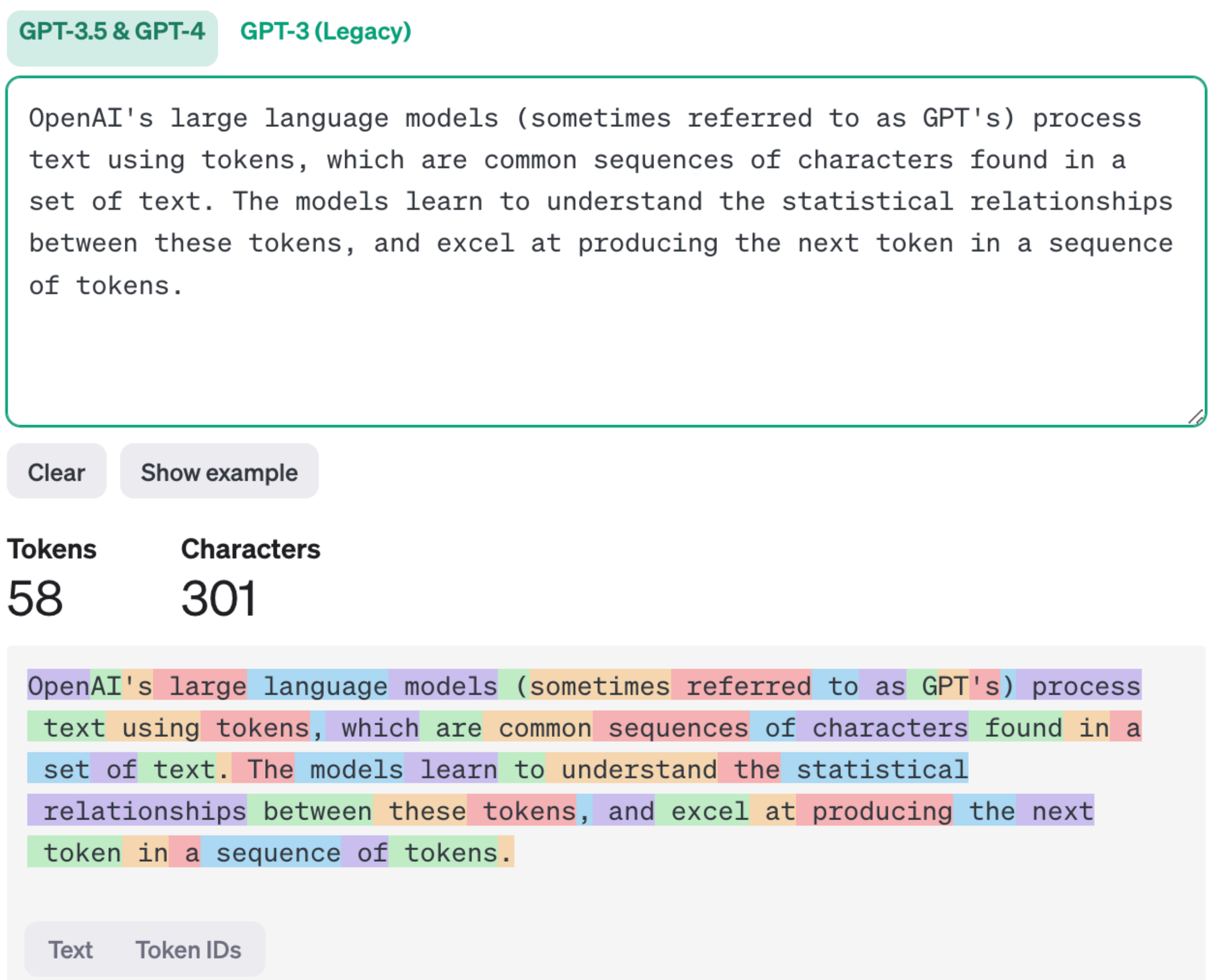


- Increasing the number of low-resource languages —> decrease in the quality of high-resource language translations (Aharoni et al, 2019)

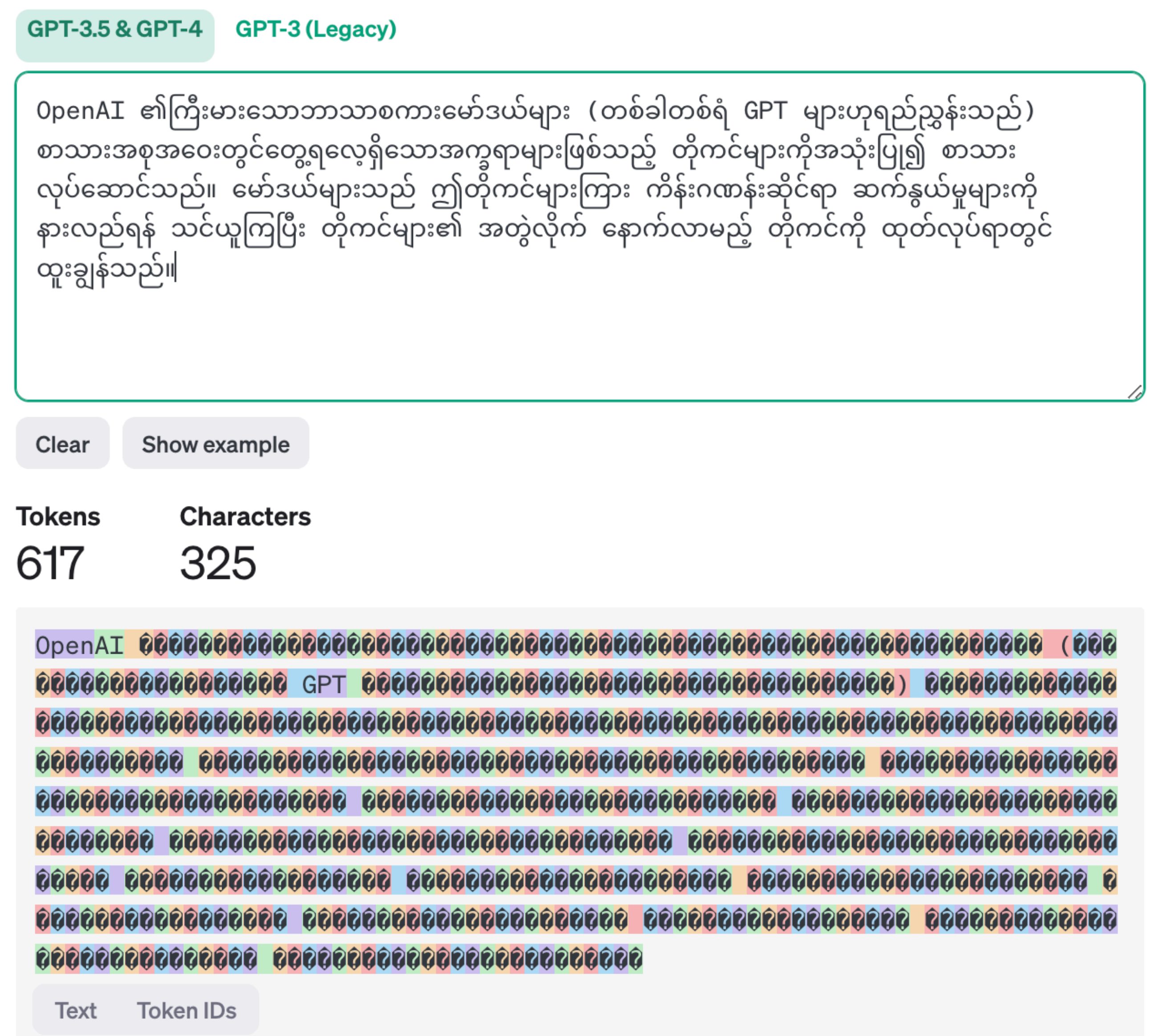
- How to mitigate? **Better data balancing, better parameter sharing**

Tokenization Disparity

English

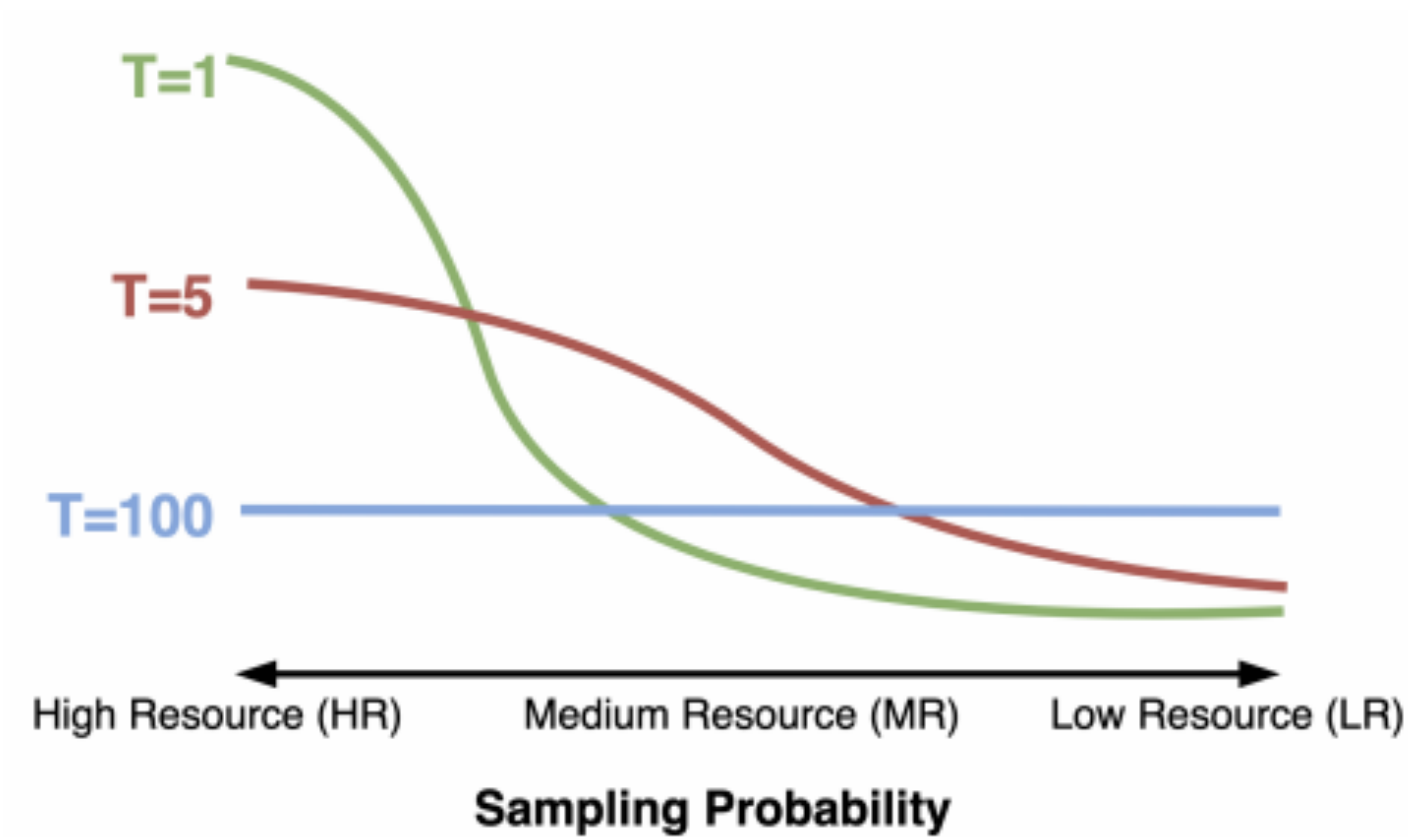


Burmese/Myanmar (Google Translated)



Similar content, 10.6x the tokens!

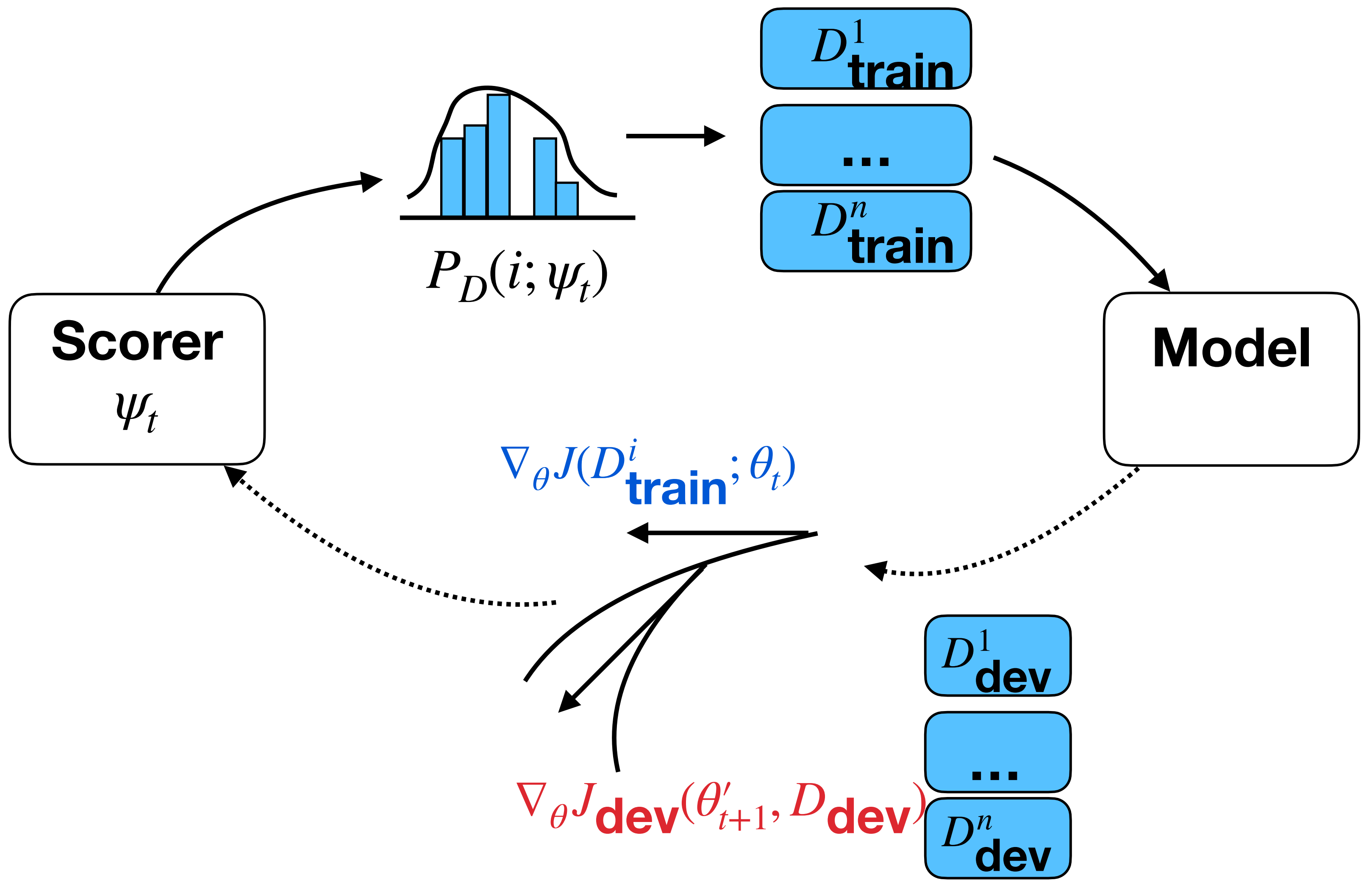
Heuristic Sampling of Data



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

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

Learning to Balance Data

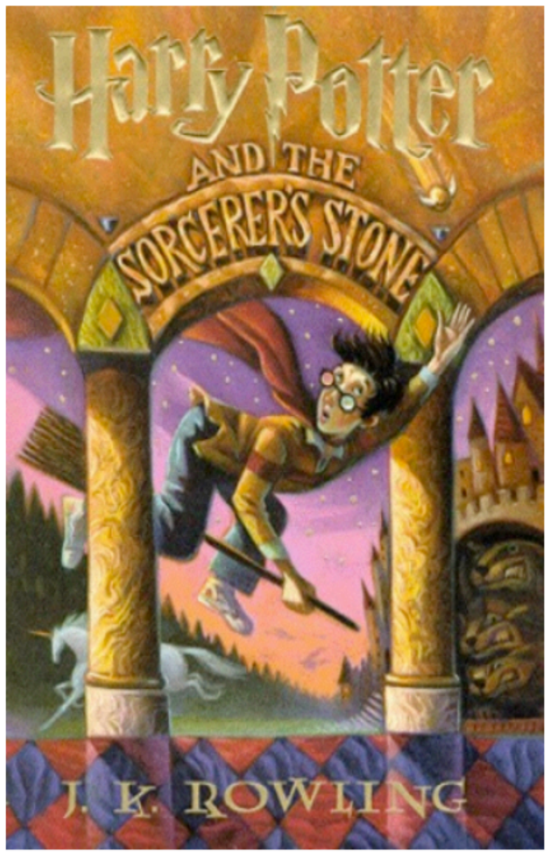


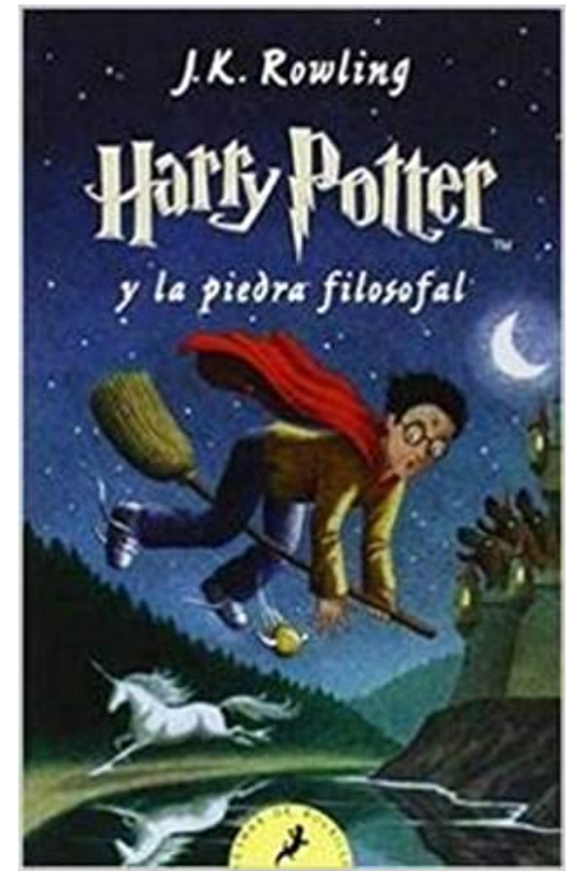
- Optimize the data sampling distribution during training
- Upweight languages that have similar gradient with the multilingual dev set

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

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.

Why is it difficult to translate?

- Syntactic divergences between languages

The development of artificial intelligence is a really big deal.

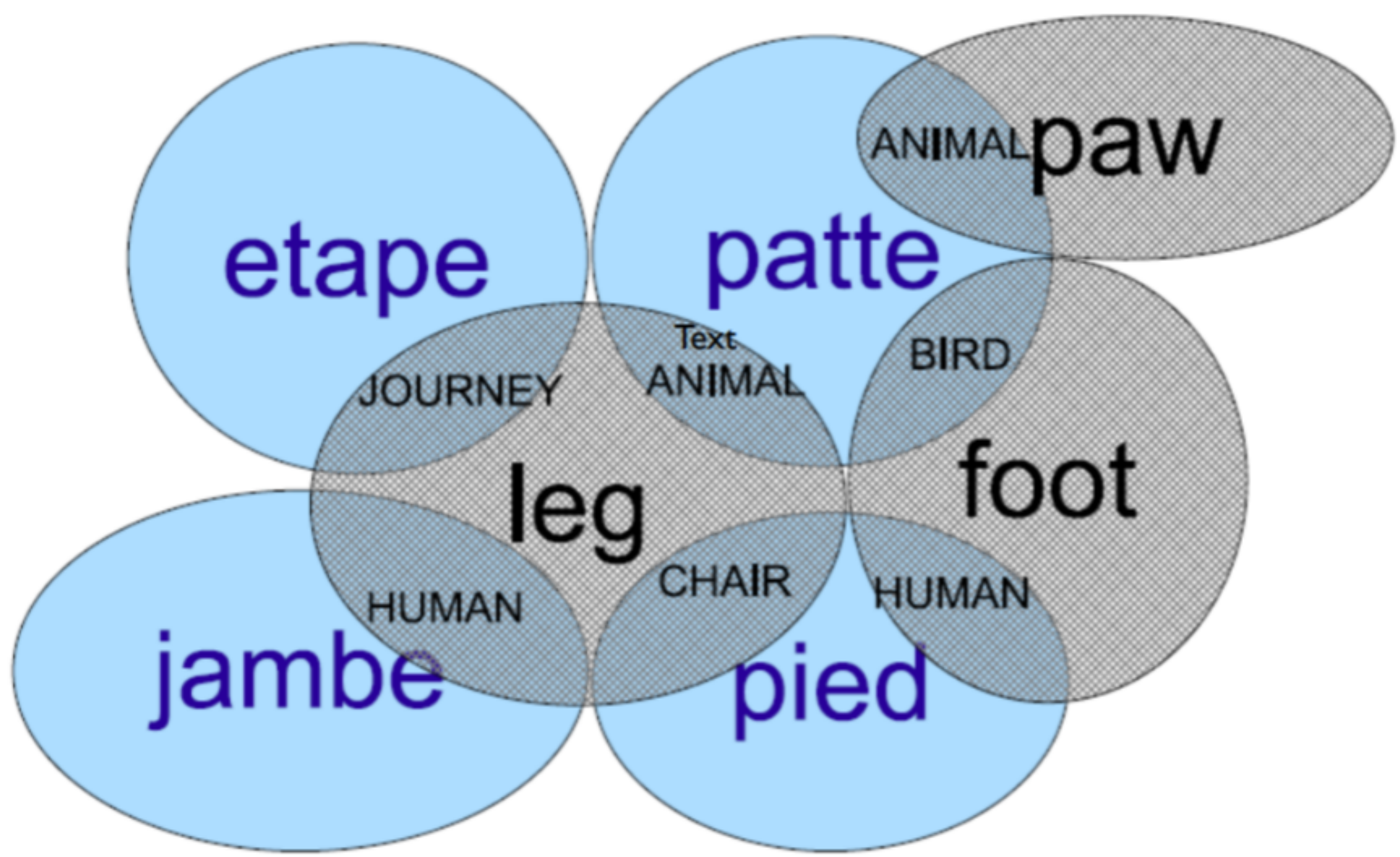
El desarrollo de la inteligencia artificial es un asunto realmente importante.

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。

Why is it difficult to translate?

- 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

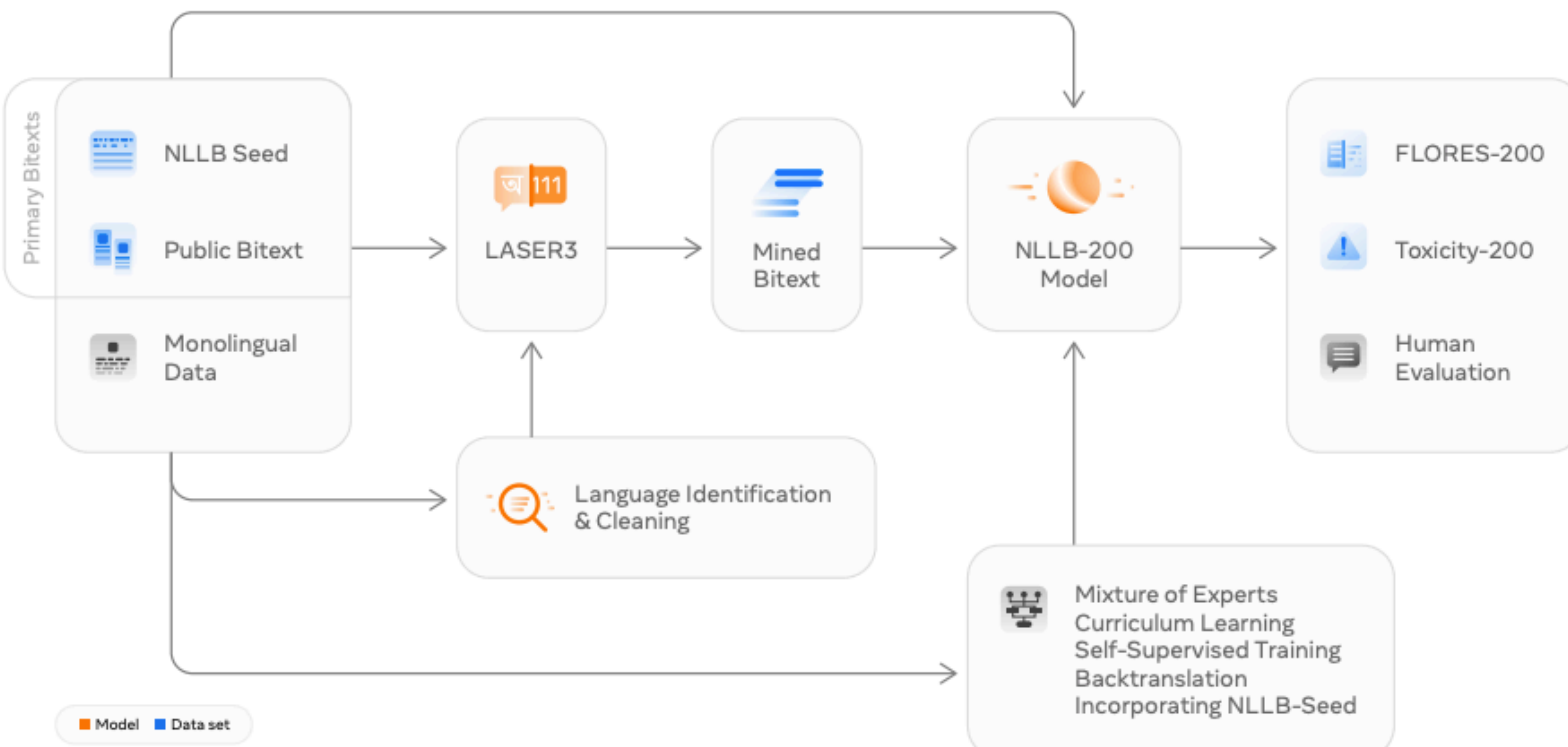
Automatically Evaluating MT

- **BLEU:** Measure overlap of token n-grams (Papineni et al. 2002)
  - Problem: doesn't consider paraphrases, morphology, etc.
- **chrF:** Based on character n-grams instead (Popovic et al. 2015)
- **COMET:** Trained based on multilingual embeddings (Rei et al. 2020)
- **GEMBA:** Ask an LM how good the translation is (Kocmi and Federmann 2023)

NLLB Translation Model

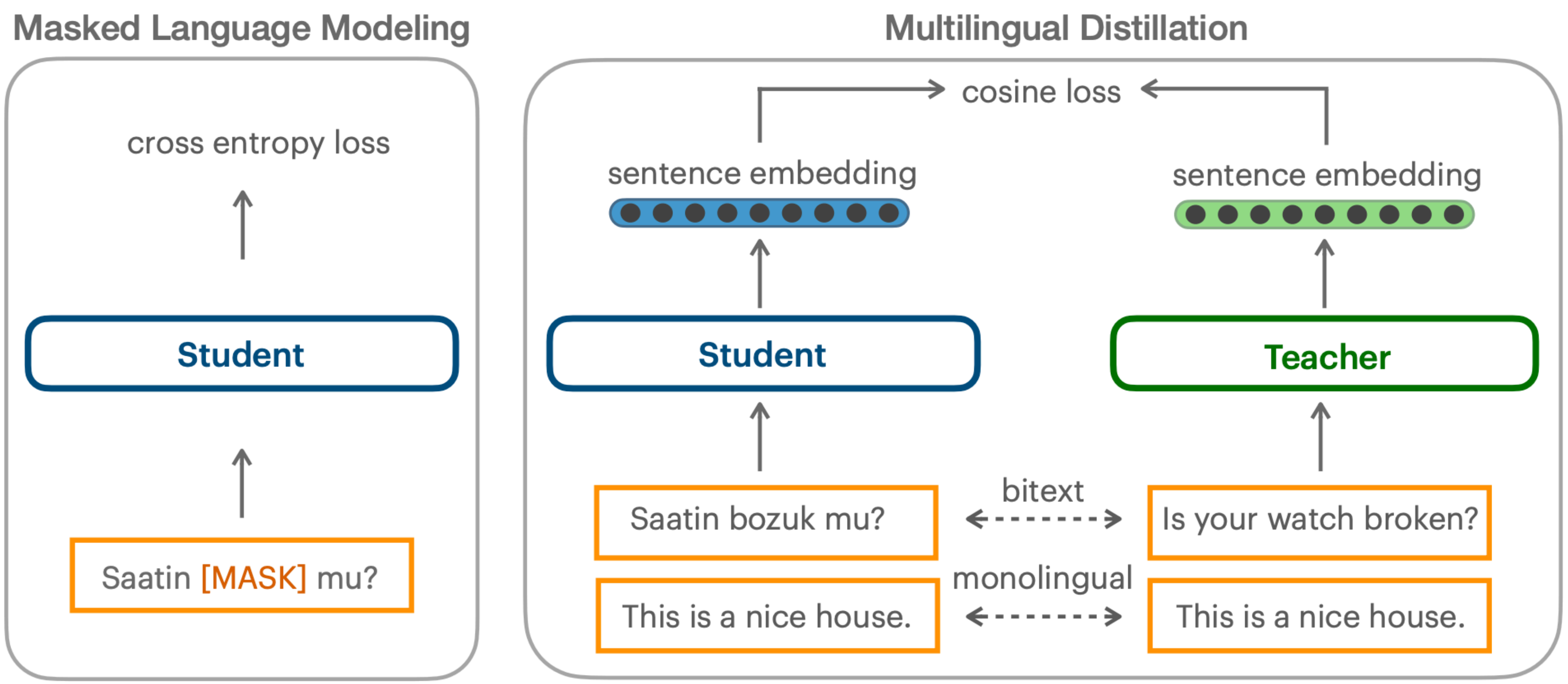
(NLLB Team 2022)

- Example of building a strong MT model



Bitext Mining w/ Sentence Embeddings (Heffernan et al. 2022)

- Take sentence representations and adapt them for similarity search

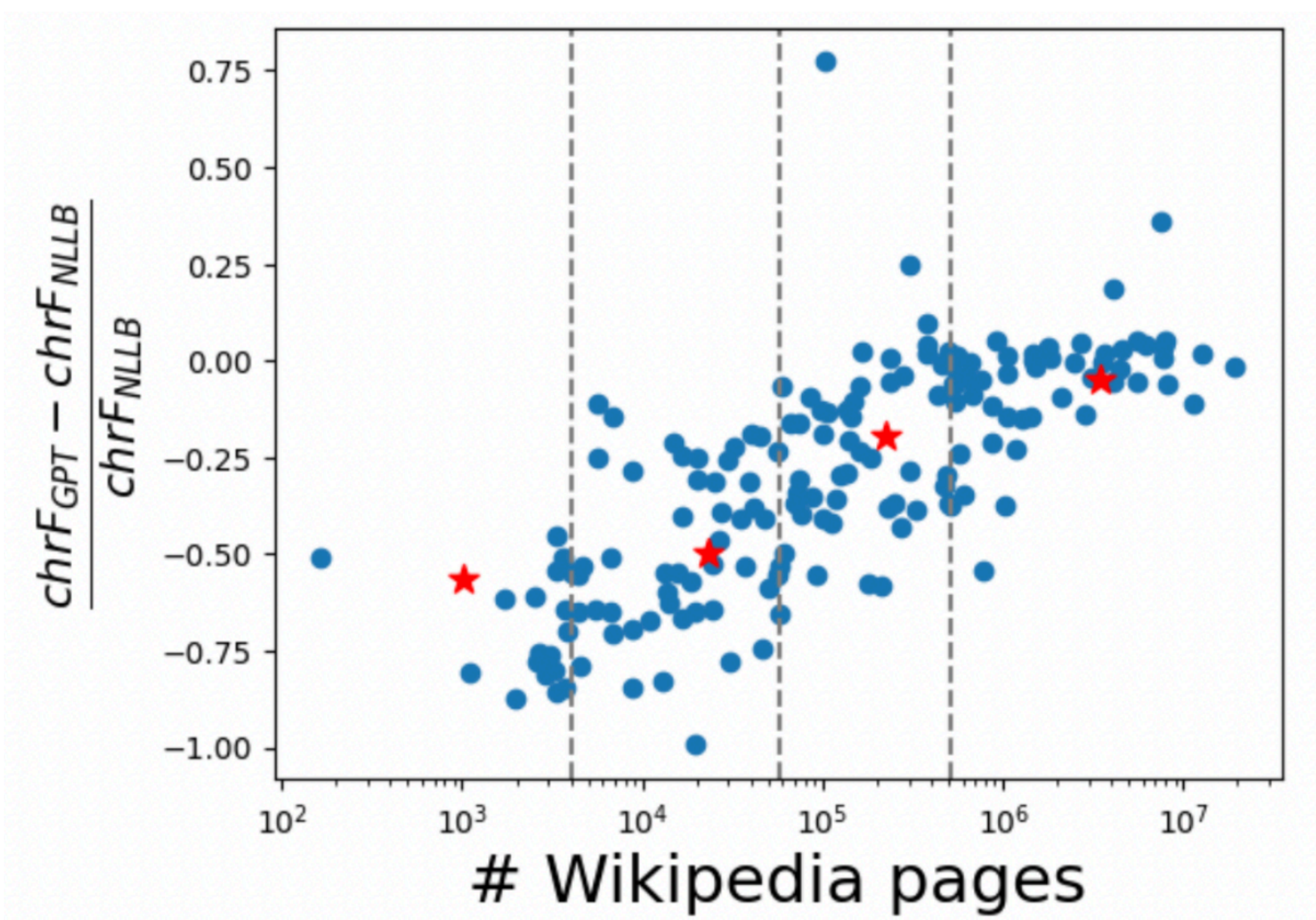


- Search for the most similar sentence, normalized to prevent "hubness"

$$\texttt{xsim}(x, y) = \textbf{margin}(cos(x, y),$$
$$\sum_{z \in NN_k(x)} \frac{cos(x, z)}{2k} + \sum_{z \in NN_k(y)} \frac{cos(y, z)}{2k})$$

Can we Use LLMs as-is for Translation?

- We can just ask an LLM to translate
- Results can be good for high-resource languages, but less so for low-resource languages (Robinson et al. 2023)



Multilingual Pre-trained Models

Multilinguality of Standard LLMs

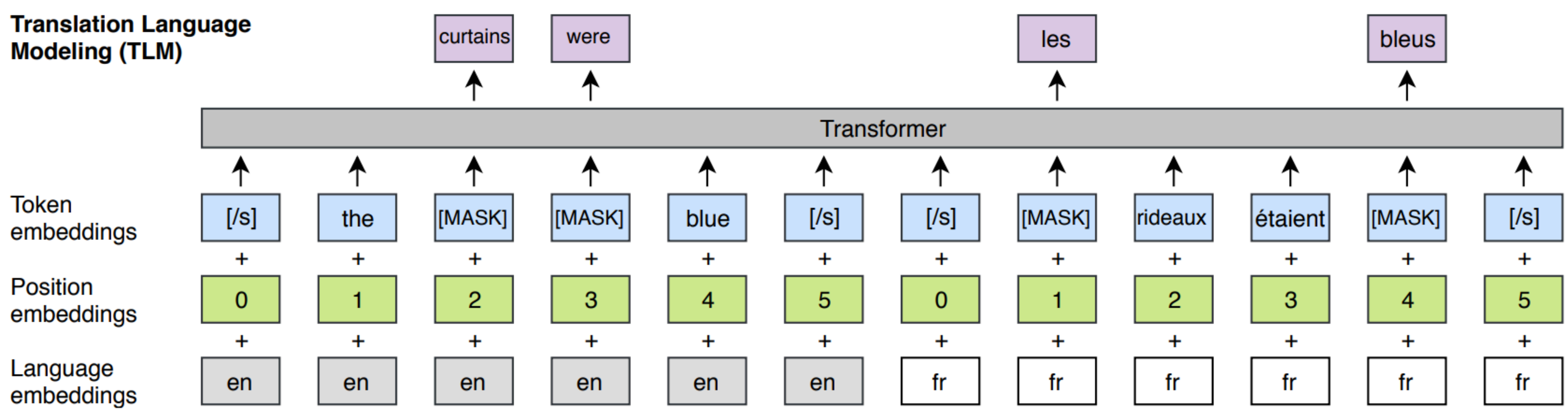
- Closed LLMs such as GPT-4 are typically incidentally multilingual due to large training data
- Open LLMs often do data filtering to allow for good performance on English, and can be less multilingual

Multi-lingual Representation Learning

- Language model pre-training has shown to be effective for many NLP tasks, eg. BERT
- BERT uses masked language model (MLM) and next sentence prediction (NSP) objective.
- Models such as mBERT, XLM, XLM-R extend BERT for multi-lingual pre-training.

Multilingual Masked Language Modeling

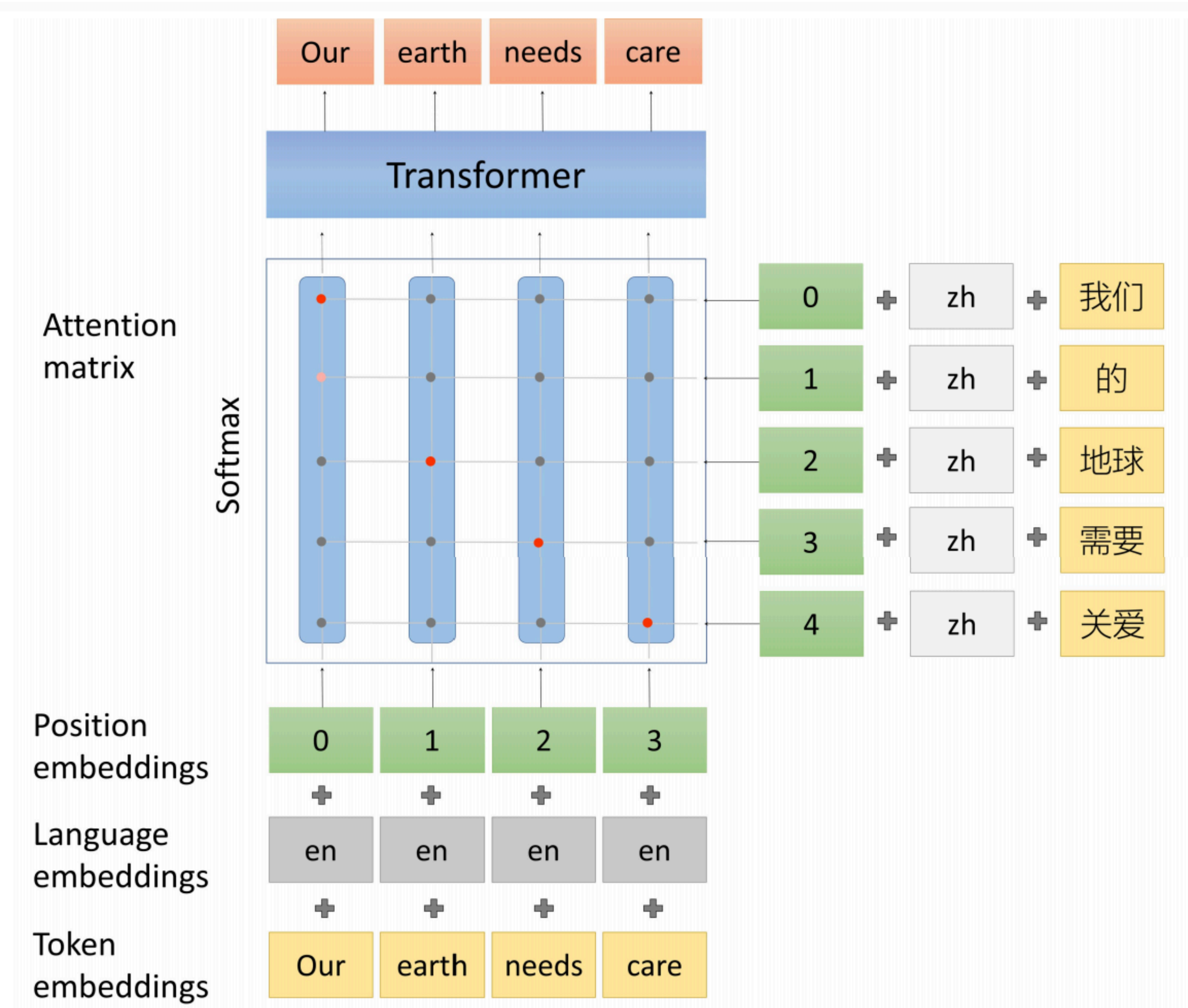
- Also called translation language modeling (Lample and Conneau 2019)



More Explicit Alignment Objectives

Unicoder (Huang et al. 2019)

"cross-lingual word recovery"



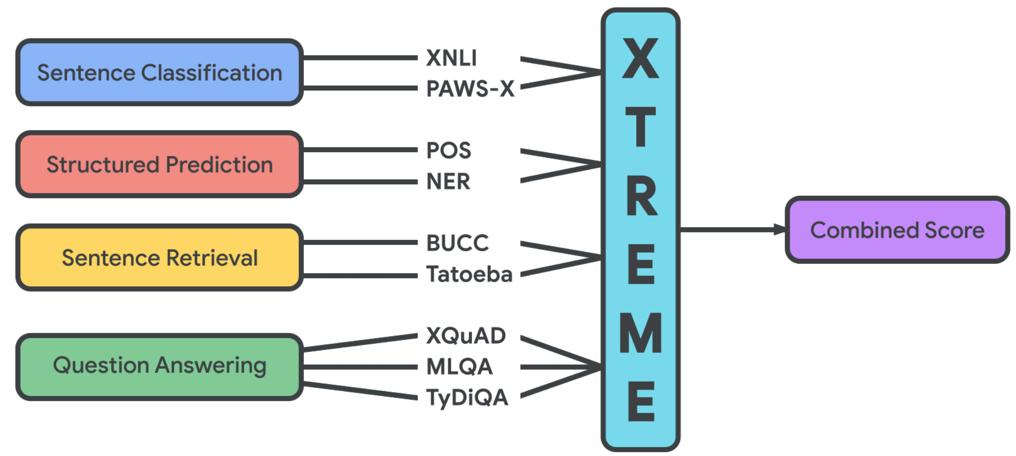
AMBER (Hu et al. 2020)

bidirectional explicit alignment objective

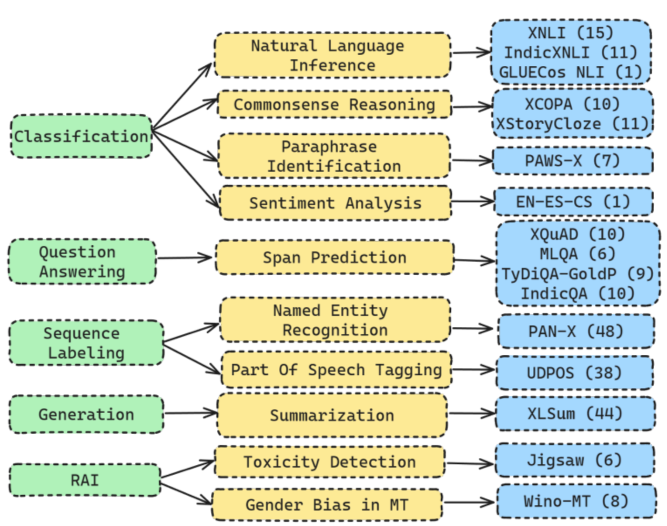
$$\ell_{\mathrm{WA}}(x, y) = 1 - \frac{1}{H}\sum_{h=1}^{H}\frac{\mathrm{tr}\left(\mathbf{A}^{h}_{y\to x}{}^{T}\mathbf{A}^{h}_{x\to y}\right)}{\min(|x|, |y|)}$$

Multilingual Understanding Evaluation

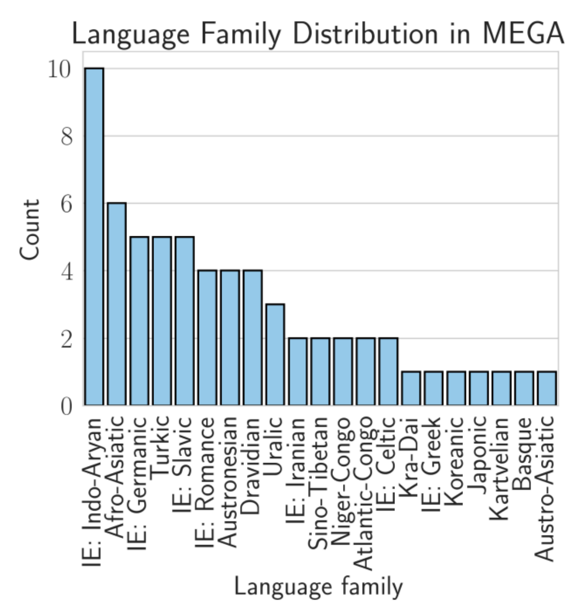
- **XTREME:** 40 languages, 9 tasks focused on representation-based models (Hu et al. 2020)



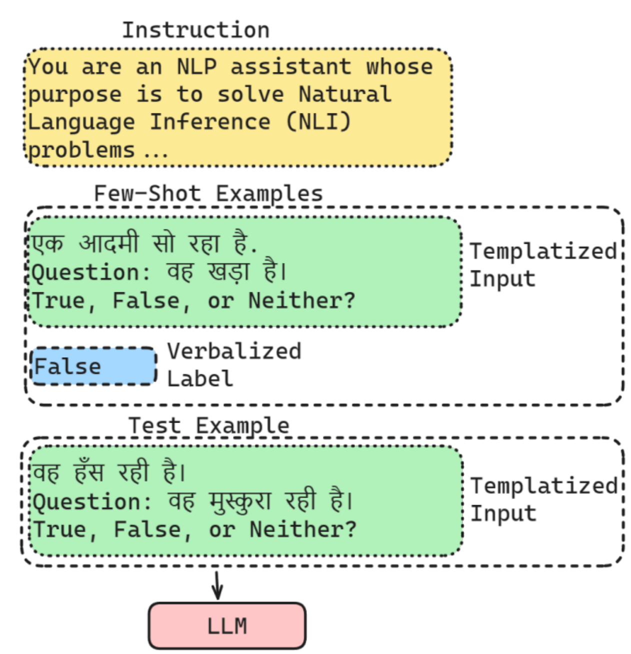
- **MEGA**: Focused more on LLMs (Ahuja et al. 2023)



(a) Tasks and Datasets included in MEGA.



(b) Language Family Distribution



(c) Example of multilingual prompting

Explicitly Multilingual Pre-training

mT5 (Xue et al 2020)

- Multilingual encoder-decoder
- Trained on many languages, high performance

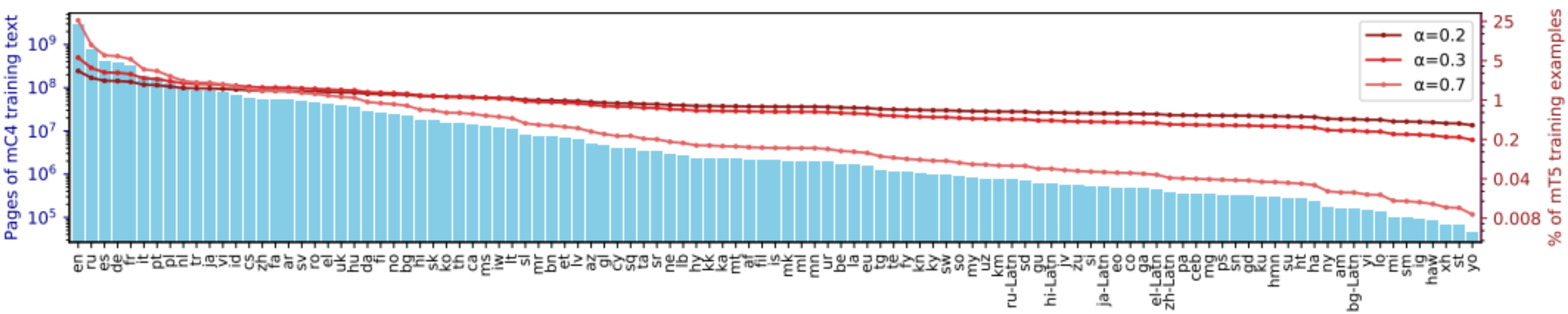


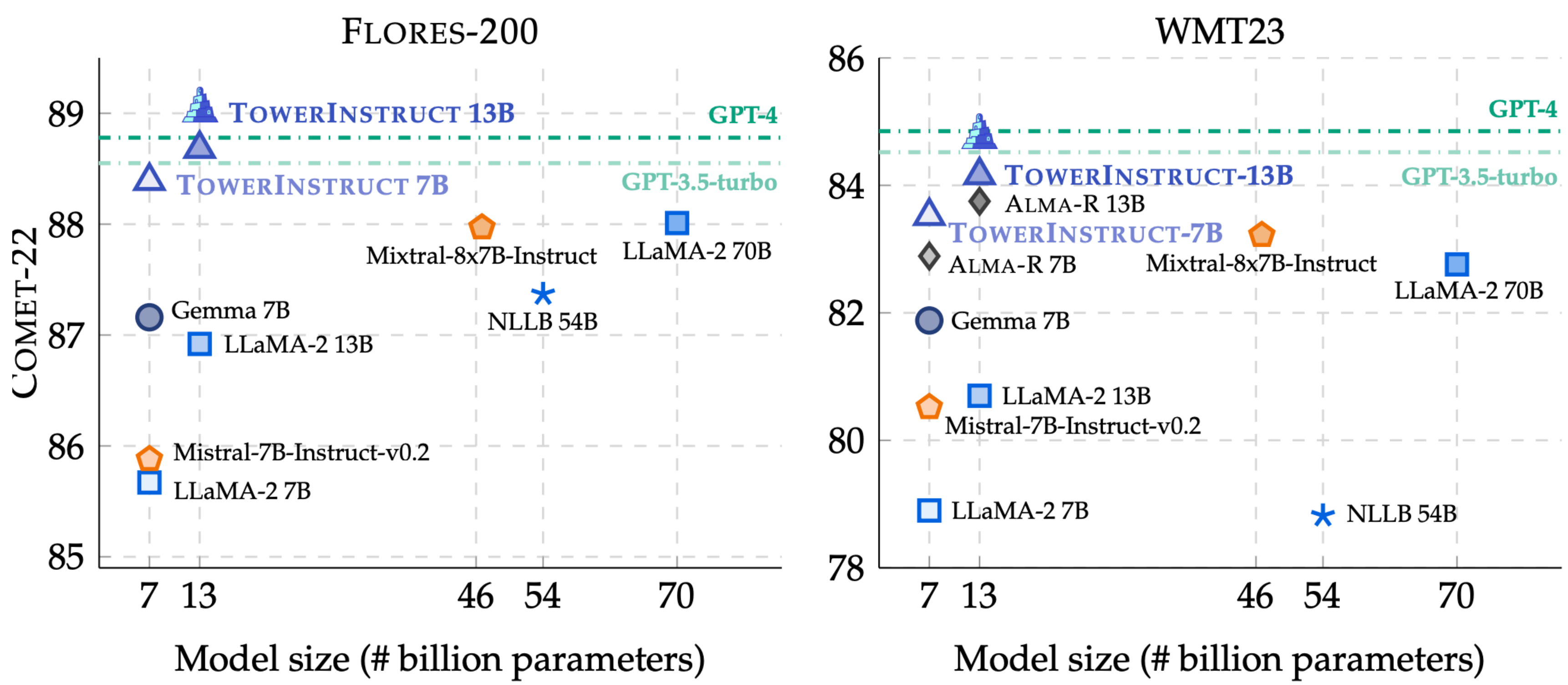
Figure 1: Page counts per language in mC4 (left axis), and percentage of mT5 training examples coming from each language, for different language sampling exponents α (right axis). Our final model uses α=0.3.

Aya 23 (Aryabumi et al. 2024)

- 8B and 35B autoregressive LMs
- **Pre-training:** based on standard pre-trained model w/ good multilingual balance (Command-R)
- **Fine-tuning:** multilingual templates, the Aya dataset of human-labeled data (204k), translated data, synthetic data

Tower (Alves et al. 2024)

- 8B autoregressive LM tailored specifically for translation
- **Pre-training:** llama
- **Continued pre-training:** On translation and filtered monolingual data
- **Fine-tuning:** multilingual instruction tuning data
- Results: strong results on translation tasks

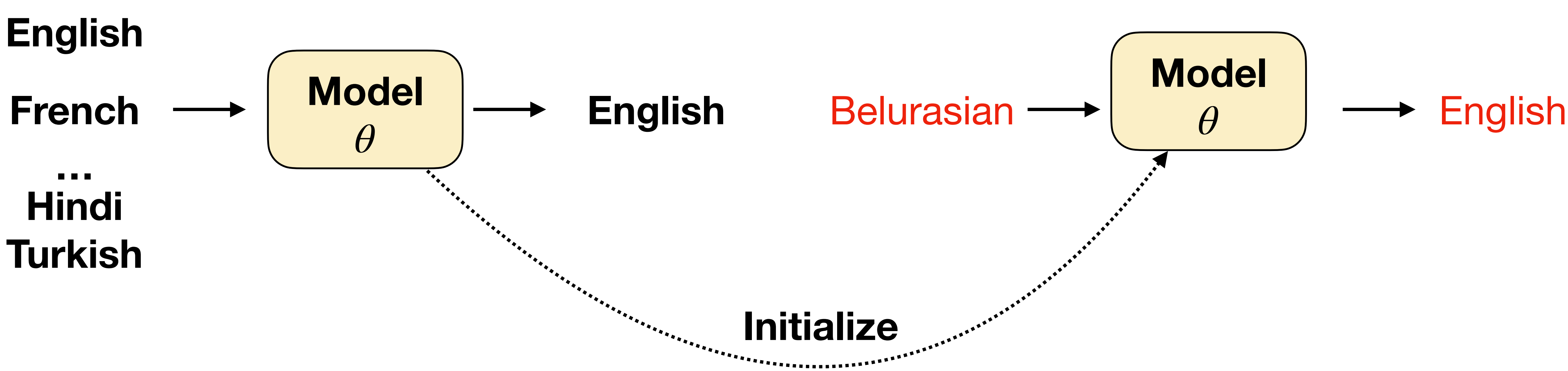


Advanced Modeling Strategies

Cross-lingual Transfer Learning

- CLTL leverages data from one or more high-resource source languages.
- **Popular strategies:**
  - Multilingual learning (above)
  - Pre-train and fine-tune
  - Zero-shot transfer
  - Annotation projection

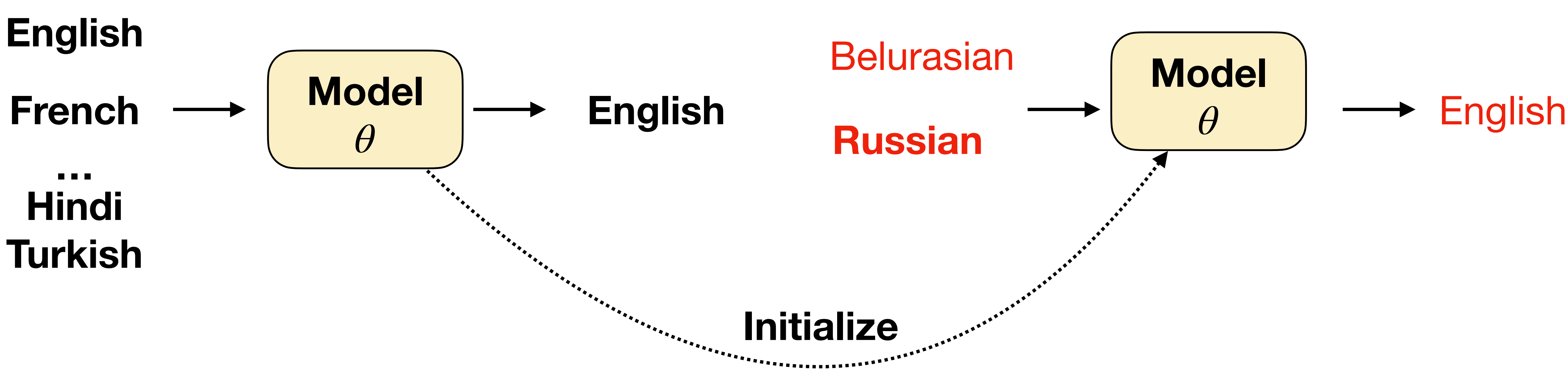
Pre-train and Fine-tune



- 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

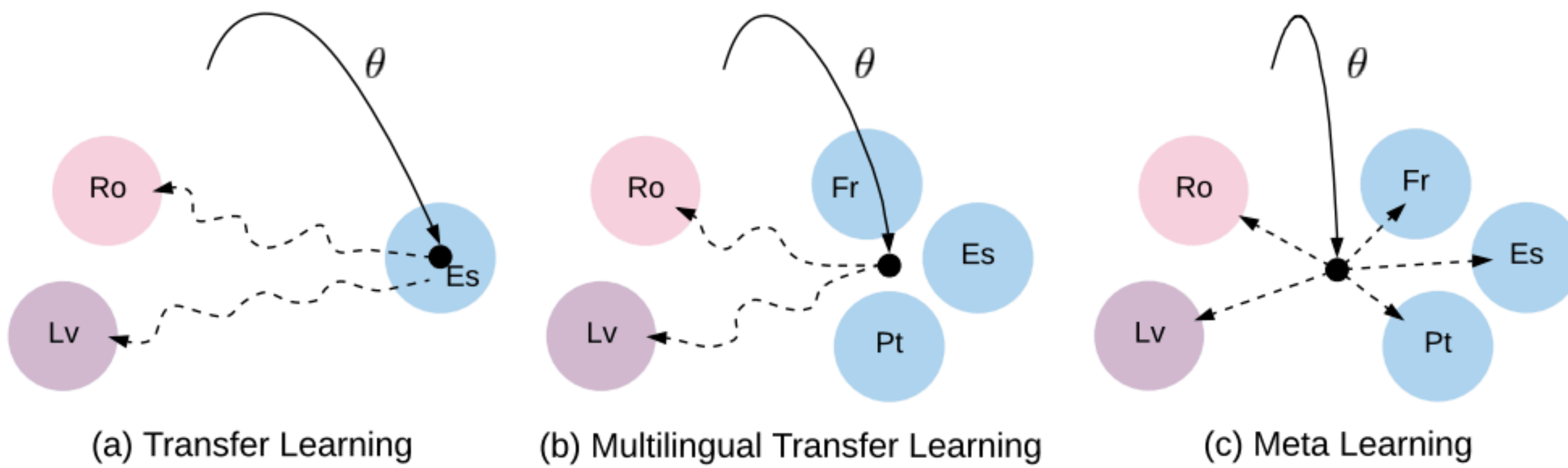
Similar Language Regularization



- 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

Meta-learning for multilingual training



- Learning a good initialization of model for fast adaptation to all languages
- Meta-learning: learn how to learn
  - Inner loop: optimize/learn for each language
  - Outter loop (meta objective): learn how to quickly optimize for each language

Meta-learning for low-resource neural machine translation. Gu et. al. 2018

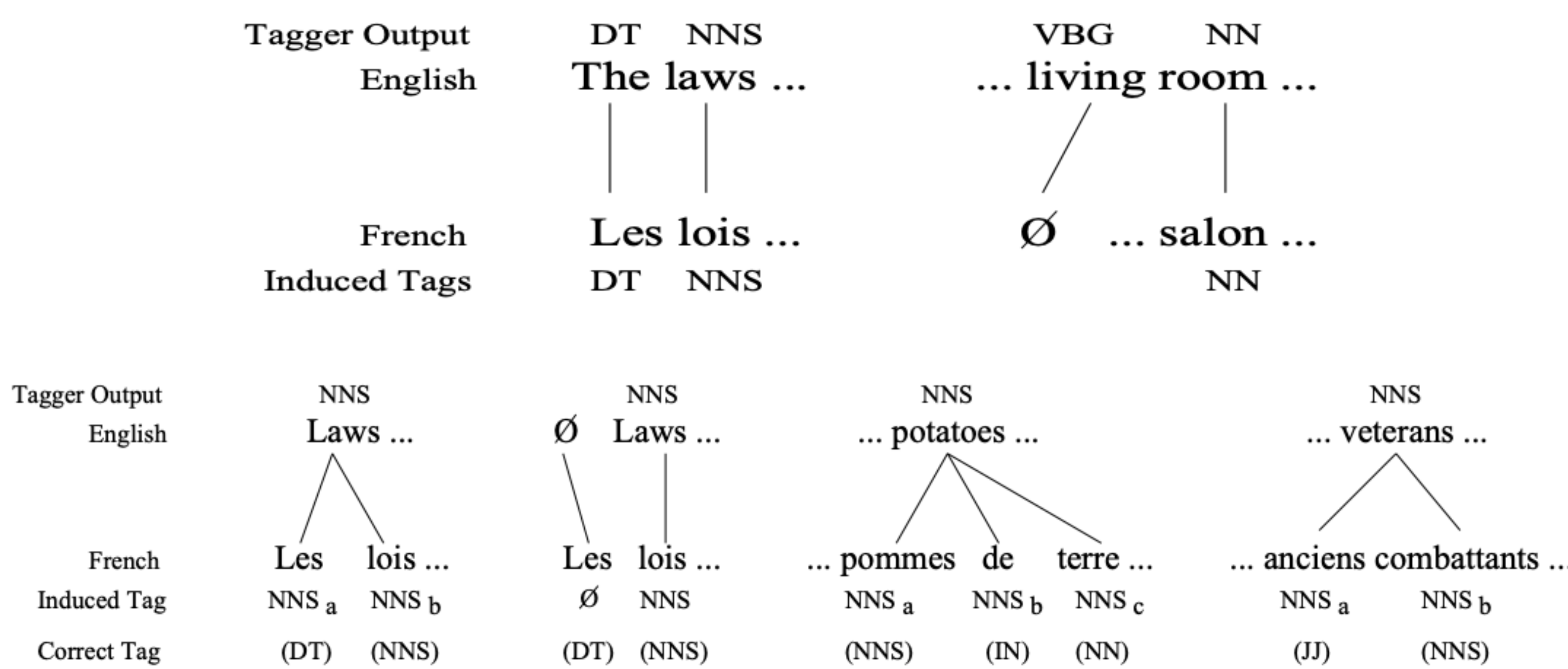
Zero-shot transfer for pretrained representations

- Pretrain: large language model using **monolingual data** from many different langauges
- Fine-tune: using **annotated data** in a given language (eg. English)
- Test: test the fine-tuned model on a **different** language from the fine-tuned language (eg. French)
- **Multilingual pretraining** learns a language-universal representation!

How multilingual is multilingual BERT? Pires et. al. 2019

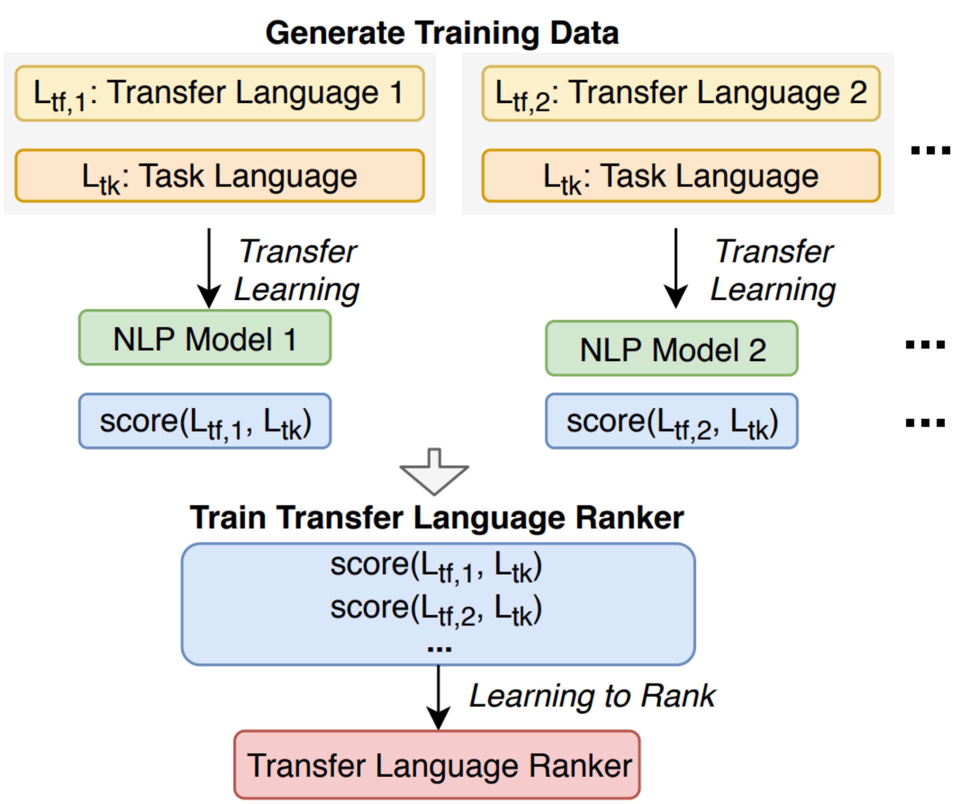
Annotation Projection

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



Which Language to Use?

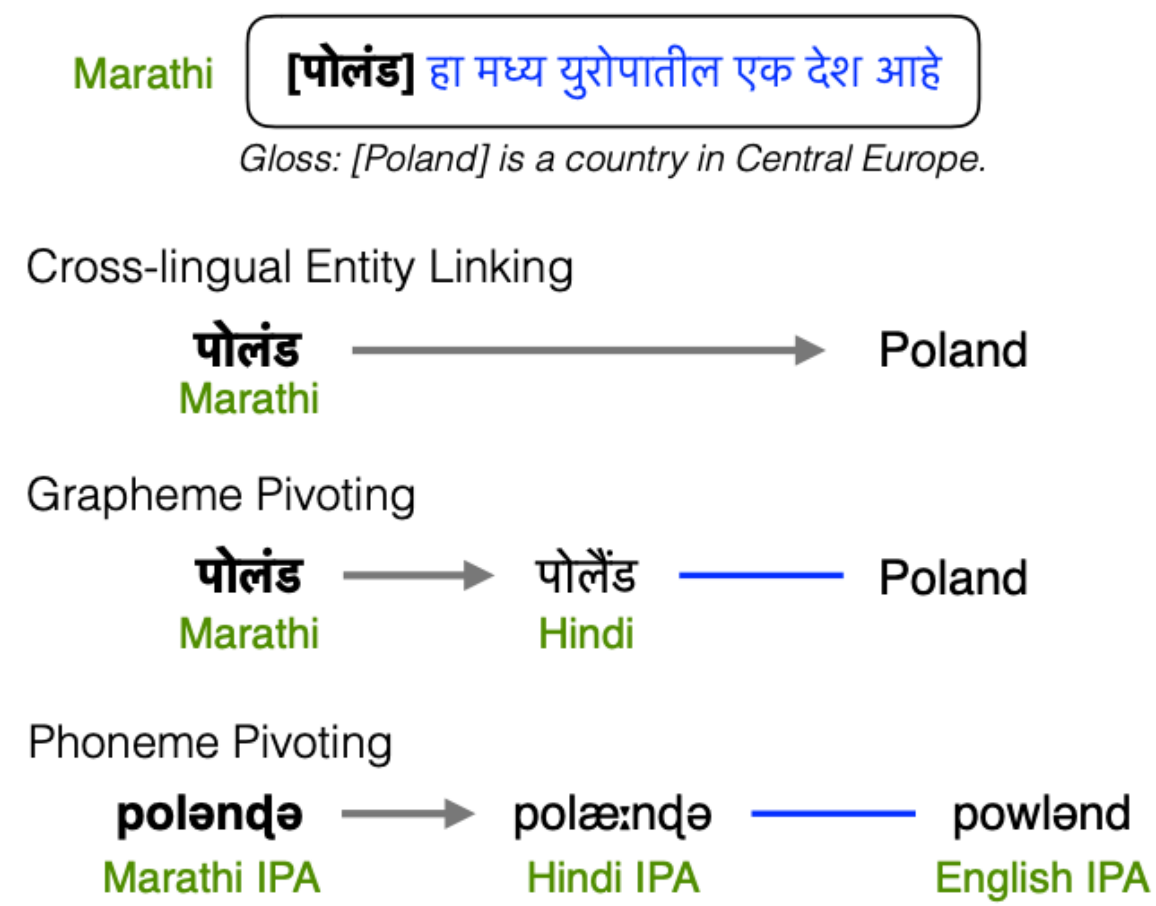
- When transferring from another language, it is ideal that it is
  - **Similar** to the target language
  - **Data-rich**
- Lin et al. (2019) examine how to identify better transfer languages



	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 the same 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.

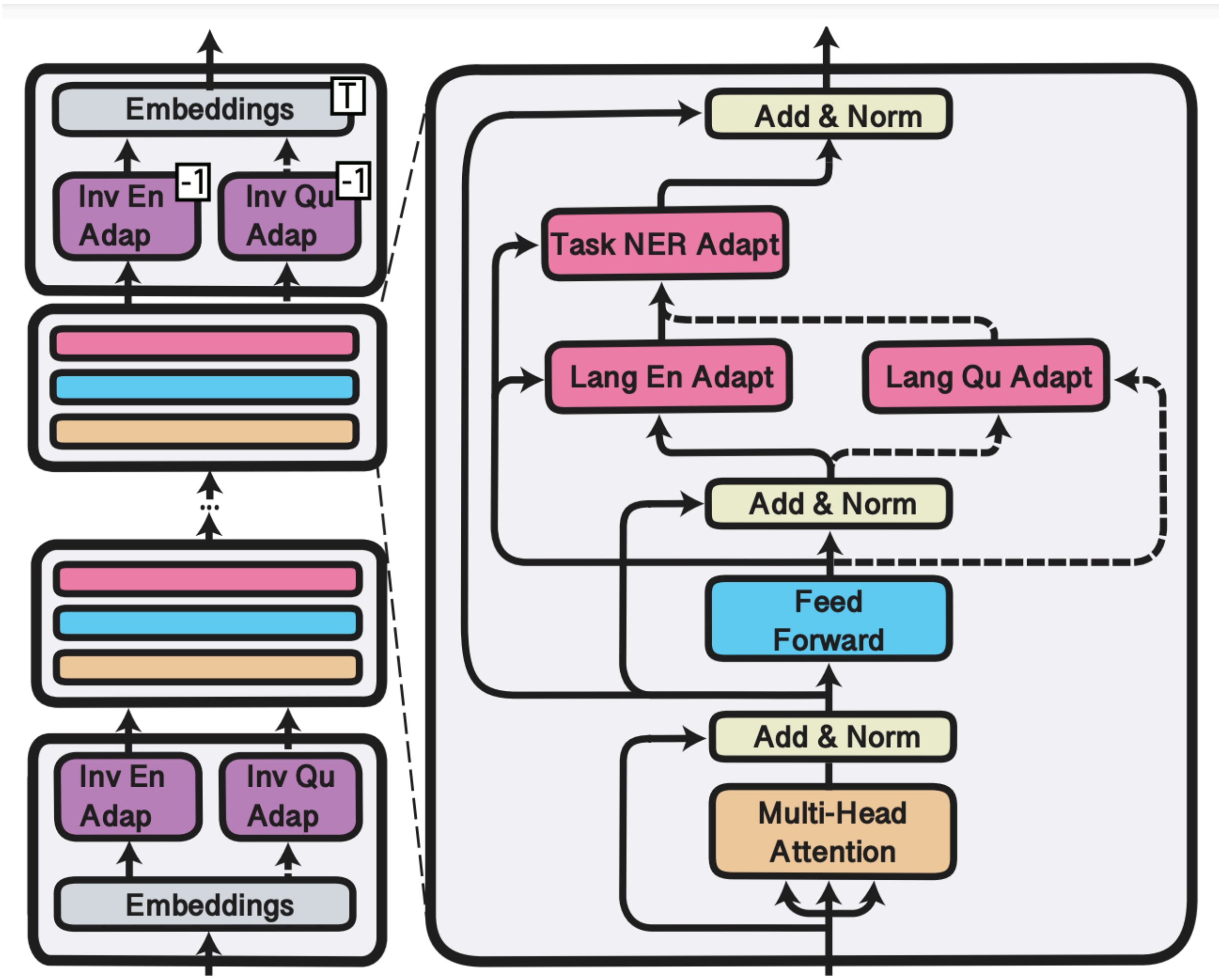


How to Share Parameters?

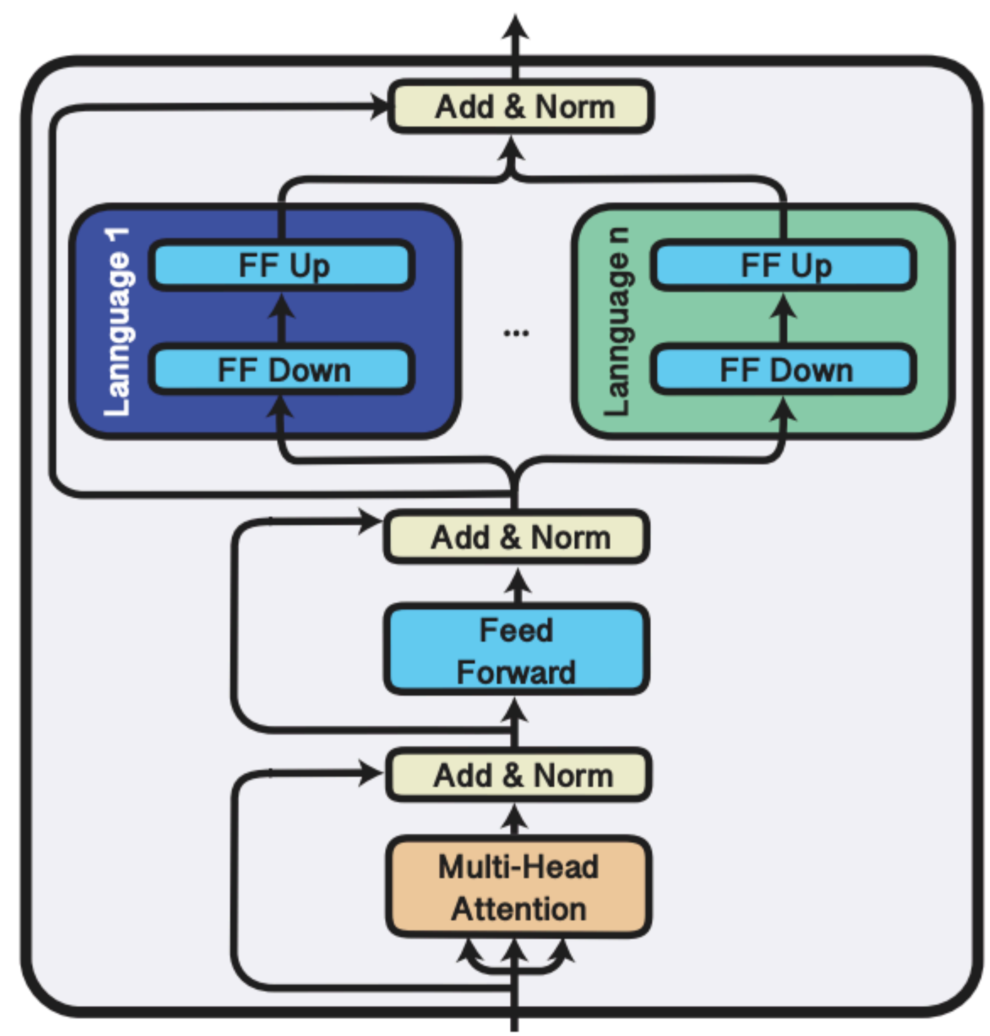
- Share all parameters (e.g. Johnson et al. 2016)
- Share only the encoder or or attention mechanism (Dong et al. 2015, Firat et al. 2016)
- Share some matrices of the Transformer model (Sachan and Neubig 2018)
- Use a parameter generator to generate parameters per language (Platonios et al. 2018)

Language Experts

- Apply language experts post-hoc for task/ language adaptation (e.g. Pfeiffer et al. 2020)

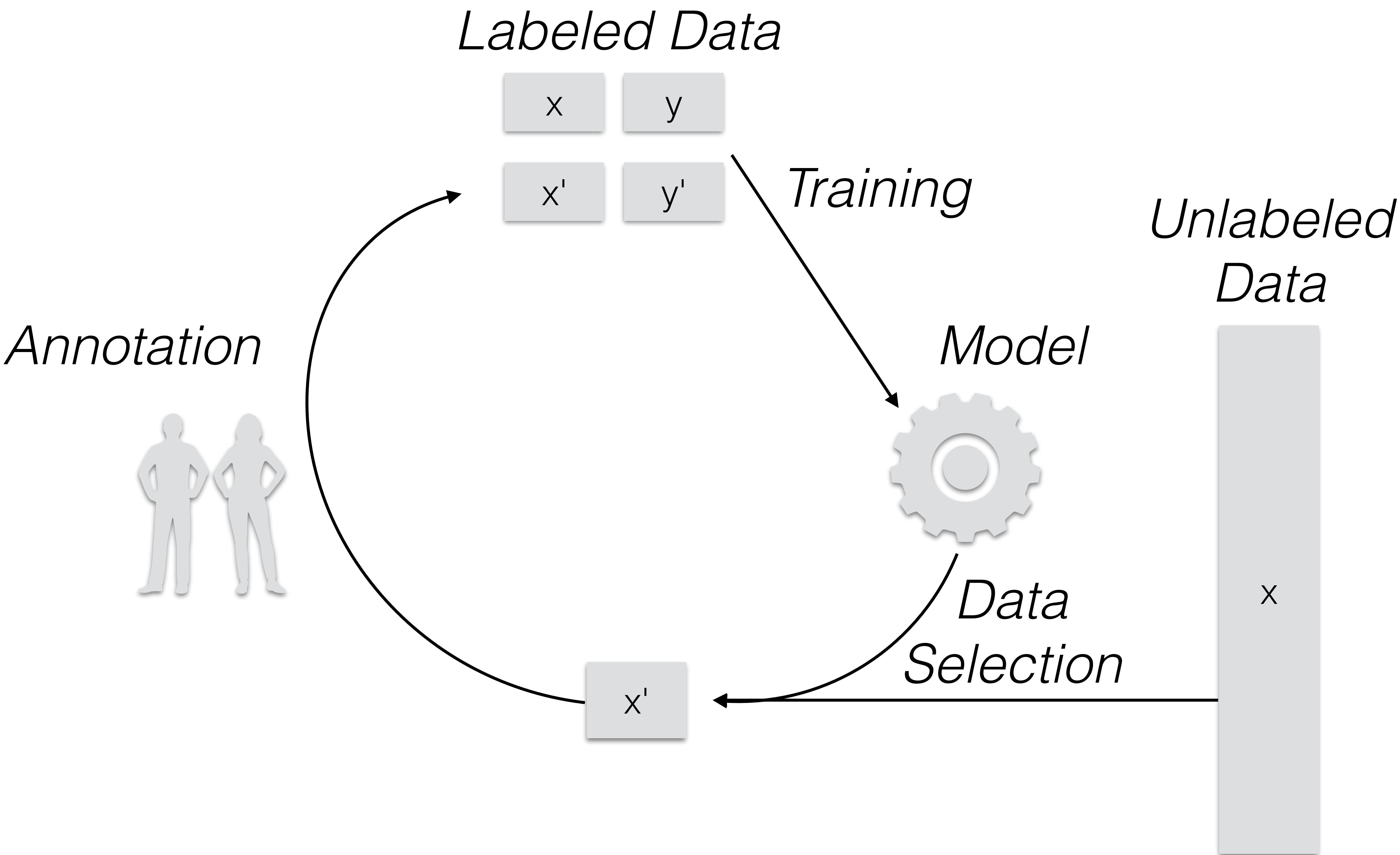


- Or pre-train with language experts (e.g. Pfeiffer et al. 2022)

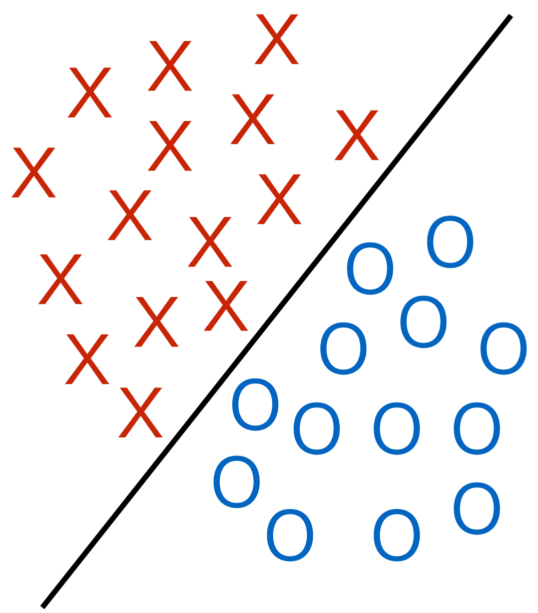


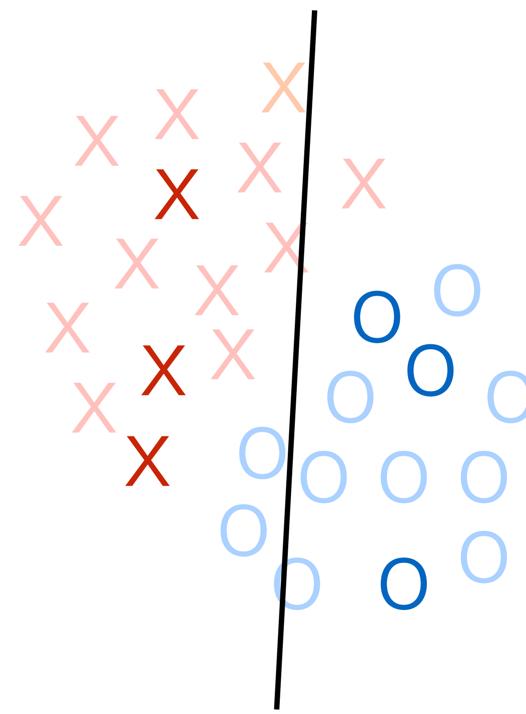
Creating New Data

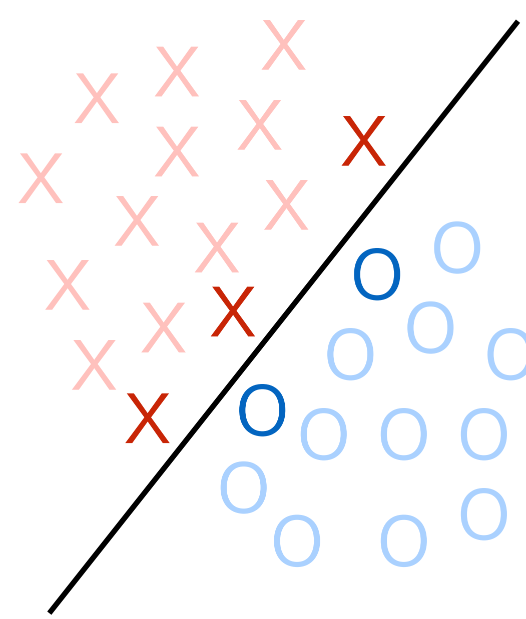
Active Learning Pipeline



Why Active Learning?







Fundamental Ideas

- **Uncertainty:** we want data that are *hard* for our current models to handle
- **Representativeness:** we want data that are *similar* to the data that we are annotating

Uncertainty Sampling Criteria

- **Entropy:** larger entropy = more uncertain

$$H(x) = -\sum_y P(y|x) \log P(y|x)$$

- **Top-1 confidence:** lower top-1 confidence = more uncertain

$$\hat{y} = \operatorname*{argmax}_y \log P(y|x)$$

$$\mathrm{top1}(x) = \log P(\hat{y}|x)$$

- **Margin:** smaller difference between first and second candidates = more uncertain

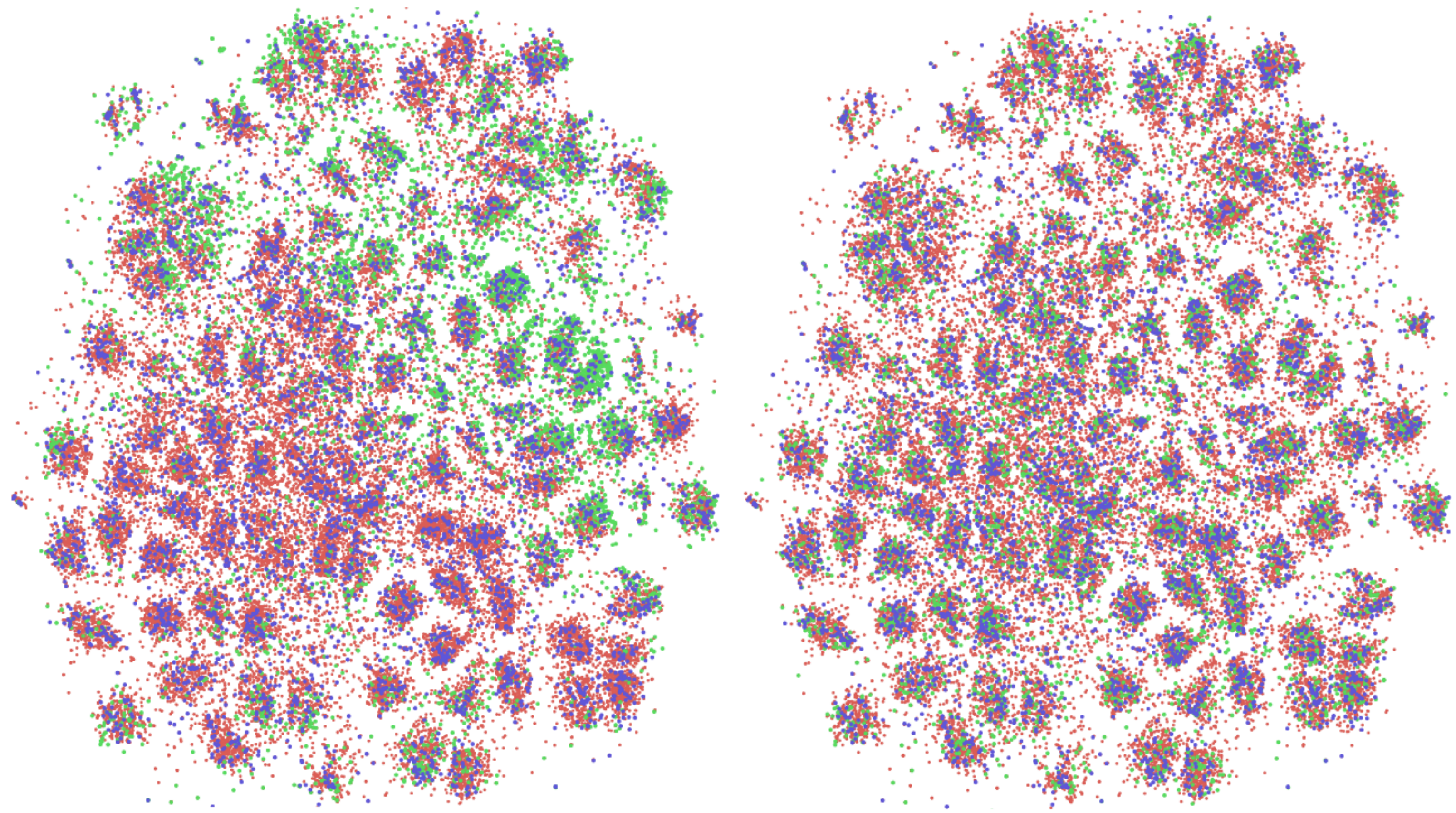
$$\mathrm{margin}(x) = \log P(\hat{y}|x) - \max_{y \neq \hat{y}} \log P(y|x)$$

Tong, Simon, and Daphne Koller. "Support vector machine active learning with applications to text classification." *Journal of machine learning research* 2.Nov (2001): 45-66.

Culotta, Aron, and Andrew McCallum. "Reducing labeling effort for structured prediction tasks." *AAAI*. Vol. 5. 2005.

Representativeness

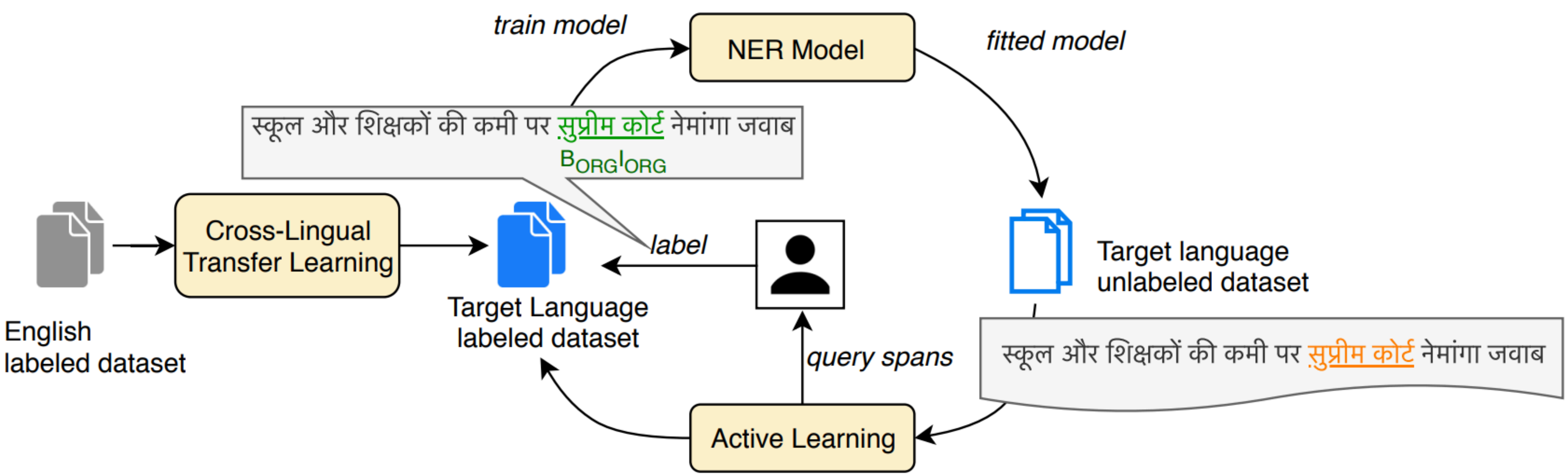
- How can we classify examples as being "similar to many others"?
- In simple feature vectors: high overlap in vector space



(a) Uncertainty Oracle (b) Our Method

Sener, Ozan, and Silvio Savarese. "Active learning for convolutional neural networks: A core-set approach." *arXiv preprint arXiv:1708.00489* (2017).

Cross-lingual Learning + Active Learning

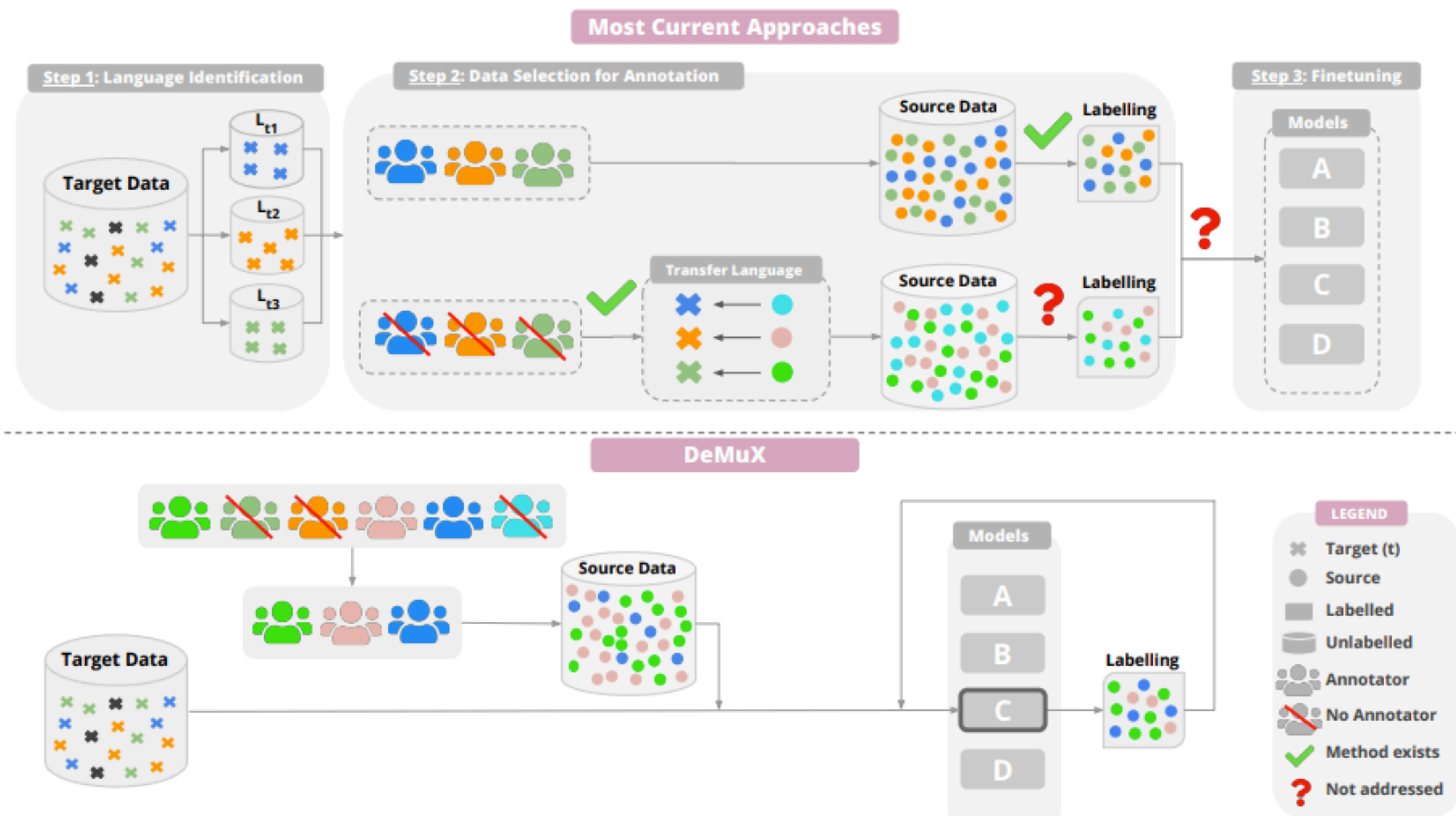


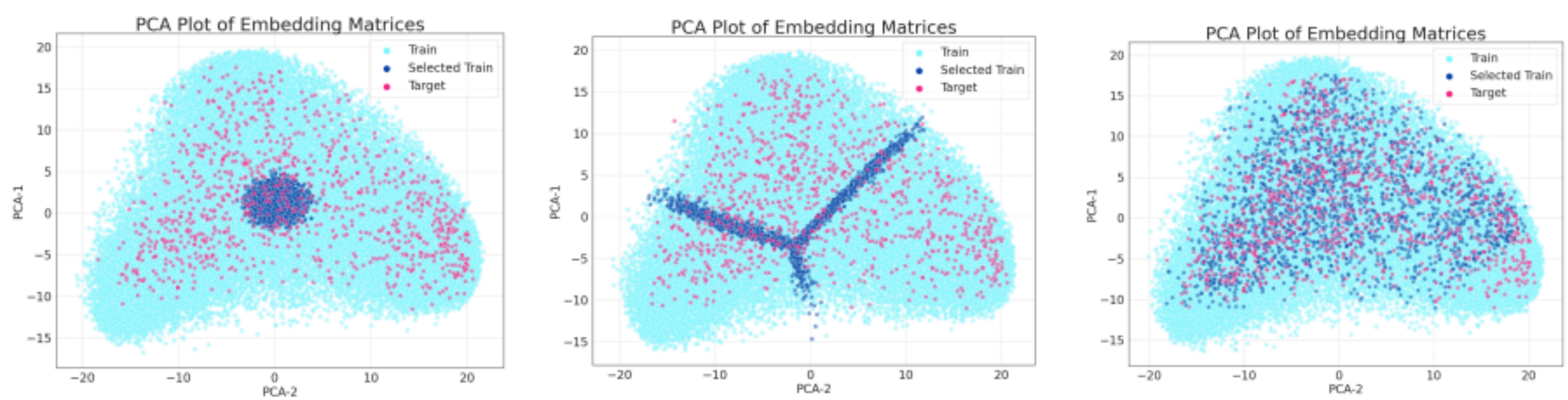
- Both perform better than either in isolation

Chaudhary, Aditi, et al. "A little annotation does a lot of good: A study in bootstrapping low-resource named entity recognizers." *EMNLP 2019*.

Active Learning for Multiple Languages

(Khanuja et al. 2023)





(a) AVERAGE-DIST (b) UNCERTAINTY (c) KNN-UNCERTAINTY

Questions?