Topics in NLP (CS6803)

(Not) lost in translation: Multilinguality

Dr. Maunendra Sankar Desarkar

Disclaimer

- The contents of these slides are mostly adapted from existing slides on similar content by several researchers/instructors.

Microsoft

ACL 2023 Tutorial

Everything you need to know about Multilingual LLMs:

Towards fair, performant and reliable models for languages of the world

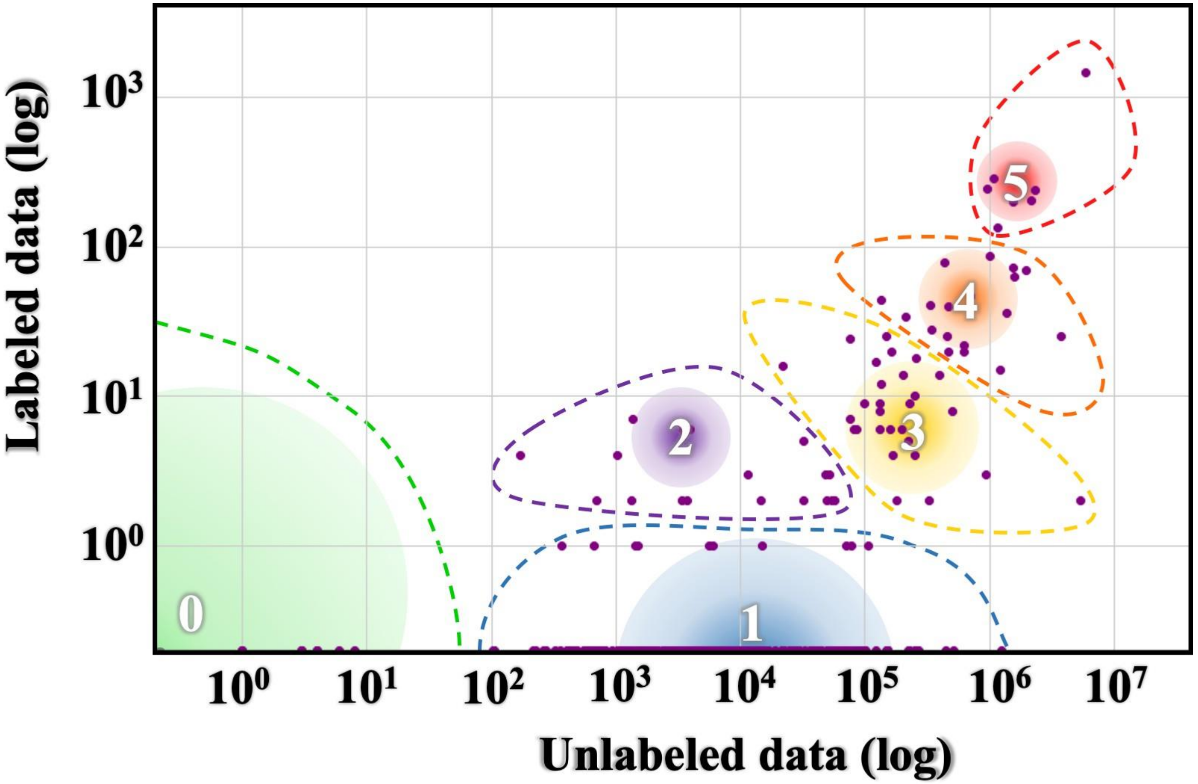
Barun Patra
Vishrav Chaudhary
Kabir Ahuja
Kalika Bali
Monojit Choudhury
Sunayana Sitaram

Microsoft Corporation

Introduction

1

How well have Language Technologies been serving the 6000+ languages of the planet?



Hierarchy of languages in terms of available resources for training NLP systems

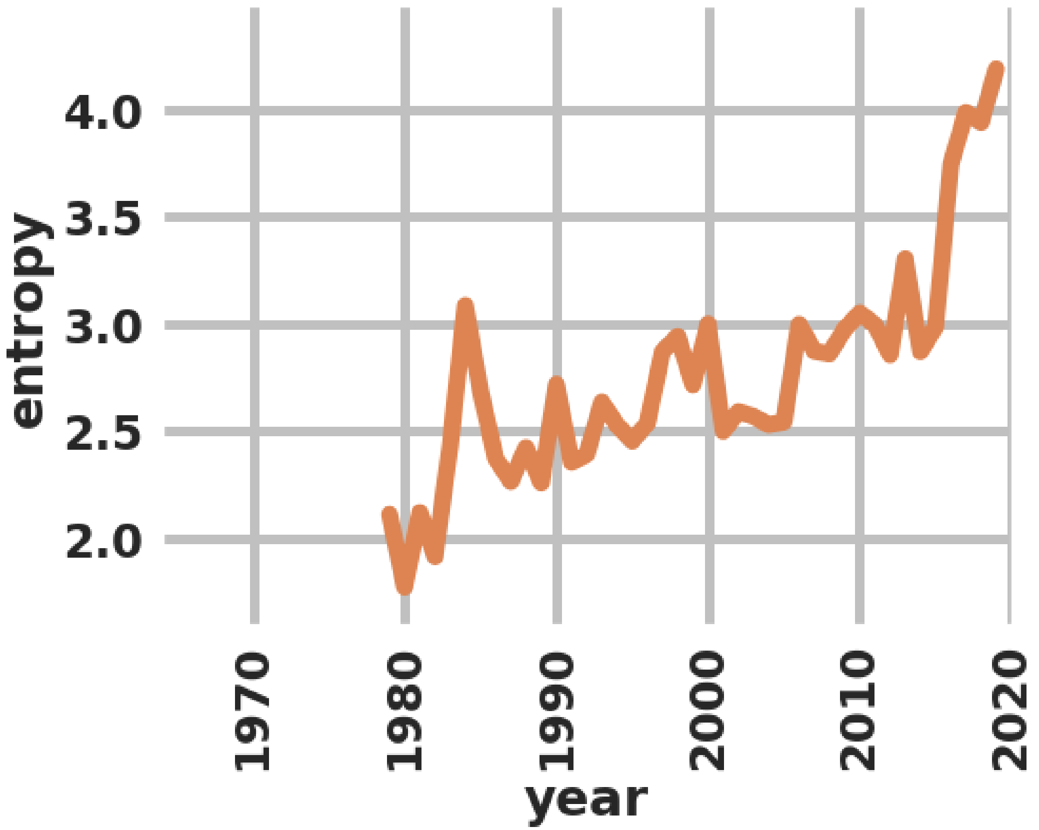
[Joshi et al. ACL 2020]

88% of the world’s languages, spoken by 1.2B people are untouched by the benefits of language technology.

Class	5 Example Languages	#Langs	#Speakers	% of Total Langs
0	Dahalo, Warlpiri, Popoloca, Wallisian, Bora	2191	1.2B	88.38%
1	Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo	222	30M	5.49%
2	Zulu, Konkani, Lao, Maltese, Irish	19	5.7M	0.36%
3	Indonesian, Ukranian, Cebuano, Afrikaans, Hebrew	28	1.8B	4.42%
4	Russian, Hungarian, Vietnamese, Dutch, Korean	18	2.2B	1.07%
5	English, Spanish, German, Japanese, French	7	2.5B	0.28%

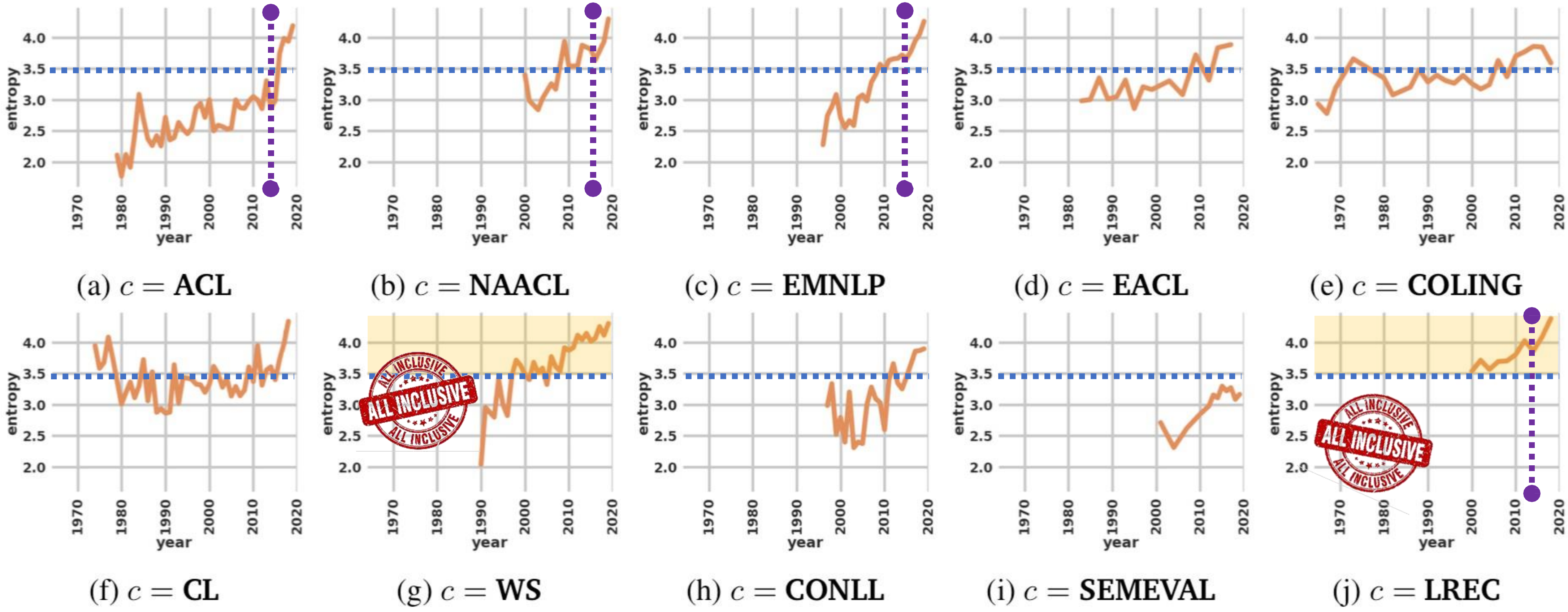
2

Are our technologies progressively getting more *linguistically inclusive and diverse*?



Entropy of the distribution of Language mentions in ACL papers over the years

[Joshi et al. ACL 2020]

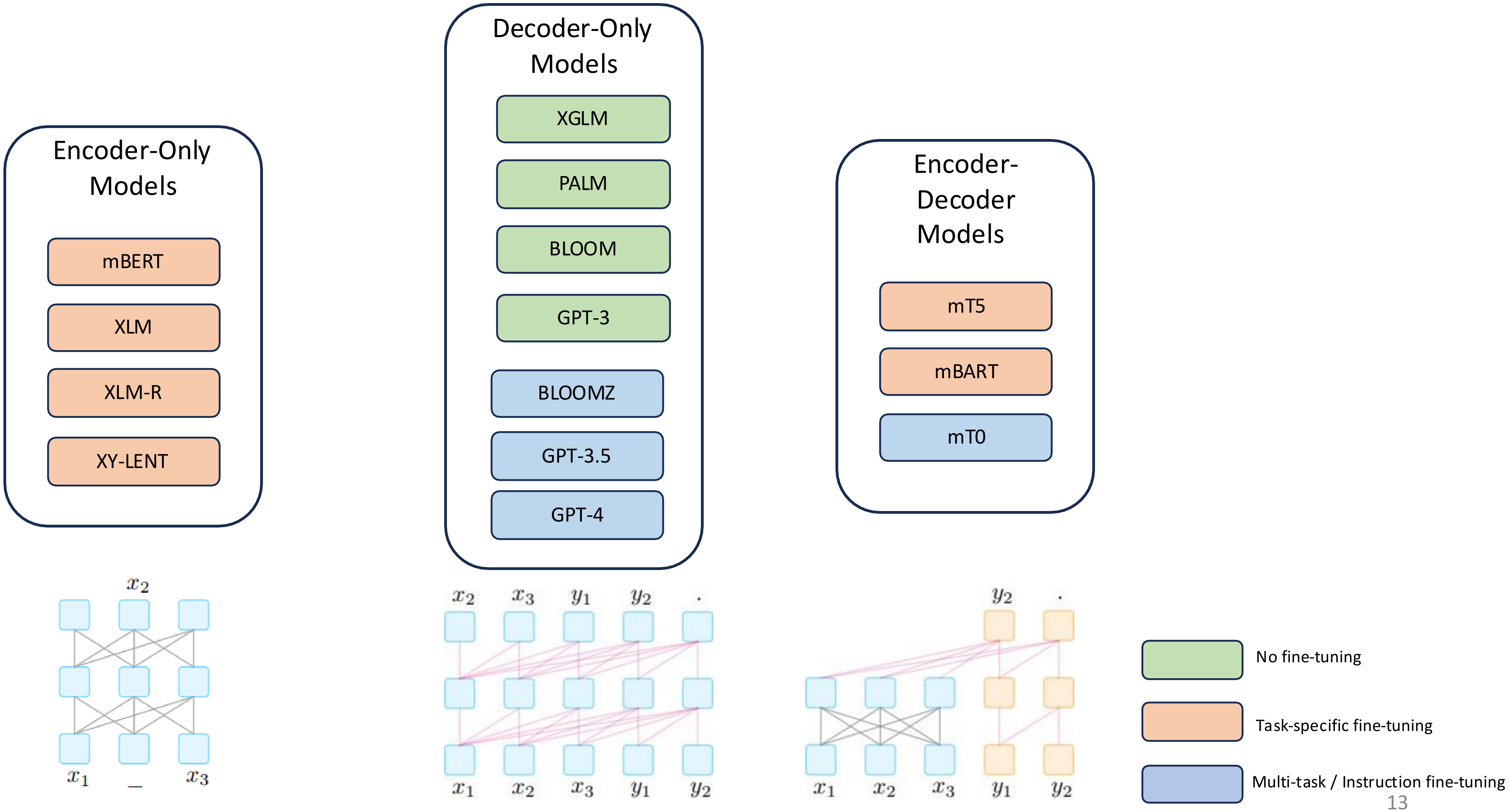


Until 2015, prestige of a conference has been inversely correlated to Linguistic D&I. Things are getting better recently.

[Doddapaneni et al. 2021. A Primer on Pretrained Multilingual Language Models](2107.00676.pdf (arxiv.org))
2107.00676.pdf (arxiv.org)

Model	Architecture				pretraining				Languages	
	N	k	d	$\#Params.$	*Objective Function*	*Mono.*	*Parallel*	*Task specific data*	*#langs.*	*vocab.*
IndicBERT (Kakwani et al., 2020)	12	12	768	33M	MLM	IndicCorp	✗	✗	12	200K
Unicoder (Huang et al., 2019)	12	16	1024	250M	MLM, TLM, CLWR, CLPC, CLMLM	Wikipedia	✓	✗	15	95K
XLM-15 (Conneau and Lample, 2019)	12	8	1024	250M	MLM, TLM	Wikipedia	✓	✗	15	95K
XLM-17 (Conneau and Lample, 2019)	16	16	1280	570M	MLM, TLM	Wikipedia	✓	✗	17	200K
MuRIL (Khanuja et al., 2021)	12	12	768	236M	MLM, TLM	CommonCrawl + Wikipedia	✓	✗	17	197K
VECO-small (Luo et al., 2021)	6	12	768	247M	MLM, CS-MLM†	CommonCrawl	✓	✗	50	250K
VECO-Large (Luo et al., 2021)	24	16	1024	662M	MLM, CS-MLM	CommonCrawl	✓	✗	50	250K
InfoXLM-base (Chi et al., 2021a)	12	12	768	270M	MLM, TLM, XLCO	CommonCrawl	✓	✗	94	250K
InfoXLM-Large (Chi et al., 2021a)	24	16	1024	559M	MLM, TLM, XLCO	CommonCrawl	✓	✗	94	250K
XLM-100 (Conneau and Lample, 2019)	16	16	1280	570M	MLM, TLM	Wikipedia	✗	✗	100	200K
XLM-R-base (Conneau et al., 2020a)	12	12	768	270M	MLM	CommonCrawl	✗	✗	100	250K
XLM-R-Large (Conneau et al., 2020a)	24	16	1024	559M	MLM	CommonCrawl	✗	✗	100	250K
X-STILTS (Phang et al., 2020)	24	16	1024	559M	MLM	CommonCrawl	✗	✓	100	250K
HiCTL-base (Wei et al., 2021)	12	12	768	270M	MLM, TLM, HICTL	CommonCrawl	✓	✗	100	250K
HiCTL-Large (Wei et al., 2021)	24	16	1024	559M	MLM, TLM, HICTL	CommonCrawl	✓	✗	100	250K
Ernie-M-base (Ouyang et al., 2021)	12	12	768	270M	MLM, TLM, CAMLM, BTMLM	CommonCrawl	✓	✗	100	250K
Ernie-M-Large (Ouyang et al., 2021)	24	16	1024	559M	MLM, TLM, CAMLM, BTMLM	CommonCrawl	✓	✗	100	250K
mBERT (Devlin et al., 2019)	12	12	768	172M	MLM	Wikipedia	✗	✗	104	110K
Amber (Hu et al., 2021)	12	12	768	172M	MLM, TLM, CLWA, CLSA	Wikipedia	✓	✗	104	120K
RemBERT (Chung et al., 2021a)	32	18	1152	, 559M‡	MLM	CommonCrawl + Wikipedia	✗	✗	110	250K

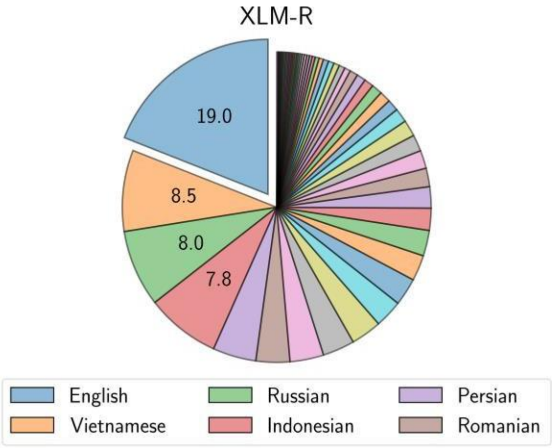
Multilingual Language Models

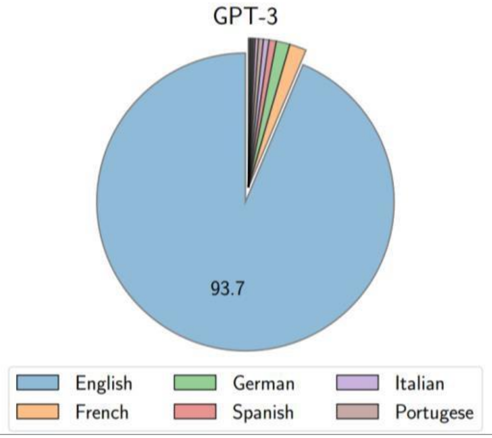


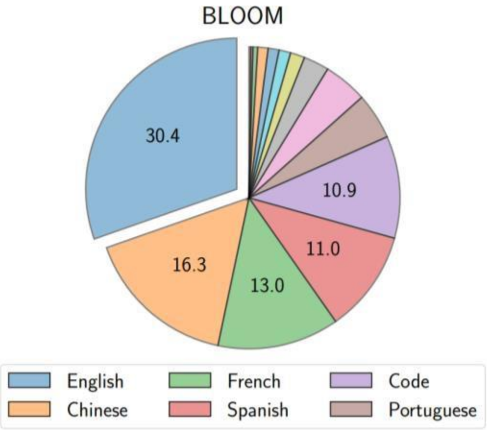
Figures from Liu et al. 2021

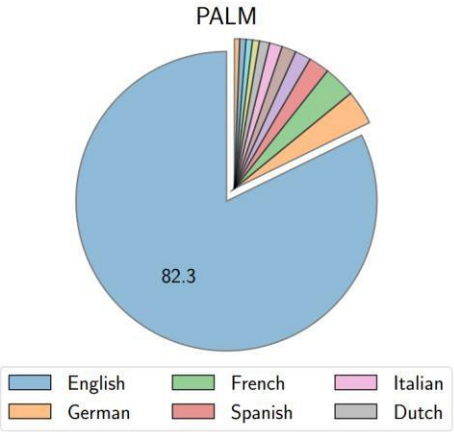
Linguistic Coverage of Different Models

- Pre-training Data of different models is predominantly English!
- However, even small percentages of non-English data can facilitate cross lingual transfer. Blevins et al. 2022 [2204.08110] Language Contamination Helps Explain the Cross-lingual Capabilities of English Pretrained Models (arxiv.org)









Microsoft

Data Collection and Training of Multilingual LLMs

Barun Patra and Vishrav Chaudhary

Data is a key component for training better performing Language Models in the Multilingual domain.

- A Multilingual LLM can enable and even revolutionize several downstream scenarios for many languages at once

- Also aid in bridging the gap between societies and pushing the frontier for technological advancements

Data is a key component for training better performing Language Models in the Multilingual domain.

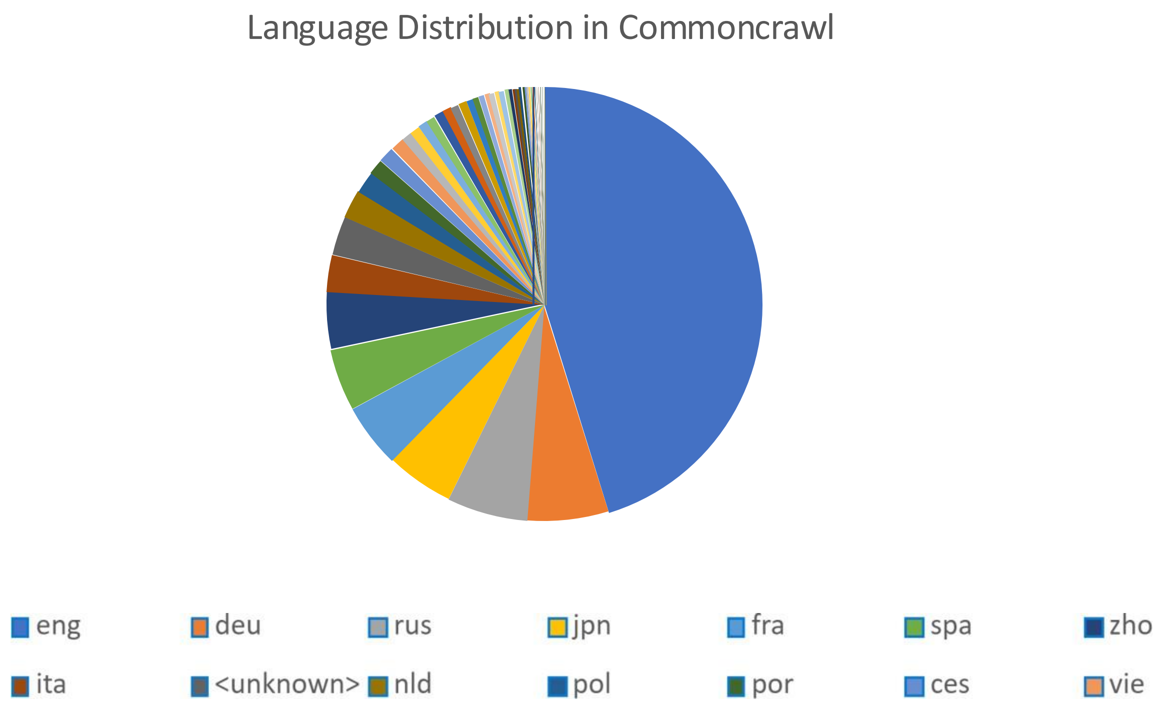
- A Multilingual LLM can enable and even revolutionize several downstream scenarios for many languages at once
- Also aid in bridging the gap between societies and pushing the frontier for technological advancements

Challenges:

- *Quantity*
- *Quality*
- *Sourcing*
- *Governance*

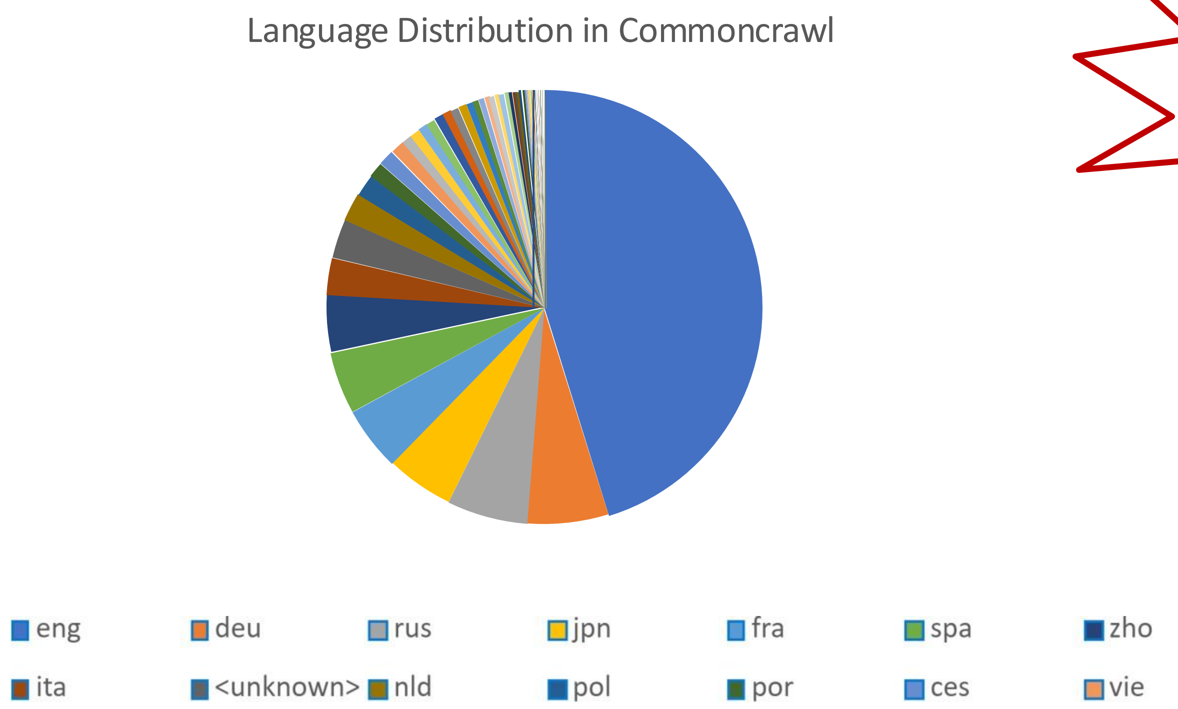
Data Collection Challenges: Quantity

- Substantial gaps in quantity across
  - Languages (commoncrawl.org)



Data Collection Challenges: Quantity

- Substantial gaps in quantity across
  - Languages (commoncrawl.org)



57 languages are < 0.001%

Data Collection Challenges: Quantity

- Substantial gaps in quantity across
  - Languages (commoncrawl.org)
  - Domains (Gao et al., 2020)

ArXiv

Conversational

Law

Medical

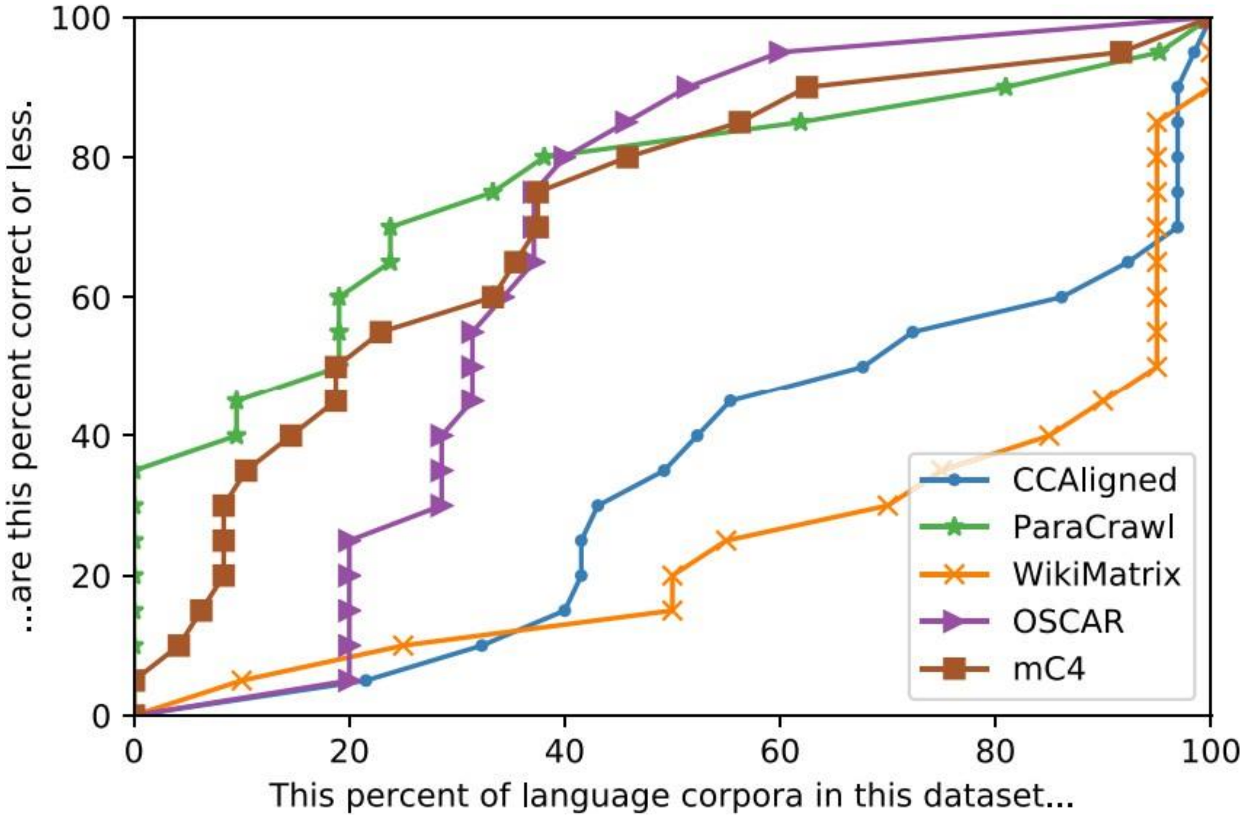
Educational

……

Data Collection Challenges: Quality

- Kreutzer et al., 2022 did a comprehensive survey covering quality issues across different datasets

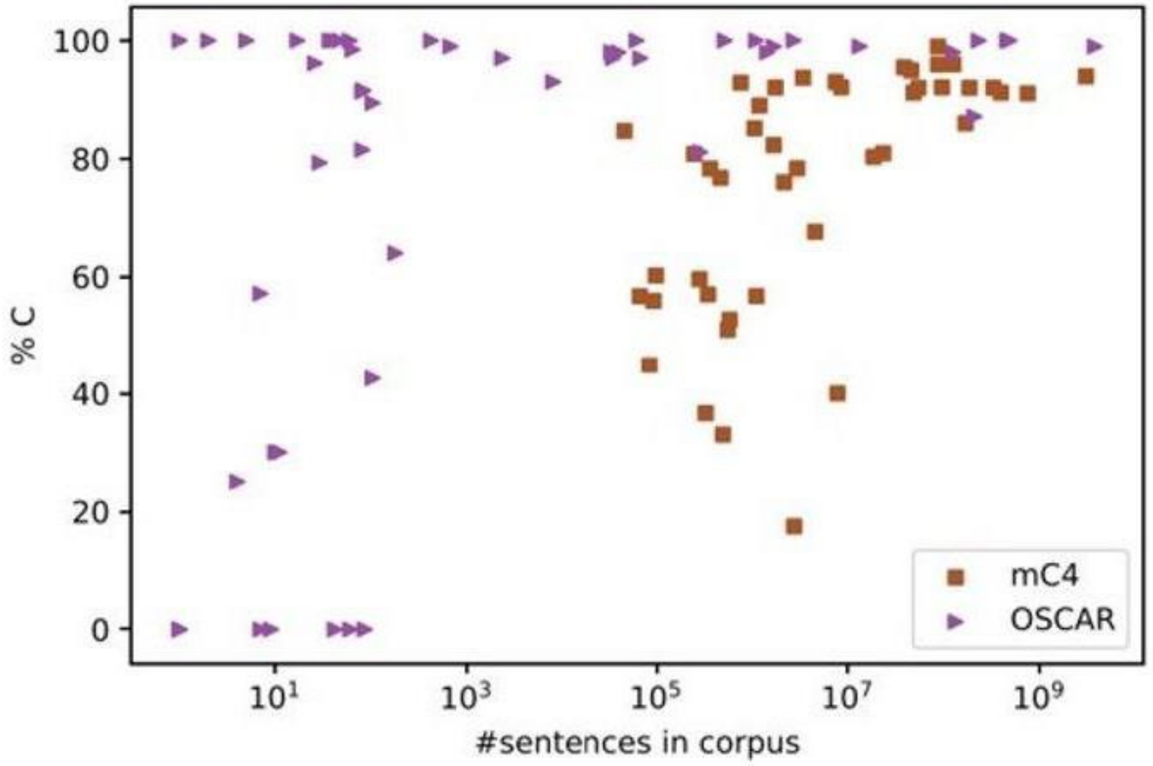
- Q1: What % of languages have good quality data?



Data Collection Challenges: Quality

- Kreutzer et al., 2022 did a comprehensive survey covering quality issues across different datasets

- Q2: Do low resource languages always have poor quality data?



Data Collection Challenges: Quality

- Reasons include
  - Incorrect Language Identification (poor quality + similar languages)
  - Machine generated data
  - Limited identification tools available for toxic/adult content

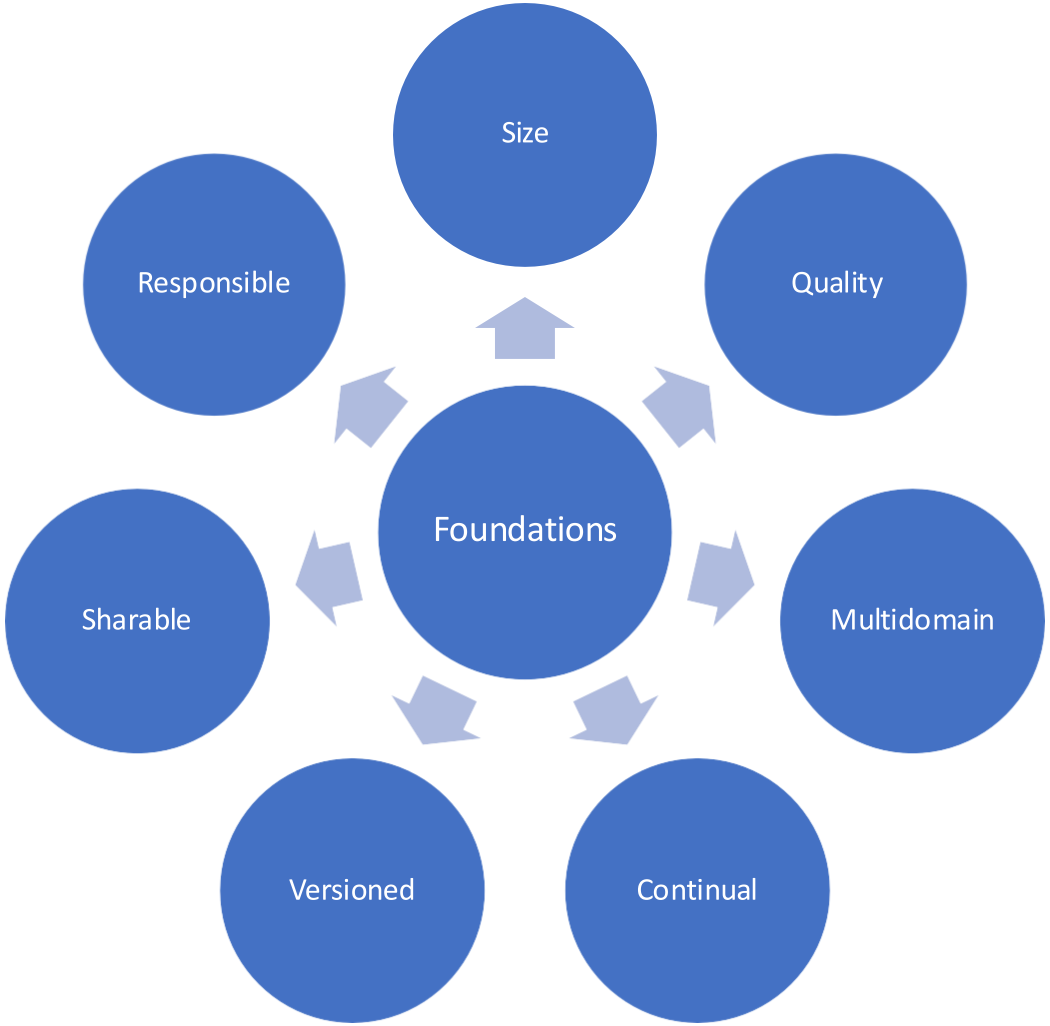
Data Collection: Sourcing & Governance

- Initiatives by government agencies

- Defining actors: data custodians, rights-holders, and other parties to appropriately govern shared data

- Designed to account for the privacy, intellectual property, and user rights of the data and algorithm subjects in a way that aims to prioritize local knowledge and expression of guiding values

Data Requirements



Data Preprocessing

Collection

- Downloading
- Text Extraction
- Simple Deduplication (URL Based)

Initial Cleaning

- Language Identification
  - Threshold based filtering
  - Multi-language documents

Deduplication

- Exact Substring based (mC4, OSCAR v*, CC100)
- Fuzzy Minhash based (GPT-3, ThePILE)
- Both (Refined Web)

Filtering

- Heuristics Based (Refined Web)
- Model Based (CC-Net, CC100)
- NSFW URL Based, PII
- Line Based, Doc based

Tokenization

- Tokenization algorithms that have a fallback to bytes (and hence produce few / no UNK tokens)
  - Most popular Sentencepiece, BPE and Wordpiece
- Larger vocabulary size usually correlated with better performance
  - At cost of training speed, inference speed and increased parameters)
- Allocating vocab capacity across different languages improves performance
  - Eg: following the VoCAP approach presented in Zheng et al. 2021
- Another alternative seems to be leveraging byte-based models
  - But seem to require deeper (encoder) models / with additional capacity (byte-T5)
  - Additionally, require models that can cover larger context windows
  - More robust to mis-spellings

Models

Wordpiece

- mBERT

Sentencepiece

- XLM-Roberta, mBART, XGLM, mT5

VoCAP

- XLM-E, XY-LENT

BPE

- GPT*, Bloom

Byte-level

- Byte-T5, Perceiver

Data Sources For Training

Monolingual Corpora

Machine learning is changing the world today with research happening at an extremely fast pace.

मशीन लर्निंग आज दुनिया को बदल रही है और अनुसंधान बहुत तेज गर्त से हो रहा है।

L'apprentissage automatique change le monde aujourd'hui avec des recherches qui se déroulent à un rythme extrêmement rapide.

기계 학습은 매우 빠른 속도로 진행되는 연구로 오늘날 세상을 변화시키고 있습니다.

Models

- mBERT, XLM-Roberta
- mT5, AlexaTM, byte-mT5

Bitext Corpora

English Centric

I love cats | J'aime les chats.

I love cats | मुझे बबल्ललयाँ पसन्द है।

I love cats | 나는 고양이를 좋아합니다.

Models

- XLM, XLM-E, DeBERTa v3, Info-XLM
- mBART
- PaLM-2

X-Y Directions

J'aime les chats. | मुझे बबल्ललयाँ पसन्द है।

나는 고양이를 좋아합니다. | I love cats

Models

- M2M 100*
- XY-LENT

General Trend of Performance Increase (within a model class type)

Sampling Techniques

Monolingual Corpora

Temperature Sampling

- P(j) = $\frac{n_j^{\alpha}}{\sigma\ n_k^{\alpha}}$, where n_j is the number of samples for j^th language
- Upsamples low resource langauges, downsamples low resource languages

Unimax

- Allocate budget as uniformly as possible
- Start with lowest resource language, and keep adding, allocating uniform budget
- *Better performance compared to Temperature Sampling*

Bitext Corpora

English Centric

Temperature Sampling

- Here, the normalization is over non-English languages

X-Y Directions

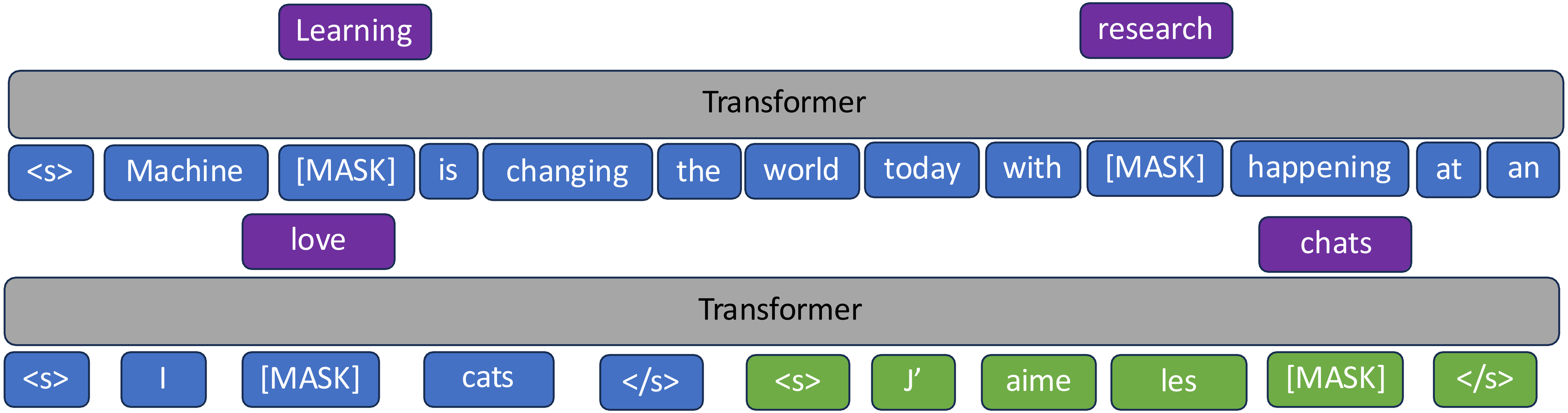
Temperature Sampling

- P(i, j) = $\frac{n_{i,j}^{\alpha}}{\sigma\ n_{k,l}^{\alpha}}$, where $n_{i,j}$ is the number of samples for i-j^th language pair

Approximating English Centric marginal distributions

- P(i, j) such that $\forall\ j$ P(j) = $\sigma_i\ P(i,j)$ is similar to English Centric distributions

Encoder Models: Cloze Infilling



- BERT style models
- X% of tokens are masked, and model uses left and right context to predict the middle token
- Can use both monolingual and bitext data

Models

- mBERT
- XLM
- XLM-Roberta

Encoder Models: Electra Models

- Electra style training paradigm
    - Predicting which tokens come from generator vs which come from data
    - But unlike a GAN, generator trained on MLM task
- More sample efficient
- In general better performance
- Variants to stop gradient flow between generator and discriminator embeddings
- Different layer-wise behavior compared to MLM
    - Higher layers better at semantic retrieval tasks

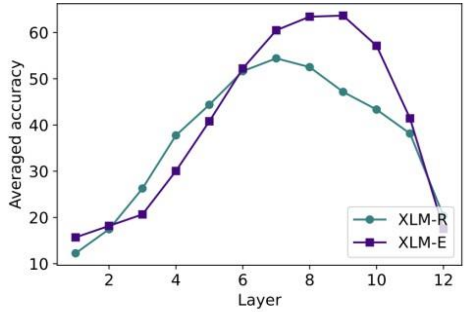
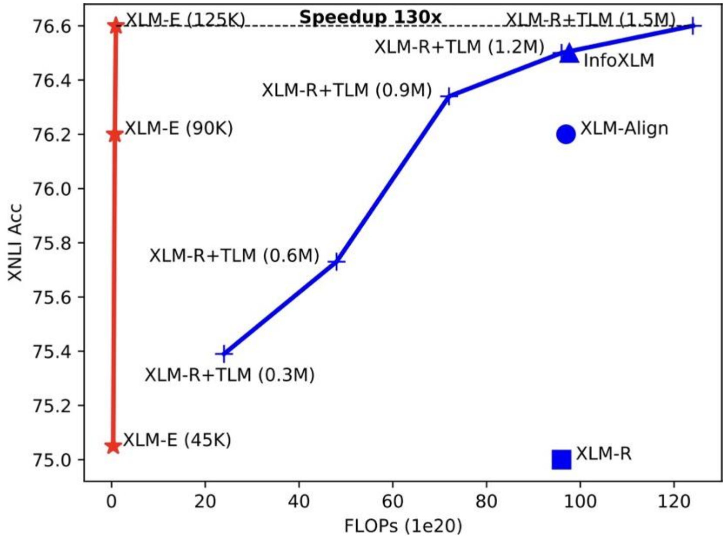
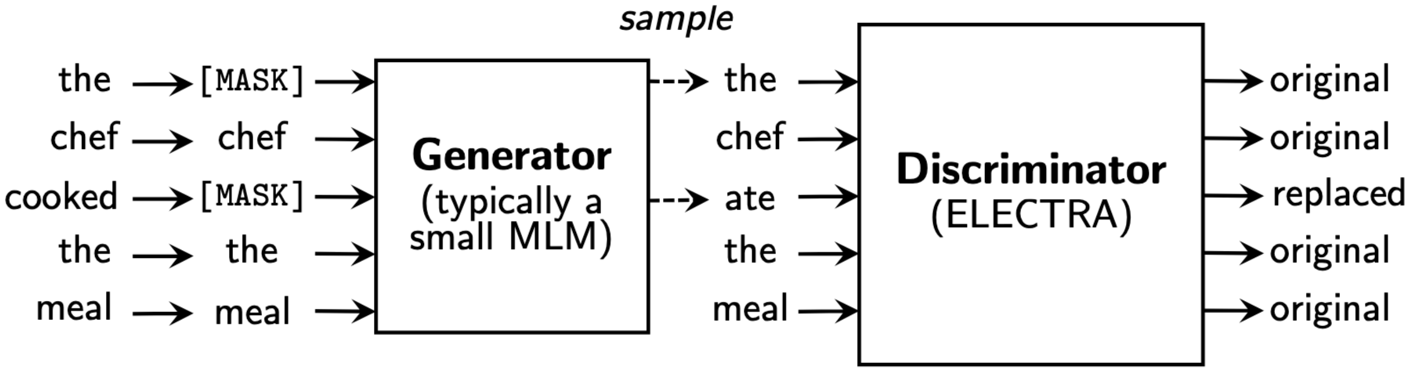


Figure 3: Evaluation results on Tatoeba cross-lingual sentence retrieval over different layers. For each layer, the accuracy score is averaged over all the 36 language pairs in both the xx → en and en → xx directions.





Models

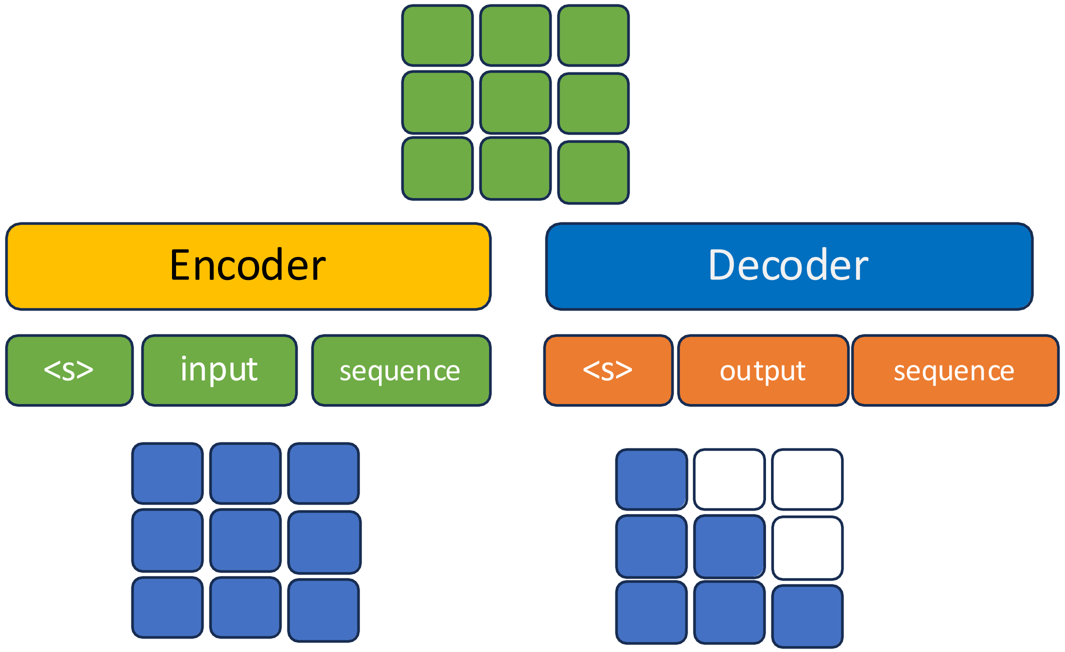
- XLM-E, XY-LENT, DEBERTAv3

ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. ICLR 2020

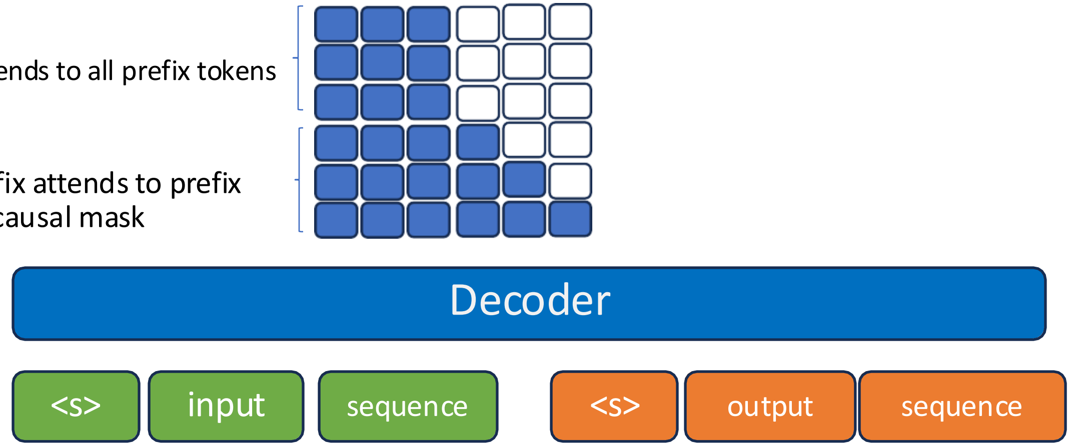
**Figures taken from Chi et al 2022*

Encoder Decoder Models

- Standard Transformer Architecture
- Two transformers one for encoder, one for decoder
- Can repurpose a decoder with prefix LM for similar purpose



Traditional Encoder Decoder



Prefix LM structure

Models

- mT5, byteT5
- mBART
- AlexaTM

Encoder Decoder Denoising Objectives

- **Token Masking:** Masking certain fraction of tokens (similar to BERT), but get the model to generate the tokens

Machine Learning is <X> the <Y> today

<S><X> changing <Y> world </s>

mT6, byteT5: using sentinel tokens for indicating what tokens / bytes to mask and get decoder to generate generate

Machine Learning is [MASK] the [MASK] today

<S>Machine Learning is changing the world today </s>

mBART: reconstructing the entire sentence, AlexaTM: no use of MASK tokens, still reconstruct entire sequence

- **Sentence Masking / Denoising:** Mask out continuation of a document, getting model to generate the continuation

[S] L'apprentissage automatique <X>

change le monde aujourd'hui

UL2, UL2R, AlexaTM: Get model to complete generation. Note the usage of prefix tokens to denote type of noise

Encoder Decoder Denoising Objectives

- **Extreme corruption:** Mask out large parts of the document, getting the model to generate them

[X] El aprendizaje <X> está <X> el <Y>

<B> automático <S> cambiando <S> mundo <E>

UL2, UL2R: Try and recover a severely noised document, using multiple sentinels

- **Combinations:** Combine different noising strategies together (using sentinel tokens to denote different masking strategies)

[R] Machine Learning is <X> the <Y> today

<S><X> changing <Y> world </s>

[S] L'apprentissage automatique <X>

change le monde aujourd'hui

[X] El aprendizaje <X> está <X> el <Y>

<B> automático <S> cambiando <S> mundo <E>

UL2 / UL2R / PaLM2:

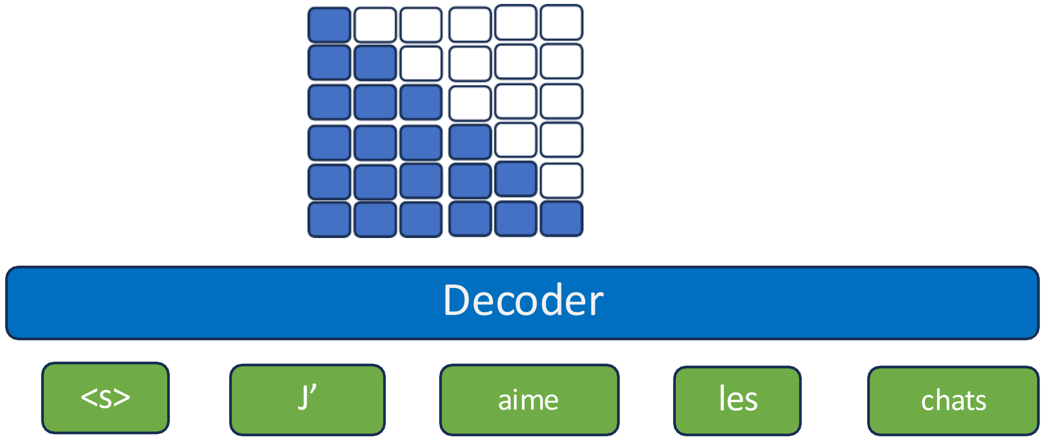
Also possible as post training step, to boost a general purpose decoder's abilities

Note the different prefix tokens to tell the model what mode to generate in

Causal Decoder Models

- Standard autoregressive decoding
- Shown in (Wang et al 2022) to have best performance for direct zero-shot adaptation
    - In contrast, encoder decoder models tend to perform better after fine-tuning on instruction datasets
- The authors recommend training decoder models followed by non decoder training followed by instruction tuning
    - Improvement using non decoder continued training also shown in (Tay et. al 2022)
    - Improvement of instruction tuning over such a model also corroborated by (Chung et al 2022)

- *Note: The previous observations are for English centric models.*
  PALM-2 report impressive multilingual performance following a similar recipe, so might be applicable for multilingual scenarios too.



Models

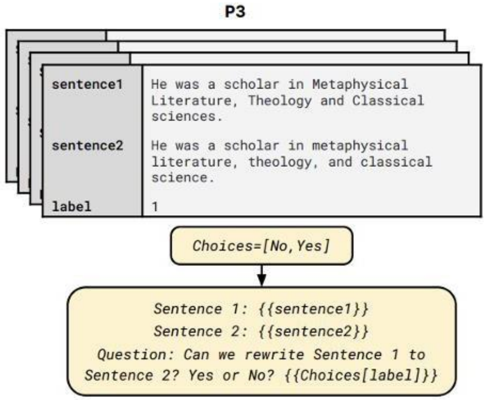
- XGLM
- Bloom

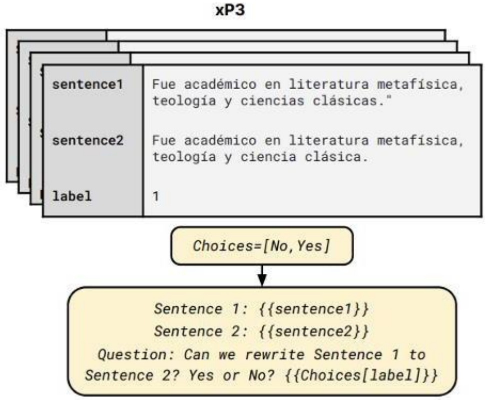
Continued Training with non decoder objectives

- UL2R

Post-Training: Instruction Finetuning

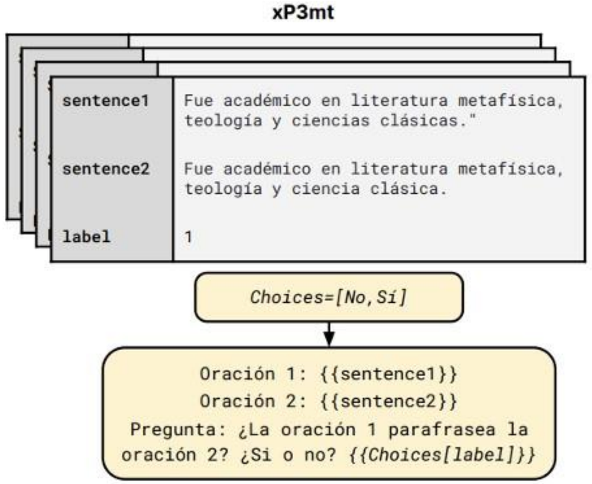
- Post training carried out on instructions dataset
- Multilingual LLM trained on
    - English only instructions (P3 dataset)
    - Multilingual datasets (but with English Prompts xP3)
    - Multilingual datasets (with prompts translated to target language xP3mt)
- Seems to improve both English and multilingual performance
- When prompts are multilingual, there seems to be a tradeoff between English and multilingual performance





Models

- BloomZ
- mT0



Figures taken from Muennighoff et al 2023

References

1. Fan, Angela, et al. "Beyond English-Centric Multilingual Machine Translation. arXiv e-prints, page." *arXiv preprint arXiv:2010.11125* (2020).
2. Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).
3. Conneau, Alexis, et al. "Unsupervised cross-lingual representation learning at scale." *arXiv preprint arXiv:1911.02116* (2019).
4. Conneau, Alexis, et al. "Unsupervised cross-lingual representation learning at scale." *arXiv preprint arXiv:1911.02116* (2019).
5. Xue, Linting, et al. "mT5: A massively multilingual pre-trained text-to-text transformer." *arXiv preprint arXiv:2010.11934* (2020).
6. Chi, Zewen, et al. "Xlm-e: Cross-lingual language model pre-training via electra." *arXiv preprint arXiv:2106.16138* (2021).
7. Liu, Yinhan, et al. "Multilingual denoising pre-training for neural machine translation." *Transactions of the Association for Computational Linguistics* 8 (2020): 726-742.
8. Patra, Barun, et al. "Beyond english-centric bitexts for better multilingual language representation learning." *arXiv preprint arXiv:2210.14867* (2022).
9. Chung, Hyung Won, et al. "Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining." *arXiv preprint arXiv:2304.09151* (2023).
10. He, Pengcheng, Jianfeng Gao, and Weizhu Chen. "Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing." *arXiv preprint arXiv:2111.09543* (2021).
11. Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." *International conference on machine learning*. PMLR, 2020.
12. He, Kaiming, et al. "Momentum contrast for unsupervised visual representation learning." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.
13. Chi, Zewen, et al. "InfoXLM: An information-theoretic framework for cross-lingual language model pre-training." *arXiv preprint arXiv:2007.07834* (2020).
14. Xue, Linting, et al. "mT5: A massively multilingual pre-trained text-to-text transformer." arXiv preprint arXiv:2010.11934 (2020).
15. Xue, Linting, et al. "Byt5: Towards a token-free future with pre-trained byte-to-byte models." *Transactions of the Association for Computational Linguistics* 10 (2022): 291-306.

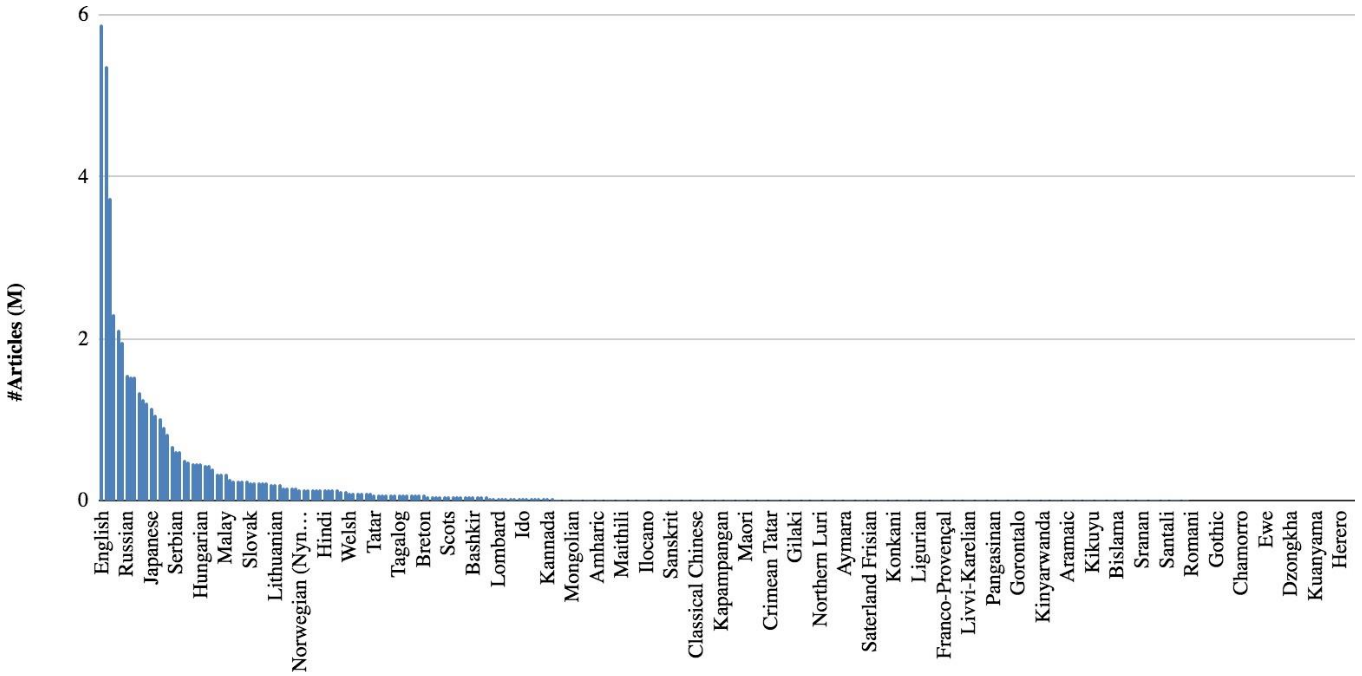
References

16. Liu, Yinhan, et al. "Multilingual denoising pre-training for neural machine translation." *Transactions of the Association for Computational Linguistics* 8 (2020): 726-742.
17. Soltan, Saleh, et al. "Alexatm 20b: Few-shot learning using a large-scale multilingual seq2seq model." *arXiv preprint arXiv:2208.01448* (2022).
18. Lin, Xi Victoria, et al. "Few-shot learning with multilingual language models." *arXiv preprint arXiv:2112.10668* (2021).
19. Wang, Thomas, et al. "What language model architecture and pretraining objective works best for zero-shot generalization?." *International Conference on Machine Learning*. PMLR, 2022.
20. Tay, Yi, et al. "Transcending scaling laws with 0.1% extra compute." *arXiv preprint arXiv:2210.11399* (2022).
21. Chung, Hyung Won, et al. "Scaling instruction-finetuned language models." *arXiv preprint arXiv:2210.11416* (2022).
22. Muennighoff, Niklas, et al. "Crosslingual generalization through multitask finetuning." *arXiv preprint arXiv:2211.01786* (2022).

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

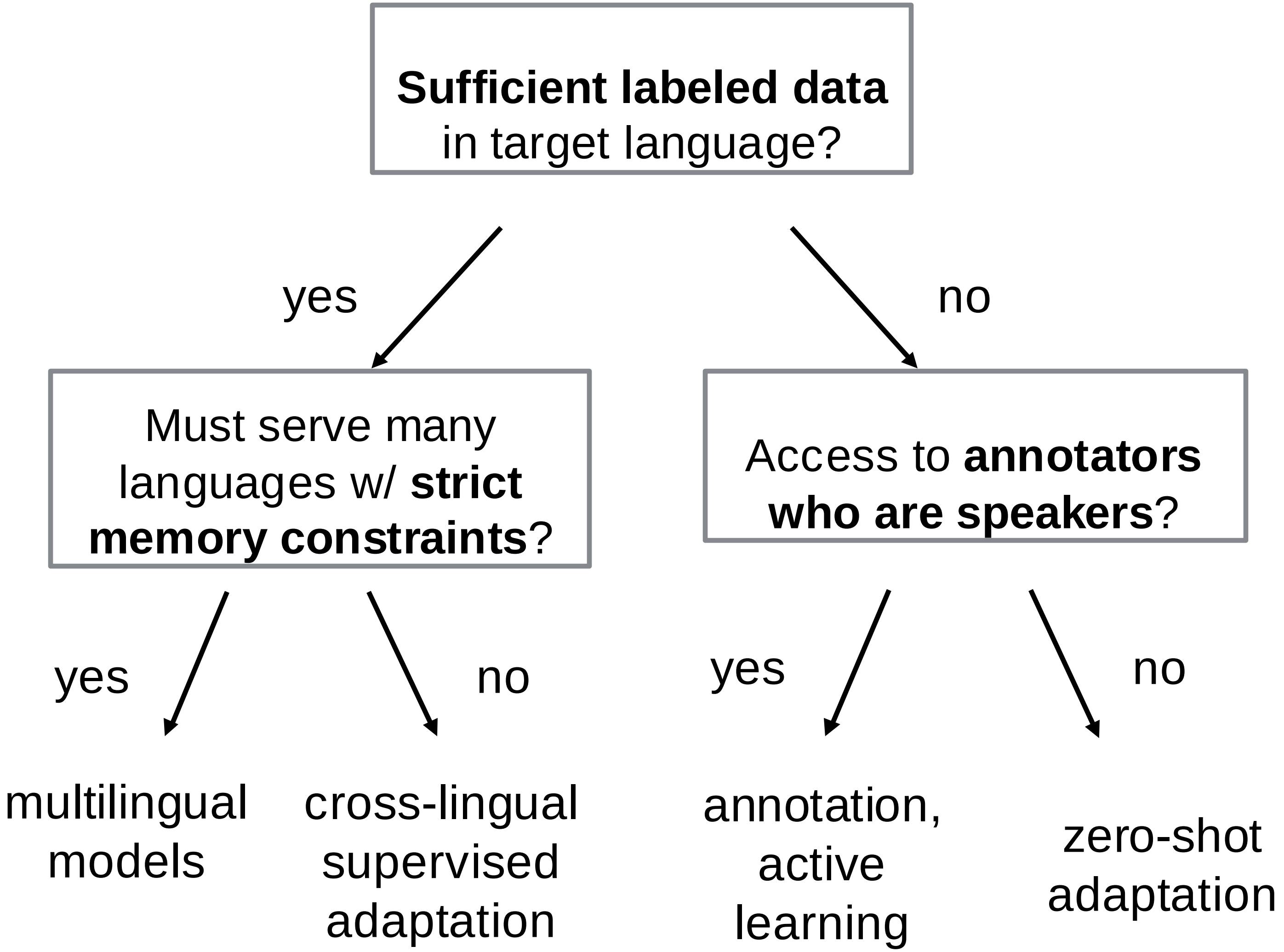
- Most methods are tested first on English, but not all languages are the same as English
- e.g.
  - Rich morphology (case, gender, etc.)
  - Accents/diacritics
  - Different scripts such as CJK
  - Dialectal language
  - Lack of formal writing systems

Data Source: Wikipedia articles from different languages

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)

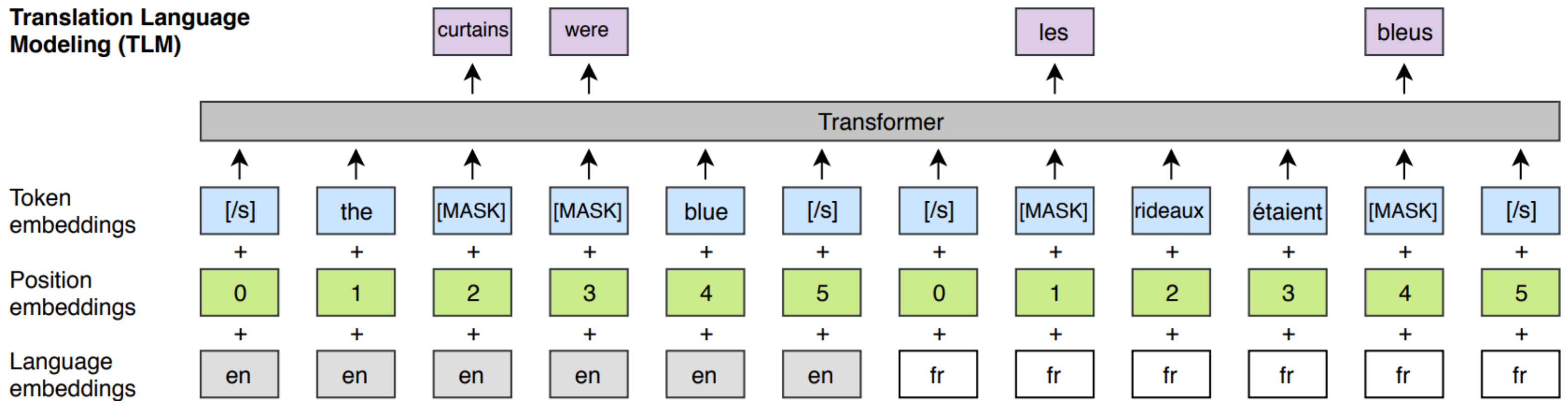
<fr> this is an example → ceci est un exemple

<ja> this is an example → これは例です

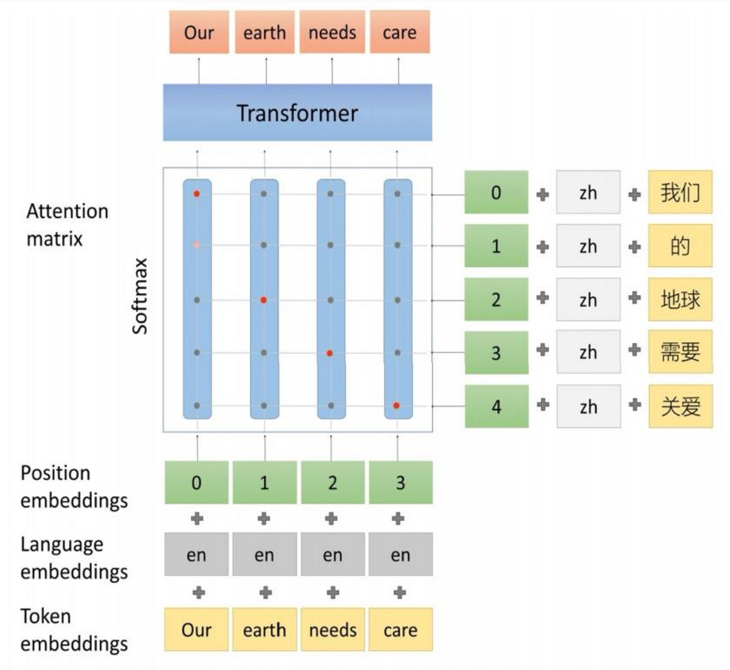
Multilingual Masked Language Modeling

-

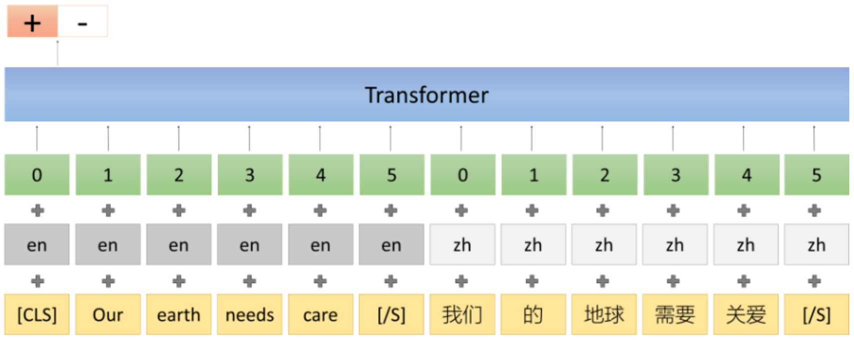
Also called translation language modeling (Lample and Conneau 2019)



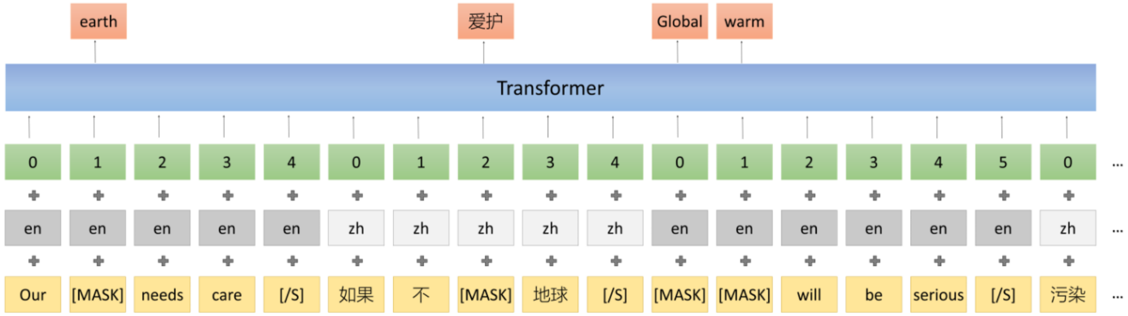
More Explicit Alignment Objectives



cross-lingual word recovery



cross-lingual paraphrase classification

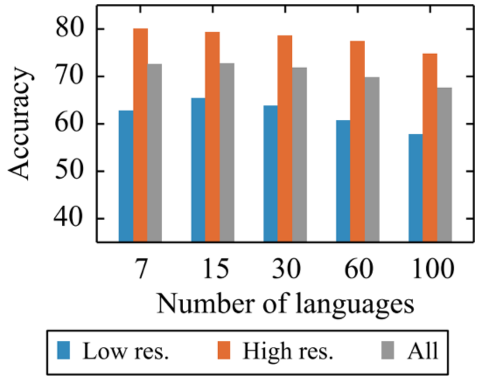


cross-lingual MLM

[Unicoder: A Universal Language Encoder by Pre-training with Multiple Cross-lingual Tasks](#) (Huang et al., EMNLP-IJCNLP 2019)

Difficulties in Fully Multi-lingual Learning

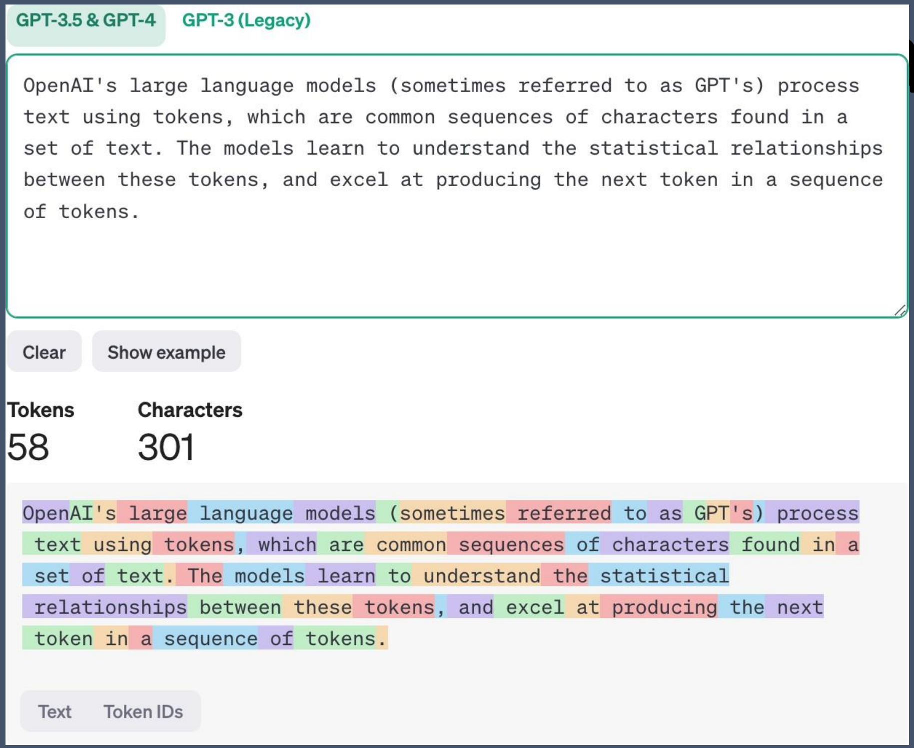
- **“Curse of Multilinguality”** For a fixed sized model, the per-language capacity decreases as we increase the number of languages. (Conneau et al, 2019)



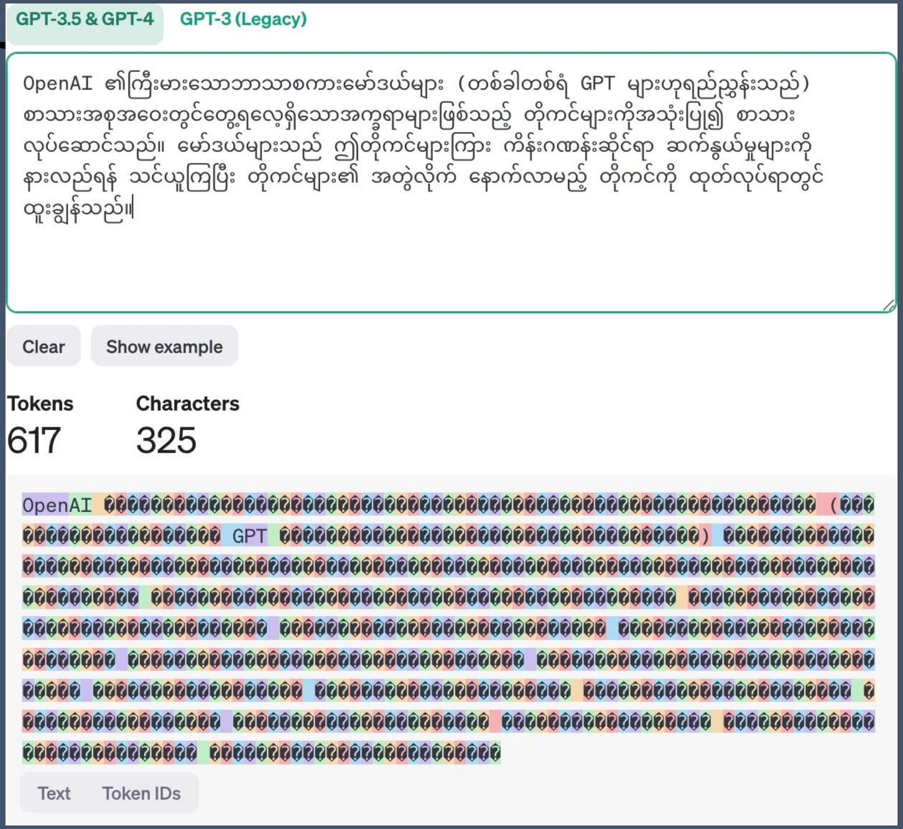
- Increasing the number of low-resource languages —> decrease in the quality of high-resource language translations (Aharoni et al, 2019)

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

English

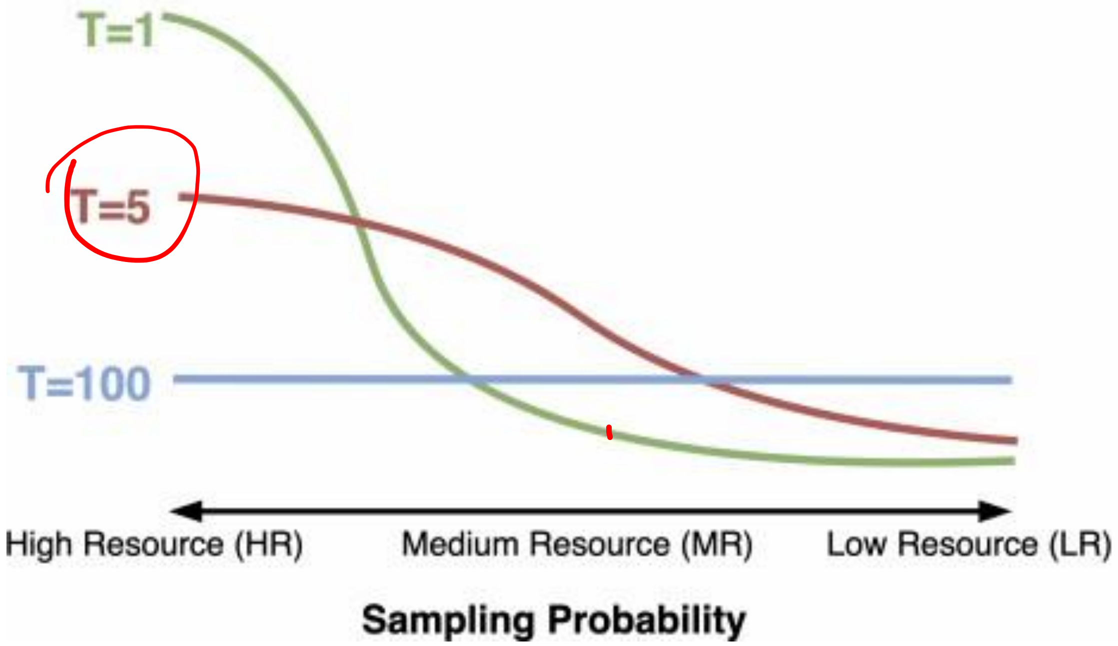


Burmese/Myanmar (Google Translated)



Similar content, 10.6x the tokens!

Heuristic Sampling of Data

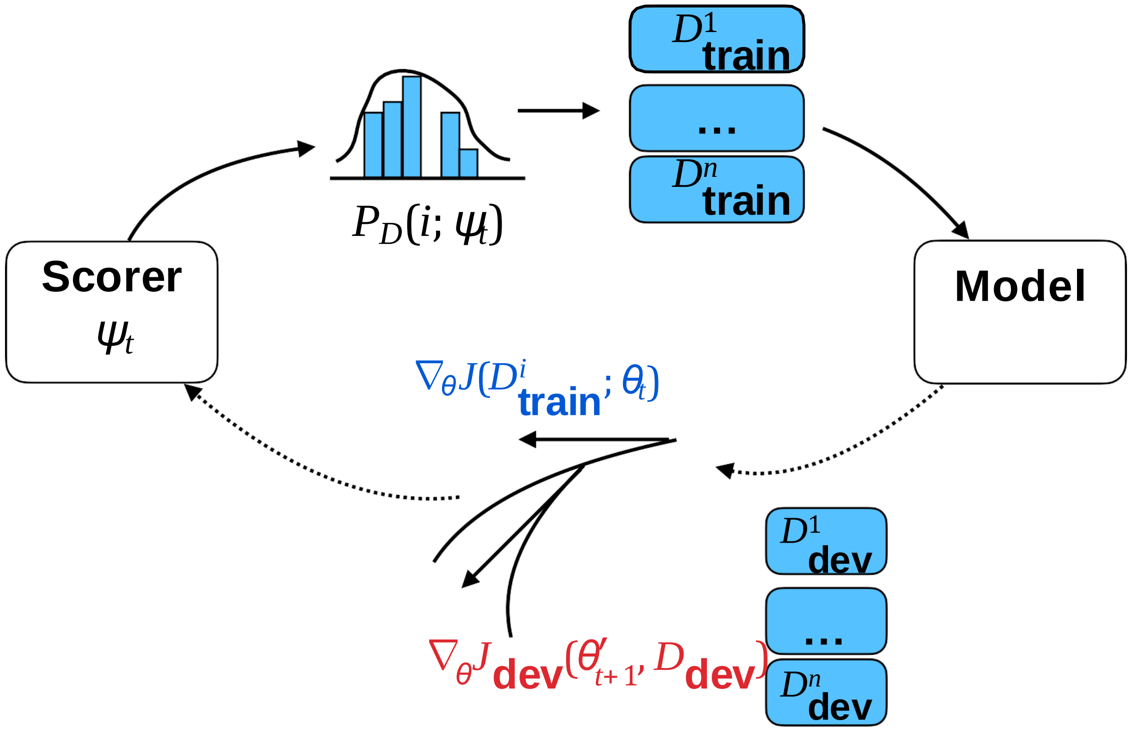


$$\frac{n_i^t}{\sum_{j=1}^{L} n_j^t}$$

- 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: indings and Challenges. 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