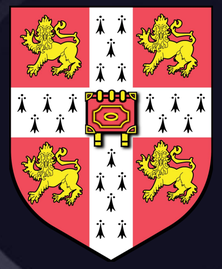
Towards Language Technology for a Truly Multilingual World?

Ivan Vulić
LTL, University of Cambridge & PolyAI



ECIR 2022, April 13 2022

Why Multilingual NLP?

Speaking **more languages** means communicating with **more people**…
…and reaching **more users and customers**…

Why Multilingual NLP?

...but there are **more profound** and **democratic** reasons to work in this area:

- decreasing **the digital divide**
- dealing with **inequality of information (access)**
- mitigating **cross-cultural biases**
- deploying language technology for **underrepresented** languages, dialects, minorities; societal impact
- understanding cross-linguistic differences

"95% of all languages in use today will never gain traction online" (Andras Kornai)

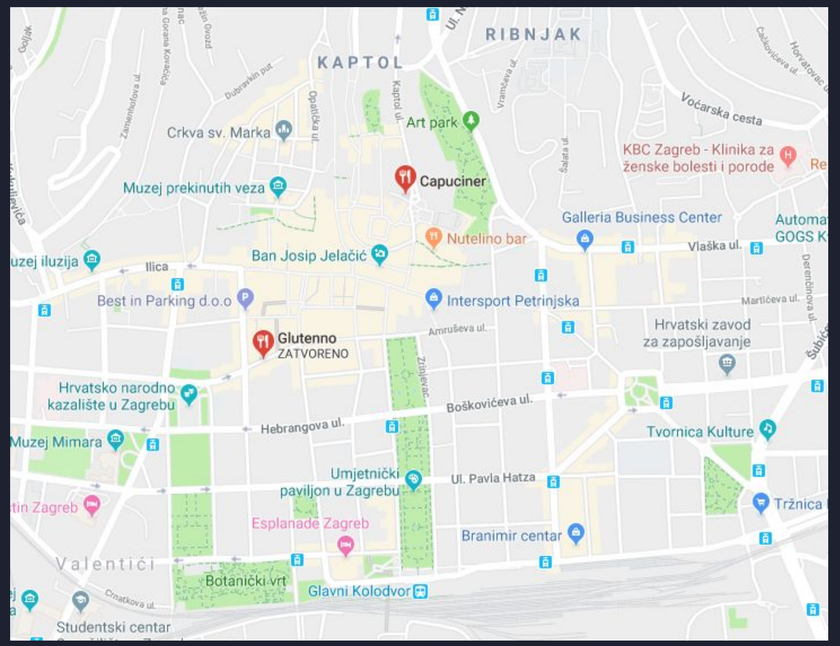
"The limits of my language *online* mean the limits of my world?"

Source: http://labs.theguardian.com/digital-language-divide/

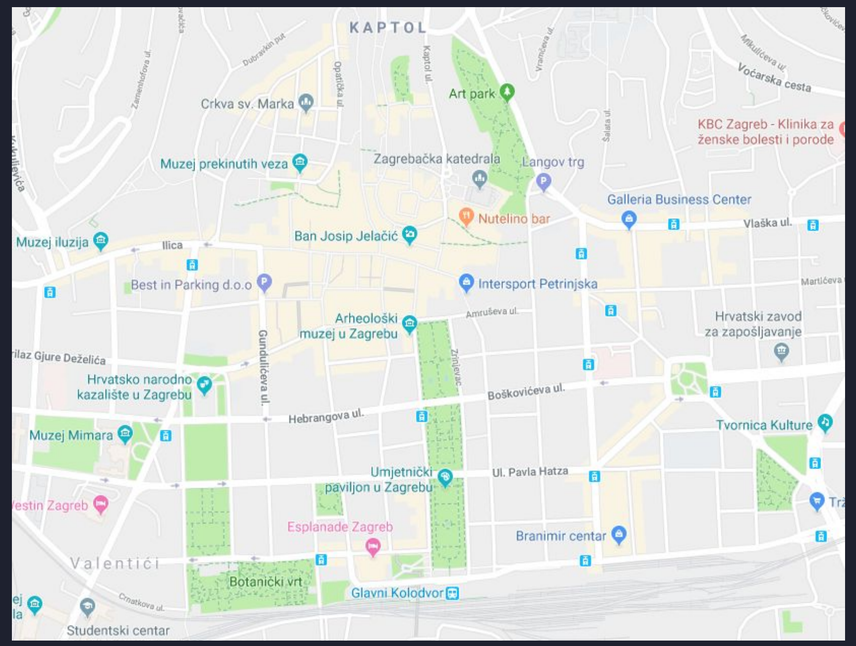
Why Multilingual NLP?

Inequality of information and representation can also affect how we understand places, events, processes...

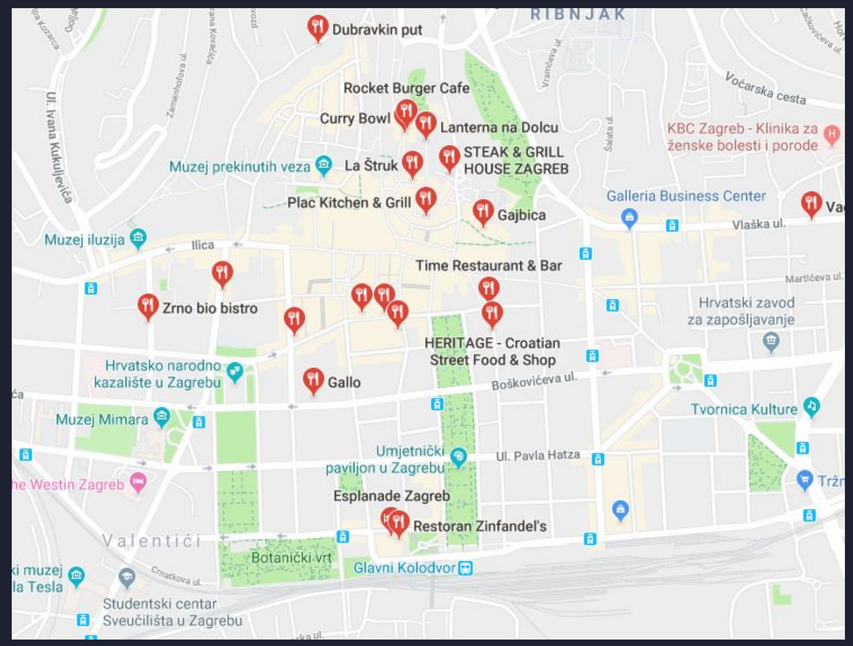
We're in Zagreb searching for...



...éttermek (HU)



...jatetxea (EU)



...restaurants (EN)

English Conversational AI

A successful conversational agent must perform:

- **Automatic speech recognition (ASR)**
- **Language analysis:**
  - Language modeling, spelling correction
  - Syntactic analysis: POS tagging, parsing
  - Semantic analysis: named entity recognition, event detection, semantic role labeling, WSD
  - Coreference resolution, entity linking, commonsense reasoning, world knowledge
- **Dialog modeling:**
  - Natural language understanding, intent detection, language generation, dialog state tracking
- **Information Search and QA**
- **Text-to-Speech**

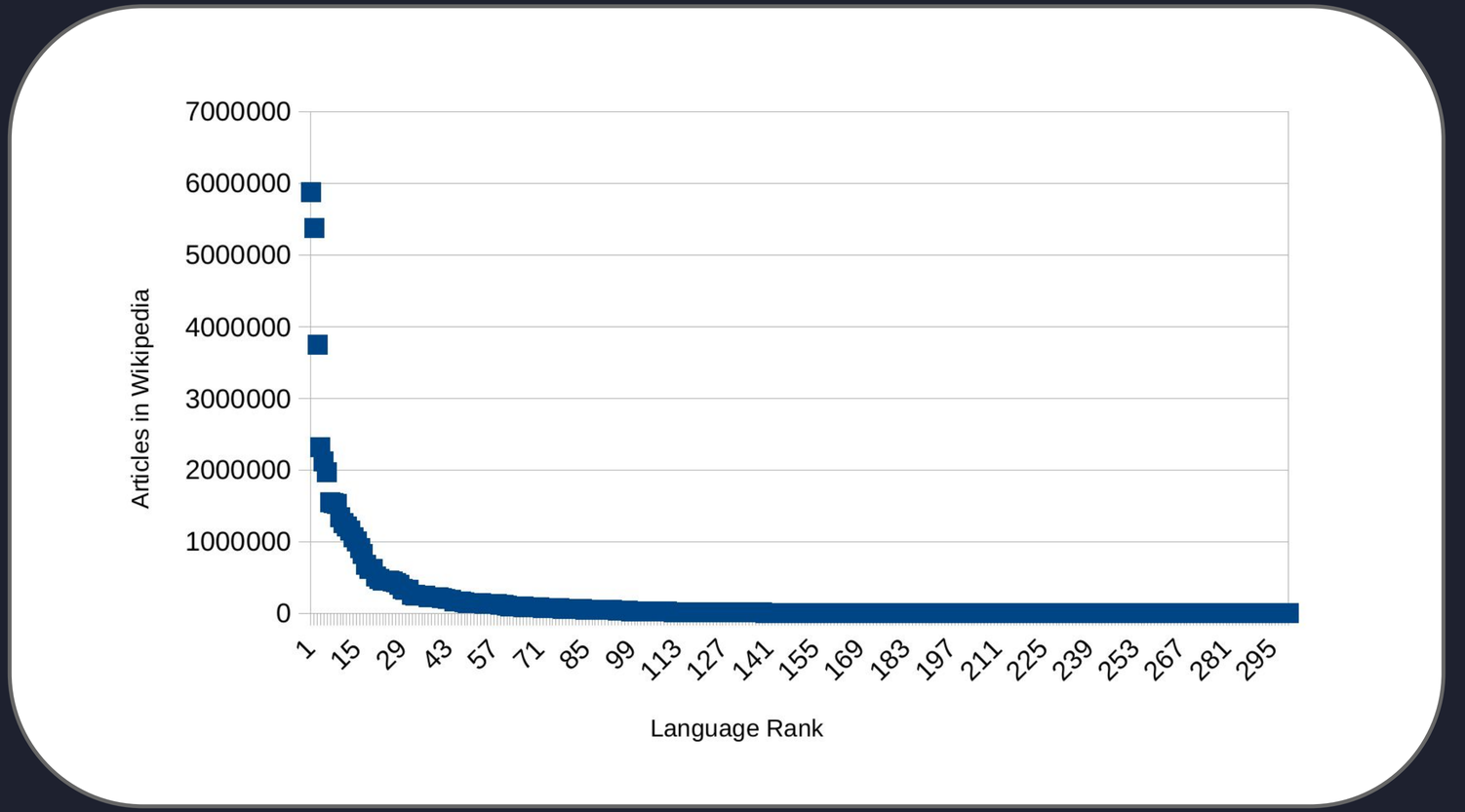
Multilingual Conversational AI?

According to Ethnologue there are **7,000+** living languages

What about **language varieties** and **dialects**?

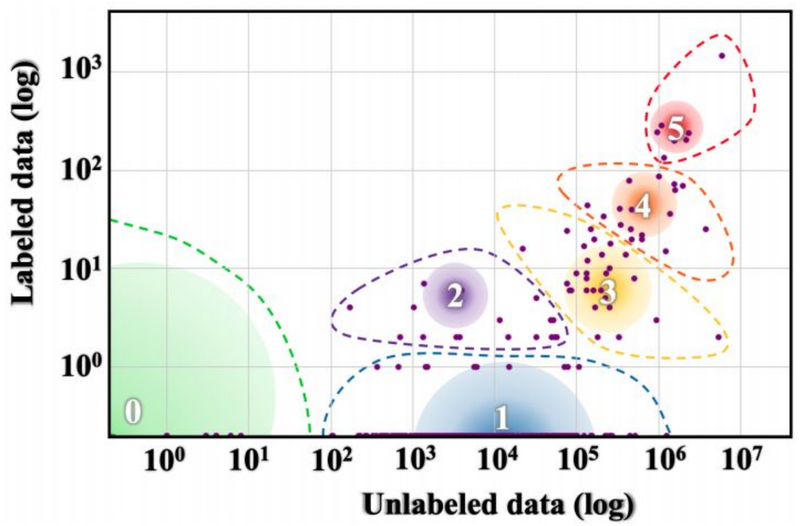
What about **“social media” languages** and **slang**?

The Long Tail of Data



Even getting “raw” unannotated data is problematic for many languages…

Are All Languages Created Equal?



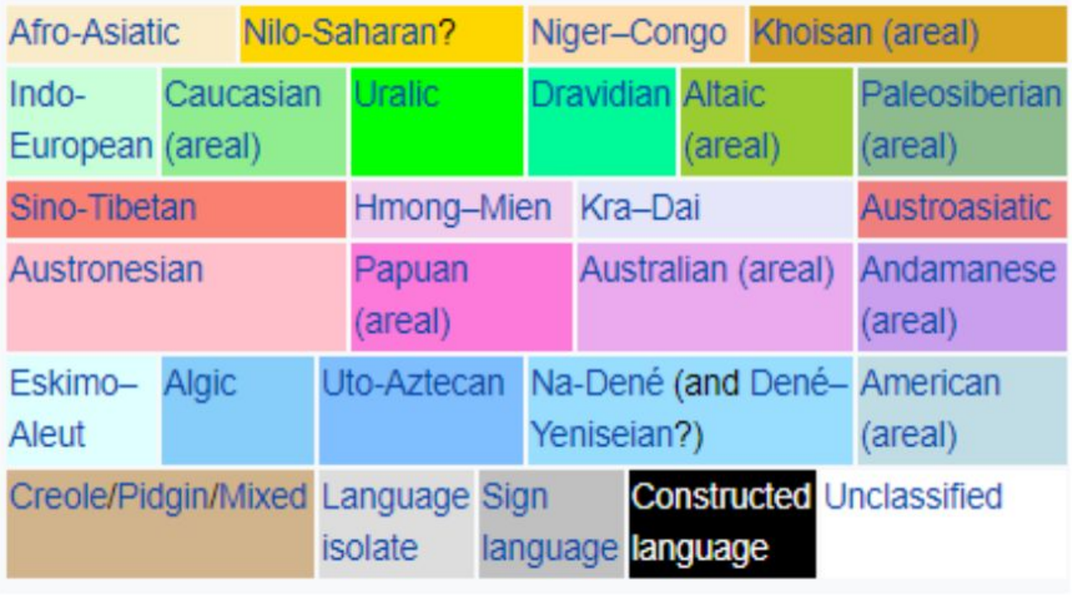
Most languages are “Left-Behinds”
[Joshi et al., ACL-20; Blasi et al., ACL-22]

Is creating equitable language technology across different languages then even possible?

Can we at least try to ‘approximate’ equality?

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%

Language Variety and Variability



[Image from: Wikipedia]

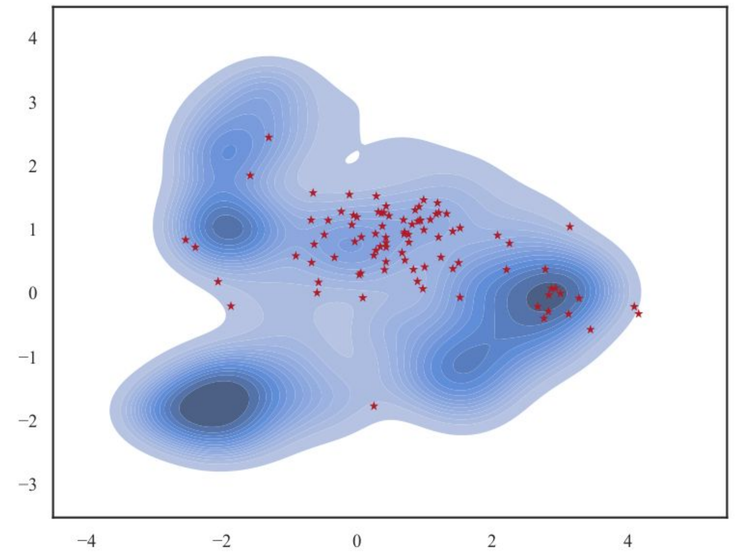


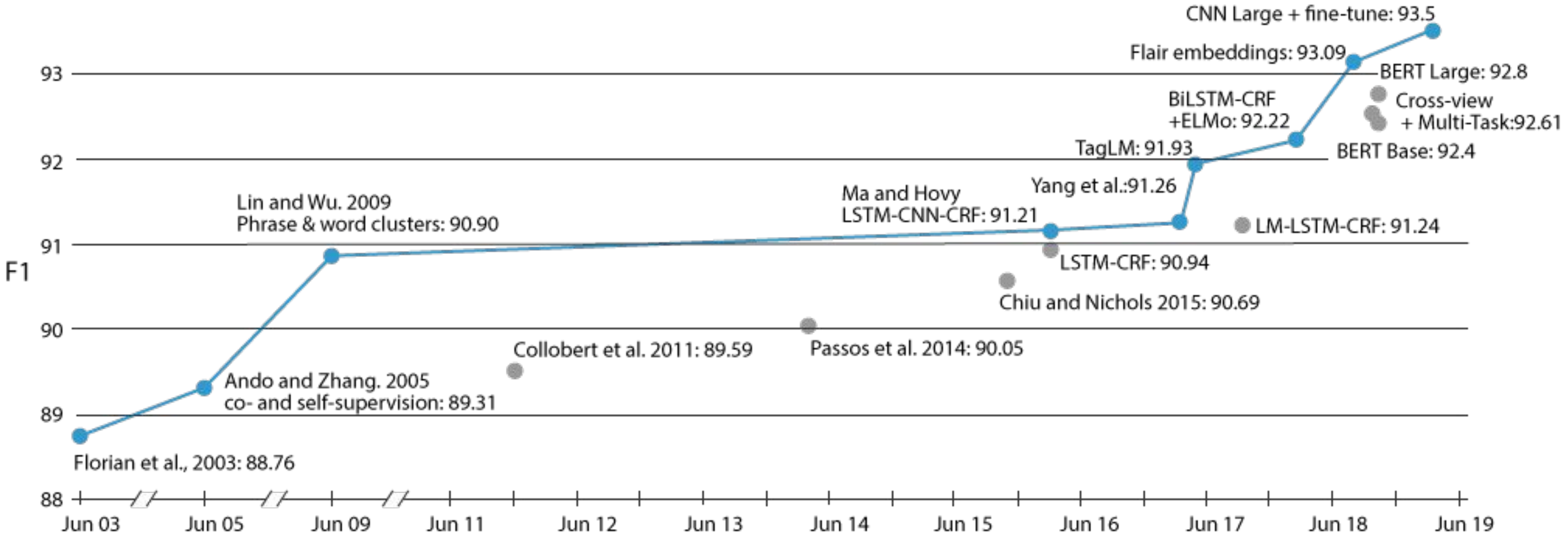
Figure 2: Density of WALS typological features of the world's languages reduced to 2 dimensions via PCA. Red dots are languages covered by UD. Darkness corresponds to more probable regions.

Image: courtesy of *Edoardo Maria Ponti*

Multilingual Representation Learning

Cross-lingual Transfer Learning

Why Cross-Lingual Transfer?



Better word representations

Cross-lingual performance follows the same trend

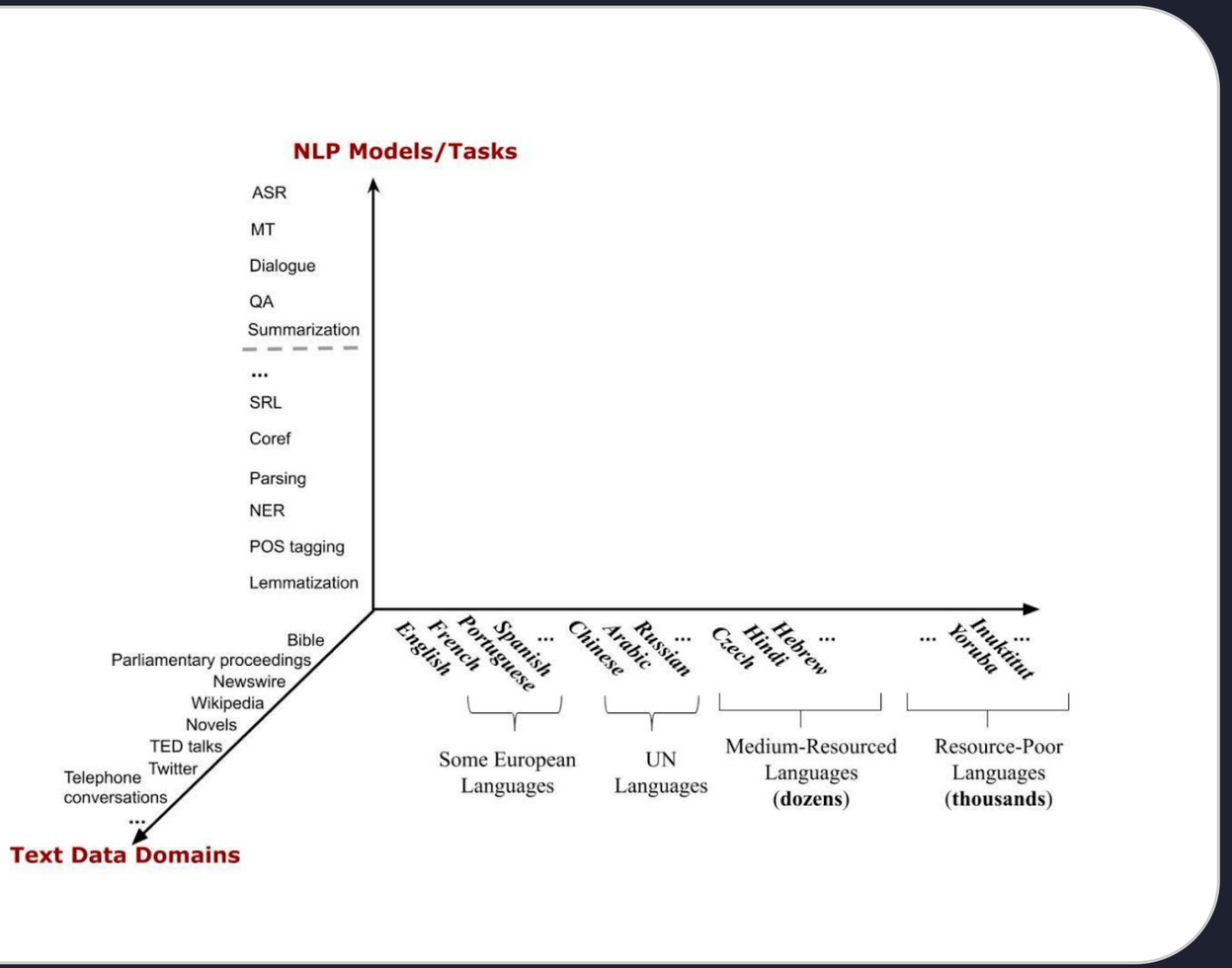
Better deep representations

Image: courtesy of *Sebastian Ruder*

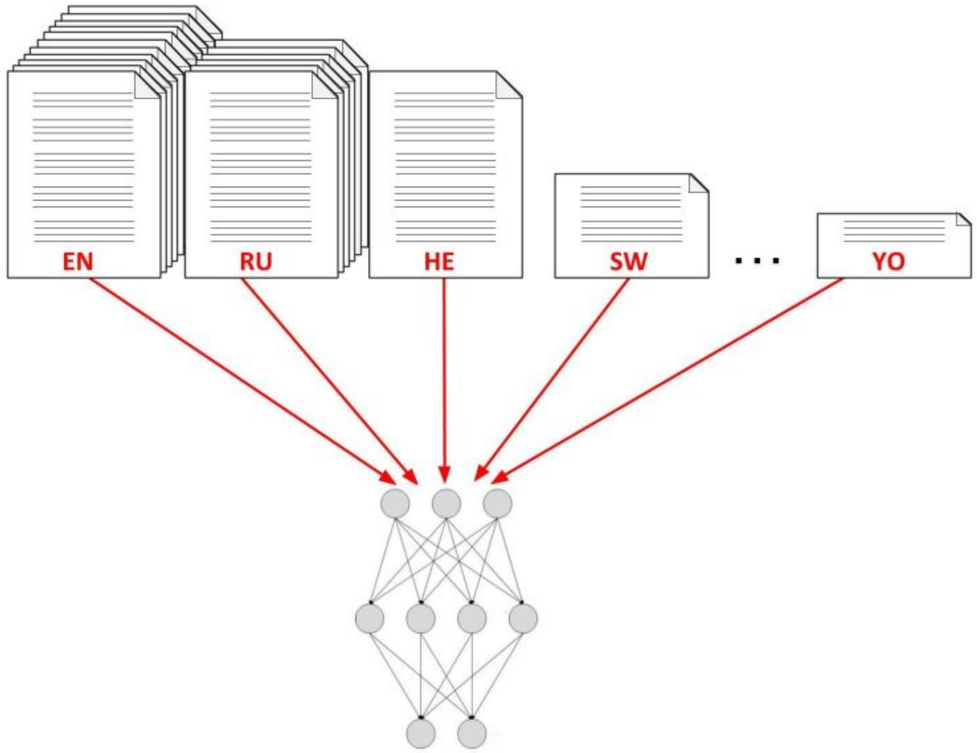
Why Cross-Lingual Transfer and Multilingual Learning?

- Many NLP tasks share common knowledge about language (e.g. linguistic representations, structural similarities)
- Languages share common structure (on the lexical, syntactic, and semantic level)
- Annotated data is rare, make use of as much supervision as available
- Empirically, transfer learning has resulted in SOTA for many supervised NLP tasks (e.g. classification, information extraction, QA, etc)

Image courtesy of *Yulia Tsvetkov*

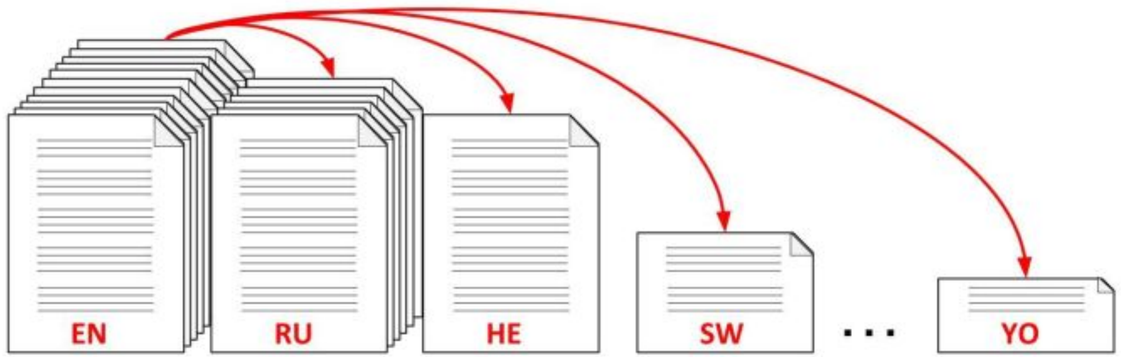


Joint Multilingual Learning in a Nutshell



Joint multilingual learning – train a single model on a mix of datasets in all languages, to enable **data and parameter sharing** where possible

Cross-Lingual Transfer in a Nutshell

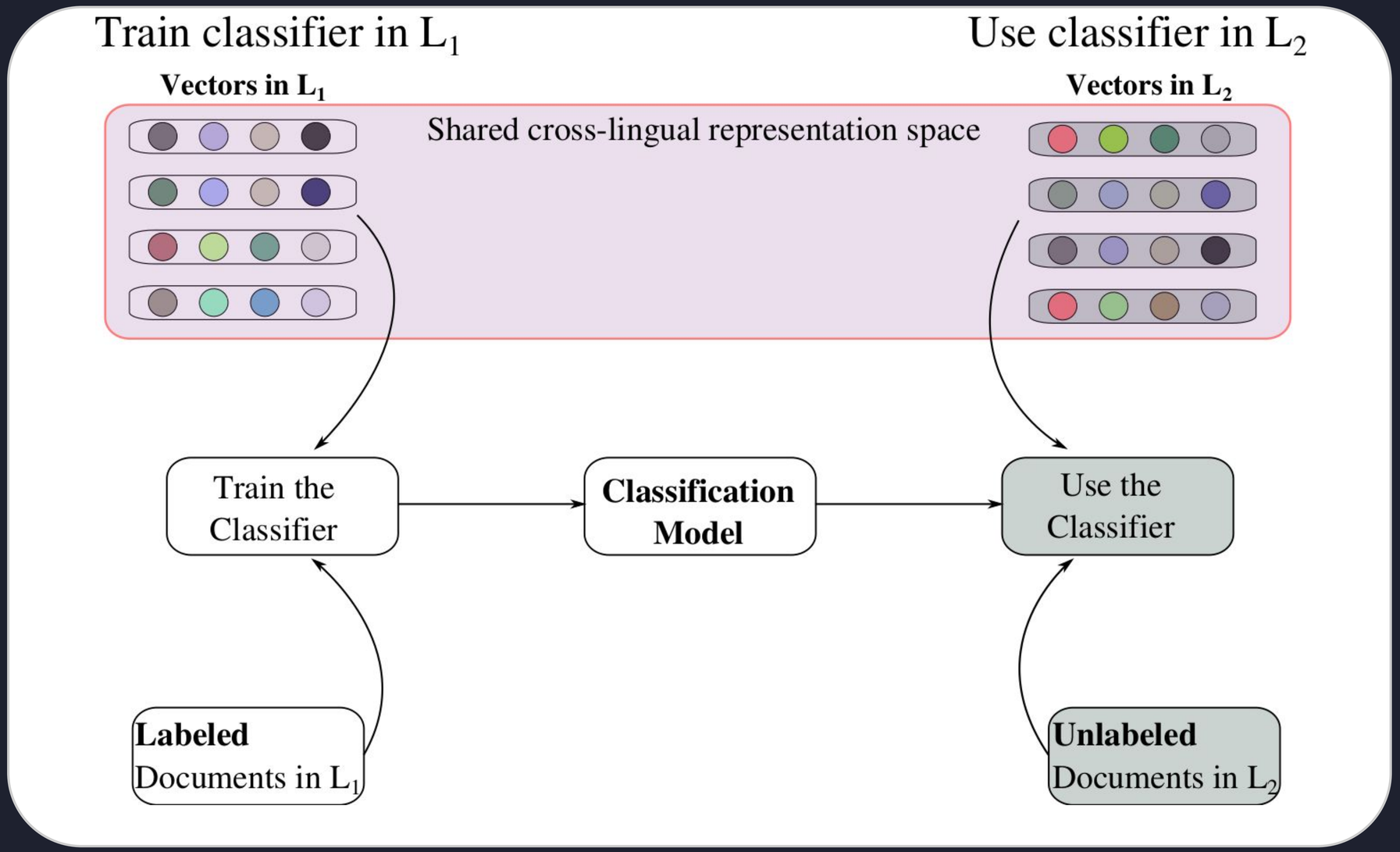


Transfer of **resources** and **models** from **resource-rich source** to **resource-poor target languages**

Zero-shot learning: train a model in one language/domain and assume it generalizes out-of-the-box in a low-resource language/domain

Few-shot learning: train a model in one language/domain and use only few examples from a low-resource language/domain to adapt it

Cross-Lingual Representations Enable Transfer



Crossing the Chasm: Representation Learning

Multilingual / cross-lingual representation of meaning

Word-Level

- Cross-lingual word embeddings
  - Words with similar meaning across languages have similar representations

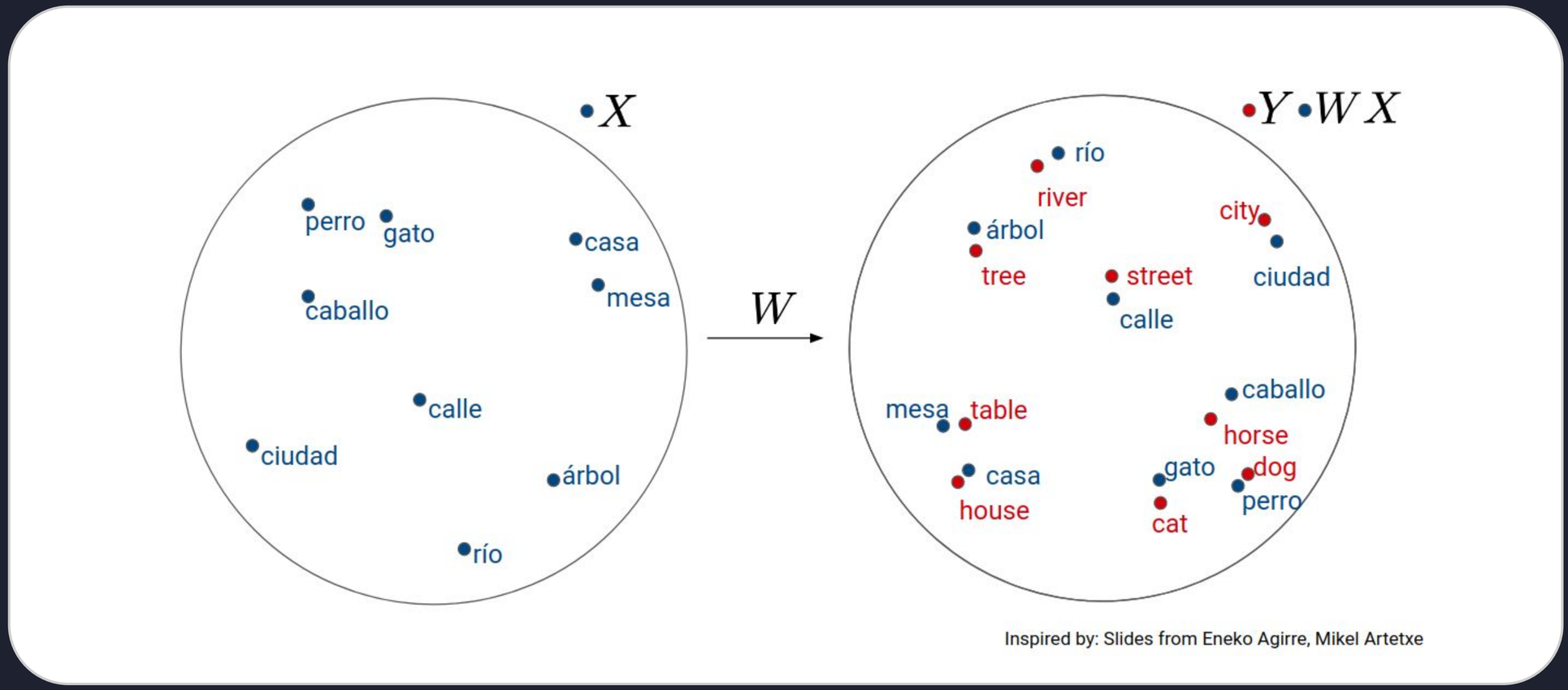
Text Encoding

- Multilingual unsupervised pretraining
  - mBERT
  - XLM(-R)
  - mT5
  - ...

Old School: Cross-Lingual Word Embeddings

A range of different methods (with different data requirements), but the same **goal:**

Induce a semantic vector space in which words with similar meaning end up with similar vectors, regardless of whether they come from the same language or from different languages.



Old School*: Cross-Lingual Word Embeddings

*Old *(NLP/ML/IR dialectal)*: dominantly used in 2019 and even in 2020...

How to use CLWEs for cross-lingual transfer for supervised tasks?

(Assumption: zero-shot transfer)

Step 1. Induce the cross-lingual (bilingual) word embedding space

Step 2. Train the (neural) model using the task-annotated data in the source language

- e.g., for NER train a BiLSTM+classifier using embeddings of the source language words as the input

Step 3. At prediction time, for texts in the target language, use embeddings of target language words as the input to the trained classifer

New School: Multilingual Language Models

Deep Transformer networks pretrained on large multilingual corpora via (masked) language modeling objectives

They work with an automatically induced shared subword vocabulary across all languages they represent

They are unsupervised from the perspective of not using any explicit cross-lingual learning signal.

At first, praised for their effective (zero-shot) cross-lingual performance

- *“Surprising cross-lingual effectiveness of mBERT”*
- *“mBERT surprisingly good at zero-shot cross-lingual transfer”*

New School: Zero-Shot Transfer to (Low-Resource) Languages

Step 1:

Train a multilingual model.

Step 2:

Fine-tune model on a **task** in a high resource **source language.**

Step 3:

Transfer and **evaluate** the model on a low resource **target language.**

Why?

Training **data** is **expensive** and not available for many languages, especially ones that are considered “low-resource”.

So... We Have Solved Zero-Shot Cross-Lingual Transfer?

No! Settings in which they were evaluated were **too simple** and **too favourable**...

Study 1. **Tasks:** NER, POS tagging
(Pires et al., ACL-19)

Target Languages: DE, NL, ES

Study 2. **Tasks:** NER, NLI
(Karthikeyan et al., ICLR-20)

Target Languages: ES, HI, RU

In most studies the selected target languages were:

1. from **the same language family** as the source (English)
2. with **large corpora in pretraining**

So… We Have Solved Zero-Shot Cross-Lingual Transfer?

Task	Model	EN	ZH Δ	TR Δ	RU Δ	AR Δ	HI Δ	EU Δ	FI Δ	HE Δ	IT Δ	JA Δ	KO Δ	SV Δ	VI Δ	TH Δ	ES Δ	EL Δ	DE Δ	FR Δ	BG Δ	SW Δ	UR Δ
DEP	B	91.2	-43.9	-46.0	-28.1	-56.4	-36.1	-50.2	-30.7	-36.1	-17.1	**-60.1**	-56.1	-14.3	-	-	-	-	-	-	-	-	-
	X	92.0	**-85.4**	-44.2	-29.7	-54.6	-39	-49.5	-26.7	-39	-23.5	-80.5	-56.0	-16.3	-	-	-	-	-	-	-	-	-
POS	B	95.8	-38.0	-35.9	-16.0	-40.1	-33.4	-34.6	-21.9	-33.4	-19.8	**-46.1**	-42.0	-9.6	-	-	-	-	-	-	-	-	-
	X	96.3	-69.2	-27.7	-14.3	-37.1	-27.3	-31.9	-17.9	-27.3	-19.0	**-77.0**	-37.3	-10.7	-	-	-	-	-	-	-	-	-
NER	B	92.4	-23.3	-11.6	-10.7	**-31.7**	-11.1	-12.8	-3.8	-11.1	-2.6	-25.7	-13.8	-6.7	-	-	-	-	-	-	-	-	-
	X	91.6	**-34.8**	-6.2	-13.7	-24.6	-16.5	-8.0	-0.9	-16.5	-2.4	-30.1	-15.6	-2.2	-	-	-	-	-	-	-	-	-
XNLI	B	82.8	-13.6	-20.6	-13.5	-17.3	-21.3	-	-	-	-	-	-	-	-11.9	-28.1	-8.1	-14.1	-10.5	-7.8	-13.3	**-33.0**	-23.4
	X	84.3	-11.0	-11.3	-9.0	-13.0	-14.2	-	-	-	-	-	-	-	-9.7	-12.3	-5.8	-8.9	-7.8	-6.1	-6.6	**-20.2**	-17.3
XQuAD	B	71.1	-22.9	-34.2	-19.2	-24.7	-28.6	-	-	-	-	-	-	-	-22.1	**-43.2**	-16.6	-28.2	-14.8	-	-	-	-
	X	72.5	**-26.2**	-18.7	-15.4	-24.1	-22.8	-	-	-	-	-	-	-	-19.7	-14.8	-14.5	-15.7	-16.2	-	-	-	-

(Lauscher et al., EMNLP-20)

B=mBERT; X=XLM-R

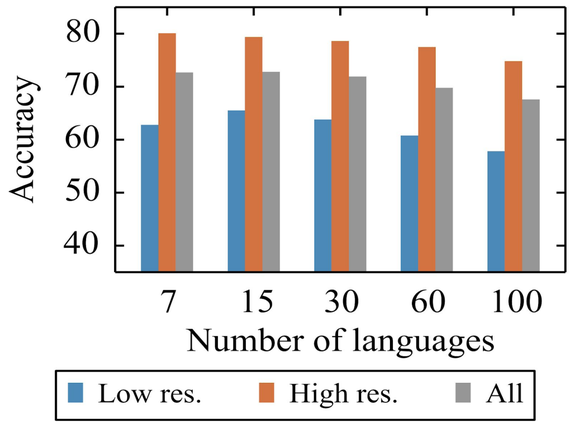
Huge drops for:

1. **Distant** target languages
2. Target languages with **small pretraining corpora**

So… We Have Solved Zero-Shot Cross-Lingual Transfer?

More problems…

“The Curse of Multilinguality” (Conneau et al., ACL-20)

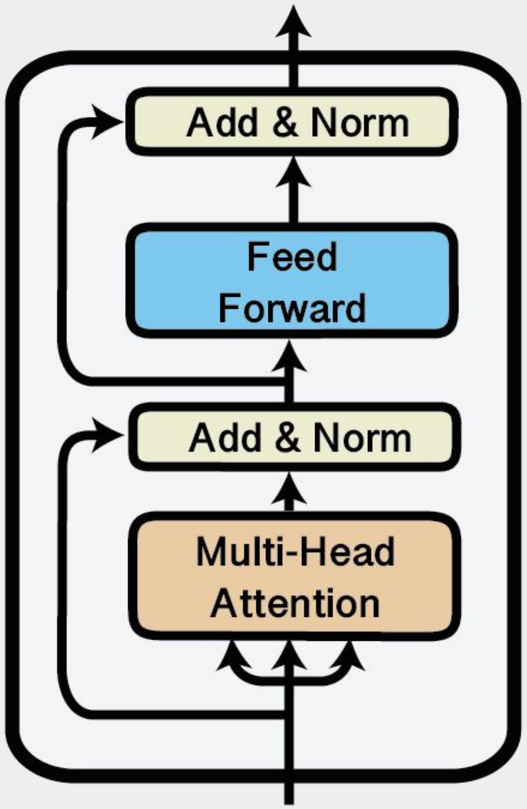


And what about low-resource languages not covered at all in the pretraining data?

Some Recap and Basics...

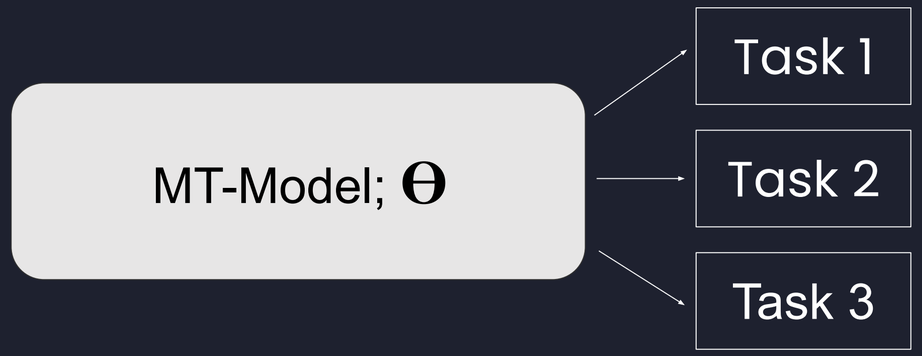
(Almost) everything I am about to cover in more detail involves:

- Transfer learning in NLP
- We **only** look at deep neural networks, specifically the **Transformer** architecture.
- We leverage pre-trained transformer-based models such as BERT/RoBERTa/XLM-R/mBERT.
- These have been trained on **massive** amount of text data using **Masked Language Modeling (MLM).**
- Transfer learning with these pretrained models usually involves stacking a prediction head on top of the model.
- **Usually** all parameters are the fine-tuned on the downstream task (e.g. using cross-entropy loss).



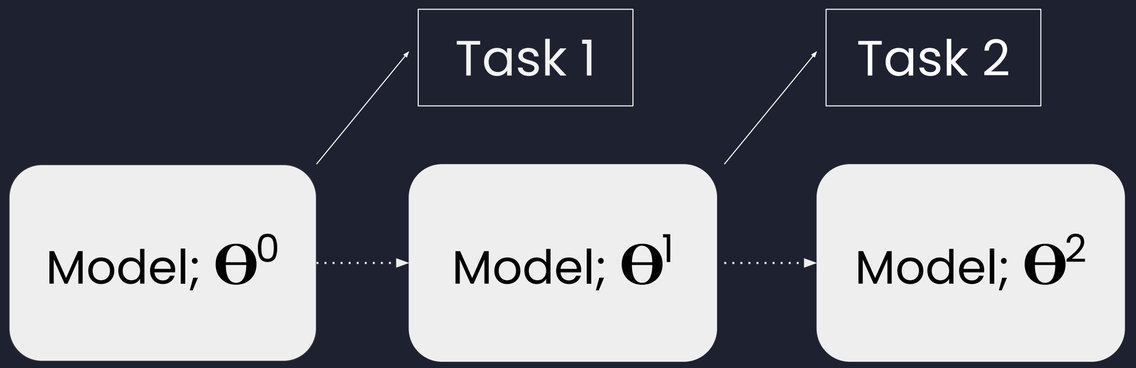
Problems of Multi-Task and Transfer Learning

Multi-Task Learning:



Catastrophic Interference: Sharing all parameters Θ between tasks results in deterioration of performance for a subset of tasks.

Sequential Fine-Tuning:

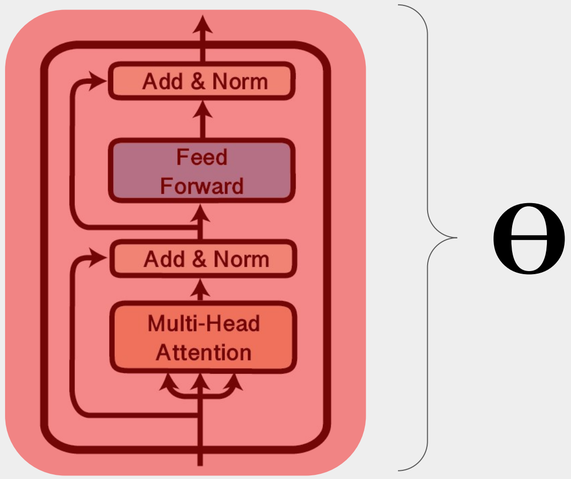


Catastrophic Forgetting: Sequential fine-tuning on tasks results in forgetting information learned in earlier stages of transfer learning.

Modular and Parameter-Efficient?

$$\theta \leftarrow \underset{\theta}{\operatorname{argmin}}\, L(D_{\text{NLI}}; \theta)$$

A single Transformer (encoder) layer



D_{NLI} = NLI Dataset

L = Loss function, e.g. cross entropy loss

θ = Parameters of the model

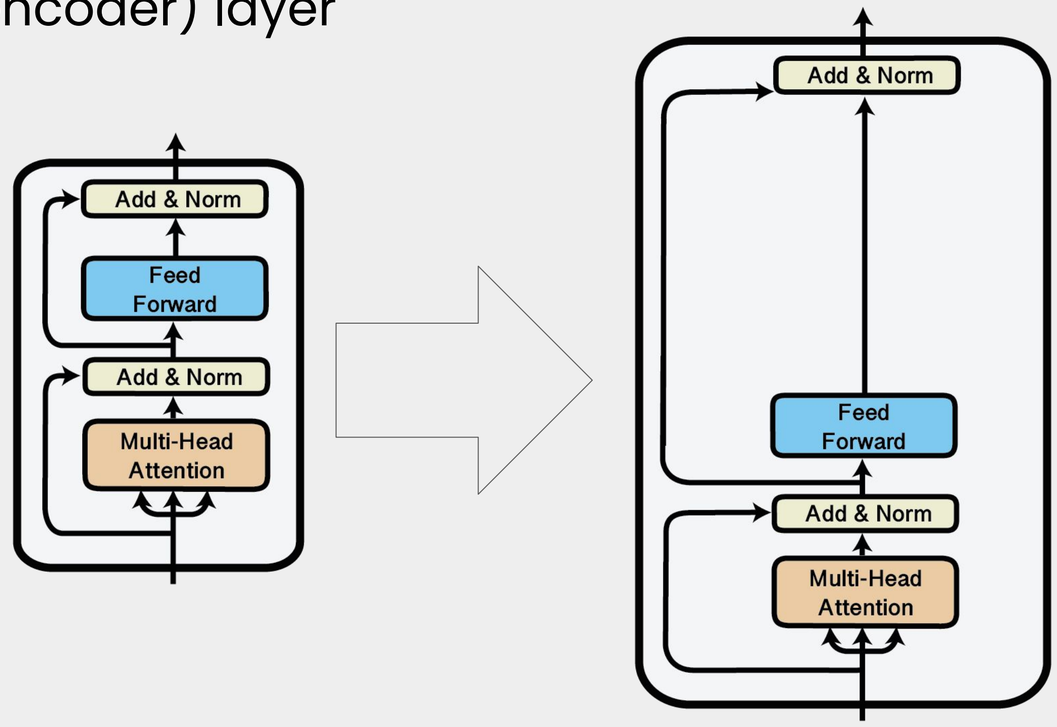
= Parameters are frozen

= Parameters are fine-tuned

Modular and Parameter-Efficient: Adapters

$$\theta \leftarrow \underset{\theta}{\operatorname{argmin}}\, L(D_{\text{NLI}}; \theta)$$

A single Transformer (encoder) layer



= Parameters are frozen

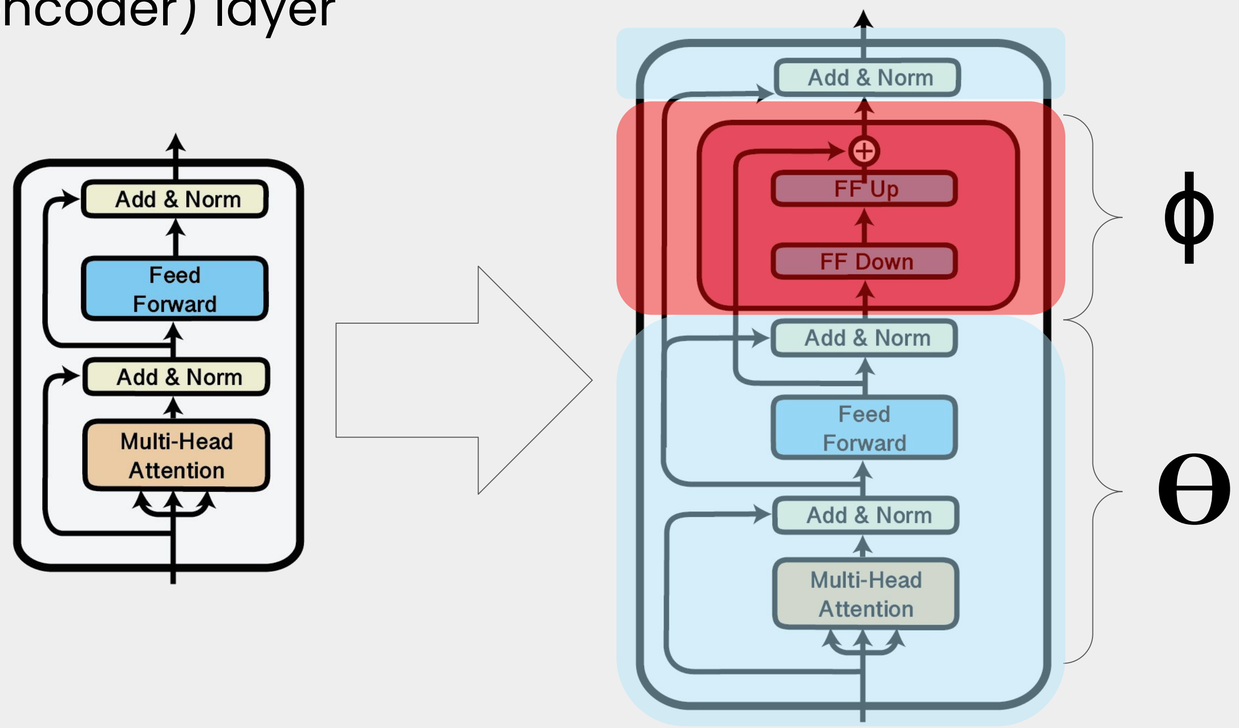
= Parameters are fine-tuned

Houlsby, Neil, et al. "Parameter-Efficient Transfer Learning for NLP." *International Conference on Machine Learning*. 2019.

Modular and Parameter-Efficient: Adapters

$$\phi \leftarrow \underset{\phi}{\operatorname{argmin}}\, L(D_{\text{NLI}}; \theta, \phi)$$

A single Transformer (encoder) layer



Adapter parameters ϕ are **encapsulated** between transformer layers with parameters θ which are frozen

= Parameters are frozen

= Parameters are fine-tuned

Houlsby, Neil, et al. "Parameter-Efficient Transfer Learning for NLP." *International Conference on Machine Learning*. 2019.

Parameter Efficiency of Adapters in Transformers

Training adapters instead of full model fine-tuning achieves similar results.

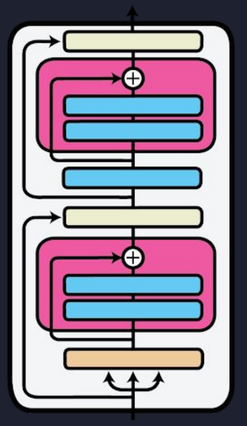
Adapters are smaller in size than training the full model.

Performance on GLUE tasks

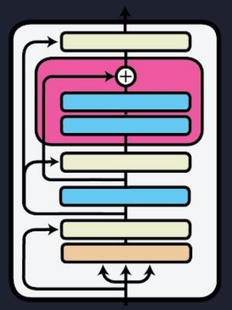
	Full	Pfeif.	Houl.
RTE (Wang et al., 2018)	66.2	70.8	69.8
MRPC (Dolan and Brockett, 2005)	90.5	89.7	91.5
STS-B (Cer et al., 2017)	88.8	89.0	89.2
CoLA (Warstadt et al., 2019)	59.5	58.9	59.1
SST-2 (Socher et al., 2013)	92.6	92.2	92.8
QNLI (Rajpurkar et al., 2016)	91.3	91.3	91.2
MNLI (Williams et al., 2018)	84.1	84.1	84.1
QQP (Iyer et al., 2017)	91.4	90.5	90.8

Number of newly introduced Parameters

	Base		**Large**	
CRate	#Params	Size	#Params	Size
64	0.2M	0.9Mb	0.8M	3.2Mb
16	0.9M	3.5Mb	3.1M	13Mb
2	7.1M	28Mb	25.2M	97Mb



Houlsby et al., 2019

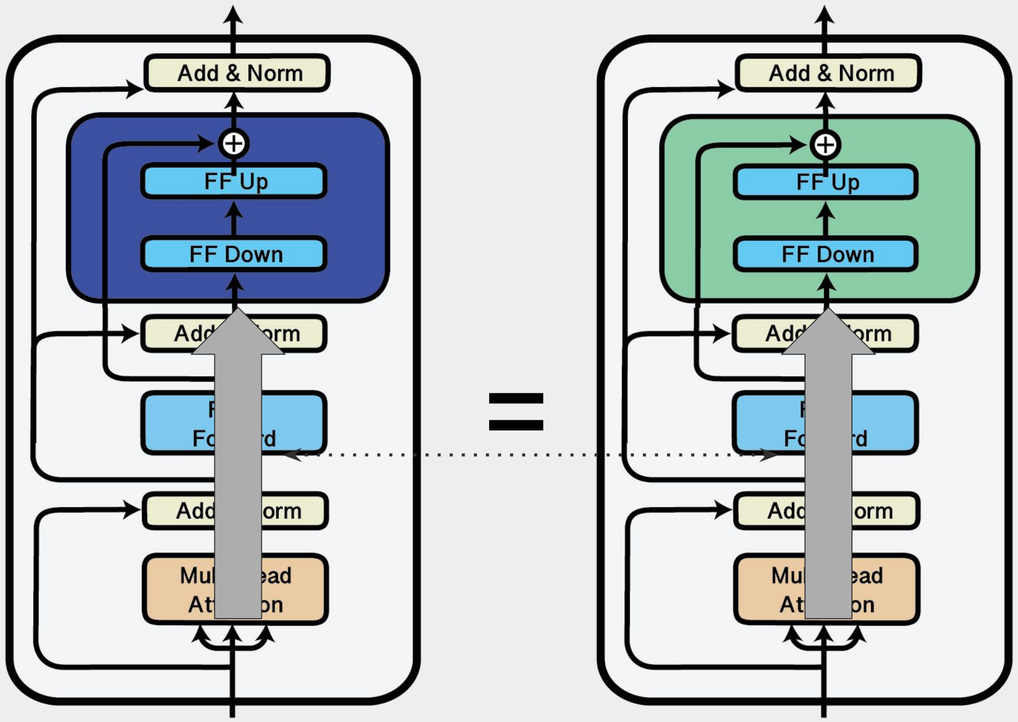


Pfeiffer et al., 2020a

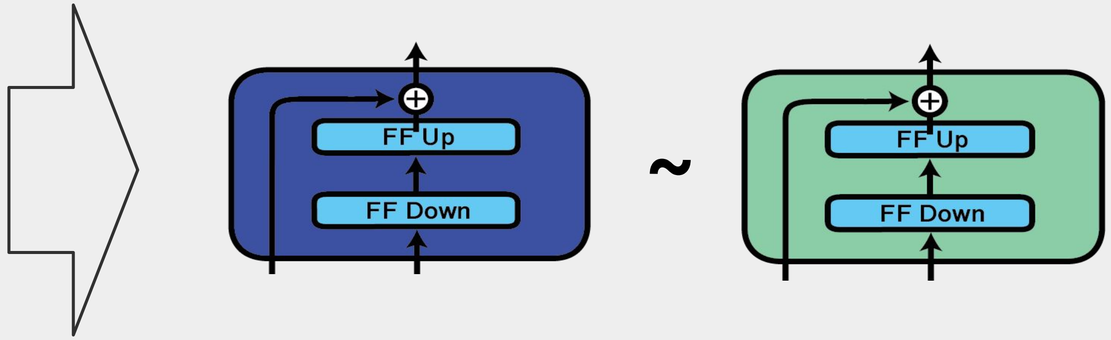
Houlsby, Neil, et al. "Parameter-Efficient Transfer Learning for NLP." *International Conference on Machine Learning*. 2019.
Pfeiffer, Jonas, et al. "AdapterFusion: Non-destructive task composition for transfer learning." *arXiv preprint* (2020a).
Pfeiffer, Jonas, et al. "Adapterhub: A framework for adapting transformers." *Proceedings of EMNLP: Systems Demonstrations* (2020b)

Encapsulated Adapters?

MLM (English) **MLM (Quechuan)**



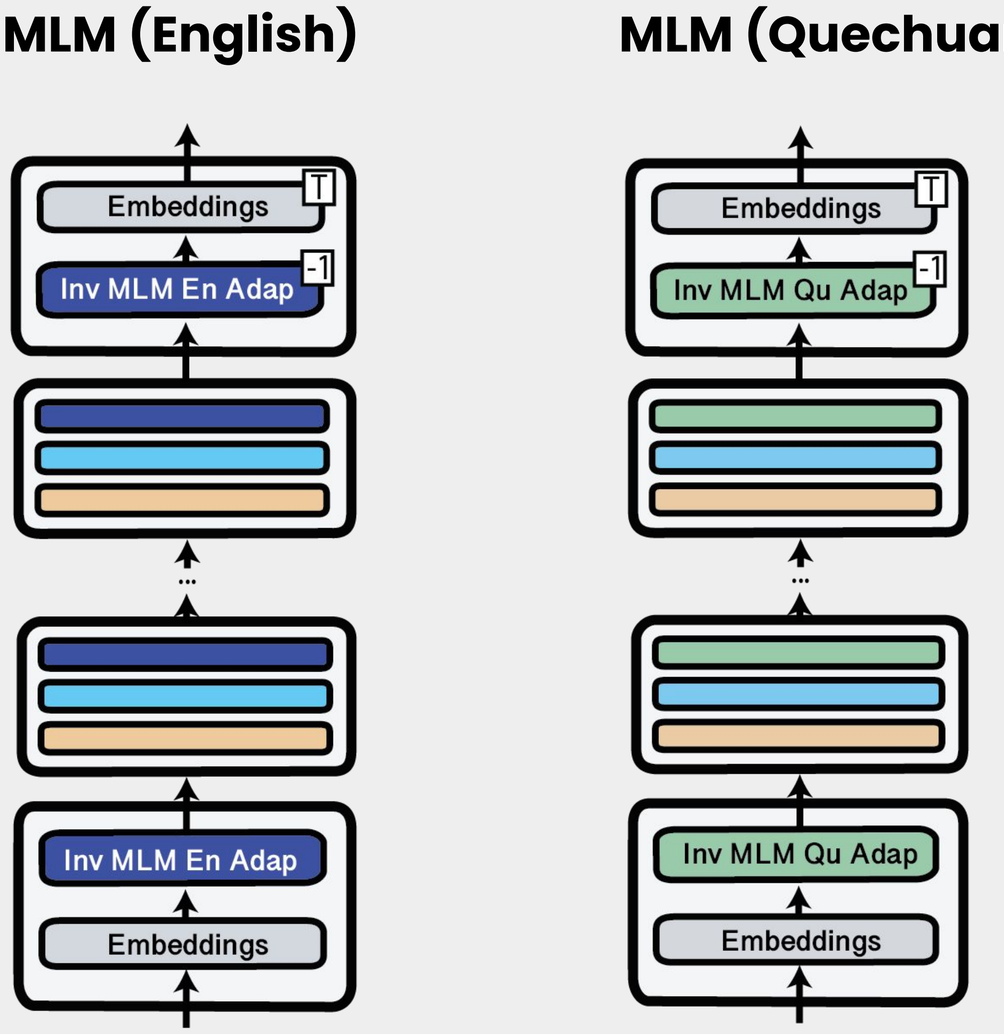
- Adapters **learn transformations** that make the underlying model **more suited** to a task or language.
- Using masked language modelling (MLM), we can learn **language-specific transformations** for e.g. **English** and **Quechua**.
- As long as the underlying model is kept fixed, these transformations are **roughly interchangeable**.



MAD-X: An Adapter-Based Framework for Transfer

Step 1: Train Language Adapters

We train **language adapters** for the **source language** and the **target language** with masked language modelling on Wikipedia.

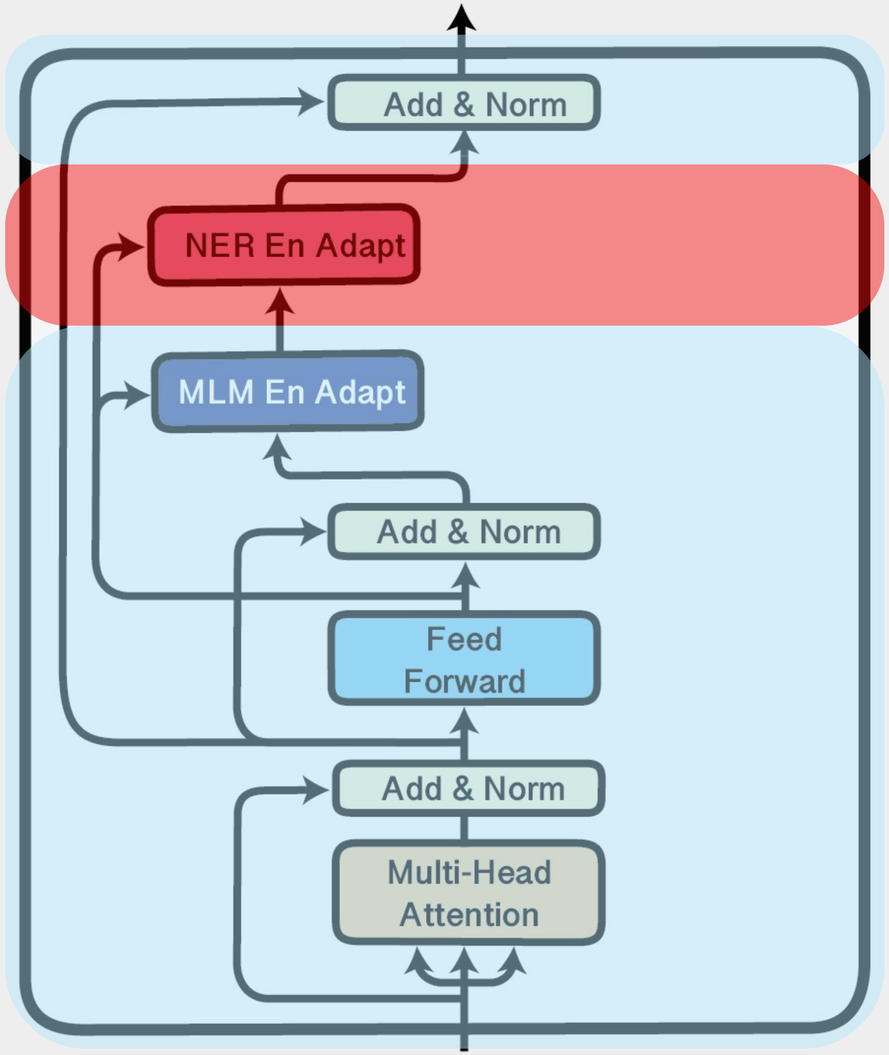


MAD-X

Step 2: Train a Task Adapter

We **train** **task adapters** in the source language **stacked** on top of the source **language adapter**.

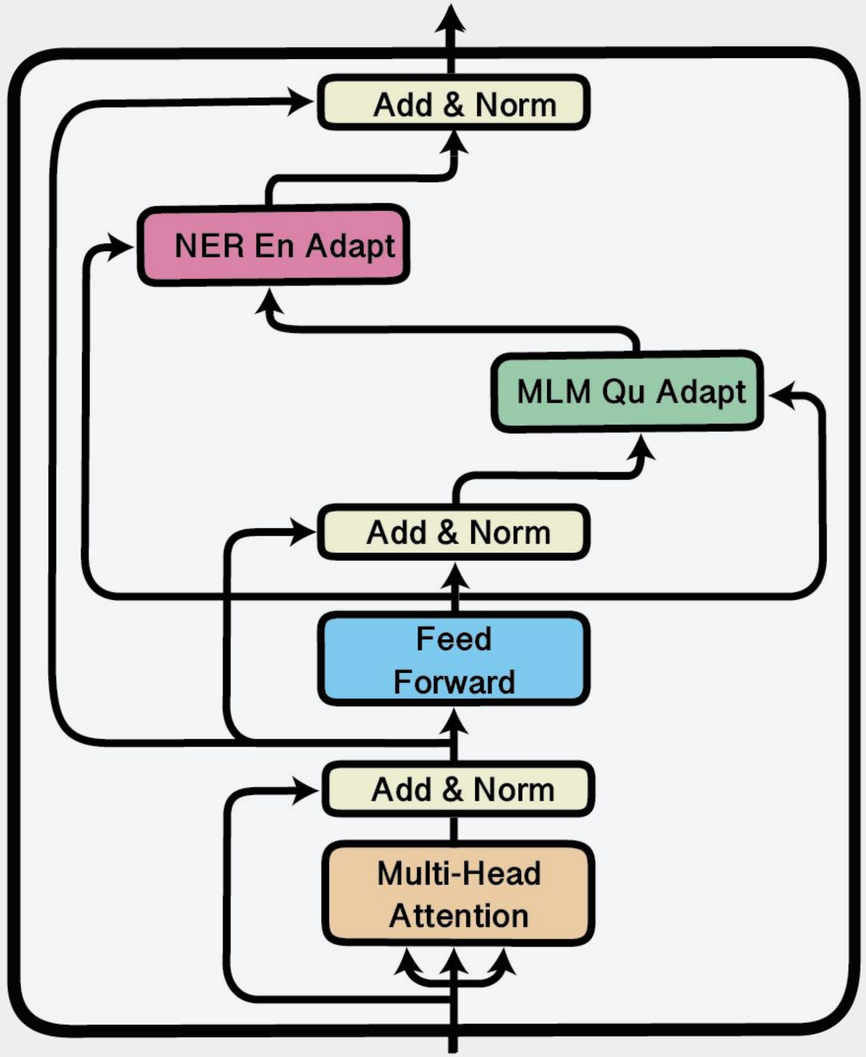
The language adapter ϕ_l as well as the transformer weights θ are frozen while only the task adapter parameter ϕ_t are trained.



MAD-X

Step 3: Zero-Shot transfer to unseen language

We **replace** the **source** language adapter with the **target** language adapter, while **keeping** the "language agnostic" **task adapter**.



Datasets: Inclusion of Diverse and Low-Resource

NER: WikiAnn Dataset We chose a diverse set of languages from **different language families**.

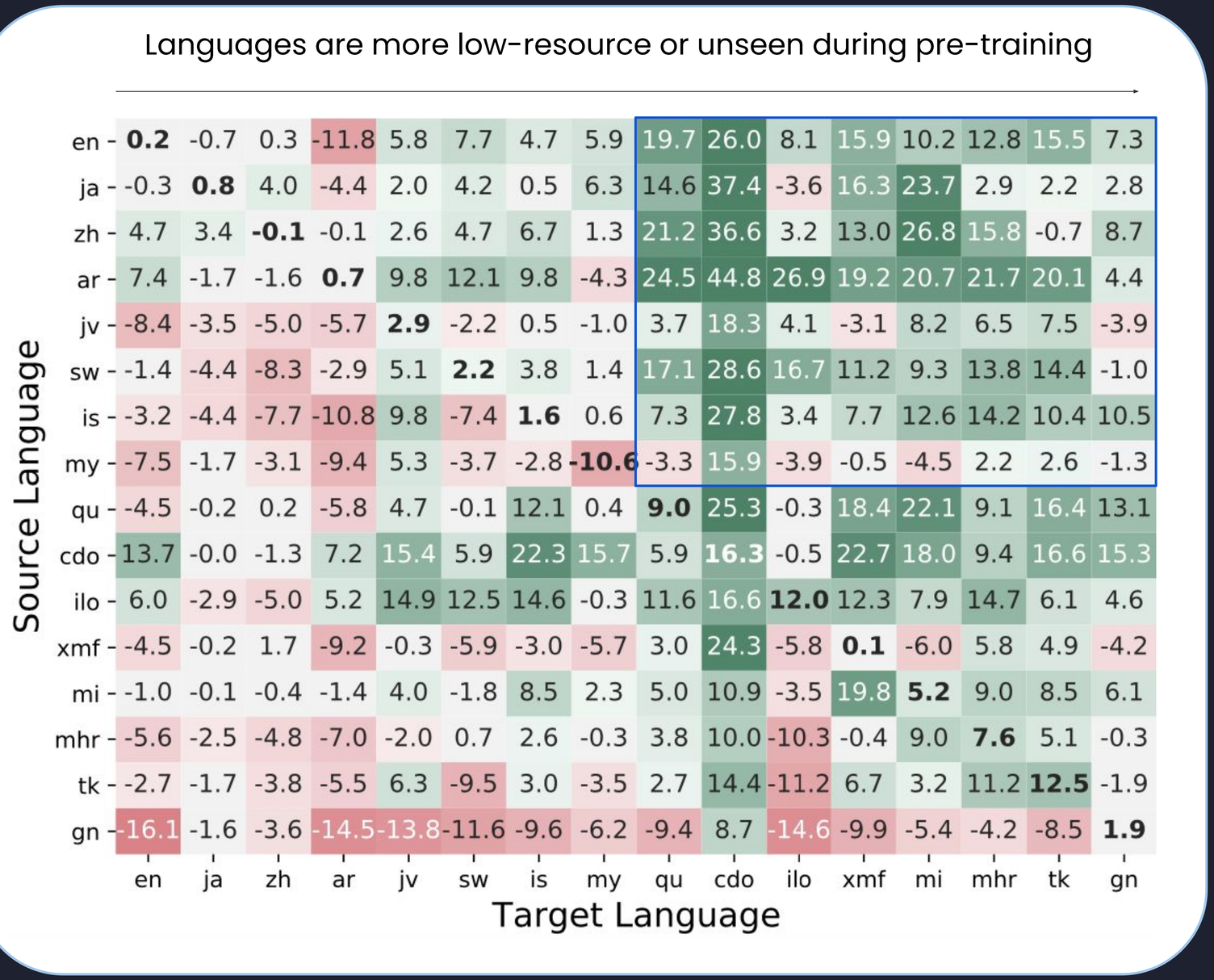
XQuAD (Cross-lingual Question Answering Dataset)

XCOPA (Ponti et al. 2020b)

Language	ISO code	Language family	# of Wiki articles	Covered by SOTA?
English	en	Indo-European	6.0M	✓
Japanese	ja	Japonic	1.2M	✓
Chinese	zh	Sino-Tibetan	1.1M	✓
Arabic	ar	Afro-Asiatic	1.0M	✓
Javanese	jv	Austronesian	57k	✓
Swahili	sw	Niger-Congo	56k	✓
Icelandic	is	Indo-European	49k	✓
Burmese	my	Sino-Tibetan	45k	✓
Quechua	qu	Quechua	22k	
Min Dong	cdo	Sino-Tibetan	15k	
Ilokano	ilo	Austronesian	14k	
Mingrelian	xmf	Kartvelian	13k	
Meadow Mari	mhr	Uralic	10k	
Maori	mi	Austronesian	7k	
Turkmen	tk	Turkic	6k	
Guarani	gn	Tupian	4k	

Relative **F1 improvement** of **MAD-X^{Large} over XLM-R^{Large}** in cross-lingual NER transfer.

Top right corner represent the **realistic scenario** of **transfering from high resource to low resource**



Towards Efficient (and Typologically Driven) Cross-Lingual Transfer

Generating Adapter Parameters

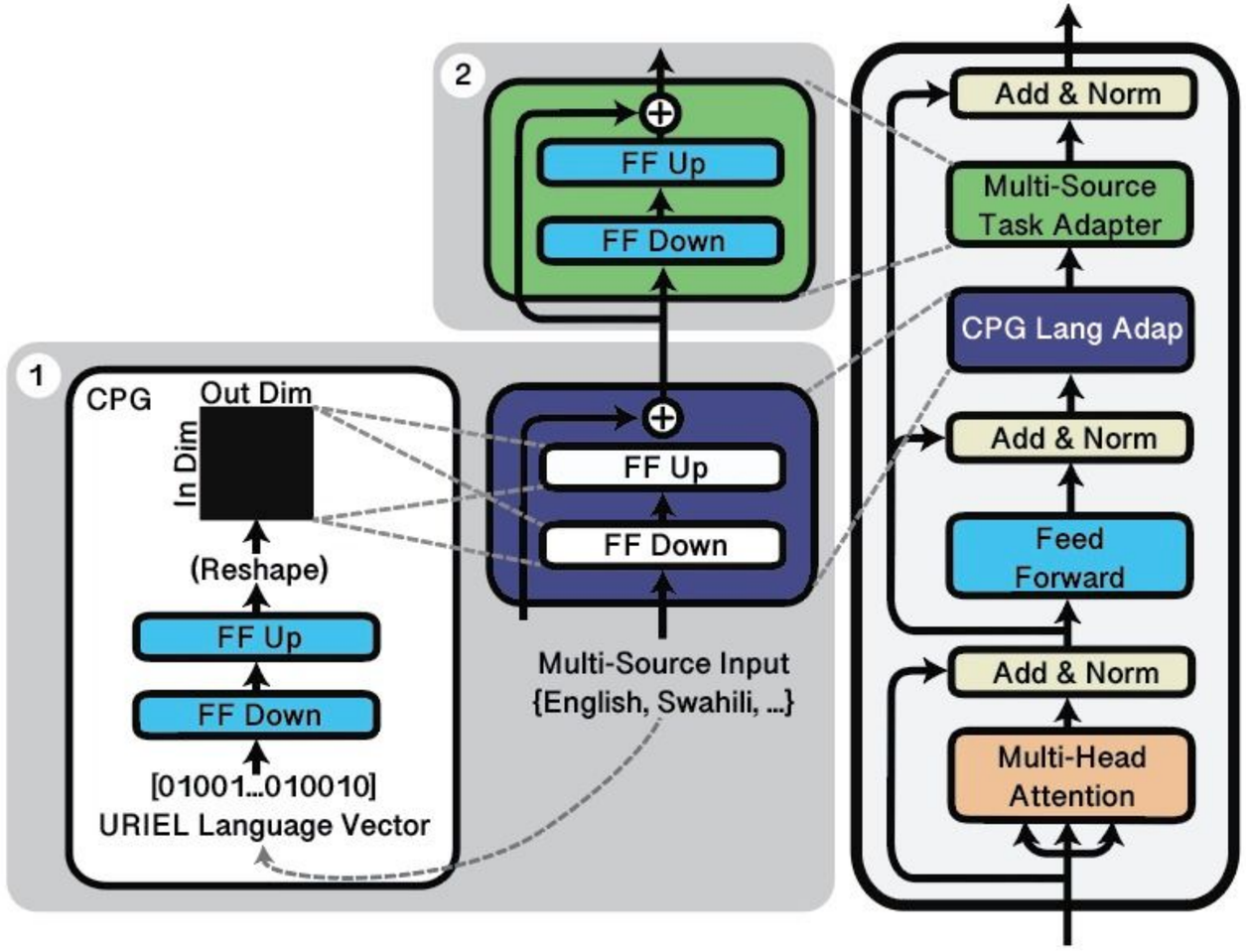
The main idea: instead of having dedicated single-language adapters, can we learn to **generate adapters** on the fly, conditioned on **language vectors?**

This can be seen as **factorising adapters** to **language** (and **layer) parameters**

More efficient than keeping dedicated language adapters

It can even work (in theory) in zero-shot (no text data whatsoever!) and few-shot setups...

Generating Adapter Parameters: *MAD-G*



(Ansell et al., 2021) learn to generate a monolithic multilingual CPG*-adapter by MLM-ing on 95 languages. A simpler and more efficient MAD-X-style transfer learning.

**CPG=Contextual Parameter Generation* (Platanios et al., 2018)

Multi-Source Transfer works much better across different tasks: the model is forced to learn more general language-invariant transfer features?

CPG-Adapters offer better initialisation for further target-specific MLM-ing in low-data scenarios...

MAD-G is efficient...

- Full fine-tuning of mBERT for 95 languages requires:
  95 * 178M = **16.91B** parameters (!!)

- Having MAD-X for 95 languages requires:
  728M additional parameters (95 single-language adapters)

- MAD-G for 95 languages conditioned on language vectors only:
  228M parameters

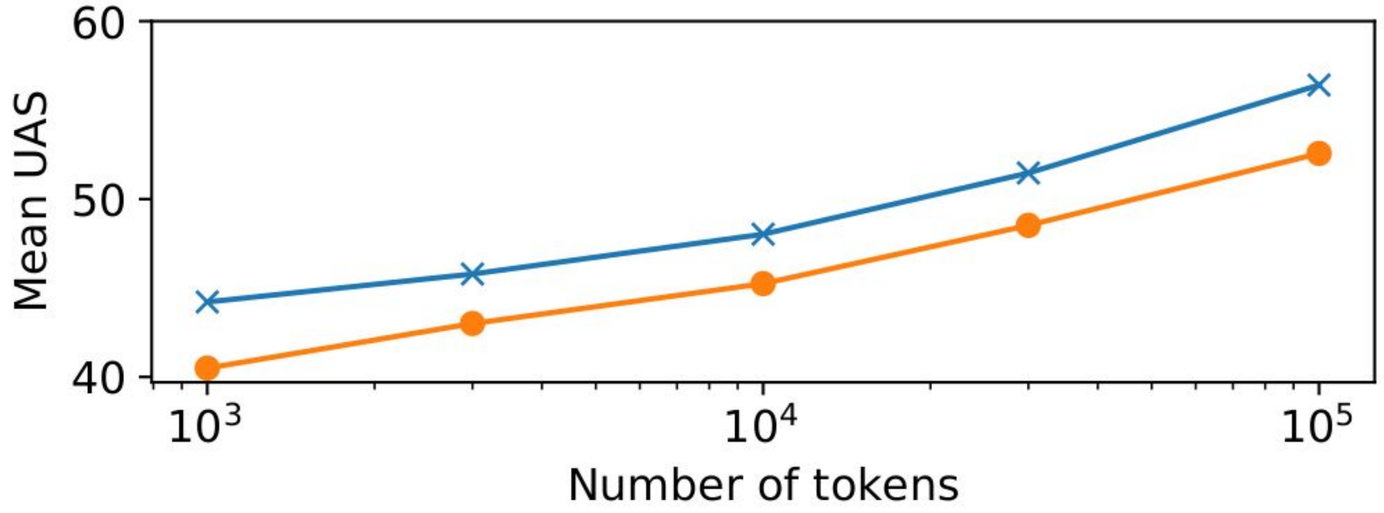
- MAD-G for 95 languages conditioned on language vectors and layer positions:
  38M additional parameters

- Average **per-language training time**: x50 shorter than for MAD-X (II)

...but also effective...

method	hau MAD-G-seen	ibo MAD-G-seen	kin MAD-G-seen	lug unseen	luo unseen	pcm unseen	swa mBERT-seen	wol unseen	yor mBERT-seen	avg.
MAD-G	**77.1**	**69.9**	**66.1**	54.2	32.5	**72.6**	72.6	32.1	**68.8**	**60.7**
MAD-G-LS	72.8	67.5	63.0	**55.7**	**33.3**	72.4	71.3	**36.7**	68.4	60.1
MAD-G-en	44.9	54.5	51.4	50.6	32.9	70.4	69.2	36.4	63.9	52.7
TA-only	43.4	55.7	52.8	47.9	32.8	72.3	68.6	32.1	65.3	52.3
mBERT-ft	43.2	45.5	49.9	49.3	31.6	70.5	65.8	28.1	54.3	48.7
XLM-R-ft	66.4	45.5	36.1	34.8	31.9	68.4	**74.5**	21.6	33.4	45.8

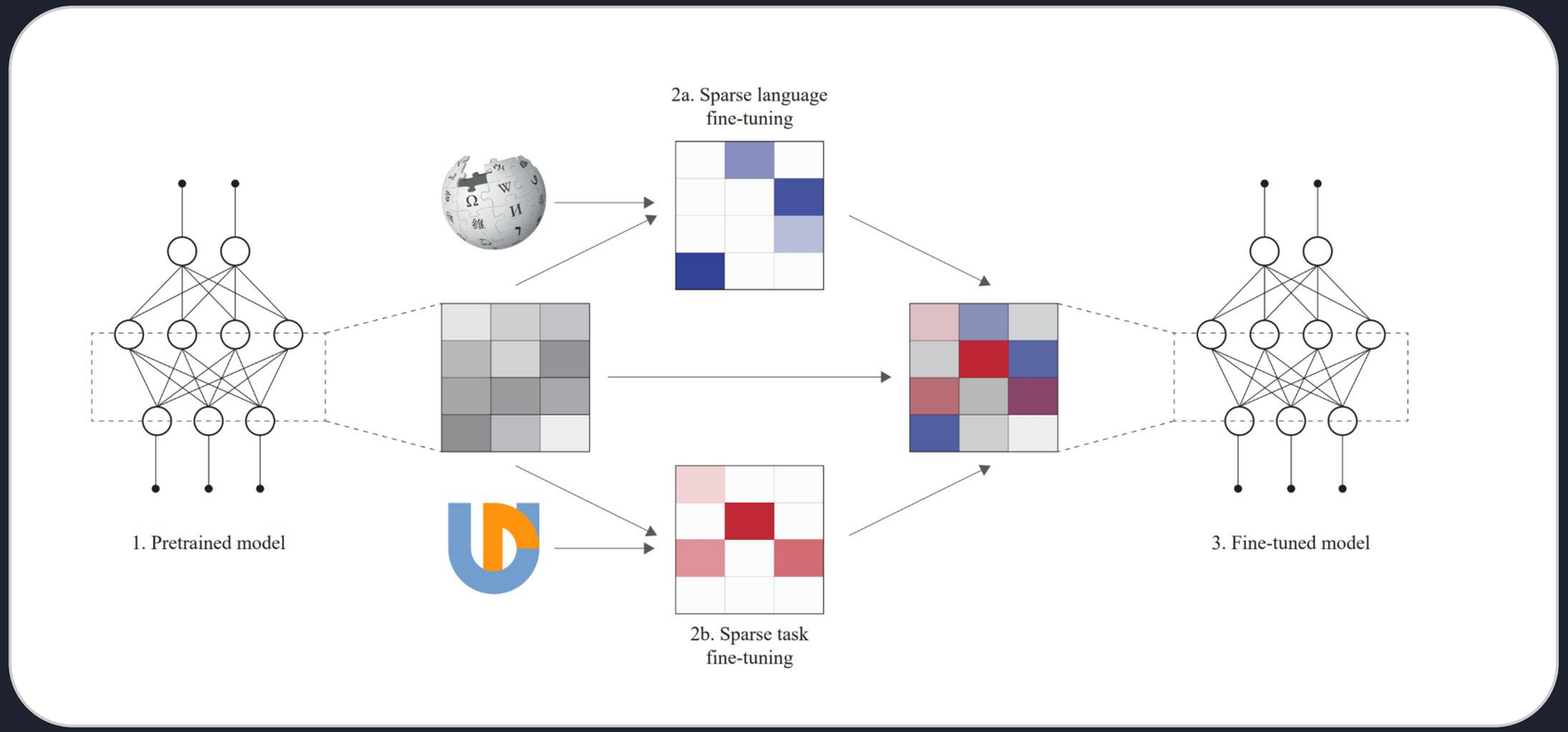
Table 3: F_1 scores on the MasakhaNER dataset for African languages. Task adapter training/model fine-tuning is conducted on the CoNLL 2003 English NER dataset. **XLM-R-ft** results are as reported by Adelani et al. (2021).



We can initialise dedicated language adapters via MAD-G for improved transfer

Another Parameter-Efficient Paradigm

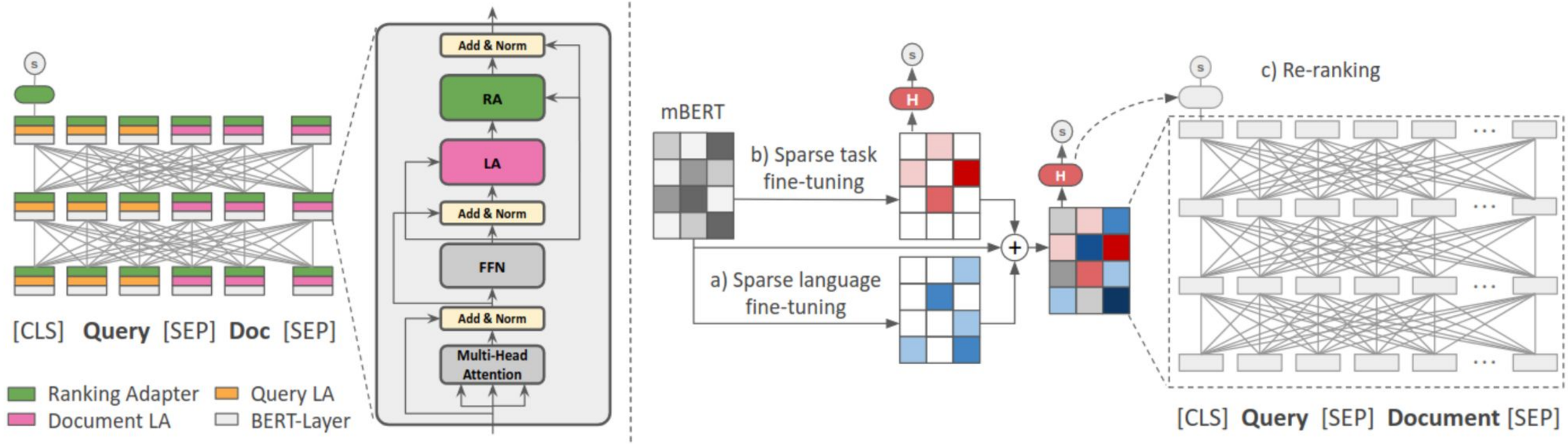
Sparse composable masks for cross-lingual transfer



(Ansell et al., ACL-2022)

Performance gains over MAD-X and MAD-G

Parameter-Efficient Cross-Lingual (Re)ranking



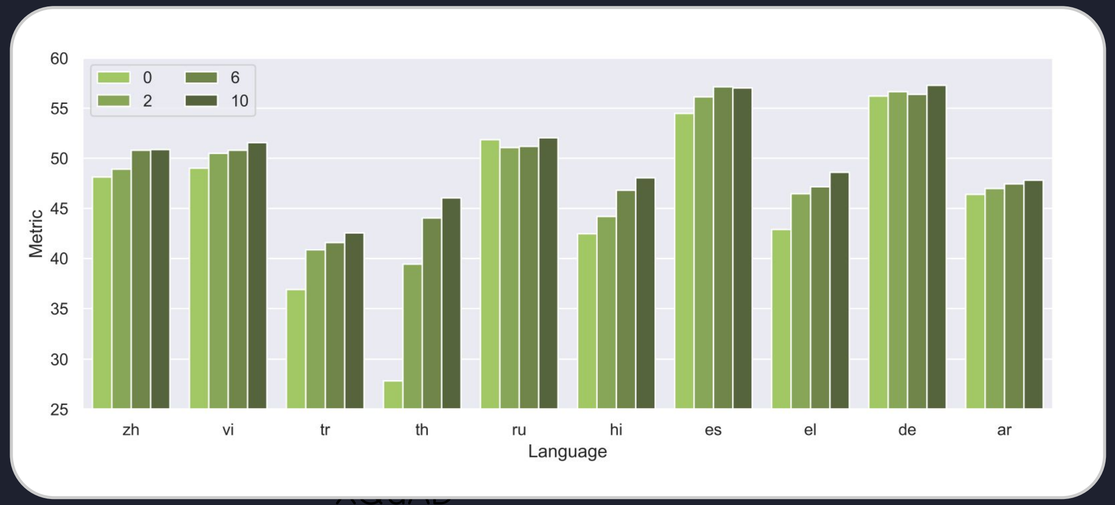
(WIP)

Notes on Zero-Shot versus Few-Shot Learning

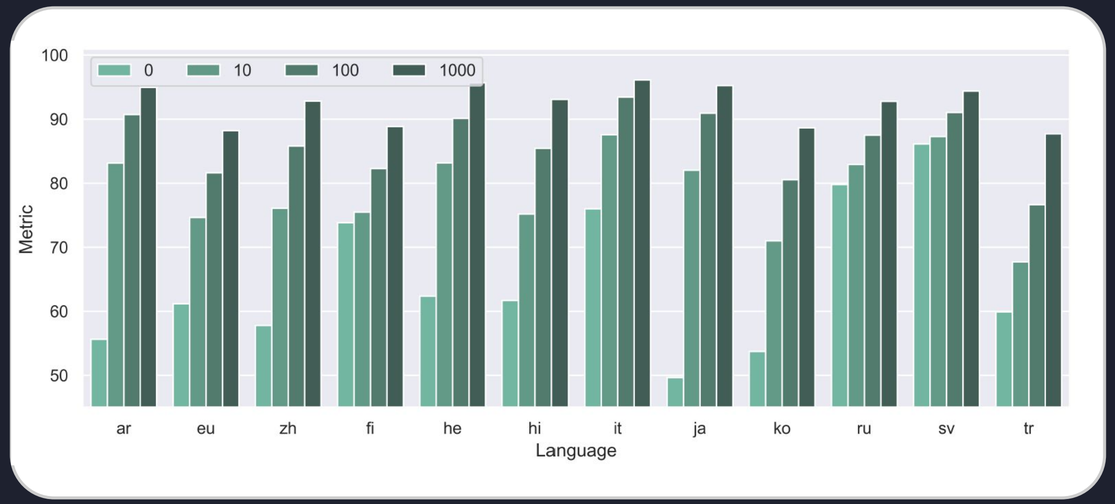
Should we focus more on few-shot transfer scenarios
and quick annotation cycles?

Multilingual Pretraining → Massively Multilingual Transformer → Fine-tuning on target task in English → Few-shot fine-tuning on target task in target language → Prediction on target task in target language

XQuAD



POS



(Lauscher et al., ACL-20; Zhao et al., ACL-21)

Few-Shot > (or >>) Zero-Shot

		K=0	K=1	K=2	K=4	K=8
MLDoc	EN	96.88	-	-	-	-
	DE	88.30	90.36 ± 1.48	90.77 ± 0.87	91.85 ± 0.83	91.98 ± 0.82
	FR	83.05	88.94 ± 2.46	89.71 ± 1.68	90.80 ± 0.88	91.01 ± 0.94
	ES	81.90	83.99 ± 2.35	85.65 ± 1.60	86.30 ± 1.85	88.46 ± 1.90
	IT	74.13	74.97 ± 2.04	75.29 ± 1.57	76.43 ± 1.41	78.12 ± 1.25
	RU	72.33	77.40 ± 4.27	80.57 ± 1.37	81.33 ± 1.33	81.91 ± 1.21
	ZH	84.38	87.18 ± 1.45	87.31 ± 1.53	88.33 ± 1.11	88.72 ± 1.05
	JA	74.58	76.23 ± 1.59	76.71 ± 2.12	78.60 ± 2.43	81.17 ± 1.72
MARC	EN	64.52	-	-	-	-
	DE	49.62	51.50 ± 1.58	52.76 ± 0.87	52.78 ± 1.00	53.32 ± 0.59
	FR	47.30	49.32 ± 1.34	49.70 ± 1.43	50.64 ± 0.94	51.23 ± 0.76
	ES	48.44	49.72 ± 1.24	49.96 ± 1.12	50.45 ± 1.22	51.25 ± 0.93
	ZH	40.40	43.19 ± 1.76	44.45 ± 1.36	45.40 ± 1.26	46.40 ± 0.93
	JA	38.84	41.95 ± 2.09	43.63 ± 1.30	43.98 ± 0.89	44.44 ± 0.69
XNLI	EN	82.67	-	-	-	-
	DE	70.32	70.58 ± 0.36	70.60 ± 0.34	70.61 ± 0.39	70.70 ± 0.50
	FR	73.57	73.41 ± 0.48	73.74 ± 0.46	73.57 ± 0.49	73.77 ± 0.44
	ES	73.71	73.84 ± 0.40	73.87 ± 0.44	73.74 ± 0.48	73.87 ± 0.46
	RU	68.70	68.81 ± 0.52	68.76 ± 0.54	68.87 ± 0.55	68.81 ± 0.77
	ZH	69.32	69.73 ± 0.94	69.75 ± 0.94	70.56 ± 0.76	70.62 ± 0.86
	AR	64.97	64.75 ± 0.36	64.82 ± 0.23	64.82 ± 0.23	64.94 ± 0.37
	BG	67.58	68.15 ± 0.69	68.19 ± 0.75	68.55 ± 0.67	68.32 ± 0.70
	EL	65.67	65.64 ± 0.40	65.73 ± 0.36	65.80 ± 0.41	66.00 ± 0.53
	HI	56.57	56.94 ± 0.82	57.07 ± 0.82	57.21 ± 1.14	57.82 ± 1.18
	SW	48.08	50.33 ± 1.08	50.28 ± 1.24	51.08 ± 0.62	51.01 ± 0.79
	TH	46.17	49.43 ± 2.60	50.08 ± 2.42	51.32 ± 2.07	52.16 ± 2.43
	TR	60.40	61.02 ± 0.68	61.20 ± 0.61	61.35 ± 0.49	61.31 ± 0.56
	UR	57.05	57.56 ± 0.85	57.83 ± 0.91	58.20 ± 0.93	58.67 ± 1.03
	VI	69.82	70.04 ± 0.59	70.14 ± 0.75	70.23 ± 0.63	70.41 ± 0.70
PAWSX	EN	93.90	-	-	-	-
	DE	83.80	84.14 ± 0.40	84.08 ± 0.42	84.04 ± 0.47	84.23 ± 0.66
	FR	86.90	87.07 ± 0.27	87.06 ± 0.37	87.03 ± 0.31	86.94 ± 0.41
	ES	88.25	87.90 ± 0.54	87.80 ± 0.56	87.84 ± 0.53	87.85 ± 0.75
	ZH	77.75	77.71 ± 0.37	77.63 ± 0.47	77.68 ± 0.51	77.82 ± 0.64
	JA	73.30	73.78 ± 0.75	73.71 ± 1.04	73.48 ± 0.69	73.79 ± 1.28
	KO	72.05	73.75 ± 1.30	73.11 ± 1.05	73.79 ± 0.92	73.31 ± 0.61

(...not only for token-level tasks)

Source Fine-Tuning Helps

	MLDoc		PAWSX			POS		NER	
	K=1	K=8	K=1	K=8		K=1	K=4	K=1	K=4
DE	-37.73	-7.67	-31.11	-30.82	RU	-15.89	-3.20	-48.19	-35.77
FR	-38.14	-13.21	-33.02	-32.34	ES	-9.51	-0.93	-63.98	-41.53
ES	-33.69	-14.38	-33.76	-33.97	VI	-7.82	-0.36	-54.41	-41.45
IT	-33.63	-12.62	-	-	TR	-15.05	-8.08	-54.35	-34.52
RU	-30.66	-11.08	-	-	TA	-13.72	-4.40	-34.70	-24.81
ZH	-37.31	-12.57	-23.74	-23.65	MR	-11.34	-3.63	-40.10	-25.68
JA	-29.82	-14.32	-20.97	-20.82	-	-	-	-	-
KO	-	-	-19.83	-19.68	-	-	-	-	-

Huge drops without source-language fine-tuning, using only target-language shots

Random Fundamental Problems with Data

We still do not have good (even) evaluation data for many tasks

- If we are working with low-resource languages, we can evaluate only on “lower-level” tasks such as POS tagging, NER, parsing, sentence matching…
- New benchmarking initiatives such as XTREME(-R) and XGLUE help…
- New datasets (e.g., NER data for 10 African languages; NLI dataset for 10 languages of the Americas; a renewed TREC interest in CLIR) help…

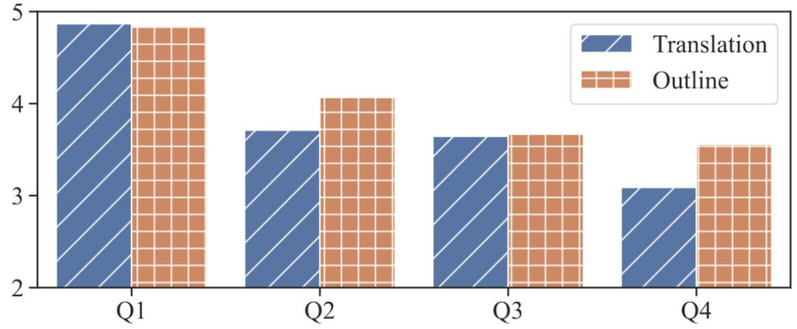
Discussion Point: Can We Make Better Training and Evaluation Data for Low-Resource Languages?
(A collective bottom-up effort?)

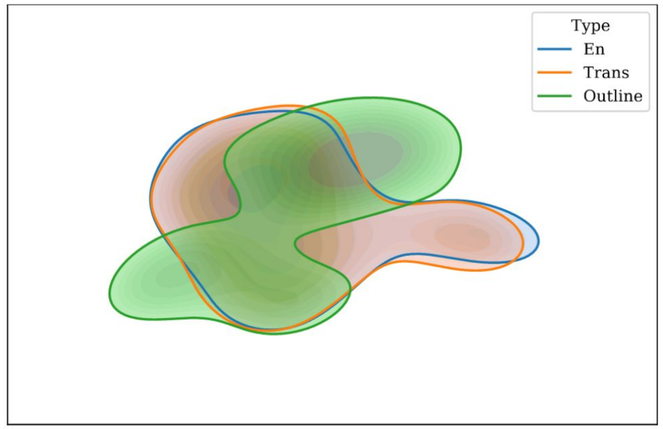
Random Fundamental Problems with Data

Translation-based data creation? Data “imperialism”? (Bird, 2020)

A curious study of dataset creation for task-oriented dialogue (Majewska et al., 2022)

Instructions
Please state to what extent you agree/disagree with each statement on the scale of 1-5 (1-Strongly disagree, 5-Strongly agree)
Questions
Q1. The ASSISTANT helps satisfy the USER’s requests. Q2. The USER speaks naturally and sounds like a Russian native speaker. Q3. The ASSISTANT speaks naturally and sounds like a Russian native speaker. Q4. I can easily imagine myself mentioning or hearing the proper names referred to in the dialogue (e.g., titles of films or songs, people, places) in a conversation with my Russian friends or family.





Take-Home Messages: Episode I
(We only scratched the surface in this talk…)

- Multilingual and cross-lingual NLP and IR are vibrant research fields in the mission of democratising language technology

- Too many domains, too many languages, dialects -> we need general and adaptable solutions, we need to learn from whatever we've got…

- We have covered (at a very shallow level) high-level approaches as well as lower-level cutting-edge approaches to multilingual and cross-lingual NLP
  - cross-lingual word embeddings
  - massively multilingual language models
  - multilingual representation learning; cross-lingual transfer methods
  - some more advanced topics: modular and parameter-efficient transfer, few-shot learning

- Despite positive trends, many languages are still left behind (and difficult to work with)

Advanced Topics
(We only scratched the surface in this talk…)

- Active learning
- Meta-learning and few-shot adaptation strategies
- Data annotation and resource creation in low-resource languages
- Model adaptation to languages with unseen scripts
- Induction of linguistic structure from pretrained multilingual LMs
- Semantic specialisation of general-purpose models
- Learning multilingual word, sentence, and document encoders
- Unsupervised and weakly supervised Neural Machine Translation
- Injection of linguistic and world knowledge into multilingual text-based models
- Multi-modal multilingual modeling
- Creative applications of multilingual models
- Multilingual speech recognition
- Speech translation

Multilingual and Cross-Lingual NLP and IR: How to Cope?

Better Models and Algorithms:

- sophisticated modeling/training methods - know NLP/ML
- linguistically informed methods - know linguistics
- task knowledge - know your task

Better Data and Evaluation:

- every piece of relevant data can help - be resourceful
- make data if necessary - be connected
- track progress with challenging (and natural!) evaluation data

Better Adaptation:

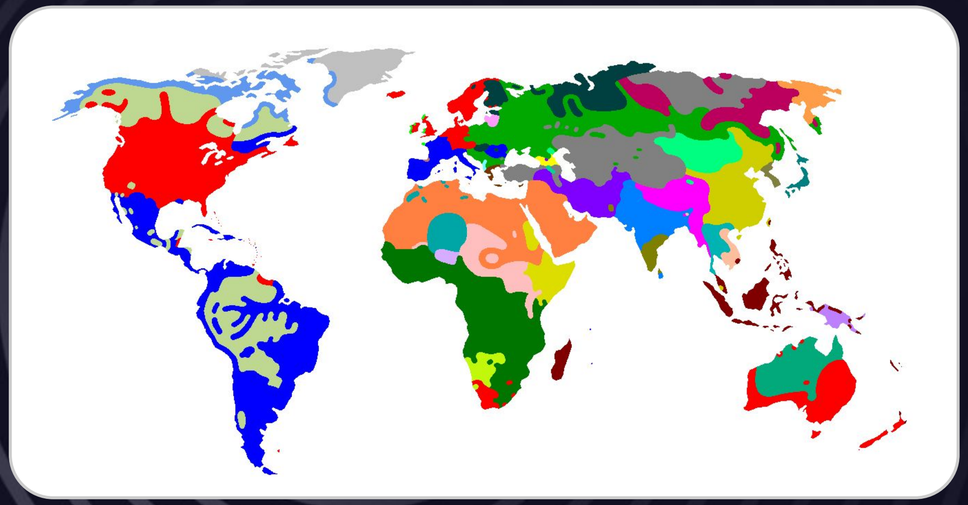
- leverage similarity between languages
- adapt quickly to low-data regimes and new domains

(Multilingual) Natural Language Processing

…and its many applications

Conversational Systems Virtual Assistants Information Search Question Answering

Digital Education Language Learning Assisted Translation Fact Checking Verification



The grand challenge of multilinguality

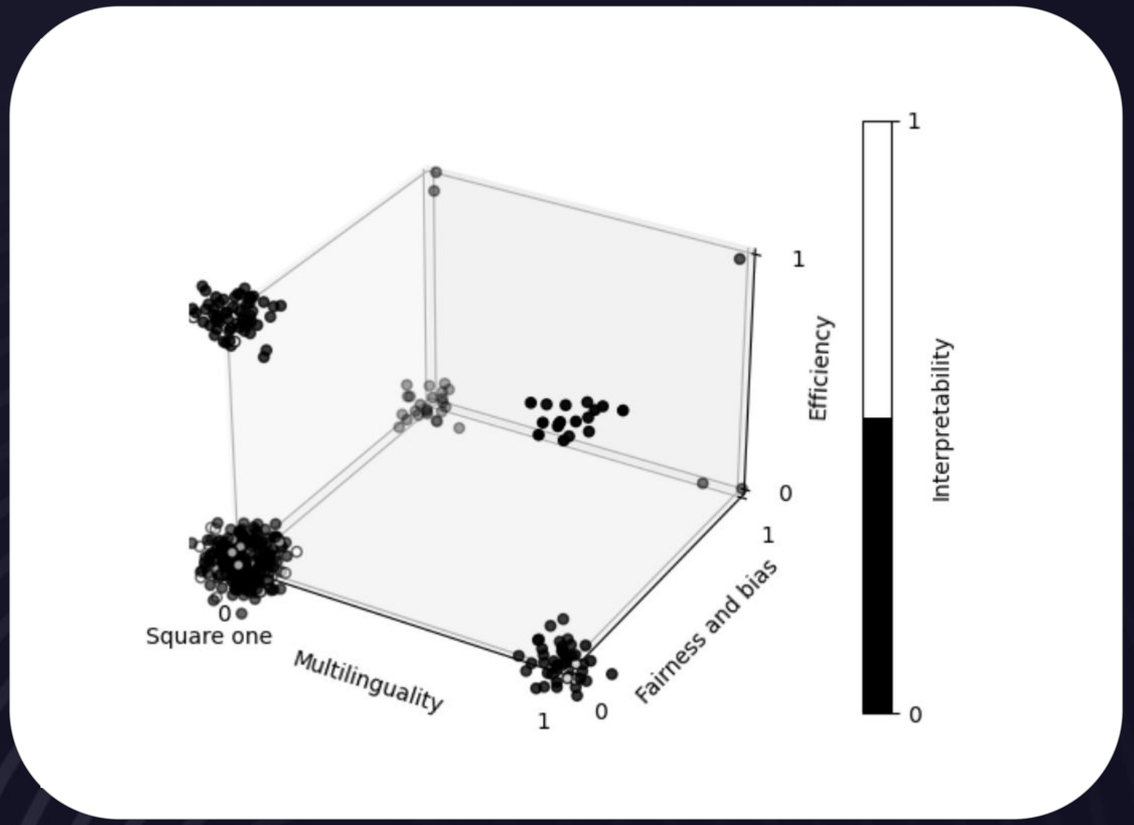
Digital language divide versus equal opportunities

7,000+ languages; a wide spectrum of tasks and domains

- High-resource languages
- Medium-resource
- Low-resource
- Endangered

Towards Inclusive, Sustainable, Equitable Multilingual NLP

Widening the global reach of NLP: Far-reaching technological and socioeconomic consequences



An analysis of ACL 2021 papers in a “research meta-space”

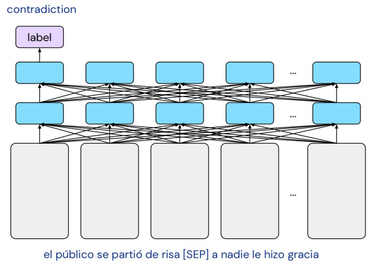
Søgaard, Vulić, Ruder: Square One Bias in NLP: Towards a Multi-Dimensional Exploration of the Research Manifold (ACL 2022)

- Are we too obsessed with task performance only and leaderboards?
- **Towards a more holistic bottom-up approach to multilingual language technology**

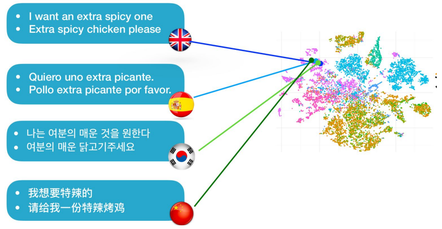
Towards Inclusive, Sustainable, Equitable Multilingual NLP

Widening the global reach of NLP: Far-reaching technological and socioeconomic consequences

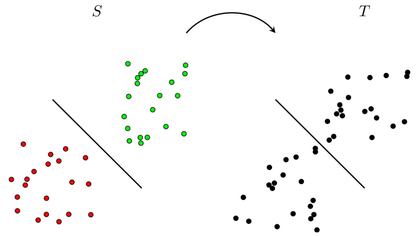
Deep (machine) learning



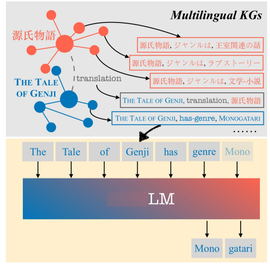
Multilingual representation learning



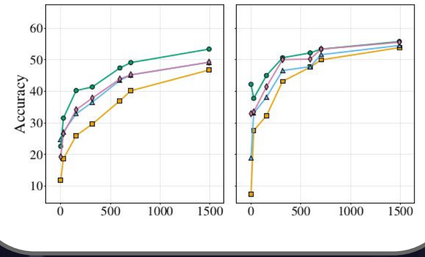
Cross-lingual knowledge transfer



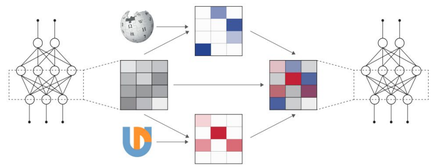
External Knowledge



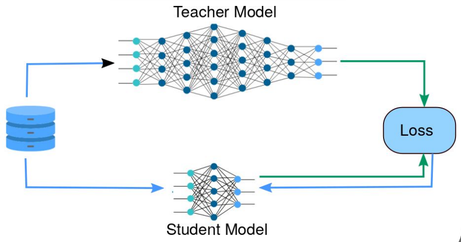
Sample Efficiency



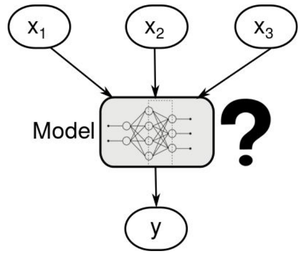
Modularity



Model Compactness



Explainability



Other Aspects of Inclusivity and Equity: *Fairness, Cross-Cultural Adaptation, Multi-Modal Learning*

Featuring:

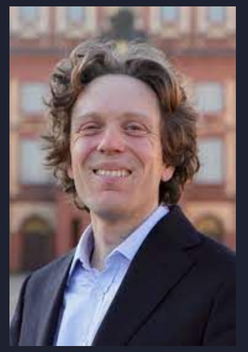

















iv250@cam.ac.uk

Once again, big big thankyous and credits to my collaborators...