Multilingual Models for Natural Language Processing

Anoop Kunchukuttan

Microsoft Translator

AI4Bharat

Winter School on Deep Learning for Vision and Language Modelling
IIT Guwahati
January 12, 2025

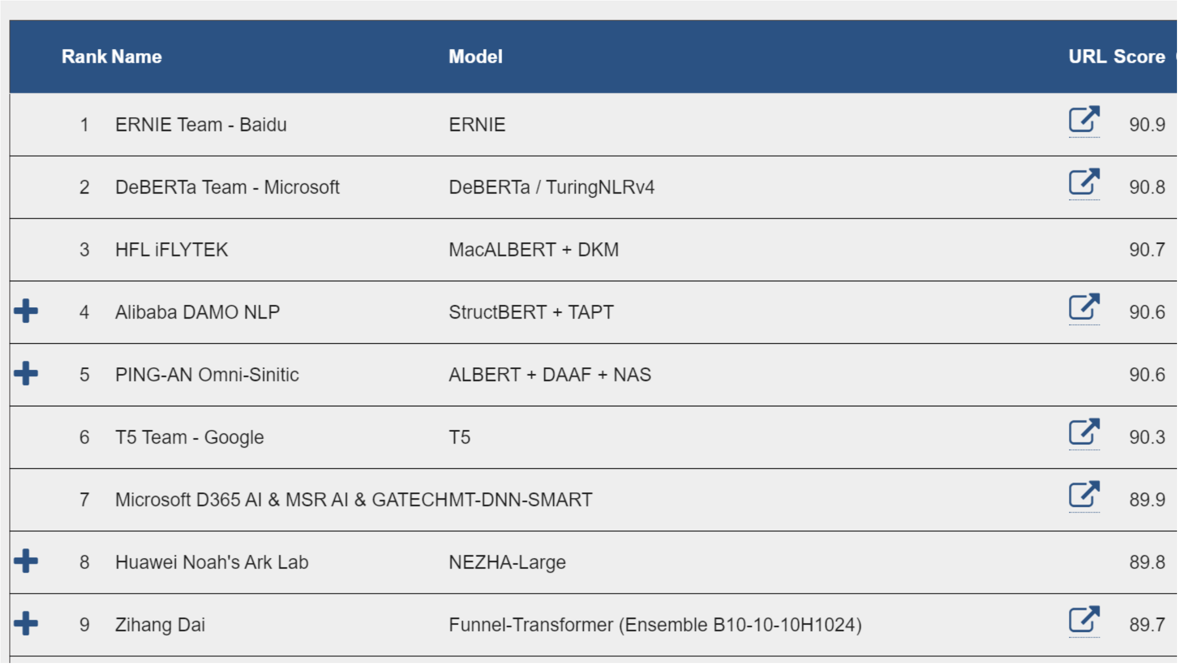
Agenda

- Introduction
- Typical Multilingual NLP Pipeline
- Cross-lingual Transfer
- Multilingual Word Embeddings
- Multilingual Pre-trained Language Models
  - Extending English LLMs to new languages
- Finetuning Pre-trained LMs
- Evaluation of Multilingual models

Deep Learning has led to tremendous success

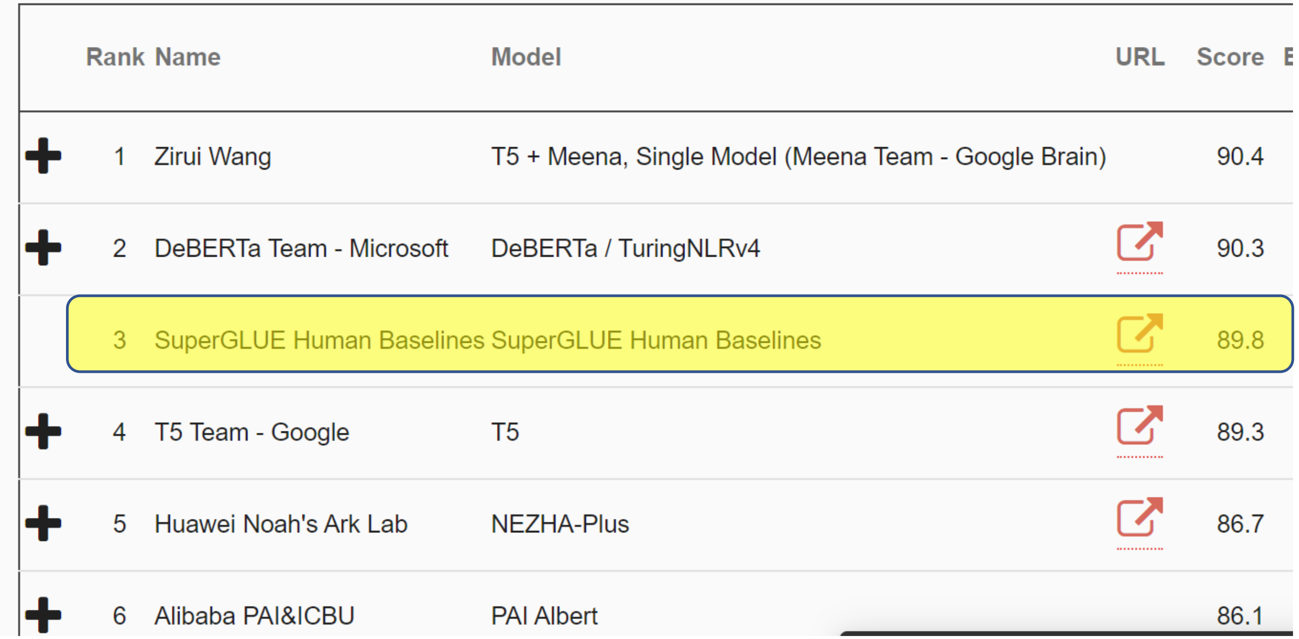
Transformer-based self-supervised pre-trained models

GLUE Benchmark



	Rank	Name	Model	URL	Score
	1	ERNIE Team - Baidu	ERNIE	↗	90.9
	2	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4	↗	90.8
	3	HFL iFLYTEK	MacALBERT + DKM		90.7
+	4	Alibaba DAMO NLP	StructBERT + TAPT	↗	90.6
+	5	PING-AN Omni-Sinitic	ALBERT + DAAF + NAS		90.6
	6	T5 Team - Google	T5	↗	90.3
	7	Microsoft D365 AI & MSR AI & GATECH	MT-DNN-SMART	↗	89.9
+	8	Huawei Noah's Ark Lab	NEZHA-Large		89.8
+	9	Zihang Dai	Funnel-Transformer (Ensemble B10-10-10H1024)	↗	89.7

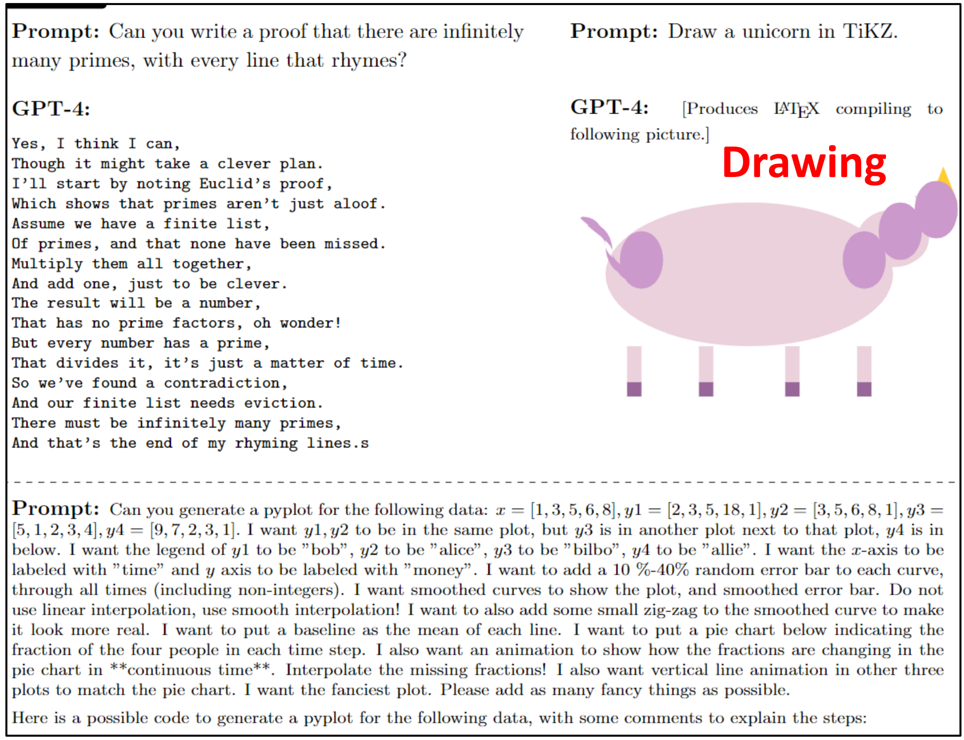
SUPERGLUE Benchmark



	Rank	Name	Model	URL	Score
+	1	Zirui Wang	T5 + Meena, Single Model (Meena Team - Google Brain)		90.4
+	2	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4	↗	90.3
	3	SuperGLUE Human Baselines	SuperGLUE Human Baselines	↗	89.8
+	4	T5 Team - Google	T5	↗	89.3
+	5	Huawei Noah's Ark Lab	NEZHA-Plus	↗	86.7
+	6	Alibaba PAI&ICBU	PAI Albert		86.1

***Tasks:* Classification, Structured Prediction, Span Prediction, Semantic Similarity**

Maths



Prompt: Can you write a proof that there are infinitely many primes, with every line that rhymes?

GPT-4:

```
Yes, I think I can,
Though it might take a clever plan.
I'll start by noting Euclid's proof,
Which shows that primes aren't just aloof.
Assume we have a finite list,
Of primes, and that none have been missed.
Multiply them all together,
And add one, just to be clever.
The result will be a number,
That has no prime factors, oh wonder!
But every number has a prime,
That divides it, it's just a matter of time.
So we've found a contradiction,
And our finite list needs eviction.
There must be infinitely many primes,
And that's the end of my rhyming lines.s
```

Prompt: Draw a unicorn in TiKZ.

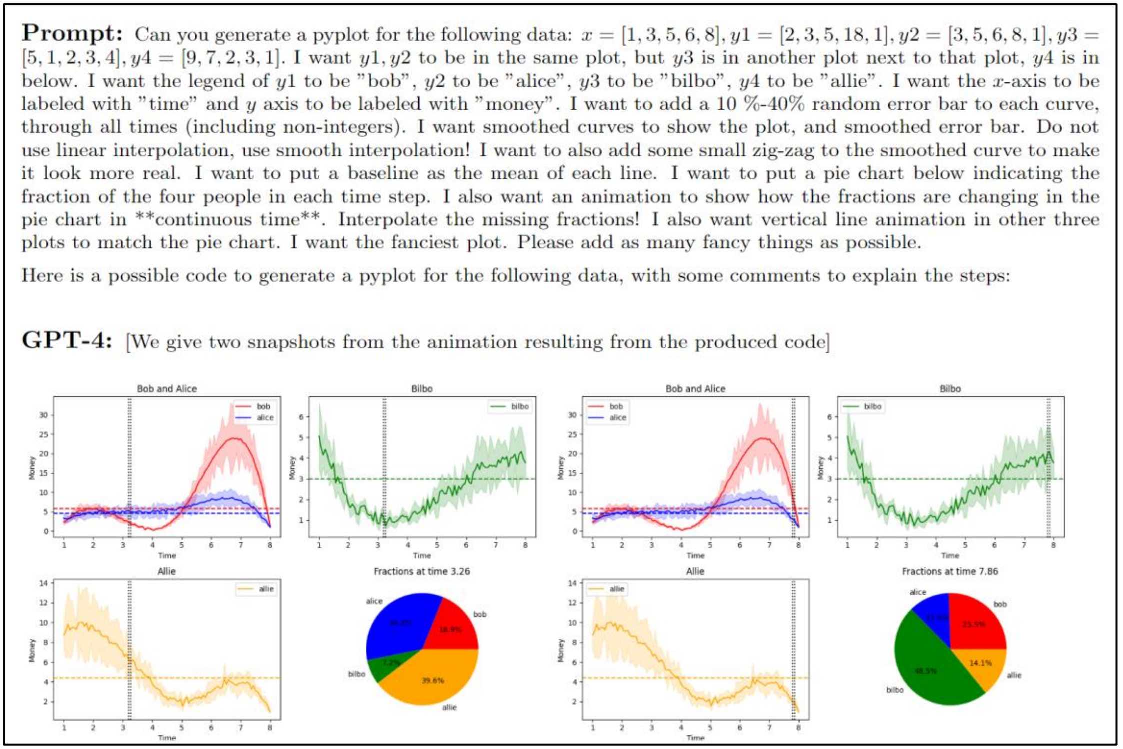
GPT-4: [Produces LaTeX compiling to following picture.]

Drawing

Prompt: Can you generate a pyplot for the following data: $x = [1, 3, 5, 6, 8]$, $y1 = [2, 3, 5, 18, 1]$, $y2 = [3, 5, 6, 8, 1]$, $y3 = [5, 1, 2, 3, 4]$, $y4 = [9, 7, 2, 3, 1]$. I want $y1, y2$ to be in the same plot, but $y3$ is in another plot next to that plot, $y4$ is in below. I want the legend of $y1$ to be "bob", $y2$ to be "alice", $y3$ to be "bilbo", $y4$ to be "allie". I want the x-axis to be labeled with "time" and y axis to be labeled with "money". I want to add a 10 %-40% random error bar to each curve, through all times (including non-integers). I want smoothed curves to show the plot, and smoothed error bar. Do not use linear interpolation, use smooth interpolation! I want to also add some small zig-zag to the smoothed curve to make it look more real. I want to put a baseline as the mean of each line. I want to put a pie chart below indicating the fraction of the four people in each time step. I also want an animation to show how the fractions are changing in the pie chart in **continuous time**. Interpolate the missing fractions! I also want vertical line animation in other three plots to match the pie chart. I want the fanciest plot. Please add as many fancy things as possible.

Here is a possible code to generate a pyplot for the following data, with some comments to explain the steps:

Plotting

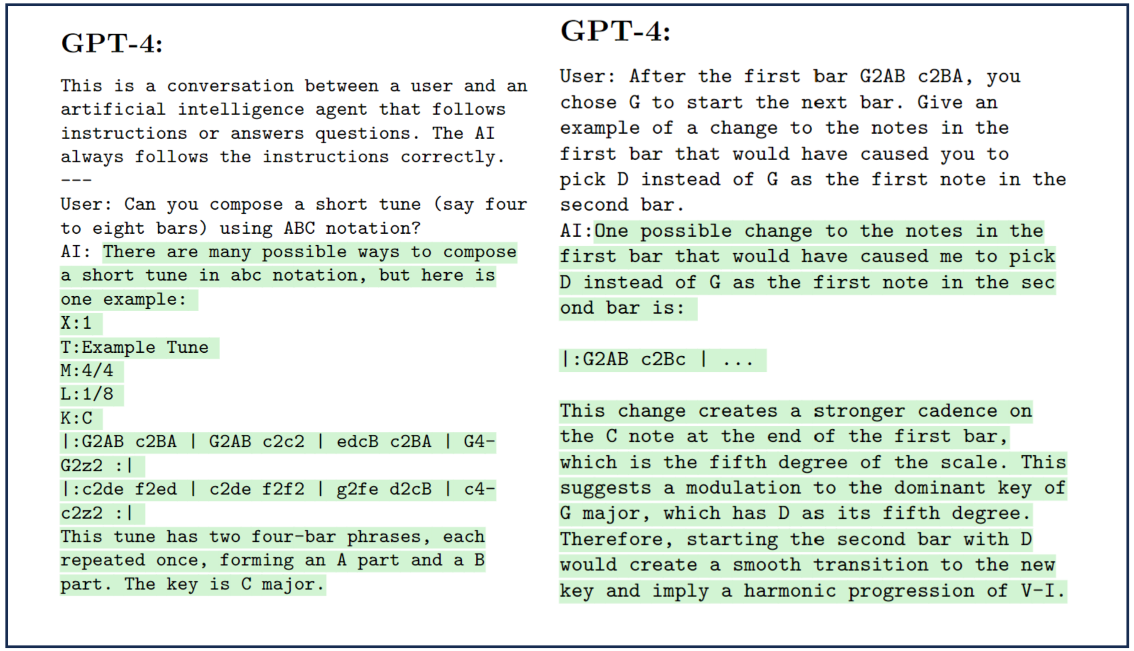


Prompt: Can you generate a pyplot for the following data: $x = [1, 3, 5, 6, 8]$, $y1 = [2, 3, 5, 18, 1]$, $y2 = [3, 5, 6, 8, 1]$, $y3 = [5, 1, 2, 3, 4]$, $y4 = [9, 7, 2, 3, 1]$. I want $y1, y2$ to be in the same plot, but $y3$ is in another plot next to that plot, $y4$ is in below. I want the legend of $y1$ to be "bob", $y2$ to be "alice", $y3$ to be "bilbo", $y4$ to be "allie". I want the x-axis to be labeled with "time" and y axis to be labeled with "money". I want to add a 10 %-40% random error bar to each curve, through all times (including non-integers). I want smoothed curves to show the plot, and smoothed error bar. Do not use linear interpolation, use smooth interpolation! I want to also add some small zig-zag to the smoothed curve to make it look more real. I want to put a baseline as the mean of each line. I want to put a pie chart below indicating the fraction of the four people in each time step. I also want an animation to show how the fractions are changing in the pie chart in **continuous time**. Interpolate the missing fractions! I also want vertical line animation in other three plots to match the pie chart. I want the fanciest plot. Please add as many fancy things as possible.

Here is a possible code to generate a pyplot for the following data, with some comments to explain the steps:

GPT-4: [We give two snapshots from the animation resulting from the produced code]

Explanations



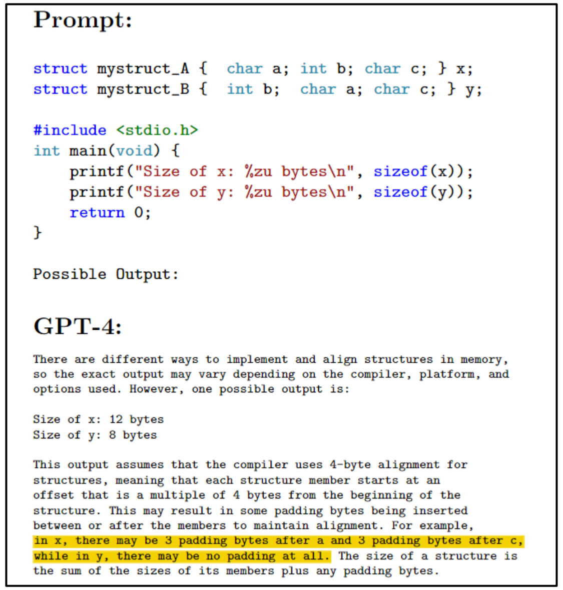
GPT-4:

```
This is a conversation between a user and an
artificial intelligence agent that follows
instructions or answers questions. The AI
always follows the instructions correctly.
---
User: Can you compose a short tune (say four
to eight bars) using ABC notation?
AI: There are many possible ways to compose
a short tune in abc notation, but here is
one example:
X:1
T:Example Tune
M:4/4
L:1/8
K:C
|:G2AB c2BA | G2AB c2c2 | edcB c2BA | G4-
G2z2 :|
|:c2de f2ed | c2de f2f2 | g2fe d2cB | c4-
c2z2 :|
This tune has two four-bar phrases, each
repeated once, forming an A part and a B
part. The key is C major.
```

GPT-4:

```
User: After the first bar G2AB c2BA, you
chose G to start the next bar. Give an
example of a change to the notes in the
first bar that would have caused you to
pick D instead of G as the first note in the
second bar.
AI:One possible change to the notes in the
first bar that would have caused me to pick
D instead of G as the first note in the sec
ond bar is:

|:G2AB c2Bc | ...

This change creates a stronger cadence on
the C note at the end of the first bar,
which is the fifth degree of the scale. This
suggests a modulation to the dominant key of
G major, which has D as its fifth degree.
Therefore, starting the second bar with D
would create a smooth transition to the new
key and imply a harmonic progression of V-I.
```

Understanding Programs



Prompt:

```
struct mystruct_A { char a; int b; char c; } x;
struct mystruct_B { int b;  char a; char c; } y;

#include <stdio.h>
int main(void) {
    printf("Size of x: %zu bytes\n", sizeof(x));
    printf("Size of y: %zu bytes\n", sizeof(y));
    return 0;
}

Possible Output:
```

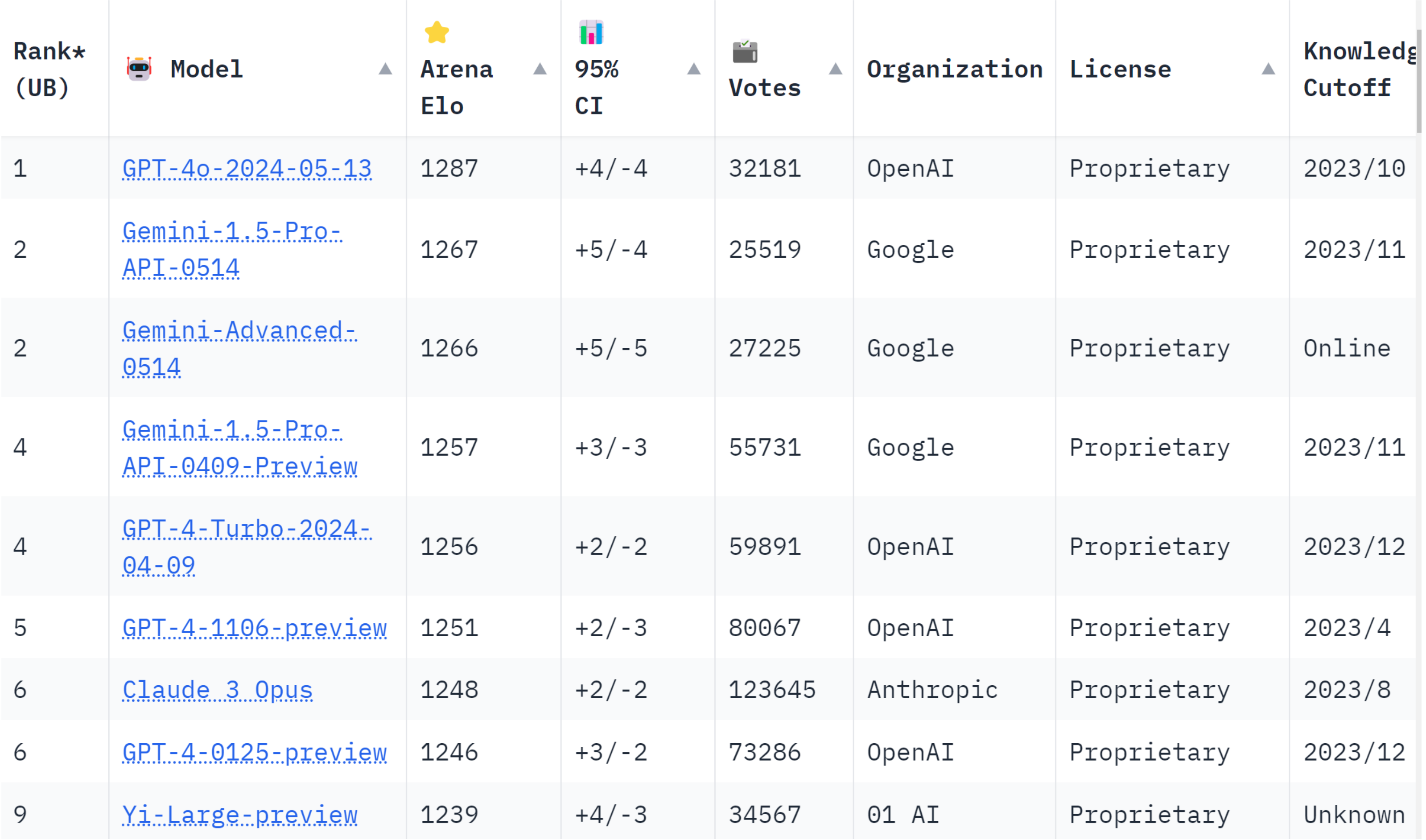
GPT-4:

```
There are different ways to implement and align structures in memory,
so the exact output may vary depending on the compiler, platform, and
options used. However, one possible output is:

Size of x: 12 bytes
Size of y: 8 bytes

This output assumes that the compiler uses 4-byte alignment for
structures, meaning that each structure member starts at an
offset that is a multiple of 4 bytes from the beginning of the
structure. This may result in some padding bytes being inserted
between or after the members to maintain alignment. For example,
in x, there may be 3 padding bytes after a and 3 padding bytes after c,
while in y, there may be no padding at all. The size of a structure is
the sum of the sizes of its members plus any padding bytes.
```

(SparksOfAGI)

Large Language Models show great performance on diverse open-ended tasks

Transformer-based self-supervised decoder only models



Rank* (UB)	Model	Arena Elo	95% CI	Votes	Organization	License	Knowledg Cutoff
1	GPT-4o-2024-05-13	1287	+4/-4	32181	OpenAI	Proprietary	2023/10
2	Gemini-1.5-Pro-API-0514	1267	+5/-4	25519	Google	Proprietary	2023/11
2	Gemini-Advanced-0514	1266	+5/-5	27225	Google	Proprietary	Online
4	Gemini-1.5-Pro-API-0409-Preview	1257	+3/-3	55731	Google	Proprietary	2023/11
4	GPT-4-Turbo-2024-04-09	1256	+2/-2	59891	OpenAI	Proprietary	2023/12
5	GPT-4-1106-preview	1251	+2/-3	80067	OpenAI	Proprietary	2023/4
6	Claude 3 Opus	1248	+2/-2	123645	Anthropic	Proprietary	2023/8
6	GPT-4-0125-preview	1246	+3/-2	73286	OpenAI	Proprietary	2023/12
9	Yi-Large-preview	1239	+4/-3	34567	01 AI	Proprietary	Unknown

(LMSys Chatbot Arena)

Tasks: Open-ended Question Answering evaluated on dynamic questions based on human preferences

What is the recipe behind the success of English NLP?

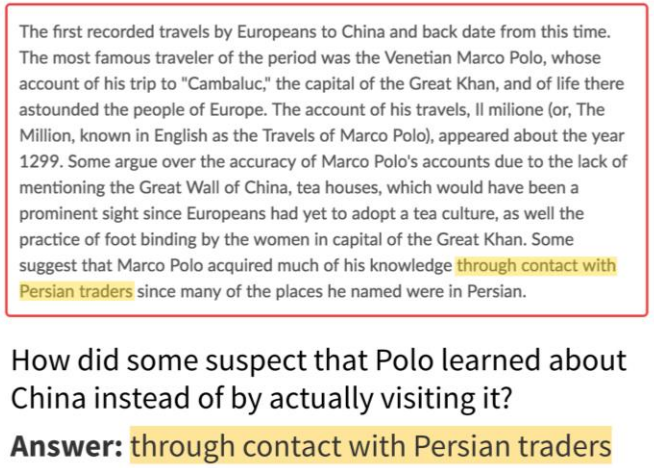
Collect a large amount of textual data

Lot of monolingual data

BERT → encoder
BART → encoder-decoder
GPT → decoder

*Learn a **language model** using **very large neural networks** trained for a **long time***

Lot of memory & computation



The first recorded travels by Europeans to China and back date from this time. The most famous traveler of the period was the Venetian Marco Polo, whose account of his trip to "Cambaluc," the capital of the Great Khan, and of life there astounded the people of Europe. The account of his travels, Il milione (or, The Million, known in English as the Travels of Marco Polo), appeared about the year 1299. Some argue over the accuracy of Marco Polo's accounts due to the lack of mentioning the Great Wall of China, tea houses, which would have been a prominent sight since Europeans had yet to adopt a tea culture, as well the practice of foot binding by the women in capital of the Great Khan. Some suggest that Marco Polo acquired much of his knowledge through contact with Persian traders since many of the places he named were in Persian.

How did some suspect that Polo learned about China instead of by actually visiting it?

Answer: through contact with Persian traders

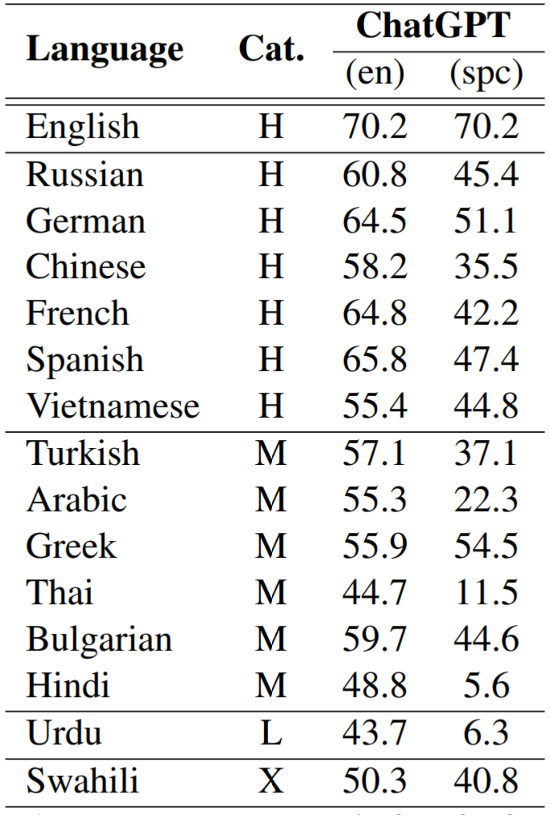
Finetune the language model on reasonable amount of data

- *Task-specific*
- *General instruction-response*
- *Human Preference*

In-context learning abilities in LLMs

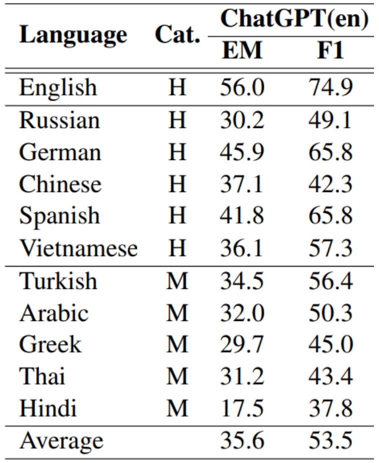
Lot of supervised data

Benefits of LLMs are mostly limited to English



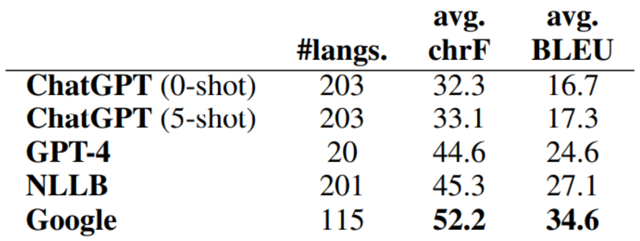
Language	Cat.	ChatGPT	
		(en)	(spc)
English	H	70.2	70.2
Russian	H	60.8	45.4
German	H	64.5	51.1
Chinese	H	58.2	35.5
French	H	64.8	42.2
Spanish	H	65.8	47.4
Vietnamese	H	55.4	44.8
Turkish	M	57.1	37.1
Arabic	M	55.3	22.3
Greek	M	55.9	54.5
Thai	M	44.7	11.5
Bulgarian	M	59.7	44.6
Hindi	M	48.8	5.6
Urdu	L	43.7	6.3
Swahili	X	50.3	40.8

Results on XNLI



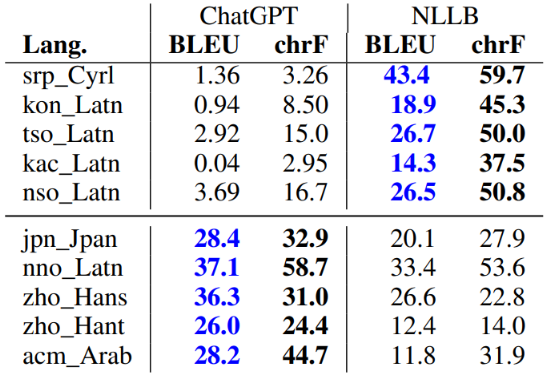
Language	Cat.	ChatGPT(en)	
		EM	F1
English	H	56.0	74.9
Russian	H	30.2	49.1
German	H	45.9	65.8
Chinese	H	37.1	42.3
Spanish	H	41.8	65.8
Vietnamese	H	36.1	57.3
Turkish	M	34.5	56.4
Arabic	M	32.0	50.3
Greek	M	29.7	45.0
Thai	M	31.2	43.4
Hindi	M	17.5	37.8
Average		35.6	53.5

Results on QnA



	#langs.	avg. chrF	avg. BLEU
ChatGPT (0-shot)	203	32.3	16.7
ChatGPT (5-shot)	203	33.1	17.3
GPT-4	20	44.6	24.6
NLLB	201	45.3	27.1
Google	115	**52.2**	**34.6**

Performance on translation averaged across languages



Lang.	ChatGPT		NLLB	
	BLEU	chrF	BLEU	chrF
srp_Cyrl	1.36	3.26	**43.4**	**59.7**
kon_Latn	0.94	8.50	**18.9**	**45.3**
tso_Latn	2.92	15.0	**26.7**	**50.0**
kac_Latn	0.04	2.95	**14.3**	**37.5**
nso_Latn	3.69	16.7	**26.5**	**50.8**
jpn_Jpan	**28.4**	**32.9**	20.1	27.9
nno_Latn	**37.1**	**58.7**	33.4	53.6
zho_Hans	**36.3**	**31.0**	26.6	22.8
zho_Hant	**26.0**	**24.4**	12.4	14.0
acm_Arab	**28.2**	**44.7**	11.8	31.9

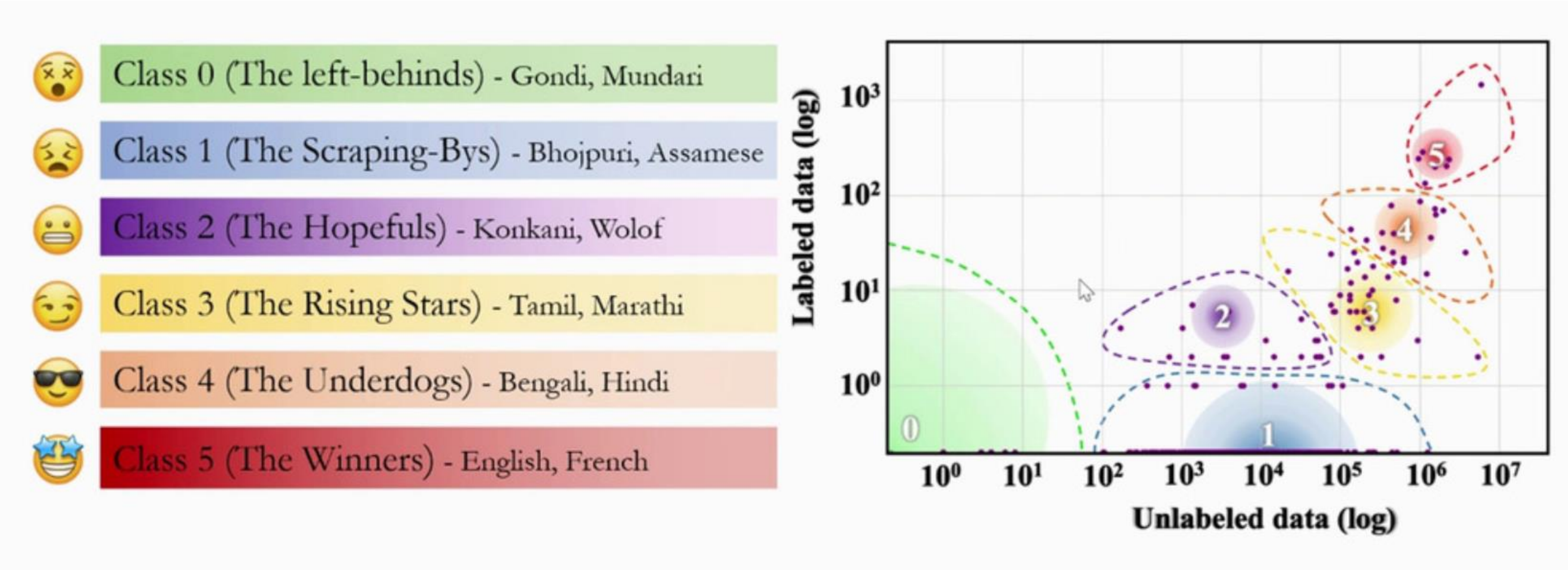
Performance on translation High vs low resource

- Significant gap between English and other languages on multiple tasks
- High-resource and Latin script languages can give good performance on GPT
- Poor performance on low-resource languages
- Translate-test is a strong baseline
- Open-source models lag behind GPT models ➔ they are very English heavy

(BUFFET, MEGA, ChatGptMT,ChatGptMLing)

Disparity in linguistic resources has always been an issue for NLP

Wikipedia/CommonCrawl data as a proxy for monolingual data availability



How do we bring the state-of-the-art NLP solutions to all languages?

Can we train such large models for all languages?

Joshi et al. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. ACL 2023.

The proposed recipe for multilingual NLP

Wikipedia
The Free ~~Encyclopedia~~
Corpus

OPUS ... the open parallel corpus

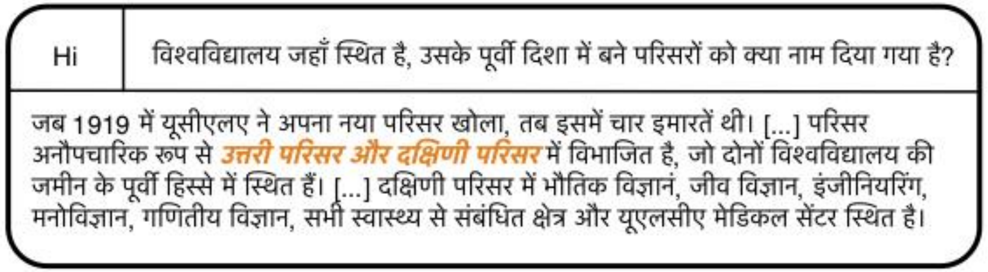
Collect a large amount of monolingual and/or parallel data

Lot of monolingual data

mBERT
mBART/mT5
AYA

Learn a multilingual ***language model***

Lot of memory & computation



Hi	विश्वविद्यालय जहाँ स्थित है, उसके पूर्वी दिशा में बने परिसरों को क्या नाम दिया गया है?

जब 1919 में यूसीएलए ने अपना नया परिसर खोला, तब इसमें चार इमारतें थी। [...] परिसर अनौपचारिक रूप से ***उत्तरी परिसर और दक्षिणी परिसर*** में विभाजित है, जो दोनों विश्वविद्यालय की जमीन के पूर्वी हिस्से में स्थित हैं। [...] दक्षिणी परिसर में भौतिक विज्ञानं, जीव विज्ञान, इंजीनियरिंग, मनोविज्ञान, गणितीय विज्ञान, सभी स्वास्थ्य से संबंधित क्षेत्र और यूएलसीए मेडिकल सेंटर स्थित है।

Finetune the language model on reasonable amount of data from one or more languages

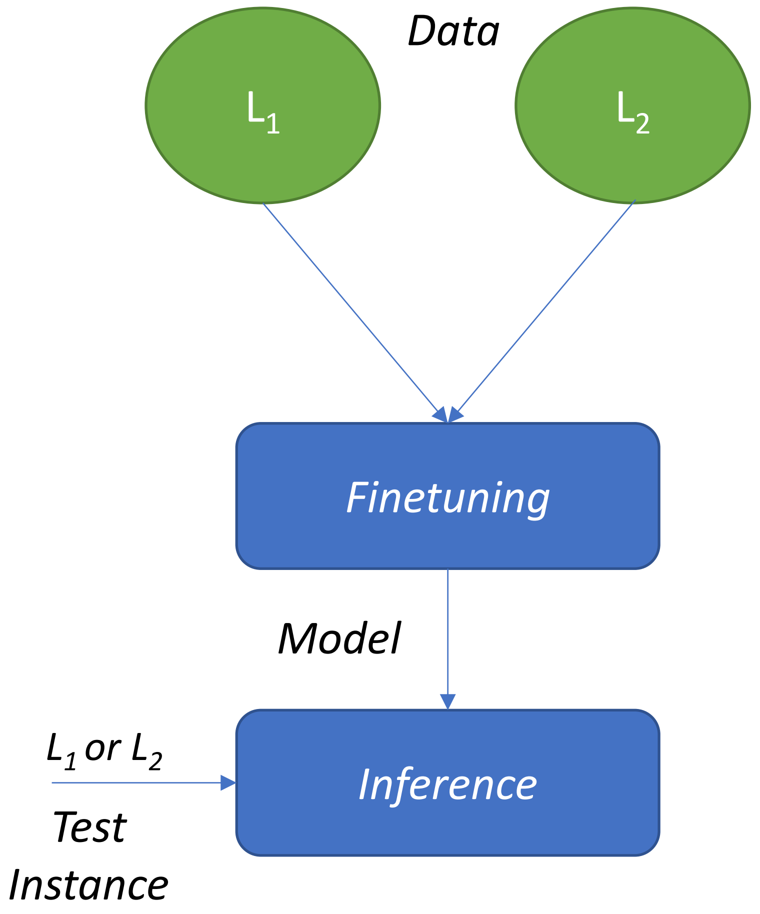
- *Task-specific*
- *General instruction-response*
- *Human Preference*

Supervised data limited for many languages

But not all languages have such data ....
Joint learning leads to transfer of knowledge across languages

Transfer Learning

Joint Learning



Data

L_1

L_2

Finetuning

Model

Inference

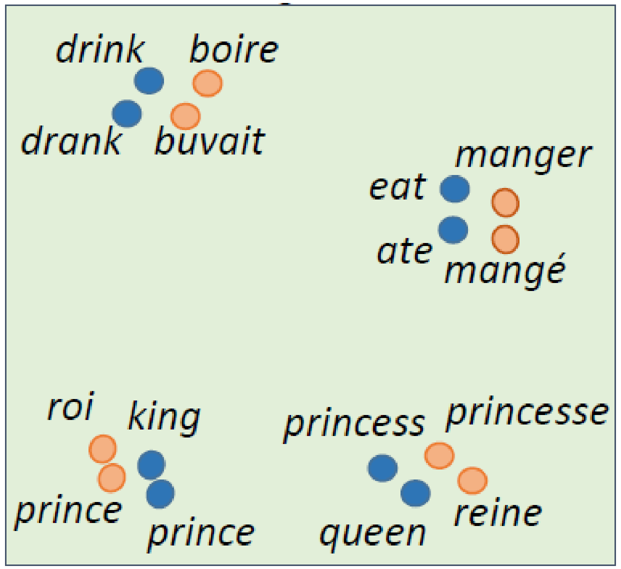
L_1 or L_2

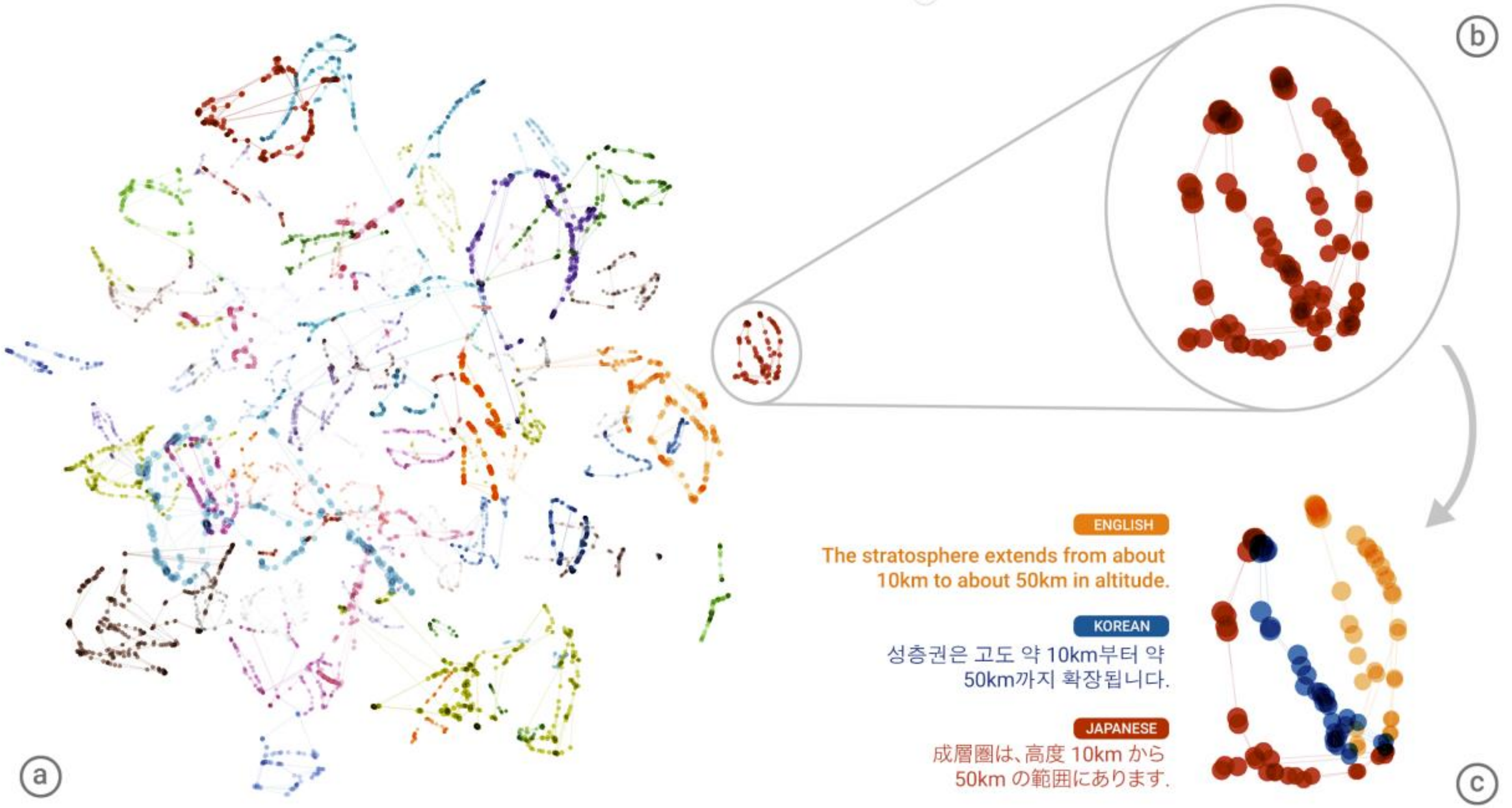
Test Instance

- *Analogy to Multi-task learning* ➔ *Task ≡ Language*
- *Related Tasks can share representations*
- *Representation Bias: Generalize over multiple languages*
- *Eavesdropping*
- *Data Augmentation*

(Caruana., 1997)

Similar words and sentences across languages have similar embeddings





The stratosphere extends from about 10km to about 50km in altitude.

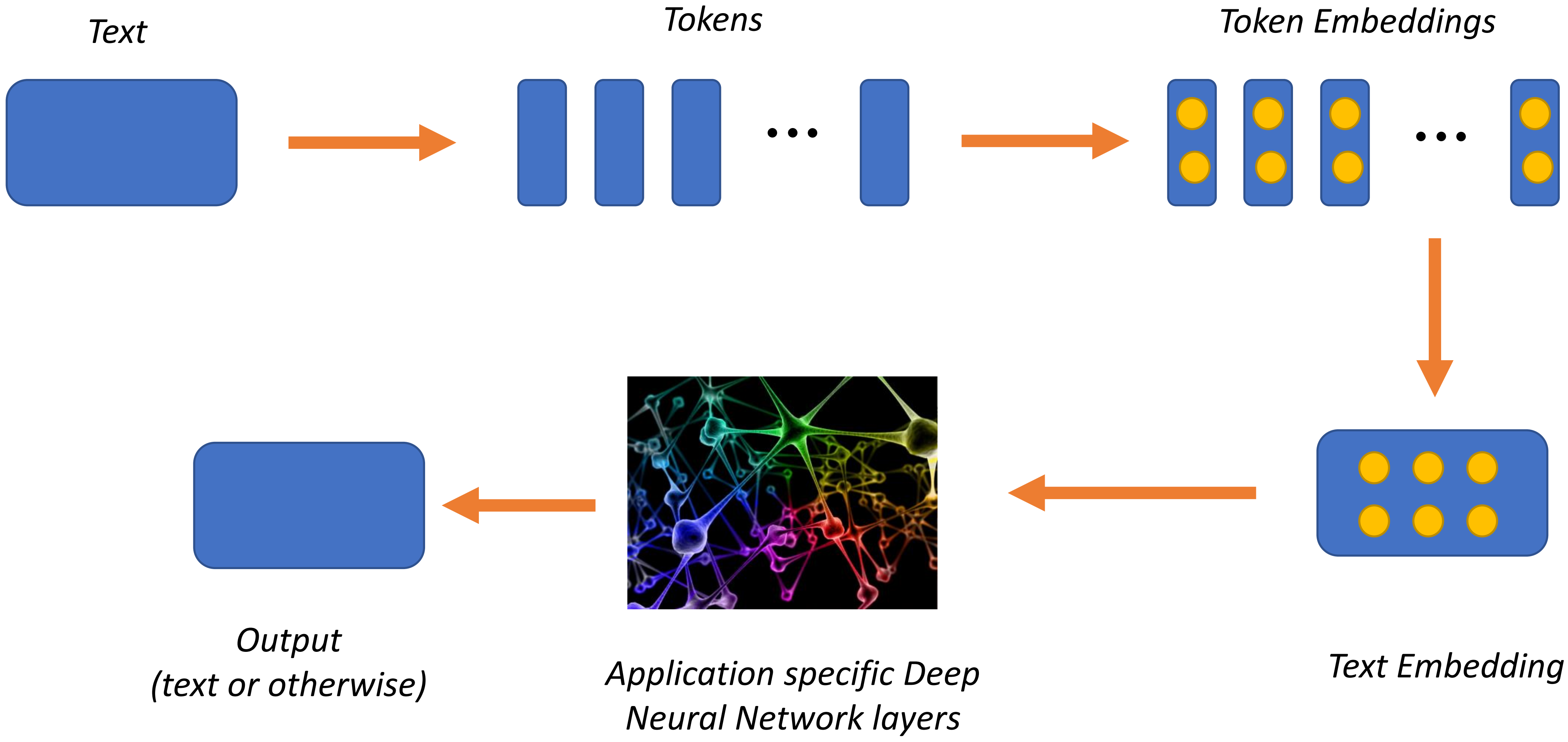
성층권은 고도 약 10km부터 약 50km까지 확장됩니다.

成層圏は、高度 10km から 50km の範囲にあります.

Agenda

- Introduction
- Typical Multilingual NLP Pipeline
- Cross-lingual Transfer
- Multilingual Word Embeddings
- Multilingual Pre-trained Language Models
    - Extending English LLMs to new languages
- Finetuning Pre-trained LMs
- Evaluation of Multilingual models

A Typical Multilingual NLP Pipeline



Text

Tokens

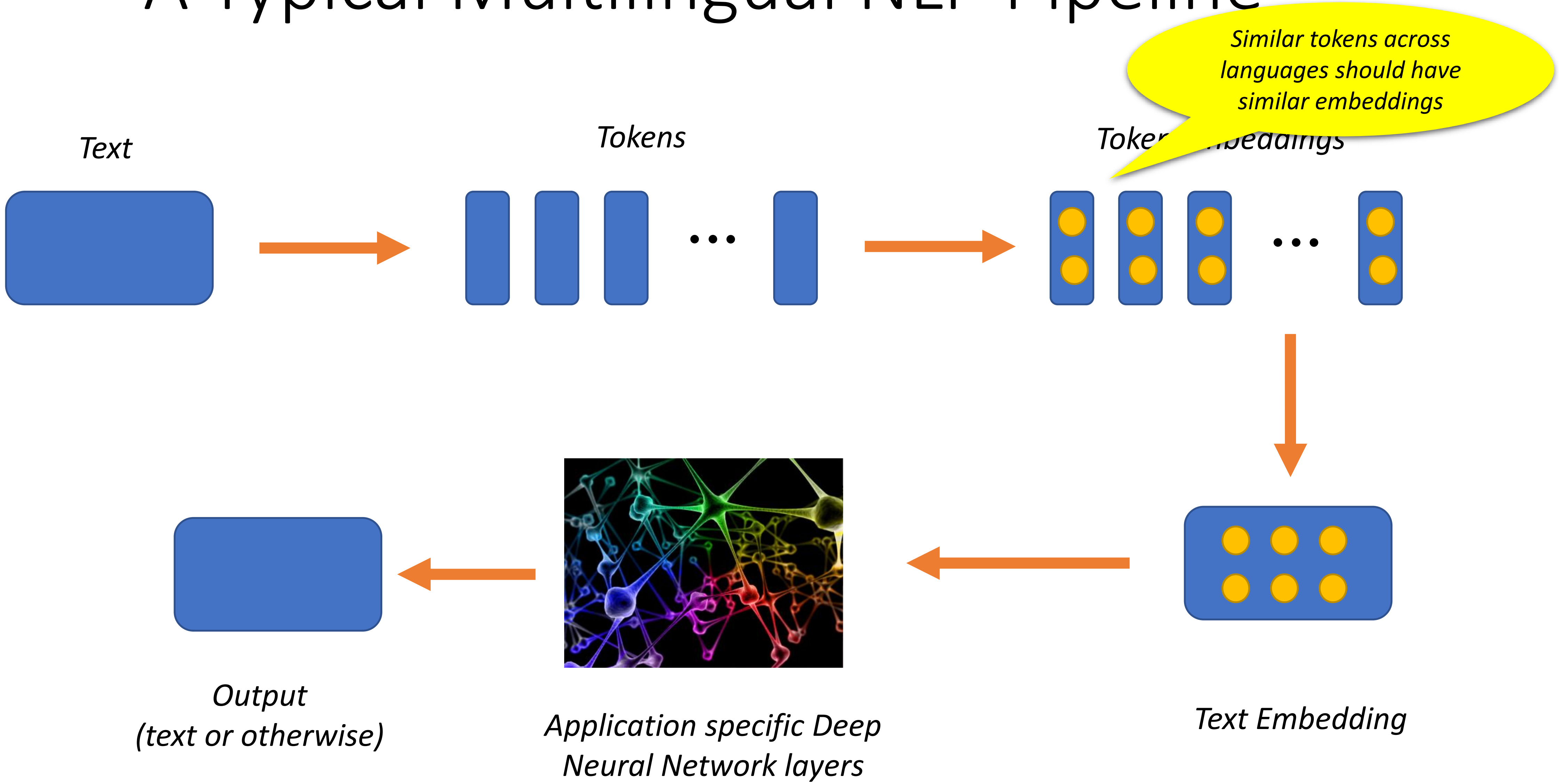
Token Embeddings

Text Embedding

Application specific Deep Neural Network layers

Output (text or otherwise)

A Typical Multilingual NLP Pipeline



Similar tokens across languages should have similar embeddings

Text

Tokens

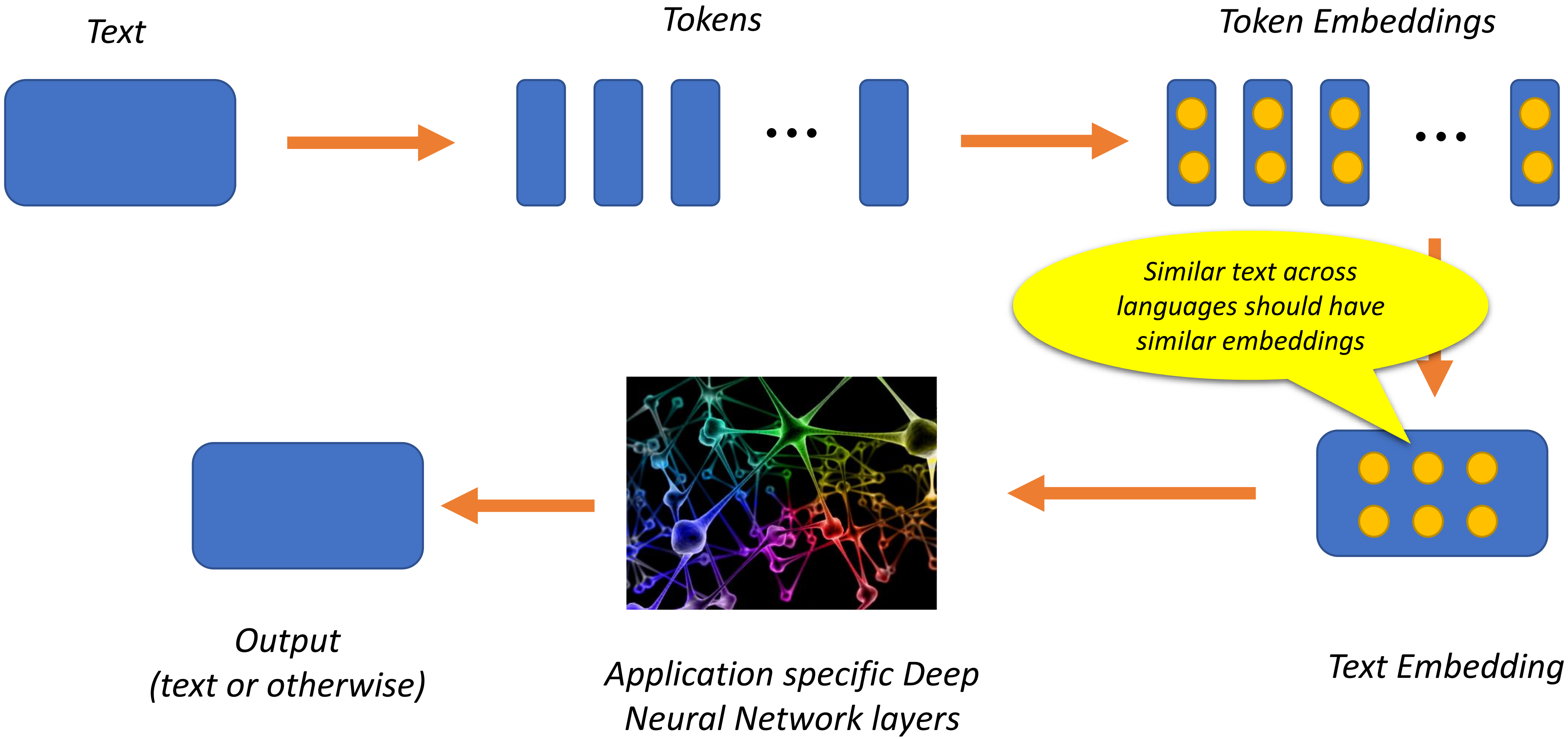
Toke mbeddings

Text Embedding

Application specific Deep Neural Network layers

Output
(text or otherwise)

A Typical Multilingual NLP Pipeline



Text

Tokens

Token Embeddings

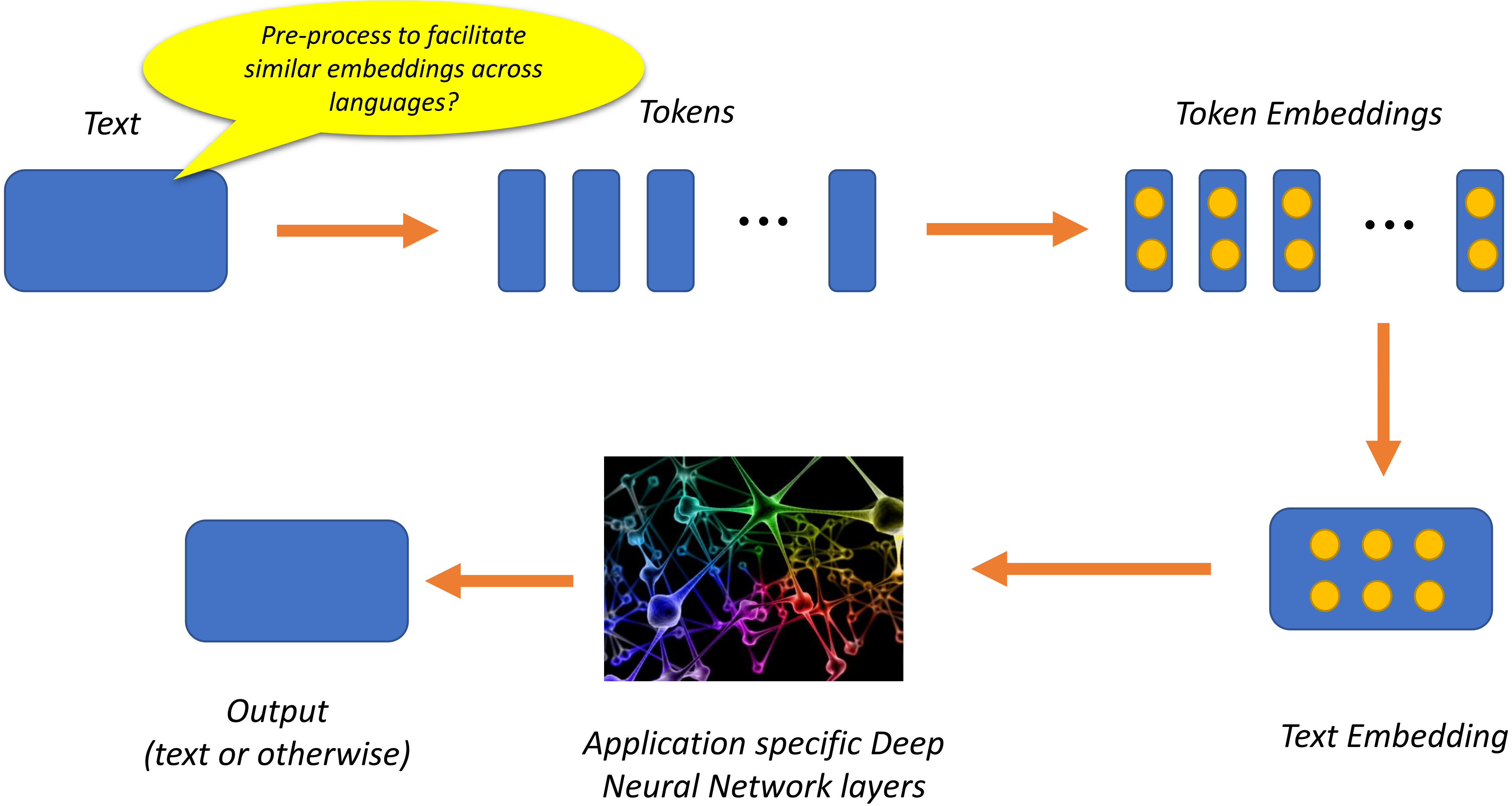
Similar text across languages should have similar embeddings

Text Embedding

Application specific Deep Neural Network layers

Output (text or otherwise)

A Typical Multilingual NLP Pipeline



Pre-process to facilitate similar embeddings across languages?

Text

Tokens

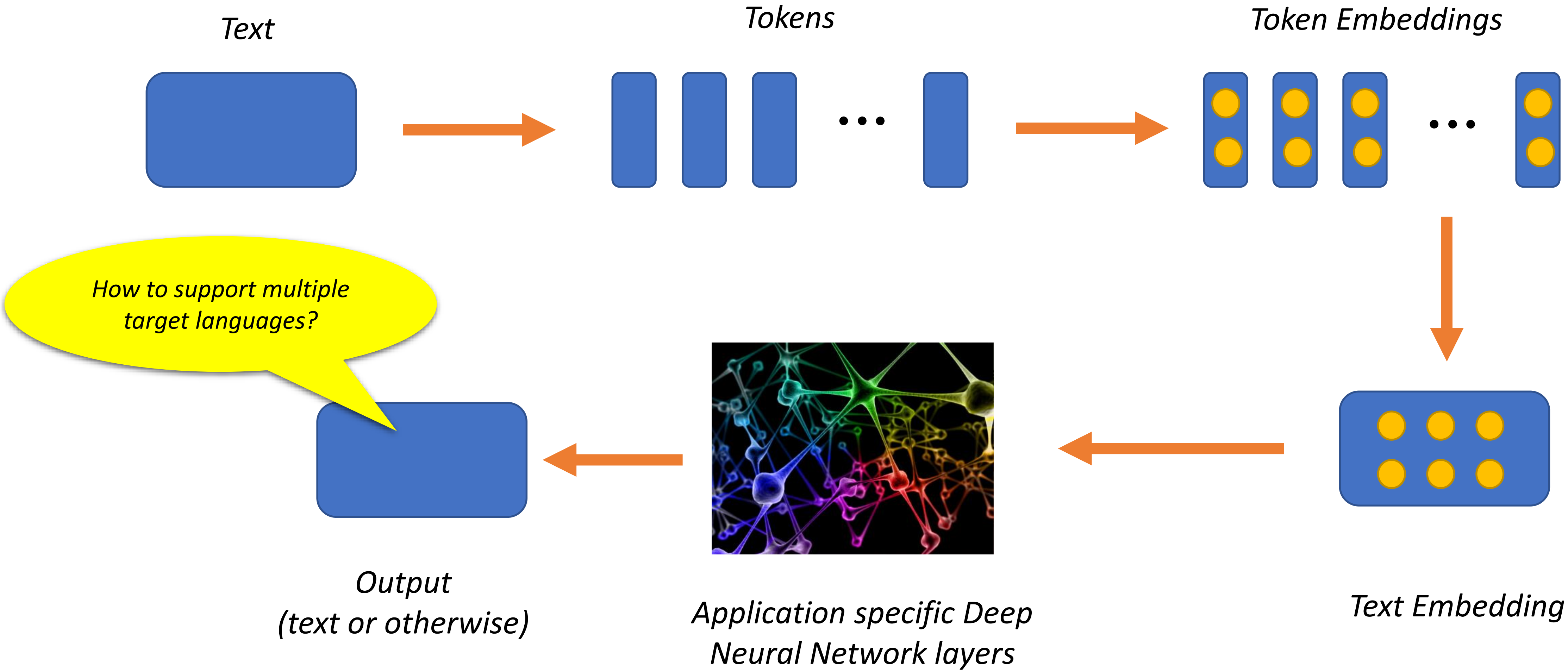
Token Embeddings

Text Embedding

Application specific Deep Neural Network layers

Output (text or otherwise)

A Typical Multilingual NLP Pipeline



Text

Tokens

Token Embeddings

Text Embedding

Application specific Deep Neural Network layers

Output (text or otherwise)

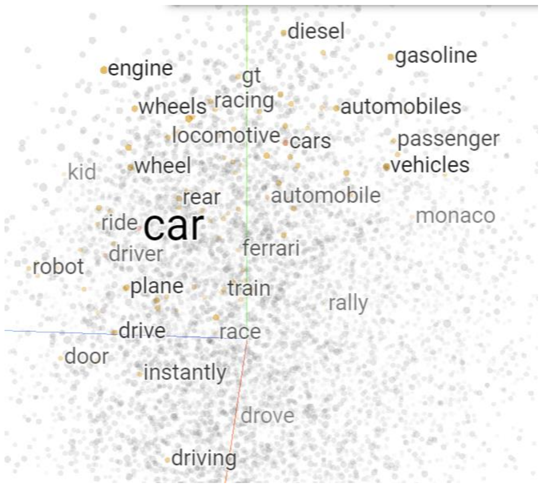
How to support multiple target languages?

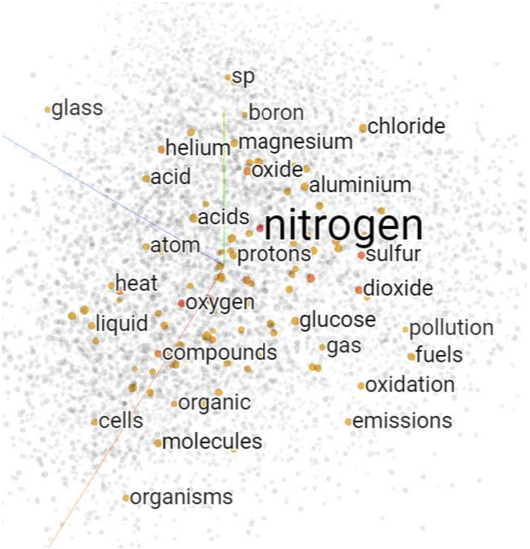
Agenda

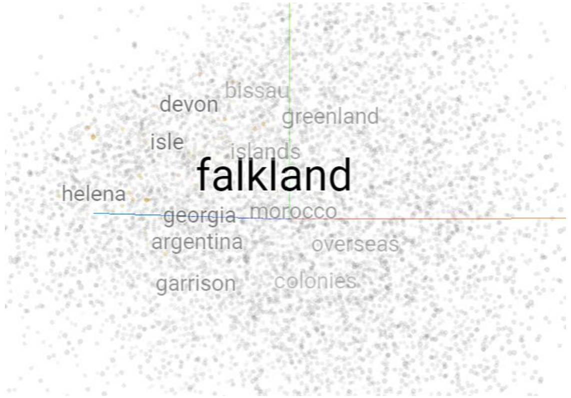
- Introduction
- Typical Multilingual NLP Pipeline
- Cross-lingual Transfer
- Multilingual Word Embeddings
- Multilingual Pre-trained Language Models
  - Extending English LLMs to new languages
- Finetuning Pre-trained LMs
- Evaluation of Multilingual models

Word Embeddings

Distributed, dense vector representations of words that capture similarities between words



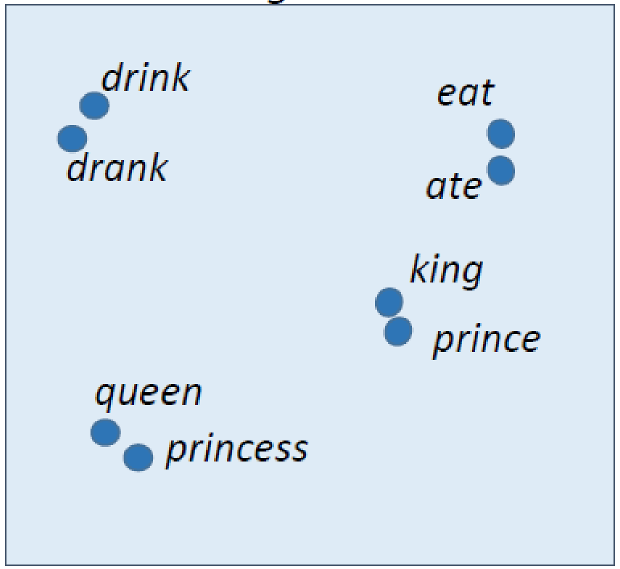




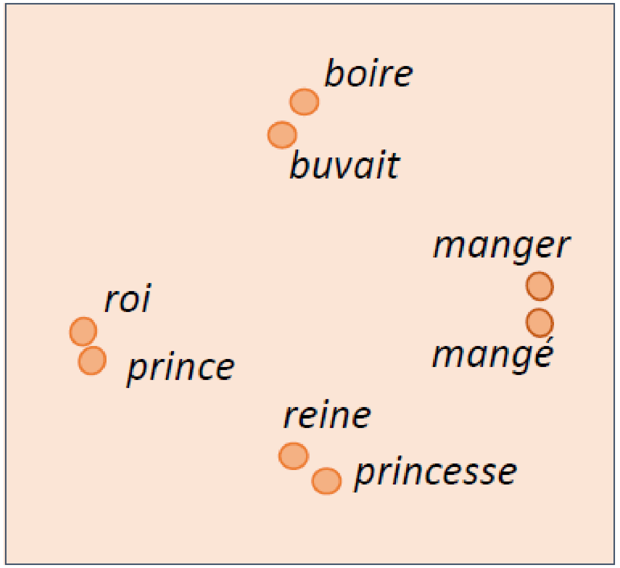
“Words that occur in similar contexts tend to have similar meanings”
- Turney and Pantel (2010)

He is unhappy about the failure of the project

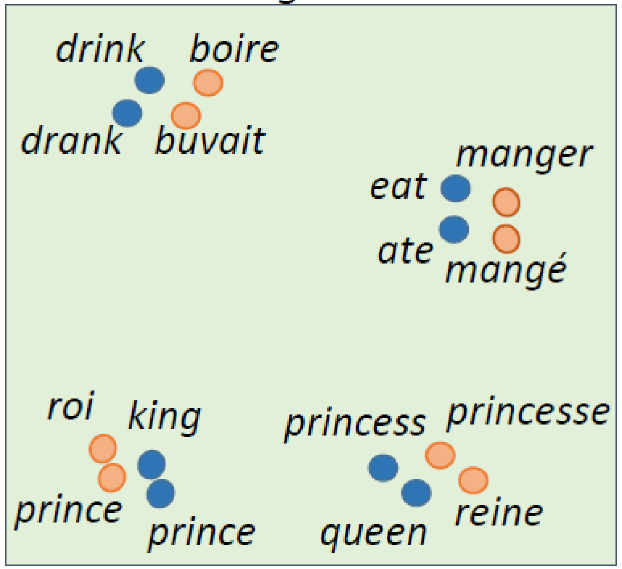
The failure of the team to successfully finish the task made him sad



English



French



Joint English French

Monolingual Word Representations

(capture syntactic and semantic similarities between words)

Multilingual Word Representations

(capture syntactic and semantic similarities between words both within and across languages)

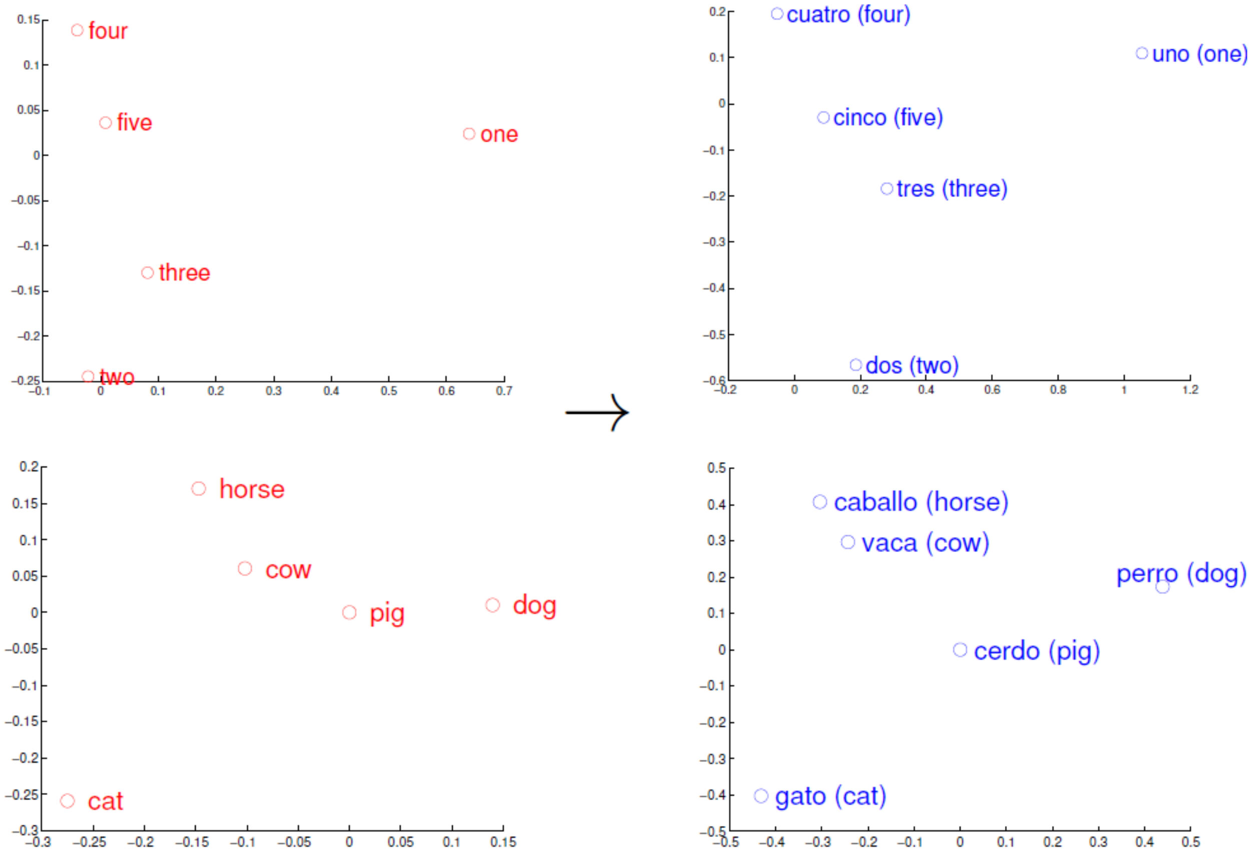
(Source: Khapra and Chandar, 2016)

$$embed(y) = f(embed(x))$$

x, y are source and target words

$embed(w)$: embedding for word w

Is it possible to learn mapping functions?



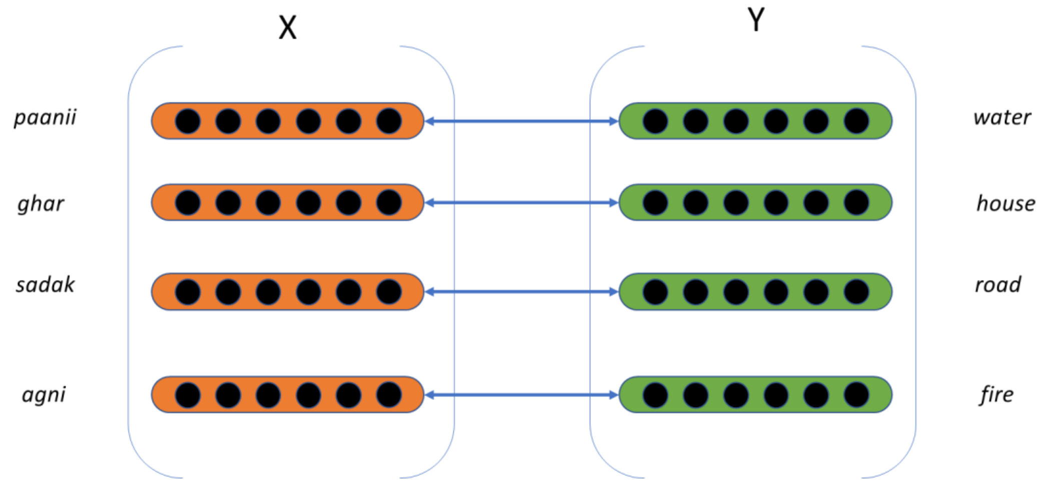
- Languages share concepts ground in the real world
- Some evidence of universal semantic structure (*Youn et al., 2016)*
- Isomorphism between embedding spaces (*Mikolov et al., 2013)*
- Isomorphism can be captured via a linear transformation

(Source: Mikolov et al., 2013)

Supervised Learning

Linear Least Square and variants

(Mikolov et al., 2013; Xing et al., 2015; Artetxe et al., 2016; Smith et al., 2017)



$$\mathrm{X}\color{red}{\mathrm{W}} = Y$$

$$W^* = \underset{W \in \mathbb{R}^d}{\operatorname{argmin}} \|XW - Y\|_2^2$$

We can have a closed form solution:

$$X^+ = (X^T X)^{-1} X^T$$

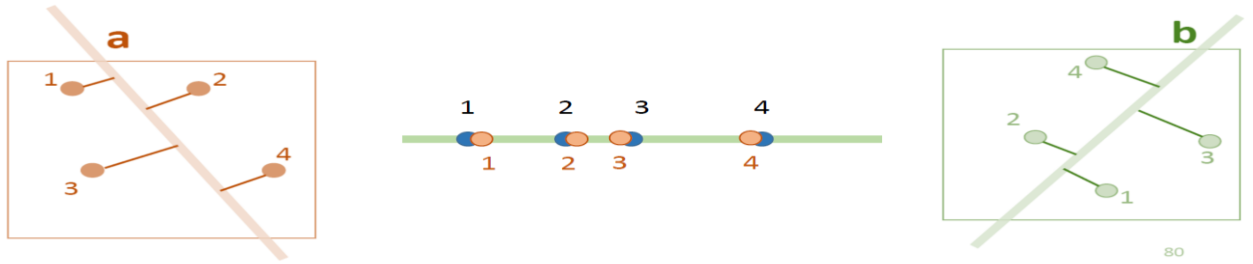
$$W^* = X^+ Y$$

Methods mapping different languages to a common space

Canonical Correlation Analysis (CCA)

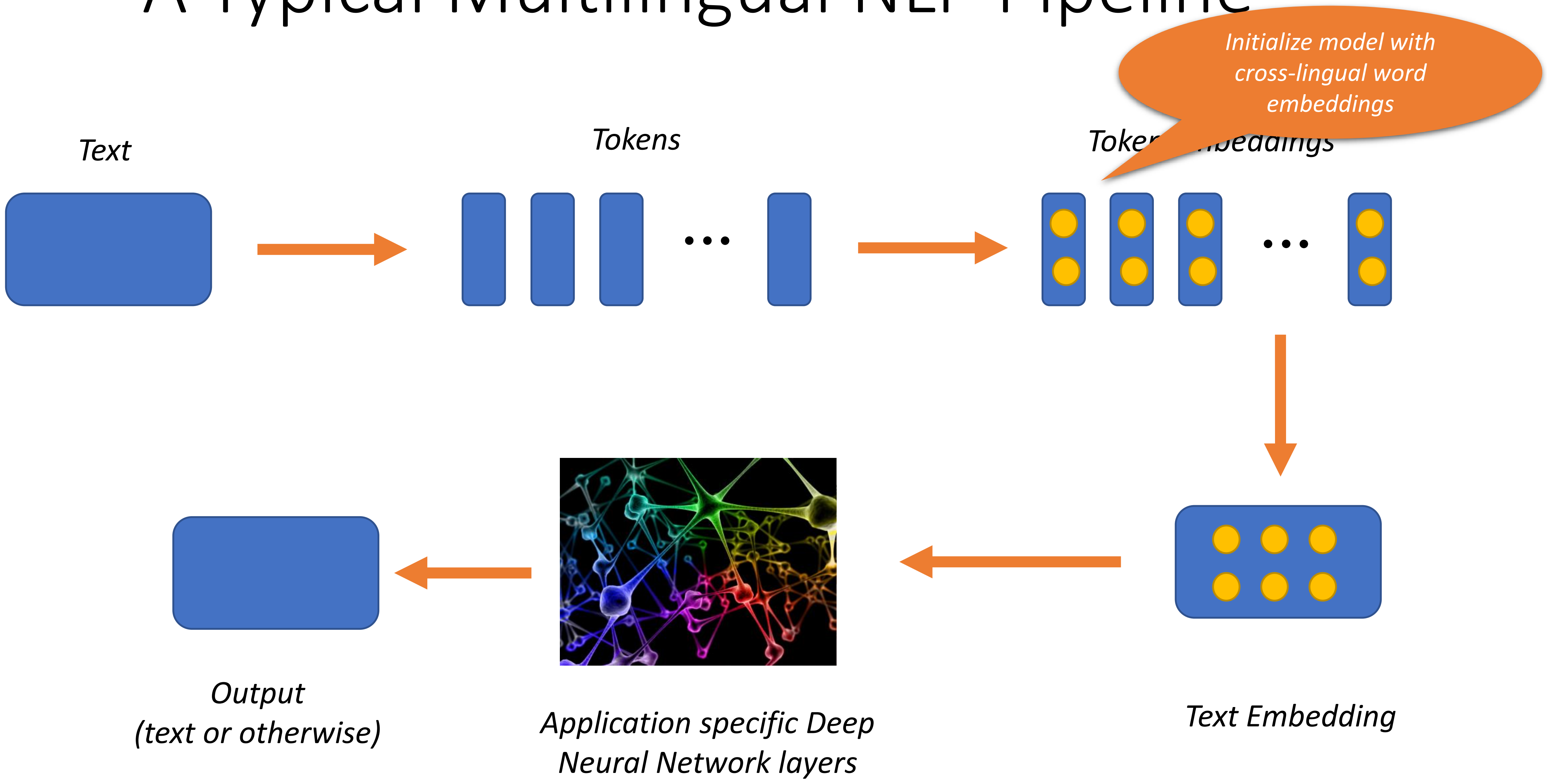
(Faruqui and Dyer, 2014;
Ammar et al. 2015)

GeoMM *(Jawanpuria et al., 2019)*



MUSE is popular toolkit to learn cross-lingual word embeddings https://github.com/facebookresearch/MUSE

A Typical Multilingual NLP Pipeline



Initialize model with cross-lingual word embeddings

Text

Tokens

Toke beddings

Text Embedding

Application specific Deep Neural Network layers

Output
(text or otherwise)

Agenda

- Introduction
- Typical Multilingual NLP Pipeline
- Cross-lingual Transfer
- Multilingual Word Embeddings
- Multilingual Pre-trained Language Models
    - Extending English LLMs to new languages
- Finetuning Pre-trained LMs
- Evaluation of Multilingual models

Multilingual Language Models

Encoder Models for NLU: mBERT (Multilingual BERT)

- Simple extension to monolingual BERT training
- BERT trained jointly on monolingual data from Wikipedias of 104 languages
- Standard MLM objective
- Shared word-piece vocabulary (110k shared WordPiece)
- No parallel resources to provide translation information across languages

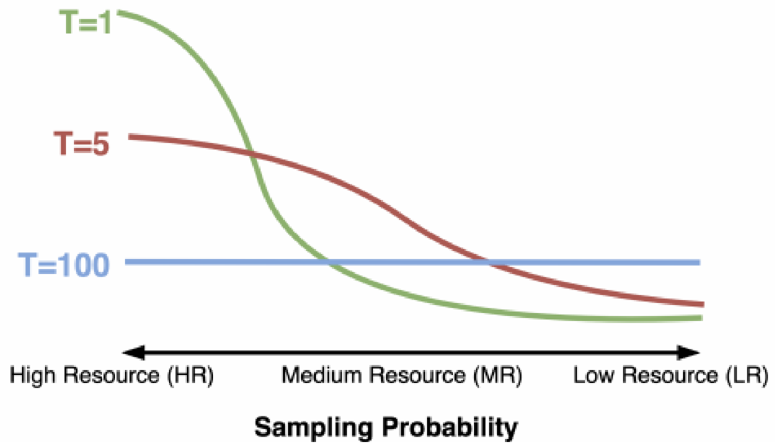
Temperature based data sampling to address data skew (during vocab construction and pre-training)

Data Distribution
Language *l*

$$p_l = \frac{D_l}{\Sigma_k D_k}$$

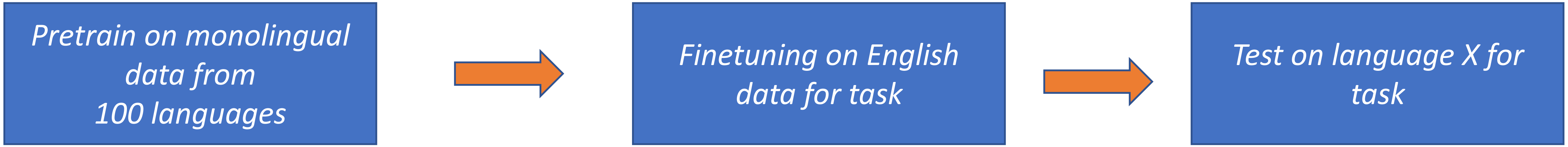
Sampling Probability
Language *l*

$$p_l^{\frac{1}{T}}$$



104 languages, 12-layer, 768-hidden, 12-heads, 172M parameters

Surprisingly good at zero-shot cross-lingual model transfer

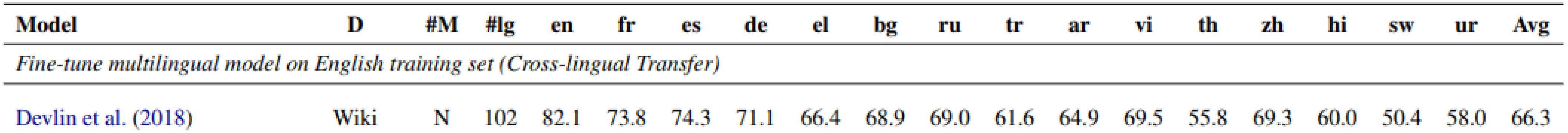


Pretrain on monolingual data from 100 languages

Finetuning on English data for task

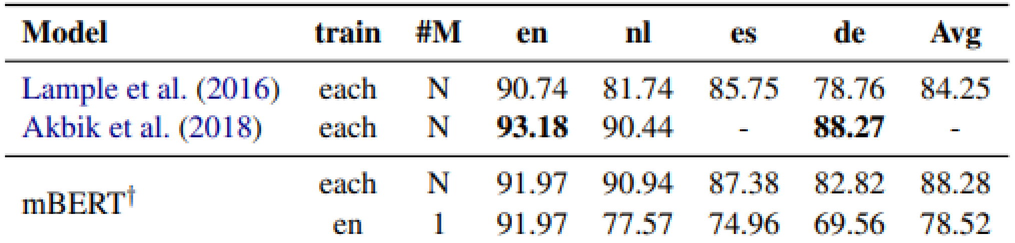
Test on language X for task

XNLI



Model	**D**	**#M**	**#lg**	**en**	**fr**	**es**	**de**	**el**	**bg**	**ru**	**tr**	**ar**	**vi**	**th**	**zh**	**hi**	**sw**	**ur**	**Avg**
Fine-tune multilingual model on English training set (Cross-lingual Transfer)																			
Devlin et al. (2018)	Wiki	N	102	82.1	73.8	74.3	71.1	66.4	68.9	69.0	61.6	64.9	69.5	55.8	69.3	60.0	50.4	58.0	66.3

NER



Model	**train**	**#M**	**en**	**nl**	**es**	**de**	**Avg**
Lample et al. (2016)	each	N	90.74	81.74	85.75	78.76	84.25
Akbik et al. (2018)	each	N	**93.18**	90.44	-	**88.27**	-
mBERT†	each	N	91.97	90.94	87.38	82.82	88.28
	en	1	91.97	77.57	74.96	69.56	78.52

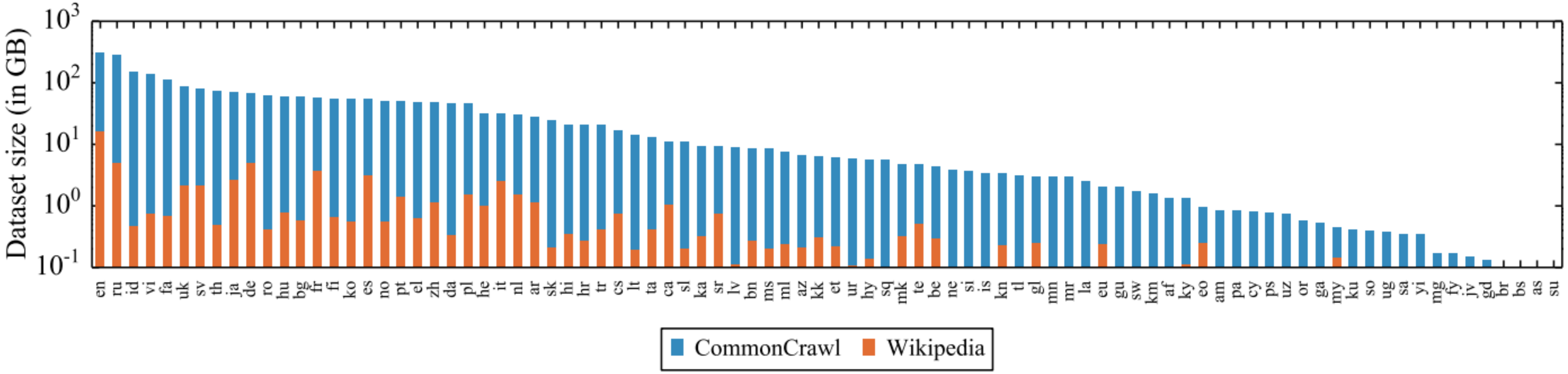
Can we do better data and larger models?

XLM-R

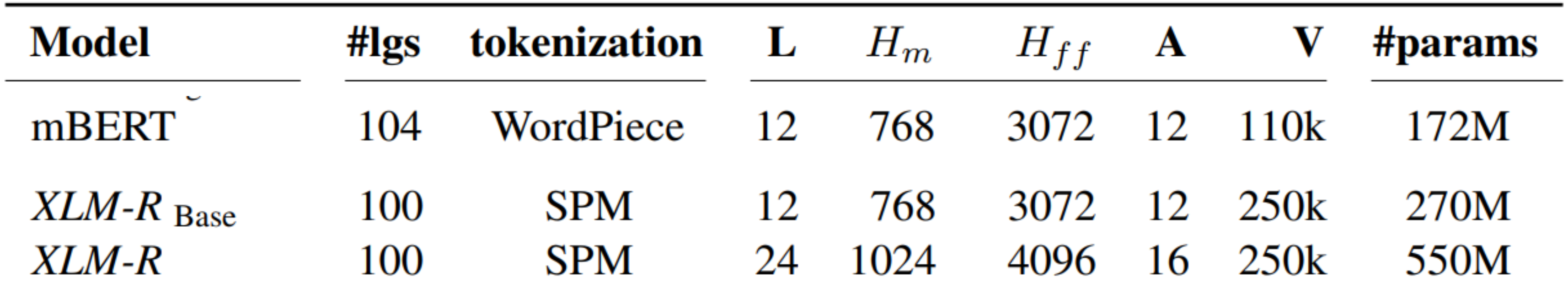
Conneau et al. Facebook. Unsupervised Cross-lingual Representation Learning at Scale. ACL 2020

Essentially mBERT architecture

Train on very large dataset (CommonCrawl derived CC-100 dataset)



Increased Model capacity



Model	**#lgs**	**tokenization**	**L**	H_m	H_{ff}	**A**	**V**	**#params**
mBERT	104	WordPiece	12	768	3072	12	110k	172M
XLM-R Base	100	SPM	12	768	3072	12	250k	270M
XLM-R	100	SPM	24	1024	4096	16	250k	550M

Tradeoff studies

Positive transfer vs. capacity dilution
High vs. low resource language performance

Factors that improve model quality

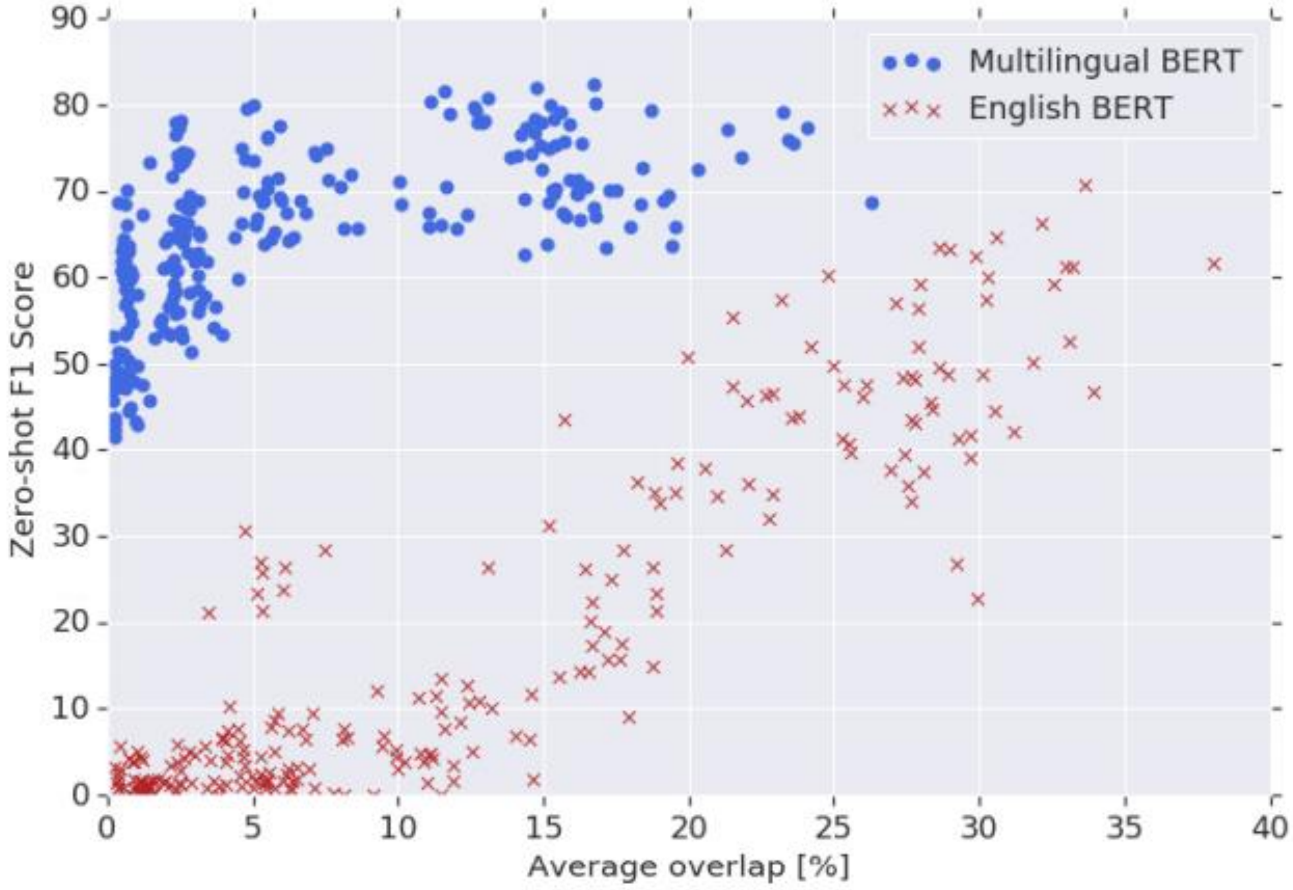
- **Larger** model capacity helps
- **More** monolingual data is better
- **Longer** Pre-training
  - Don’t use validation perplexity for early stopping of pre-training
- **Larger** vocabulary
- **Larger batch size** while pre-training

- Performance degrades with increasing number of pretraining languages
- Right trade-off of using high vs low resource languages via data sampling

mBERT does not use any cross-lingual signals for learning …
Why does multilingual BERT work?

What is the nature of the multilingual representations?

Does word-piece overlap explain zeroshot performance of mBERT?



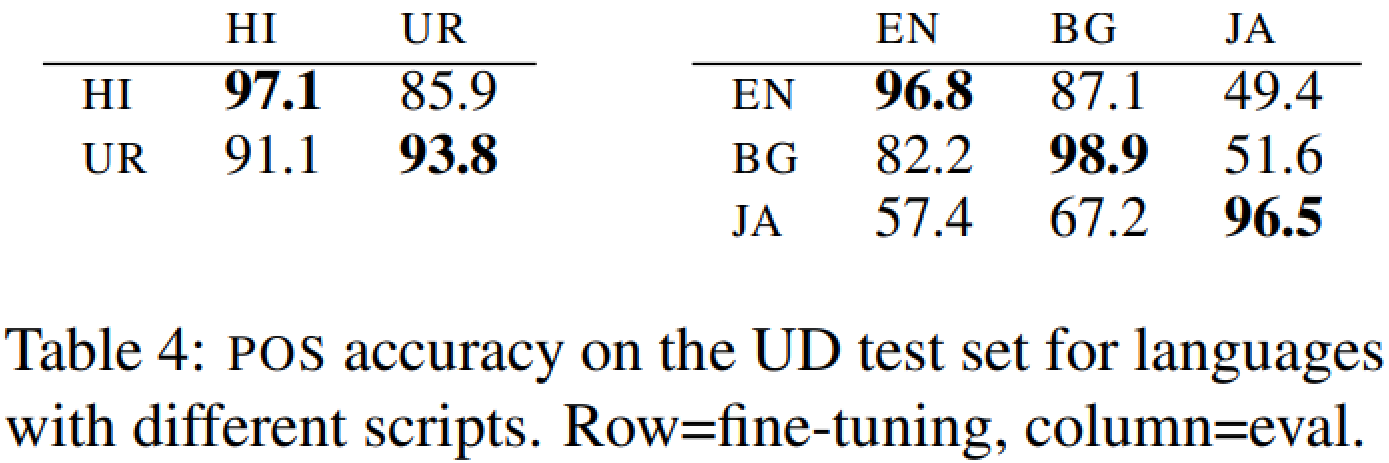
For English-BERT, the NER performance increases with word overlap

For M-BERT, the NER performance is not strongly correlated to wordpiece overlap

M-BERT learns about language structure beyond word-piece overlap

Pires et al., Google. How multilingual is Multilingual BERT? ACL. 2019.

Zero-short transfer happens even hold even though the scripts are different



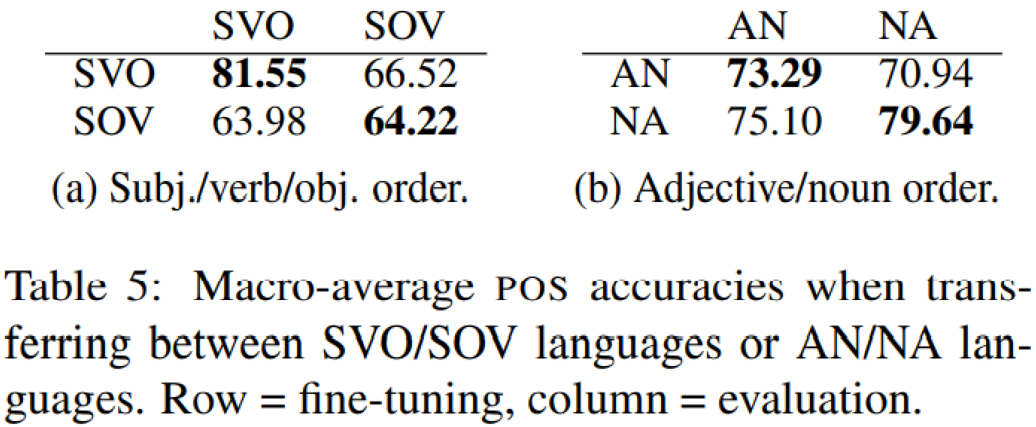
	HI	UR
HI	**97.1**	85.9
UR	91.1	**93.8**

	EN	BG	JA
EN	**96.8**	87.1	49.4
BG	82.2	**98.9**	51.6
JA	57.4	67.2	**96.5**

Table 4: POS accuracy on the UD test set for languages with different scripts. Row=fine-tuning, column=eval.

Does not generalize for all languages – see Japanese which has a different word order from English

How does typological similarity affect M-BERT's ability to generalize?



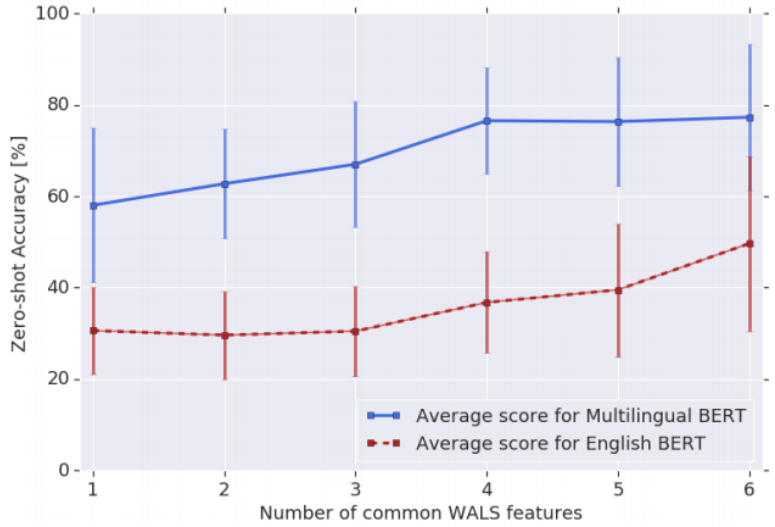
	SVO	SOV
SVO	**81.55**	66.52
SOV	63.98	**64.22**

(a) Subj./verb/obj. order.

	AN	NA
AN	**73.29**	70.94
NA	75.10	**79.64**

(b) Adjective/noun order.

Table 5: Macro-average POS accuracies when transferring between SVO/SOV languages or AN/NA languages. Row = fine-tuning, column = evaluation.



Study of effect of grammatical features related to ordering

Performance is best when transferring between languages that share word order features

Cannot learn syntactic transformations

Why does multilingual BERT work?

Hypothesis:

Word pieces used in all languages (numbers, URLs, etc) which have to be mapped to a shared space forces the co-occurring pieces to also be mapped to a shared space, thus spreading the effect to other word pieces, until different languages are close to a shared space

Similar findings in

Shijie Wu, Mark Dredze. ***Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT. EMNLP 2019.***

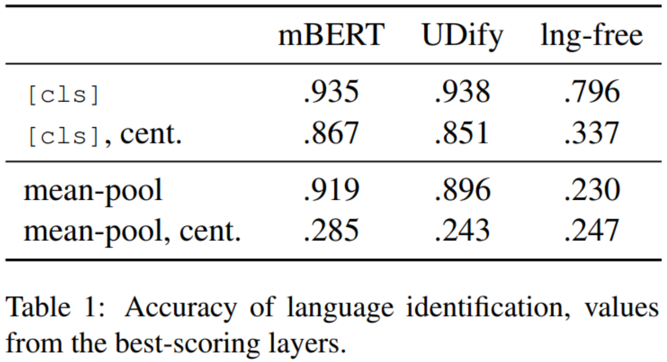
Somewhat different findings in:

Karthikeyan K, Zihan Wang, Stephen Mayhew, Dan Roth. Cross-Lingual Ability of Multilingual BERT: An Empirical Study. ICLR. 2020.

- Lexical overlap does not matter (Use FakeEnglish to destroy lexical similarity)
- Structural similarity matters

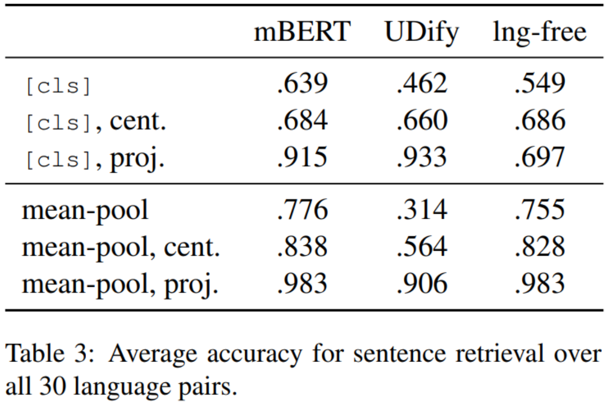
How language neural is mBERT?

- Representations cluster by language family
- Language information is still maintained at each layer
- Mean centering the embeddings per language can improve language invariant behaviour



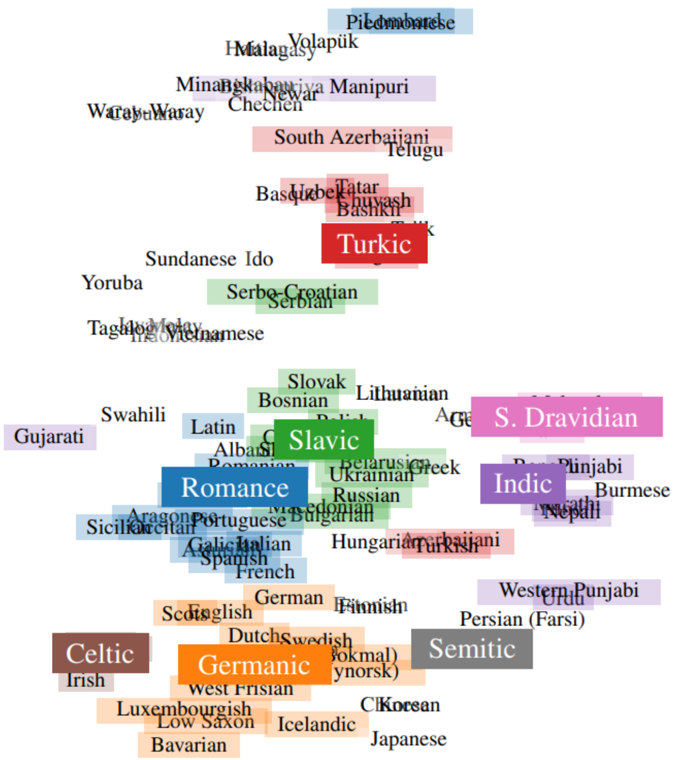
	mBERT	UDify	lng-free
`[cls]`	.935	.938	.796
`[cls]`, cent.	.867	.851	.337
mean-pool	.919	.896	.230
mean-pool, cent.	.285	.243	.247

Table 1: Accuracy of language identification, values from the best-scoring layers.



	mBERT	UDify	lng-free
`[cls]`	.639	.462	.549
`[cls]`, cent.	.684	.660	.686
`[cls]`, proj.	.915	.933	.697
mean-pool	.776	.314	.755
mean-pool, cent.	.838	.564	.828
mean-pool, proj.	.983	.906	.983

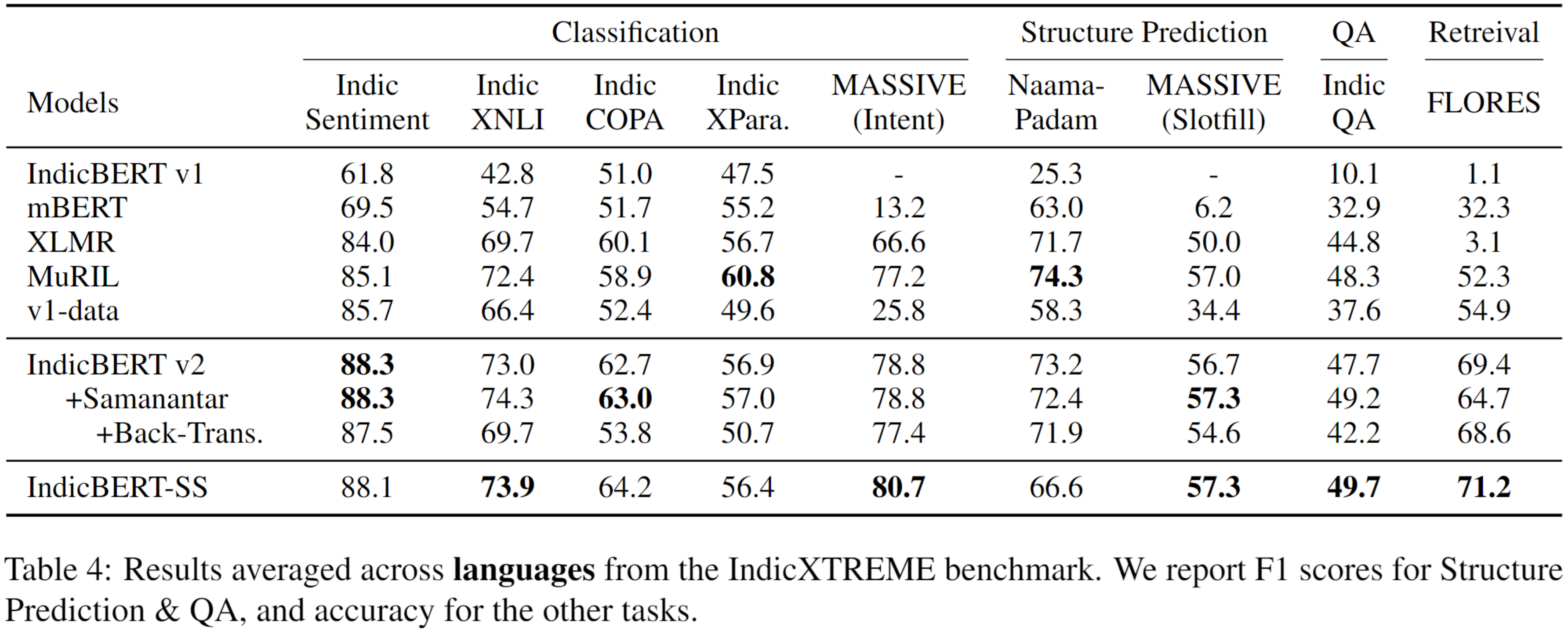
Table 3: Average accuracy for sentence retrieval over all 30 language pairs.



Libovicky et al., On the Language Neutrality of Pre-trained Multilingual Representations. EMNLP-Findings. 2020.

Language-family specific pre-trained models

IndicBERT➔ model for 22 Indian languages + English)



Models	Classification					Structure Prediction		QA	Retreival
	Indic Sentiment	Indic XNLI	Indic COPA	Indic XPara.	MASSIVE (Intent)	Naama-Padam	MASSIVE (Slotfill)	Indic QA	FLORES
IndicBERT v1	61.8	42.8	51.0	47.5	-	25.3	-	10.1	1.1
mBERT	69.5	54.7	51.7	55.2	13.2	63.0	6.2	32.9	32.3
XLMR	84.0	69.7	60.1	56.7	66.6	71.7	50.0	44.8	3.1
MuRIL	85.1	72.4	58.9	**60.8**	77.2	**74.3**	57.0	48.3	52.3
v1-data	85.7	66.4	52.4	49.6	25.8	58.3	34.4	37.6	54.9
IndicBERT v2	**88.3**	73.0	62.7	56.9	78.8	73.2	56.7	47.7	69.4
+Samanantar	**88.3**	74.3	**63.0**	57.0	78.8	72.4	**57.3**	49.2	64.7
+Back-Trans.	87.5	69.7	53.8	50.7	77.4	71.9	54.6	42.2	68.6
IndicBERT-SS	88.1	**73.9**	64.2	56.4	**80.7**	66.6	**57.3**	**49.7**	**71.2**

Table 4: Results averaged across **languages** from the IndicXTREME benchmark. We report F1 scores for Structure Prediction & QA, and accuracy for the other tasks.

Compact models can outperform large models built for 100+ languages

1. Doddapaneni et al. Towards Leaving No Indic Language Behind: Building Monolingual Corpora, Benchmark and Models for Indic Languages. ACL. 2023
2. Khanuja et al. MuRIL: *Multilingual Representations for Indian Languages*. https://arxiv.org/abs/2103.10730. 2021.
3. Divyanshu Kakwani, etal., AI4Bharat/Microsoft India. *IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages*. EMNLP-Findings . 2020.

How do we use parallel data to improve multilingual models?

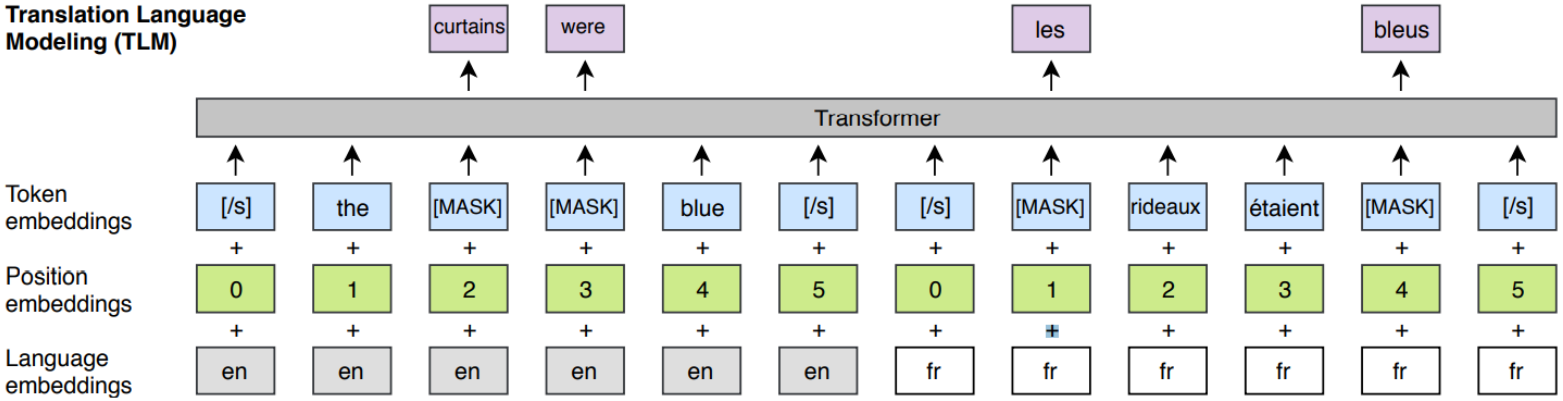
XLM

Conneau., et al. Facebook. Cross-lingual Language Model Pretraining. NeurIPS. 2019.

Input two sentence from different languages

Model alternates with MLM and TLM objectives

TLM ➔ the model can look at both sentences to predict masked token

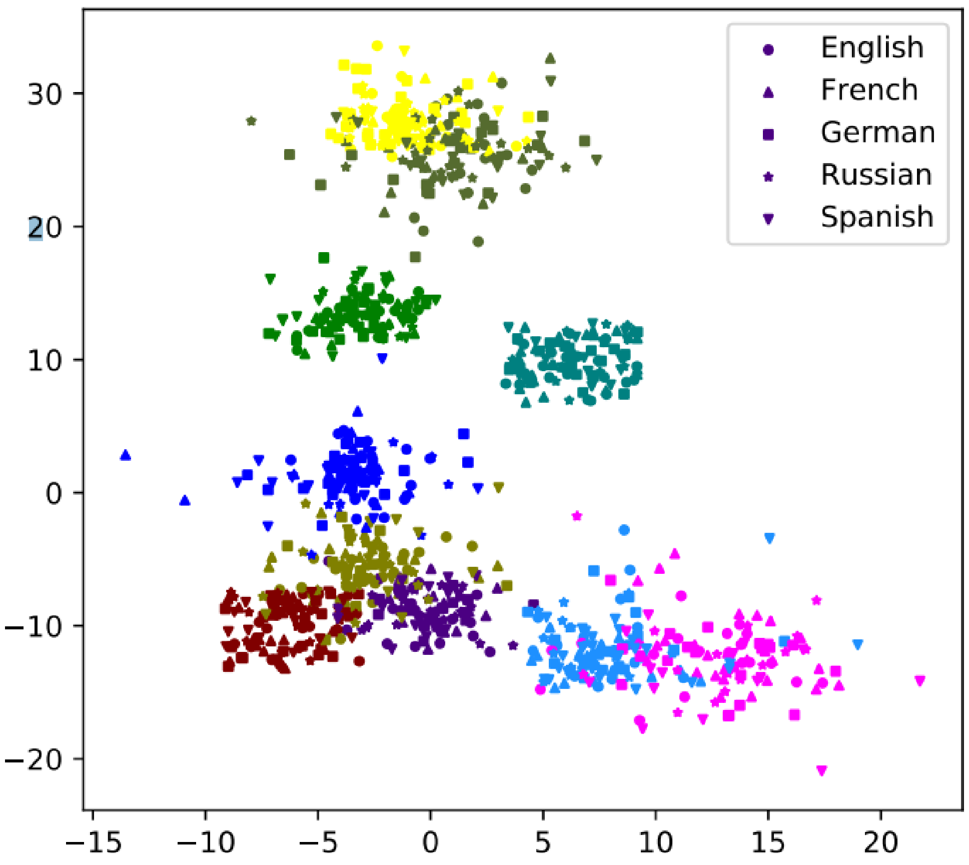


	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Δ
XLM (MLM)	83.2	76.5	76.3	74.2	73.1	74.0	73.1	67.8	68.5	71.2	69.2	71.9	65.7	64.6	63.4	71.5
XLM (MLM+TLM)	**85.0**	**78.7**	**78.9**	**77.8**	**76.6**	**77.4**	**75.3**	**72.5**	**73.1**	**76.1**	**73.2**	**76.5**	**69.6**	**68.4**	**67.3**	**75.1**

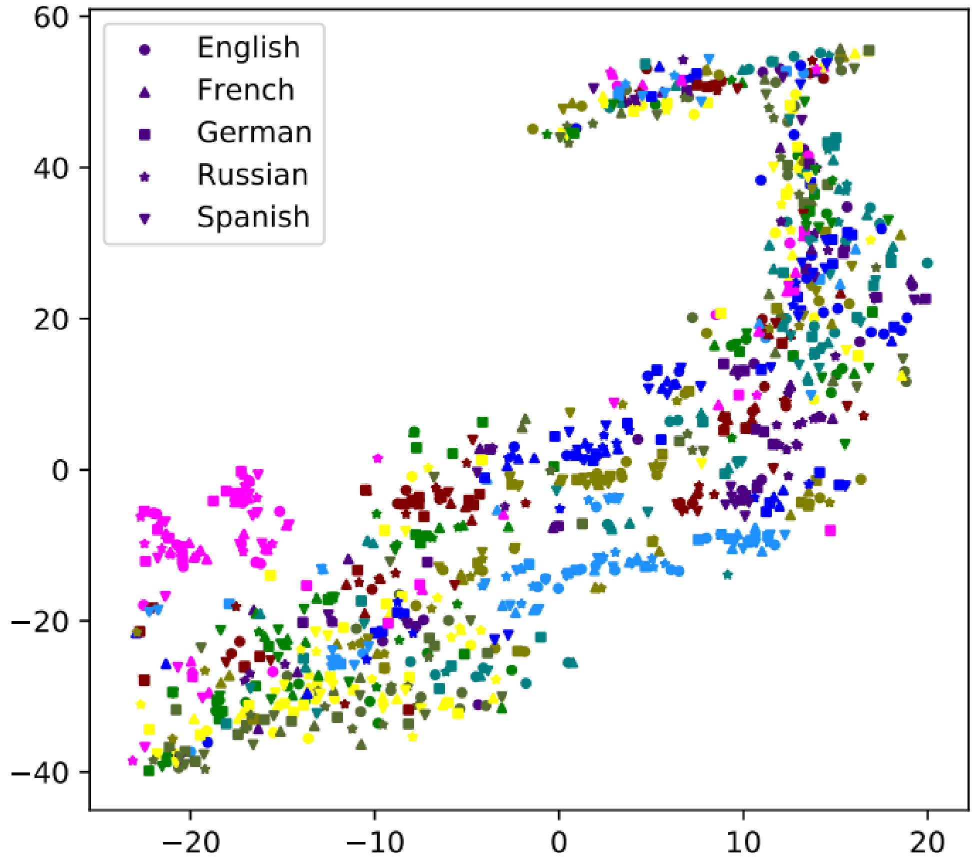
Cross-lingual objectives → *Huang et al., Microsoft. Unicoder: A Universal Language Encoder by Pre-training with Multiple Cross-lingual Tasks. EMNLP. 2019.*

Contrastive objectives → Wei et al. Alibaba. ON LEARNING UNIVERSAL REPRESENTATIONS ACROSS LANGUAGES. ICLR 2021.

Models utilizing parallel data seem to generate more language agnostic representations



XLM-R



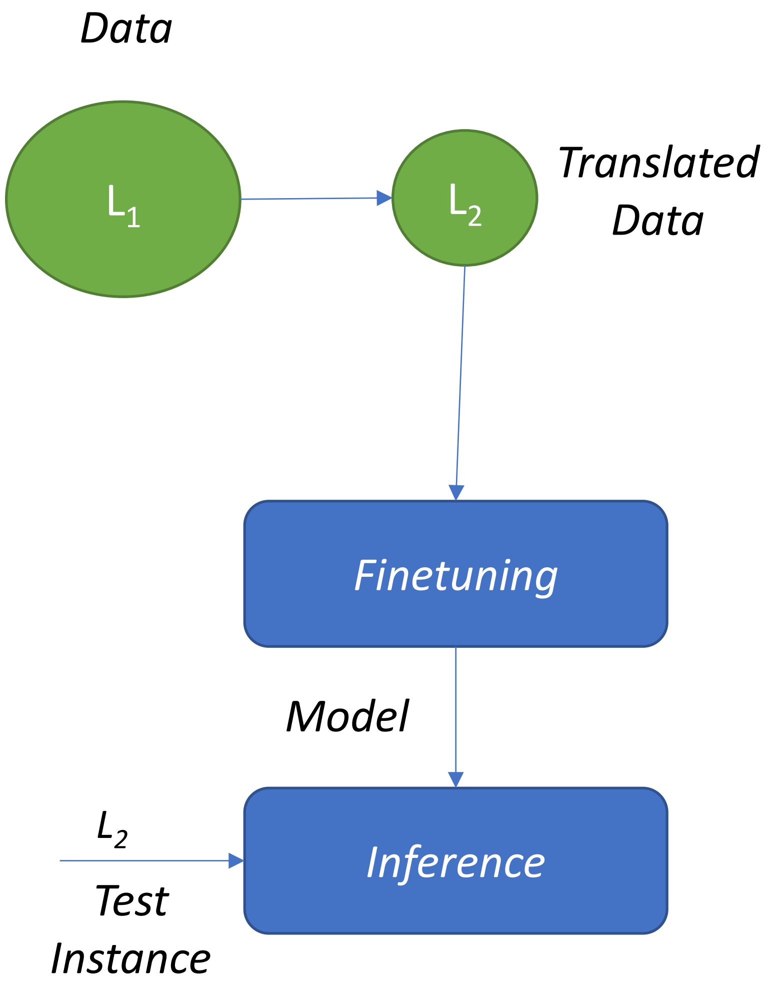
HiCTL

In terms of downstream tasks, retrieval tasks gain the most – gains are modest for other tasks

Finetuning Strategies

Translation Baselines

Translate Train



Data

L_1

L_2

Translated Data

Finetuning

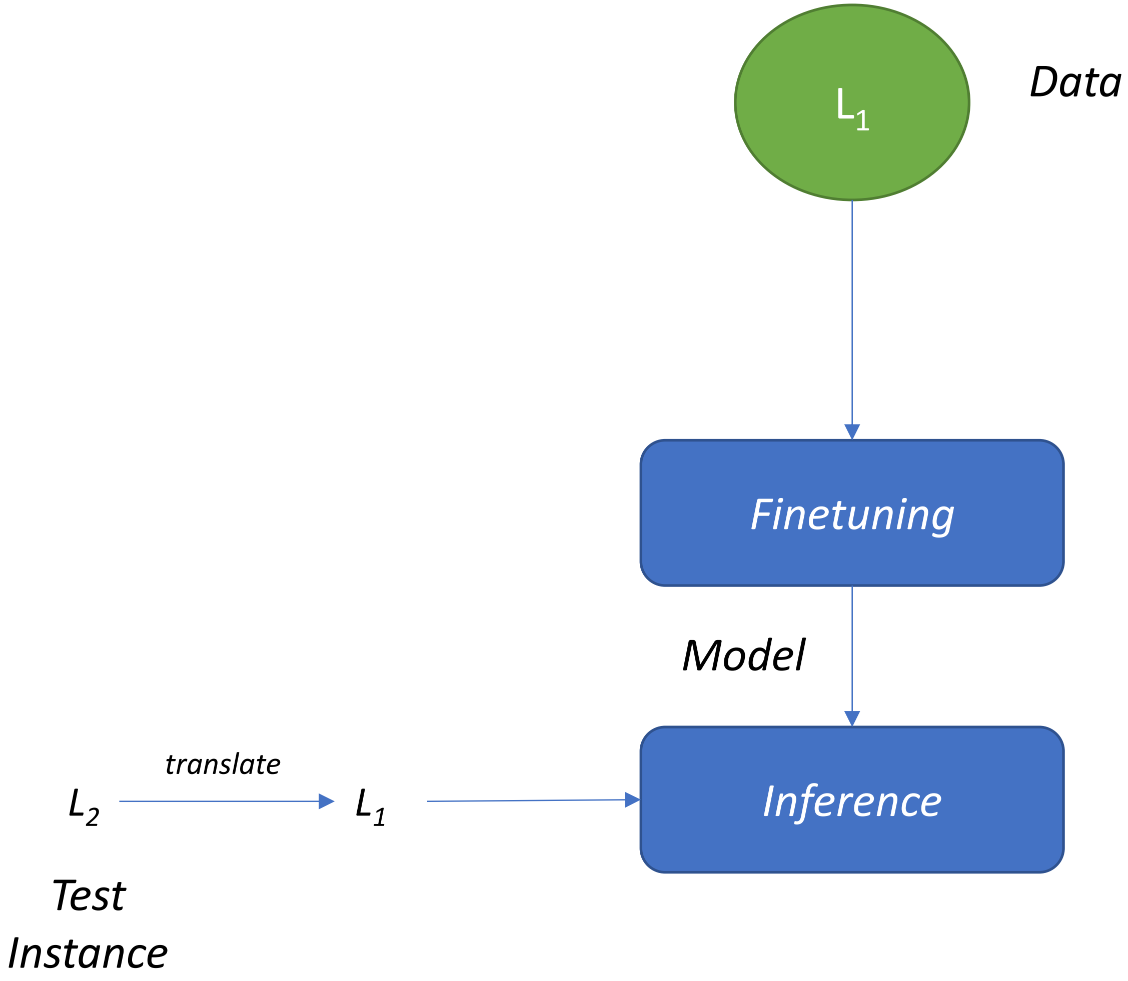
Model

L_2

Test Instance

Inference

Translate Test



L_1

Data

Finetuning

Model

translate

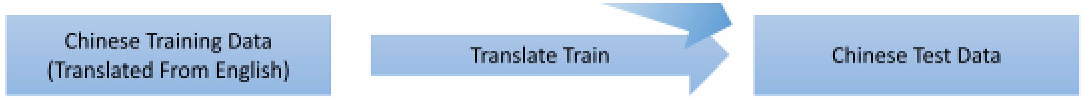
L_2

L_1

Test Instance

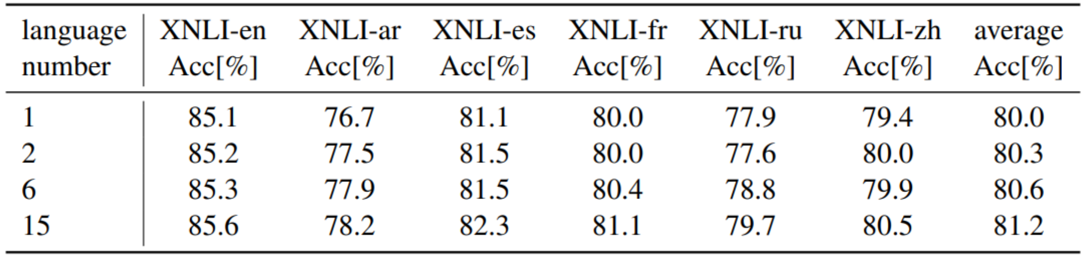
Inference

Translate-Train


Chinese Training Data (Translated From English)

Translate Train

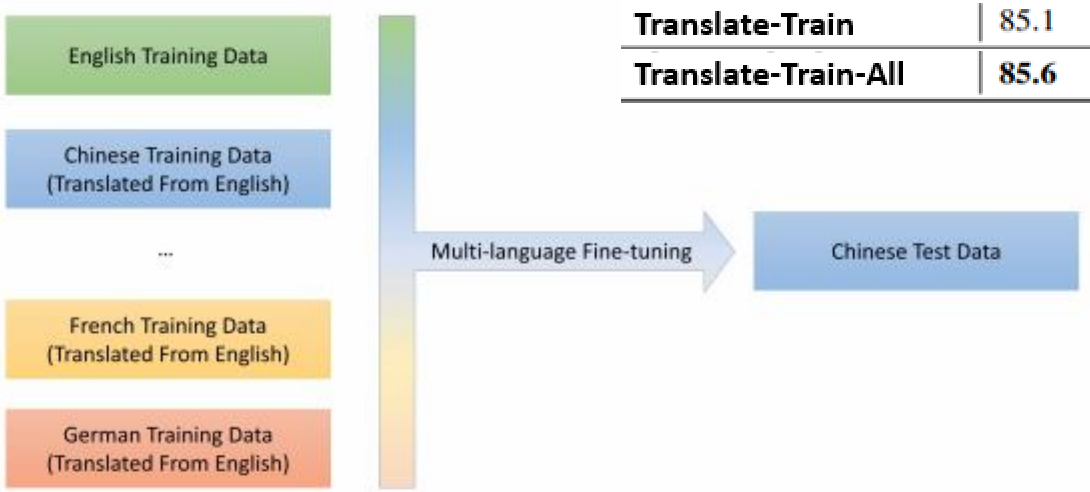
Chinese Test Data



language number	XNLI-en Acc[%]	XNLI-ar Acc[%]	XNLI-es Acc[%]	XNLI-fr Acc[%]	XNLI-ru Acc[%]	XNLI-zh Acc[%]	average Acc[%]
1	85.1	76.7	81.1	80.0	77.9	79.4	80.0
2	85.2	77.5	81.5	80.0	77.6	80.0	80.3
6	85.3	77.9	81.5	80.4	78.8	79.9	80.6
15	85.6	78.2	82.3	81.1	79.7	80.5	81.2

Translate-Train-All

	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw
Zeroshot	85.1	79.0	79.4	77.8	77.2	77.2	76.3	72.8	73.5	76.4	73.6	76.2	69.4	69.7
Translate-Train	85.1	80.0	81.1	79.9	77.7	80.2	77.9	75.3	76.7	76.4	75.2	79.4	71.8	71.8
Translate-Train-All	**85.6**	**81.1**	**82.3**	**80.9**	**79.5**	**81.4**	**79.7**	**76.8**	**78.2**	**77.9**	**77.1**	**80.5**	**73.4**	**73.8**


English Training Data

Chinese Training Data (Translated From English)

...

French Training Data (Translated From English)

German Training Data (Translated From English)

Multi-language Fine-tuning

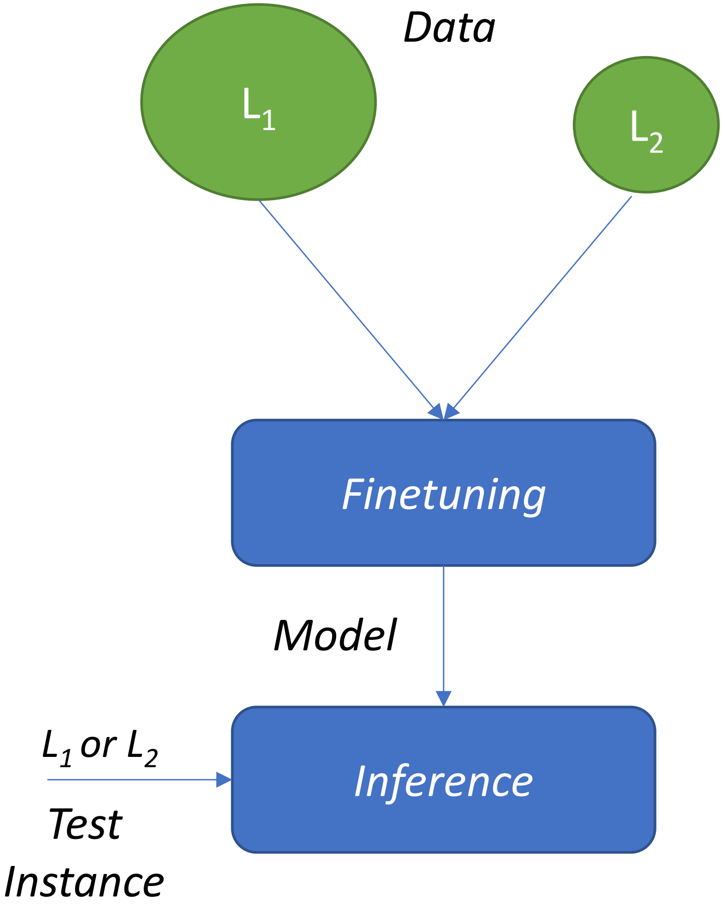
Chinese Test Data

Translating all training data works best

Above results on Unicoder for XNLI

Finetuning Scenarios

Joint Learning



L_1

Data

L_2

Finetuning

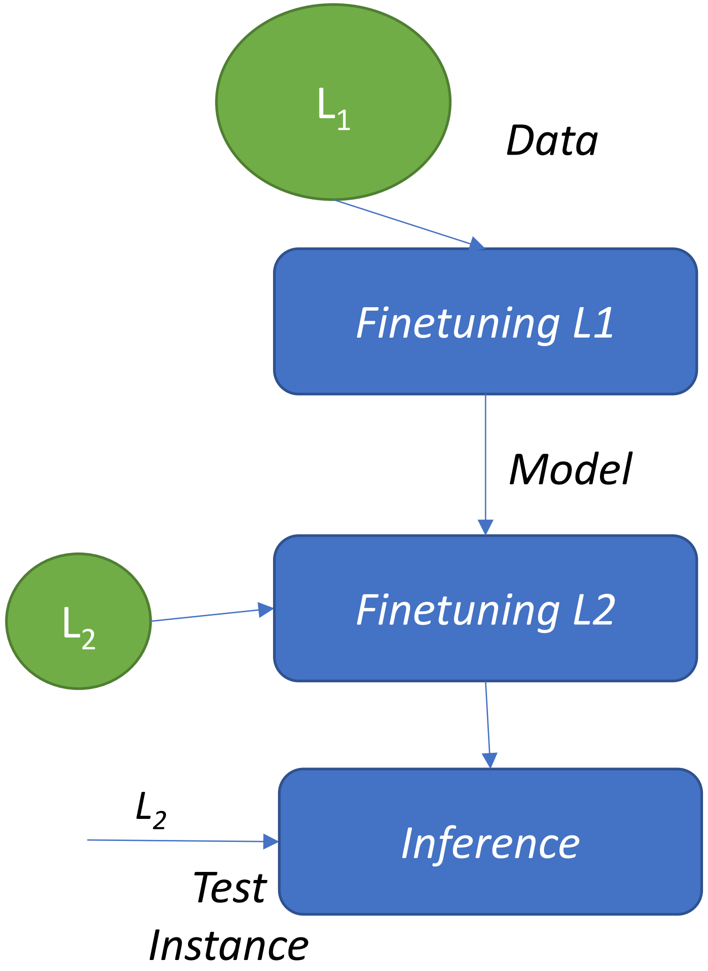
Model

L_1 or L_2

Test Instance

Inference

Multi-stage Learning



L_1

Data

Finetuning L1

Model

L_2

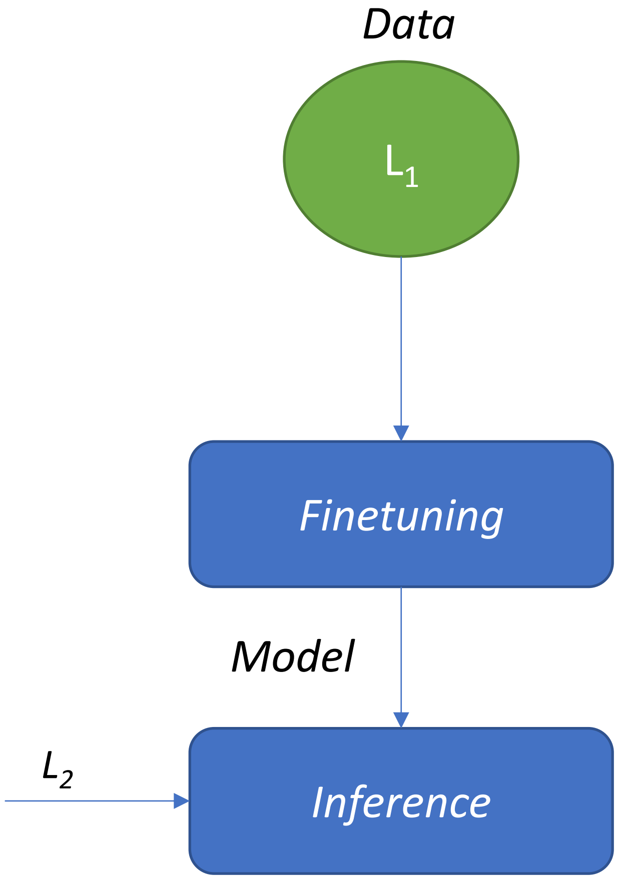
Finetuning L2

L_2

Test Instance

Inference

Zeroshot Learning



Data

L_1

Finetuning

Model

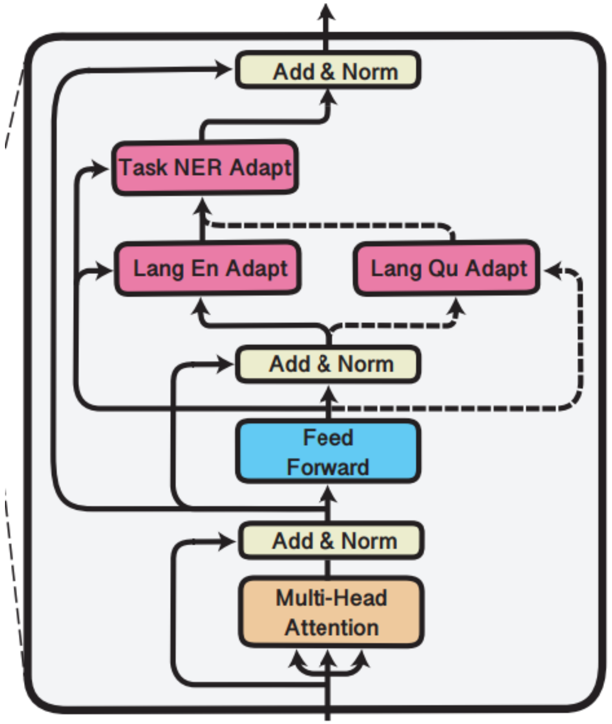
L_2

Inference

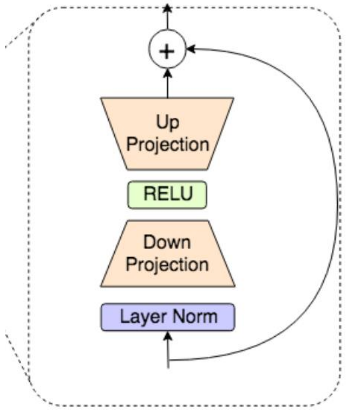
Adapters

Multilingual adapters

Pfeffer et al. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. EMNLP 2020.



Adapter layer



<u>Limited capacity to share amongst multiple languages</u>

- *Low-resource language don't get enough representation*
- *Performance on high resource languages suffers*

<u>Introduce small adapter networks for specific languages/tasks</u>

- *Better utilization of model capacity*
- *Adapt the model to new languages*
- *Parameter efficient adapters*

Train baseline model
➔ *Freeze parameters*
➔ *Introduce adapter modules*
➔ *Finetune adapter parameters*

***adapter-transformers + Adapter Hub*: https://adapterhub.ml**

Agenda

- Introduction
- Typical Multilingual NLP Pipeline
- Cross-lingual Transfer
- Multilingual Word Embeddings
- Multilingual Pre-trained Language Models
  - Extending English LLMs to new languages
- Finetuning Pre-trained LMs
- Evaluation of Multilingual models

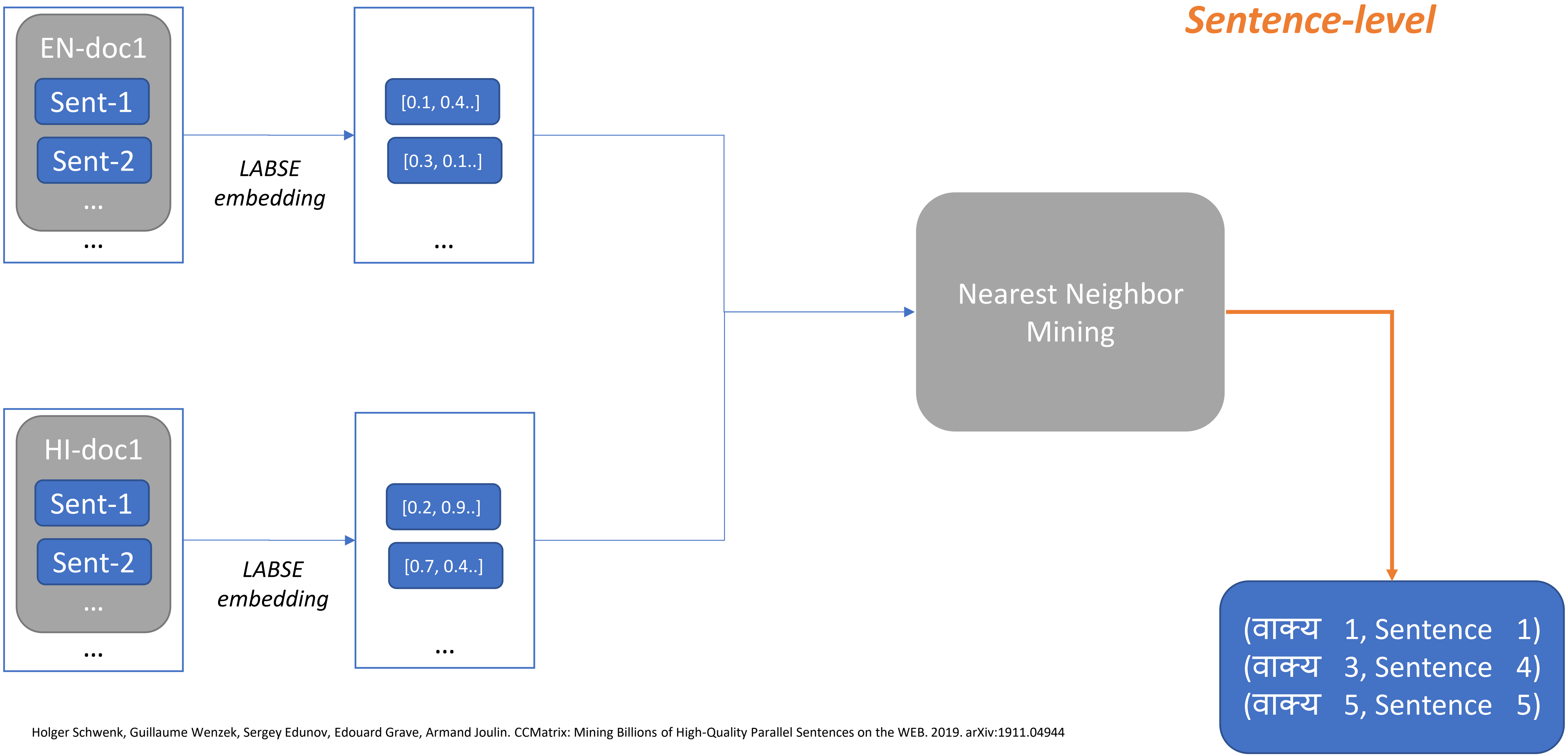
Mining Task-specific Datasets

In-language datasets are always useful ➔ expensive to create

Transfer learning does not work well for some problems. POS tagging, NER, etc.

Can datasets be mined from public sources?

Parallel Corpus Mining from Machine Translation



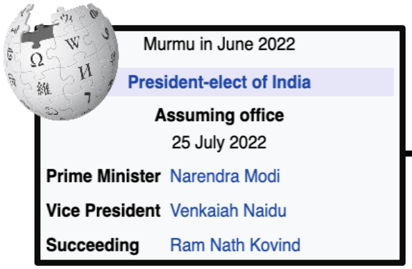
(वाक्य 1, Sentence 1)
(वाक्य 3, Sentence 4)
(वाक्य 5, Sentence 5)

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin. CCMatrix: Mining Billions of High-Quality Parallel Sentences on the WEB. 2019. arXiv:1911.04944

Mining NLU/NLG Datasets

Creativity is the limit for mining data of different kinds!

BIOGRAPHY GENERATION



Murmu in June 2022

President-elect of India

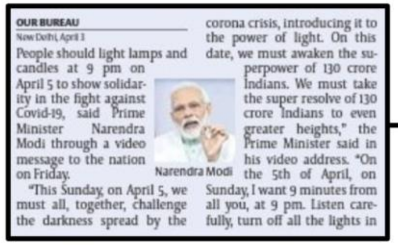
Assuming office
25 July 2022

Prime Minister	Narendra Modi
Vice President	Venkaiah Naidu
Succeeding	Ram Nath Kovind

→

Murmu assumed office on 25th July and succeeded Ram Nath Kovind

HEADLINE GENERATION



OUR BUREAU
New Delhi, April 3

People should light lamps and candles at 9 pm on April 5 to show solidarity in the fight against Covid-19, said Prime Minister Narendra Modi through a video message to the nation on Friday.

"This Sunday, on April 5, we must all, together, challenge the darkness spread by the corona crisis, introducing it to the power of light. On this date, we must awaken the superpower of 130 crore Indians. We must take the super resolve of 130 crore Indians to even greater heights," the Prime Minister said in his video address. "On the 5th of April, on Sunday, I want 9 minutes from all you, at 9 pm. Listen carefully, turn off all the lights in

Narendra Modi

→

Dispel the virus darkness with light: PM to people

SENTENCE SUMMARIZATION

India 's financial markets are closed on Monday for a public holiday.

→

India markets closed for holiday

PARAPHRASE GENERATION

दिल्ली विश्वविद्यालय, भारत में उच्च शिक्षा के लिए एक प्रतिष्ठित संस्थान है।

→

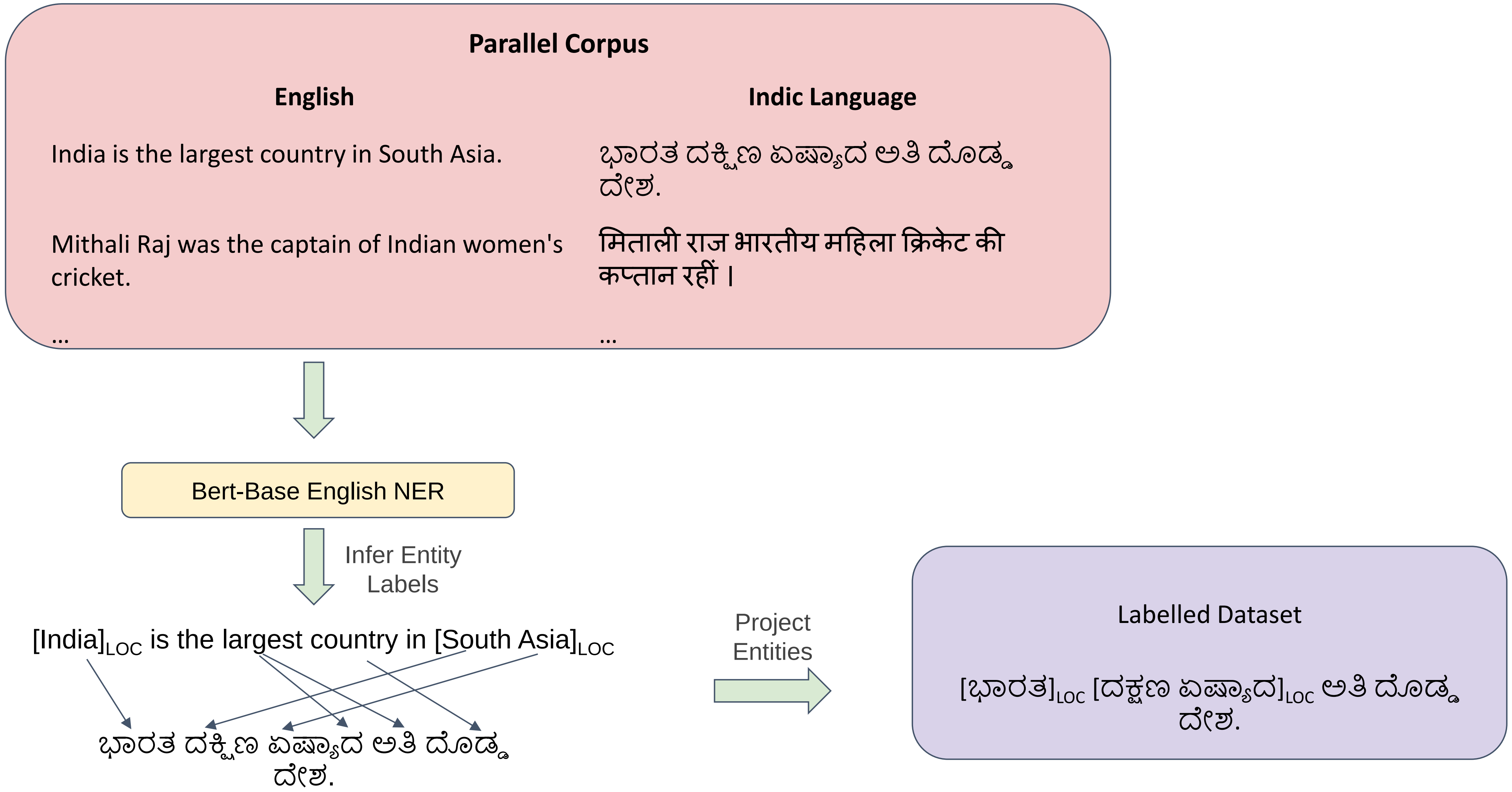
The University of Delhi is a prestigious institution for higher education in India.

Delhi University is one of the famous universities of the country.

QUESTION GENERATION

SQuAD → INDICTRANS → HI SQuAD

Mining Named Entities



Parallel Corpus

English	**Indic Language**
India is the largest country in South Asia.	ಭಾರತ ದಕ್ಷಿಣ ಏಷ್ಯಾದ ಅತಿ ದೊಡ್ಡ ದೇಶ.
Mithali Raj was the captain of Indian women's cricket.	मिताली राज भारतीय महिला क्रिकेट की कप्तान रहीं ।
…	…

Bert-Base English NER

Infer Entity Labels

[India]$_{\text{LOC}}$ is the largest country in [South Asia]$_{\text{LOC}}$

ಭಾರತ ದಕ್ಷಿಣ ಏಷ್ಯಾದ ಅತಿ ದೊಡ್ಡ ದೇಶ.

Project Entities

Labelled Dataset

[ಭಾರತ]$_{\text{LOC}}$ [ದಕ್ಷಣ ಏಷ್ಯಾದ]$_{\text{LOC}}$ ಅತಿ ದೊಡ್ಡ ದೇಶ.

Yarowsky, D., Ngai, G., and Wicentowski, R. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In Proceedings of the first International Conference on Human language technology research.

Summary

- Large model, large data, correct parametrization of models helpful
- Transfer learning works best between related languages
- Use of parallel data, sentence-level objectives to improve representation alignment
- Effect of parallel data
    - Retrieval tasks improve
    - Other tasks do not improve as much
    - Helpful for languages with different scripts

- Translate-train-all performs very well

- Have *expert* model components can improve performance

- Mine in-language data where you can ➔ very useful

Multilingual NLG

Introduction

- Multilingual extension to pre-trained NLG models
- Supporting tasks like Machine Translation, Summarization, Free-form Question Answering, Grammar Correction, Paraphrasing, etc.
- Desired Model Capabilities
  - Joint models for multiple languages
  - Transfer learning for low-resource scenarios
  - Zero-shot model performance
  - Cross-lingual tasks *e.g. document in one language, summary in another*
- *Generation stage needs to balance between*
  - *Language agnostic representation*
  - *Language specific surface realization*

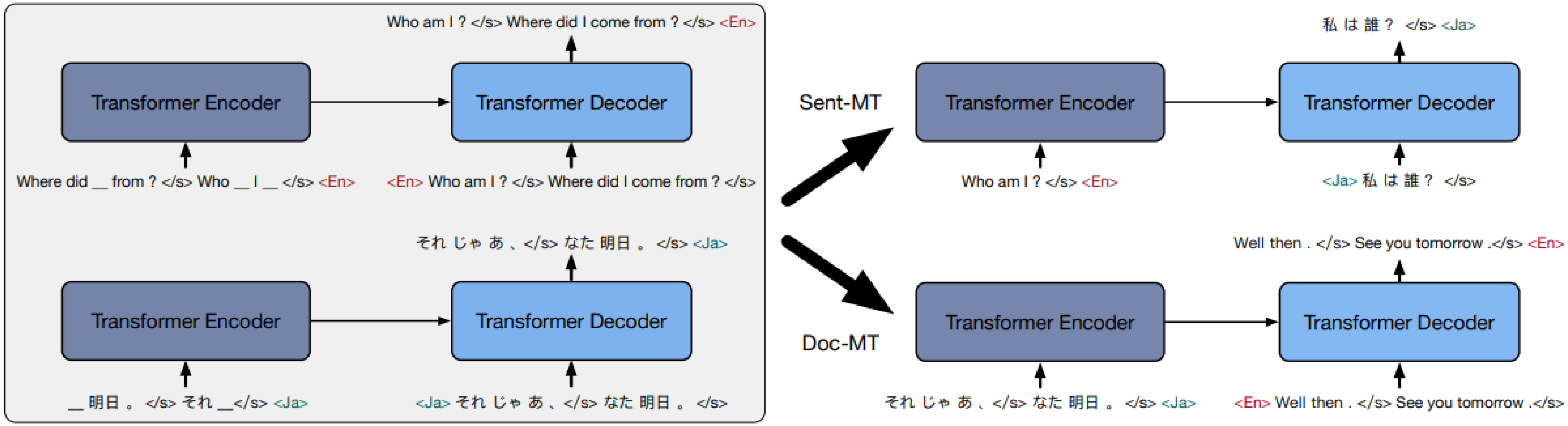
Encoder-Decoder Models: mBART/mT5

Additional Objectives
(1) MLM on encoder side
(2) Cross-lingual MLM & DAE objectives

- Simple modification to mBERT training ➔ Jointly train monolingual Denoising Auto-Encoder (DAE) objective
- Target language tag: Special token as input to identify target language to generate. Token can be:

  (1) In the input stream OR (2) initial “forced” decoder output

- Source language tag: input sequence *(optional)*

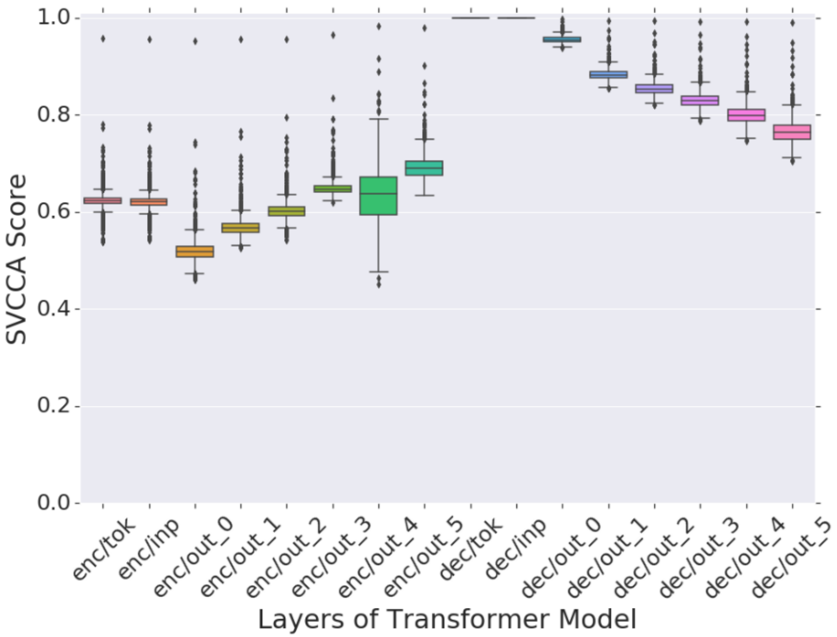

Who am I ? </s> Where did I come from ? </s> <En>
Transformer Encoder
Transformer Decoder
Where did __ from ? </s> Who __ I __ </s> <En>
<En> Who am I ? </s> Where did I come from ? </s>
それ じゃ あ 、</s> なた 明日 。 </s> <Ja>
Transformer Encoder
Transformer Decoder
__ 明日 。 </s> それ __</s> <Ja>
<Ja> それ じゃ あ 、</s> なた 明日 。 </s>
Sent-MT
私 は 誰 ？ </s> <Ja>
Transformer Encoder
Transformer Decoder
Who am I ? </s> <En>
<Ja> 私 は 誰 ？ </s>
Doc-MT
Well then . </s> See you tomorrow .</s> <En>
Transformer Encoder
Transformer Decoder
それ じゃ あ 、</s> なた 明日 。 </s> <Ja>
<En> Well then . </s> See you tomorrow .</s>

Multilingual Denoising Pre-Training (mBART)

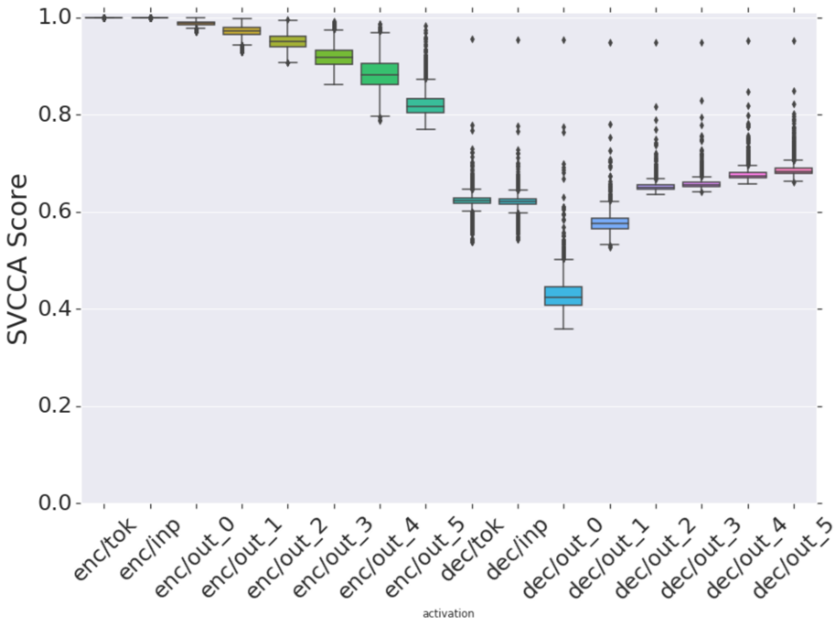
Fine-tuning on Machine Translation

Liu et al., *Multilingual denoising pre-training for neural machine translation*. TACL. 2020.
Xue et al. *mT5: A massively multilingual pre-trained text-to-text transformer.* 2020. https://arxiv.org/abs/2010.11934
Chi et al. *Cross-Lingual Natural Language Generation via Pre-Training. AAAI 2020.*
ZCODE

Cross-lingual Representations



(a) X-En Language Pairs



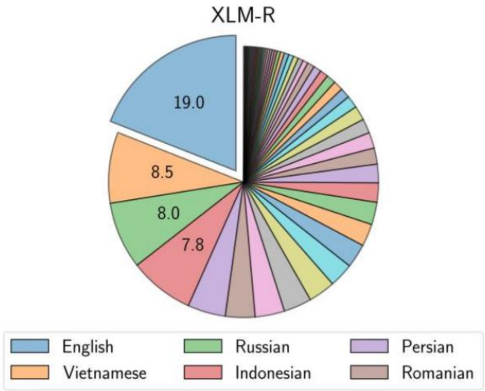
(b) En-X Language Pairs

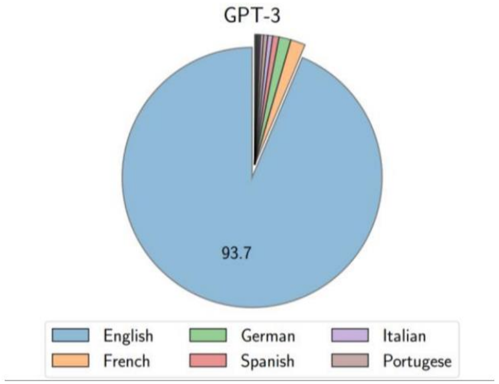
- Encoder and decoder representations become more language agnostic in higher layers
- English representations differ depending on the language on the other side
  - Encoder-decoder representation boundary is blurry

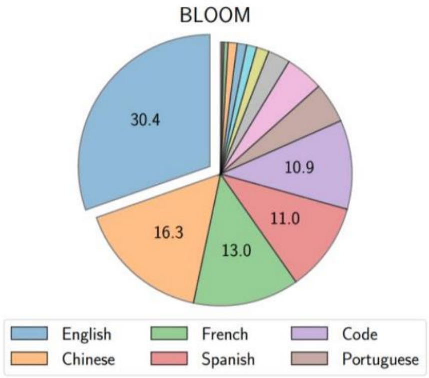
Kudugunta et al. Investigating multilkingual NMT representations at Scale. EMNLP 2019.

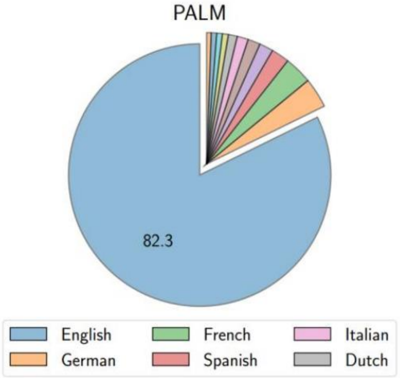
Decoder Models: BLOOM/xGLM/AYA/GPT

- Simple modification to mBERT training ➔
  - Jointly train Causal Language Modeling/Next word prediction objective
- Standard Architecture for LLMs
- Target language tokens not used for generation
- Flexible instruction following capabilities used to specify target language in prompt
- Consideration: English data far outstrips data from other languages



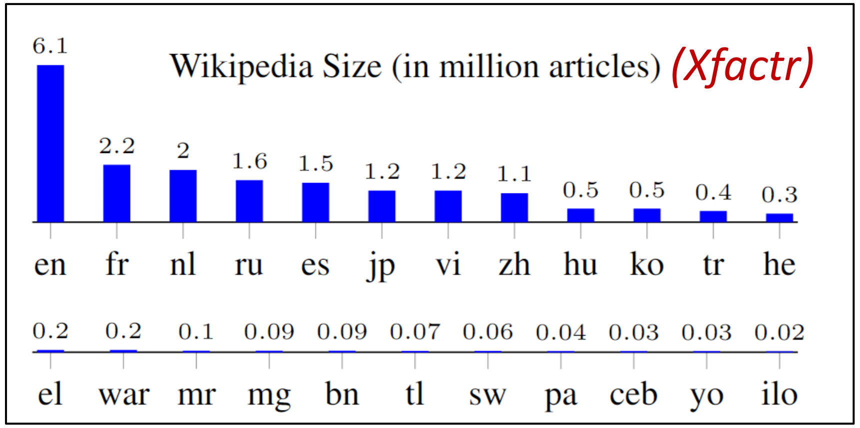






Multilingual LLMs

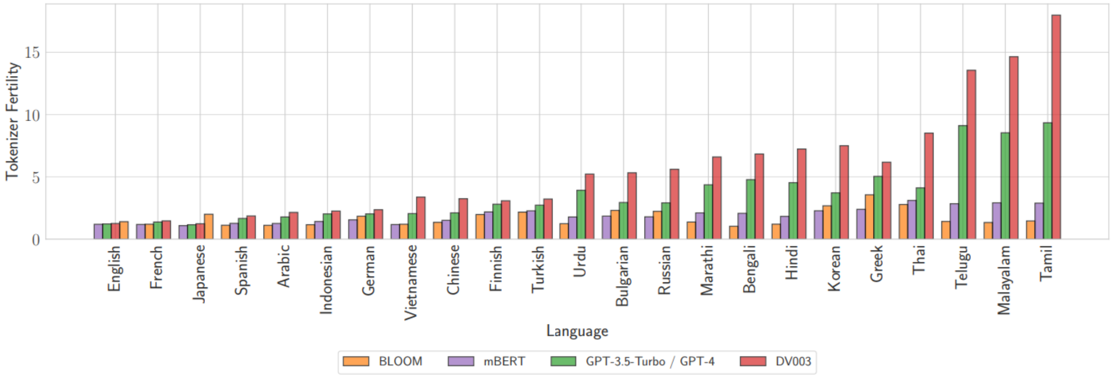
- Very few multilingual LLMs
- Most LLMs are very English heavy
  - Focus on English
  - English data far outstrips LLMs for other languages



(Xfactr)

Limited tokenizer representation for most non-English languages

Fertility ➔ number of tokens per word
High fertility ➔ low-efficiency, suboptimal representations



(BUFFET, MEGA, ChatGptMT)

- Do English-heavy LLMs have any non-English capabilities? How?
- Prohibitively expensive to train multilingual LLMs from scratch, is it possible to extend English LLMs to new languages?

Do English LLMs have some inherent multilingual capabilities?

Yes, to some extent …

Why? – during training they might have been exposed to some non-English data

- Documents with multiple languages
- Incorrect LID

How good are the multilingual capabilities?

- Might be ok at language understanding *e.g. classification, sentiment analysis*
- Bad at generation
- Better on Latin script languages
- Languages with better pre-training representation perform better

How do English LLM achieve multilingual capabilities?

- *Do LLMs think in English?*
- *Do LLM use English as a pivot for decision making?*

Bottom layers: Feature learning

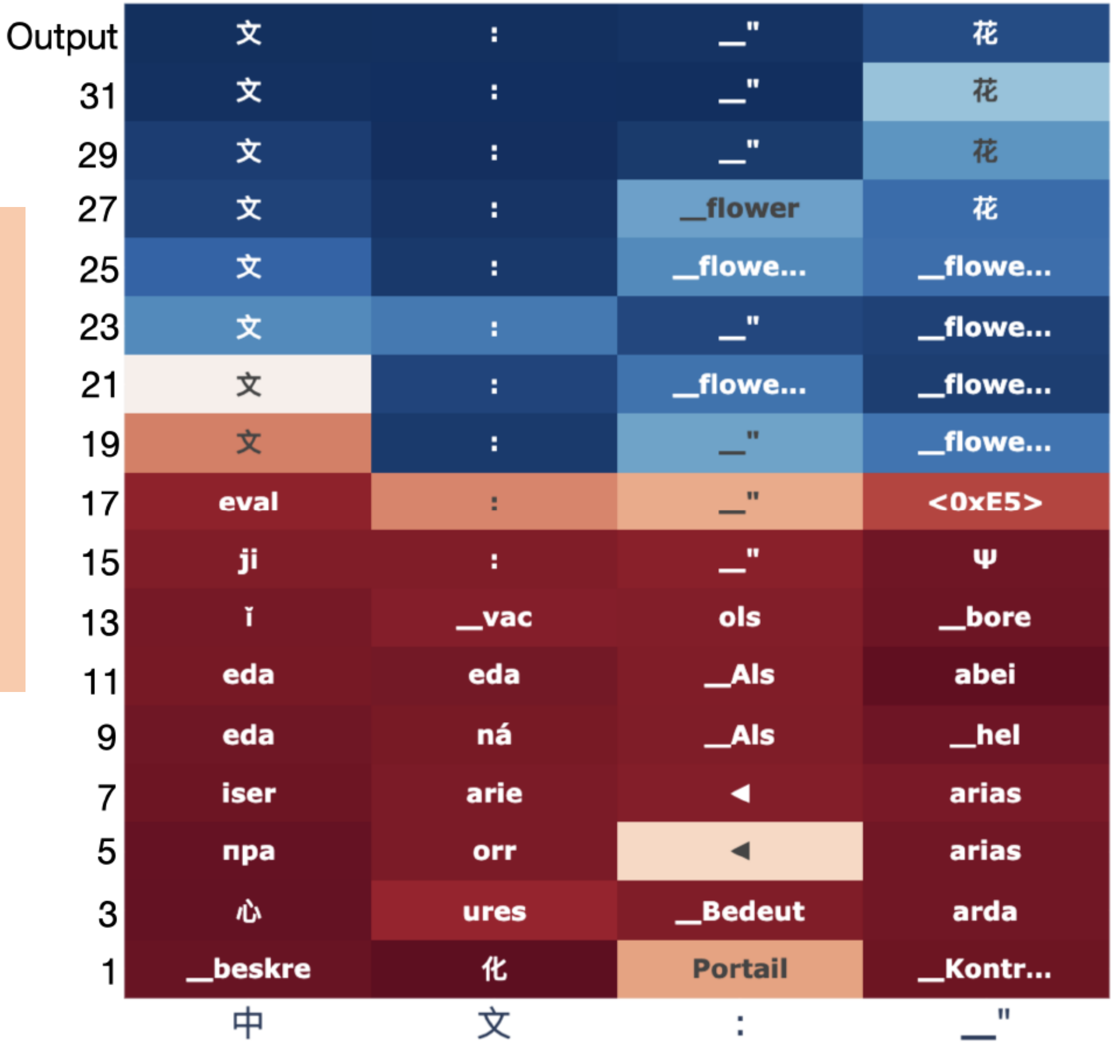
Middle layers: Concept mapping to language tokens (with English bias)

Top layers: Language generation in target language

There are language-specific neurons

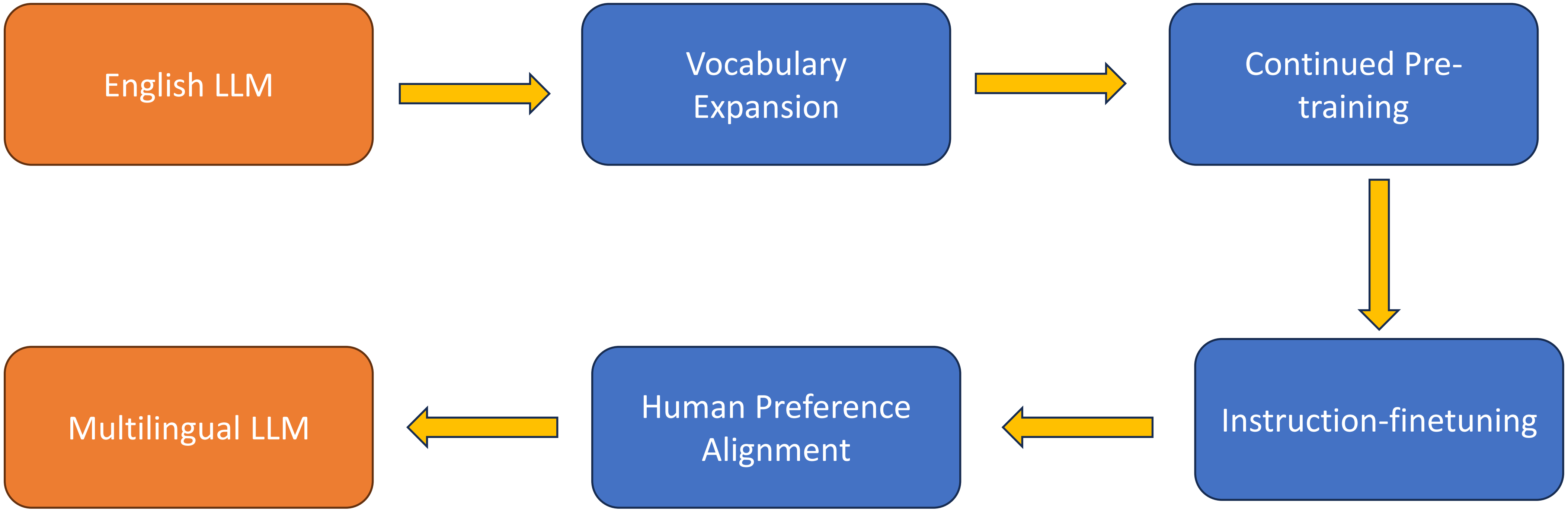
The central question in building multilingual LLM is to bring representations of English and other languages closer to achieve good cross-lingual transfer

(LmaLatent,PNLD,LSP)



Output	文	:	_"	花
31	文	:	_"	花
29	文	:	_"	花
27	文	:	_flower	花
25	文	:	_flowe...	_flowe...
23	文	:	_"	_flowe...
21	文	:	_flowe...	_flowe...
19	文	:	_"	_flowe...
17	eval	:	_"	<0xE5>
15	ji	:	_"	ψ
13	ĭ	_vac	ols	_bore
11	eda	eda	_Als	abei
9	eda	ná	_Als	_hel
7	iser	arie	◀	arias
5	пра	orr	◀	arias
3	心	ures	_Bedeut	arda
1	_beskre	化	Portail	_Kontr...
	中	文	:	_"

Extending English LLMs to Non-English Languages



English LLM

Vocabulary Expansion

Continued Pre-training

Instruction-finetuning

Human Preference Alignment

Multilingual LLM

✓ Large-scale, Document-level Datasets

? High Quality Documents

? Wide coverage of topics

✗ Representation of culture-specific data, native literature

✗ Capture data in different modalities and genres

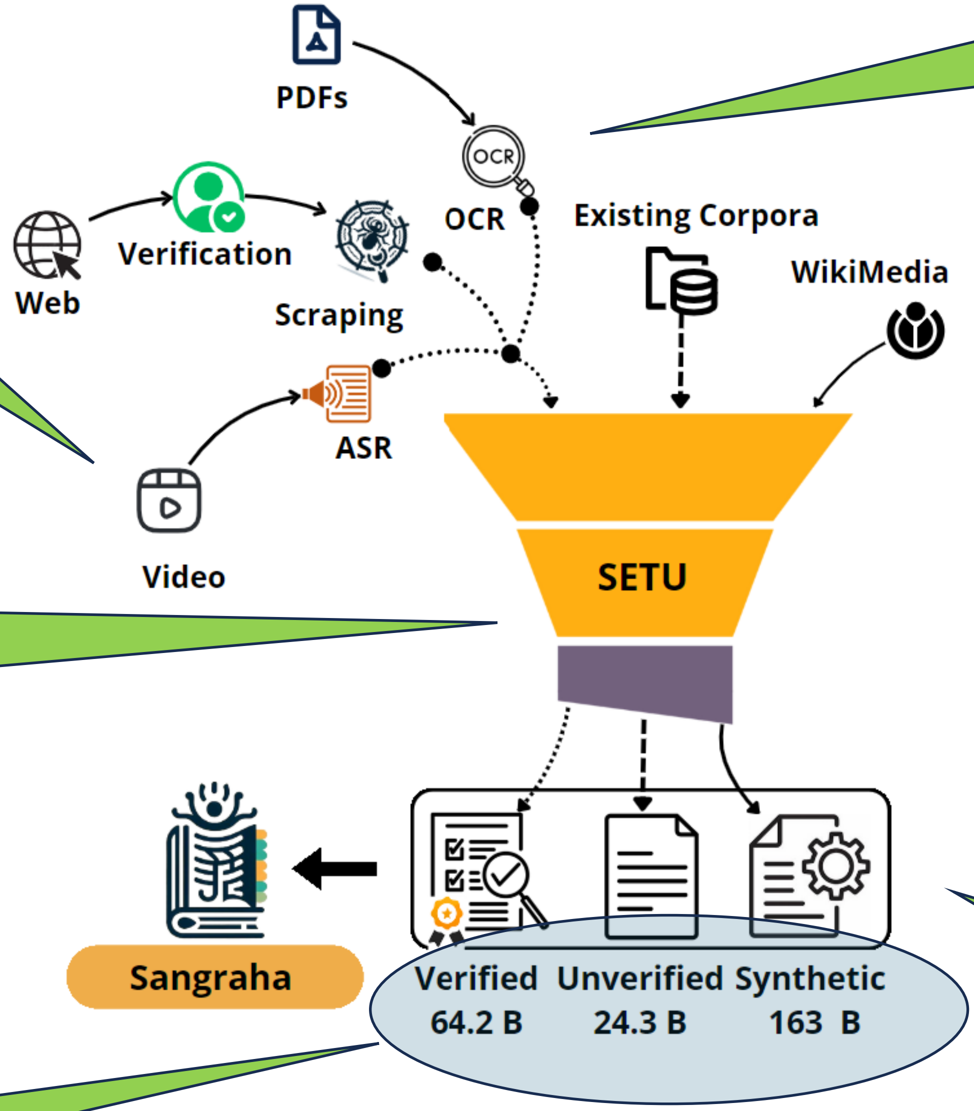
✗ Data to Help Cross-lingual transfer with English

Multilingual corpora like mC4, CC-100, CulturaX are good starting points

Build custom language (group) specific collections to address gaps

SANGRAHA

Creating high-quality pre-training data at scale for Indic languages



Data from different modalities – audio/OCR

OCR to unlock diverse regional and cultural content in PDFs

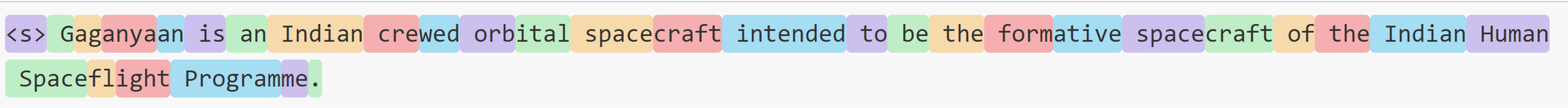
Filtering pipeline to ensure high-quality

High –quality sources

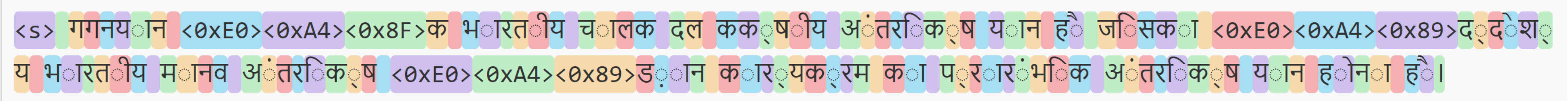
Translations and Transliterations to enable cross-lingual transfer

(IndicLLMSuite)

What if vocabulary is under-represented?



<s> Gaganyaan is an Indian crewed orbital spacecraft intended to be the formative spacecraft of the Indian Human Spaceflight Programme.



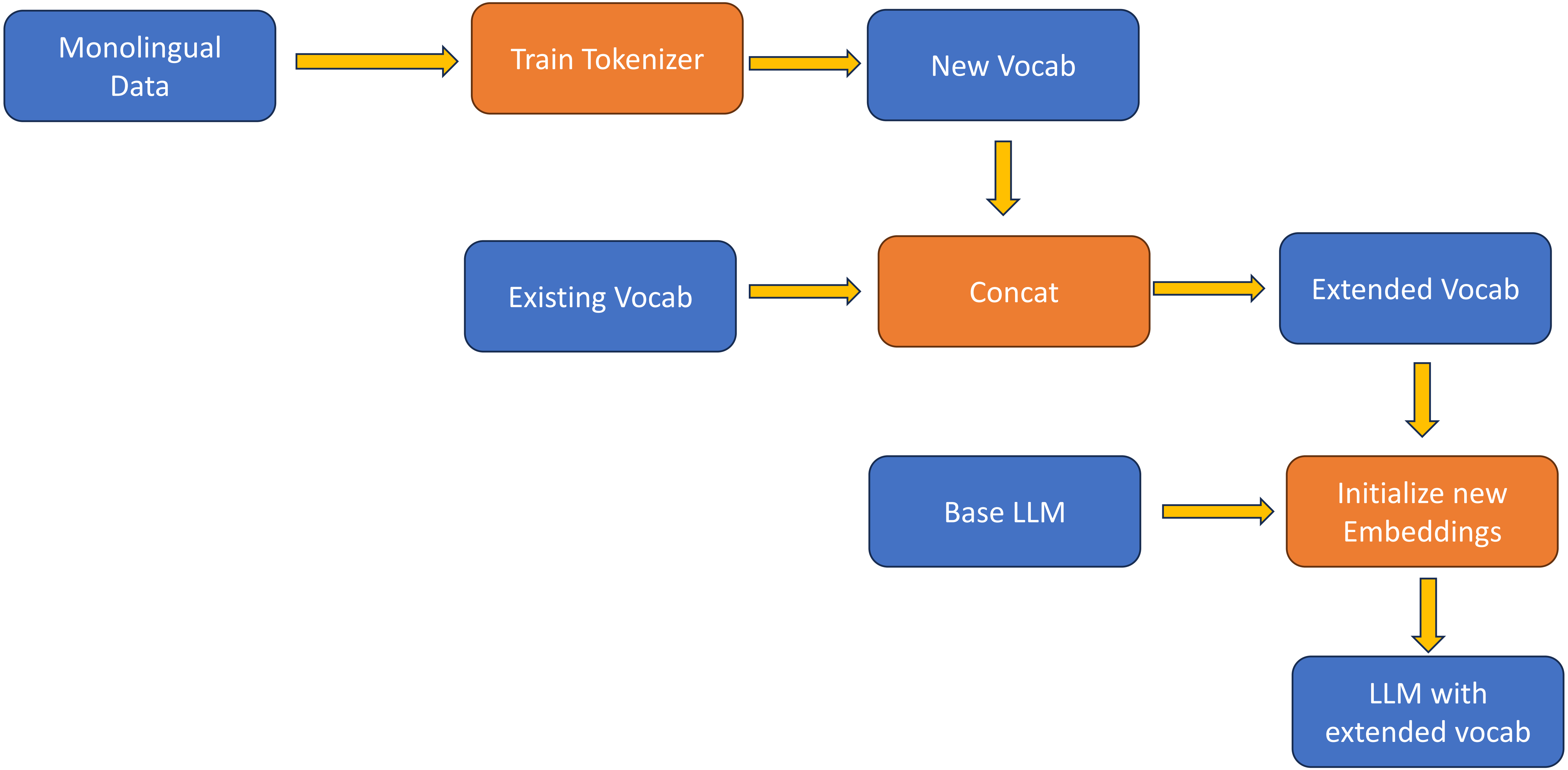
<s> गगनयान <0xE0><0xA4><0x8F>क भारतीय चालक दल कक्षीय अंतरिक्ष यान है जिसका <0xE0><0xA4><0x89>द्देश्य भारतीय मानव अंतरिक्ष <0xE0><0xA4><0x89>ड़ान कार्यक्रम का प्रारंभिक अंतरिक्ष यान होना है।

Fertility = Average number of tokens per word

High Fertility → More memory consumption, more decoding time, limit on longest processable sequence

Model	Hindi Fertility
GPT4	5.32
Llama2	5.83
Mistral	5.60
BLOOM	1.38

How to extend tokenizer vocabulary?



Monolingual Data

Train Tokenizer

New Vocab

Existing Vocab

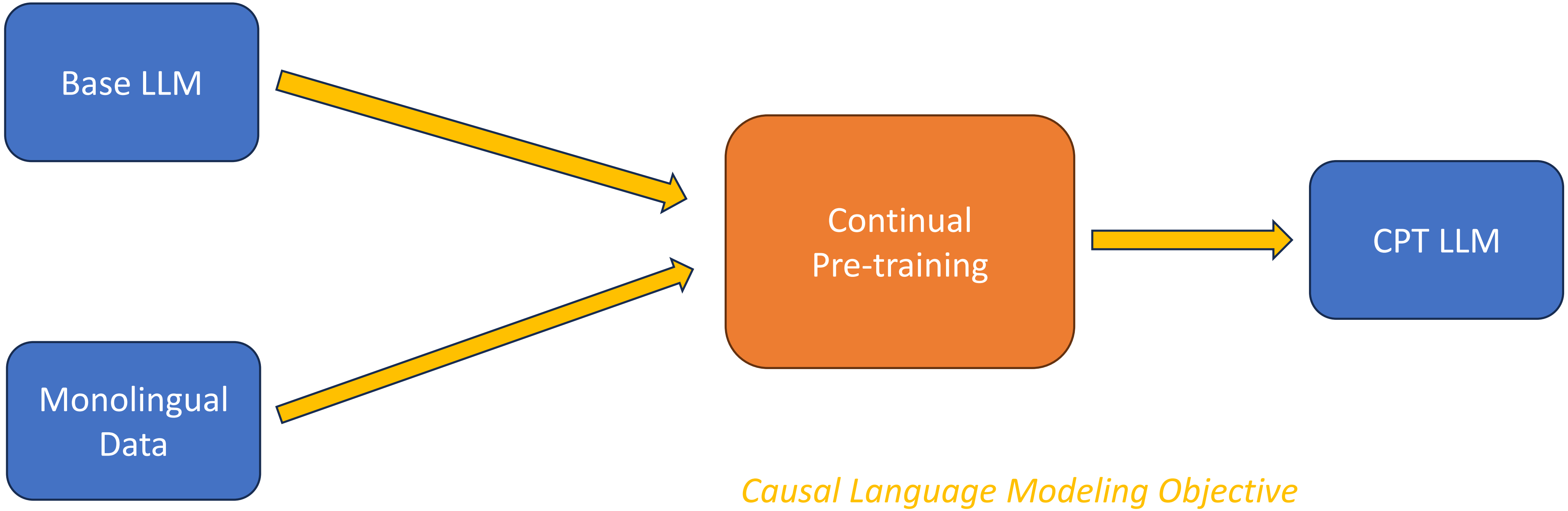
Concat

Extended Vocab

Base LLM

Initialize new Embeddings

LLM with extended vocab



Train on document-level data

Finetuning on long, coherent sequences helps model learn and correlate different pieces of knowledge

Causal Language Modeling Objective

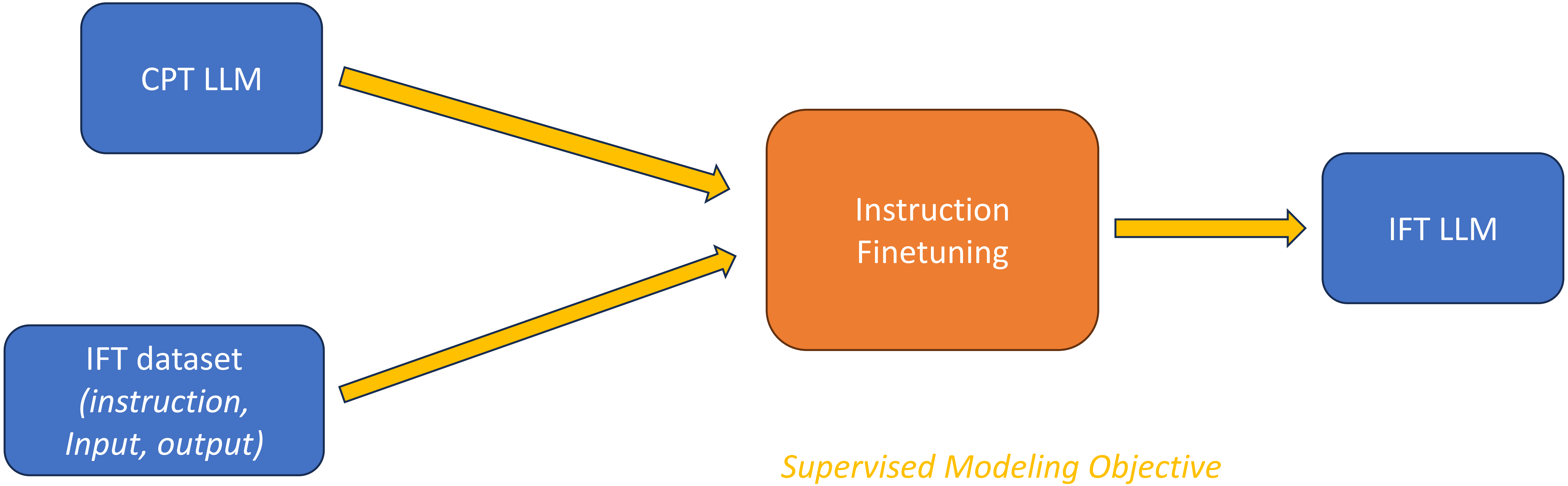
$$p(\mathbf{x}) = p(x_1, x_2, ..., x_T) = \prod_{t=1}^{T} p(x_t|\mathbf{x}_{<t})$$

To avoid forgetting English competence and knowledge

- Include English in the pre-training data

To align English and new language

- Pre-train on parallel data
- Pre-train using romanized data
- *(ChineseLLama, OpenHathi,RomanSetu,IndicMonoDoc,Tower,Palm2)*



CPT LLM

IFT dataset
*(instruction,
Input, output)*

Instruction
Finetuning

IFT LLM

Train on in-language IFT dataset

Sources of IFT dataset
Quality and diversity of IFT dataset

Supervised Modeling Objective

$$\ell_{\mathrm{CE}}(\mathbf{y},\hat{\mathbf{y}}) = -\sum_{j=1}^{|\mathcal{V}|} y_j \log(\hat{y}_j) \qquad \mathcal{L}_{\mathrm{SFT}} = \frac{1}{N}\sum_{i=1}^{N} \ell_{\mathrm{CE}}(\mathbf{y}_i, \mathcal{M}_\theta(\mathbf{x}_i))$$

To retain English task performance

- Include English in the IFT training

Instruction Tuning Tasks

Variety of tasks/objectives to improve non-English performance

Generating IFT Data

- English Data IFT *(SDRRL)*
- In-language IFT with Machine Translated Data *(Airavat,Okapi)*
- Locally/Culturally relevant IFT data *(SeaLLM,IndicLLMSuite)*

Auxiliary Tasks

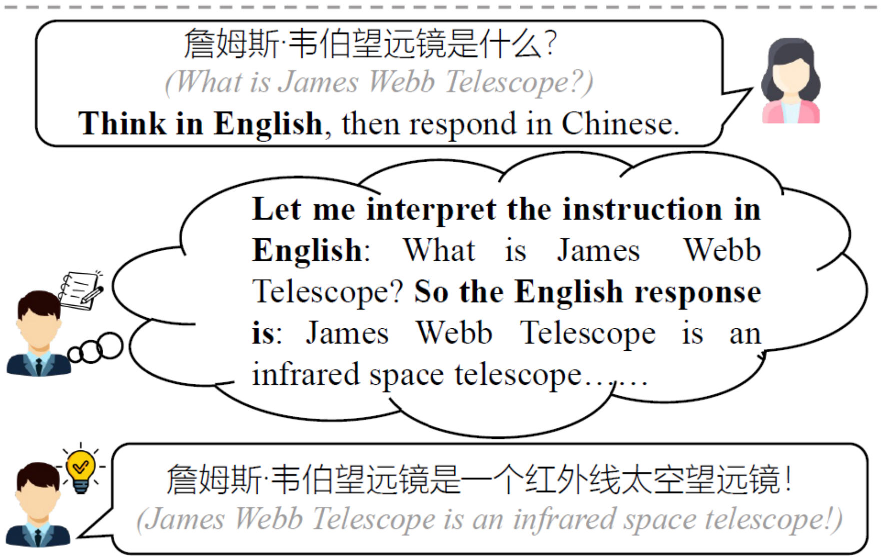
- Parallel Data
- Monolingual Data *(SDRRL)*

Transforming IFT Datasets

- Romanized IFT Data *(RomanSetu)*
- Cross-lingual Thought Data *(PLUG)*
- Cross-lingual IFT Data *(SDDRL)*
- Code-switched IFT Data

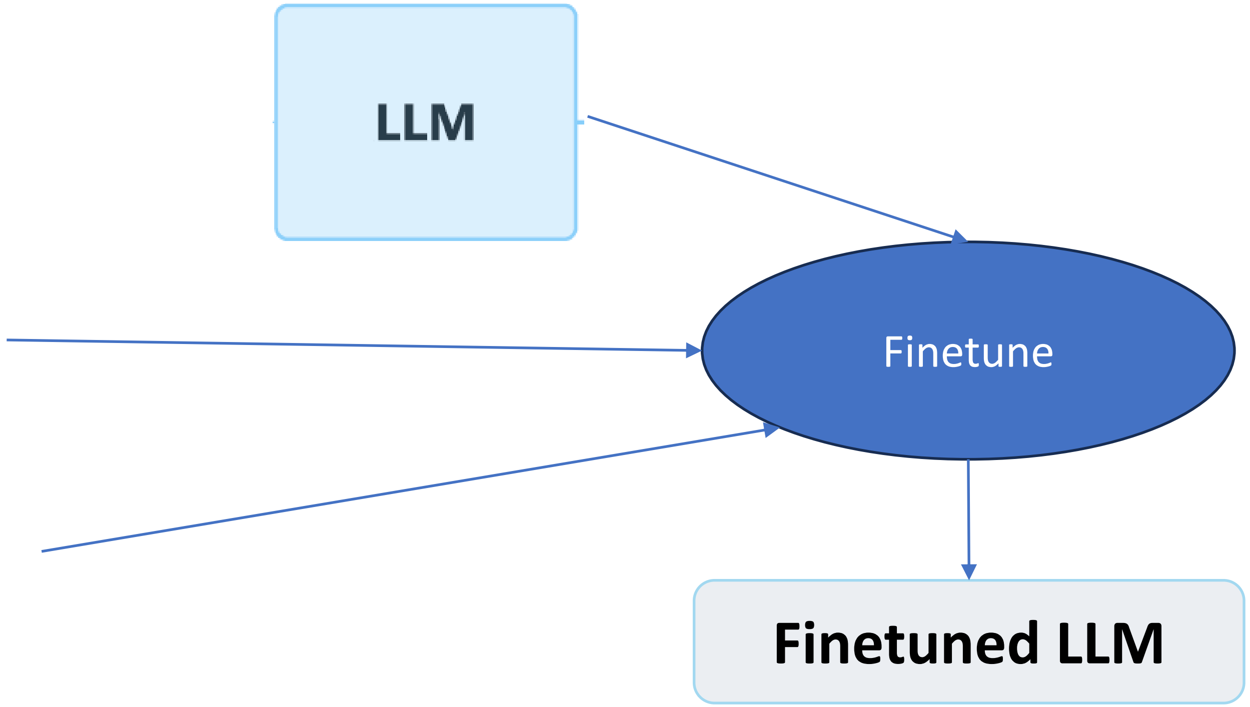
Let’s look at some of these tasks in detail

Cross-Lingual Thought data


詹姆斯·韦伯望远镜是什么?
(What is James Webb Telescope?)
Think in English, then respond in Chinese.

Let me interpret the instruction in English: What is James Webb Telescope? **So the English response is**: James Webb Telescope is an infrared space telescope……

詹姆斯·韦伯望远镜是一个红外线太空望远镜!
(James Webb Telescope is an infrared space telescope!)



Limitations

- Reduced maximum sequence length
- Increased latency

	Standard IFT	TaCo
Nepali	39.3	88.1
Persian	46.5	83.9

Evaluation on translated Vicuna Benchmark with GPT4

The PLUG paper also shows improvements (discussed later)

(TaCo,PLUG)

Cross-Lingual Instruction Tuning

	English	*Non-English*
In-language instruction sets	(Q_{en}, A_{en})	(Q_{hi}, A_{hi})
Cross-language instruction sets	(Q_{hi}, A_{en})	(Q_{en}, A_{hi})

Jointly optimize all these objectives

Q_{hi}: मैं अपने समय प्रबंधन कौशल को कैसे सुधार सकता हूँ? मुझे 5 संक्षिप्त बिंदु दीजिये।

A_{en}:
1. Prioritize tasks based on importance and urgency.
2. Set specific goals with deadlines (SMART goals).
3. Use time management tools like calendars and apps.
4. Minimize distractions to maintain focus.
5. Regularly review and adjust your schedule for better efficiency.

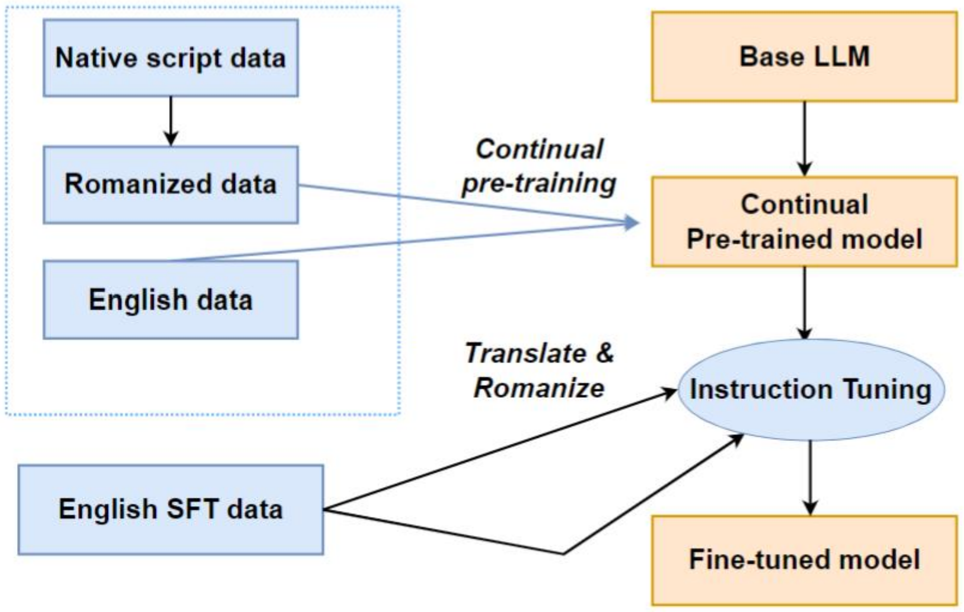
Cross-lingual instruction help the model to align languages and think in English

Avoids intermediate English generation of Cross-lingual Thought Prompting

(SDDRL,XCOT,CIT)

Romanized Representation

Just like pre-training, use romanized representation for IFT too



Native script data

Romanized data

English data

Continual pre-training

Base LLM

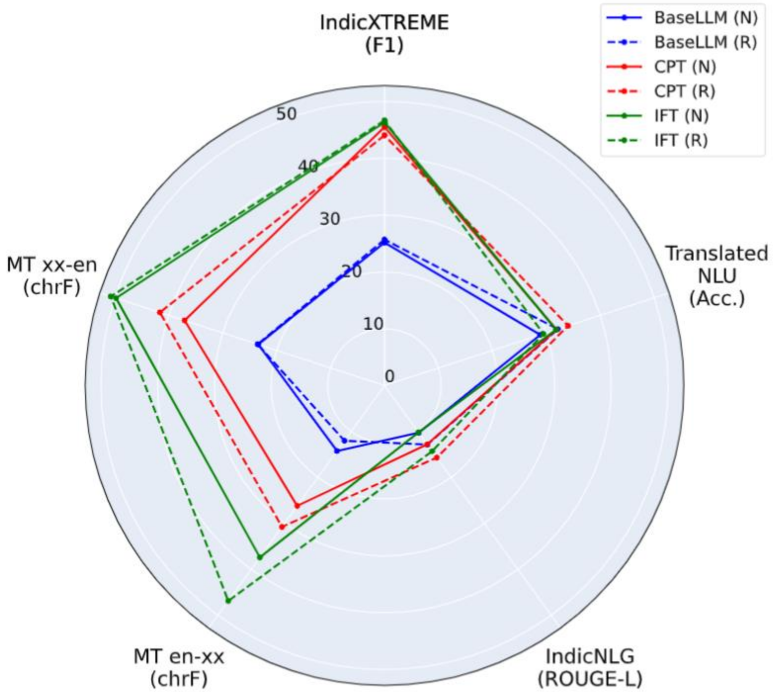
Continual Pre-trained model

Translate & Romanize

Instruction Tuning

English SFT data

Fine-tuned model



- *Continual Pre-training with romanized data is crucial*
- *NLG task performance improves with romanized data*
- *NLU task performance is on par, though more efficient*

(RomanSetu)

Agenda

- Introduction
- Typical Multilingual NLP Pipeline
- Cross-lingual Transfer
- Multilingual Word Embeddings
- Multilingual Pre-trained Language Models
  - Extending English LLMs to new languages
- Finetuning Pre-trained LMs
- Evaluation of Multilingual models

Benchmarks for Multilingual Learning

XTREME

Cross-lingual TRansfer Evaluation of Multilingual Encoders (XTREME) benchmark

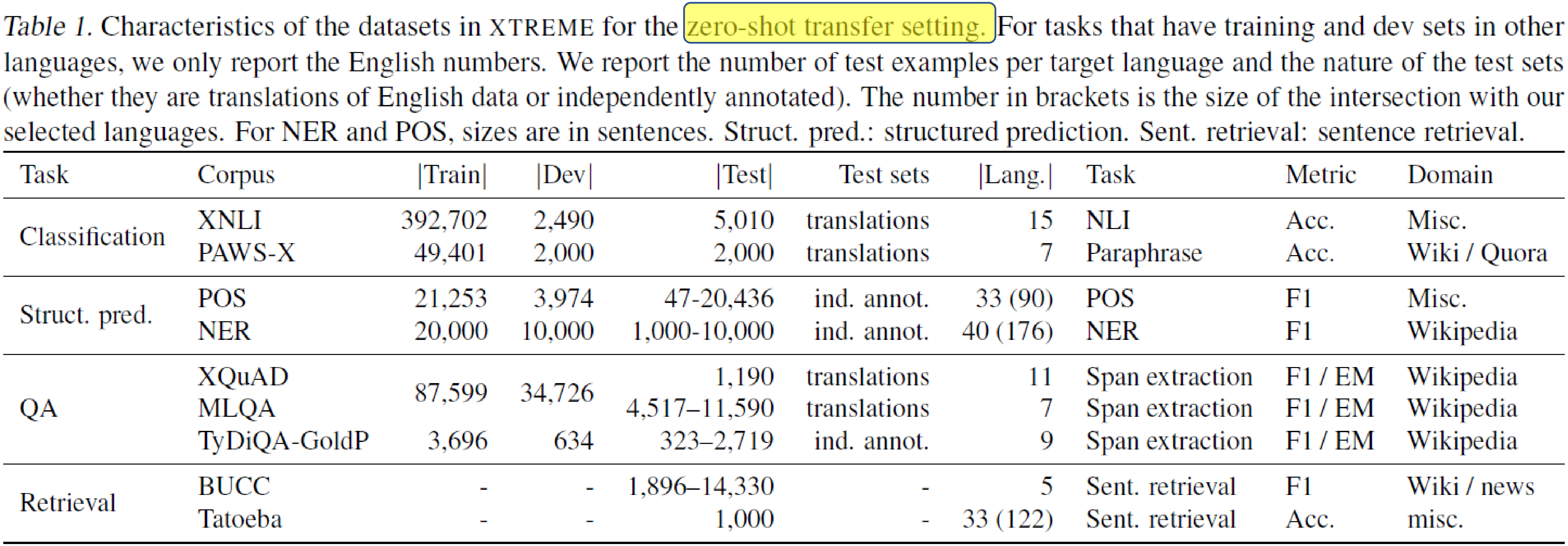


Table 1. Characteristics of the datasets in XTREME for the zero-shot transfer setting. For tasks that have training and dev sets in other languages, we only report the English numbers. We report the number of test examples per target language and the nature of the test sets (whether they are translations of English data or independently annotated). The number in brackets is the size of the intersection with our selected languages. For NER and POS, sizes are in sentences. Struct. pred.: structured prediction. Sent. retrieval: sentence retrieval.

Task	Corpus	|Train|	|Dev|	|Test|	Test sets	|Lang.|	Task	Metric	Domain
Classification	XNLI	392,702	2,490	5,010	translations	15	NLI	Acc.	Misc.
	PAWS-X	49,401	2,000	2,000	translations	7	Paraphrase	Acc.	Wiki / Quora
Struct. pred.	POS	21,253	3,974	47-20,436	ind. annot.	33 (90)	POS	F1	Misc.
	NER	20,000	10,000	1,000-10,000	ind. annot.	40 (176)	NER	F1	Wikipedia
QA	XQuAD	87,599	34,726	1,190	translations	11	Span extraction	F1 / EM	Wikipedia
	MLQA			4,517–11,590	translations	7	Span extraction	F1 / EM	Wikipedia
	TyDiQA-GoldP	3,696	634	323–2,719	ind. annot.	9	Span extraction	F1 / EM	Wikipedia
Retrieval	BUCC	-	-	1,896–14,330	-	5	Sent. retrieval	F1	Wiki / news
	Tatoeba	-	-	1,000	-	33 (122)	Sent. retrieval	Acc.	misc.

Task Difficulty, Task Diversity, Language Diversity, Sufficient Monolingual data, Efficiency, Accessibility

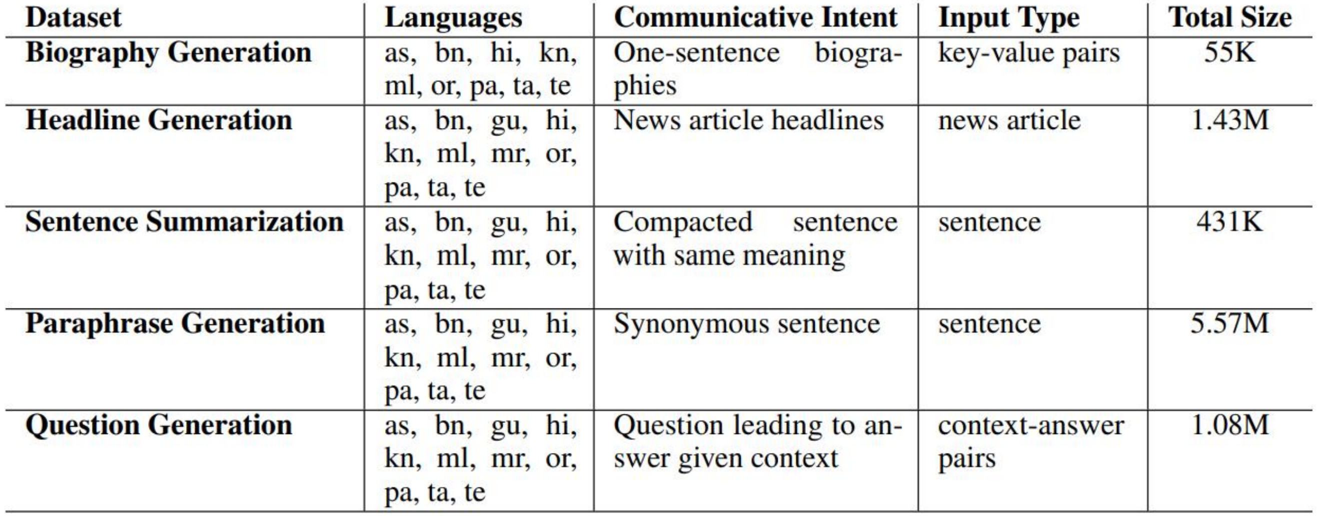
40 languages, 12 language families, 9 tasks, 4 task types

NLG Benchmarks

IndicNLG Benchmark

5 datasets in 11 Indian languages

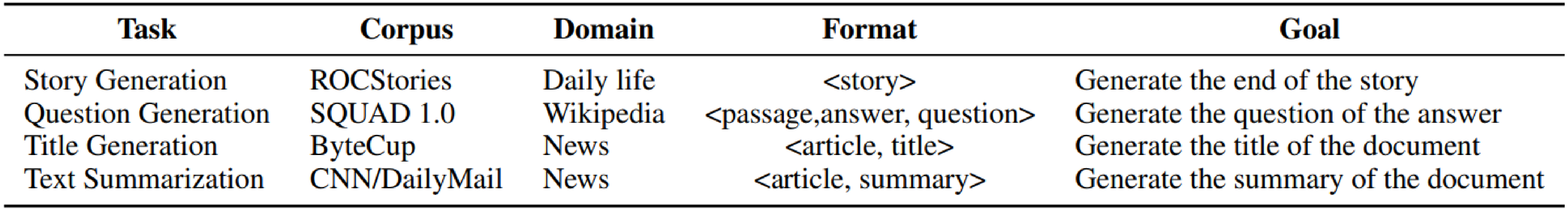
Aman Kumar, Himani Shrotriya, Prachi Sahu, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, Mitesh M. Khapra, Pratyush Kumar. *IndicNLG Suite: Multilingual Datasets for Diverse NLG Tasks in Indic Languages*. arxiv preprint 2203.05437. 2022.



Dataset	**Languages**	**Communicative Intent**	**Input Type**	**Total Size**
Biography Generation	as, bn, hi, kn, ml, or, pa, ta, te	One-sentence biographies	key-value pairs	55K
Headline Generation	as, bn, gu, hi, kn, ml, mr, or, pa, ta, te	News article headlines	news article	1.43M
Sentence Summarization	as, bn, gu, hi, kn, ml, mr, or, pa, ta, te	Compacted sentence with same meaning	sentence	431K
Paraphrase Generation	as, bn, gu, hi, kn, ml, mr, or, pa, ta, te	Synonymous sentence	sentence	5.57M
Question Generation	as, bn, gu, hi, kn, ml, mr, or, pa, ta, te	Question leading to answer given context	context-answer pairs	1.08M

MTG Benchmark

4 datasets in 5 languages



Task	**Corpus**	**Domain**	**Format**	**Goal**
Story Generation	ROCStories	Daily life	<story>	Generate the end of the story
Question Generation	SQUAD 1.0	Wikipedia	<passage,answer, question>	Generate the question of the answer
Title Generation	ByteCup	News	<article, title>	Generate the title of the document
Text Summarization	CNN/DailyMail	News	<article, summary>	Generate the summary of the document

Yiran Chen, Zhenqiao Song, Xianze Wu, Danqing Wang, Jingjing Xu, Jiaze Chen, Hao Zhou, and Lei Li. 2022. MTG: A Benchmark Suite for Multilingual Text Generation *NAACL-Findings. 2022*

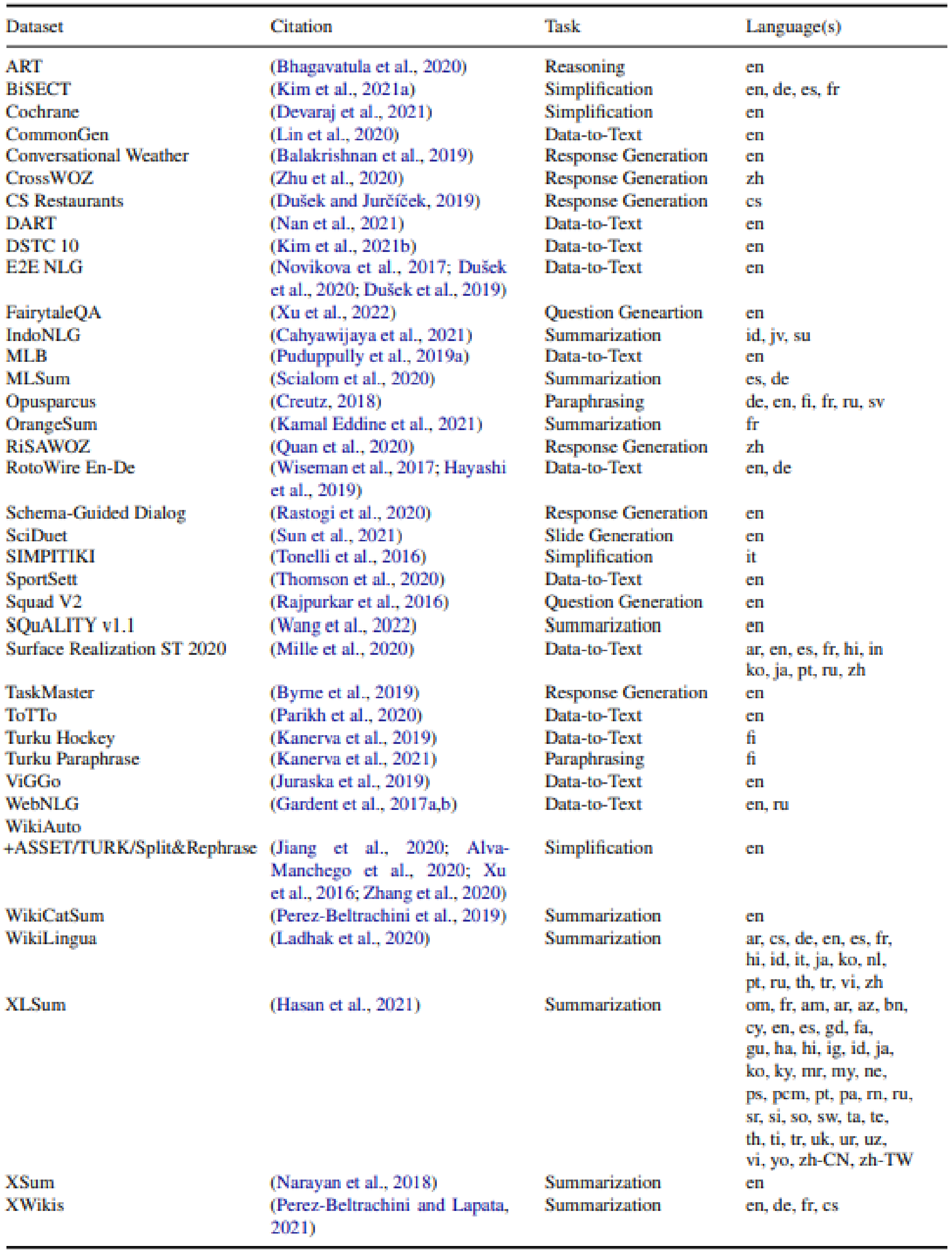
GEMv2 Benchmark

GEMv2 aggregates 40 documented datasets in 51 languages

Sebastian Gehrmann, Abhik Bhattacharjee et al.. *GEMv2: Multilingual NLG Benchmarking in a Single Line of Code. Arxiv 2206.11249. 2022*

Machine Translation Benchmarks

- *FLORES-200*
- *NTREX*
- *BPCC-IN22*



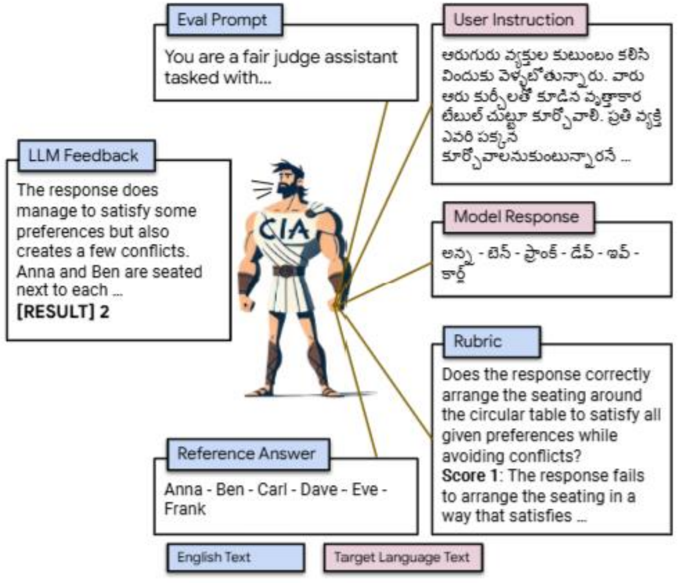
Dataset	Citation	Task	Language(s)
ART	(Bhagavatula et al., 2020)	Reasoning	en
BiSECT	(Kim et al., 2021a)	Simplification	en, de, es, fr
Cochrane	(Devaraj et al., 2021)	Simplification	en
CommonGen	(Lin et al., 2020)	Data-to-Text	en
Conversational Weather	(Balakrishnan et al., 2019)	Response Generation	en
CrossWOZ	(Zhu et al., 2020)	Response Generation	zh
CS Restaurants	(Dušek and Jurčíček, 2019)	Response Generation	cs
DART	(Nan et al., 2021)	Data-to-Text	en
DSTC 10	(Kim et al., 2021b)	Data-to-Text	en
E2E NLG	(Novikova et al., 2017; Dušek et al., 2020; Dušek et al., 2019)	Data-to-Text	en
FairytaleQA	(Xu et al., 2022)	Question Geneartion	en
IndoNLG	(Cahyawijaya et al., 2021)	Summarization	id, jv, su
MLB	(Puduppully et al., 2019a)	Data-to-Text	en
MLSum	(Scialom et al., 2020)	Summarization	es, de
Opusparcus	(Creutz, 2018)	Paraphrasing	de, en, fi, fr, ru, sv
OrangeSum	(Kamal Eddine et al., 2021)	Summarization	fr
RiSAWOZ	(Quan et al., 2020)	Response Generation	zh
RotoWire En-De	(Wiseman et al., 2017; Hayashi et al., 2019)	Data-to-Text	en, de
Schema-Guided Dialog	(Rastogi et al., 2020)	Response Generation	en
SciDuet	(Sun et al., 2021)	Slide Generation	en
SIMPITIKI	(Tonelli et al., 2016)	Simplification	it
SportSett	(Thomson et al., 2020)	Data-to-Text	en
Squad V2	(Rajpurkar et al., 2016)	Question Generation	en
SQuALITY v1.1	(Wang et al., 2022)	Summarization	en
Surface Realization ST 2020	(Mille et al., 2020)	Data-to-Text	ar, en, es, fr, hi, in ko, ja, pt, ru, zh
TaskMaster	(Byrne et al., 2019)	Response Generation	en
ToTTo	(Parikh et al., 2020)	Data-to-Text	en
Turku Hockey	(Kanerva et al., 2019)	Data-to-Text	fi
Turku Paraphrase	(Kanerva et al., 2021)	Paraphrasing	fi
ViGGo	(Juraska et al., 2019)	Data-to-Text	en
WebNLG	(Gardent et al., 2017a,b)	Data-to-Text	en, ru
WikiAuto +ASSET/TURK/Split&Rephrase	(Jiang et al., 2020; Alva-Manchego et al., 2020; Xu et al., 2016; Zhang et al., 2020)	Simplification	en
WikiCatSum	(Perez-Beltrachini et al., 2019)	Summarization	en
WikiLingua	(Ladhak et al., 2020)	Summarization	ar, cs, de, en, es, fr, hi, id, it, ja, ko, nl, pt, ru, th, tr, vi, zh
XLSum	(Hasan et al., 2021)	Summarization	om, fr, am, ar, az, bn, cy, en, es, gd, fa, gu, ha, hi, ig, id, ja, ko, ky, mr, my, ne, ps, pcm, pt, pa, rn, ru, sr, si, so, sw, ta, te, th, ti, tr, uk, ur, uz, vi, yo, zh-CN, zh-TW
XSum	(Narayan et al., 2018)	Summarization	en
XWikis	(Perez-Beltrachini and Lapata, 2021)	Summarization	en, de, fr, cs

Multilingual LLM Benchmarks for Open-ended tasks

- Very few such benchmarks
- Some examples: covers only a few languages
  - Aya Evaluation Suite
  - MGSM for Math Problems
  - Multilingual MMLU from OpenAI
- Machine Translated versions of various English datasets are available
  - Vicuna, ALPACA, MMLU, ARC, Dolly

Evaluation Metrics

- Evaluation of text is a challenging task
- How good is an evaluation metric?
- Particularly, methods that are applicable to multiple languages
  - Morphology and grammar of languages vary
  - Resource available for building and meta-evaluating evaluation models vary
- Popular evaluation metrics
  - String-based: BLEU, ROUGE, chrF
  - Embedding-based: BERTScore
  - Learned Metrics: COMET, BLEURT
  - LLM as Judge: GEMBA, G-Eval, Prometheus, **CIA**



Eval Prompt

You are a fair judge assistant tasked with...

User Instruction

ఆరుగురు వ్యక్తుల కుటుంబం కలిసి విందుకు వెళ్ళబోతున్నారు. వారు ఆరు కుర్చీలతో కూడిన వృత్తాకార టేబుల్ చుట్టూ కూర్చోవాలి. ప్రతి వ్యక్తి ఎవరి పక్కన కూర్చోవాలనుకుంటున్నారనే ...

LLM Feedback

The response does manage to satisfy some preferences but also creates a few conflicts. Anna and Ben are seated next to each ...
[RESULT] 2

Model Response

అన్న - బెన్ - ఫ్రాంక్ - డేవ్ - ఇవ్ - కార్ల్

Rubric

Does the response correctly arrange the seating around the circular table to satisfy all given preferences while avoiding conflicts?
Score 1: The response fails to arrange the seating in a way that satisfies ...

Reference Answer

Anna - Ben - Carl - Dave - Eve - Frank

https://github.com/AI4Bharat/CIA

Summary

- Deep Learning has revolutionized multilingual representation learning
  - Word representations → Pre-trained LM with monolingual objectives → bilingual objectives
- Opened up possibilities
  - Multilingual training, Zero-shot performance, Compact models
  - Support low-resource languages and domains via transfer Learning
- Both multilingual NLU and NLG made possible
  - Multilingual NLG is more challenging
- Effectiveness of multilingual transfer depends on task complexity, language relatedness
- Efficient Multilingual Evaluation is a challenge

Resources

AI4Bharat: Resources and tools for Multilingual NLP for Indian languages

Survey Papers & Tutorials:

Sumanth Doddapaneni, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M. Khapra. A Primer on Pretrained Multilingual Language Models. 2021. https://arxiv.org/abs/2107.00676

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Yuqi Ye, Hanwen Gu. A Survey on Multilingual Large Language Models: Corpora, Alignment, and Bias. 2023. https://arxiv.org/abs/2404.00929

Raj Dabre, Chenhui Chu, Anoop Kunchukuttan. Multilingual Machine Translation. [Survey paper at ACM Computing Surveys 2020] [Tutorial at COLING 2020] [website]

Anoop Kunchukuttan. Extending English Large Language Models to New Languages: A Survey. 2024.
https://anoopkunchukuttan.gitlab.io/publications/presentations/extend_en_llms_aug2024.pdf
(upcoming tutorial at EMNLP 2025)

Anoop Kunchukuttan. Multilingual Learning. Tutorial at Summer School on Machine Learning. IIIT Hyderabad. 2018.
https://anoopkunchukuttan.gitlab.io/publications/presentations/IIIT-Hyderabad-ML-Summer-School-2018.pdf

Thank You!

anoop.kunchukuttan@gmail.com

https://anoopkunchukuttan.gitlab.io/