

THE PROMISE OF TRANSFER LEARNING

... for Improving NLP
for Low Resource
Languages

Sibonelo Dlamini (UKZN)
Digital Humanities Colloquium
17 August 2022

CONTEXT

Great HL technologies have been developed:

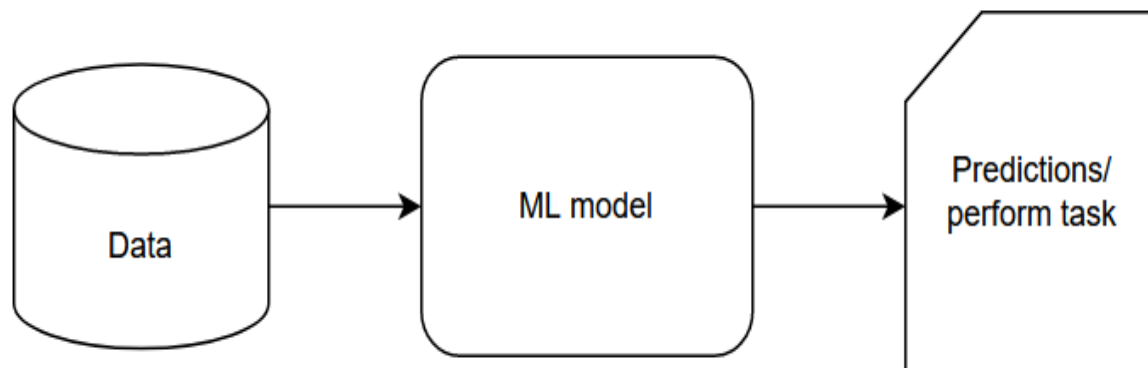
- Machine Translation (Google Translate)
 - Speech-to-text (Siri)
 - Autocompletion (Gmail)
- But mostly available in English (other well-resourced languages)
- **Reason:** ML models need data to perform well

WHY WE NEED HLTS IN INDIGENOUS LANGUAGES

- We are increasingly interfacing with computers via speech (inclusivity monolingual non-English speakers)
- Can assist with implementation of language policies
- Continue to attract new speakers
- Access and document oral history

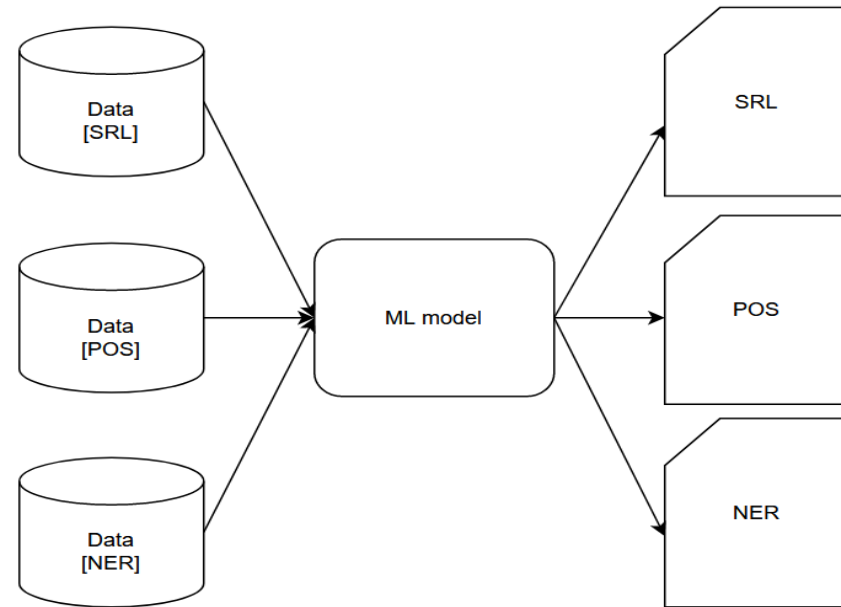
NORMAL VS. TRANSFER LEARNING

One model trained on data to perform a single task



[1] MULTI-TASK TRANSFER LEARNING

One model trained on multiple tasks (Semantic Role Labelling, POS tagging, Named Entity Recognition)



MULTI-TASK TRANSFER LESSONS

- The tasks you choose matter
- Interleaving vs. sequential
- Impact on unequal datasets?
- The model you choose matters

Remember! [2008 paper](#).

A Unified Architecture for Natural Language Processing:
Deep Neural Networks with Multitask Learning

Ronan Collobert
Jason Weston
NEC Labs America, 4 Independence Way, Princeton, NJ 08540 USA

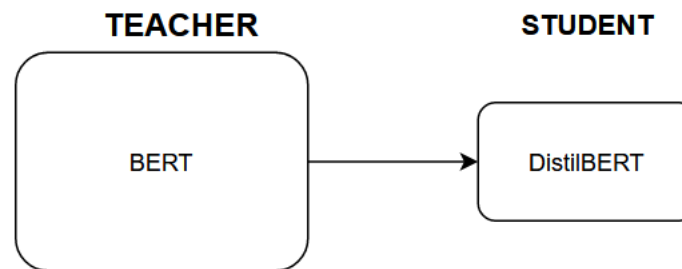
COLLOBER@NEC-LABS.COM
JASONW@NEC-LABS.COM

[2] KNOWLEDGE DISTILLATION

[KNOWLEDGE DISTILLATION: A SURVEY – GOU ET AL.]

In order to reduce latency, train smaller student model from much larger parent model

[DistilBERT] “has **40% less parameters** than *bert-base-uncased*, runs **60% faster** while **preserving over 95%** of BERT’s performances as measured on the GLUE language understanding benchmark.” – huggingface.co



[3] CROSS-LINGUAL TRANSFER

Usually, ML models are trained on a single language

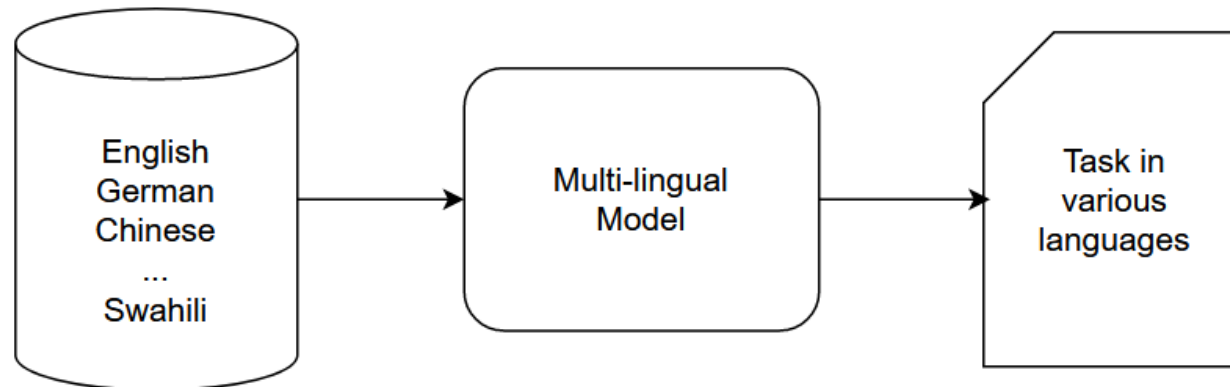
Disadvantages:

- We need to train one model per language
- Maintain each languages model
- No leveraging of similarity between languages
- Wide disparity in performance between languages, given available data

MULTI-LINGUAL MODELS

Solution: multilingual models

- Single model trained on 100+ languages
- Can perform task in multiple languages



MONOLINGUAL TRANSFORMER MODELS

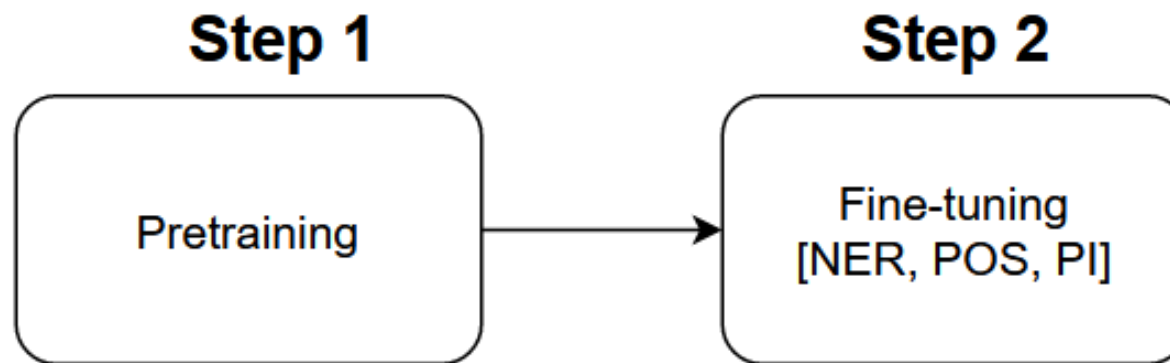
[[THE ILLUSTRATED TRANSFORMER: JAY ALAMMAR](#)]

Progress: statistical models => neural networks => deep learning => transformers

Transformers:

- pretrained on large (multilingual) corpus [MLM, NSP] [**language acquisition**]
- Fine-tuned for specific task [**skill acquisition**]

BERT, etc...



NATURAL LANGUAGE UNDERSTANDING TASKS

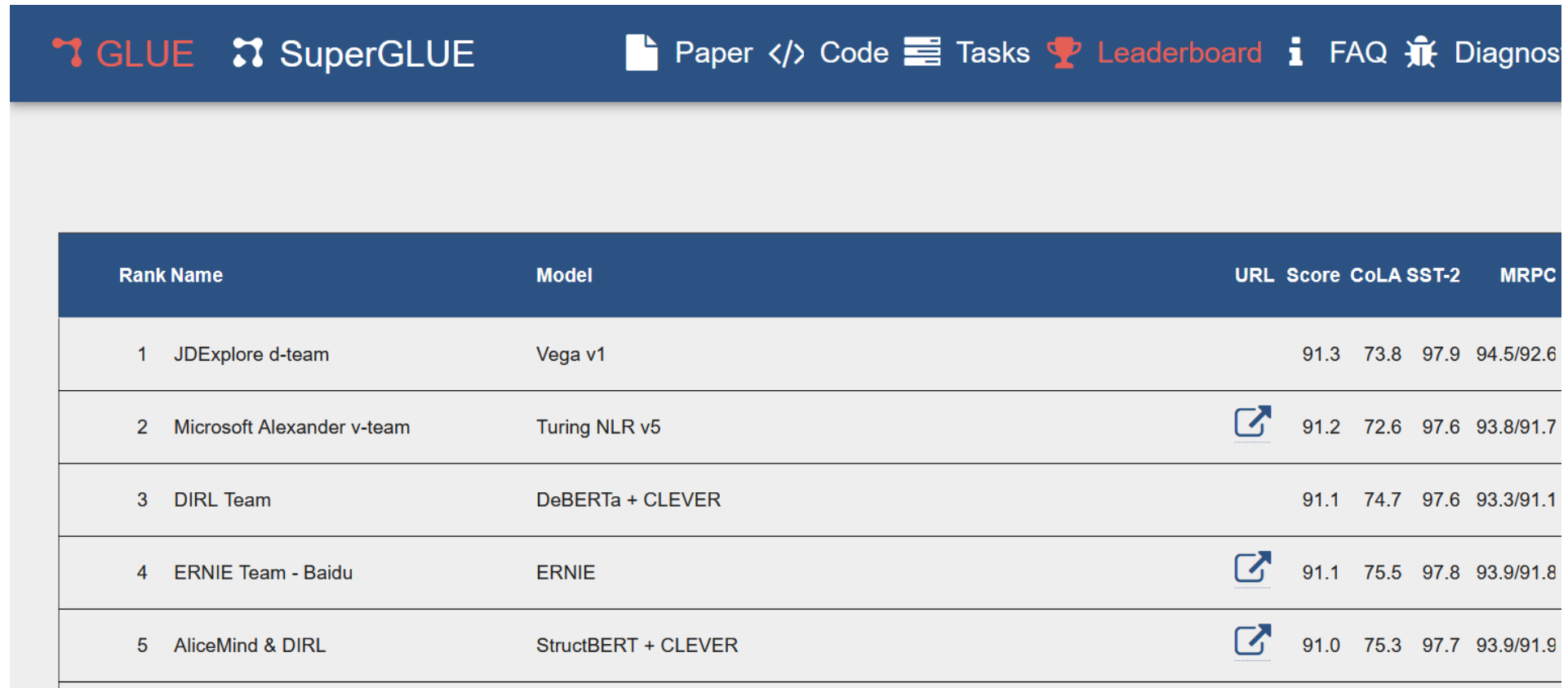
- Part-of-Speech Tagging
- Named Entity Recognition [Location, Person, Organisation]
- Paraphrase Identification
- Natural Language Inference [(premise, hypothesis) \Rightarrow entailment, contradiction, neither]
- Document classification [news articles]
- Question Answering [Single word, span in running text]
- **“Diverse set of tasks helps us learn the characteristics of model”**

BASIC TASKS ARE IMPORTANT

User Application	Transcription of speech in multiple languages
High-level NLP Tasks	Speech-to-text AND automatic machine translation
Basic NLP Tasks	POS Tagging, Morphological Analysis, NER

DOMINANCE OF TRANSFORMER MODELS

[[GLUE BENCHMARK](#)]

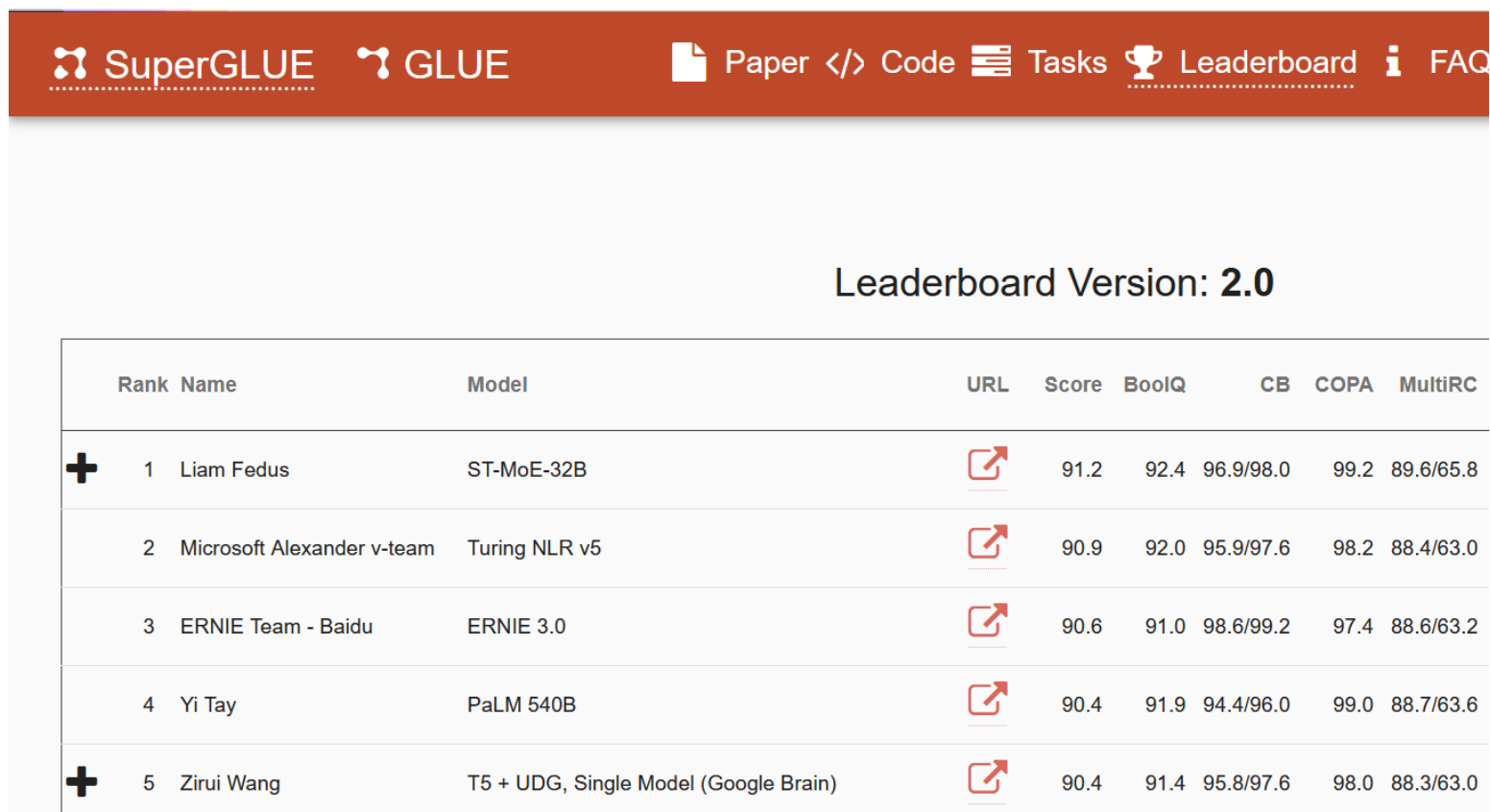







The screenshot shows the GLUE SuperGLUE Leaderboard interface. At the top, there are navigation links for GLUE, SuperGLUE, Paper, Code, Tasks, Leaderboard, FAQ, and Diagnostics. Below the navigation is a table listing the top 5 teams and their models, along with their scores on various tasks.

Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC
1	JDExplore d-team	Vega v1		91.3	73.8	97.9	94.5/92.6
2	Microsoft Alexander v-team	Turing NLR v5	🔗	91.2	72.6	97.6	93.8/91.7
3	DIRL Team	DeBERTa + CLEVER		91.1	74.7	97.6	93.3/91.1
4	ERNIE Team - Baidu	ERNIE	🔗	91.1	75.5	97.8	93.9/91.8
5	AliceMind & DIRL	StructBERT + CLEVER	🔗	91.0	75.3	97.7	93.9/91.9

ALSO TRUE FOR SUPERGLUE

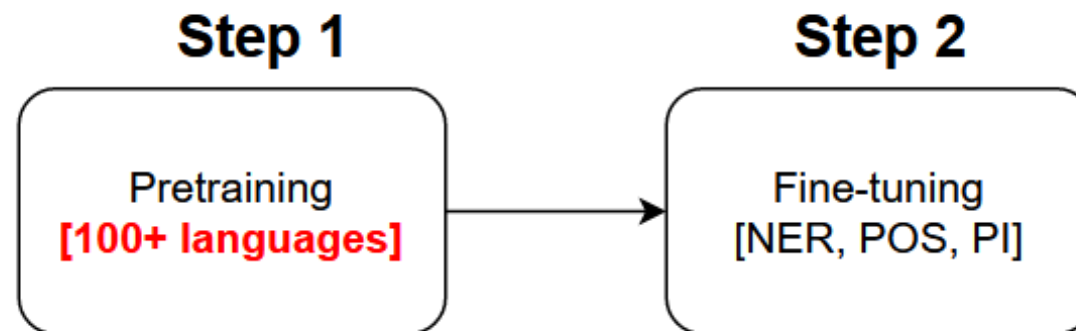
[[SUPERGLUE BENCHMARK](#)]



Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC
+ 1	Liam Fedus	ST-MoE-32B		91.2	92.4	96.9/98.0	99.2	89.6/65.8
2	Microsoft Alexander v-team	Turing NLR v5		90.9	92.0	95.9/97.6	98.2	88.4/63.0
3	ERNIE Team - Baidu	ERNIE 3.0		90.6	91.0	98.6/99.2	97.4	88.6/63.2
4	Yi Tay	PaLM 540B		90.4	91.9	94.4/96.0	99.0	88.7/63.6
+ 5	Zirui Wang	T5 + UDG, Single Model (Google Brain)		90.4	91.4	95.8/97.6	98.0	88.3/63.0

MULTILINGUAL TRANSFORMER MODELS

- mBERT
- mT5 (Google)
- M2M100 (Facebook)
- XLM-R



XTREME BENCHMARK

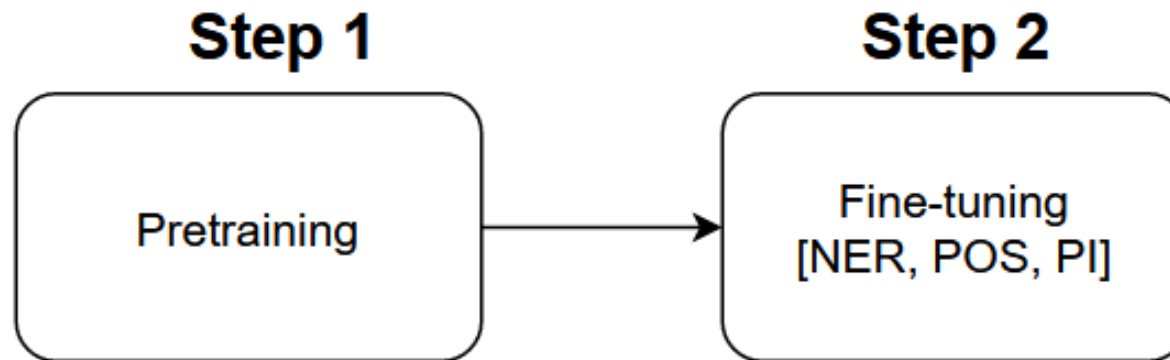
[XTREME + XTREME-R]

Leaderboard results

Rank	Model	Participant	Affiliation	Attempt Date	Avg	Sentence-pair Classification	Structured Prediction	Question Answering	Sentence Retrieval
0		Human	-	-	93.3	95.1	97.0	87.8	-
1	MShenNonG+TDT	Cloud Xiaowei AI	Tencent	May 29, 2022	85.0	90.4	83.1	76.3	94.4
2	Turing ULR v5	Alexander v-team	Microsoft	Nov 24, 2021	84.5	90.3	81.7	76.3	93.7
3	CoFe	HFL	iFLYTEK	Oct 26, 2021	84.1	90.1	81.4	75.0	94.2
4	InfoXLM-XFT	Noah's Ark Lab	Huawei	Oct 5, 2021	82.2	89.3	75.5	75.2	92.4
5	VECO + HICTL	AliceMind + MT	Alibaba	Sep 21, 2021	82.0	89.0	76.7	73.4	93.3

OUR 2 PROBLEMS WITH USING TRANSFORMERS

1. No corpus – running text – to pretrain the transformer model
2. No data to fine-tune the model for specific task



PROBLEM 1: WIKIPEDIA

[[HTTPS://META.WIKIMEDIA.ORG/WIKI/LIST OF WIKIPEDIAS](https://meta.wikimedia.org/wiki/List_of_Wikipedias), ACCESSED 16 AUGUST 2022]

Table 1: The rank of various languages in terms of size of their Wikipedias measured by the number of articles they have [1]. Accessed 16th of August 2022.

Rank	Language	# Articles
1	English	6,486,854
2	Cebuano	6,125,900
3	German	2,683,425
50	Lithuanian	202,703
69	Afrikaans	104,366
100	Tagalog	43,042
.		
.		
.		
156	isiZulu	10,538
271	isiXhosa	1,232
290	Sesotho	823

PROBLEM 1: COMMCRAWL

[[STATISTICS OF COMMON CRAWL MONTHLY ARCHIVES](#), ACCESSED 16 AUGUST 2022]

Table 2: Language size in CommonCrawl corpus as percentage. Accessed 16th of August 2022.

Language	Percentage of Corpus
English	46.5384
Russian	5.8779
German	5.4824
Chinese	4.6777
Japanese	4.8135
Afrikaans	0.0119
isiZulu	0.0018
isiXhosa	0.0018
Sesotho	0.0007

PROBLEM 2: XTREME BENCHMARK LANGUAGES

Family	Languages
Afro-Asiatic	Arabic, Hebrew
Austro-Asiatic	Vietnamese
Austronesian	Indonesian, Javanese, Malay, Tagalog
Basque	Basque
Dravidian	Malayalam, Tamil, Telugu
Indo-European (Indo-Aryan)	Bengali, Marathi, Hindi, Urdu
Indo-European (Germanic)	Afrikaans, Dutch, English, German
Indo-European (Romance)	French, Italian, Portuguese, Spanish
Indo-European (Greek)	Greek
Indo-European (Iranian)	Persian
Japonic	Japanese

PROBLEM 2: XTREME-R

The tasks included in XTREME cover a range of paradigms, including sentence classification, structured prediction, sentence retrieval, cross-lingual question answering, and common sense reasoning. Consequently, in order for models to be successful on the XTREME benchmarks, they must learn representations that generalize to many standard cross-lingual transfer settings.

Each of the tasks covers a subset of the 50 languages. XTREME-R adds the following ten languages to XTREME: Haitian Creole, Cusco Quechuan, Wolof, Lithuanian, Punjabi, Gujarati, Polish, Ukrainian, Azerbaijani, and Romanian.

PROBLEM 2: XGLUE BENCHMARK LANGUAGES

[XGLUE BENCHMARK]

Task	ar	bg	de	el	en	es	fr	hi	it	nl	pl	pt	ru	sw	th	tr	ur	vi	zh
NER			✓		✓	✓				✓									
POS	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓
NC*			✓		✓	✓	✓						✓						
MLQA	✓		✓		✓	✓		✓										✓	✓
XNLI	✓	✓	✓	✓	✓	✓	✓	✓					✓	✓	✓	✓	✓	✓	✓
PAWS-X			✓		✓	✓	✓												
QADSM*			✓		✓		✓												
WPR*			✓		✓	✓	✓		✓			✓							✓
QAM*			✓		✓		✓												
QG*			✓		✓	✓	✓		✓			✓							
NTG*			✓		✓	✓	✓						✓						

Table 3: The 19 languages covered by the 11 downstream tasks: *Arabic* (ar), *Bulgarian* (bg), *German* (de), *Greek* (el), *English* (en), *Spanish* (es), *French* (fr), *Hindi* (hi), *Italian* (it), *Dutch* (nl), *Polish* (pl), *Portuguese* (pt), *Russian* (ru), *Swahili* (sw), *Thai* (th), *Turkish* (tr), *Urdu* (ur), *Vietnamese* (vi), and *Chinese* (zh). All these 6 new tasks with * are labeled by human, except es, it and pt datasets in QG (80+% accuracy) are obtained by an in-house QA ranker.

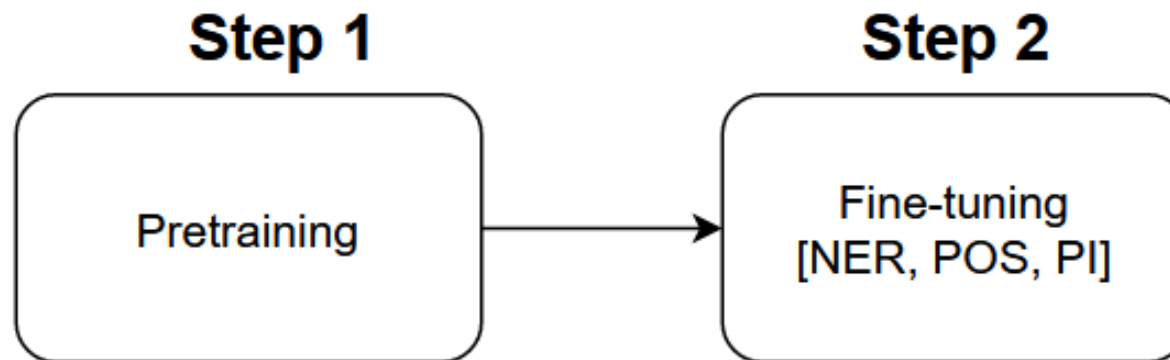
THE PROMISE!

Zero-shot transfer produces good results

- Fine-tune multilingual transformer model on source language(s) samples, test on target language

Few-shot [10 - 100] transfer enhances results

- Fine-tune multilingual transformer model on source language(s) **and a few** target language samples, test on target language



GOOD RESULTS BEING REPORTED

Table 3: Performance results reported in the literature for cross-lingual zero-shot transfer learning.

Paper	Model	Dataset	Result
Natural Language Inference (NLI)			
Lauscher et. al. 2020 [2]	mBERT	XNLI [3]	68.16
Lauscher et. al. 2020 [2]	XLM-R	XNLI	75.84
Vidoni et. al. 2020 [4]	Ortho-adapters [4]	XNLI	71.43
Hong et. al. 2020 [5]	mBERT	XNLI	59.68
Keung et. al. 2020 [6]	mBERT	XNLI	71.22
Soltan et. al. 2021 [7]	mBERT	XNLI	78.91
Huang et. al. 2021 [8]	mBERT	XNLI	67.6
Turc et. al. 2021 [9]	mBERT	XNLI	96.3
Turc et. al. 2021 [9]	mT5-Base	XNLI	95.6
Named Entity Recognition (NER)			
Lauscher et. al. 2020 [2]	mBERT	WikiANN [10]	89.31
Lauscher et. al. 2020 [2]	XLM-R	WikiANN	93.82
Liu et. al. 2020 [11]	mBERT	CoNLL 2002 [12]	75.46
Vidoni et. al. 2020 [4]	Ortho-adapters [4]	CoNLL 2003 [13]	58.99
Keung et. al. 2020 [6]	mBERT	WikiANN	74.6
Soltan et. al. 2021 [7]	mBERT	CoNLL 2002	80.76
		CoNLL 2003	
		WikiANN	
Part-of-Speech Tagging (POS)			
Lauscher et. al. 2020 [2]	mBERT [2]	Universal Dependencies 2.1 [14]	92.64
Lauscher et. al. 2020 [2]	XLM-R [2]	Universal Dependencies 2.1	92.80
Liu et. al. 2020 [11]	mBERT [11]	Universal Dependencies 2.0	86.95
Vidoni et. al. 2020 [4]	Ortho-adapters [4]	Universal Dependencies 2.2	66.77
Question Answering (QA)			
Lauscher et. al. 2020 [2]	mBERT	XQuAD [15]	50.19
Lauscher et. al. 2020 [2]	XLM-R	XQuAD	55.78

NO SUBSTITUTE FOR DATA

Overcome Problem 1: Build larger corpora

Overcome Problem 2: Translate datasets into SA languages [can't test cross-lingual transfer without this!]

NO ESCAPE FROM FUNDING DATASET CREATION

Table 2: Statistics on the PAWS-Wiki corpus to be translated: number of samples and words; along with the cost of in Rands for translating the training, development and test sets.

	Number of Samples	Number of Words	Cost (R)
Train	49,401	1,825,511	1,460,408,80
Development	8,000	296,266	237,012,80
Test	8,000	297,006	213,604,80
Total	65,401	2,418,783	2,297,843,85



THANK YOU!

Question?

And comments...