Challenges and Opportunities in Multilingual Evaluation

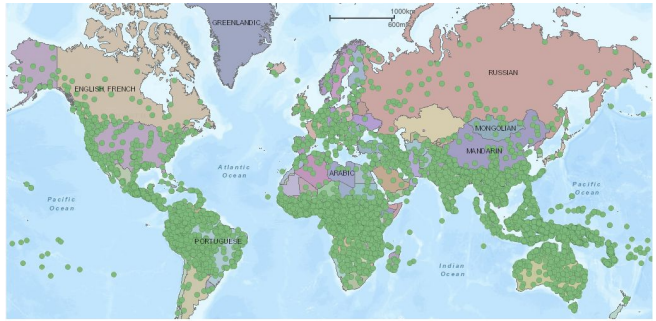
Sebastian Ruder

Eval4NLP Workshop at EMNLP 2021

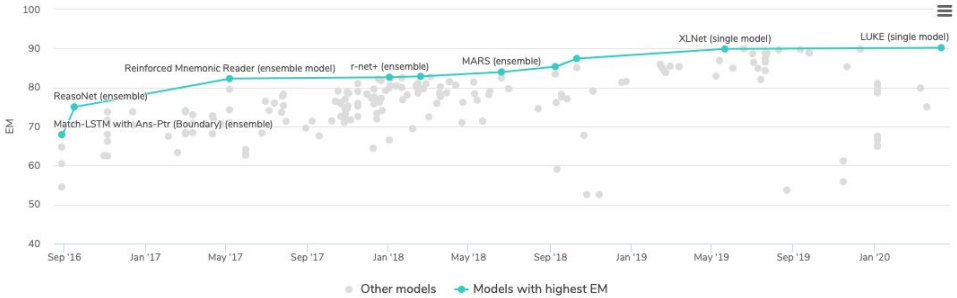
Google Research

Language Diversity in NLP

- There are more than 7,000 languages spoken around the world
- In NLP research, models are typically developed to work well only for English
- Two big implications:
  - Lack of technological inclusion of more than 3 billion speakers
  - Overfitting to English



http://langscape.umd.edu/map.php



Performance on SQuAD 1.1 Question Answering dataset (Papers with Code)

Agenda

01 Resources*

02 Evaluation Setting

*: Main focus is on datasets for tasks other than MT

01

Resources

a Biases and limitations

b Scaling to Many Languages

Google Research

Translation-based Bias

- Many multilingual datasets (XNLI, XQuAD, PAWS-X, etc) are **based on translations**
- **"Translationese"**: translations differ in many aspects from natural language [Volansky et al., 2015]
- **Inherits artefacts** from existing datasets
  - Train-test overlap for answers in NQ [Lewis et al., 2020]
  - Language-specific replications may **improve upon annotation methodology** [Watarai & Tsuchiya, 2020]
- Leads to **new artefacts**
  - Bias towards models trained on translations in XNLI [Artetxe et al., 2020]
- Translated text is **different from text "naturally" generated text** by speakers of different languages
  → English and Western-centric bias

Lang	Context paragraph w/ answer spans	Questions
en	The heat required for boiling the water and supplying the steam can be derived from various sources, most commonly from **[burning combustible materials]**$_1$ with an appropriate supply of air in a closed space (called variously **[combustion chamber]**$_2$, firebox). In some cases the heat source is a nuclear reactor, geothermal energy, **[solar]**$_3$ energy or waste heat from an internal combustion engine or industrial process. In the case of model or toy steam engines, the heat source can be an **[electric]**$_4$ heating element.	1. What is the usual source of heat for boiling water in the steam engine? 2. Aside from firebox, what is another name for the space in which combustible material is burned in the engine? 3. Along with nuclear, geothermal and internal combustion engine waste heat, what sort of energy might supply the heat for a steam engine? 4. What type of heating element is often used in toy steam engines?
es	El calor necesario para hervir el agua y suministrar el vapor puede derivarse de varias fuentes, generalmente de **[la quema de materiales combustibles]**$_1$ con un suministro adecuado de aire en un espacio cerrado (llamado de varias maneras: **[cámara de combustión]**$_2$, chimenea...). En algunos casos la fuente de calor es un reactor nuclear, energía geotérmica, **[energía solar]**$_3$ o calor residual de un motor de combustión interna o proceso industrial. En el caso de modelos o motores de vapor de juguete, la fuente de calor puede ser un calentador **[eléctrico]**$_4$.	1. ¿Cuál es la fuente de calor habitual para hacer hervir el agua en la máquina de vapor? 2. Aparte de cámara de combustión, ¿qué otro nombre que se le da al espacio en el que se quema el material combustible en el motor? 3. Junto con el calor residual de la energía nuclear, geotérmica y de los motores de combustión interna, ¿qué tipo de energía podría suministrar el calor para una máquina de vapor? 4. ¿Qué tipo de elemento calefactor se utiliza a menudo en las máquinas de vapor de juguete?
zh	让水沸腾以提供蒸汽所需热量有多种来源，最常见的是在封闭空间（别称有 **[燃烧室]**$_2$ 、火箱）中供应适量空气来 **[燃烧可燃材料]**$_1$ 。在某些情况下，热源是核反应堆、地热能、 **[太阳能]**$_3$ 或来自内燃机或工业过程的废气。如果是模型或玩具蒸汽发动机，还可以将 **[电]**$_4$ 加热元件作为热源。	1. 蒸汽机中让水沸腾的常用热源是什么？ 2. 除了火箱之外，发动机内燃烧可燃材料的空间的别名是什么？ 3. 除了核能、地热能和内燃机废气以外，还有什么热源可以为蒸汽机供能？ 4. 玩具蒸汽机通常使用什么类型的加热元件？

English examples in SQuAD and translations in XQuAD [Artetxe et al., 2020]

English and Western-centric Bias

- Crowd-sourced content is **biased towards an English and Western-centric viewpoint**
- Cultures differ in **what type of content is relevant** to them
  - Speakers outside the US probably don't care about famous American football and baseball players
  - In COPA [Roemmele et al., 2011], many referents have no language-specific terms in some languages, e.g. bowling ball, hamburger, lottery [Ponti et al., 2020]
  - Concepts in ImageNet are Western-centric [Liu et al., 2021]
  - Commonsense knowledge, social norms, taboo topics, social distance, etc are **culture-dependent** [Thomas, 1983]

	Value	NQ		QB		SQuAD		TriviaQA	
		Train	Dev	Train	Dev	Train	Dev	Train	Dev
Country	US	59.62	58.66	29.70	26.28	32.74	24.93	31.32	30.91
	UK	15.76	15.78	17.92	17.68	19.66	16.83	41.92	41.32
	France	1.79	1.18	10.06	10.34	7.76	10.57	4.37	4.84
	Italy	1.83	1.88	8.07	10.50	9.00	3.88	3.75	3.48
	Germany	1.52	2.12	7.21	6.71	4.77	6.61	3.01	3.00
	No country	4.82	4.36	7.12	6.79	3.48	2.56	6.19	6.10

Coverage (% of examples) of countries across examples with people in QA datasets [Gor et al., 2021]

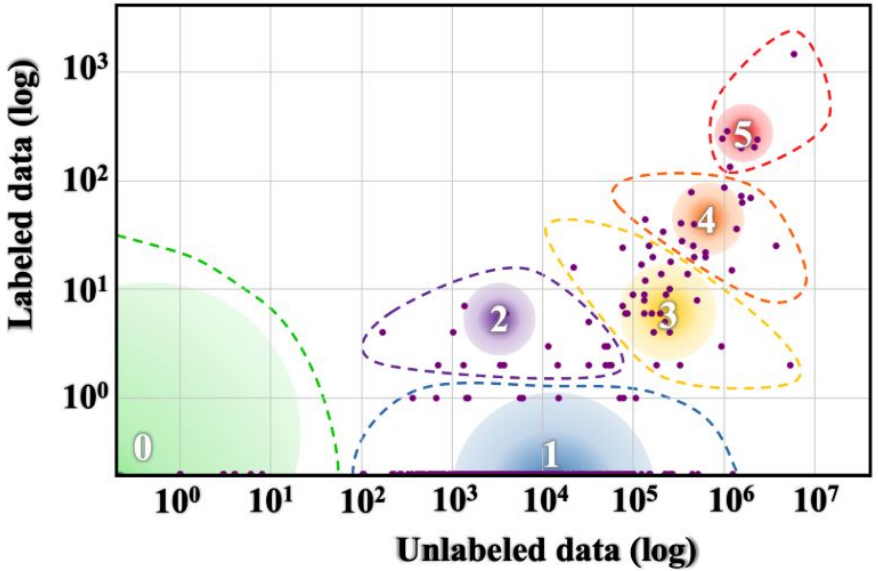
(a) இரு படங்களில் ஒன்றில் இரண்டிற்கும் மேற்பட்ட மஞ்சள் சட்டை அணிந்த வீரர்கள் காளையை அடக்கும் பணியில் ஈடுப்பட்டிருப்பதை காணமுடிகிறது. ("In one of the two photos, more than two yellow-shirted players are seen engaged in bull taming."). Label: TRUE.

A Tamil example in MaRVL [Liu et al., 2021]

Opportunity: Create datasets that capture knowledge and information natively in different languages

Scaling to Many Languages

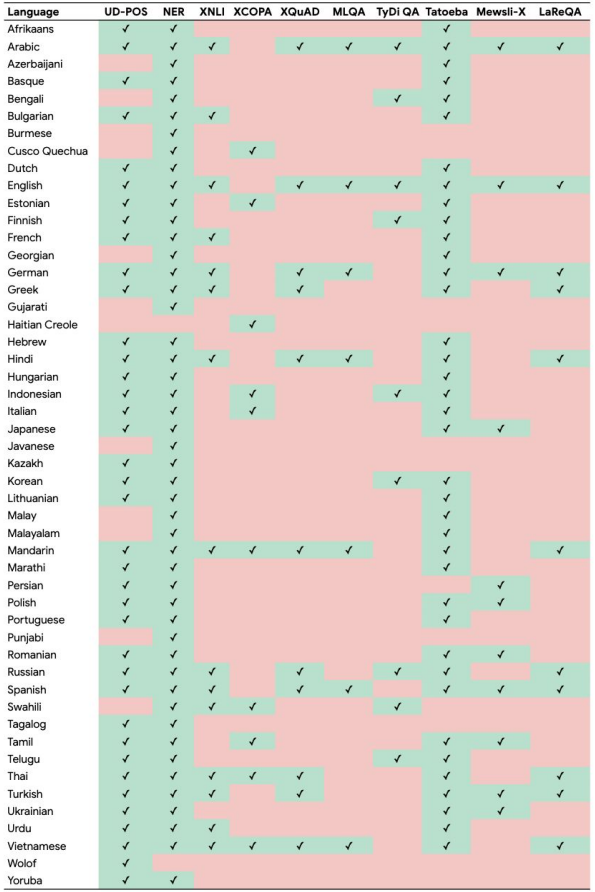
- Labelled data for evaluation is **only available in a small number of languages**
- Annotation or generation of data in many languages is **expensive**



Distribution of resources across languages [Joshi et al., 2020]

Scaling to Many Languages

- Labelled data for evaluation is **only available in a small number of languages**
- Annotation or generation of data in many languages is **expensive**
- How can we fill in the gaps and **efficiently assess performance** in many languages?
- Translation is comparatively cheap but **introduces biases**
- Alternatives:
  - Generate **targeted evaluation datasets**
  - Create **few-shot datasets**
  - Cross-lingual **performance prediction**



Language	UD-POS	NER	XNLI	XCOPA	XQuAD	MLQA	TyDi QA	Tatoeba	Mewsli-X	LaReQA
Afrikaans	✓	✓						✓		
Arabic	✓	✓	✓		✓	✓	✓	✓	✓	✓
Azerbaijani		✓						✓		
Basque	✓	✓						✓		
Bengali		✓					✓	✓		
Bulgarian	✓	✓	✓					✓		
Burmese		✓								
Cusco Quechua		✓		✓						
Dutch	✓	✓						✓		
English	✓	✓	✓		✓	✓	✓	✓	✓	✓
Estonian	✓	✓		✓				✓		
Finnish	✓	✓					✓	✓		
French	✓	✓	✓					✓		
Georgian		✓						✓		
German	✓	✓	✓		✓	✓		✓	✓	✓
Greek	✓	✓	✓		✓			✓		✓
Gujarati		✓								
Haitian Creole				✓						
Hebrew	✓	✓						✓		
Hindi	✓	✓	✓		✓	✓		✓		✓
Hungarian	✓	✓						✓		
Indonesian	✓	✓		✓			✓	✓		
Italian	✓	✓		✓				✓		
Japanese	✓	✓						✓	✓	
Javanese		✓								
Kazakh	✓	✓								
Korean	✓	✓					✓	✓		
Lithuanian	✓	✓						✓		
Malay		✓						✓		
Malayalam		✓						✓		
Mandarin	✓	✓	✓	✓	✓	✓		✓		✓
Marathi	✓	✓						✓		
Persian	✓	✓							✓	
Polish	✓	✓						✓	✓	
Portuguese	✓	✓						✓		
Punjabi		✓								
Romanian	✓	✓						✓	✓	
Russian	✓	✓	✓		✓		✓	✓		✓
Spanish	✓	✓	✓		✓	✓		✓	✓	✓
Swahili		✓	✓	✓			✓			
Tagalog	✓	✓								
Tamil	✓	✓		✓				✓	✓	
Telugu	✓	✓					✓	✓		
Thai	✓	✓	✓	✓	✓			✓		✓
Turkish	✓	✓	✓		✓			✓	✓	✓
Ukrainian	✓	✓						✓	✓	
Urdu	✓	✓	✓					✓		
Vietnamese	✓	✓	✓	✓	✓	✓		✓		✓
Wolof	✓									
Yoruba	✓	✓								

Language coverage of tasks in XTREME-R [Ruder et al., 2021]

Creating Targeted Evaluation Data

- Create **template-based test cases**, e.g. using CheckList [Ribeiro et al., 2020]
- A **small number of templates** can cover many different model capabilities
- Scaling across languages still **requires native speaker expertise or translation**
- So far have been used for **evaluating reading comprehension** [Ruder et al., 2021] and **closed-book QA** [Jiang et al., 2020; Kassner et al., 2021]

Test	Template	Generated test
Comparisons	`{first_name}` is `{adj[0]}` than `{first_name1}`. Who is less `{adj[1]}`?	C: Ben is smaller than Frank. Q: Who is less small?
Intensifiers	`{first_name}` `{state}` `{very}` בקשר לפרויקט. `{first_name1}` `{state}` בקשר לפרויקט. מי הכי פחות `{state}` בקשר לפרויקט?	C: עמנואל שמח בקשר לפרויקט. יצחק שמח ביותר בקשר לפרויקט. Q: מי הכי פחות שמח בקשר לפרויקט?
Properties	`{obj[0]}` يوجد `{obj[1]}` في الغرفة. `{attribute1}` هو و`{attribute2}`. أي `{property2}` هو `{obj[1]}`؟	C: يوجد ورق حائط في الغرفة. ورق الحائط هو ضئيل ومربع. Q: أي شكل هو ورق الحائط؟
Job vs Nationality	`{first_name}` একজন `{profession}` এবং `{nationality}`। `{first_name}` এর জাতীয়তা কী?	C: হালিম একজন ওয়েট্রেস এবং চীনা। Q: হালিম এর জাতীয়তা কী?

Lang.	Comparisons	Intensifiers	Properties	Job vs Nationality	Animal vs Vehicles	Animal vs Vehicles 2
af	42.4	66.3	36	0	54.5	10.7
ar	24.4	97.5	100	0	100	26.6
az	96	67.3	72	5	35.5	98
bg	41.3	80.9	18.1	0	20	8
bn	93	92.8	91	7	49	30.2
de	13.6	91.1	34.3	1.5	16.5	6.1
el	7.6	99.5	42.8	16.4	47.5	23.2
en	7	92.4	39	0	15	1
es	9.6	98.5	62	0	33.5	4.5
et	13.2	87.9	37.6	11	42	0
eu	100	99	59.5	16	19	32.5
fa	12.1	83.5	52.2	6.5	45.5	16
fi	11.2	83.4	29.8	4.5	10.3	1.5
fr	3.6	94.5	18.9	0.8	12.5	18
gu	100	100	100	34.5	87	27.8
ha	100	100	100	100	100	85.8
he	100	100	100	100	100	28.4
hi	28.3	57.7	66.7	6	30	8.6
ht	94.4	100	100	100	100	91.4
hu	7.6	98	38.5	10.5	29	13.1
id	6	97	78	0	39	33.7
it	2.5	99	56.7	6.7	16	1.5
ja	95.5	100	65.4	99	35.5	93
jv	7.7	99	83.5	7	100	100
ka	0	95.5	36.5	9.5	7.5	14.9
kk	93	100	79.5	0.5	100	11.1
ko	14	98	41.5	8.5	43.5	45.7
lt	14.3	84	66.2	25.6	23.5	1.5
ml	9.6	74.9	70.5	8.5	33.5	22.7
mr	0	82	72.3	100	42.5	96
ms	4.6	97.9	87.5	6.5	18	0
my	99	89.5	78.5	11.5	91	0
nl	19.6	94	26.3	0	28.7	10.3
pa	99.5	67.2	100	0	40	10.1
pl	18.4	100	22.2	0	16.5	0.5
pt	50.8	99.5	58.5	2.5	32.5	7.1
qu	93.3	100	100	97.5	96.5	93.5
ru	29.5	97.5	36.8	8.8	30.5	7.5
sw	100	100	94.5	8	99.5	98.5
ta	57.5	70.9	62.5	11.5	14.5	13.6
te	42.9	94	69.5	26	43.5	6.1
th	91.4	81.2	100	100	100	50.8
tl	0.5	100	70.5	12	58	16.2
tr	100	82.4	68.5	1.5	18	11.2
uk	12.8	95	36.8	24.8	56.5	5.1
ur	90.5	53	79.5	16	100	18.6
vi	18.1	99.5	84	9.5	100	4.1
wo	100	100	100	100	100	100
yo	100	100	100	100	100	98
zh	98.5	100	34	100	95	78.8
Avg	47.5	90.8	65.2	26.4	52.5	32.7

Templates and generated tests for different capabilities in English, Hebrew, Arabic, and Bengali (top) and Multilingual CheckList evaluation of XLM-R (right) [Ruder et al., 2021]

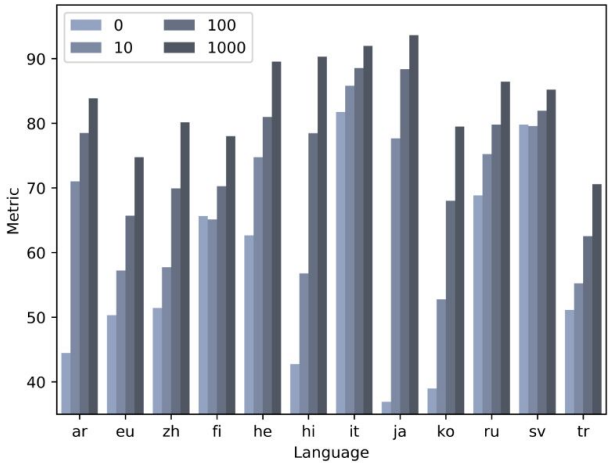
Creating Targeted Evaluation Data

- Create **template-based test cases**, e.g. using CheckList [Ribeiro et al., 2020]
- A **small number of templates** can cover many different model capabilities
- Scaling across languages still **requires native speaker expertise or translation**
- So far have been used for **evaluating reading comprehension** [Ruder et al., 2021] and **closed-book QA** [Jiang et al., 2020; Kassner et al., 2021]

Opportunity: Create targeted evaluation datasets in many languages

Few-shot Learning

- For many tasks, **fine-tuning on a small number of examples** can significantly improve performance in the target language [Hu et al., 2020; Hedderich et al., 2020; Lauscher et al., 2020] compared to zero-shot transfer
- **Fine-tuning on many languages** is better than on a few, with the same number of examples [Debnath et al., 2021]
- **Caveat**: More examples are needed for more challenging tasks [Kirstain et al., 2021]

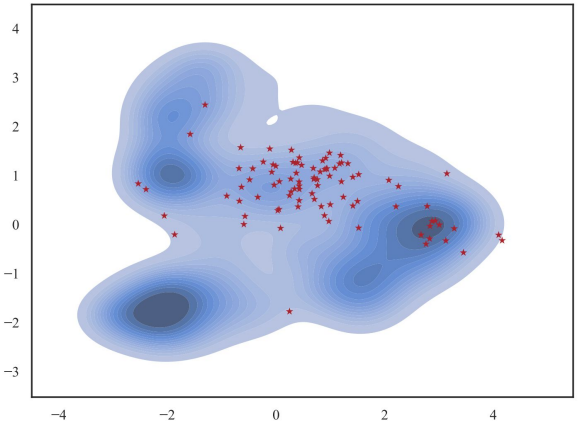


Dependency parsing results across different numbers of examples [Lauscher et al., 2020]

Opportunity: Focus on creating small training datasets with larger test sets in many languages

Cross-lingual Performance Prediction

- Instead of creating labelled data, **extrapolate performance** to languages without data [Ye et al., 2021]
- Evaluating models on many languages is expensive; could save costs by only **evaluating on a representative subset of languages** [Xia et al., 2020]
- Could also inform on which languages to **focus annotation efforts**
- Performance prediction methods have been evaluated on **languages in existing datasets such as UD**
- **Chicken-and-egg problem**: need labelled data in order to evaluate benefit of performance prediction for unseen languages

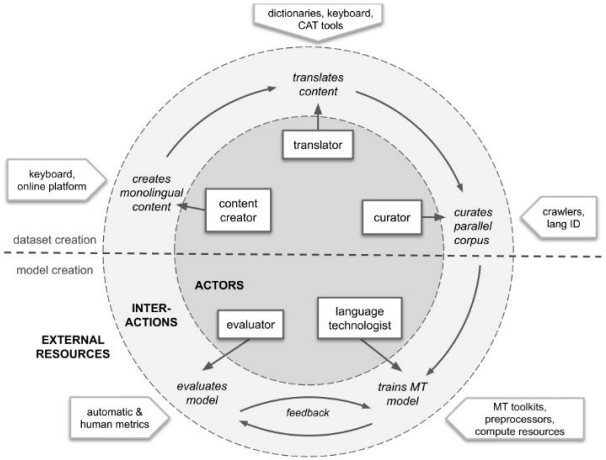


Density of WALS typological features of the world's languages. Red dots are languages in UD [Ponti et al., 2021]

Opportunity: Develop performance prediction methods that generalize to unseen languages with different linguistic characteristics

Participatory Research

- In order to create **high-quality, naturalistic data** in under-represented languages, we **need to work with language communities** directly
- Involvement of native speakers is beneficial beyond annotation:
  - Ensures that a **task formulation is realistic and beneficial** for a language community
  - **Prevents biases**, e.g., a Western-centric viewpoint
  - Enables **covering language varieties** such as dialects or different styles
  - See also Steven Bird's EMNLP 2021 keynote talk



Different stakeholders involved in the MT process [∀ et al., 2020]

Opportunity: Participatory research with grassroots communities such as Masakhane

02

Evaluation Setting

a Evaluation Protocol

b Evaluation Metrics

Google Research

Evaluation Protocol

- For cross-lingual transfer, there is a **bias towards the source language** (often English)
  - Favours languages **similar to the source language**
  - Other source languages often perform better [Lin et al., 2019; Anastasopoulos & Neubig, 2020; Turc et al., 2021]
- **Training on translations** helps particularly for some translated tests sets such as XNLI [Artetxe et al, 2020]
- Evaluation across many source languages enables a more fine-grained evaluation (but is also more expensive)

Opportunity: Consider the evaluation protocol and associated biases

Evaluation Metrics

- **Token-based metrics** (e.g., **F1, EM** in QA tasks) are **not appropriate for languages without whitespace separation** (e.g. Japanese, Thai, Chinese)
- Require a **language-specific segmentation method**, which introduces a dependence on the evaluation
- Metrics based on string matching such as **BLEU** are **not appropriate for morphologically rich languages**

English:
Her village is large.

Translations in Shipibo (spoken in Peru) [Valenzuela, 2003]:
Jawen jemara ani iki.
Jawen jemaronki ani iki.

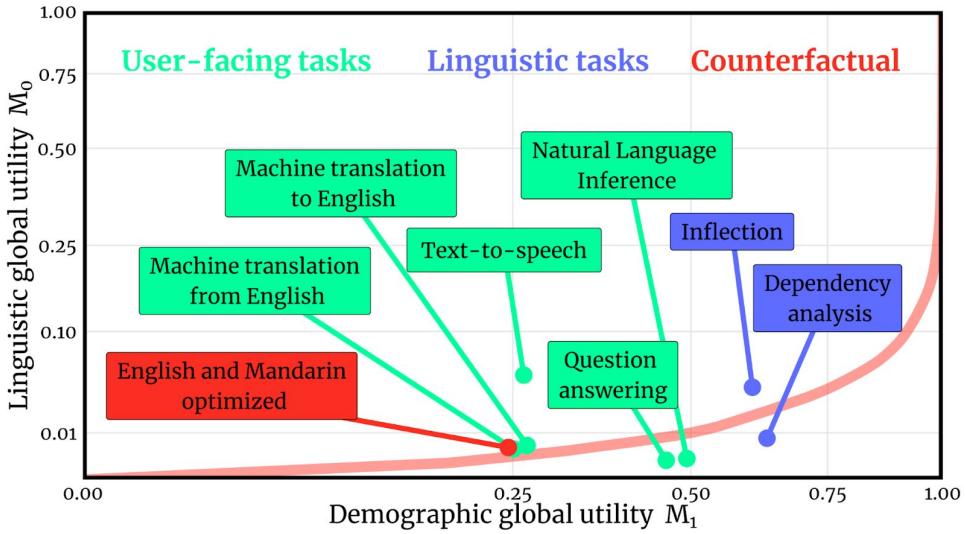
Evaluation Metrics

- **Token-based metrics** (e.g., **F1, EM** in QA tasks) are **not appropriate for languages without whitespace separation** (e.g. Japanese, Thai, Chinese)
- Require a **language-specific segmentation method**, which introduces a dependence on the evaluation
- Metrics based on string matching such as **BLEU** are **not appropriate for morphologically rich languages**

Opportunity: Consider the bias of your evaluation metric; use character-level evaluation, e.g. chrF [Popović, 2015]

Aggregating Evaluation Metrics

- **Averaging performance** only over languages in existing datasets provides a distorted view of progress
- Most languages in existing datasets are **high-resource**
- If we **consider all languages** of the world or **languages with large speaker populations**, the outlook is much more pessimistic [Blasi et al., 2021]

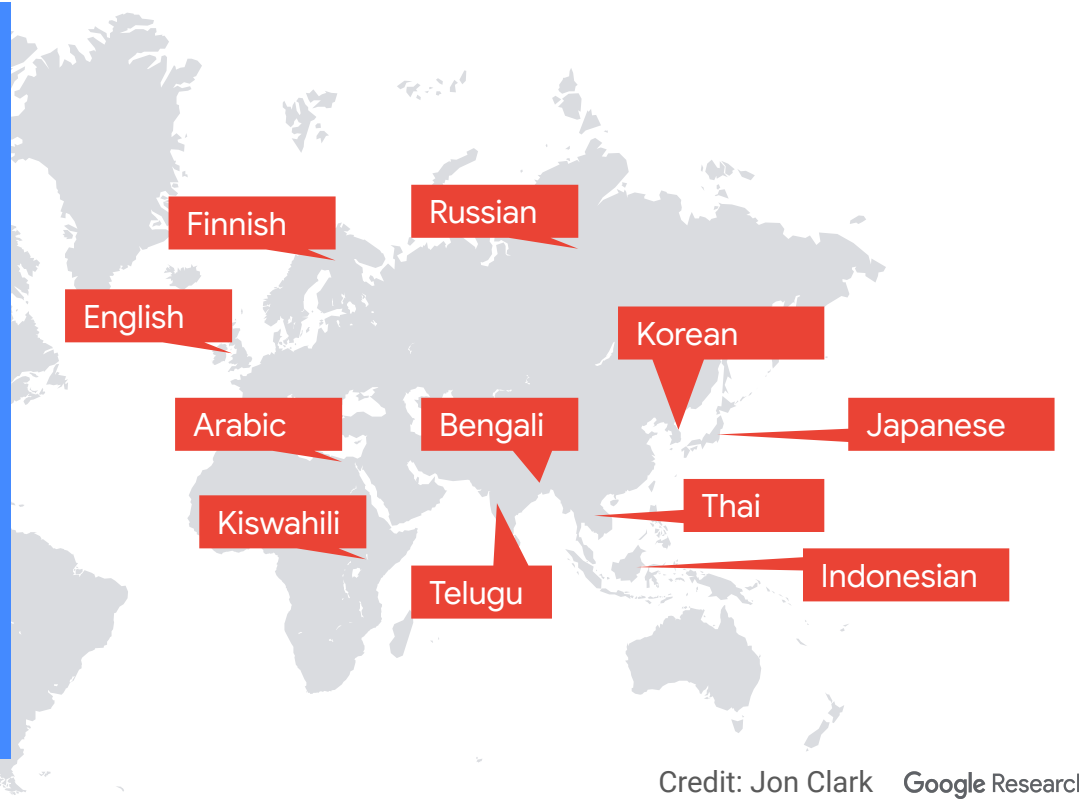


Linguistic vs demographic global utility of different NLP applications [Blasi et al., 2021]

Opportunity: Be cautious with aggregating performance; highlight performance on different language families, etc

Languages are diverse, not only linguistically but culturally

English	Fascinating language! ...but *no credit!*
Arabic	كتب
Bengali	সফেদা ফল খেতে কেমন
Finnish	jälleenrakennustöihin
Kiswahili	inayozungumzwa katika
Korean	우리 모두가 만들어가는
Indonesian	kecilkecilan
Japanese	24時間でのサーキット周回数
Russian	микроскопический
Telugu	ఖండాలలో అతిపెద్ద
Thai	เรือยิงตอร์ปิโดยูโกสลาเวีย ที5



Credit: Jon Clark Google Research

Takeaways

- Be **aware of biases** in existing multilingual datasets
- Aim to create datasets that **depart from a Western-centric viewpoint**
- To scale evaluation to many languages, we can…
  - Create **targeted evaluation datasets**
  - Create **small datasets across many languages** for few-shot learning
  - Develop **better cross-lingual performance prediction** methods
- **Participatory research** with native speaker communities can help to generate more high-quality, naturalistic data on tasks that native speakers care about

- Consider **biases in the evaluation protocol and in the evaluation metric**
- Be **cautious with aggregating performance**
- Highlight performance **across different language families and geographies**

Thank You

Google Research