ACL - 2022

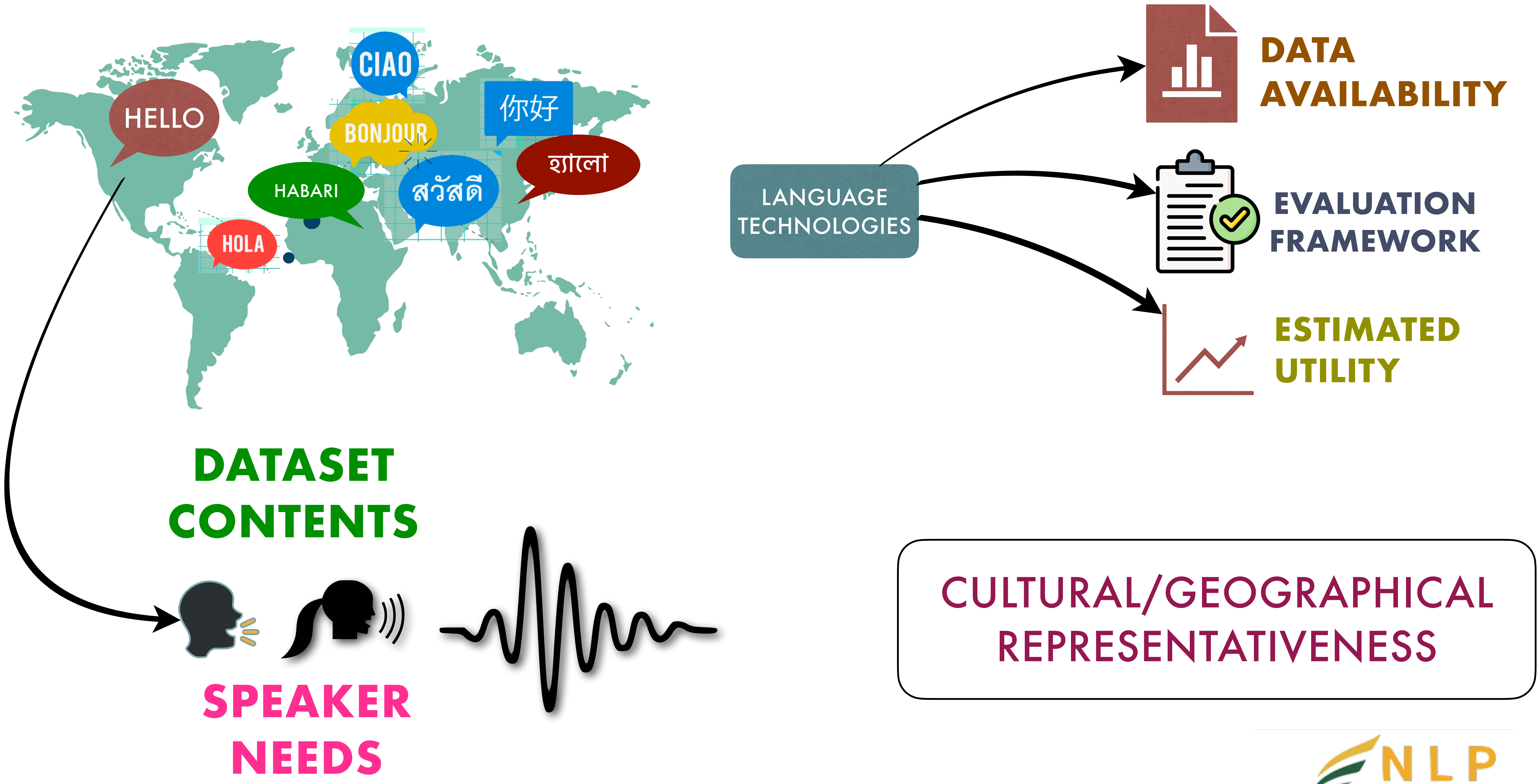
Dataset Geography: Mapping Language Data to Language Users

Fahim Faisal, Yinkai Wang, Antonios Anastasopoulos

ffaisal@gmu.edu

https://nlp.cs.gmu.edu/

GEORGE MASON UNIVERSITY


HELLO
CIAO
BONJOUR
你好
হ্যালো
HABARI
สวัสดี
HOLA
DATASET CONTENTS
SPEAKER NEEDS
LANGUAGE TECHNOLOGIES
DATA AVAILABILITY
EVALUATION FRAMEWORK
ESTIMATED UTILITY
CULTURAL/GEOGRAPHICAL REPRESENTATIVENESS

Our Contributions

MAP NLP DATASETS ↔ GEOGRAPHICAL AREAS — *Countries*

EVALUATE DATA REPRESENTATIVENESS OF LANGUAGE USERS — *Varying Results depending on Data Collection Process*

ANALYSIS SOCIO-ECONOMIC CORRELATES TO DATASET MAPS — *GDP, Geo-Distance, Population*

APPROACH ENTITY LINKING BYPASSING NER — *up to 85% accuracy*

Assumptions

1 Data Locality Matters

To Learn $p(\texttt{L1}|\texttt{text})$ (i.e. $p(\texttt{L1} = \texttt{Finnish}|\texttt{Finland})$)

Avoid Learning $p(\texttt{Finland}|\texttt{L1} = \texttt{Finnish})$

2 Capture locality by focusing on entities

English	ireland irish british britain russia scotland england states american london brexit
Finnish	finland finnish finns helsinki swedish finn nordic sweden sauna nokia estonian
French	french france paris sarkozy macron fillon hollande gaulle hamon marine valls breton

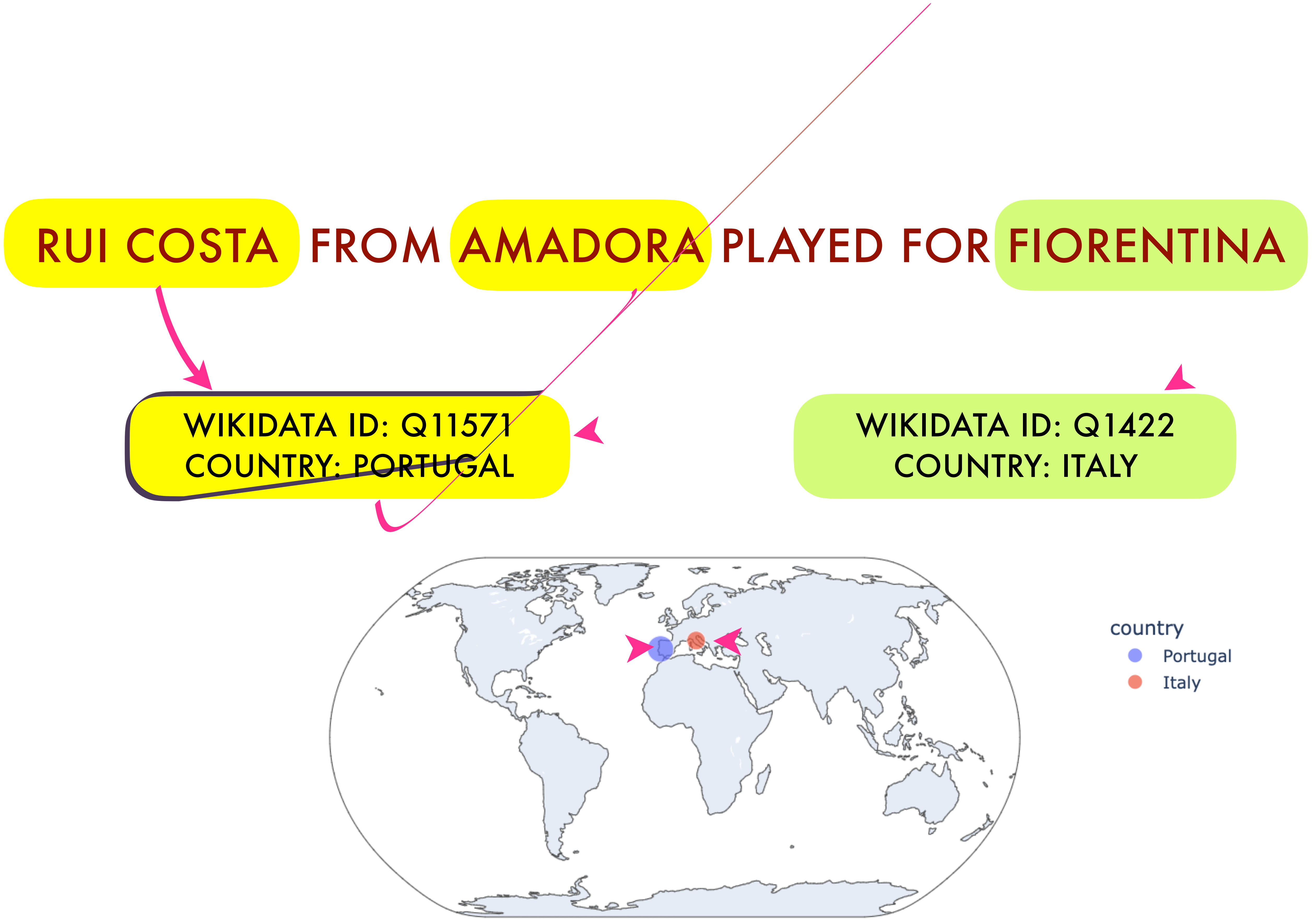
Top words based on log-odds scores for each label in the L2-Reddit dataset.

(Kumar et al. 2019)

Proposed Approach

For a given dataset

- ☑ Identify named entities
- ☑ Link entities to wikidata
- ☑ Aggregate through dataset
  - ☑ Representativeness measure
  - ☑ Fairness measure
  - ☑ Visualization



Proposed Approach

Mapping Dataset to countries

- Entity recognition-linking pipeline
- mGENRE (Cao et al. 2021): multilingual, seq2seq, auto-regressive entity linker
- Links to wikidata IDs

NER-INFORMED:

NER: [S]*Rui Costa*[E] from [S]*AMADORA*[E] played for [S]*FIORENTINA*[E]

NE-Link: {*Rui Costa*} from {*AMADORA*} played for {*FIORENTINA*}

- Bypassing NER Step to perform recognition & linking altogether

NER-RELAXED

[S]*Rui Costa from AMADORA played for FIORENTINA*[E]

{*Rui, score:-1*}, {*Costa, score:-1*}, {*Rui Costa,score:2*}, {*AMADORA,score:3*}, {*FIORENTINA,score:4*}

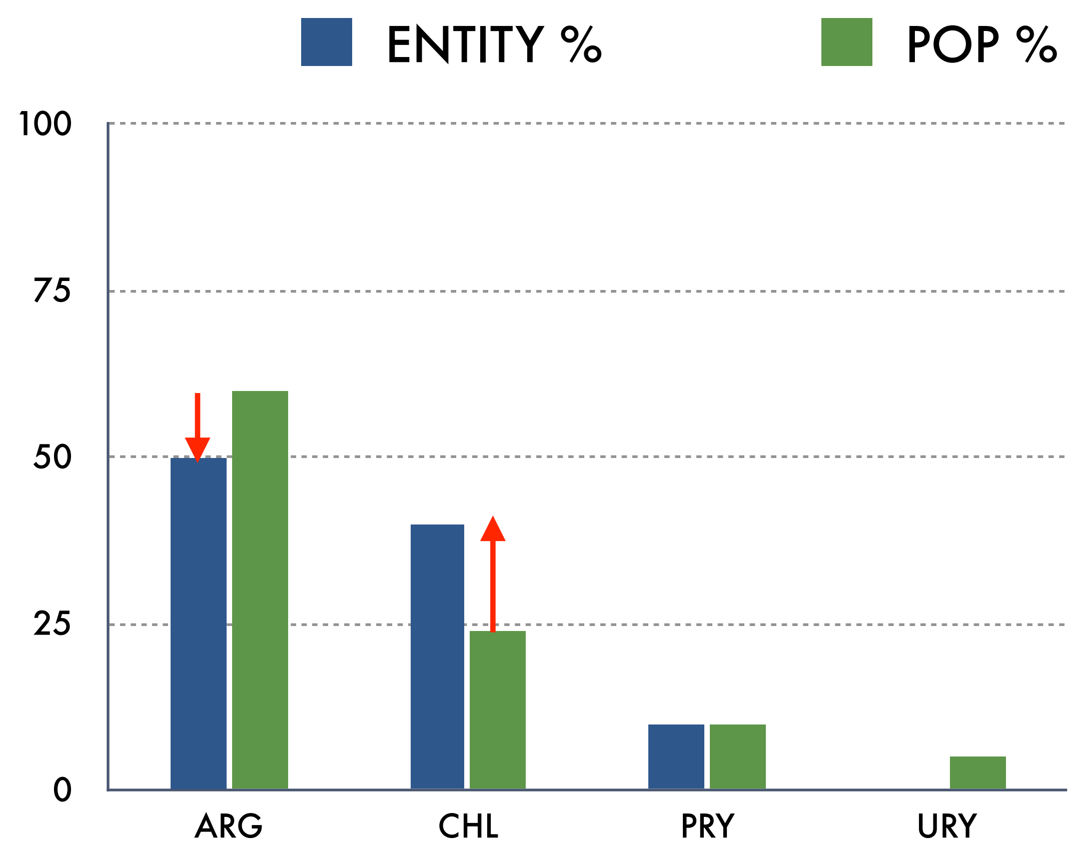
NLP GEORGE MASON

Proposed Approach

Representativeness measures from Dataset-Country Maps

- Entity percentage in Language speaking countries
  - country [SPANISH] = {ARG, CHL, PRY, URY}
  - entity [SPANISH] = (50+40+10+0)/total = 0.67
- Fairness indices
  - Country population
  - Country missing(1) or underrepresented (eg. URY~25%)
- In-country representativeness for widely spoken languages
  - Distribution Difference in speaker population & Observed entity

LANGUAGE: SPANISH

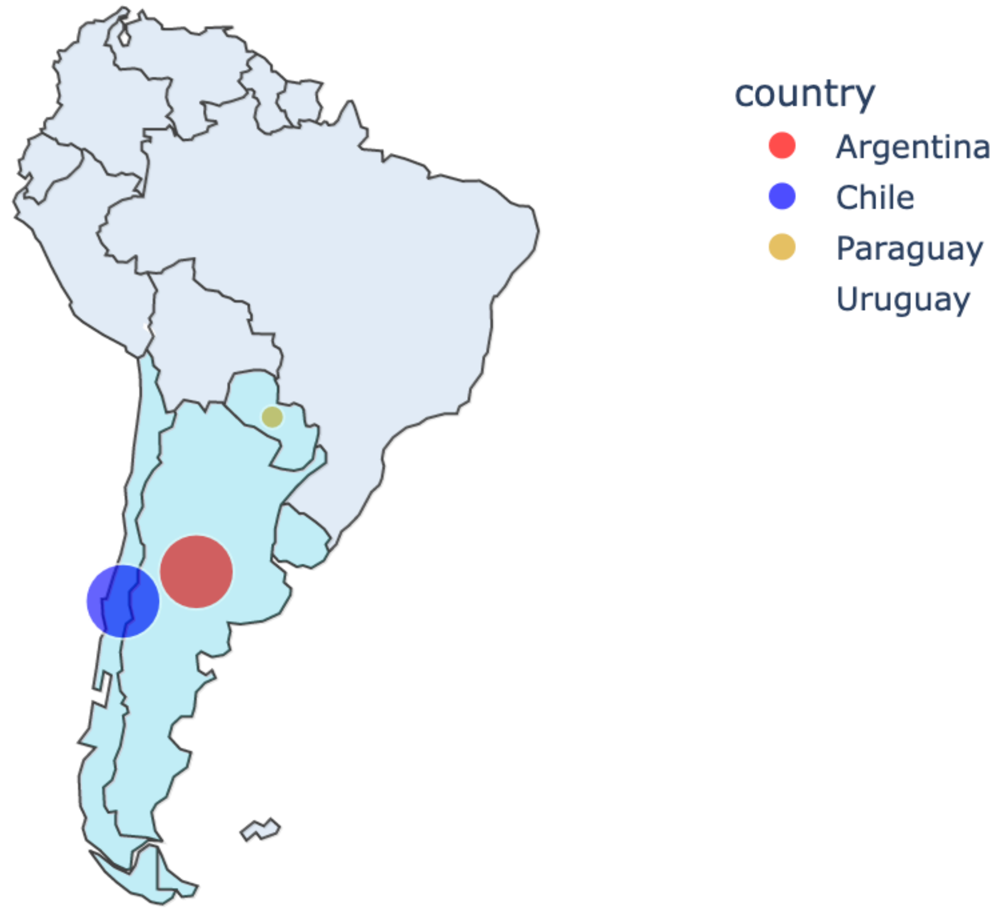


ARG -- 50

CHL -- 40

PRY -- 10

URY -- 0



Datasets and Settings

NER DATASETS

- WikiANN (Pan et al. 2017)
- Masakhaner (Adelani et al. 2021)

QA DATASETS

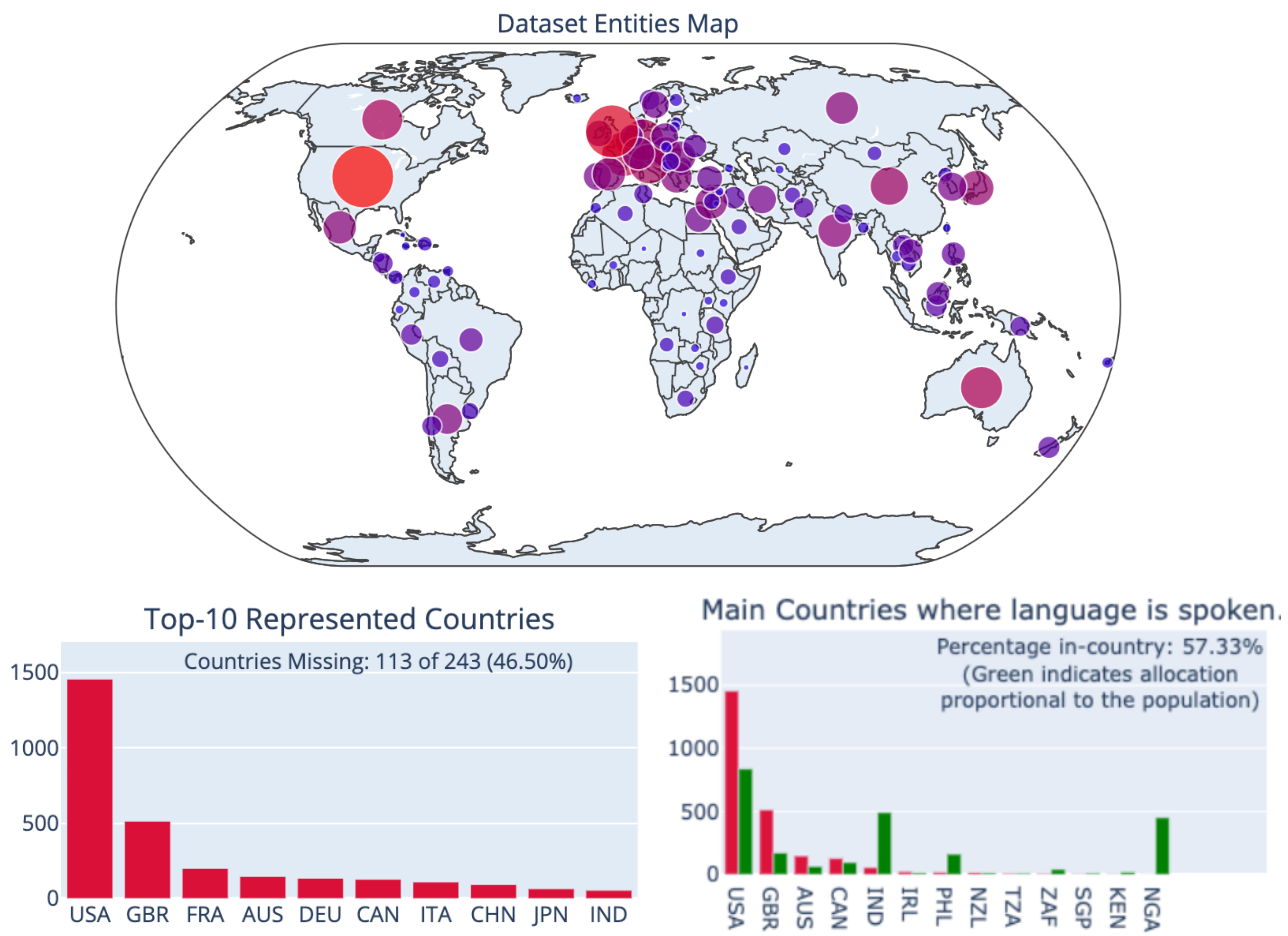
- SQuAD (Rajpurkar et al. 2016)
- MLQA (Lewis et al. 2020)
- TyDi-QA (Clark et al. 2020)
- Natural Questions (Kwiatkowski et al. 2020)

ADDITIONAL DATASETS

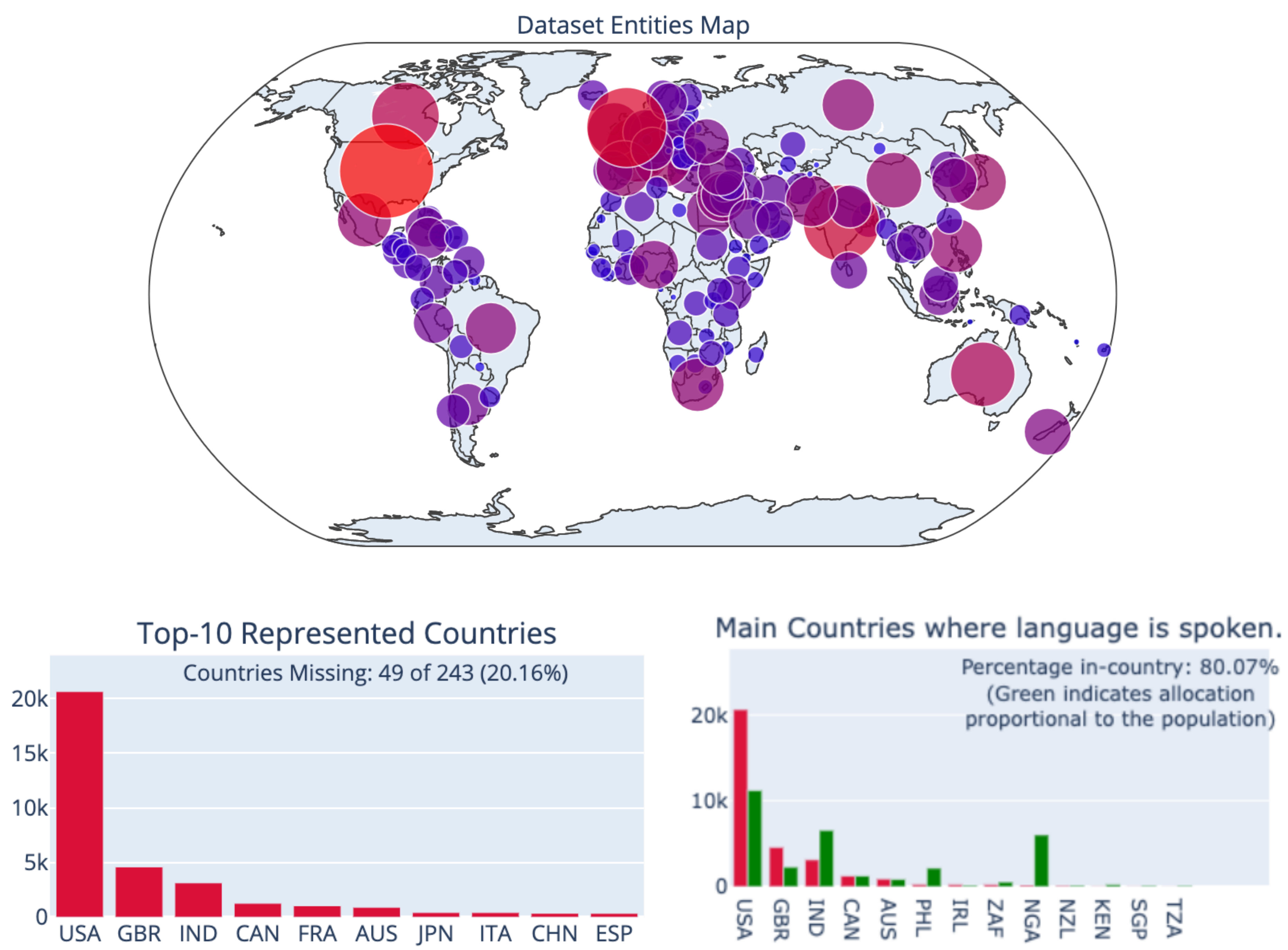
- Visualizations (X-FACTR benchmark~Jiang et al. 2020, WMT datasets) available in project webpage

Dataset Comparison (QA)

TyDi-QA(EN)



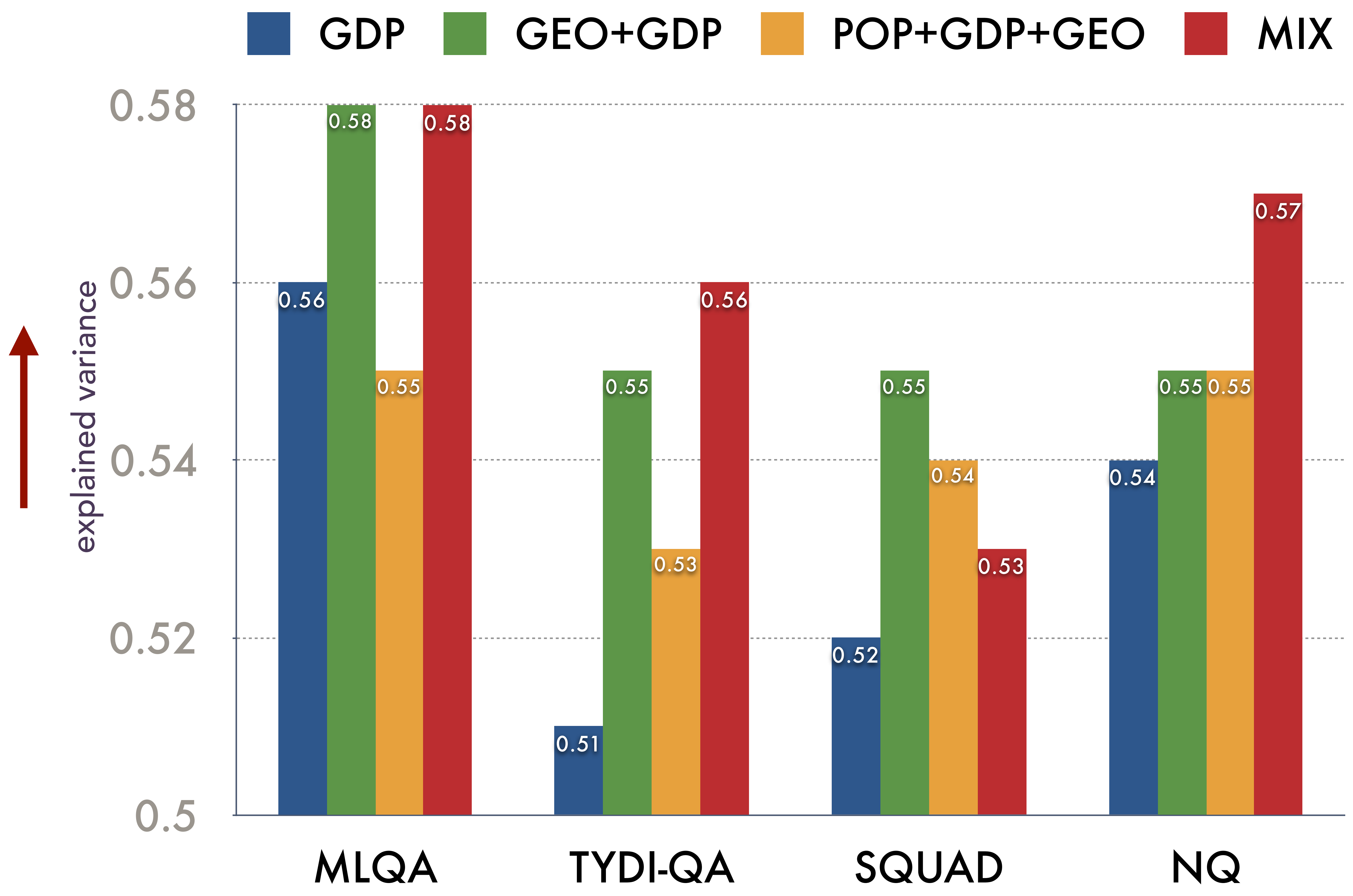
Natural Questions



Under-represented English Speakers in TyDi-QA(EN): Global South
(eg. Kenya, South Africa, Nigeria)

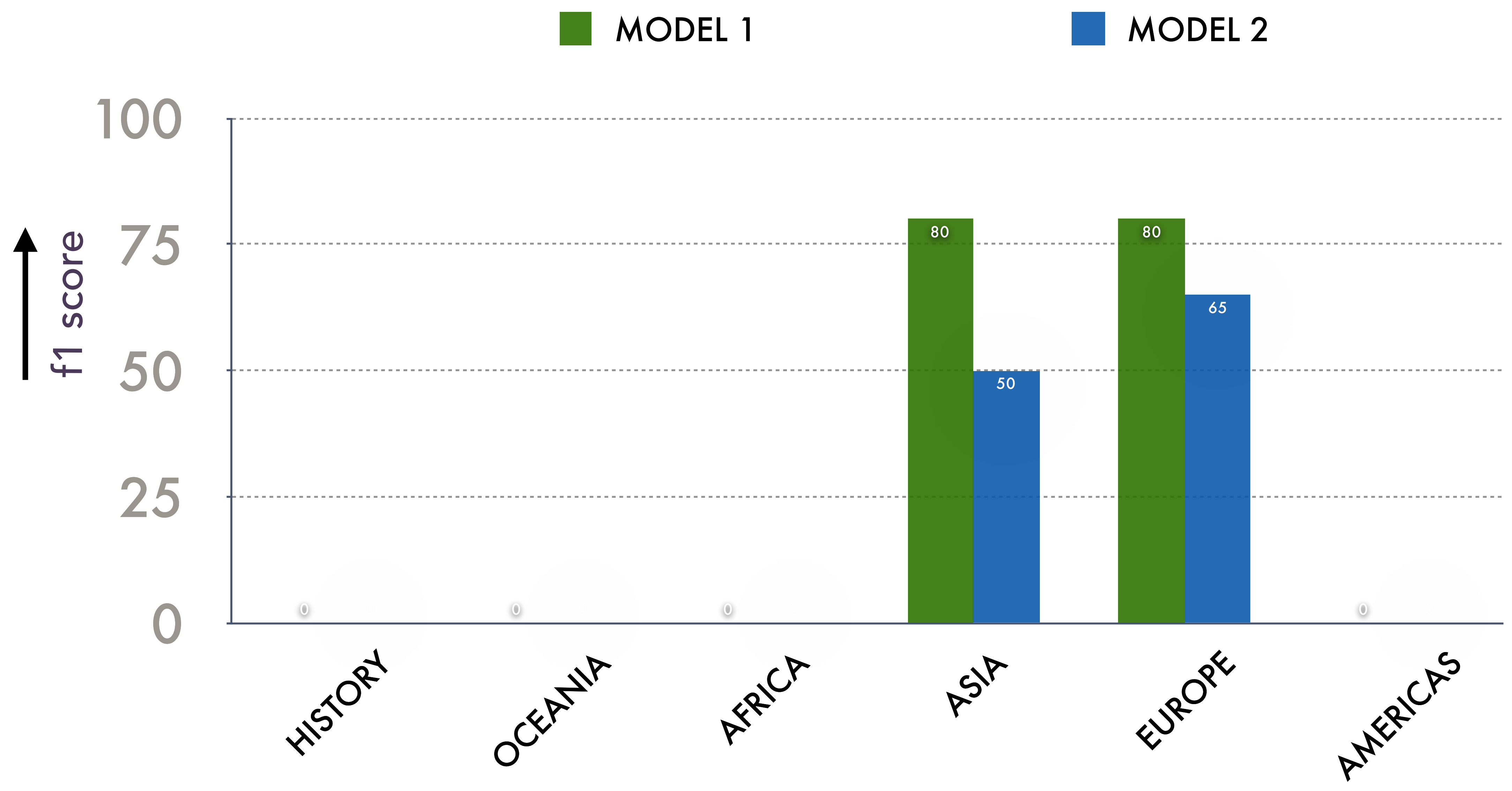
Socioeconomic correlates

- Single best predictor: GDP (all dataset over-representing wealthy countries)
- Including population statistics impact negatively except NQ (exemplar of representativeness)
- Mix of factors explain variance well.



- GEO: Distance from Language Spoken Country
- POP: population average
- MIX: combination of GDP, GDP/CAPITA, GEO, POP and Land-Mass

Geographical Breakdown (QA)



Model 1
- train: TyDiQA gold
- test: telugu TyDiQA dev

Model 2
- train: SQuAD english+translation,
- test: telugu TyDiQA dev

Model 2 performs worse on Asia-related data than Europe-related ones, unlike Model 1

A recipe for representativeness visualization for NLP datasets

- Country-language mapping: inherently lossy
- Granularity level smaller than country: higher cultural relevance
- Wikidata: western country bias
- Ideal combination of socioeconomic factors: subjective

- Robustness of NER/EL model
- Expansion of dataset and task coverage
- Inspect other granularity level

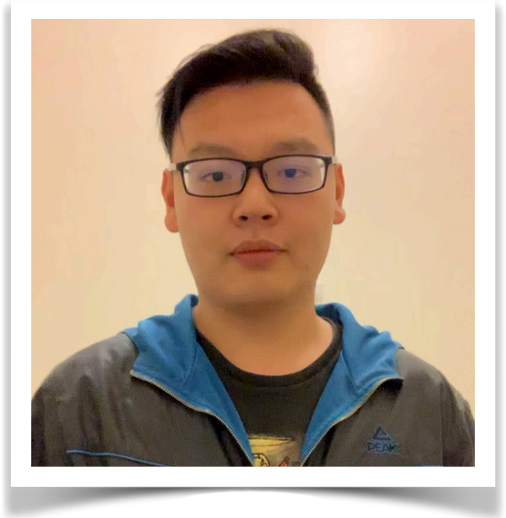
Code
&
Dataset

https://github.com/ffaisal93/dataset_geography

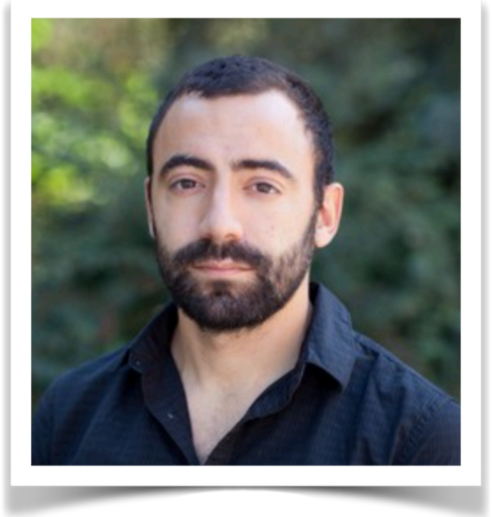
Project Webpage
&
Additional Visualizations

https://nlp.cs.gmu.edu/project/datasetmaps

FAHIM FAISAL (ffaisal@gmu.edu)



Yinkai Wang (ywang88@gmu.edu)



Antonios Anastasopoulos (antonis@gmu.edu)

NLP GEORGE MASON