Dataset Geography: Mapping Language Data to Language Users

Fahim Faisal, Yinkai Wang, Antonios Anastasopoulos

{ffaisal,yang88,antonis}@gmu.edu

Highlights

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

Email us for any questions!

Highlights

- ❑ **Task: Measuring Dataset-Representativeness.**
- ❑ Current NLP Research:
  - Not enough language coverage.
  - we should focus on language systems ***utility, not only accuracy*** (Blasi et al, 2022).
- ❑ **Are our datasets representative of the underlying language speakers?**
- ❑ We develop **Dataset-Geography**: the cultural representativeness of NLP datasets by mapping those onto geographical space.

Our Contributions

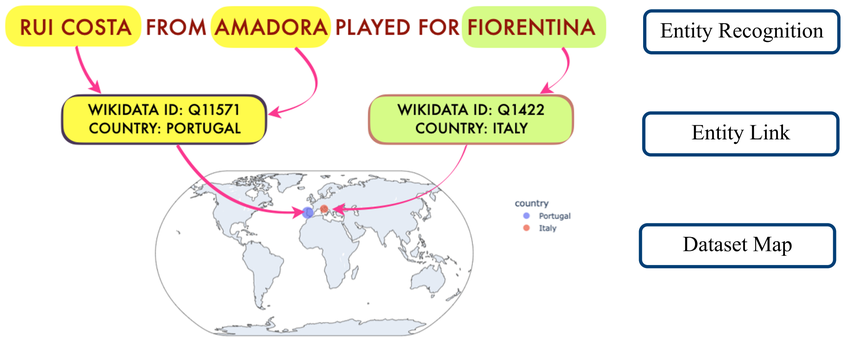
- Map: NLP Datasets-Geography (country).
- Evaluate: Data representativeness of language users
- Analysis: Explaining dataset maps through socio-economic correlates
- Approach: Entity-linking bypassing NER with upto 85% accuracy

Takeaways

- Significant disparity in terms of geographical mapping across datasets.
- Over-representation of wealthy countries.
- Dataset-Map Visualization reveals inherent biases
- Dataset building process influence system fairness.

Dataset Geography

Step 1: Mapping Dataset to countries



- Entity recognition-linking pipeline
- mGENRE (Cao et al. 2021): multilingual, seq2seq, auto-regressive entity linker
- Links to wikidata IDs
- We use NER-RELAXED approach with a small trade-off

NER-INFORMED: Separate NER & Entity Linking

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

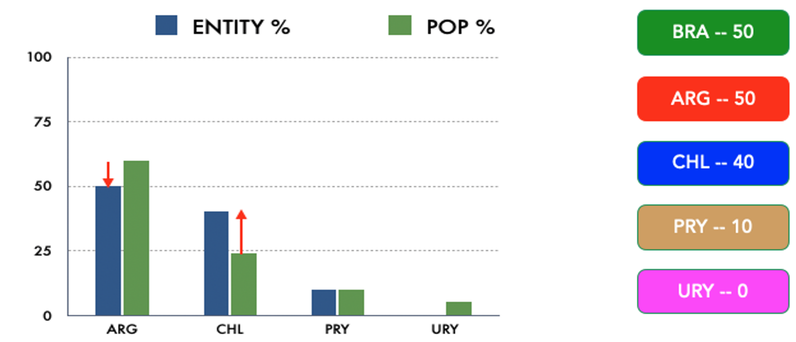
NE-Link: {*Rui Costa*} from {*AMADORA*} played for {*FIORENTINA*}

NER-RELAXED: NER & Entity Linking altogether using unconstrained mGENRE

[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*}

Step 2: Representativeness measures from Dataset-Country Maps



- **Entity percentage**
  - country [SPANISH] = {ARG, CHL, PRY, URY}
  - entity [SPANISH] = (50+40+10+0) /total = 0.67
- **Fairness indices**
  - Country population
  - Country missing (e.g. URY~25%)
- **In-country representativeness**
  - Distribution Difference in speaker population & Observed entity

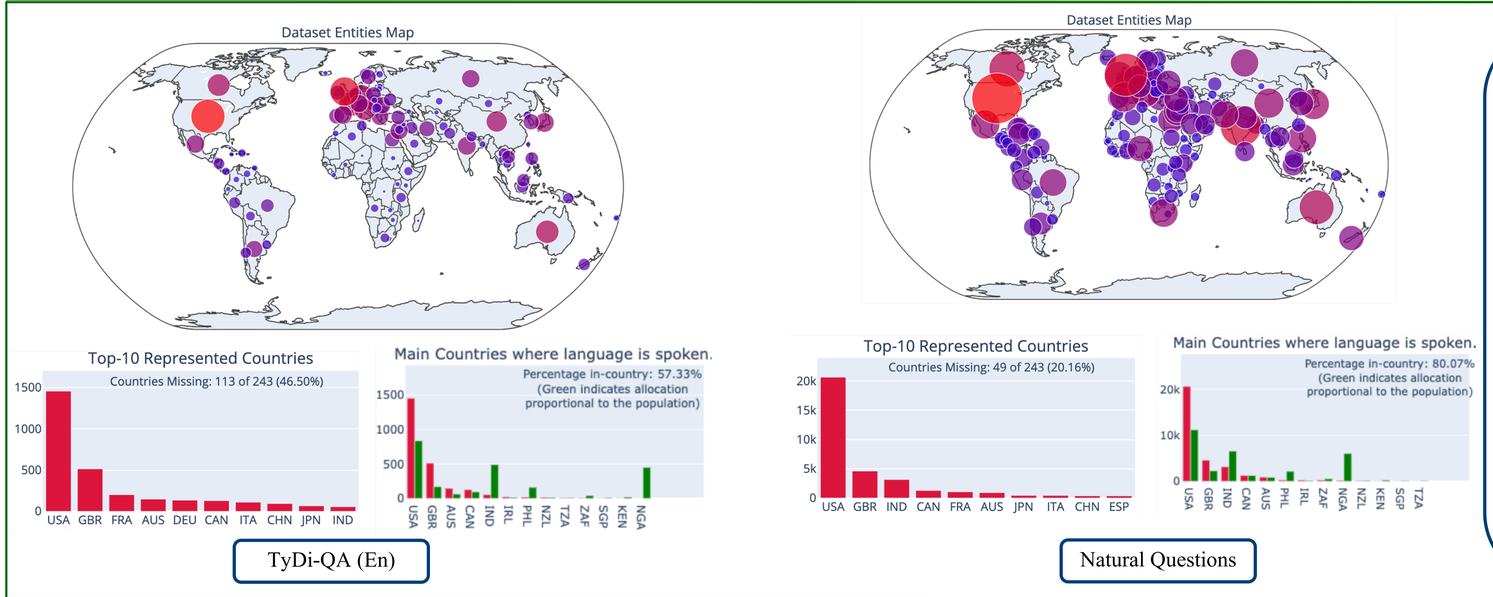
Datasets and Settings

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)

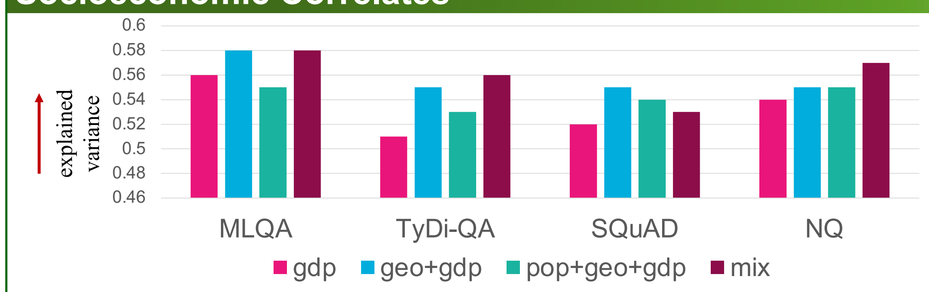
ADDITIONAL DATASETS
- X-FACTR benchmark (Jiang et al. 2020)
- WMT datasets

Dataset Map Comparison (QA)



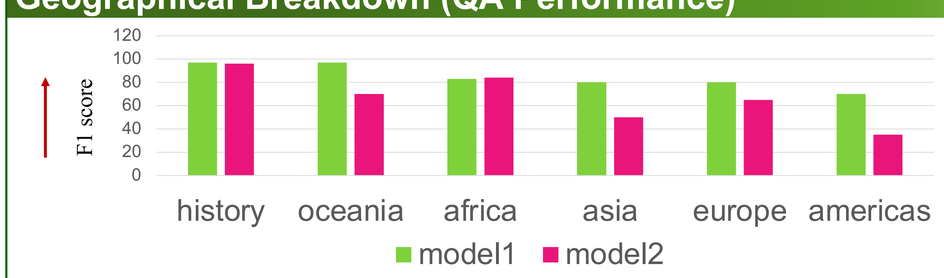
- **TyDi-QA (EN):** Under-represented English Speakers in Global South. (eg. Kenya, South Africa, Nigeria)
- **Natural Questions:** Well representation, built from global search queries

Socioeconomic Correlates



- Single best predictor: GDP (over-representing wealth countries)
- Including population statistics impact negatively
- Mix of factors explain variance well

Geographical Breakdown (QA Performance)



- Model 2 performs worse on Asia-related data than Europe-related ones, unlike Model: Unfairness because of representative entity lacking

Model1: TyDiQA, **Model2:** SQuAD~translate-train, **Evaluation:** TyDi-QA telugu