Geo-Cultural Representation and Inclusion in Language Technologies

Sunipa Dev
Google Research
sunipadev@google.com

Rida Qadri
Google Research
ridaqadri@google.com

Abstract

Training and evaluation of language models are increasingly relying on annotations by humans to judge questions of representation and safety. While techniques such as RLHF are being broadly applied, there is less consideration of how socio-cultural identity and positionality of the annotators involved in this process play a key role in what is taken as ground truth by our models. Yet, we currently do not have ways to integrate rich and diverse community perspectives into our language technologies.

Accounting for such cross-cultural differences in interacting with technology is an increasingly crucial step for evaluating AI harms holistically. Without this, the state of the art of the AI models being deployed is at risk of causing unprecedented biases at a global scale. This tutorial uses interactive exercises to illustrate how cultural identity of annotators and varying methods of human feedback influence evaluations of appropriate representations of global concepts.

1 Introduction

Increasingly, researchers and engineers are relying on human annotation to train, develop, and shape language models. However, as language models are being integrated into global systems of social and cultural importance such as search, education, and even creativity, the annotation tasks veer into increasingly culturally subjective questions of evaluating representation, toxicity, abusive language, stereotyping, and more. Measuring such representational quality of generated content requires significant culturally situated expertise and nuanced judgment on specific signifiers and social connotations of language (Qadri et al., 2023). Who you ask and how you ask them also changes the content of such subjective evaluations(Denton et al., 2021; Dev et al., 2023). To highlight this contingency of our existing evaluation methods, in this tutorial we will work through the following questions together:

1. How do we account for socio-cultural identities and perspectives of the annotators training our models?

2. How do we resolve disagreements in annotations when they come from culturally different raters for a subjective task?

3. What do qualitative and open-ended methods offer us as a mode of evaluation?

4. How can new research on understanding socially subjective data annotation tasks help build more robust, generalizable, and safe models?

1.1 Relevance at LREC-COLING

NLP research and development has seen immense, fast-paced progress in recent years, with a large growth in generative language models both in size, and number. Their capabilities have also increased and diversified, making their evaluations that much harder, but also more critical. However, as has been seen, these evaluations of models mostly focus on Western perspectives across a board of tasks from language fluency to NER. When we consider tasks closely related to experienced biases and harms, this concern magnifies (Davani et al., 2023). Harms faced by people in different parts of the world goes unchecked, and populations are often misrepresented or not represented at all in the model outputs. This major gap hints at a need for advancements in existing evaluation paradigms, and a recalibration of the approaches towards data annotation and aggregation.

We will discuss this pressing topic through emerging, state-of-the-art research in the area. With methodologies such as RLHF, and human centered AI fast developing, and cutting edge AI technologies being integrated into lives globally, these discussions at computational linguistics venues will be imperative towards fostering inclusive practices

LREC-COLING 2024 Tutorials, pages 9–12

around data resource creation, model building, and evaluations.

2 Outline

2.1 Tutorial Content

This tutorial will adopt three interactive annotation exercises and discuss approaches and the results obtained from them. All participants will together rate some sample questions in each exercise.

The first two exercises will ask for binary or categorical answers in response to first, a culturally under-specified question for instance quality of a response on music or film without a cultural locale specified, and then a statement with cultural specificity, such as a text quality of a model generated paragraph about people from a nationality or culturally specific facts about an area or population. These two exercises will open space for discussion on the varying forms of expertise annotations require and whether binary or closed-ended questions capture this cultural expertise

The third exercise will pair up individuals of different cultural backgrounds to evaluate generated text from each other's cultural background. The mode of evalution will be open ended.

The tutorial will the discuss the pros and cons of these approaches, the subjectivity of annotations, and ways to incorporate them into our NLP pipelines. In doing so, it will demonstrate the importance of culturally situated, and deeply engaged strategies of data collection and annotation. It will discuss the need for well documented, distributed, and diversely annotated data for ensuring data (for both training and measurement) quality.

3 Tutorial Structure

We have structured the tutorial into the following parts. Each part will be interactive and we will encourage questions throughout the tutorial. We will also keep aside at least the last 7 minutes of each of the following sessions to be just for Q/A.

Part 1 - Context and Motivation [45 mins] We will begin with a short, introductory talk by the presenters where we will motivate the problem setup and give examples of how cultural subjectivity and expertise can shape evaluation outcomes. We will also demonstrate how these differences impact what is treated as 'ground truth' by our AI pipelines. Specifically, in tasks that check for model safety and beneficence, these discrepancies can lead to representational as well as quality of service harms.

Part 2: Live rater annotation [45 mins] This segment of the tutorial will be extremely hands on, and aimed at investigating together how our experiences shape the way we annotate presence or absence of certain features in text or image data points. The task will be shared through a web link during the tutorial.
The total time for this segment will be split in the following way:

- Annotation [15 mins] Introduce text snippets and do two exercises to have the audience evaluate the two types of generated text : culturally under specified and culturally specified.
- Review of what was annotated [30 mins] Collective review of results of annotation exercise to discuss what kinds of knowledge did the person leverage to answer and if a binary rating was able to capture their feedback?

Coffee Break: 30 mins

Part 3: Cross-Cultural Annotation [45 mins]

- Annotation [15 mins] Cross Open ended questions on cultural quality of the text and explanations of what the models did well what it did poorly
- Discussion of the specificity of cultural expertise needed to evaluate text and what annotators of other identities missed or picked up on[30 mins]

Discussion and Closing [30 mins] We will spend the last 30 minutes summarizing the tutorial and answering any additional questions.

4 Target Audience

The target audience for this could be NLP researchers, engineers, and practitioners at any career stage. They could be actively using annotated data to rain or evaluate models, or creating the datasets for these purposes. With the discussions and exercises at the tutorial, they will collectively reflect on the range of impacts each rater assumption and rating task structure choice has.

Prerequisite Knowledge: No specific prerequisite knowledge is needed. However, a general knowledge of data annotations and/or evaluation tasks in NLP could be helpful.

Equipment needed: Venue with wifi so participants can engage with material.
Attendees are recommended to bring their laptops for better experience.

5 Diversity Statement

The topic of the tutorial is very tightly linked with the mission of diverse representation of people in NLP. The tutorial highlights how differing lived experiences across the globe impact what is 'ground truth' in data annotations for different people. Unilateral decisions or tasks only considering majority over categorical ratings do not do justice to the subjective tasks that LLMs built on these datasets perform. Through this tutorial we will elaborate the importance of global inclusion into NLP technologies for equitable model development and deployment.

6 Other Information

The presenters have experience introducing and leading discussions on cultural considerations in AI pipelines. Some other venues where we have co-organized and conducted tutorials and workshops with a similar goals include FAccT 2023 (Tutorial on Cross Cultural Considerations in AI; 50 attendees), EACL 2023 (Cross Cultural Considerations in NLP Workshop; 75 attendees), NeurIPS 2022 (Cultures in AI Workshop; 50 attendees), CVPR 2023 (Ethical Considerations in Creative Applications of Computer Vision).

With this track record of successful events on this theme at multiple venues, we expect a similar range of attendees at COLING. We will also be advertising the tutorial through multiple channels including social media, and mailing lists.

7 Reading List

1. Whose Ground Truth? Accounting for Individual and Collective Identities Underlying Dataset Annotation; Emily Denton, Mark Díaz, Ian Kivlichan, Vinodkumar Prabhakaran, Rachel Rosen; Data Centric AI Workshop at NeurIPS 2021 ((Denton et al., 2021))

2. SeeGULL: A Stereotype Benchmark with Broad Geo-Cultural Coverage Leveraging Generative Models; Akshita Jha, Aida Mostafazadeh Davani, Chandan K Reddy, Shachi Dave, Vinodkumar Prabhakaran, Sunipa Dev; ACL 2023 ((Jha et al., 2023))

3. Probing pre-trained language models for cross-cultural differences in values; Arnav Arora, Lucie-Aimée Kaffee, Isabelle Augenstein; Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP) at EACL 2023 ((Arora et al., 2023))

4. Cultural Incongruencies in Artificial Intelligence; Vinodkumar Prabhakaran, Rida Qadri, Ben Hutchinson; Cultures and AI Workshop at NeurIPS 2022 ((Prabhakaran et al., 2022))

5. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study; Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, Daniel Hershcovich; Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP) at EACL 2023 ((Cao et al., 2023))

8 Presenter Bios

Sunipa Dev (she/her, Google Research, sunipadev@google.com) is a Senior Research Scientist at Google Research working towards fair, inclusive, and socio-culturally aware NLP. Her research centers around inclusion of global perspectives in different pipelines in NLP, particularly in model evaluations to better understand and mitigate potential risks and harms. Prior to this, she was an NSF Computing Innovation Fellow at UCLA, before which she was awarded her PhD at the School of Computing at the University of Utah.

She has taught guest lectures and given talks centered on inclusive NLP at multiple places including University of Utah (2023), University of Southern California (2023), University of Bocconi (2021), and a keynote at TrustNLP Workshop (ACL 2023). She is currently a program chair for WINLP (organizing across different NLP venues including NAACL, ACL, and EMNLP), and was the affinity workshop chair at NeurIPS 2022 and a workflow chair for AAAI 2022. She has also co-organized tutorials and workshops at various venues including KDD 2021, NeurIPS 2022, EACL 2023, and FAccT 2023.

Rida Qadri (she/her, Google Research, ridaqadri@google.com) Rida Qadri is a Senior Research Scientist at Google Research.

Her research interrogates the cultural assumptions underpinning the design and deployment of generative AI systems. She specifically focuses on the harms produced by culturally inappropriate AI design choices and documents how communities resist and repair these technologies.

She has given guest lectures on cultural failures of AI at MIT, University of North Carolina, Maastricht University and spoken on keynote panels at FAccT 2022 and IEEE world AI IOT Congress. She has co-organized workshops at the intersection of AI and Culture at NeurIPS 2022, CHI 2021 and CVPR 2023. She has a PhD in Computational Urban Studies from the Massachusetts Institute of Technology.

9 Ethics Statement

This workshop will help draw attention towards the ethics of globally deploying models which incorporate world views of only few parts of the world, both in its training and evaluations. It will urge deeper reflections of how each data instance that we use to build or evaluate a model can have different interpretations by different people and communities globally. By doing so, this tutorial will be actively fighting against further marginalizations or erasure of people from different communities and cultures.

References

Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2023. Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia. Association for Computational Linguistics.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study.

Aida Mostafazadeh Davani, Mark Diaz, Dylan Baker, and Vinodkumar Prabhakaran. 2023. Disentangling disagreements on offensiveness: A cross-cultural study. In *The 61st Annual Meeting of the Association for Computational Linguistics*.

Emily Denton, Mark Díaz, Ian Kivlichan, Vinodkumar Prabhakaran, and Rachel Rosen. 2021. Whose ground truth? accounting for individual and collective identities underlying dataset annotation. *arXiv preprint arXiv:2112.04554*.

Sunipa Dev, Jaya Goyal, Dinesh Tewari, Shachi Dave, and Vinodkumar Prabhakaran. 2023. Building socio-culturally inclusive stereotype resources with community engagement.

Akshita Jha, Aida Mostafazadeh Davani, Chandan K Reddy, Shachi Dave, Vinodkumar Prabhakaran, and Sunipa Dev. 2023. SeeGULL: A stereotype benchmark with broad geo-cultural coverage leveraging generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9851–9870, Toronto, Canada. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Rida Qadri, and Ben Hutchinson. 2022. Cultural incongruencies in artificial intelligence.

Rida Qadri, Renee Shelby, Cynthia L Bennett, and Emily Denton. 2023. Ai's regimes of representation: A community-centered study of text-to-image models in south asia. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 506–517.