Ling/CSE 472:
Introduction to Computational Linguistics

Data and Model Documentation

Overview

- Questions about term projects?
- Data statements - v1
- Data statements workshop => v2, guide
- Reading questions
- NLP/Compling in the news

Bias and associated impacts

- **Bias**: cases where computer systems "*systematically* and *unfairly discriminate* against certain individuals or groups of individuals in favor of others" (Friedman & Nissenbaum 1996:332)

- **Pre-existing bias**: Bias with roots in social institutions, practices and attitudes

- **Technical bias**: Seemingly neutral technical decisions producing bias in real-world contexts

- **Emergent bias**: When a system designed for one context is deployed in another

2018: A flourishing of work on standards for documentation of models, systems, datasets

- Gebru et al 2018: Datasheets for datasets
- Chmielinski et al (MIT Media Lab): Dataset nutrition labels
- Yang et al 2018: Ranking facts
- Mitchell et al 2019: Model cards
- Diakopoulos et al 2016, Shneiderman 2016, AI Now Institute 2018: Algorithmic Impact Statements

Value tensions

- Transparency v. privacy
    - Strive for as much transparency as possible without exposing information about particular individuals
    - Plan ahead: Ask for permission to include demographic information
- Thoroughness v. ubiquity
    - Data statements should accompany all datasets and all models/ experiments built on them
    - Long form and short form (pointing to long form)

Proposed Schema: Long Form

- A. Curation Rationale
- B. Language Variety
- C. Speaker Demographic
- D. Annotator Demographic
- E. Speech Situation
- F. Text Characteristics
- G. Recording Quality
- H. Other
- I. Provenance Appendix

A. Curation Rationale

- Which texts were included and what were the goals in selecting texts, both in the original collection and in any further sub-selection?

- Especially important in datasets too large to thoroughly inspect by hand.

- Can help dataset users make inferences about what other kinds of texts systems trained with them could conceivably generalize to.

C. Speaker Demographic

- What demographic groups do the speakers represent?

- Variation in pronunciation, prosody, word choice, and grammatical structures also correlates with speaker demographic characteristics (Labov, 1966)

- Speakers use linguistic variation to construct and project identities (Eckert and Rickford, 2001)

- Transfer from native languages (L1) can affect the language produced by non-native (L2) speakers (Ellis, 1994, Ch 8)

- Disordered speech (e.g. dysarthria) leads to further variation

C. Speaker Demographic

- Age
- Gender
- Race/ethnicity
- Native language
- Socio-economic status
- Number of different speakers represented
- Presence of disordered speech

E. Speech Situation

- Time and place
- Modality (spoken/signed, written)
- Scripted/edited vs. spontaneous
- Synchronous vs. asynchronous interaction
- Intended audience

F. Text Characteristics

- Genre: “Text categorizations made on the basis of external criteria relating to author/speaker purpose”

- Topic: What the text is about

Genres of text

Total Results: 0

Won’t It Get Repetitive?

- Include an NLP data statement in every paper?
  - Really?
  - Even for things like the PTB (Marcus et al 1993) that are familiar to everyone?
- Yes!
  - Always consider how the dataset fits the current study
  - Always consider how the results of the current study do & don’t generalize

How do data statements help?

- Emergent bias: Procurers, consumers and advocates can check whether a system is trained on appropriate data for its deployed use case

- Emergent bias: As a field, we can track what speaker populations are underserved

- Pre-existing bias: Knowing what kind of texts a system is trained on can be key to working out the source of bias, as in Speer's (2017) study of word embeddings and sentiment analysis

Data statements alone won't 'solve' bias, but if we do not make a commitment to data statements or a similar practice for making explicit the characteristics of datasets, then we will single-handedly undermine the field's ability to address bias.

Tech Policy: Proposed Best Practice

- If NLP data statements turn out to be as useful as predicted, we see two implications for tech policy:

    - For academia, industry and government, inclusion of *long-form* data statements should be a required part of system documentation. As appropriate, inclusion of long-form data statements should be a requirement for ISO and other certification. Even groups that are creating datasets that they don't share (e.g. NSA, IARPA) would be well advised to make internal data statements.

    - For academic publication in journals and conferences, inclusion of *short-form* data statements should be a requirement for publication. Implement with care to avoid barriers to access.

Tech Policy: Sensitive Information

- There may also be security and secrecy concerns for some groups in some situations.

- There may be groups who are willing to share datasets but not demographic information (e.g. for fear of public relations backlash or to protect the safety of contributors to the dataset).

As consumers of datasets or products trained with them, NLP researchers, developers and the general public would be well advised to use systems **only** if there is access to the type of information we propose should be included in data statements.

Reading questions

- What is the current state of data statements in NLP? Have they been successfully adopted?

- How successful has the proposed schema been? Is there any kind of information mentioned in this paper that has turned out to be less important? Or is there a kind of information not mentioned in this paper that might be important to include in a data statement?

2020: Workshop “at” LREC
McMillan-Major, Bender & Friedman (in press)

- Three-days, online
- Researchers from every continent except Antarctica
- Develop data statements (v1) in pairs (interview technique)
- Group reflections on process, best practices

2020: Workshop “at” LREC
McMillan-Major, Bender & Friedman (in press)

- Analysis of participants’ data statements
- Analysis of group discussions
- Comparison with Datasheets for Datasets (Gebru et al 2018, 2021)
- => version 2 of the data statements schema
- => Guide to Writing Data Statements (Bender, Friedman & McMillan-Major 2021)

A Guide to Writing Data Statements for Natural Language Processing

- Bender, Friedman & McMillan-Major 2021

From McMillan-Major et al (in press)

Revisions	Phase 1: Workshop	Phase 2: Datasheet Comparison
General Best Practices	New	-
Key Terms	New	-
Schema Elements		
1 Header	New	Updated c
2 Executive Summary	New	-
3 Curation Rationale	Updated b, c, d	Updated c
4 Documentation for Source Datasets	Updated a, b, c, d	Updated c
5 Language Varieties	Updated a, b, c, d	-
6 Speaker Demographic	Updated b, c, d	-
7 Annotator Demographic	Updated b, c, d	-
8 Speech Situation and Text Characteristics	Merged and updated a, b, c, d	
9 Preprocessing and Data Formatting	New	Updated c
10 Capture Quality	Updated a, b, c, d	-
11 Limitations	New	-
12 Metadata	New	Updated c
13 Disclosures and Ethical Review	New	-
14 Other	Updated b, c, d	-
15 Glossary	New	-

Table 2. Revisions by source of change. Each element is comprised of a: (a) title, (b) rationale, (c) description, and (d) best practices. "New" refers to the addition of an entirely new element.

Schema Version 1	Schema Version 2	Changes
	⋮	
A. Curation Rationale Which texts were included and what were the goals in selecting texts, both in the original collection and in any further sub-selection? This can be especially important in datasets too large to thoroughly inspect by hand. An explicit statement of the curation rationale can help dataset users make inferences about what other kinds of texts systems trained with them could conceivably generalize to.	**3 Curation Rationale** *Why* For dataset creators, a curation rationale can help to promote intentionality in data selection and ensure representativeness. In addition, as difficult decisions arise, an explicit rationale can help to structure and resolve discussions about the data collection process and select pathways going forward. For data statement readers, an explicit statement of why and how the dataset was curated can help with inferences about the domain of generalizability of systems trained on the dataset. Knowing which texts were included, and what the goals were in selecting texts, can be especially important in datasets too large to thoroughly inspect by hand. *What* The curation rationale should answer questions including: Why was this dataset created? What is the task or research question the dataset is intended to address? Which texts were included and what were the goals in selecting texts, both in the original collection and in any further sub-selection? What is the internal organization of the dataset? What constitutes a data instance?	Element moved to third position after analysis of workshop participants' data statements Elaboration of motivation added after analysis of workshop results Motivation for the element moved to the first ('Why') part of the description Elaboration after analysis of data statements produced by workshop participants
⋮	⋮	
G. Recording Quality For data that include audiovisual recordings, indicate the quality of the recording equipment and any aspects of the recording situation that could impact recording quality.	**10 Capture Quality** *Why* For dataset creators, documenting quality issues can help inform decisions about preprocessing. For data statement readers, accurate descriptions of the recording quality are important for at least two reasons: first, to assess if the dataset would be well-matched for a particular intended use case (e.g., a corpus of collected speech may have word level transcription, but may not include disfluencies or mistakes made in the speech); and second, to enable future third party technology developers or adopters to make similar assessments of match to quality needs at a future time. *What* A description of quality issues in data capture should be provided. This includes all types of quality issues that arise across a broad range of collection methodologies for capturing an otherwise impermanent event.	Elaboration after comparison with datasheets Element generalized in response to broader use by workshop participants Elaboration of motivation added after analysis of workshop results Elaboration of content of element added after analysis of workshop participants' data statements
⋮		

Summary

- NLP datasets come from people (speakers, annotators, curators)
- Those people aren't representative of the full populations our technology impacts
- This mismatch leads to potential real-world harms
- Practical suggestion: NLP data statements
- Anticipated results: Better science and more ethical practice

Who's job is this?

- **Speech/language tech researchers & developers:** build better systems, promote systems appropriately, educate the public

- **Procurers:** choose systems/training data that match use case, align task assigned to speech/language tech system with goals

- **Consumers:** understand speech/language tech system output as the result of pattern recognition, trained on some dataset somewhere

- **Members of the public:** learn about benefits and impacts of speech/language tech and advocate for appropriate policy

- **Policy makers:** consider impacts of pattern matching on progress towards equity, require disclosure of characteristics of training data

Case: Direct stakeholders whose varieties aren't well represented

- **Speech/language tech researchers & developers:** Map out underrepresented language varieties and direct effort appropriately; test approaches more broadly

- **Procurers:** Is this trained model likely to work for our clientele?

- **Consumers:** Is this trained model likely to work for me?

- **Members of the public:** Advocate for models trained on datasets that are responsive to the community of users

- **Policy makers:** Require automated systems to be *accessible* to speakers of all language varieties in the community

Case: Indirect stakeholders whose varieties aren't well represented

- **Speech/language tech researchers & developers:** Map out underrepresented language varieties and direct effort appropriately; test approaches more broadly

- **Procurers:** What information is this system going to expose and what is it going to miss?

- **Consumers:** Is this software being transparent about how well it can work and under what circumstances it works better/worse?

- **Members of the public:** Advocate for transparency regarding system performance across representative samples

- **Policy makers:** Require broad testing of systems and transparency regarding system confidence/failure modes

Data statements are not a panacea!

- Mitigation of the negative impacts of speech/language technology will require on-going work and engagement (and cost/benefit analysis)

- Data statements are intended as one practice among others that position us (in various roles) to anticipate & mitigate some negative impacts

- Probably won't help with e.g.:

  - impacts of gendering virtual agents

  - privacy concerns around classification of identity characteristics

- Can help with problems stemming from lack of representative data sets and possibly also 'automation bias' (Skitka et al 2000)

Reading questions

- What is metadata and how is it used?
- What exactly is ‘mentoring’ (Section 7.3) and what would that practically look like in the NLP world?

Reading questions

- Since using data statements in NLP research is quite obviously beneficial in many aspects, I wonder what prevents it from being standardized as regular practice when using datasets in NLP work. What might be some underlying costs of implementing data statements for all out datasets? A scenario that I could think of (but not sure if it's valid) is when researchers or developers overgeneralize the characteristics of a dataset, like claiming that it represents certain populations without providing enough proof. The process of looking for proof, however, may bring about privacy issues for those who contributed to the dataset. Is this something we should worry about when constructing accurate data statements for NLP?

Reading questions

- Even if you wanted to go back and ask those writers of those Tweets questions about themselves, I would think that would almost be a privacy issue to some people, as to them these random researchers have suddenly found their Tweet, are using it in their research really without their consent, and are now wanting to ask them personal questions that would possibly be revealed in the paper when it is complete. Even just the fact that you are putting 'The writers of these comments weren't approached' and yet are using their content anyway seems a little.... at least to me it would make me second-guess if I should even be using their data for a study in the first place.

- Furthermore, if it was left up to the researcher in these studies to then try and guess what the characteristics of the individuals they got their data from, I would think this could leave the potential for the bias of the researcher to creep in.

Reading questions

- Is the implementation of data statements primarily for the sake of transparency, or is there also an assumption that NLP researchers have historically not been thinking critically about the kinds of data they are using, and data statements cause researchers to actually start thinking about the limits of their systems and their data?

Reading questions

- So if the implementation of the use of data statements is a relatively new idea in the field of NLP, what has historically and currently been used as a way of understanding the datasets being used? How was what's in datasets being accounted for?

- What do researchers/practitioners who do not include data statements in their work have to gain from doing such a thing?

- Is the main reason that most people don't include data statements just the time commitment it takes to do it and the fact that people are lazy?

Reading questions

- Machine learning is often being modeled as a black box as you have control of the input, but you can't really understand the process in the middle. In this case, it is extremely important to make sure the input is as transparent as possible to ensure the final result can be trusted. So why don't more people do it? It is because fear of other people reproducing the same result and improve on it? (I thought this is what scientific research is all about, but companies may think differently.) Is is because of other might use it against them or reverse-engineer them like Google would never publish their search algorithm?

Reading questions

- Was surprised to hear that including data statements wasn't already an enforced standard! Seems like a no-brainer with how varied/biased data can be, especially large datasets containing internet-sourced data. I was curious as to what specific next steps would be taken by a researcher reviewing the data statement. Is the idea to simply gain context about the research conducted with this extra information, or are there specific steps they may would be able to take to further mitigate bias?

Reading questions

- Would these data statements have an impact on how companies which use large data sets market/describe their products? The example I'm thinking of is of course ChatGPT; we don't know what kind of data OpenAI is using to model ChatGPT, so how would the public perception change around it if we had a long form data statement for it? It seems like they're keeping their data sources secret to protect it from others who may copy it, creating similar products from the same data. Would a data statement allow 3rd parties to create a sort of ChatGPT clone, or would we simply be better informed about where the information is coming from, and still unable to replicate the model?

Reading questions

- Considering that companies like OpenAI have not been transparent about what is in their training data, how would one go about implementing data statements in NLP systems such as ChatGPT that have such enormous amounts of training data?

NLP/Compling in the news

- https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/