Ling/CSE 472: Introduction to Computational Linguistics

4/23
Data and Model Documentation

Overview

- Announcements:
  - Midterm course feedback
  - Assignments 3 & 4 posted
- Data statements (largely from NAACL 2019 presentation)
- Reading questions

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

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.

Next steps

- Develop data statements for more datasets (two in paper)
  - If you’re interested in trying this, I’m happy to help!
- Workshop/working group of best practices for developing data statements
  - ==> May 11-13 2020 (previously associated to LREC 2020)
- Tutorials promoting best practices
- Propose policy (both academic and legal) building on the emergent best practices

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

- We've come across "training systems" in different texts now, but what does it actually mean? Is it running training data and adjusting software accordingly as problems of processing arise? So you "train" the system by giving input and then manually adjusting the code/software?

- How are data statements different from normal methodology description? Is it through their specific purpose of mitigating bias that they differ?

- For implementation, how do the writers decide which data sets to use when writing the data statements? How would the writers determine the reliability/usability of the data sets?

- Would the larger corpora out there, that people presumably use to train NLP systems on, pose an issue for this kind of framework? Or are these already well-documented enough that most of the required information for a data statement would be available?

Reading questions

- By providing a data statement, a resulting systems that used a particular dataset and annotations can be associated to the data statement of the data, which then can be used to assess a possible bias that it might have. But how does one assess this bias? Is there an objective metric that can be used to determine the level of bias that a system has? Can one assume that if a system was constructed using a set of input data with wide range of dataset and annotation, it will be less biased than a system constructed with a set of data with a narrow range? What can a researcher do in order to assess this?

- What might be a good way to remove biases in existing models? Is it possible to “edit” the vector representations in any way to remove certain features? Or would adding examples that contradict the biases to the data set and retraining be of any use?

Reading questions

- Wow, it sounds like implementing data statements would increase researcher workload, but also be very useful in understanding, analyzing, and handling biases in training data. There are several scenarios in the reading, how does one know how much/little to imagine in creating a scenario that supports an argument in a paper, it is also itself not too biased as to ignore crucial information or put the argument in too positive of a light?

- For example, a lot of research use Wikipedia dump, which will not have (and should not have due to privacy issue) information about speech demographic. The same thing goes with datasets from social media, united nation and a lot of existing treebanks. A lot of paper use one of these kinds of datasets. Is it still necessary to include speaker demographic if we cannot get such information majority of the time?

Reading questions

- Why is it important to detail the quality of any recording(s) in a data statement? Is this primarily an indication of transcription quality, or does a recording's clarity have a different impact on its use that I'm not aware of?

- What would be the best method to determine what language varieties are included in a large data collection like a twitter scrape? In one of the blog posts this week it was mentioned that a model was used to decide what language variety an utterance belonged to. However, it seems like this introduces even more potential biases. Has there been any research on this?

Reading questions

- The paper also did not discuss details about privacy, only saying that "how to find an appropriate level of detail given privacy concerns". So by collecting such sensitive information from both speaker and annotator, how can we protect their privacy?

- Data statements seem like they can lead to better precision when used as regular practice in NLP research. One potentially limiting factor is that long form data statements are required to provide information on speaker demographic. Not all metadata is provided (Race/ethnicity, Socioeconomic status, etc.) and not all metadata is necessarily ethical to examine or collect.

Reading questions

- When does the small form description of the data become filler? For example, datasets that are commonly used within a field will receive little more than a footer description. Would the small form description start to look identical across papers using the same dataset?

- If the "long form" contains the necessary information what is the reason for including a "short form"? I thought the idea of data statements was to scrutinize the data to uncover any biases. If it's shortened to only cover most of the main points then it's not as thorough. Nevertheless, the long form is included whether or not there is a short form. Then, if the short form is just there for readability, why call it "short form" and not a summary of the long form?

Reading questions

- How easy would it be to have the ACL actually implement this policy? Is that a realistic goal or more of a long-shot?

- I saw that the paper has been published in 2018, and I wonder if there have been changes in recent NLP research papers? Have there been improvements in the inclusion of data statements? If so, to what degree did the data statements reduce the bias in such systems?

- What might be a rough timeline for a practice like data statements to become widespread and integrate into existing practices? Or, what are similar examples of practices that have (relatively recently) achieved widespread adoption within NLP literature?

Reading questions

- How should a discussion of the ethics of crowdworking platforms might be incorporated into the data statement? I remember reading Fort et al 2011 a year ago that was very critical of Amazon Mechanical Turk, both as a source of data and as an employer. In the first example it was acknowledged that some of the annotations were sourced from crowdworkers, but it didn’t articulate a position on that source. Is that kind of explicit concern (as opposed to context) something that belongs in the data statement? How do computational linguists generally approach the ethical issues associated with platforms like Amazon Mechanical Turk?

- https://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00057

Reading questions

- As a society, we are progressively pushing for more gender-neutral and language such as the standardization of singular 'they', the term 'latinx,' campaigning to stop using ableist words such as 'retarded' in casual settings, and other exclusive language. As we push for this more exclusive and progressive language, so would the corpora of language and datasets that NLP systems be able to train off of. How much would said NLP systems and researchers need to catch up in order to mitigate bias?

Reading questions

- Since data statements are characterizations of a dataset written by professionals (humans as well), does that mean there can be potential for data statements to also contain bias in them? And if so, how would we go about handling the additional bias formed?

- Even if data statements become more widely implemented in publications, how do we know that over time, they won't also be subject to biases and changes in society unforseen?