Ethics in NLP seminar recap

Emily M. Bender
Oct 10, 2017

UW Linguistics Treehouse

Paging Dr. Bender

Christopher Phipps
@lousylinguist

Following

Paging the Ethics in NLP crew @dirk_hovy @emilymbender

> **The Information** @theinformation
> Amazon considering giving developers acc
> when using Alexa apps. $AMZN theinforma

8:08 AM - 13 Jul 2017

Emily M. Bender @emilymbender · Jul 14

Replying to @lousylinguist @dirk_hovy

I'm totally happy to be associated with serious thought about #ethNLP, but something has been bugging me about being "paged" >>

1 2

Emily M. Bender @emilymbender · Jul 14

and I think I've figured it out. The other #ethNLP organizers and I aren't the ethics & #NLProc police, nor is it our job to keep >>

1 1

Emily M. Bender @emilymbender · Jul 14

everyone in line. These issues are for *everyone* to be aware of & work on. >>

1 3

Emily M. Bender @emilymbender · Jul 14

I realize that's probably not what you (@lousylinguist) meant, but I still wanted to say all this :)

2

Goals of this talk

- Synthesize what I learned from the course
- Bring more of you into the discussion
- Pointers to shorter readings

Four types of questions

- What problems have occurred/might occur in the future?
- What frameworks are available for analyzing those problems in ethical terms?
- What best practices are out there for mitigating such problems?
- How do we engage the field in these discussions?

Course syllabus

- http://faculty.washington.edu/ebender/2017_575/

Week 1: Framing and getting started

- Spruit & Hovy 2016 call for discussion of ethics in NLP, and introduce five key concepts:

- Exclusion (leaving groups of people out of the training data)

- Overgeneralization (false positives)

- Overexposure (feeds human biases)

- Underexposure (most resources focus on a few languages)

- Dual use (e.g. NLP used to detect fake reviews, or to generate them)

Week 1: Framing and getting started

- Obstacles to more ethical practice in NLP more broadly:
    - Maverick-y culture in CS (esp in start ups)
    - No single avenue for enforcement (no analogue to IRB)
        - Buy-in: “It’s just common sense”, “bottom line considerations don’t leave the time/resources”
- Suggested solution: Lead by example; broad, active discussion
- *If you read just one thing:* Sourour, B. (Nov 13, 2016). The code I'm still ashamed of. medium.com

Weeks 2 & 3: Philosophical underpinnings

- Two items from Philosophical Foundations below, at least one of which comes from an author whose perspective varies greatly from your own life experience. Be prepared to discuss the following:

    - What is the main thesis of the reading?
    - What is their definition of ethics?
        - In what ways do they contrast their definition with others?
        - How does this reading relate to ethics in NLP?

- => Various systems for thinking about what constitutes ethical behavior/systems

- *If you read just one thing:* Vallor, Shannon, "Social Networking and Ethics", The Stanford Encyclopedia of Philosophy (Winter 2016 Edition)

Week 4: Exclusion/Discrimination/Bias

- Angwin, J., & Larson, J. (Dec 30, 2016). Bias in criminal risk scores is mathematically inevitable, researchers say. ProPublica.
- boyd, d. (2015). What world are we building? (Everett C Parker Lecture. Washington, DC, October 20)
- Brennan, M. (2015). Can computers be racist? big data, inequality, and discrimination. (online; Ford Foundation)
- Clark, J. (Jun 23, 2016). Artificial intelligence has a `sea of dudes' problem. Bloomberg Technology.
- Crawford, K. (Apr 1, 2013). The hidden biases in big data. Harvard Business Review.
- Daumé III, H. (Nov 8, 2016). Bias in ML, and teaching AI. (Blog post, accessed 1/17/17)
- Emspak, J. (Dec 29, 2016). How a machine learns prejudice: Artificial intelligence picks up bias from human creators--not from hard, cold logic. Scientific American.
- Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. ACM Transactions on Information Systems (TOIS), 14(3), 330-347.
- Guynn, J. (Jun 10, 2016). `Three black teenagers' Google search sparks outrage. USA Today.
- Hardt, M. (Sep 26, 2014). How big data is unfair: Understanding sources of unfairness in data driven decision making. Medium.
- Jacob. (May 8, 2016). Deep learning racial bias: The avenue Q theory of ubiquitous racism. Medium.
- Larson, J., Angwin, J., & Parris Jr., T. (Oct 19, 2016). Breaking the black box: How machines learn to be racist. ProPublica.
- Morrison, L. (Jan 9, 2017). Speech analysis could now land you a promotion. BBC capital.
- Rao, D. (n.d.). Fairness in machine learning. (slides)
- Sweeney, L. (May 1, 2013). Discrimination in online ad delivery. Communications of the ACM, 56 (5), 44-54.
- Zliobaite, I. (2015). On the relation between accuracy and fairness in binary classification. CoRR, abs/1505.05723.

- *If you read just one thing:* Sweeney, L. (May 1, 2013). Discrimination in online ad delivery. Communications of the ACM, 56 (5), 44-54.

Week 5.1: Word Embeddings and Language Behavior as Ground Truth

- Much work in NLP assumes that large text collections are a good source from which to extract information about the world
- But text collections reflect biases, including what people choose to talk about but also biased perspectives
- Li Zilles: Problems in applying ‘descriptive models prescriptively’
- *If you read just one thing:* Bolukbasi, T., Chang, K., Zou, J. Y., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. CoRR, abs/1607.06520
- *If you read another:* Herbelot, A., Redecker, E. von, & Müller, J. (2012, April). Distributional techniques for philosophical enquiry. In Proceedings of the 6th workshop on language technology for cultural heritage, social sciences, and humanities (pp. 45-54). Avignon, France: Association for Computational Linguistics.
- *Another:* https://blog.conceptnet.io/2017/04/24/conceptnet-numberbatch-17-04-better-less-stereotyped-word-vectors/

Week 5.2: Chat Bots

- Fessler, Leah. (Feb 22, 2017). SIRI, DEFINE PATRIARCHY: We tested bots like Siri and Alexa to see who would stand up to sexual harassment. Quartz.
- Fung, P. (Dec 3, 2015). Can robots slay sexism? World Economic Forum.
- Mott, N. (Jun 8, 2016). Why you should think twice before spilling your guts to a chatbot. Passcode.
- Paolino, J. (Jan 4, 2017). Google home vs Alexa: Two simple user experience design gestures that delighted a female user. Medium.
- Seaman Cook, J. (Apr 8, 2016). From Siri to sexbots: Female AI reinforces a toxic desire for passive, agreeable and easily dominated women. Salon.
- Twitter. (Apr 7, 2016). Automation rules and best practices. (Web page, accessed 12/29/16)
- Yao, M. (n.d.). Can bots manipulate public opinion? (Web page, accessed 12/29/16)

- Issues with both privacy and reinforcing gender stereotypes

- *If you read just one thing*: Paolino, J. (Jan 4, 2017). Google home vs Alexa: Two simple user experience design gestures that delighted a female user. Medium.

Week 6: Proposed Code of Ethics for NLP

- From Hal Daumé III's blog: http://nlpers.blogspot.jp/2016/12/should-nlp-and-ml-communities-have-code.html

    - What is missing and why?
    - Which points shouldn't be there and why?
    - Should the ACL adopt a code of ethics of this general sort? Why or why not?
    - What are some cases that seem to contradict one or more points in the code that you think are nonetheless ethical?

Week 7: Value sensitive design

- Set of practices to identify and integrate values of stakeholders in the design process
- Better not best
- Both direct and indirect stakeholders
- Ex: Stakeholder interviews
- Ex: Design noir
- Ex: Envisioning cards
- *If you read just one thing:* Nathan, L. P., Klasnja, P. V., & Friedman, B. (2007). Value scenarios: a technique for envisioning systemic effects of new technologies. In CHI'07 extended abstracts on human factors in computing systems (pp. 2585-2590).

Week 8: Other best practices

- Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. CoRR, abs/1606.06565.
- Markham, A. (2012). Fabrication as ethical practice: Qualitative inquiry in ambiguous Internet contexts. Information, Communication & Society, 15(3), 334-353.
- Ratto, M. (2011). Critical making: Conceptual and material studies in technology and social life. The Information Society, 27 (4), 252-260.
- Russell, S., Dewey, D., & Tegmark, M. (2015). Research priorities for robust and benefcial artifcial intelligence. AI Magainze.
- Shilton, K., & Anderson, S. (2016). Blended, not bossy: Ethics roles, responsibilities and expertise in design. Interacting with Computers.
- Shilton, K., & Sayles, S. (2016). "We aren't all going to be on the same page about ethics": Ethical practices and challenges in research on digital and social media. In 2016 49th Hawaii international conference on system sciences (HICSS) (pp. 1909-1918).

Week 9: Privacy

- What is privacy and why is it valued?

- To what extent do we have a moral obligation to inform the public about how NLP + massive online presentation of self (and others) impinge on privacy?

- Michael Strube has a whole class on this: The Dark Side of NLP: Gefahren automatischer Sprachverarbeitung

- *If you read just one thing:* Solove, D. J. (2007). 'I've got nothing to hide' and other misunderstandings of privacy. San Diego Law Review, 44 (4), 745-772.

Week 10: NLP Applications Addressing Ethical Issues

- Fokkens, A. (2016). Reading between the lines. (Slides presented at Language Analysis Portal Launch event, University of Oslo, Sept 2016)
- Gershgorn, D. (Feb 27, 2017). NOT THERE YET: Alphabet's hate-fighting AI doesn't understand hate yet. Quartz.
- Google.com. (2017). The women missing from the silver screen and the technology used to find them. Blog post, accessed March 1, 2017.
- Greenberg, A. (2016). Inside Google'S Internet Justice League and Its AI-Powered War on Trolls. Wired.
- Kellion, L. (Mar 1, 2017) Facebook artificial intelligence spots suicidal users. BBC News.
- Munger, K. (2016). Tweetment effects on the tweeted: Experimentally reducing racist harassment. Political Behavior, 1-21.
- Munger, K. (Nov 17, 2016). This researcher programmed bots to fight racism on twitter. It worked. Washington Post.
- Murgia, M. (Feb 23, 2017). Google launches robo-tool to flag hate speech online. Financial Times.
- The times is partnering with jigsaw to expand comment capabilities. (Sep 20, 2016). The New York Times.

- Fake News Challenge
- Jigsaw Challenges
- Perspective (from Jigsaw)
    - But see: Hosseini, H, S. Kannan, B. Zhang and R. Poovendran. 2017. Deceiving Google's Perspective API Built for Detecting Toxic Comments. ArXiv.
- Textio See also:
    - CEO Kieran Snyder's posts on medium.com
    - Recording of Kieran Snyder's NLP Meetup talk from Aug 15, 2016

Week 10: NLP Applications Addressing Ethical Issues

- What was the social issue addressed?
- How well did it work/how could you carry out an evaluation if one wasn't done?
- Design noir: What could go wrong?

My next steps

- This talk
- Paper with Batya Friedman (in progress)
- Two lectures on ethics in Intro to CL course
- Think about how to bring ethics into CLMS curriculum more broadly

Your next steps?