Chapter 3

Ethics Intro

3.1 Why Discuss Ethics in NLP Now?

A lot (all?) of this material comes from UW and CMU courses in NLP ethics. This is not a full course in the subject, so much will be skipped.

See `http://demo.clab.cs.cmu.edu/ethical_nlp/` and `http://faculty.washington.edu/ebender/2017_575/index.html`.

We just discussed probability...this seems like a topic jump. It's important as we get into the actual details to consider the ethical implications of:

- What we create (power of the models and tasks to do good or harm)
- How we create it (data sets and other ways that biases are implicitly baked into models)
- Why we create it (who is funding the work? How will it be used after it's made?)

We are dealing with *human language* which means we are dealing with people.

Nice quote: "The common misconception is that language has to do with words and what they mean. It doesn't. It has to do with people and what they mean. " (Herbert H. Clark and Michael F. Schober, 1992)

Ethics is broadly 'what is good/right.' However, what is good or right? Squishy!!

3.1.1 Example from CMU overview

- Should we design a classifier to predict if a chicken is male or female, while in the egg?
  - + Lowers cost of hatching/raising chicks you don't want (females for eggs, males for meat)
  - + Destroying the eggs may be less painful than killing the unwanted chicks
  - - it's not the chicken's fault
  - ...
- Should we design a classifier to predict IQ of an adult? Of a child? Of a fetus?

- Positives? Negatives?
- Let's stick with adults for now.
  - Who benefits from this classifier?
  - Who can be harmed, even if the classifier is *never wrong?*
  - If the classifier is more accurate for e.g. white women than other groups, who is responsible for the failings? Developer? Manager/Reviewer? University/Company? Society?
- NLP increasingly used to make real-world decisions.
  - Credit-worthy
  - Recommendations (based on some kind of stereotypes)
  - self-driving car decisions (more vision than NLP but commands)
  - whether to grant parole
- Data collection and annotation is a big part of NLP. Human language necessarily means human collection
  - Unsupervised data: intentional publication (news)? Unintentional publication (text messages)? Semi-intentional publication (social media)? Monetized publication (copyright violation)?
  - Annotation: Is this considered human subjects research (HSR)? Can it be distressing to annotators?
  - Are providers of data being fairly compensated? (Mechanical Turk, scraping against copyright, etc.)
  - Are these limits altering the demographic distribution of the data, and what are the consequences with regard to model performance?
- Issues to consider throughout this course:
  - How can this be used?
  - How might this be used?
  - What are the consequences of this use? Who will be affected? Who won't be able to take advantage?
  - Who has interest (ownership or otherwise) of the data you will use?
- The answer is probably not 'abandon everything you are doing here.' And I don't have all the answers (and in many cases there is no answer). But it is important to be aware of these issues.