Socially Responsible NLP

As language technologies have become increasingly prevalent, there is a growing awareness that decisions we make about our data, methods, and tools are often tied up with their impact on people and societies. This tutorial will provide an overview of real-world applications of language technologies and the potential ethical implications associated with them. We will discuss philosophical foundations of ethical research along with state-of-the art techniques. Discussion topics include:

- **Philosophical foundations:** what is ethics, history, medical and psychological experiments, IRB and human subjects, ethical decision making.
- **Misrepresentation and bias:** algorithms to identify biases in models and data and adversarial approaches to debiasing.
- **Civility in communication:** monitoring explicit abusive language and implicit microaggression.

Through this tutorial, we intend to provide the NLP researcher with an overview of tools to ensure that the data, algorithms, and models that they build are socially responsible. These tools will include a checklist of common pitfalls that one should avoid (e.g., demographic bias in data collection), as well as methods to adequately mitigate these issues (e.g., adjusting sampling rates or debiasing through regularization).

The tutorial is based on a new course on Ethics and NLP (http://demo.clab.cs.cmu.edu/ethical_nlp/) developed at Carnegie Mellon University.

You can learn more about the tutorial content and outline at https://sites.google.com/view/srnlp.

Additional relevant courses in the intersection of Ethics and NLP:

- Emily Bender at Univ. of Washington: http://faculty.washington.edu/ebender/2017_575/
- Graham Hirst at Univ. of Toronto: http://www.cs.utoronto.ca/~gh/cscD03/index.shtml

Readings relevant to tutorial preparation:

- https://goo.gl/7hA9D

Proceedings of NAACL-HLT 2018: Tutorial Abstracts, pages 24–26
New Orleans, Louisiana, June 1, 2018. ©2018 Association for Computational Linguistics

Outline

Foundations

- Motivation
- Philosophical foundations
- History: medical, psychological experiments, IRB and human subjects

Bias and Misrepresentation in NLP

- Psychological foundations of implicit bias
- Quantifying stereotypes, prejudice, and discrimination
- Debiasing

Modeling Civility in Communication

- Hate speech
- Implicit negativity: condescension
- Respect and formality in police-community communications

Instructors

Yulia Tsvetkov, Carnegie Mellon University
ytsvetko@cs.cmu.edu
http://www.cs.cmu.edu/~ytsvetko/

Yulia Tsvetkov is an assistant professor in the Language Technologies Institute at Carnegie Mellon University. Her research interests lie at or near the intersection of natural language processing, machine learning, linguistics, and social science. Her current research projects focus on NLP for social good, including advancing language technologies for resource-poor languages spoken by millions of people, developing approaches to promote civility in communication (e.g., modeling gender bias in texts and debiasing), identifying strategies that undermine the democratic process (e.g., political framing and agenda-setting in digital media). Prior to joining CMU, Yulia was a postdoc in the Stanford NLP Group; she received her PhD from Carnegie Mellon University.

Vinodkumar Prabhakaran, Stanford University
vinod@cs.stanford.edu
www.cs.stanford.edu/~vinod

Vinodkumar Prabhakaran is currently a postdoctoral fellow at the Stanford NLP lab, and prior to this, received his PhD in Computer Science from Columbia University in 2015. In the fall, he will be starting as a research scientist at Google to work on issues around Ethics in AI and ML Fairness. His research falls in the interdisciplinary field of computational social sciences, with a focus on applying NLP for social good. He combines NLP techniques with social science methods in order to identify and address large scale societal issues, such as racial bias and disparities in law enforcement, manifestations of power and gender at workplace, and online incivility such as condescension and gender bias.

Rob Voigt, Stanford University
robvoigt@stanford.edu
https://nlp.stanford.edu/robvoigt/

Rob Voigt is a PhD student in the Linguistics Department at Stanford University, working on topics in computational sociolinguistics with Dan Jurafsky. His research focuses on using computational methods to understand how social context and social factors subtly influence linguistic behavior at a large scale. His dissertation is focused on techniques for extracting and analyzing linguistic implicit bias, including respectfulness in police-community interaction, gender bias in online communications, and “othering” in historical media representations of immigrant groups.

Estimate of Audience Size

~50 people.

Description of Special Requirements

- A data projector
- A computer with PowerPoint and Acrobat Reader
- Poster boards and adhesive tape
- Tables, power sockets and Internet connection, in case presenters want to give demonstrations

Venue Preference

ACL > NAACL > EMNLP > COLING