

Upcoming Colloquiums

January 22 Po-Ling Loh UC Berkeley

January 29 Tamara Broderick UC Berkeley

February 5 Adel Javanmard Stanford University

March 26 Debashis Ghosh Penn State University

April 2 Donald Rubin Harvard University

May 7 Thomas Louis Johns Hopkins University

Department of Statistics Colloquium



Brendan O'Connor Machine Learning Department Carnegie Mellon University

Statistical Text Analysis for Social Science: Learning to Extract International Relations from the News

4:30 pm, Wednesday, January 15 F60 Jon M. Huntsman Hall

ABSTRACT

What can text analysis tell us about society? Corpora of news, books, and social media encode human beliefs and culture. But it is impossible for a researcher to read all of today's rapidly growing text archives. My research develops statistical text analysis methods that measure social phenomena from textual content, especially in news and social media data. For example: How do changes to public opinion appear in microblogs? What topics get censored in the Chinese Internet? What character archetypes recur in movie plots? How do geography and ethnicity affect the diffusion of new language?

In this talk I will focus on a project that analyzes events in international politics. Political scientists are interested in studying international relations through *event data*: time series records of who did what to whom, as described in news articles. To address this event extraction problem, we develop an unsupervised Bayesian model of semantic event classes, which learns the verbs and textual descriptions that correspond to types of diplomatic and military interactions between countries. The model uses dynamic logistic normal priors to drive the learning of semantic classes; but unlike a topic model, it leverages deeper linguistic analysis of syntactic argument structure. Using a corpus of several million news articles over 15 years, we quantitatively evaluate how well its event types match ones defined by experts in previous work, and how well its inferences about countries correspond to real-world conflict. The method also supports exploratory analysis; for example, of the recent history of Israeli-Palestinian relations.

BIO: Brendan O'Connor (<u>http://brenocon.com/</u>) is a 5th year Ph.D. candidate in Carnegie Mellon University's Machine Learning Department. He is interested in statistical machine learning and natural language processing, especially when informed by or applied to the social sciences. In the past he has interned in the Facebook Data Science group, and worked on crowdsourcing (Crowdflower / Dolores Labs) and "semantic" search (Powerset). His undergraduate degree was Symbolic Systems.

Refreshments will be served at 4:00 pm in The Stat Lounge (4th floor of JMHH).

Check out our website for details about upcoming seminars: https://statistics.wharton.upenn.edu/research/seminars-conferences/.