MIDAS MICHIGAN INSTITUTE FOR DATA SCIENCE UNIVERSITY OF MICHIGAN

Seminar Series

Friday, October 26, 2018 | **4:00pm** | **West Hall, Room 340**

The Promise of Crowdsourcing for Natural Language Processing and Other Data Sciences

Chris Callison-Burch, PhD

Associate Professor of Computer and Information Science
University of Pennsylvania

Abstract: Hidden Mark Crowdsourcing is a new tool for data scientists that allows us to collect data and annotations on a large scale and at low cost. This offers new possibilities for research in economics, linguistics and other social sciences, as well as for computer vision, natural language processing (NLP) and other machine learning applications.

Dr. Callison-Burch will discuss how he uses crowdsourcing to create speech and language data for NLP. He will detail a number of his own recent experiments using Amazon Mechanical Turk for NLP, including
* Building quality control models to achieve professional translation quality from non-professional translators
* Taking a census of the language skills of 4000 Turkers from more than 100 countries
* Collecting sufficient volumes of data to train statistical translation models that beat the state of the art translation systems.
He will also present some of his preliminary studies into collecting political science data.

Dr. Callison-Burch will also discuss the general challenges of crowdsourcing, including quality control, conveying complex tasks to lay users, professional v. non-professional annotation, and the advantages, including scalability and access to a worldwide workforce with diverse language skills. Based on his own experience, he will attempt to give general guidance about when crowdsourcing works and when it doesn't, and how to customize your annotation schemes to be more appropriate to the Mechanical Turk crowdsourcing platform.

Bio: Chris Callison-Burch is an associate professor of Computer and Information Science at the University of Pennsylvania. Before joining Penn, he was a research faculty member at the Center for Language and Speech Processing at Johns Hopkins University for 6 years. He served as the General Chair of the ACL 2017 conference, and the Program Co-Chair for the EMNLP 2015 conference. He has served as the Chair of the NAACL Executive Board, the Secretary-Treasurer for SIGDAT, and on the editorial boards of TACL and Computational Linguistics. He has more than 100 publications, which have been cited over 14,000 times. He is a Sloan Research Fellow, and he has received faculty research awards from Google, Microsoft, Amazon and Facebook in addition to funding from DARPA and the NSF. His research interests include natural language processing and crowdsourcing.

MIDAS gratefully acknowledges Wacker Chemie AG for supporting the MIDAS Seminar Series.

WACKER

For more information see: http://midas.umich.edu/events