# **Crowdsourcing for NLP**

Chris Callison-Burch, Lyle Ungar, and Ellie Pavlick Computer and Information Science Department University of Pennsylvania

## 1 Introduction

Crowdsourced applications to scientific problems is a hot research area, with over 10,000 publications in the past five years. Platforms such as Amazons Mechanical Turk and CrowdFlower provide researchers with easy access to large numbers of workers. The crowds vast supply of inexpensive, intelligent labor allows people to attack problems that were previously impractical and gives potential for detailed scientific inquiry of social, psychological, economic, and linguistic phenomena via massive sample sizes of human annotated data. We introduce crowdsourcing and describe how it is being used in both industry and academia. Crowdsourcing is valuable to computational linguists both (a) as a source of labeled training data for use in machine learning and (b) as a means of collecting computational social science data that link language use to underlying beliefs and behavior. We present case studies for both categories: (a) collecting labeled data for use in natural language processing tasks such as word sense disambiguation and machine translation and (b) collecting experimental data in the context of psychology; e.g. finding how word use varies with age, sex, personality, health, and happiness.

We will also cover tools and techniques for crowdsourcing. Effectively collecting crowdsourced data requires careful attention to the collection process, through selection of appropriately qualified workers, giving clear instructions that are understandable to non-?experts, and performing quality control on the results to eliminate spammers who complete tasks randomly or carelessly in order to collect the small financial reward. We will introduce different crowdsourcing platforms, review privacy and institutional review board issues, and provide rules of thumb for cost and time estimates. Crowdsourced data also has a particular structure that raises issues in statistical analysis; we describe some of the key methods to address these issues.

No prior exposure to the area is required.

#### 2 Presenters' Background

Dr. Chris Callison-Burch is the Aravind K Joshi term assistant professor in the Computer and Information Science Department 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 was the Chair of the Executive Board of the North American chapter of the Association for Computational Linguistics (NAACL) from 2011?2013, and he has served on the editorial boards of the journals Transactions of the ACL (TACL) and Computational Linguistics. He is a Sloan Research Fellow, and he has received faculty research awards from Google, Microsoft and Facebook in addition to funding from DARPA and the NSF. Chris teaches a semseter-long course on Crowdsourcing at Penn (http://crowdsourcing?class.org/)

Dr. Lyle Ungar is a Professor of Computer and Information Science at the University of Pennsylvania. He also holds appointments in several other departments in the Engineering, Medicine, and Business Schools. Dr. Ungar received a B.S. from Stanford University and a Ph.D. from M.I.T. He has published over 200 articles and is co-inventor on eight patents. His current research includes machine learning, data mining, and text mining, and uses social media to better understand the drivers of physical and mental well-being. Lyles research group collects MTurk crowdsourced labels on natural language data such Facebook posts and tweets, which they use for a variety of NLP and psychology studies. Lyle (with collaborators) has given highly successful tutorials on information extraction, sentiment analysis, and spectral methods for NLP at conferences including NAACL, KDD, SIGIR, ICWSM, CIKM, and AAAI. He and his student gave a tutorial on crowdsourcing last year at the Joint Statistical Meetings (JSM)

Ellie Pavlick is a Ph.D. student at the University of Pennsylvania. Ellie received her B.A. in economics from the Johns Hopkins University, where she began working with Dr. Chris Callison-?Burch on using crowdsourcing to create low? cost training data for statistical machine translation by hiring nonprofessional translators and post-editors. Her current research interests include entailment and paraphrase recognition, for which she has looked at using MTurk to provide more difficult linguistic annotations such as discriminating between fine-grained lexical entailment relations and identifying missing lexical triggers in FrameNet. Ellie TAed and helped design the curriculum for the Crowdsourcing and Human Computation course at Penn.

## 3 Learning Objectives

Participants will learn to:

- identify where crowdsourcing is and is not useful
- use best practices to design MTurk applications for creating training sets and for conducting natural language experiments
- analyze data collected using MTurk and similar sources
- · critically read research that uses crowdsourcing

## 4 Topics

• Taxonomy of crowdsourcing and human computation. Categorization system: motivation, quality control, aggregation, human skill, process flow. Overview of uses of crowdsourcing

- The Mechanical Turk crowdsourcing platform. Terminology and mechanics: Turkers, Requesters, HITs, micropayments. Demographics and motivation of Mechanical Turk workers.
- The human computation process. Design of experiments, selection of software, cost estimation, privacy/IRB considerations.
- Designing HITs. Writing clear instructions, using qualifications, pricing HITs, approving/rejecting work.
- Quality control. Agreement-?based methods, embedded quality control questions, applying the EM algorithm to find the correct label. When to invest extra funds in quality control versus when to collect more singularly labeled data
- Statistical analysis of MTurk results. Accounting for the block structure and non random sampling of the data
- Case Studies in NLP. Word sense disambiguation, machine translation, information extraction, computational social science