Workshops Held at the First AAAI Conference on Human Computation and Crowdsourcing: A Report

Tatiana Josephy, Matt Lease, Praveen Paritosh, Markus Krause, Mihai Georgescu, Michael Tjalve, Daniela Braga

■ *The first AAAI Conference on Human Computation and Crowdsourcing (HCOMP-2013) was held November 6–9, 2013, in Palm Springs, California. Three workshops took place on Saturday, November 9: Crowdsourcing at Scale (full day), Disco: Human and Machine Learning in Games (full day), and Scaling Speech, Language Understanding, and Dialogue through Crowdsourcing (half day). This report summarizes the activities of those three events.*

The first AAAI Conference on Human Computation and Crowdsourcing (HCOMP-2013) was held November 6–9, 2013, in Palm Springs, California. Three workshops took place on Saturday, November 9: Crowdsourcing at Scale, Disco: Human and Machine Learning in Games, and Scaling Speech, Language Understanding, and Dialogue through Crowdsourcing.

The goal of the Crowdsourcing at Scale workshop was to identify the biggest challenges in very large-scale crowdsourcing and to reduce the chasm between industrial and academic crowdsourcing research.

The aim of the Disco: Human and Machine Learning in Games workshop was to extend upon the focus of two past workshops and explore the intersection of entertainment, learning and human computation. The goal of the workshop was to examine both human learning and machine learning in games and human computation. Human computation methods let machines learn from humans where games can provide humans the opportunity to learn. The workshop was thus devoted to I learn, in Latin *disco,* for machines and humans alike.

The First AAAI Conference on Human Computation and Crowdsourcing Was Held in the Southern California Desert Community of Palm Springs.

The goal of the Scaling Speech, Language Understanding and Dialogue through Crowdsourcing workshop was to bring together industry and academe in a discussion and share experiences and findings. Another goal was to create more awareness in the speech and language community to crowdsourcing as a powerful and effective approach to scale speech and language technology.

Crowdsourcing at Scale

The goal of the first Crowdsourcing at Scale workshop was to identify the biggest challenges in very large-scale crowdsourcing and to reduce the chasm between industrial and academic crowdsourcing research by sharing problems and data sets to further crowdsourcing research. The workshop brought together 31 attendees from industry and academia, with a strong representation from universities (42 percent of the attendees), the largest crowdsourcing customers like Google, Microsoft, and Xerox (39 percent of the attendees), and the leading crowdsourcing platforms like Amazon's Mechanical Turk, CrowdFlower, Galaxyzoo, reCAPTCHA, and Mobileworks (19 percent of the attendees).

There were three main parts to the day: keynotes from industry experts, presentations from participants of the shared task challenge using data sets released by Google and CrowdFlower, and collaborative brainstorming based on submitted position papers.

Our call for papers asked for brief position papers identifying and clearly articulating problems, even if there aren't satisfactory solutions proposed. We encouraged submissions to clearly describe a problem of scale, why it matters, why it is hard, existing approaches, and desired properties of effective solutions. Twelve position papers were submitted, written by 32 authors. The purpose was to generate a seed set of ideas for collective discussion and brainstorming, that is, to crowdsource the process of generating a research agenda!

Introductions and brief presentations on position papers allowed the attendees to organize into four focused groups in the following areas: (1) training contributors for complex tasks, (2) improving the contributor experience, (3) combining human and machine computation effectively, and (4) defining and detecting experts. Each group independently brainstormed and discussed the area of focus over the breakout session and lunch, and generated a list of the most pressing problems they faced that they shared with the groups. We collected the questions using Google Moderator, and all the attendees collectively voted on problems that they thought were

important. This generated a list of 49 problems with 181 votes from the workshop attendees. The top questions included How can we systematically debug, revise, and improve task guidelines for a diverse set of tasks? and What instrumentation for collecting behavioral data from workers might help improve worker training and learning outcomes? We will be publishing a longer report with the results of the workshop as a useful resource for researchers and funding agencies.

The keynotes during the day were delivered by Anand Kulkarni from Mobile Works and Rajesh Patel from Microsoft. Kulkarni presented a radical method of managing tasks at very large scale with cascading layers of online management. In this system, the best workers in the system interact with clients and then manage others in the crowd who complete work. They pose the crowdsourcing problem as that of designing a large virtual, collaborative, and trainable workforce with a diverse set of skills. Patel engaged the group in a thought experiment: that we move away from the word *crowdsourcing*, which implies a big, anonymous crowd, toward a targeted, intelligently routed workforce.

Studying and solving problems at scale requires large human judgment data sets. Toward this end, we organized a shared task challenge based on two massive data sets of human judgments released by Google and CrowdFlower. The data sets totaled almost 1 million human judgments from fact evaluation and sentiment analysis tasks. The challenge was to optimize the aggregation of crowd labels for the data sets, with a $1500 prize from Google for the winning algorithms. The participants took very different approaches and showed substantial improvements over the baseline majority by using other signals and doing domain-specific empirical tuning. The winning entries were tied with no significant differences across the UCI/MIT team (Q. Liu, J. Peng, and A. Ihler, 2013, Report of Crowdscale Shared Task Challenge 2013), MSR/Southampton team (M. Venanzi, J. Guiver, G. Kazai, and P. Kohli, 2013, Bayesian Combination of Crowd-Based Tweet Sentiment Analysis Judgments) and University of Kalyani (M. Bhattacharyya, 2013, Opinion Ensembling: Learning from Dependent Judgements of the Crowd, CrowdScale 2013). Team MSR's use of signals from text was noteworthy for improving the quality of human computation for natural language related tasks. Team UCI and Kalayani's approach showed the value of empirically tuning how judgments were modeled and scored. The shared task challenge and the collaborative discussion made it clear and obvious to the participants that sharing data sets, problems, and pain points can spur valuable research.

Tatiana Josephy, Matt Lease, and Praveen Paritosh served as the organizers of this workshop. No technical report was published.

Disco: Human and Machine Learning in Games

With the Internet, the way we think about communication, computation, artificial intelligence, and research is changing. Human computation has emerged as a powerful approach to solve problems intractable without humans in the loop. Within human computation, games are a successful approach to incite people to collaborate in human computation. Games are also for human means to learn.

Digital games are interaction machines and, however implicit it might be, always contain a learning component. Whether one is stacking blocks, exploring dungeons, or building cities, games provide a variety of human-machine interactions in complex problem spaces. The challenges that emerge through these mechanics are how to foster human learning during the course of a game and leveraging this newly gained knowledge to solve the underlying problem by observing, and automatically learning from, the interactions between players and machines.

The Disco workshop brought together researchers and practitioners from a variety of fields such as AI, data analysis, economics, and game design. This workshop served to continue the progress made at two earlier workshops: The First International Workshop on Systems with Homo Ludens in the Loop, held at the 11th International Conference on Entertainment Computing (ICEC), and the First International Workshop on Human Computation in Digital Entertainment, held at the 2012 AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE). These earlier workshops addressed serious games and games with a purpose. The AIIDE workshop had an interdisciplinary focus on artificial intelligence and digital media; the ICEC conference focuses on digital entertainment.

The workshop had two distinct sessions. In the first session, authors from industry and academe presented their original work. The acceptance rate for paper presentations was 60 percent. The overall theme of the papers was to explore, understand, and utilize complex player strategies and learn from their behavior.

The second session was a group activity. Participants presented one of their current research projects and stated a particular complex challenge of this project. In the spirit of "practice what you preach," the participants, in fact being an expert crowd, then discussed this challenge to propose possible solutions. A recurrent research topic was natural language understanding from different perspectives, for example, machine translation, teaching a new language to human learners, and finding synergetic effects of these two challenges. A controversial question was how to evaluate learning success of human learners with adaptive systems that learn from the human learner. Another topic was the integration of lan-

guage learning into games. Especially giving reliable real-time feedback to players was an extensively discussed question. The feedback to this collaborative workshop style was very positive. The multitude of fascinating ideas developed during this session also highlight the validity of the approach.

The workshop was organized by Markus Krause, François Bry, and Mihai Georgescu. The papers of the workshop are part of the AAAI Press Technical Report WS-13-24.

Scaling Speech, Language Understanding, and Dialogue through Crowdsourcing

Crowdsourcing has rapidly grown and expanded across different scientific areas as one of the most successful strategies to scaling businesses and processes in a rapid and cost effective way. Since 2005 (when Amazon launched its microtask crowdsourcing platform, Mechanical Turk) the community of researchers in computer science, linguistics, speech technology, and so on have been changing their data collection, data labeling/annotation, data analysis, and user studies paradigms. Through crowdsourcing, research can scale and allow more robust models, better results, more accurate assessments and conclusions, and more sophisticated user studies. This workshop brought to light some of the most recent research and findings in the use of the crowdsourcing strategies for speech, language understanding, and dialogue tasks and to kick off a discussion on what works and what doesn't.

The workshop brought together crowdsourcing researchers and enthusiasts from both industry and academe. The workshop included three invited talks. Gina-Anne Levow (University of Washington) gave a talk on the predictive aspect of crowdsourcing spoken dialog systems evaluation. A talk given by Jeanne Parson (Microsoft) focused on how to use the crowd to identify the best human text-to-speech (TTS) voice talent. Nancy Chang (Google) gave a talk on the challenges around crowdsourcing semantic data annotation. The workshop also included three presentations of peer-reviewed papers on research on alternatives to Mechanical Turk for crowdsourcing (Haofeng Zhou, Denys Baskov, Matthew Lease), using the crowd for gathering task-oriented dialog learning data (Walter S. Lasecki, Ece Kamar, Dan Bohus, Eric Horvitz), and the impact of context in crowdsourced annotation (Elnaz Nouri). The workshop concluded with a panel discussion on the topic Does crowdsourcing really work?—successful and less successful stories, challenges, and learnings with Dan Bikel (Google Research), Michael Tjalve (Microsoft/University of Washington), Daniela Braga (VoiceBox Technologies), Ece Kamar (Microsoft Research), Annika Hämäläinen (Microsoft).

The workshop participants presented and discussed different approaches to make the best use of crowdsourcing for data quality and validation, for data collection and annotation, and for extracting semantic information. The discussion highlighted the need for industry and academia working together on the future of crowdsourcing and the desire to have an open-source, free crowdsourcing platform where researchers can build on top of shared human intelligence task (HIT) templates and antispam shared HIT templates and antispam rules, automation pipelines, and best practices, for example, by extending an existing HIT app. Several presentations highlighted the importance of the design of crowdsourcing tasks and the strong correlation between the framing of the task and the outcome of the study. This includes the importance of creating tasks that are engaging and personally relevant. Approaches to dealing with the risk of bias were presented and the value vs. the complexity of open-ended tasks was also discussed.

The workshop was organized by Daniela Braga (VoiceBox), Michael Tjalve (Microsoft and University of Washington), Ece Kamar (Microsoft), Gina-Anne Levow (University of Washington), Daniel Bikel (Google), Maxine Eskenazi (Carnegie Mellon University), Jon Barker (University of Sheffeid), Nikko Ström (Amazon), and Christoph Draxler (Ludwig Maximilian University of Munich). The papers of the workshop were published as AAAI Press Technical Report WS-13-25.

Tatiana Josephy is vice president of product at CrowdFlower.

Matt Lease is an assistant professor at the University of Texas at Austin.

Praveen Paritosh is an engineer and researcher at Google.

Markus Krause is a research associate in the Department of Computer Science at the Leibniz University Hannover, Germany.

Mihai Georgescu is a research associate at the L3S Research Center at the Leibniz University Hannover, Germany.

Michael Tjalve is a senior program manager at Microsoft and an affiliate assistant professor at University of Washington.

Daniela Braga is a senior speech scientist at VoiceBox Technologies.