Crowdsourcing for Success: Motivations, Design, & Ethics

Matthew Lease
School of Information
The University of Texas at Austin
ml@ischool.utexas.edu

ABSTRACT

While use of Mechanical Turk (AMT) studies [2, 22] is perhaps the best known example of crowdsourcing in the NLP community, not only do other paid platforms exist with different capabilities [23], but successful volunteer-based crowdsourcing initiatives are also bountiful, ranging from games [25, 16, 8] to to citizen science or humanities [21, 9, 3, 15], to less common avenues such as enabling self-discovery [17] or mining additional value from existing activities [26]. Motivation [18], design [1], and localization [10] all provide value. This paper draws on earlier work with Omar Alonso [13].

1. MOTIVATION

Different people are motivated by many different incentives, which can operate independently or in combination. Design of appropriate incentives for the given task at hand is referred to as *incentive engineering*. To motivate people to perform work (and do it well), incentive mechanisms explored inlucde: pay, fun, prestige, socializing, altruism, barter, and learning. For example, with citizen science, scientists or non-profits distribute work to volunteers motivated by altruism, learning, and/or opportunities to socialize (e.g., galaxyzoo.org, ebird.org). Games with a Purpose (www.gwap.com), on the other hand, motivate work via opportunities for fun, socialization, and prestige. The re-Captcha project (recaptcha.net) re-purposes an existing, ongoing activity (solving captchas) to extract useful work as a by-product. von Ahn's most recent DuoLingo project (duolingo.com) incentivizes translation work by the opportunity to learn a foreign language.

Many AMT studies have found distribution of HITs completed across the workforce tends to follow a power-law distribution, with a few workers doing much of the work and many workers doing very little work [22]. It is not clear to what extent that is a natural product of online crowd work, of AMT in particular, of the type of tasks being posted, or due to other factors. This has led some requesters to view crowdsourcing not as utilizing a large crowd to perform work, but filtering a large crowd to find a few people to do the work. For statistical approaches to quality assurance, a resulting challenge is estimating quality of individual worker contributions for those workers who contribute few labels, representing a sparse data problem.

2. TASK DESIGN VS. AGGREGATION

As crowdsourcing is inherently a human-centric enterprise, attention (or inattention) to human factors will clearly impact the quality of data collected from the crowd. While poor quality of crowd data has been often blamed on lazy and/or ignorant workers, task instructions and/or interfaces poorly designed for non-experts may be equally culpable. Moreover, if task instructions are unclear, or a task interface poorly designed, the inevitable low quality of crowd data can at best be ameliorated by statistical methods. It is somewhat remarkable, given this, that machine learning approaches to quality assurance represent one of the most studied areas of research in crowdsourcing [12, 19].

Approaches to quality assurance based on human factors vs. statistical methods are largely complementary, allowing each to be independently studied in controlled experimentation. However, the quality of data collected from the crowd may vary significantly under different task designs. Consequently, investigation of statistical quality assurance algorithms should take into consideration the highly variable quality of crowd data to be encountered in practice. This impacts both comparative benchmarking of alternative techniques, as well as measuring robustness [19].

A very important aspect of any crowdsourcing activity involves asking the right questions. Intuitively, this step seems very easy and straightforward to implement but in practice it could produce undesirable results if it is not taken seriously. Humans design a task that needs to be completed by another human so a precondition for proper communication is the use of simple language to convey expectations.

A worker is part of multilingual and multicultural distributed workforce and they need to understand questions consistently. The requester of the task has to make sure that all workers have a shared, common understanding of the meaning of the question. What constitutes a good answer or good work should be communicated to the workers. Examples of good and bad responses are encouraged. There is no need to use specific terminology unless all workers are expected to be experts on a particular subject. A common mistake when setting up crowdsourcing experiments is to condense too many questions on a single task. At any given day, requesters compete for workers so a simple and precise task is a much better technique for attracting the right crowd. Partitioning a larger task into atomic micro-tasks that can be easily completed in little time is important.

Some interesting recent work has shown workers responses either predicted or generated by others workers and found this to be quite valuable [4, 5]. We have found annotator rationales valuable in crowdsourcing task design [14], beyond their original proposed value for dual supervision.

3. ETHICS

While optimizing worker behaviors and investigating technological opportunities and challenges is clearly important and valuable work, how might we best wrestle also with questions about what is ethical, legal, and sustainable economic practice in crowd work [20, 7]. For example, is it

truly preferable to pay people nothing, i.e., having workers play online games which generate valuable work products as an output (which the workers may not be aware of, and which may generate revenue which the workers do not benefit from) [11], than to pay them low wages and clearly communicate to workers they are performing work [7, 6]? The recent case of prisoners compelled to perform gold farming in online games provides another example of what is ostensibly a game developing significant revenue, and thereby leading to forced labor [24]. As the industry of crowd work grows, these pressures will continue to grow in force. How might we monitor and mitigate the corresponding growth of such practices with at-risk populations?

4. REFERENCES

[1] Harini Alagarai Sampath, Rajeev Rajeshuni, and Bipin Indurkhya. Cognitively inspired task design to improve user performance on crowdsourcing platforms. In *32nd annual ACM conference on Human factors in computing systems*, pages 3665–3674. ACM, 2014.

[2] Chris Callison-Burch and Mark Dredze. Creating speech and language data with amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 1–12, 2010.

[3] Tim Causer, Justin Tonra, and Valerie Wallace. Transcription maximized; expense minimized? crowdsourcing and editing The Collected Works of Jeremy Bentham. *Literary and Linguistic Computing*, 27(2):119–137, 2012.

[4] Nancy Chang, Praveen Paritosh, David Huynh, and Collin F Baker. Scaling semantic frame annotation. In *The 9th Linguistic Annotation Workshop held in conjuncion with NAACL 2015*, page 1, 2015.

[5] Ryan Drapeau, Lydia B Chilton, Jonathan Bragg, and Daniel S Weld. Microtalk: Using argumentation to improve crowdsourcing accuracy. In *Proceedings of the 4th AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2016. 10 pages.

[6] Alek Felstiner. Sweatshop or paper route?: Child labor laws and in-game work. In *Proceedings of the 1st Annual Conference on the Future of Distributed Work (CrowdConf)*, San Francisco, September 2010.

[7] Karën Fort, Gilles Adda, and K Bretonnel Cohen. Amazon mechanical turk: Gold mine or coal mine? *Computational Linguistics*, 37(2):413–420, 2011.

[8] Panagiotis G Ipeirotis and Evgeniy Gabrilovich. Quizz: targeted crowdsourcing with a billion (potential) users. In *23rd World Wide Web (WWW) Conference*, pages 143–154. ACM, 2014.

[9] Steve Kelling, Jeff Gerbracht, Daniel Fink, Carl Lagoze, Weng-Keen Wong, Jun Yu, Theodoros Damoulas, and Carla Gomes. A human/computer learning network to improve biodiversity conservation and research. *AI magazine*, 34(1):10, 2012.

[10] Shashank Khanna, Aishwarya Ratan, James Davis, and William Thies. Evaluating and improving the usability of mechanical turk for low-income workers in india. In *Proceedings of the first ACM symposium on computing for development*, page 12. ACM, 2010.

[11] Edith Law and L. von Ahn. Human computation. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 5(3):1–121, 2011.

[12] Matthew Lease. On Quality Control and Machine Learning in Crowdsourcing. In *3rd Human Computation Workshop*, pages 97–102, 2011.

[13] Matthew Lease and Omar Alonso. Crowdsourcing and human computation, introduction. *Encyclopedia of Social Network Analysis &Mining*, pages 304–315, 2014.

[14] Tyler McDonnell, Matthew Lease, Mucahid Kutlu, and Tamer Elsayed. Why Is That Relevant? Collecting Annotator Rationales for Relevance Judgments. In *4th AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2016. 10 pages.

[15] Atsuyuki Morishima, Sihem Amer-Yahia, and Senjuti Basu Roy. Crowd4u: An initiative for constructing an open academic crowdsourcing network. In *Second AAAI Conference on Human Computation and Crowdsourcing*, 2014.

[16] Massimo Poesio, Jon Chamberlain, Udo Kruschwitz, Livio Robaldo, and Luca Ducceschi. Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 3(1):3, 2013.

[17] Katharina Reinecke and Krzysztof Z Gajos. Labinthewild: Conducting large-scale online experiments with uncompensated samples. In *18th ACM Conference on Computer Supported Cooperative Work (CSCW)*, pages 1364–1378, 2015.

[18] Dana Rotman, Jenny Preece, Jen Hammock, Kezee Procita, Derek Hansen, Cynthia Parr, Darcy Lewis, and David Jacobs. Dynamic changes in motivation in collaborative citizen-science projects. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 217–226. ACM, 2012.

[19] Aashish Sheshadri and Matthew Lease. SQUARE: A Benchmark for Research on Computing Crowd Consensus. In *1st AAAI Conference on Human Computation (HCOMP)*, pages 156–164, 2013.

[20] M Silberman, Lilly Irani, and Joel Ross. Ethics and tactics of professional crowdwork. *XRDS: Crossroads, The ACM Magazine for Students*, 17(2):39–43, 2010.

[21] Robert Simpson, Kevin R Page, and David De Roure. Zooniverse: observing the world's largest citizen science platform. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 1049–1054. ACM, 2014.

[22] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 254–263. ACL, 2008.

[23] Donna Vakharia and Matthew Lease. Beyond mechanical turk: an analysis of paid crowd work platforms. *Proceedings of the iConference*, 2015.

[24] Danny Vincent. China used prisoners in lucrative internet gaming work. *The Guardian*, May 25, 2011.

[25] Luis Von Ahn and Laura Dabbish. Designing games with a purpose. *ACM Comm.*, 51(8):58–67, 2008.

[26] Luis Von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum. recaptcha: Human-based character recognition via web security measures. *Science*, 321(5895):1465–1468, 2008.