

Towards Large-Scale Processing of Simple Tasks with Mechanical Turk

Paul Wais, Shivaram Lingamneni, Duncan Cook, Jason Fennell,
Benjamin Goldenberg, Daniel Lubarov, David Marin, and Hari Simons *

Yelp, Inc.

706 Mission Street

San Francisco, CA 94103

{pwais, shivaram, duncan, jfennell, benjamin, daniel, dave}@yelp.com hari.khalsa@gmail.com

Abstract

Crowdsourcing platforms such as Amazon’s Mechanical Turk (AMT) provide inexpensive and scalable workforces for processing simple online tasks. Unfortunately, workers participating in crowdsourcing tend to supply inconsistent quality. We report on our experiences using AMT to verify hundreds of thousands of local business listings for the online directory Yelp.com. Using expert-verified changes, we evaluate the accuracy of our workforce and present the results of preliminary experiments that work towards filtering low-quality workers and correcting for worker bias. We seek to inform the community of practical and financial constraints that are critical to understanding the problem of quality control in crowdsourcing systems.

Introduction

Crowdsourcing services such as Amazon Mechanical Turk¹ (AMT) provide useful platforms for processing simple web-based tasks. Several studies have reported that many workers contribute biased or even malicious answers. In this paper, we focus on strategies that redundantly assign several workers to each task in order to recover high-accuracy answers. We endeavor to design a system that can process a *large number* of tasks with high accuracy subject to reasonable financial constraints. We present the results of an AMT-based system that has processed hundreds of thousands of simple business listing verification tasks (see Table 1). We discuss an approach for actively filtering low-quality workers as well as techniques that correct for biased answers.

Studies have demonstrated that worker effort is sensitive to the balance of task difficulty and pay (Horton and Chilton 2010) as well as the number of tasks that must be completed before payment is made (Mason and Watts 2009). Though attention to workers’ financial incentives is crucial (especially for large-scale studies such as ours), our study’s primary focus is on recovering highly accurate answers from our workforce at a fixed pay rate. We paid workers upon completion of 2-4 small business listing verification tasks, and workers in our study earned between USD\$6/hr

to USD\$9/hr. Our pay rate and pay frequency are similar to those used in other AMT studies.

Several studies have proposed techniques that infer correct item labels from redundant worker-contributed labels of variable quality (Ipeirotis, Provost, and Wang 2010) (Snow et al. 2008) (Ghosh, Kale, and McAfee 2011). Each of these approaches processes labels *offline* after work has been completed. Furthermore, the experimental results published in these studies show that these methods tend to require assigning at least 10 workers per individual task for 90% or better accuracy. The cost of such high redundancy is impractical for large-scale systems.

In order to make large-scale processing of our simple tasks financially feasible, we find that we can only afford to assign *at most three* workers per task. For about the same cost of assigning more than this many workers, we were able to hire a team of on-site experts capable of higher accuracy. In contrast to the previous studies discussed above, our system filters inaccurate workers *online* in order to reduce costs. To complement our online vetting process, we also wish to compensate for the biases of workers who process real tasks (that have no expert-verified labels). We presented a preliminary analysis of the accuracy we could recover through redundancy at a NIPS Workshop (Wais et al. 2010). In this paper, we present additional results and discussion of applying the methods of Ipeirotis et al (Ipeirotis, Provost, and Wang 2010) and Ghosh et al (Ghosh, Kale, and McAfee 2011) to our dataset.

Tasks and Observations

Our system supports five different kinds of business listing changes summarized in Table 1. Each task required the worker to accept or reject a proposed business listing change. For example, a Category Task might ask a worker to accept or reject the *Aquatic Parks* category for a *Seafood* restaurant. (Many workers indeed erroneously voted to accept this change). For each task type, we generated tens of thousands of *test* tasks from a limited pool of expert-verified listing changes. Due to the limited number of qualifying expert-verified changes available, we allowed up to 15 workers to complete each test task.

We observed a high variance in worker quality. Despite the simplicity of our tasks, the median worker accuracy for all but our Category Task was at or below 50%. We observed a median task completion time of about 2 minutes (modal

*The majority of the work presented here was completed while this author was at Yelp, Inc.

Copyright © 2011, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹<http://www.mturk.com>

time of 60 seconds) and a very high variance in completion times overall. We studied the relationships between worker accuracies and completion times as well as accuracies and locale but found no significant correlations. Finally, workers had to complete CAPTCHAs after every few tasks, and we observed that a relatively small number of workers subverted the CAPTCHA protection and submitted work to AMT that included incorrect CAPTCHA solutions.

Filtering and Correcting for Biased Workers

Before allowing a worker to complete test or real tasks, we required each worker to pass an AMT *Qualification Test* consisting of multiple choice questions similar in nature to our real tasks. Though we confirmed with workers that the test had reasonable difficulty, only 35.6% of 4,660 applicants passed the test. Next, we observed qualifying workers’ performances on test tasks and filtered low-performing workers *online* as tasks were completed. As a result, we vetted a final workforce of 79 workers who achieved at least 80% accuracy on test tasks. Unfortunately, this small workforce could not meet our throughput goals; however, this preliminary filtering step dramatically reduced the funds we spent on very poor work.

In Table 2, we present the results of applying four different methods to our dataset that leverage redundant labels in order to infer a task’s true label. For each method, we used worker labels from test tasks so that we could establish true accuracies and estimate a plausible upper bound on the performance of workers vetted to work on real tasks. The **Majority** method consists of choosing the answer with the simple majority vote among workers. The **Best Worker** method consists of choosing only the answer of the worker with the highest accuracy for the task type. The results of this method serve as a baseline since for real tasks we could not estimate true worker accuracy with perfect confidence. The **EM** approach is an application of the Expectation-Maximization-based method of Ipeirotis et al (Ipeirotis, Provost, and Wang 2010).² Finally, the **SR** approach is an application of the Spectral-Rating algorithm devised by Ghosh et al (Ghosh, Kale, and McAfee 2011) for a binary labeling problem similar to ours.³ The more complicated **EM** and **SR** methods perform well but do not outperform simpler methods. Due to space restrictions, we do not report the results of testing each method on only subsets of tasks (to simulate an online setting) or subsets of workers (to test limiting redundancy). Synthetic results for **EM** and **SR** published by their respective authors suggest that these methods may require greater redundancy to achieve higher accuracy. Nevertheless, the comparisons presented here underscore the impact of budget limitations on existing methods.

We examined the application of traditional machine learning techniques to our tasks and note the success of applying a simple Naïve Bayes classifier (Sahami et al. 1998)

²We used their open source implementation available at <http://code.google.com/p/get-another-label/> and ran their algorithm for 100 iterations.

³**SR** requires as input the identity of a worker with accuracy greater than 50%. We chose the worker with the greatest measured task accuracy.

to fulfill our Category Task. We trained a our classifier on the text of over 12 million user-contributed reviews of local businesses on Yelp.com and tested the classifier on 1,615 of the reviewed businesses used in our test Category Tasks. The classifier achieved an average per-category accuracy of 79.4% and outperformed our workforce as well as the methods listed in Table 2. We note that not all of our task types may be fulfilled using traditional machine learning methods.

Our experiences present exciting new directions for further research. Though task redundancy is necessary in order to recover high accuracy from crowdsourced labels, balancing redundancy and frugality is very challenging. In order to address this tradeoff, future work might study hybrid approaches that augment automated methods (e.g. classifiers) with human supervision. Furthermore, restructuring the task may help incite more worker involvement and improve accuracy rates.

Acknowledgments We are very grateful to Jennifer Wortman Vaughan and John Horton for very helpful discussions of this work.

Task Type	Real Tasks	Test Tasks	\$USD/Task
Phone #	28131	1548	0.05
Category	138222	3185	0.05
Hours	93336	5080	0.05
URL	88222	9834	0.025
Address	47962	3088	0.05

Table 1: Summary of tasks processed and pay rates.

Task Type	Majority	Best Worker	EM	SR
Phone #	73.3%	89.0%	78.1%	55.1%
Category	79.0%	66.9%	67.1%	56.8%
Hours	59.1%	76.0%	52.7%	66.3%
URL	92.7%	75.0%	91.6%	50.1%
Address	68.5%	77.7%	67.7%	63.5%

Table 2: Accuracies of methods that compensate for bias.

References

Ghosh, A.; Kale, S.; and McAfee, P. 2011. Who moderates the moderators? crowdsourcing abuse detection in user-generated content. In *ACM Electronic Commerce*.

Horton, J. J., and Chilton, L. B. 2010. The labor economics of paid crowdsourcing. In *ACM Electronic Commerce*.

Ipeirotis, P. G.; Provost, F.; and Wang, J. 2010. Quality management on amazon mechanical turk. In *ACM SIGKDD Workshop on Human Computation*, 64–67. New York, NY, USA: ACM.

Mason, W., and Watts, D. J. 2009. Financial incentives and the “performance of crowds”. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, 77–85. New York, NY, USA: ACM.

Sahami, M.; Dumais, S.; Heckerman, D.; and Horvitz, E. 1998. A bayesian approach to filtering junk e-mail. In *AAAI Workshop on Learning for Text Categorization*.

Snow, R.; O’Connor, B.; Jurafsky, D.; and Ng, A. Y. 2008. Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks. In *EMNLP*, 254–263.

Wais, P.; Lingamneni, S.; Cook, D.; Fennell, J.; Goldenberg, B.; Lubarov, D.; Marin, D.; and Simons, H. 2010. Towards building a High-Quality workforce with mechanical turk. In *NIPS Workshop on Computational Social Science and the Wisdom of Crowds*.