What is a Question? Crowdsourcing Tweet Categorization

Sharoda A. Paul, Lichan Hong, Ed H. Chi
Palo Alto Research Center
3333 Coyote Hill Rd, Palo Alto, CA 94304.
{spaul, hong}@parc.com, chi@acm.org

ABSTRACT
One major way in which Amazon Mechanical Turk has been used is in the human labeling (or coding) of data, such as the relevance of search results or quality of Wikipedia articles. Recently, we used Amazon Mechanical Turk for classifying or labeling Twitter updates as questions or not. We present the design of our study and the steps that we took to address the challenges we faced in using Mechanical Turk for this labeling task. We also present our findings and some lessons learnt about the utility and effectiveness of using micro-task markets for conducting large-scale studies involving human-intelligence tasks.

Author Keywords
Social Q&A, social search, Twitter, Mechanical Turk.

ACM Classification Keywords
H5.3. Group and organizational interfaces: Web-based interaction.

INTRODUCTION
Researchers are increasingly using Amazon Mechanical Turk to source micro-tasks that require human intelligence, such as categorization of text or labeling of images [1, 4]. However, there are various challenges to using Mechanical Turk that raise questions about whether it is a valid method for labeling and acquiring research data. Workers have been found to "spam" the system and hence controls must be designed carefully to obtain valid data [1]. Also, the diversity and unknown nature of the worker pool raises questions about the usefulness of data collected using this approach. We report our experiences with using Mechanical Turk to perform a large-scale text categorization task. We describe the study design and the challenges faced in collecting data as well as the results of our study.

The broad goal of our research was to characterize natural question-asking behavior on Twitter. With the rising popularity of social networking sites like Facebook and Twitter, people are turning to their social networks to fulfill their information needs. Recent studies have found that people are using their status messages on online social networking sites for conversation [2], often specifically for asking questions to their friends [3]. We were interested in a large-scale study of Twitter status updates (tweets) to understand the types of questions people are asking their followers on Twitter, naturally during their normal use of the service.

One of the main challenges for us was to identify tweets as questions due to several reasons:

- First, tweets have developed a unique convention for the use of language, which combined with their 140-character length limitation, can make them hard to understand.
- Second, tweets often contain little context, since they are targeted at friends who typically already know a lot of context about the user. Hence, from a third person perspective, it was hard to determine which tweets were questions asked by the Twitter user to their social network.
- Third, we were concerned about introducing our own bias regarding what is a question and what is not.
- Finally, the manual classification of thousands of tweets was time-consuming, and hence not scalable.

We hypothesized that Mechanical Turk could help us deal with these challenges. We crowd-sourced the classification of tweets as questions and tried to leverage the scalability and low-cost of using Mechanical Turk.

METHOD

Candidate tweets
The goal of our Mechanical Turk study was to identify tweets as questions and we performed several steps to achieve this goal. First, we used Twitter's API to randomly sample the public Twitter stream and collected about 1.2 million tweets. We next performed a sequence of screening on these tweets to arrive at a set of candidate tweets that we provided to Turkers. During our screening, we removed retweets, tweets that were directed to specific users (starting with @username), tweets containing URLs, tweets containing certain inappropriate words, and tweets that were non-English. We also removed tweets that did not contain a '?' as previous literature suggests that most questions asked on social networking sites [3] end with a '?'. To verify this, we examined 25,000 questions posted on the Q&A site, Yahoo! Answers and found that 100% of those questions contained at least one '?'. Thus, the inclusion of a '?' in the tweet suggests that it was likely to be a question asked by the Twitter user.

This screening process helped us boost the 'signal-to-noise' ratio so that we could reduce the number of tweets Turkers had to examine to find a reasonable number of questions.

At the end of the screening process, we obtained a set of candidate tweets for the Turkers to classify.

HIT design

Next, we designed a task (HIT) on Mechanical Turk in which we presented Turkers with the candidate tweets and asked them to classify those tweets. We designed an external HIT that was hosted on our own server. The HIT provided workers with instructions about the task and a link to our website that hosted the task. The website was designed to collect demographic information about Turkers and then present them with the following instructions: "*Please read each of the following tweets and tell us whether you think this tweet is a question posed by the Twitter user to his/her followers with the expectation of getting a response.*"

The instructions were followed by a list of 25 tweets that we asked the Turkers to classify as 'question', 'not a question', or 'not sure'. Each candidate tweet was rated by two Turkers. If both Turkers rated a tweet as a 'question', we then classified the tweet as a true question.. At the end of the task, workers were provided a code that they had to submit to Mechanical Turk to get paid. Each worker could only do our HIT once.

Quality Controls

We designed several controls to ensure the validity of the data collected. First, Turkers were required to be Twitter users; this helped ensure that they were familiar with the language of Twitter and hence could understand the tweets in order to classify them correctly. Turkers had to enter a valid Twitter user id, which was then verified with the Twitter service. Further, to deal with the problem of spam responses, we inserted some control tweets along with the candidate tweets. Control tweets were tweets that we deemed easy to understand and were obviously 'questions' or 'not questions'. Each HIT consisted of 25 tweets, 20 candidate tweets and 5 control tweets. We only included data from those Turkers who rated all control tweets correctly in our subsequent analysis.

Worker demographics

Along with recording the ratings for tweets, we collected some basic usage data about every Turker who visited our website. This data consisted of demographic information, such as age and gender, as well as information about Turkers' Twitter usage such as their Twitter username (which we checked to see if it was valid), how often and how long they had been using Twitter, and whether they had ever asked or answered questions on Twitter. We also recorded how much time it took each Turker to do our task.

Our participants ranged in age from 18 to 68 years and were 41% female. The largest proportion (31%) of our participants had been using Twitter for 1-2 years. Our participant pool was well-represented in terms of frequency of Twitter-use with almost equal proportion of participants reporting that they used Twitter once a week (23%), once every few days (25%), 1-2 times a day (25%), and several times a day (21%). The rest (5%) reported that they used Twitter "constantly". Participants had experienced Q&A behavior on Twitter; 57% of participants had asked a question on Twitter, 63% had answered a question, and 51% had done both.

RESULTS

Performance of the Turkers

We were interested in examining how scalable our method of classifying tweets was, and whether the crowdsourced labeling approach could help us identify tweets as questions.

We created 3700 HITs on Mechanical Turk in the period of Dec 2010 – Jan 2011. These HITs were published in batches at different times of the day and during different days of the week. According to Mechanical Turk, 40% (1497/3700) of HITs posted were completed and we approved 83% (1248/1497) of the completed HITs.

Of the Turkers who completed the task (1497), 439 (29%) rated all control tweets correctly. Thus, we received valid data from only 29% of Turkers who completed the task. These 439 Turkers rated 4140 tweets (where each tweet got two ratings). Of these tweets, 1351 (32%) tweets were rated as questions by both Turkers.

We paid workers 25 cents for each HIT that was approved. Since it took about 5 minutes to complete the HIT, the equivalent rate was about $3/hour. Initially we approved all HITs but soon realized that we were paying spammers too and hence modified our study so that we only paid those Turkers who rated at least 4 out of 5 control tweets correctly. Obtaining 1351 questions for our study cost a total of $312.00.

Time taken by Turkers

We were interested in how much time it took Turkers to complete our task, both as a measure of the difficulty of the task and to examine if completion time could be an indicator of spam responses [1]. We designed the HIT in Mechanical Turk to automatically expire after 10 minutes as we expected our task to take about 5-7 minutes to complete. Indeed, for Turkers who rated the control tweets correctly, the average time for completing the task was 4.3 minutes.

We expected spammers to quickly click through the task and hence their completion time would be much less than the average completion time. Therefore, we examined those Turkers who took less than 2 minutes to complete the task, which is about 17% of the total number of Turkers. In contrast, of the Turkers who rated control tweets correctly, 13% of Turkers took less than 2 minutes. This suggests that completion time alone may not be a good measure for identifying spam responses. This is an interesting finding, when compared to prior research that suggested that completion time might be a good indicator of spam [1].

Labeling Performance

Turkers rated 4140 tweets where each tweet was rated twice and each rating was one of the following: question (Q), not question (NQ), or not sure (NS). Table 1 provides a breakdown of the ratings. Turkers rated 32% (1351/4140) of the candidate tweets presented to them as questions.

Rating	No. of tweets	Rating	No. of tweets
Q-Q	1351	**Q-NQ**	1029
NQ-NQ	1152	**Q-NS**	256
NS-NS	66	**NQ-NS**	286

Table 1. Breakdown of Turker ratings.

Characterization of questions

In previous work, Morris et al. [3] conducted a survey of 624 Microsoft employees and interns to study the kinds of questions they were asking via their status messages on social networks. In that study, 249 self-reported questions from Facebook and Twitter were analyzed by type and topic. They found that the majority of the questions were of the type recommendations (29%) and opinions (22%) and pertained to the topics of technology (29%) and entertainment (17%). Participants said that they were uncomfortable asking questions about health, religion, politics, and dating since they were too personal.

We were interested in comparing our results with those found in their study. Therefore, for tweets labeled as 'question' by both turkers, we classified them by type and topic.

Type	%	Example
Rhetorical	42%	How can I love and respect someone who doesn't love and respect herself?
Factual knowledge	16%	In UK when you need to see a specialist do you need special forms or permission?
Poll	15%	Who watched Harry Potter last night?
Opinion	10%	How do you all feel about interracial dating? You feel it's ok to date outside your race?
Recommendation	7%	Any suggestions from fellow artists about getting sponsorship and funds to keep being able to pursue my music???
Invitation	5%	I really want to go to a Georgetown bball game. Anyone down to go with me?
Favor	2%	I am with out a ride to the school this morning can anyone please drive me?
Offer	1%	I just bought 10 pounds of potatoes. Would anyone like some free potatoes?
Social connection	0.1%	Recruiting 4 #internship (#Rome #Milan) for my team. Would you reference any new graduate for this role?

Table 2. Breakdown of questions by type.

Question types and topics

We manually coded the question tweets for type and topic using the coding scheme in Morris et al. [3]. Two of the researchers independently coded the tweets and then discussed their differences to reach consensus. We modified the coding scheme to accommodate new types and topics of questions that emerged during our coding. Tables 2 and 3 show the breakdown of questions by type and topic. We were unable to categorize 1% of questions by type and 5% by topic.

Our results differ significantly from those of Morris et al. [3]. We found that the most popular question type on Twitter is rhetorical questions (42%), followed by factual questions (16%), and polls (15%).

Also, examining questions by topic, we found that most often people asked questions about entertainment (32%). Also in contrast with Morris et al. [1], surprisingly, we found a significant amount of personal and health questions (11%). Indeed, we also found questions related to additional topics not reported by Morris et al. such as greetings, time, weather, and general knowledge. These results suggest natural Q&A behavior in a conversational social network

Topic	%	Example
Entertainment	32%	Which team is better raiders or steelers?
Personal & health	11%	Any idea how to lose weight fast?? In a healthy way. Other than exercise, healthy diet?? Any other method can help??
Technology	10%	Any good iPad app recommendations?
Ethics & philosophy	7%	If you were to die right now, how would you feel about your life?
Greetings	7%	What's everyone doing on this amazing warm day??
Restaurants & food	4%	Anyone have any restaurant recommendations for Medford/Ashland area?
Current events	4%	What's going on in the stock markets today?
Twitter	4%	Anyone know how to delete a twitter account?
Professional	4%	When's a good time to go to #law school?
Home	3%	Getting my daughter a ferret for her Christmas present! What color pattern should I get? Male or female?
Time	3%	It's Friday everyone!!! What's your plans for the weekend?
Places	2%	Palm Beach County residents: Where can I find salt water taffy locally?
Shopping	2%	Looking to buy a used #canon #430ex ii. Are you or anyone you know selling??
General knowledge	1%	In physics what does fusion and vaporisation mean?
Weather	1%	Would u rather have it be 80 degrees in November or have seasons?

Table 3. Breakdown of questions by topic.

like Twitter is quite different from self-reported Q&A behavior in surveys.

DISCUSSION

In future work we plan to discuss the implications of our findings for Q&A in social networks. Here we focus on what we learnt about using Mechanical Turk as a method for doing text classification tasks as well as general insights gained about collecting data using this platform.

Task difficulty

Surprisingly, we found that certain kinds of tasks, such as characterizing short snippets of text, are difficult for Turkers. Admittedly, this difficulty might be due to the lack of clarity in our instructions to the Turkers. One of the main challenges in our task design was to phrase the instructions for the task. Task instructions must be short since we do not expect Turkers to spend too much time or attention on reading instructions.

On the one hand, our task was easy to explain to Turkers – identify questions from among the tweets presented. On the other hand, it was a difficult-to-explain task as there are many nuances in identifying text as a question (some questions are rhetorical, some contain little context, and some are too short to understand). We left it to the Turkers to use their judgment and designed the 'not sure' category to allow for the fact that it might be hard for Turkers to make that judgment with certainty for all candidate tweets. For ill-defined or hard-to-define tasks, it is challenging for researchers to convey task expectations to workers. So careful thought must be given to phrasing instructions and HITs should be designed to accommodate uncertainty on the part of workers.

Designing Controls

Identifying spam responses was also a challenge and we found that designing good controls was important. We inserted a high number of control tweets (5 for every 20 tweets rated). Kittur et al. [1] found that designing verifiable questions as part of the task helps researchers identify spam. Consistent with this prior finding, our experience suggests that if the task does not have a verifiable answer, inserting verifiable control questions helps researchers verify the results.

Payment

Since Mechanical Turk does not have much support for requesters to identify spam responses, so the onus lies on requesters to carefully design their task so as to be able to identify spam responses. In our study, we initially approved all HITs but we realized that a significant percentage of Turkers were not rating the control tweets correctly and had very short task completion time (<1 minute), suggesting that they may be spammers.

Even though we were rejecting data from spammers by not including it in our analysis, we were still paying them. In order to reject work from spammers, we modified our task design such that Turkers who did not rate at least 4 out of 5 control Tweets correctly received a second chance to do the task. In the second attempt, they were presented with a different set of tweets and if they still failed to rate at least 4 out of 5 control tweets correctly, their work was rejected and we did not pay them. The burden lies with the researchers to design controls carefully to discourage spamming behavior. However, overall as a service, Mechanical Turk should have other quality controls to ensure the long-term viability of the platform.

Variability of the worker pool

One of the challenges we faced was to attract enough workers for our task in order to classify a large number of tweets. We wanted unique Turkers to work on our HIT and we wanted them to be Twitter users. This combination of eligibility constraints made it hard for us to find workers for our HIT. Finally, seasonal trends in worker availability made it a challenge to obtain quick results. Since we published our HIT during the holiday season (Dec-Jan) we saw a drop-off in Turker response during the holidays. Thus, researchers using this kind of platform for data collection must keep in mind the limitations of finding participants who have the required characteristics, as well as plan for seasonal variations in worker availability.

CONCLUSION

Based on our experiences with Mechanical Turk, we are interested in better understanding how we can leverage this platform for tasks that require human-processing of large data sets. Rating of text snippets, such as status updates generated by users of social networks, is a hard task for computers due to the need to understand a variety of syntactic and semantic idiosyncrasies contained in the text. However, the challenges associated with using an unknown and variable worker pool must be addressed before we can use platforms like Mechanical Turk for such research.

At the workshop, we hope to discuss the challenges faced in our study, as well as the utility of our method, in order to gain insight into how to effectively leverage micro-task markets for research involving human-intelligence tasks.

ACKNOWLEDGMENTS

We thank Kevin Canini for helping us design the Mechanical Turk study and for technical support.

REFERENCES

1. Kittur, A., Chi, E.H., and Suh, B. Crowdsourcing user studies with Mechanical Turk. In *Proc. CHI 2008*, ACM Press (2008), 453-456.

2. Honeycutt, C. and Herring, S. Beyond microblogging: Conversation and collaboration via Twitter. In *Proc. HICSS 2009*. ACM Press (2009), 1-10.

3. Morris, M.R., Teevan, J., and Panovich, K. What do people ask their social networks and why? A survey study of status message Q&A behavior. In *Proc. CHI 2010*. ACM Press (2010), 1739-1748.

4. Heer, J. and Bostock, M. Crowdsourcing graphical perception: Using Mechanical Turk to assess visualization design. In *Proc. CHI 2010*. ACM Press (2010). 203-212.

Biographies of workshop attendees

Sharoda A. Paul
Sharoda Paul is a Computing Innovation Fellow in the Augmented Social Cognition group at the Palo Alto Research Center. She received her Ph.D. in Information Sciences and Technology from Penn State University. Her dissertation focused on collaborative information seeking and sensemaking in two domains: healthcare and Web search. Her research interests are in collaborative and social search, social computing, micro-task markets, and healthcare informatics.

Ed Chi
Ed Chi is Area Manager and Principal Scientist in the Augmented Social Cognition group at the Palo Alto Research Center. He leads the group in understanding how Web 2.0 and social computing systems help groups of people to remember, think, and reason. Ed has published widely in the fields of human-computer interaction, social computing, social networks, intelligent interfaces, and information visualization. Ed has a Ph.D. in Computer and Information Sciences from the University of Minnesota.