Crowdsourcing Breakthroughs for Language Technology Applications
AAAI Technical Report WS-15-24

Job Complexity and User Attention in Crowdsourcing Microtasks

Spencer Rothwell, Steele Carter, Ahmad Elshenawy, Daniela Braga
VoiceBox Technologies, Bellevue, WA, USA
spencerr, steelec, ahmade, danielab, {@voicebox.com}

Abstract

This paper examines the importance of presenting simple, intuitive tasks when conducting microtasking on crowdsourcing platforms. Most crowdsourcing platforms allow the maker of a task to present any length of instructions to crowd workers who participate in their tasks. Our experiments show, however, most workers who participate in crowdsourcing microtasks do not read the instructions, even when they are very brief. To facilitate success in microtask design, we highlight the importance of making simple, easy to grasp tasks that do not rely on instructions for explanation.

1 Introduction

Crowdsourcing has garnered a lot of attention from academics in recent years. Since 2008, there has been a marked increase in the number of NLP conference papers that use crowdsourcing to achieve various NLP goals (Sabout et. al. 2012), including translation, tagging, transcription, and more. As research into crowdsourcing matures, researchers are finding that crowdsourcing is a viable and affordable way to accomplish what were previously difficult and expensive tasks (Zaidan and Callison-Burch 2011). However it is not without its shortcomings. A valid and frequently discussed issue when discussing crowdsourcing is that of worker performance and data quality, and the methods of mitigating said issues [(Ipeirtois et. al. 2010), (Eickhoff and De Vries, 2011), (Genarro et. al. 2010), (Rutherford et. al. 2014)].

Quality concerns generally arise from the presence of illegitimate responses from workers, often in the form of spam from people attempting to game the system and earn as much money as is possible from crowdsourcing microtasks. Thus, quality control must be properly addressed by any researcher seeking to properly utilize crowdsourcing. Work has been published that discusses the different types of cheaters (Gadiraju et. al. 2015), the different ways they cheat, and the different motivations for why (Eickhoff and De Vries, 2013).

In general, researchers address quality control with two techniques: gold-standard questions, and platform reputation. Gold-standard questions provide researchers a means of keeping workers honest by periodically and randomly testing them with work items which have already been reviewed by experts. Platform reputation is a method whereby researchers are able to restrict the access of their crowdsourcing tasks only to those workers who have achieved a certain level of positive repute on their respective platform. These *vetted crowds* achieve distinction by completing a certain amount of work on the platform and by maintaining a certain accuracy rating on gold-standard test questions.

Even with careful and well thought-out quality control measures in place, crowdsourced data often remains tainted with low quality work [(Sabout et. al. 2012), (Eickhoff and De Vries, 2013)]. The prevalence of low quality work in microtasking environments, even after filtering out malicious workers, is likely due to inattention. Research shows that media multitasking leading to split attention is exceedingly common in computer users (Papper, Holmes, & Popovich 2004 and Kaisar Family Foundation 2005). Haider and Frensch (1996) have shown that participants in a task will gradually learn to ignore elements that are unrelated to task performance as a part of information reduction associated with skill acquisition. They show that it is not always the case that reading instructions is the most efficient means of completing a task. Many microtasking participants have likely learned that trial and error is a more effective strategy than reading instructions. Therefore task designers should rely on intuitive designs that lend themselves to a priori learning and assume that workers will allocate the minimum attention required to complete the task.

Kittur et. al. (2008) demonstrate that thoughtful microtask design can significantly improve results, increasing the accuracy and consistency of the crowdsourced data. Sabout et. al. (2012) show that a well-designed task can be more attractive to legitimate workers

as well as detracting to illegitimate ones. Bontcheva et. al. (2014) discuss the merit of concise instruction and clear, intuitive interface design, further supported by Khanna et. al. (2010) which suggests that thoughtful design can be just as, if not more important than an interface presented in a worker's native language.

Our work in this paper aims to demonstrate that microtask workers generally do not read instructions carefully and that therefore task instructions, either verbose or concise, are not the best means of teaching workers about the task they are to undertake. We also seek to demonstrate that *vetted crowds* are not necessarily more attentive or legitimate workers than their counterparts of lesser repute.

In this paper we describe two experiments; the first uses a fake survey to measure how many workers in vetted and unvetted crowds read instructions carefully. The second experiment compares the performance of workers when using two different interface designs to complete an entity tagging microtask. Our results show that a majority of workers in both vetted and unvetted crowds do not read instructions and that intuitive task design reduces the barrier to entry allowing more workers to complete the task accurately.

2 Attention Experiments

A set of three faux tasks were designed in order to measure how many participants actually read the instructions in the crowdsourcing microtasks we deployed. All three conditions of the faux task were tested with both relatively new members of the platform as well as more experienced members. The first crowd, which we refer to as unvetted, consisted mostly of inexperienced workers or workers with poor performance records. The second crowd, which we refer to as vetted, had previously demonstrated their competence in crowdsourcing microtasks by answering at least 100 test questions with a minimum overall accuracy of 70% in prior tasks. It is worth noting, however, that the unvetted crowd had no restrictions and therefore included a small subset of participants who had qualified as being vetted.

Attention Experiment 1

Our first experiment took the form of a false geography quiz consisting of ten questions. In this experiment, we tested the condition that required the highest level of attention to detail to pass. We hereafter refer to this as the High Level Attention Task (HLAT). In order to complete this task successfully, workers had to read through the instructions until they reached the second paragraph. The second paragraph reveals the purpose of the experiment to the participants and informs them of a secret answer that they must enter to successfully complete the task. Users that entered the secret answer to every question, as instructed, were considered to have passed the task. If participants read part of the instructions but skipped the second paragraph, either by reading only the first or third paragraph or jumping from the first to the third, they would not be exposed to the secret instructions and would therefore have no inclination that the task was actually an experiment in disguise.

This task was run with both the vetted and unvetted crowds. 509 unvetted workers completed the task, of which 129 passed (25%). Of 297 vetted workers that participated, 80 passed (27%). A chi-square test was used to determine if the difference in pass rate between the two crowds was significant (Pearson 1900). A p-value of .618 was found, indicating that the difference was not statistically significant.

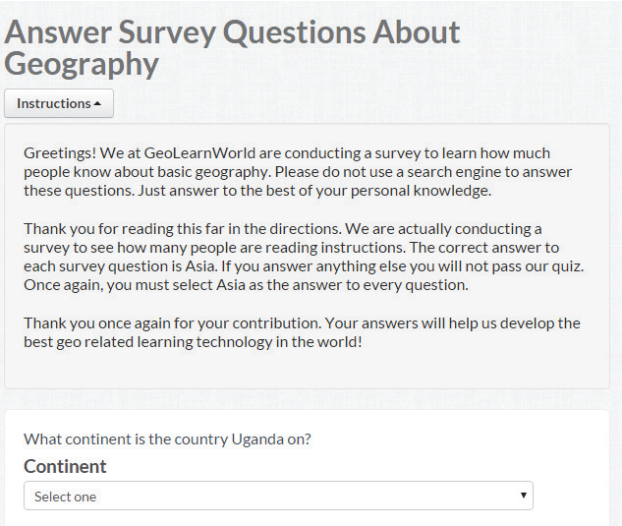

Figure 1: Screenshot of our crowdsourcing task in the HLAT condition.

Attention Experiment 2

The second experiment, like the first, was a false geography quiz. This task differed from the first in that a very visible warning was displayed at the beginning of the instructions which told participants that they needed to read on, lest they fail the task. There was no content in the instructions that would lead the participants to believe that this was a legitimate geography quiz.

We refer to this condition as a Medium Level Attention Task (MLAT). 159 of 339 unvetted users passed (47%) while 121 of 292 (41%) vetted users completed the task successfully. Although the unvetted workers here actually performed slightly better than the vetted workers, it is important to note that the difference was not found to be statistically significant, with a p-value of .168.

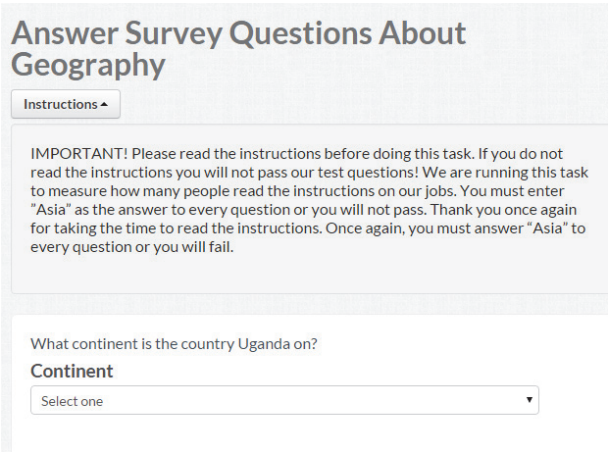

Figure 2: Screenshot of our crowdsourcing task in the MLAT condition.

Attention Experiment 3

In our third and last attention experiment, we wanted to test how many workers would pass if they were given the secret answer to our question in the most obvious possible way. We deployed another geography quiz, with the answer explicitly provided in the title of the job. The instructions consisted only of an explanation that the job was not a real quiz, and gave instructions to the worker on what they needed to do to pass. We refer to this as a Low Level Attention Task (LLAT). 187 of 328 unvetted workers taking this task passed (57%) while 114 of 143 (80%) vetted workers passed. The P-value in this condition was below 0.05, meaning the difference was statistically significant.

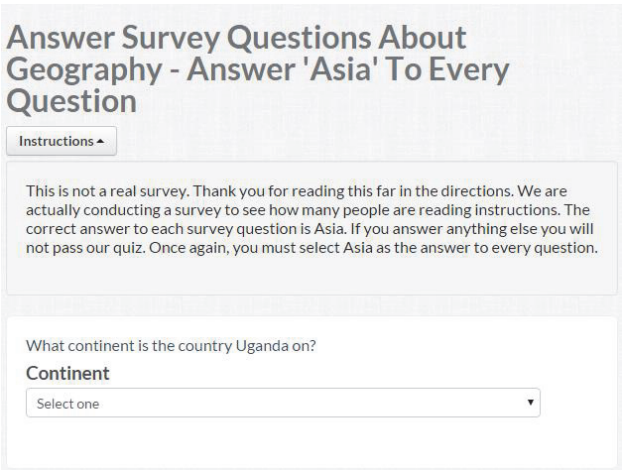

Figure 3: Screenshot of the crowdsourcing task in the LLAT condition.

Filtering of Participants

It was important in this task to ensure that each user counted in these tests had never encountered an earlier version of the experiment. In order to prevent this situation, participants were identified by a user ID and their results were filtered from those who had never been exposed to some form of our attention experiment before. No two experiments were ever running at the same time as this might introduce selection bias based on which task was chosen when two options were available. Additionally, workers were filtered out if their responses were not a legal response from the drop down list of available answers. It was possible for workers to enter answers that were not legal because of embedded translation programs in their browsers which translated the drop down options into their native language.

Discussion of Attention Experiment Results

The results in Table 2 indicate that vetted crowds are no more likely to read instructions than unvetted crowds. The unvetted and vetted crowds had 25% and 27% passing rates respectively when the answer was obscured by standard looking instructions (HLAT). This shows that the majority of crowd workers do not read instructions carefully. Both groups had a significant improvement in performance when the instructions contained a more visible warning to grab their attention (MLAT) with unvetted and vetted pass rates increasing to 47% and 41% respectively. This suggests that about half of crowd workers skim the directions if they do not read them carefully and will notice words that stand out. This result shows the importance of formatting instructions in a way that isolates and emphasizes the most relevant information rather than allowing those details to be lost in a wall of uniform text. In both the HLAT and MLAT pass rates, no statistically significant difference was found between the vetted and unvetted groups. Therefore the results show that vetted crowds are just as unlikely as unvetted crowds to read instructions carefully.

	Unvetted		Vetted		
	Pass Rate	Sample Size	Pass Rate	Sample Size	P-value
HLAT	25%	509	27%	297	0.61848687
MLAT	47%	339	41%	292	0.16830235
LLAT	57%	328	80%	143	0.00000238

Table 1. Crowd pass rates for various conditions of Attention Experiments

The third experiment (LLAT) showed a statistically significant improvement in the pass rate of vetted participants when compared to unvetted participants (80% compared to 57%). This suggests that vetted participants

were more likely to read and understand the obvious warning in the title rather than skipping straight to the task. A possible explanation for this improvement is that vetted workers are more likely to take their results seriously. Vetted workers may care more about their performance and maintaining their statistics on crowdsourcing tasks than unvetted workers, thus making them more likely to pay attention to the task and title if not the instructions. Another possible explanation is that vetted workers are much less likely to be automated bots entering random incorrect answers. In a future experiment, it may be worthwhile to include a form of CAPTCHA with the test to ensure that workers are in fact human.

As one might expect that the vetted users perform better in all conditions, a possible explanation for the similarity in performance is that the status of being vetted gives these workers a degree of confidence significant enough that they tend to feel experienced enough to pick up a task without reading the instructions. They assume they can learn the task through the test questions and only take the extra time to read the instructions when the benefit of doing so is easily recognizable.

In cases where a worker failed the task, they were given an explanation after-the-fact which indicated that the 'real' answers to the tasks were contained in the instructions and suggested they go back and read them. Workers also had a chance to dispute our questions once they received the notification that they had answered incorrectly. Approximately 1% of users disputed and wrote a justification of why they thought they were right. In reading through the users responses, we found no indication that any respondents had actually gone back to read the instruction retroactively. It is impossible to know how many users who answered incorrectly went back to read the instructions and did not dispute. These findings suggest, however, that there will be a subset of workers in any microtask who will not read the instructions, even when given the chance after being given negative feedback on their performance.

3 Entity Annotation Experiment

In this experiment, we created two separate tasks to evaluate how variations in task design impact results when instructions are held constant between them.

Entity Annotation Experimental Design

We created an entity tagging task where users were asked to annotate media-related sentences, annotating for the following entities: *Channel*, *Genre*, *Person*, and *Program*. Simple annotation instructions and examples were provided, gold-standard questions were created and inter-annotator agreement was assigned for quality-control. Both tasks were given the exact same input data. The tasks differed only in the layout and format of the work itself.

The first task was designed to be as straightforward and streamlined as possible. Text entry fields were provided for each entity tag that was expected, and users were to enter the appropriate entity into its respective text entry field. If a particular tag is not present in a given sentence, workers are expected to enter an arbitrary string, in this case "<N/A>", to denote that no such tag can be found. Tooltips are given next to each tag, which show users a popup containing rules and examples for that specific tag. This makes it so that workers are able to quickly reference guidelines and examples throughout the task instead of having to navigate to the instructions at the top of the page whenever they have a question. A screenshot of the task has been provided in Figure 4. Results were post-processed to insert the tags into the data.

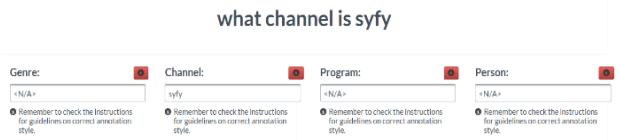

Figure 4: Screenshot of streamlined microtask.

The second task was designed to be more basic, relying on workers to manually insert tags into the data. In this version of the task, workers were asked to re-write the original sentence, manually inserting xml-style tags where appropriate. Unlike the previous task, no tools were provided to make it easier for the worker to reference the guidelines for the task. This means workers of this basic task are required to remember what tags are expected, since the task itself does not show the tags as it does in the streamlined task. Furthermore, workers for this task are forced scroll to the top of the task and reference the instructions if they have any questions. In all, it makes it such that a worker must be willing to work harder to successfully complete the task.

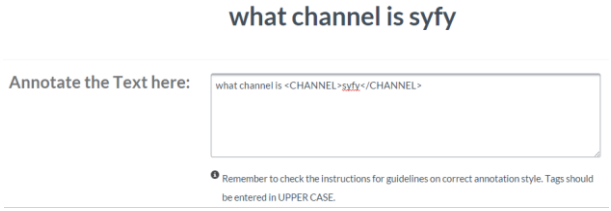

Figure 5: Screenshot of basic microtask.

Table 1 shows the results of these two tasks. The results show that the streamlined approach was considered more agreeable by workers, with many more participants passing the pre-test. It is important to note, however, that the

accuracy of the end results differed little between the two groups.

	Streamlined	Basic
Unique Workers	16	10
Avg Judgments/Worker	19.72	26.8125
Quiz Failure Rate	11.10%	45.50%
Mean User Rating	4.2/5	2.7/5
Collection Duration (min)	143	228
Mean Time/Sentence (sec)	27	24
Accuracy	92%	89%

Table 2. Results of microtask design implementation

Entity Annotation Experiment Discussion

It is likely that the reason behind why the quality of the end results of the two tasks differed little is because of quality control measures, namely the entry quiz that users had to take before starting the task which consisted of test questions. Since 45% of participants failed this quiz in the basic implementation and only 11% in the streamlined, we expect the actual results of the basic implementation would have been significantly worse had such a quiz not been present to act as a gatekeeper.

The streamlined implementation had the added benefit in that it attracted a greater number of workers to the task, thus allowing it to finish significantly faster than the basic variant. Many workers struggled to pass the preliminary quiz of the basic task, resulting in a smaller pool of qualified workers actually moving on to complete work. This made it so that the completion of the task was reliant upon a smaller group of people, each completing a larger amount of work units. This, without a doubt, contributed to the extra time needed for the task to complete.

4 Conclusions

We can conclude from the results of these experiments that makers of microtasks should not rely on users reading directions. Regular and informative test questions embedded within a task work better as a teaching method for crowds. Vetting users based on previous performance does not seem to increase the chances of users reading the instructions carefully. Although there is some evidence that vetted users will more often read the instruction when the incentive to is made extremely obvious to them. There will always be a subset of both vetted and unvetted crowd workers who do not read instructions, no matter how strongly they are encouraged. In the world of crowdsourcing microtasks, very simple, intuitive jobs work best as they allow a user to quickly pick up the task without needing to read instructions. Easy to understand tasks not only result in increased quality, they also finish faster and are rated higher by their participants.

References

Bontcheva, K., Derczynski, L., & Roberts, I. 2014. Crowdsourcing Named Entity Recognition and Entity Linking Corpora. *The Handbook of Linguistic Annotation. Forthcoming.*

Eickhoff, C., & De Vries, A. 2011. How crowdsourcable is your task. In Proceeding of the workshop on crowdsourcing for search and data mining *at the fourth ACM international conference on web search and data mining*. 2011., 11-14. Hong Kong, China: ACM.

Eickhoff, C., & De Vries, A. P. 2013. Increasing Cheat Robustness of Crowdsourcing Tasks. *Information retrieval* 16(2):121–137.

Gadiraju, U., Kawase, R., Dietze, S., & Demartini, G. 2015. Understanding Malicious Behavior in Crowdsourcing Platforms: The Case of Online Surveys. In Proceedings of CHI Conference on Human Factors in Computing Systems. Seoul, Korea: CHI.

Genarro, R., Gentry, C., & Parno, B. 2010. Non-interactive verifiable computing: Outsourcing computation to untrusted workers. In Proceedings of the 30th Annual Cryptology Conference of the International Association for Cryptologic Research, 465-482.

Haider, H., & Frensch, P. A. 1996. The role of information reduction in skill acquisition. Cognitive Psychology, 30(3), 304-337.

Ipeirtois, P.G., Provost, F., & Wang, J. 2010. Quality Management on Amazon Mechanical Turk. In Proceeding of the ACM SIGKDD workshop on human computation, 64-67. Washington DC, United States: ACM.

Kaiser Family Foundation. 2005. Generation M: Media in the lives of 8-18 year olds. CA: Menlo Park.

Khanna, S., Ratan, A., Davis, J., & Thies, W. 2010. Evaluation and Improving the Usability of Mechanical Turk for Low-Income Workers in India. In Proceedings of the first ACM symposium on computing for development, 12-21. London, United Kingdom.

Kittur, A., Chi E. H. & Suh, B. 2008. Crowdsourcing User Studies with Mechanical Turk. In Proceedings of the SIGCHI conference on human factors in computing systems, 453-456. Florence, Italy.

Papper, R.A., Holmes, M.E., & Popovich, M.N. 2004. Middletown media studies: Media multitasking...and how much people really use the media. The International Digital Media and Arts Association Journal, 1(1), pp. 9-50

Pearson, Karl. 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine Series 5 50 (302): 157–175.

Rutherford, A., Peng, F., & Beufays, F. 2014. Pronunciation Learning for Named-Entities through Crowdsourcing. In Proceedings of the 15th Annual Conference of the International Speech Communication Association, 1448-1452. Singapore: Interspeech.

Sabout, M., Bontcheva, K., & Scharl, A. 2012. Crowdsourcing Research Opportunities: Lessons from Natural Language Processing. In Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies, p.17. Graz, Austria: International Conference on Knowledge Management and Knowledge Technologies

Zaidan, O., & Callison-Burch, C. 2011. Crowdsourcing Translation: Professional Quality from Non-Professionals. In Proceeding of the 49th Annual Meeting of the Association for Computational Linguistics, 1220-1229. Portland, Oregon: Association for Computational Linguistics.