Use of internal testing data to help determine compensation for crowdsourcing tasks

Mike Lauruhn, Helena Deus, Paul Groth, and Corey Harper

Elsevier Labs, Philadelphia, PA 19103, USA
m.lauruhn@elsevier.com

Abstract. Crowdsourcing is a popular means for developing datasets that can be used to build machine learning models including classification and entity recognition from text as well as tasks related to images, tables, charts, and graphics. As the use of crowdsourcing grows, concerns about appropriate and fair compensation of contributors are also increasing. However, estimating the correct compensation levels can be a challenge a priori. In this paper, we will describe the input and data that inform considerations for how much to pay workers for various tasks. We will give an overview of three separate crowdsourcing tasks and describe how internal testing processes and qualification tasks contribute to the end user (Worker) experience and how we attempt to gauge the effort required to complete a task.

Keywords: Ethics of Crowdsourcing, Crowdsourcing, Subject Matter Experts, Microtasks.

1 Introduction

Crowdsourcing is growing in popularity as a means to accelerate creating datasets that are used to build and train machine learning models including classification and entity recognition from text and tasks related to images, tables, charts, and graphics. However, as the use of crowdsourcing expands, some researchers are bringing to light issues about workers earning fair payment in exchange for their contributions to the tasks [Fort et al., 2011]. In addition to being paid adequately, crowdsourcing workers have also voiced concerns about other aspects of the platforms including having to return work due to unclear instructions or having requester underestimate the time it takes to complete a task [Semuels, 2018].

While our management with Elsevier Technology Services encourages us to pursue and continue to evaluate crowdsourcing as a means to create datasets, we are also given guidance on proper compensation and the types of tasks we are permitted to make available via crowdsourcing. As a target, we try to price tasks in order to be above United States minimum wage of $7.25 USD per hour [U.S. Dept. of Labor, 2018]. In addition, we intentionally attempt to create HITs that take around 10 to 15 minutes or less to complete. This is to allow some flexibility for Workers and also to

mitigate risk of losing a payment for being on a task for a long time only to experience technical difficulties as a result of connectivity.

However, estimating the appropriate compensation levels can be a challenge. This paper describes the input and data that inform considerations for how much to pay workers for various tasks. We will give an overview of three separate crowdsourcing tasks and describe how internal testing processes and qualification tasks contribute to the end user (Worker) experience and how we attempt to gauge the effort required to complete a task.

2 Project Descriptions

This section describes three sample crowdsourcing projects that Elsevier Labs completed on the Amazon Mechanical Turk platform.

2.1 Citing Sentences

One of the first projects that we did as a proof of concept for crowdsourcing was designed to build a sentence classifier for sentences in scientific and medical articles that cite previous articles. Specifically, we wanted to build a classifier of intent of the citation. That is to say, what is the purpose of this citation and why did the author include it? For this, we came up with a list of high-level citation types. They are: Contextual, Disputes, Evidence, Extends, Supports, and Resource. For the purpose of providing instructions, as well as clarity within our own team, we subsequently created a short definition for each:

Contextual – cited entity provides general information for this domain.
Disputes – disagreement with the cited entity.
Evidence - cited entity provides material or data accepted as fact.
Extends – builds on cited entity's work.
Supports – Author expresses agreement with the cited entity.
Resource – cited entity provides method, data, etc. that was utilized.

In addition to these categories, the Workers could also select, "None of these, or, cannot be determined."

For the task, Workers would be presented with a sentence from a scientific article that contained at least one citation. The actual citation, typically an author or authors' names and a year in parentheses, would be highlighted. This was to disambiguate the citation in cases where sentences had multiple citations. The Worker also saw the preceding and following sentences in order to provide additional context that might aid in categorizing it accurately.

The end result was to create a process that would provide a certain number of qualified Workers hundreds (or perhaps thousands) of citations that they would categorize. When there is consensus among workers as to what the citation type of a sentence is, then that sentence would be collected as part of a Machine Learning classifier.

2.2 Image Similarity

Another pilot that was conducted was called Image Similarity. A researcher from Elsevier Labs was working on a Machine Learning algorithm for ranking image similarity. The purpose of the Image Similarity exercise was to compare similarity scores generated from that algorithm to a group of humans' perception of similarity. The researcher could then identify significant differences -- where the algorithm scored an image pair as being similar, but the humans scored them as being different and vice versa. The researcher could the use this information to make adjustments to their training process and algorithm.

For the Image Similarity task, Workers were presented with a pair of images asked to describe how similar the two images were to one another on small scale. There were 250 image pairs selected. One set were actually duplicate images, a larger set has distributed scores of similarity as generated by the algorithm, and images for a third set were randomly selected, but had a score associated with them for comparison. Workers were given a limited scale of four options to choose from when describing similarity: *Almost Identical*, *Very Similar*, *Similar*, and *Different*. The Worker instructions included examples along with definitions of how to use each category:

> ***Almost Identical*** *means that it is difficult to find any differences in the images. They are possibly interchangeable if not the same image. In some cases the same image could be inverted or rotated.*
>
> ***Very Similar*** *means that the two images share many visual AND subject (semantic) characteristics. That is to say, it is clear that they are two different images, but they are probably using the same medium and are likely to be depicting the same subject. The images could be part of the same series, composite, or sequence of images.*
>
> ***Similar*** *means that while the images share some similarities, but are quite different. For example, two different media of the same subject – like a sketch and an image of the same thing would be considered similar. Or two of the same style of graphics depicting different data would be similar.*
>
> ***Different*** *means that the images share almost no characteristics. Images that share a similar orientation or color palette but nothing else would also be considered different.*

As with the Citation Type example, we also included a Cannot Determine option and because of the manner in which the Image Similarity template was populated during tasks, we anticipated that *Cannot Determine* would be used primarily to indicate broken links that prohibited an image from rendering properly.

2.3 Triple Extraction Evaluation

A third crowdsourcing project that was conducted involved the evaluation of semantic triples from scientific corpora. The project was testing the quality of different Open Information Extraction (OIE) processes in generating triples from sentences. More specifically, the OIE processes were compared in their abilities to generate triples in scientific text and non-scientific encyclopedic text.

For these tasks, Workers were presented with a sentence that was either from scientific text or non-scientific encyclopedic text (the Worker was unaware of the source) and a series of semantic triples that were extracted from the sentence by various OIE methods. The number of triples ranged from 1 to *n*. Workers were asked to select each valid triple my selecting checkboxes in the template.

The output of this project was for the OIE evaluation reported in "Open Information Extraction on Scientific Text: An Evaluation" [Groth et al 2018].

3 Task Process overview

As appropriate for the Citation Types and Triple Extraction Evaluation tasks, we followed recommended practices for the use of crowdsourcing in linguistics [Erlewine and Kotek, 2016]. We used Amazon Mechanical Turk as the crowdsourcing platform. Within Mechanical Turk, we made extensive use of the sandbox test environment. This has the advantages of easier onboarding of colleagues to participate in testing and it is an environment where no money is exchanged between the annotator (Worker) and the researchers (Requester). Within Mechanical Turk tasks are called Human Intelligence Tasks (HITs). A HIT is comprised of a series of Microtasks (referred to as Items) i.e. a sentence with a citation is an Microtask, an image pair is an Microtask, and a sentence and its triples are an Microtask.

An early step of each of the three projects was to create a gold set of known correct answers to Microtasks for the HITs. To do this, the principal researcher on the project worked closely with the Mechanical Turk implementation lead and created a series of HITs. The HIT would be launched in the sandbox environment the two of them would test it amongst themselves. This step helped validate the functionality of the template and made certain that the required data was being capture correctly. This also gave an initial indication of how long each Microtasks was taking to complete and how many Microtasks should be in each HIT.

After the template was modified to contain the desired number of Microtasks per HIT, a set of HITs was loaded into the Sandbox environment and made visible to a group of internal Workers (Elsevier Labs colleagues) who were encouraged to take some of the HITs over the course of a few days. This step of the process was critical for collecting information and data that would inform the HITs that would be presented to the actual Workers. First, our internal Workers were instructed to read the instructions and report back on their clarity and the examples provided. Second, the amount of time it took the Workers to complete the HITs was noted. Finally, after the HITs had been completed, the principal researcher could review annotated Microtasks and use the results to construct a gold set of answers. In our design, gold set answers serve two primary functions: first, the can be used to create a qualification HIT as we are recruiting Workers. A qualification HIT is a HIT that we made available to the public with the understanding the Workers will be evaluated based on how closely they matched the responses of the internal annotators. Second, gold set Microtasks can be distributed across HITS as a means to monitor quality of the individual Workers over the course of the project.

When assembling the gold set, an additional strategy was to ensure that quality Workers were identified. Using the Microtasks that the internal Workers achieved the highest consensus on increases the risk of only including obvious or "easy" responses which might result in several of the people taking the Qualification HIT getting high scores and not differentiating themselves from one another. Aside from the high consensus Microtasks, researchers were encouraged to include more challenging Microtasks – especially where understanding the difference between two similar choices is important to the data collection outcome.

4 Collect Data from HITs

This section describes size and makeup of the HITs we created and the metrics we were able to collect during three separate phases of the three crowdsourcing projects described above. Those phases were the internal testing, the qualification HIT, and the actual HIT.

4.1 Citing Sentences Data

For the Citing Sentences task, ten HITs were prepared for the internal annotating phase. Each HIT had ten items (citations) to be categorized. We had a requirement that each HIT would be seen by at least three unique members of the internal annotators group. A total of five different annotators participated in the task with one completing nine HITs, two completing six HITs, one completing three HITs, and one completing two. This gave us a total of 26 HITs to analyze when monitoring efficiency, but fewer when looking for consensus Citations to consider for the inclusion in the Gold Set.

Based on data collected from Mechanical Turk administration tools, we observed that the average Worker 422 seconds, or roughly seven minutes to complete a single HIT. Further review of the individual times revealed that the Worker who had completed three HITs had outlier times that were significantly higher than the others. Their average time to complete HITs was over 17 minutes. We were able to confirm that this Worker was a user experience resource that had been tasked with taking notes about the template and other aspects of usability that could be improved before creating a public-facing task. In fact the usability resource actually completed one of the HITs while on the phone with the Mechanical Turk Administration. These outlier times were taken into account when considering the anticipated time to complete a HIT.

From the completed HITs from the internal annotators, citations that received a consensus category by way of inter-annotator agreement were used to create a qualification set of ten Microtasks. The qualification HIT was made available to 14 public Workers. These 14 Workers averaged 925.29 seconds or 15.42 minutes to complete the HITs. We had expected the HITs to take significantly longer when compared to the internal annotators as the public Workers were reading the instructions and the content for the first time. The four Workers who matched the most annotations to those from the Gold Set were selected (qualified) to do additional HITs. These four

Workers averaged 931 seconds or 15.51 minutes to complete the task, comparable to the times of the other workers. They did not complete the tasks particularly faster or slower than the entire group.

For the next set of HITs, the qualified Workers were given HITs with 15 Microtasks – five more than Microtasks than in the Qualification. The HIT would be priced at $1.50 USD each. Based on the average times it took to complete HITs with ten Microtasks in the qualification, while removing the time needed to read the instructions for the first time, we estimated that Workers would finish HITs in around 12 minutes – earning $7.50 to complete five HITs in an hour.

For the HITs, 12 of the Microtasks were randomly selected Citations, while three were Citations that were part of our Gold Set. A total of ten HITs were prepared with a requirement that each would be annotated by at least three individual Workers. One Worker completed all ten HITs, one Worker completed nine HITs, one Worker completed seven, and the fourth Worker completed one HIT. The Workers' times to complete each HIT averaged 658.88 seconds, or 10.98 minute. When compared to the Qualification times, the Workers were significantly faster. Even though they were presented with 5 more Microtasks, the average times went from 15.51 minutes for the Qualification HIT to 10.98 minutes for the standard one. Measuring how many seconds Workers took per Microtasks, it went from 92.5 to 43.9 (see Figure 1). This is most likely because the Workers were now familiar with the instructions and developed personal techniques to perform more efficiently. Based on the pricing, a Worker who happened to hit that average would have completed 5.46 HITs in one hour and earned $8.19. If a Worker was able to complete 6 HITs in an hour, they would have earned $9.00 USD, while a Worker who completed 5 HITs in an hour earned $7.50.

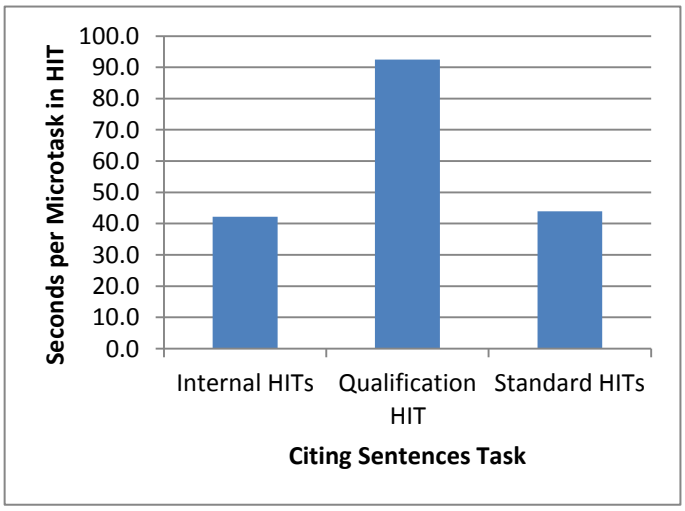

Fig. 1. Average time in Microtasks per second taken to complete the three types of Citing Sentence HITs.

4.2 Image Similarity Data

For the Image Similarity internal testing, four HITs were prepared. Each contained 15 image pairs. Five internal annotators completed all four HITs, one completed three, and a sixth one completed two of them. On average, each HIT took 161.03 seconds or 2.68 minutes. As with the Citation Type task, the user experience resource was participating in an effort to provide guidance on the task design and had outlier times. Without their times included the averages were lowered to 115.37 seconds or 1.92 minutes.

For the Qualification HIT, 15 public Workers took one HIT with 30 Microtasks (image pairs). The HIT was completed on average in 532.2 seconds or 8.87 minutes. The four who matched to most annotations to the Gold Set were Qualified for future work. They averaged 495.25 seconds or 8.25 minutes to complete the task.

Twenty Image Similarity HITs were prepared. Each had 25 Image Pairs, five fewer than in the Qualification. The Workers were paid $1.00 for each HIT. We anticipated with the smaller collection of images, Workers times would be closer to 7.5 minutes, earning $8 per hour. Three of the Qualified Workers complete Image Similarity tasks: One completed 20, a second completed 17, and the third completed 13. The HITs were completed on average in 424.51 seconds or about 7.08 minutes. This time is roughly consistent with time it took to complete the Qualification HIT, 17.7 seconds per item vs. 17.0 seconds per item (see Figure 2). Workers categorized 3.63 Image pairs per minute in the Qualification HIT and 3.53 Image pairs per minute in standard HITs. Workers who completed 8 HITs in an hour earned $8. Workers who complete 9 HITs in an hour earned $9.

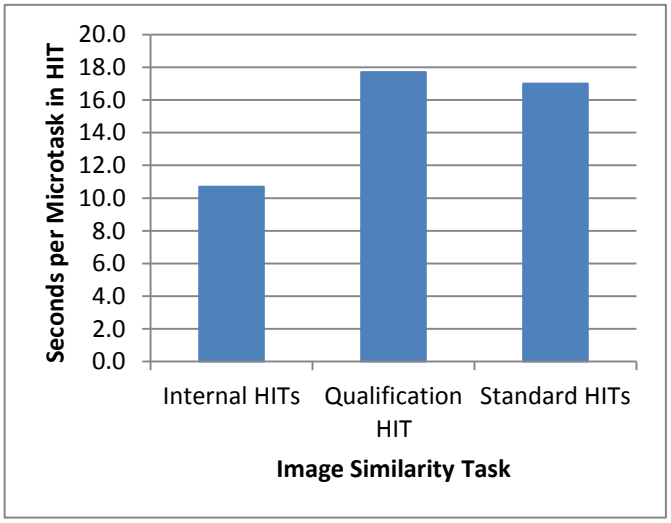

Fig. 2. Average time in Microtasks per second taken to complete the three types of Image Similarity HITs.

4.3 Triple Extraction Data

For the Triple Extraction tasks, eight HITs were prepared for internal testing. Each one had ten sentences and the series of triples that were extracted from them. Ten different internal annotators participated. A total of 44 HITs were taken. Two annotators completed eight HITs, one completed seven, two completed six, four completed two, and one completed one. On average, the HITs took 412.80 seconds or 6.88 minutes to complete. Removing two HITs from the user experience resource lowered the average slightly to 391.17 seconds or 6.52 minutes.

The qualification HIT for the Triple Extraction had 25 sentences in it and was made available to 15 public Workers on Mechanical Turk. Workers took an average of 1546.60 seconds or 25.78 minutes to complete the Qualification HIT. Ten Workers who performed well compared to the gold set answers collected during internal testing were Qualified and invited to participate in more HITs. Their average times to complete the qualification were only slightly faster (1489.30 seconds, 24.82 minutes) than those of the complete field.

For the standard HITs, 75 were prepared, each with ten sentences and their extracted triples. The HITs were initially priced at $1each. This was due to the HITs being much smaller than the Qualification task. We anticipated the Workers would complete a HIT in around eight minutes. This was a significant underestimate. Each HIT was seen by five unique workers. It took 1792.60 second or 29.88 minutes on average for the Workers to complete these HITs. The number of seconds per Microtask jumped from 61.9 in the Qualification to 179.3 in the actual HIT (see Fig. 3). At one point, Work had stalled on completing the tasks and we raised the price to $1.25 and offered additional bonuses to encourage completion. Payment here was significantly lower than the previous two tasks. Based on the averages, workers earned between $2 and $5 an hour for completing two HITs.

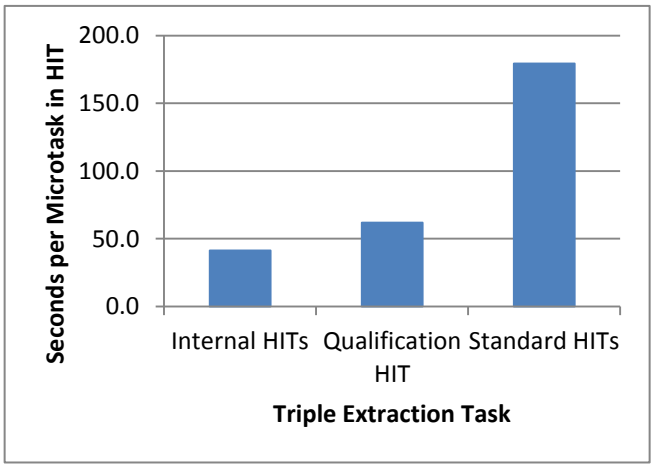

Fig. 3. Average time in Microtasks per second taken to complete the three types of Image Similarity HITs.

Upon reviewing the individual numbers further, we observed that two Workers averaged over 60 minutes to complete the tasks and two more averaged 34 minutes, while five finished between 19 and 26 minutes. While some indicate that underestimates on completion time are sometime a result of unclear instructions [Silberman, 2018], these Workers had already seen the instructions during the qualification and were working on several small tasks.

5 Future Application & Discussion

We are currently preparing to launch a new HIT where Workers will be asked to categorize sentence types similar to the Citation Type project described earlier. We have completed internal testing a made the Qualification HIT available to the public. Eight Workers took the HIT and four will be invited to participate in future HITs. On the Qualification, the Workers categorized 30 sentences. Workers took 10.87, 15.7, 17.5, and 17.67 minutes to complete the HIT for an average of 15.43 minutes. This HIT contained 30 sentences. The plan is to publish HITs with 20 sentences and we are estimating 12 minutes to complete the HIT. One additional consideration that we are planning to make is Amazon's 20 percent fee on what workers are paid as described in [Semuels, 2018]. Instead of targeting the $7.25 minimum wage number, we will likely shoot for $8.70 per hour. In this instance where we are assuming five HITs in an hour, we will pay close to $1.75 per HIT.

Task	**Worker Count**	**Microtasks per HIT**	**HITs Completed**	**Avg Time per HIT**
Citing Sentences ($1.50 per hit, Average $8.19 per hour)				
Internal	5	10	26	422 sec
Qualification	14	10	14	925.29 sec
Final	4	15	27	655.88 sec
Image Similarity ($1.00 per hit, Average pay of $8.47 per hour)				
Internal	5	15	25	161.03 sec
Qualification	15	30	15	532.2 sec
Final	3	25	50	424.51 sec
Triple Extraction ($1.00, then $1.25 per hit, Average pay between $2 and $5 per hour)				
Internal	10	10	44	412.80 sec
Qualification	15	25	15	1546.60 sec
Final	5	10	375	1792.60 sec

Fig. 4. Overall breakdown of calculations from each of the three HITs. All payment information only applies to final, production run of Tasks. Internal rounds were-unpaid while and Qualifying HITs were fixed price.

In two of our three tasks described, we were able to meet our goals for compensating Workers relatively closely. However the third task was significantly off, where Workers were, on average, taking more than three times as long to finish tasks than we anticipated. This brings up a few discussion points. First, as described above, these numbers are on average. If a Worker completes the same task in ten minutes that it takes another Worker 60 minutes to complete, should prices be adjusted to come clos-

er to the average? Or is that fact that the more efficient Worker can complete more tasks faster fair as long as the quality of their work remains satisfactory? Our projects have fairly generous time limits as we do not want someone to lose work if they do not meet a specific timeline. We have no way of knowing if a Worker accepts a HIT and then walks away from their space and are actually not working exclusively on our HIT. This makes the use of minutes spent to return a HIT a bit of a questionable indicator to rely on.

The Triple Extraction task also illustrates a need to facilitate better immediate communications between the Workers and the requesters. Workers might take a HIT and decide that they are not compensated enough or cannot understand the instructions, they may simply leave the project. Crowdsourcing platforms should consider means to provide for immediate communication channels between Workers and Requesters. Even if it is simply a rating alert if a task is priced too low. Requesters should be open to immediate feedback where they can make modifications to an active task or provide specific clarification as appropriate.

6 Conclusion

Demand for datasets to support machine learning and other artificial intelligence tasks means that many researchers, institutions, and companies are turning to crowdsourcing as a means to accelerate the creation of datasets for both testing and training processes. Many researchers, institutions, and companies are turning to crowdsourcing as a means to accelerate the creation of datasets for both testing and training processes. As this field expands, attention to the human element of the work involved must also grow so that workers are compensated fairly for their contributions. In this paper, we described three Mechanical Turk projects and shared processes of internal testing and qualification tasks that create data the help us inform how to design tasks and how to appropriately pay workers. In two of the three, we were able to achieve the desired target. For the third, we learned that sometimes the estimates can underpay and that in some cases, work effort needs to be monitored more closely as a project is getting underway. We also believe that more immediate communication channels between Worker and researcher would improve the situation greatly. In future work, we plan to take advantage of the Mechanical Turk APIs and other tools to help facilitate these types of communications.

References

1. [Erlewine and Kotek, 2016] Erlewine, M., Kotek, H. 2016. A streamlined approach to online linguistic surveys. Natural Language & Linguistic Theory, 34(2):481–495.
2. [Fort et al 2011] Fort, K., Adda, G., Cohen, K.. 2011. Amazon Mechanical Turk: Gold Mine ¨ or Coal Mine? Computational Linguistics, 37(2):413–420.
3. [Groth et al 2018] Groth, P., Daniel, R., Lauruhn, M., Scerri, A.: Open Information Extraction on Scientific Text: An Evaluation. In: [forthcoming proceedings of COLING 2018 conference; preprint available at: https://arxiv.org/pdf/1802.05574.pdf].

4. [Semuels, 2018] Semuels, A., "The Internet Is Enabling a New Kind of Poorly Paid Hell." *The Atlantic*. January 23, 2018 https://www.theatlantic.com/business/archive/2018/01/amazon-mechanical-turk/551192/ last accessed 2018/06/01.
5. [Silberman et al 2018] S. Silberman, M & Tomlinson, B & LaPlante, R & Ross, J & Irani, L & Zaldivar, A. (2018). Responsible research with crowds: pay crowdworkers at least minimum wage. Communications of the ACM. 61. 39-41. 10.1145/3180492.
6. [U.S. Dept. of Labor, 2018] United States Department of Labor Wage and Hour Division Homepage, https://www.dol.gov/whd/, last accessed 2018/06/01.