Assessing Crowdsourcing Quality through Objective Tasks

Ahmet Aker*, Mahmoud El-Haj†, M-Dyaa Albakour†, Udo Kruschwitz†

*Department of Computer Science, University of Sheffield, United Kingdom
†School of Computer Science and Electronic Engineering, University of Essex, United Kingdom

Email: a.aker@dcs.shef.ac.uk, melhaj@essex.ac.uk, malbak@essex.ac.uk, udo@essex.ac.uk

Abstract

The emergence of crowdsourcing as a commonly used approach to collect vast quantities of human assessments on a variety of tasks represents nothing less than a paradigm shift. This is particularly true in academic research where it has suddenly become possible to collect (high-quality) annotations rapidly without the need of an expert. In this paper we investigate factors which can influence the quality of the results obtained through Amazon's *Mechanical Turk* crowdsourcing platform. We investigated the impact of different presentation methods (free text versus radio buttons), workers' base (USA versus India as the main bases of MTurk workers) and payment scale (about $4, $8 and $10 per hour) on the quality of the results. For each run we assessed the results provided by 25 workers on a set of 10 tasks. We run two different experiments using objective tasks: maths and general text questions. In both tasks the answers are unique, which eliminates the uncertainty usually present in subjective tasks, where it is not clear whether the unexpected answer is caused by a lack of worker's motivation, the worker's interpretation of the task or genuine ambiguity. In this work we present our results comparing the influence of the different factors used. One of the interesting findings is that our results do not confirm previous studies which concluded that an increase in payment attracts more noise. We also find that the country of origin only has an impact in some of the categories and only in general text questions but there is no significant difference at the top pay.

Keywords: Mechanical Turk, Objective Metrics, Evaluation

1. Introduction

Crowdsourcing has become a serious alternative in creating resources for natural language processing, e.g. (Kaisser and Lowe, 2008; Alonso and Mizzaro, 2009; Yang et al., 2009; El-Haj et al., 2010; Albakour et al., 2010).

However, obtaining reliable results from the crowd remains a challenging task (Kazai et al., 2009), which requires a careful design of the experiment and also pre-selection of crowdsourcers. Different ways for obtaining reliable results have been investigated, two of the main techniques are to use aggregation over many assessments and the injection of tasks for which the correct result is known (Zaidan and Callison-Burch, 2011).

To determine the different factors that might affect annotation quality and their possible interdependence we conducted a study that explored a range of different variables. Specifically, using Mechanical Turk (MTurk)[1] as a crowdsourcing platform, we investigated the impact of different presentation methods, workers' base and payment scale on the quality of the results. For each run we assessed the results provided by 25 workers on a set of 10 tasks. For our experiments we used two types of objective tasks: maths tasks and general text questions. In both tasks the answers are unique, which eliminates the uncertainty usually present in subjective tasks, where it is not clear whether the unexpected answer is caused by a lack of worker's motivation, the worker's interpretation of the task or genuine ambiguity.

2. Related Work

In recent years, MTurk has begun to be recognised as a very promising platform for crowdsourcing. For example, Su et al. (2007) conducted four different experiments on attribute extraction and entity resolution. An average of precision above 80% was obtained. Snow et al. (2008) submitted one TREC topic with a number of irrelevant and relevant documents to MTurk for relevance judgement. The results showed high agreement between the average of assessments by MTurk workers and TREC assessors. In some cases, workers were even more precise. Similarly, Alonso and Mizzaro (2009) explored the use of MTurk for five categories of natural language processing tasks. They discovered that there existed a high agreement between the gold standard labels by experts and non-expert MTurk annotations, and that the average of a small number of workers can emulate an expert. As a cost-effective and fast service, MTurk is now being used by an increasingly large number of people for a variety of tasks, such as relevance judgement (Alonso et al., 2008; Alonso and Mizzaro, 2009), image annotation (Sorokin and Forsyth, 2008), natural language annotation (Snow et al., 2008), Wikipedia article quality assessment (Kittur et al., 2008), document facets extraction (Dakka and Ipeirotis, 2008), customising summary length for improving searching (Kaisser et al., 2008), building datasets for question answering (Kaisser and Lowe, 2008) and summarization (El-Haj et al., 2010), extracting key phrases from documents (Yang et al., 2009), user preference studies (Tang and Sanderson, 2010), email annotation (Albakour et al., 2010), and so on. Although requesters have full control over how much a worker is paid on completing a task, many of them seem to pay between $0.01 to $0.10 for a task taking "a few minutes". Assessing the quality of the work done through crowdsourcing is an important step. Quality control can be achieved using confidence score and gold-standards (Donmez et al., 2009; Bhardwaj et al., 2010) and empirical and cost constrains (Wais et al.,

[1] http://www.mturk.com

2010), which have been proven to be critical to understanding the problem of quality control in crowdsourcing.

3. Experiments

3.1. Objective Tasks

In our experiments we use maths questions and questions related to travel and history categories (general text questions). In general both types of questions can be answered if there is a motivation in performing the task. We investigated the impact of the following variables on the quality of the results:

1. Presentation method: free text versus radio buttons,
2. Workers' base: as the vast majority of workers come from either the US or from India and they make up more than 80% of all workers (Ipeirotis, 2010), we selected these two countries of origin for our investigation,
3. Payment scale: an estimated $4, $8 and $10 per hour.

For each run we assessed the results provided by 25 workers on a set of 10 tasks. Both experiments were conducted by accessing *MTurk* through *CrowdFlower*[2]. The two experiments were submitted with limitations on the workers' origins where we include only the two selected countries, i.e. US and India. US workers are from now on referred to as *Group 1* and workers from India make up *Group 2*. There was no limitation on the confidence rate as we wanted to include real workers and spammers in our experiment as we did not reject any of the submitted hits except for 3 hits that were submitted by workers from countries other than US and India.

Maths Questions These questions are word problems and are collected from different online learning sites.[3] The level of the questions vary from primary to high school (up to 6th grade) and thus the problems should be relatively easy to solve. However, because the questions are word problems the workers have to read the questions carefully in order to give the correct answers. In general these questions test the willingness of workers to answer the questions. In total we have 10 such questions. The questions vary in text length (min: 4, max: 75 and ave: 40 words). Table 1 shows a short and an average example question.

General text questions A number of 10 multiple choice questions have been selected for this experiment. The questions fall in the travel and history categories with the intent to measure the workers' knowledge and ability in reading questions and selecting correct answers. Compared to the maths questions these questions in general are not straightforward to be solved because the answer is not easily derivable from the text itself. Rather it requires some general knowledge of the task solver or the willingness to search for the correct answer on the web. All questions have exactly one correct answer. The length of the questions in average is 13 words with a maximum of 29 and a minimum of 3 words. The questions are collected from a suitable Web site[4]. Table 2 shows two examples of history and travel questions along with the choices of answers.

3.2. Experimental Design

For each question type we use two different designs: radio buttons and free text input. For the radio button design we offer the choice of four potential answers. Figure 1 shows a sample of one of the experimental designs we implemented for the text questions, which in this case is the free text presentation method. We first present the task that is to be performed, then list the criteria for a successful payment and finally list the questions. In Figure 1 there is only one question shown. However, in each HIT we show 10 questions. The worker's answers are dependent on the presentation method used. Workers are supposed to write the answer in the case of the free text presentation method or select one of the provided choices when the radio button presentation method is used. An example radio button design is shown in Figure 2 for the maths questions.

The reason of presenting two answering methods is to investigate the impact of the presentation method on the quality of answers. Our assumption is that in a design with check boxes or radio buttons the workers have some probability to guess the correct answer. However, in case of an empty text field where the worker has to write an answer the probability to write the correct answer without reading the task is very low.

4. Results

We performed several runs of the same experiment with different user settings and different payment incentives. We run our experiments with about $4, $8 and $10 per hour payments. In total we have 12 runs for each question type where each run differs from the others either in design (radio buttons versus free text), or in payment ($4, $8 or $10) or in origin of country (Group 1 or Group 2). In each experiment we use 10 questions (either maths or general text) with 25 different workers.

In each experiment we count for every worker the number of correct answers given by him/her. This means that every experiment has 25 such fields where each field corresponds to a different worker. The results of both maths and text questions are shown in Table 3 and 4 respectively.

From the tables we see that for maths questions the results tend to be generally better when radio button design is preferred over the free text field design. Furthermore, Table 3 and 4 also show that the payment can be an important factor that affects the results. For the Group 2 workers we can see for the maths questions that the higher the payment is the better results are obtained.

However, to see whether there are any significant differences between these results we compute significance tests between the different settings of the same experiment type (maths or general questions) using two-tailed paired t-tests.

[2] http://crowdflower.com/

[3] http://edhelper.com/math.htm, http://www.kidzone.ws/math/wordproblems.htm and http://www.amblesideprimary.com/

[4] http://www.triviaplaying.com/

What is double 80?
There was a fire in the building down the street. It was so large that our city had to call in 6 fire trucks. Each truck had 9 firemen riding on it. How many firemen arrived to fight the fire?

Table 1: Short and an average example maths question.

Genre	Questions	Answers
Travel	Which country is also called the Hellenic Republic?	(A)Sweden, (B)Denmark, (C)Greece, (D)Finland.
History	What U.S. president was born William Jefferson Blythe IV?	(A)Richard Nixon, (B)Bill Clinton, (C)Andrew Johnson, (D)Grover Cleveland.

Table 2: Example of History and Travel questions

Experiment	Average Score
Group1_4_RB	9.88
Group1_4_TF	9.28
Group2_4_RB	8.30
Group2_4_TF	8.24
Group1_8_RB	9.64
Group1_8_TF	9.28
Group2_8_RB	9.16
Group2_8_TF	8.52
Group1_10_RB	9.44
Group1_10_TF	9.40
Group2_10_RB	9.80
Group2_10_TF	9.44

Table 3: Average scores of the maths questions. TF stands for text field design and RB for the radio button design. 4, 8 and 10 are the payments per hour in US dollars.

Experiment	Average Score
Group1_4_RB	9.25
Group1_4_TF	9.32
Group2_4_RB	8.23
Group2_4_TF	8.12
Group1_8_RB	9.30
Group1_8_TF	8.88
Group2_8_RB	7.69
Group2_8_TF	9.40
Group1_10_RB	9.07
Group1_10_TF	9.00
Group2_10_RB	9.42
Group2_10_TF	9.29

Table 4: Average scores of the general text questions. TF stands for text field design and RB for the radio button design. 4, 8 and 10 are the payments per hour in US dollars.

Of interest are only pairs of experiments that differ in a single variable setting. Table 5 shows the t-test results for the maths questions and Table 6 for the text questions. We only report significant results.

From the maths questions results shown in Table 5 we see that the country of origin does not have any impact on the results. Our results show that there is no statistically measurable impact of the country of workers' origin on the quality of the results – we indicate this by "nil". In contrast to the country of origin the design and the payment do in some cases have a significant impact on the quality of the results. We can see in a few cases that when the radio button design is used the results can be significantly better compared to the results obtained with the choice of free text design. From the table we can also see that the payment incentives seem to have also a significant positive impact on the results. This does not confirm the findings of Mason and Duncan (2010) and Feng et al. (2009) who found that increased financial incentives improved the quantity, but not the quality, of work performed by participants. It was explained that workers who were paid more were no more motivated than workers paid less.

Table 6 illustrates the significance test results of the general text questions experiment. In contrast to the results of the maths questions experiments, the table shows some significant impact of the country of origin. At the lower end of the payment scale Group 1 workers produce significantly better results than Group 2 workers. On the other hand, the design did not significantly affect the workers' ability to answer the questions, except for one example as seen in the table. An explanation to this could be that the users do surf the Internet to answer the questions no matter what design we use. As in the maths questions the payment tends to be a significant factor in the quality of the results. Participants tend to make more effort in solving the questions when higher payments are made, where again and in contrast to Mason and Duncan (2010) and Feng et al. (2009), the quality did improve.

Experimental Pair
Impact of country of origin
nil
Impact of design
Group1_4_TF – Group1_4_RB
Group1_8_TF – Group1_8_RB
Group2_10_TF – Group2_10_RB
Impact of payment
Group2_4_TF – Group2_10_TF
Group2_8_RB – Group2_10_RB

Table 5: Results of the maths question. The significantly better (at level $p < 0.05$) results are on the right of "–". "nil" indicates the absence of any significantly different result. TF stands for text field design and RB for the radio button design. 4, 8 and 10 are the payments per hour in US dollars.

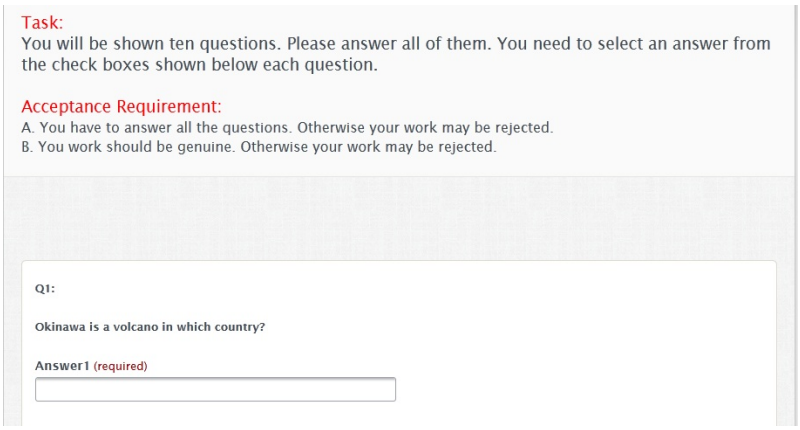

Figure 1: Text questions with free text design.

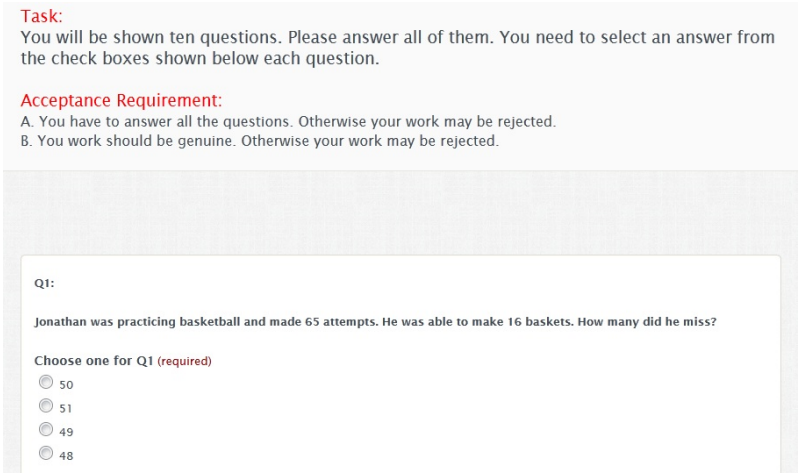

Figure 2: Maths questions with radio button design.

Experimental Pair
Impact of country of origin
Group2_4_TF – Group1_4_TF
Group2_8_RB – Group1_8_RB
Impact of design
Group2_8_RB – Group2_8_TF
Impact of payment
Group2_4_TF – Group2_8_TF
Group2_4_RB – Group2_10_RB
Group2_8_RB – Group2_10_RB
Group2_4_TF – Group2_10_TF

Table 6: Results of the general text question. The significantly better (at level $p < 0.05$) results are on the right of "–". "nil" indicates the absence of any significantly different result. TF stands for text field design and RB for the radio button design. 4, 8 and 10 are the payments per hour in US dollars.

5. Conclusion

Crowdsourcing has become a major phenomenon to address a growing number of problems. The creation of language resources is one such area and there is a huge potential in exploring the use of collective intelligence to build more resources. There is no doubt that expert quality can be achieved by aggregating the results obtained from a number of workers, e.g. (Snow et al., 2008). However, the exploration into what factors do (or do not) affect annotation quality has only just started.

In this paper we conducted a simple study to explore how different parameters in crowdsourcing such as payment levels and country of origin of workers affect the quality of results for two types of tasks. We experimented with objective tasks: maths and general text questions. We see this work as contributing to getting an overall picture of how different factors influence the quality of work produced by *individual* workers.

Our results indicate that in general higher payment is better when the aim is to obtain high quality results. In the maths questions the radio button design seems to lead to better results compared to the free text design. A qualitative analysis of the results shows however that incorrect answers are due to small imprecision rather than being completely wrong. We think that for any real task where high precision is required objective questions such as the maths questions in combination with the free text design can be used to filter out "unprecise workers". In the text questions the country of origin played an important factor for obtaining better results. It was shown that Group 1 workers performed better than the workers from Group 2. However, for maths questions the country did not play an important role.

An experiment on this scale obviously has a number of limitations. We see the results as a starting point for more extensive experimentation to uncover how different variables affect the quality of results obtained through crowdsourcing experiments.

There are a number of other important issues not discussed in this paper. First of all, we limited our experiments to the use of Mechanical Turk. There is a range of alternative crowdsourcing platforms that have recently been introduced. More importantly, we did not touch on the issue of ethicality which has started to attract more interest in the research community recently (Fort et al., 2011). Often it is not easy to distinguish whether workers contribute because they see it as a *"fruitful way to spend free time"* or whether they see this work as their *"primary source of income"* (Ipeirotis, 2010). One possible alternative to this dilemma is to collect judgements from players of online games, namely games with a purpose (GWAP) which involve no payment whatsoever and which have been shown to produce high-quality labels (von Ahn, 2006). One difficulty is to attract players into tasks that are perhaps not intuitively appealing such as linguistic tasks, e.g. Chamberlain et al. (2008). Using general knowledge and maths questions has limited the experimental work, as the provided questions might work well for carrying out the experiments but it does not explicitly reflect the use of crowdsourcing to obtaining linguistic judgments or performing natural language processing tasks.

Finally, any such experimental results are difficult to generalize as task will differ from each other (even if only slightly). Therefore we would also argue that comparisons of findings in this study with previous work can be difficult.

6. Future Work

As for the future work we are planning for additional experiments that will involve classification following the process of creating annotated resources. We are also planning to strengthen the work in the future by conducting additional experiments that will involve linguistic judgments. We are continuing our current work and aim to convert it to a framework for highlighting whether a worker is motivated or not. We believe that this framework can be used by other researchers to set up experiments and subsequently to obtain better and more accurate results for their experiments.

Acknowledgements

This research is part of the AutoAdapt research project. AutoAdapt is funded by EPSRC grants EP/F035357/1 and EP/F035705/1.

7. References

M-Dyaa Albakour, Udo Kruschwitz, and Simon Lucas. 2010. Sentence-level Attachment Prediction. In *Proceedings of the 1st Information Retrieval Facility Conference*, volume 6107 of *Lecture Notes in Computer Science*, pages 6–19, Vienna. Springer.

Omar Alonso and Stefano Mizzaro. 2009. Can we get rid of TREC Assessors? Using Mechanical Turk for Relevance Assessment. In *SIGIR '09: Workshop on The Future of IR Evaluation*.

Omar Alonso, Daniel E. Rose, and Benjamin Stewart. 2008. Crowdsourcing for relevance evaluation. *SIGIR Forum*, 42(2):9–15.

Vikas Bhardwaj, Rebecca Passonneau, Ansaf Salleb-Aouissi, and Nancy Ide. 2010. Anveshan: a framework for analysis of multiple annotators' labeling behavior. In *Proceedings of the Fourth Linguistic Annotation Workshop*, LAW IV '10, pages 47–55, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jon Chamberlain, Massimo Poesio, and Udo Kruschwitz. 2008. Phrase Detectives - A Web-based Collaborative Annotation Game. In *In Proceedings of I-Semantics*.

Wisam Dakka and Panagiotis G. Ipeirotis. 2008. Automatic extraction of useful facet hierarchies from text databases. In *ICDE*, pages 466–475. IEEE.

Pinar Donmez, Jaime Carbonell, and Jeff Schneider. 2009. Efficiently learning the accuracy of labeling sources for selective sampling. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 259–268, New York, NY, USA.

Mahmoud El-Haj, Udo Kruschwitz, and Chris Fox. 2010. Using Mechanical Turk to Create a Corpus of Arabic Summaries. In *Proceedings of the LREC Workshop on Semitic Languages*, pages 36–39, Valletta, Malta.

Donghui Feng, Sveva Besana, and Remi Zajac. 2009. Acquiring High Quality Non-expert Knowledge from On-demand Workforce. In *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, People's Web '09, pages 51–56, Morristown, NJ, USA. Association for Computational Linguistics.

Karën Fort, Gilles Adda, and K. Bretonnel Cohen. 2011. Amazon Mechanical Turk: Gold Mine or Coal Mine? *Computational Linguistics*.

Panagiotis G. Ipeirotis. 2010. Demographics of Mechanical Turk. *SSRN eLibrary*.

Michael Kaisser and John Lowe. 2008. Creating a Research Collection of Question Answer Sentence Pairs with Amazons Mechanical Turk. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Michael Kaisser, Marti A. Hearst, and John B. Lowe. 2008. Improving search results quality by customizing summary lengths.

Gabriella Kazai, Natasa Milic-Frayling, and Jamie Costello. 2009. Towards methods for the collective gathering and quality control of relevance assessments. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 452–459, New York, NY, USA. ACM.

Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with mechanical turk. In *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 453–456, New York, NY, USA. ACM.

Winter Mason and Duncan J. Watts. 2010. Financial Incentives and the "Performance of Crowds". *SIGKDD Explor. Newsl.*, 11:100–108, May.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and Fast—But is it Good?: Evaluating Non-expert Annotations for Natural Language Tasks. In *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Morristown, NJ, USA. Association for Computational Linguistics.

Alexander Sorokin and David Forsyth. 2008. Utility Data Annotation with Amazon Mechanical Turk. In *Proceedings of the First IEEE Workshop on Internet Vision at CVPR 08*, pages 1–8.

Qi Su, Dmitry Pavlov, Jyh-Herng Chow, and Wendell C. Baker. 2007. Internet-scale collection of human-reviewed data. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 231–240, New York, NY, USA. ACM.

Jiayu Tang and Mark Sanderson. 2010. Evaluation and user preference study on spatial diversity. In *ECIR*, volume 5993 of *Lecture Notes in Computer Science*, pages 179–190. Springer.

Luis von Ahn. 2006. Games With A Purpose. *IEEE Computer Magazine*, pages 96–98.

Paul Wais, Shivaram Lingamneni, Duncan Cook, Jason Fennell, Benjamin Goldenberg, Daniel Lubarov, David Marin, and Hari Simons. 2010. Towards Building a High-Quality Workforce with Mechanical Turk. In *NIPS Workshop on Computational Social Science and the Wisdom of Crowds*.

Yin Yang, Nilesh Bansal, Wisam Dakka, Panagiotis Ipeirotis, Nick Koudas, and Dimitris Papadias. 2009. Query by Document. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 34–43, New York, NY, USA. ACM.

Omar Zaidan and Chris Callison-Burch. 2011. Crowdsourcing Translation: Professional Quality from Non-Professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1220–1229, Portland, Oregon, USA. Association for Computational Linguistics.