Methods for Engaging and Evaluating Users of Human Computation Systems

Jon Chamberlain, Udo Kruschwitz, and Massimo Poesio

Introduction

One of the most significant challenges facing some Human Computation Systems is how to encourage participation on a scale required to produce high quality data. This is most relevant to systems where non-expert volunteers perform tasks, with the system aggregating the result. Issues relating to participant psychology are applicable to any system where humans (and subsequently human error) are involved.

The willingness of Web users to collaborate in the creation of resources is clearly illustrated by Wikipedia[1]: allowing users free reign of encyclopaedic knowledge not only empowers mass participation but the resulting creation is high quality. This can be seen as a good example of the broad term **collective intelligence** where groups of individuals do things collectively that seem intelligent (Malone et al. 2009).

The utility of collective intelligence became apparent when it was proposed to take a job traditionally performed by a designated employee or agent and outsource it to an undefined large group of Internet users through an open call. This approach, called **crowdsourcing** (Howe 2008), revolutionised the way traditional tasks could be completed and made new tasks possible that were previously inconceivable due to cost or labour limitations.

One use for crowdsourcing can be as a way of getting large amounts of human work hours very cheaply as an alternative to producing a computerised solution that may be expensive or complex. However, it may also be seen as a way of utilising human processing power to solve problems that computers, as yet, cannot solve, termed **human computation** as defined by von Ahn (2006).

[1] http://www.wikipedia.org

J. Chamberlain (✉) • U. Kruschwitz • M. Poesio
University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, England
e-mail: jchamb@essex.ac.uk; udo@essex.ac.uk; poesio@essex.ac.uk

P. Michelucci (ed.), *Handbook of Human Computation*,
DOI 10.1007/978-1-4614-8806-4_54,

An application of collective intelligence, crowdsourcing and human computation is to enable a large group of collaborators to work on tasks normally done by a few highly skilled (and paid) workers and to aggregate their work to produce a complex dataset that is robust and allows for ambiguity. Enabling groups of people to work on the same task over a period of time in this way is likely to lead to a collectively intelligent decision (Surowiecki 2005).

Using this method of collecting and aggregating decisions from a large, distributed group of non-expert contributors it is possible to approximate a single expert's judgements (Albakour et al. 2010; Feng et al. 2009; Snow et al. 2008).

User Motivation in Collaborative Systems

Three variations of collaboration over the Internet have been successful in recent years and are distinguished by the motivations of the participants.

1. The first variation is where the motivation for the users to participate already exists. This could be because the user is **inherently interested** in contributing, for example in the case of Wikipedia or citizen science projects such as GalaxyZoo[2] and Open Mind Commonsense[3] (now ConceptNet[4]). Users may also be intrinsically motivated because they need to accomplish a different task, for example the reCAPTCHA[5] authentication system.
2. As most tasks are neither interesting nor easy to integrate into another system, a second variation of crowdsourcing called **microworking** (or microtasking) was developed, for example Amazon Mechanical Turk.[6] Participants (sometimes called Turkers) are paid small amounts of money to complete HITs (Human Intelligence Tasks) uploaded by Requesters. The tasks can be completed very quickly, however this approach cannot be scaled up for large data collection efforts due to the cost.
3. A third approach for collecting and validating data used in human computation is to entertain the user whilst they complete the tasks, typically using games. The **games-with-a-purpose (GWAP)** approach has been used for many different types of crowdsourced data collection including text, image, video and audio annotation, biomedical applications, transcription, search results and social bookmarking (Chamberlain et al. 2013; Thaler et al. 2011).

[2] http://www.galaxyzoo.org

[3] http://openmind.media.mit.edu

[4] http://conceptnet.media.mit.edu

[5] http://www.google.com/recaptcha

[6] https://www.mturk.com

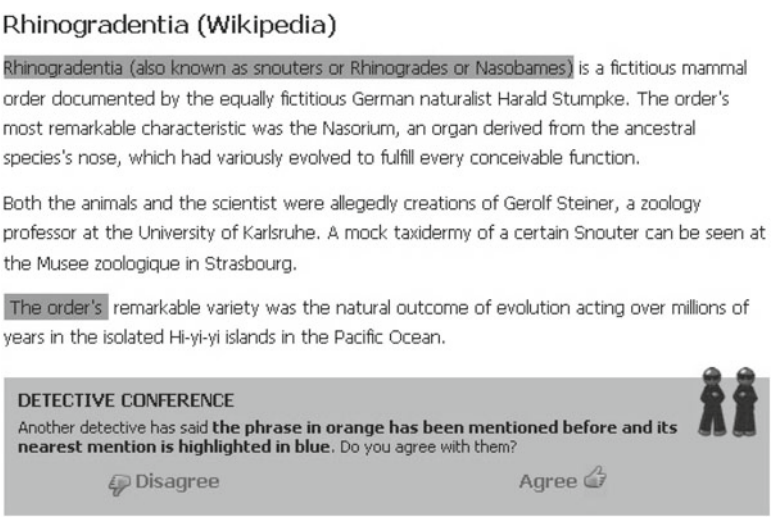

Fig. 1 Detail of a task in Phrase Detectives

There is huge potential for the general public to become engaged in Human Computation Systems and to collaborate in producing resources that would not be possible to achieve using other methods.

This chapter discusses methods that can be used to motivate and engage users. As an example, we look at how these methods were used in Phrase Detectives,[7] a Human Computation System developed by the University of Essex (England) to annotate text documents with a crowd. The conclusion summarises the benefits and limitations of using such methods in Human Computation Systems.

Phrase Detectives

Phrase Detectives (PD) is primarily a GWAP designed to collect data about English (and subsequently Italian) anaphoric co-reference (Chamberlain et al. 2008; Poesio et al. 2013).[8]

The architecture is structured around a number of tasks that use scoring, progression and a variety of other mechanisms to make the activity enjoyable (see Fig. 1).

[7] http://www.phrasedetectives.com

[8] Anaphoric coreference is a type of linguistic reference where one expression depends on another referential element. An example would be the relation between the entity *'Jon'* and the pronoun *'his'* in the text *'Jon rode his bike to school.'*

The aim of the project is not only to annotate large amounts of text, but also to collect a large number of judgements about each linguistic expression to preserve ambiguity that can be used to improve language processing algorithms.

A version of PD was developed for Facebook[9] in order to investigate the utility of social networking sites in collaborative annotation systems.

Methods to Engage and Evaluate Users

There have been several recent attempts to define and classify collaborative approaches in collective intelligence and distributed human computation (Quinn and Bederson 2011; Malone et al. 2009; Wang et al. 2010). We focus on four main areas:

1. Designing the Task
2. Attracting Users
3. Motivating Users
4. Evaluating Users

Designing the Task

Whilst design considerations can be somewhat generalised, it is worth noting a fundamental challenge for human computation systems. The goal here is to **collect data and reward users without directly knowing the quality of their work** (either by the system knowing the answer beforehand or by manual correction after the data is collected). Methods for motivating users without being able to provide specific feedback are discussed in detail later in the chapter.

Using an Appropriate Interface for Your Users

When designing any interface it is essential to **know your target audience**. Individual, social and socio-technical factors will all determine how successful the interface is at engaging users and what type of data will be contributed.

Wikipedia style open interfaces will invite a different type of user experience than a microworking or gaming approach and the expectations of the users need to be met in order for them to continue using the interface. Game interfaces should be graphically rich, although not at the expense of usability, and aimed at engaging a specific audience (i.e., a game aimed at children may include more cartoon or stylised imagery in brighter colours than a game aimed at adults).

[9] http://www.facebook.com

Interfaces should **provide a consistent metaphor and work flow**. For this PD used a detective metaphor, with buttons stylised with a cartoon detective character and site text written as if the player was a detective solving cases. The tasks should be integrated in such a way that task completion, user evaluation and work flow form a seamless experience.

Interfaces deployed on the Web should observe the normal guidelines regarding browser compatibility, download times, consistency of performance, spatial distance between click points, etc.[10]

Designing the Tasks

Whilst the design of the interface is important, it is the design of the task that determines how successfully the user can contribute data. The task design has an impact on the speed at which users can complete tasks, with clicking being faster than typing. For example, a design decision to use radio buttons or freetext boxes can have a significant impact on performance (Aker et al. 2012).

In PD the player is constrained to a set of predefined options to make annotations, with freetext comments allowed (although this is not the usual mode of contribution in the game). The pre-processing of text allows the game play in PD to be constrained in this way but is subject to errors in processing that also need to be fixed.

Considering Task Difficulty

The inherent difficulty of the task can provide a challenge to more experienced users and they need to be motivated to rise to the challenge of difficult tasks.

There is a clear difference in the performance of users when we consider the difficulty of tasks in GWAP (Chamberlain et al. 2009a). One way to measure this is to **use a Gold Standard** (a set of tasks that you have the answers for) or to **use inter-annotator agreement** that is created by experts or by the users themselves.

PD compared the responses from 2 experts across a range of tasks and found that they mostly agreed with each other (average 94 %). When comparing the responses produced by users of the game, the agreement would be in a similar range to expert agreement for simple tasks (average 90 %) but much lower for more difficult tasks (average 71 %) (Chamberlain et al. 2009a).

Setting Time Limits

A time limitation will elicit spontaneous answers from users, whereas no limitations gives users time to make a more considered response. The design of the task must

[10] http://www.usability.gov/guidelines

balance the increase in excitement a timed element can offer with the need to allow users time to give good quality answers.

The timing of tasks is usually required in the game format, either as a motivational feature or as a method of quality control (or both) (von Ahn and Dabbish 2008). In PD there are no timing constraints, although the time taken to perform a task is used to assess the quality of annotations. As the task in PD is text based, it was considered important to give players time to read documents at a relatively normal speed whilst completing tasks and this was confirmed by usability studies of the interface.

Measuring System Performance

System performance can be measured by the speed at which the users can process the input source (e.g. text, images) and deliver their response (e.g. a click, typing). This measure is called **throughput**, the number of labels (or annotations) per hour (von Ahn and Dabbish 2008). As well as measuring how well the task is presented in the interface, throughput is also an indication of task difficulty and cognitive load on the users.

Related to throughput is the **wait time** for tasks to be done. Most crowdsourcing systems allow data collection in parallel (i.e., many participants can work at once on the same tasks), although validation requires users to work in series (i.e., where one user works on the output of another user). Whilst the throughput gives us a maximum speed from the system, it is worth bearing in mind that the additional time spent waiting for a user to be available to work on the task may slow the data collection. Some systems deployed on Amazon Mechanical Turk pay workers a small retainer to be act as an on demand workforce (Bernstein et al. 2012).

Attracting Users

In order to attract the number of participants required to make a success of the system, it is not enough to develop an attractive interface; it is also necessary to develop effective forms of advertising. The number of websites competing for attention is huge and without some effort to raise the profile, it will never catch the attention of enough users.

Advertising

Not all advertising methods are equally successful and it is important to evaluate which works best for the task interface, delivery platform and target audience demographics. Traditional banner or pay-per-click advertising may go some way to attracting users, however in a rapidly changing landscape of Internet habits it would be worth investigating novel methods of delivery. For example, with a system that produces lots of content a dynamic and active Facebook news feed would engage more users in a social network rather than a static banner advert.

PD had a modest budget for pay-per-click advertising and considerable effort was made to promote the project in local and national press, on science websites, blogs, bookmarking websites, gaming forums, special interest email lists, conferences, tutorials and workshops.

The importance of promoting an interface should not be underestimated and **an advertising budget (both time and money) should be allocated** at an early stage.

The success of advertising methods can be analysed with user tracking tools such as Google Analytics.[11] This can be used to not only investigate the most successful venues for advertising to your audience, but also to analyse their behaviour when they come to your site. A useful figure is the bounce rate (the percentage of single-page visits, where the user leaves on the page they entered on) which shows how many casual users are being converted to users of the interface. Analysis of PD traffic data showed that Facebook pay-per-click banner adverts had a very high bounce rate (90 %), meaning that 9 out of 10 users that came from this source did not play the game. For this reason advertising budget was redirected to other sources of users.

Using Social Networks

Given the social nature of Human Computation it seems logical to deploy systems on platforms where the users are already networked. In recent years social networking has become the dominant pastime online. As much as 22 % of time online is spent on social networks like Facebook, Twitter and others. This is three times the amount of time spent emailing and seven times the amount of time spent searching the Internet.[12]

The success of social network games such as Cityville, with over 50 million active players each month, or The Sims, Farmville and Texas HoldEm Poker, with over 30 million active monthly players each, show that the potential for large scale participation is possible using social networking platforms.[13]

Social incentives can be made more effective when the interface is embedded within a social networking platform such as Facebook. In such a setting, users motivated by the desire to contribute to a communal effort may share their efforts with their friends, whereas those motivated by a competitive spirit can compete against each other. Surveys have shown that the majority of social game players start to play because of a friend recommendation.[14, 15]

[11] http://www.google.co.uk/analytics

[12] http://mashable.com/2010/08/02/stats-time-spent-online

[13] http://www.appdata.com

[14] http://www.infosolutionsgroup.com/2010_PopCap_Social_Gaming_Research_Results.pdf

[15] http://www.lightspeedresearch.com/press-releases/it's-game-on-for-facebook-users

Motivating Users

There are three main incentive structures that can be used to motivate users: personal; social; and financial (Chamberlain et al. 2009b). These directly relate to other classifications of motivations in previous research: Love; Glory; and Money (Malone et al. 2009). All incentives should be applied with caution as rewards have been known to decrease annotation quality (Mrozinski et al. 2008).

It is important to distinguish between **motivation to participate** (why people start doing something) and **motivation to contribute** (why they continue doing something) (Fenouillet et al. 2009). Once both conditions are satisfied we can assume that a user will continue contributing until other factors such as fatigue or distraction break the cycle. This has been called **volunteer attrition**, where a user's contribution diminishes over time (Lieberman et al. 2007).

Personal Incentives

Personal incentives are evident when simply participating is enough of a reward for the user. Generally, the most important personal incentive is that the user feels they are contributing to a worthwhile project; however personal achievement and learning can also be motivating factors.

Projects may initially attract collaborators because they are contributing to a resource from which they may directly benefit and these are usually the people that will be informed first about the research. However, in the long term, most contributors will never directly benefit from the resources being created. It is therefore essential to provide some more generic way of expressing the benefit to the user.

This was done in PD with a BBC radio interview by giving examples of natural language processing techniques used for Web searching. Although this is not a direct result of the language resources being created by the project, it is the case for efforts of the community as a whole, and this is what the general public can understand and be motivated by.

People who contribute information to Wikipedia are motivated by personal reasons such as the desire to make a particular page accurate, or the pride in one's knowledge in a certain subject matter (Yang and Lai 2010). This motivation is also behind the success of *citizen science* projects, such as the Zooniverse collection of projects (Raddick et al. 2010) (see also the chapter on citizen science participation by Reed, et al.), where the research is conducted mainly by amateur scientists and members of the public.

When users become more interested in the purpose of the project than the incentives it becomes more like a citizen science approach where users are willing to work on harder tasks, provide higher quality data and contribute more.

Social Incentives

Social incentives reward users by improving their standing amongst their peers (their fellow users and friends). By tracking the user's effort they can compete in leaderboards and see how their efforts compare to their peers. Assigning named levels for points awarded for task completion can be an effective motivator, with users often using these as targets i.e., they keep working to reach a level before stopping (von Ahn and Dabbish 2008), however results from PD do not support this (Chamberlain et al. 2012).

News feed posts are a simple way users can make social interactions from an interface that is integrated into social networks such as Facebook or Twitter. PD allows its players to make an automatically generated post to their news feed which will be seen by all of their friends.[16]

These posts include a link back to the game and has been a very important factor in recruiting more users, as well as motivating existing users by social incentives.

Financial Incentives

Financial incentives reward effort with money. Direct financial incentives reward the user for the completion of a task or for successfully competing against other users (for example, achieving a high score). The former is the main method of motivating users of microworking systems. The per-task reward however may encourage users to manipulate the system, to do minimum work for maximum reward.

Indirect financial incentives reward the user irrespective of the work they have done such as entering each completed task into a lottery where the winner is randomly selected (although doing more tasks would increase your chance of winning).

In PD and other games indirect financial incentives were sent as Amazon vouchers by email to the winners as this allows the prize to be invoiced, tracked and collected with minimum administrative effort.

Whilst financial incentives seem to go against the fundamental idea behind GWAP (i.e., that enjoyment is the motivation), it actually makes the enjoyment of potentially winning a prize part of the motivation. Prizes for high scoring players will motivate hard working or high quality players but the prize soon becomes unattainable for the majority of other players. By using a lottery style financial prize the hard working players are more likely to win, but the players who only do a little work are still motivated. Prize-based financial incentives present a risk that not enough work will be collectively done by the conclusion of the prize period, however if the users are correctly motivated it should prove much more cost-effective than pay-per-task incentives.

[16] Since the initial development of PD Facebook has changed how posts are displayed. Posts from the game now appear on the user's profile and in a news ticker.

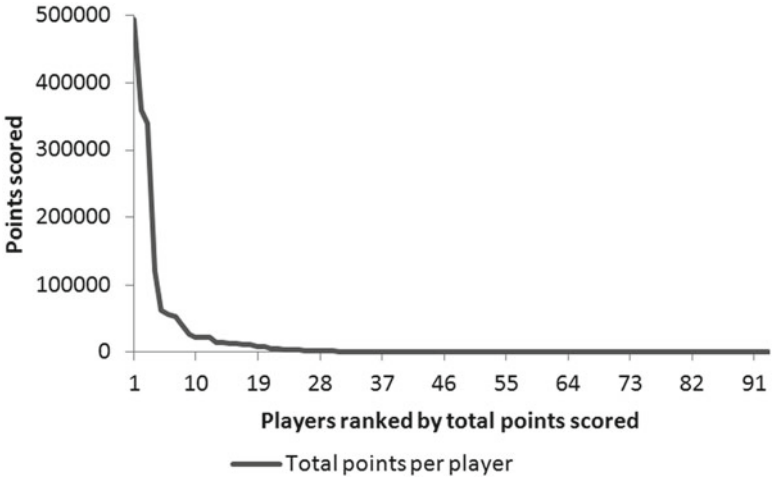

Fig. 2 Chart showing each player on the x-axis ranked by total points scored (approximately equivalent to workload) in Phrase Detectives

Whilst financial incentives are important to recruit new users, a combination of all three types of incentives is essential for the long term success of a project (Smadja 2009).

Evaluating Participation and Contribution

We can measure the success of advertising and the motivation to join the project (motivation to participate) by how many users have registered over the period of time. However, this may not be a good predictor of how much work will be done, how fast it will be completed or of what quality it will be.

Participation of users to contribute is a way to assess whether the incentives of an approach are effective. We measure motivation to contribute by the average lifetime participation.

One observation that is apparent in most crowdsourcing systems is the uneven distribution of contribution per person, often following a Zipfian power law curve—see Fig. 2 (Chamberlain et al. 2012).

An approach to improve data quality would be to focus training and incentives on the few users that are contributing significantly. However, the influence of users who only contribute a little should not be undervalued as in some systems it can be as high as 30 % of the workload (Kanefsky et al. 2001) and this is what makes the collective decision making robust. Increasing the participation from the "long tail" is key to improving the quality of the human computation.

Evaluating Users

The strategies for quality control address five main issues:

1. Training Users
2. Reducing Genuine Mistakes
3. Allowing for Genuine Ambiguity
4. Controlling Malicious Behaviour
5. Identifying Outliers

Training Users

A training stage is usually required for users to practice the task and to show that they have sufficiently understood the instructions to do a real task. The task design needs to **correlate good user performance with producing good quality data**. The level of task difficulty will drive the amount of training that a user will need and the training phase has been shown to be an important factor in determining quality and improvement in manual annotation (Dandapat et al. 2009).

Training should assume a layman's knowledge of the task and should engage the participant to increase their knowledge to become a pseudo-expert. The more they participate, the more expert they become. This graduated training makes a rating system (where the user is regularly judged against a gold standard) essential to give appropriately challenging tasks.

Most projects, at least initially, will have a core of collaborators to test and perform tasks and these are most likely to be friends or colleagues of the task designers. It can therefore be assumed that this base of people will have prior knowledge of the task background, or at least easy access to this information. These pre-trained collaborators are not the "crowd" that crowdsourcing needs if it is to operate on a large scale nor are they the "crowd" in the wisdom of the crowd.

Reducing Genuine Mistakes

Users may occasionally make a mistake and press the wrong button. Attention slips need to be identified and corrected by validation, where users can examine other users' work and evaluate it. Through validation, poor quality interpretations should be voted down and high quality interpretations should be supported (in the cases of genuine ambiguity there may be more than one). The validation process is a second stage to the data collection, that allows the task to be more varied, to make the data collection more efficient (validation is only required when there is disagreement) and to create a sense of user community and responsibility. Validation thus plays a key role as a strategy for quality control.

Unlike open collaboration in Wikipedia, it is not advisable to allow players of GWAP to go back and correct their mistakes, otherwise a player could try all possible variations of an answer and then select the one offering the highest score. In this sense the way players work together is more "collective", where individual work is aggregated after collection, than "collaborative", where users work more directly with each other.

Allowing for Genuine Ambiguity

The strength of Human Computation Systems is the ability to capture ambiguity in the data. Systems should not only aim to select the best, or most common, answer or annotation from users but also to preserve all inherent ambiguity, leaving it to subsequent processes to determine which interpretations are to be considered spurious and which instead reflect genuine ambiguity.

Collecting multiple judgements about linguistic expressions is a key aspect of PD. In the current configuration, eight players are asked to express their judgements on a task. If they do not agree on a single interpretation, four more players are then asked to validate each interpretation.

Validation has proven very effective at identifying poor quality interpretations. The value obtained by combining the player annotations with the validations for each interpretation tends to be zero or negative for all spurious interpretations.

Controlling Malicious Behaviour

Controlling cheating may be one of the most important factors in Human Computation System design. All crowdsourcing systems attract spammers, which can be a very serious issue (Feng et al. 2009; Mason and Watts 2009; Kazai 2011). However, in a game context we can expect spamming to be much less of an issue as the work is not conducted on a pay-per-task basis.

Nevertheless, several methods are used in PD to identify players who are cheating or who are providing poor annotations. These include checking the player's IP address (to make sure that one player is not using multiple accounts), checking annotations against known answers (the player rating system), preventing players from resubmitting decisions (Chklovski and Gil 2005) and keeping a blacklist of players (von Ahn 2006).

A method of profiling players was also developed for PD to detect unusual behaviour. The profiling compares a player's decisions, validations, skips, comments and response times against the average for the entire game—see Fig. 3. It is very simple to detect players who should be considered outliers using this method (this may also be due to poor task comprehension as well as malicious input) and their data can be ignored to improve the overall quality.

	System	Good player	Bad player
ANNOTATIONS			
Total Annotations:	1423078	4587	11018
Average Annotation Time:	00:00:07	00:00:07	00:00:04
Total (Ratio) DN:	955520 (0.67)	1495 (0.33)	10935 (0.99)
Total (Ratio) DO:	378256 (0.27)	2696 (0.59)	58 (0.01)
Total (Ratio) PR:	79172 (0.06)	334 (0.07)	24 (0)
Total (Ratio) NR:	13395 (0.01)	64 (0.01)	2 (0)
VALIDATIONS			
Total Validations:	608982	3848	5256
Total (Ratio) Agree:	200174 (0.33)	1186 (0.31)	8 (0)
Ave Agree Time:	00:00:09	00:00:08	00:00:18
Total (Ratio) Disagree:	408808 (0.67)	2662 (0.69)	5248 (1)
Ave Disgree Time:	00:00:08	00:00:07	00:00:02
OTHER			
Total Skips:	51616	142	26
Skip per annotation:	0.04	0.03	0
Total Comments:	26593	229	0
Comment per annotation:	0.02	0.05	0

Fig. 3 Player profiling in Phrase Detectives, showing the game totals and averages (*left*), a good player profile (*centre*) and a bad player profile (*right*) taken from real game profiles. The bad player in this case was identified by the speed of annotations and that the only responses were DN in Annotation Mode and Disagree in Validation Mode. The player later confessed to using automated form completion software

Identifying Outliers

It would be possible to ignore contributions from users who have a low rating (judged against a gold standard) however without a gold standard it is difficult to judge the performance of a user.

Variables such as annotation time could be a factor in filtering the results. An annotation in PD takes between 9 and 11 seconds and extreme variation from this may indicate that a poor quality decision has been made.

A different approach could be to identify those users who have shown to provide high quality input. A knowledge source could be created based on input from these users and ignore everything else. Related work in this area applies ideas from citation analysis to identify users of high expertise and reputation in social networks by, for example, adopting the HITS algorithm (Yeun et al. 2009) or Google's PageRank (Luo and Shinaver 2009).

Conclusion

This chapter discussed methods that can be used to engage, motivate and evaluate users of crowdsourced Human Computation Systems.

Interfaces should be attractive enough to encourage users to contribute. The design of the task itself will be determined in part by the complexity of the data being collected. By identifying the difficult or ambiguous tasks, the pre- and post-processing can be improved and the human input can be maximised to produce the highest quality resource possible given the inherent difficulty of the task. The task design should be streamlined for efficient collection of data and the throughput (annotations per hour) of the system is a good measure of this. The additional time spent waiting for a user to be available to work on the task may also slow the system.

Most users will not benefit directly from their participation, however their connection to the project and sense of contribution to science are strong motivating factors with the citizen science approach, where users are willing to work on harder tasks, provide higher quality data and contribute more. Motivational issues are less of a concern when users are intrinsically motivated to participate, as they will directly benefit from their contribution.

It is common for the majority of the workload to be done by a minority of users. Motivating the right kind of users is a complex issue and is as important as attracting large numbers of users. Controlling cheating may be one of the most important factors in crowdsourcing design and is especially problematic for a microworking approach where users are paid on a per-task basis.

The issue of data quality is an area of continuous research. The ultimate goal is to show that resources created using Human Computation Systems potentially offer higher quality and are more useful by allowing for ambiguity. By quantifying the complexity of the tasks, human participants can be challenged to solve computationally difficult problems that would be most useful to machine learning algorithms.

Acknowledgements The original Phrase Detectives game was funded as part of the EPSRC AnaWiki project, EP/F00575X/1.

References

Aker A, El-Haj M, Albakour D, Kruschwitz U (2012) Assessing crowdsourcing quality through objective tasks. In: Proceedings of LREC'12, Istanbul

Albakour M-D, Kruschwitz U, Lucas S (2010) Sentence-level attachment prediction. In: Proceedings of the 1st information retrieval facility conference. Volume 6107 of lecture notes in computer science, Vienna. Springer, pp 6–19

Bernstein MS, Karger DR, Miller RC, Brandt J (2012) Analytic methods for optimizing realtime crowdsourcing. CoRR

Chamberlain J, Poesio M, Kruschwitz U (2008) Phrase detectives: a web-based collaborative annotation game. In: Proceedings of the international conference on semantic systems (I-Semantics'08), Graz

Chamberlain J, Kruschwitz U, Poesio M (2009a) Constructing an anaphorically annotated corpus with non-experts: Assessing the quality of collaborative annotations. In: Proceedings of the 2009 workshop on the people's web meets NLP: collaboratively constructed semantic resources, Singapore

Chamberlain J, Poesio M, Kruschwitz U (2009b) A new life for a dead parrot: incentive structures in the phrase detectives game. In: Proceedings of the WWW 2009 workshop on web incentives (WEBCENTIVES'09), Madrid

Chamberlain J, Kruschwitz U, Poesio M (2012) Motivations for participation in socially networked collective intelligence systems. In: Proceedings of CI2012, Boston

Chamberlain J, Fort K, Kruschwitz U, Mathieu L, Poesio M (2013) Using games to create language resources: successes and limitations of the approach. In: ACM transactions on interactive intelligent systems, volume The People's Web Meets NLP: collaboratively constructed language resources. Springer pp 3–44

Chklovski T, Gil Y (2005) Improving the design of intelligent acquisition interfaces for collecting world knowledge from web contributors. In: Proceedings of K-CAP '05, Banff

Dandapat S, Biswas P, Choudhury M, Bali K (2009) Complex linguistic annotation – no easy way out! a case from Bangla and Hindi POS labeling tasks. In: Proceedings of the 3rd ACL linguistic annotation workshop, Singapore

Feng D, Besana S, Zajac R (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, Singapore

Fenouillet F, Kaplan J, Yennek N (2009) Serious games et motivation. In: 4eme conference francophone sur les environnements informatiques pour l'apprentissage humain (EIAH'09), vol. Actes de l'Atelier "Jeux Serieux: conception et usages", Le Mans

Howe J (2008) Crowdsourcing: why the power of the crowd is driving the future of business. Crown Publishing Group, New York

Kanefsky B, Barlow N, Gulick V (2001) Can distributed volunteers accomplish massive data analysis tasks? In: Lunar and planetary science, XXXII, Houston

Kazai G (2011) In search of quality in crowdsourcing for search engine evaluation. In: Proceedings of the 33^{rd} european conference on information retrieval (ECIR'11), Dublin

Lieberman H, Smith DA, Teeters A (2007) Common consensus: a web-based game for collecting commonsense goals. In: Proceedings of IUI, Honolulu

Luo X, Shinaver J (2009) MultiRank: reputation ranking for generic semantic social networks. In: Proceedings of the WWW 2009 workshop on web incentives (WEBCENTIVES'09), Madrid

Malone T, Laubacher R, Dellarocas C (2009) Harnessing crowds: mapping the genome of collective intelligence. Research Paper No. 4732-09, Sloan School of Management, Massachusetts Institute of Technology, Cambridge

Mason W, Watts DJ (2009) Financial incentives and the "performance of crowds". In: Proceedings of the ACM SIGKDD workshop on human computation, Paris

Mrozinski J, Whittaker E, Furui S (2008) Collecting a why-question corpus for development and evaluation of an automatic QA-system. In: Proceedings of ACL-08: HLT, Columbus

Poesio M, Chamberlain J, Kruschwitz U, Robaldo L, Ducceschi L (2013) Phrase detectives: utilizing collective intelligence for internet-scale language resource creation. ACM transactions on interactive intelligent systems 3:1–44

Quinn A, Bederson B (2011) Human computation: a survey and taxonomy of a growing field. In: CHI, Vancouver

Raddick MJ, Bracey G, Gay PL, Lintott CJ, Murray P, Schawinski K, Szalay AS, Vandenberg J (2010) Galaxy zoo: exploring the motivations of citizen science volunteers. Astronomy Educ Rev 9(1):010103

Smadja F (2009) Mixing financial, social and fun incentives for social voting. World Wide Web Internet And Web Information Systems

Snow R, O'Connor B, Jurafsky D, Ng AY (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, Honolulu

Surowiecki J (2005) The wisdom of crowds. Anchor, New York

Thaler S, Siorpaes K, Simperl E, Hofer C (2011) A survey on games for knowledge acquisition. Technical report STI TR 2011-05-01, Semantic Technology Institute

von Ahn L (2006) Games with a purpose. Comput 39(6):92–94

von Ahn L, Dabbish L (2008) Designing games with a purpose. Commun ACM 51(8):58–67

Wang A, Hoang CDV, Kan MY (2010) Perspectives on crowdsourcing annotations for natural language processing. Language resources and evaluation, pp 1–19

Yang H, Lai C (2010) Motivations of wikipedia content contributors. Comput Hum Behav 26:1377–1383

Yeun CA, Noll MG, Gibbins N, Meinel C, Shadbolt N (2009) On measuring expertise in collaborative tagging systems. In: Proceedings of WebSci'09, Athens