...

Jon Chamberlain*, Udo Kruschwitz, and Massimo Poesio

Optimising Crowdsourcing Efficiency: Amplifying Human Computation with Validation

DOI ..., Received ...; accepted ...

Abstract: Crowdsourcing has revolutionised the way tasks can be completed but the process is frequently inefficient, costing practitioners time and money. This research investigates whether crowdsourcing can be optimised with a validation process, as measured by four criteria: quality; cost; noise; and speed. A validation model is described, simulated and tested on real data from an online crowdsourcing game to collect data about human language. Results show that by adding an agreement validation (or a like/upvote) step fewer annotations are required, noise and collection time are reduced and quality may be improved.

Keywords: Crowdsourcing; Empirical studies in interaction design; Interactive games; Social networks; Natural language processing

PACS: ...

Communicated by: ...
Dedicated to ...

1 Introduction

Crowdsourcing [10] has revolutionised the way traditional tasks can be completed and made new tasks possible that were previously inconceivable due to cost or labour limitations. More specifically, distributed human intelligence tasking [5] combines collective intelligence, crowdsourcing and human computation to enable a large group of collaborators to work on tasks normally completed by highly-

***Corresponding author: Jon Chamberlain, Udo Kruschwitz,** Univeristy of Essex, Wivenhoe Park, Colchester CO4 3SQ UK
Massimo Poesio, Queen Mary University, Mile End Rd, London E1 4NS UK

skilled (and highly-paid) annotators and aggregates their collective answers to produce a more complex dataset that is robust and allows for ambiguity.

Several approaches to this type of crowdsourcing have been successful: in **peer production** [2] users are self-organised and inherently interested in contributing towards a shared outcome, such as Wikipedia[1]; in **microworking**[2] participants are paid small amounts of money per task, for example Amazon Mechanical Turk[3]; and a third approach is to entertain the user whilst they complete tasks, typically using games or gamification. This **game-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 and social bookmarking [13].

Such methods may require users to complete tasks preset by an administrator or organisation (called a ‘requester’ in microworking); however, the problem-solving abilities of a crowd can also been seen in **Community Question Answering (cQA)** websites such as StackOverflow[4] in which an active online community present and resolve problems without a central administrative structure. Similarly, social networks such as Facebook[5] are being used to organise data, to pose problems, and to connect with people who may have solutions [6, 18], and therefore could be viewed as a type of crowdsourcing system.

Requesters using crowdsourcing approaches that collect data from users can filter, aggregate and check for quality; however, this process can also be crowdsourced. A validation step exists in a number of system workflows; however, more commonly this is seen in social networks and cQAs that feature ‘liking’ or ‘upvoting’ of content which can also be viewed as validation. ‘Liking’ as a social media activity is one of the most common forms of activities on social networks [8].

This paper tests a validation model using data from a crowdsourcing game to collect information about human language to discover whether a validation stage is more efficient than simply adding more annotations and what might the optimal configuration be to reduce noise and increase efficiency. Applications for this approach beyond traditional crowdsourcing systems is discussed, in particular its use in analysing social network data.

[1] https://en.wikipedia.org/wiki/Main_Page

[2] https://www.economist.com/blogs/schumpeter/2011/04/digital_economy

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

[4] https://stackoverflow.com

[5] https://www.facebook.com

2 Related work

Data collection approaches not only provide ways for the user to submit information but also describe how data is aggregated in some way to produce a best answer, or a set of plausible answers, to the task. The goal of aggregation is to use the contributions to approximate a single expert's answer, although crowd-created data allow for more complex probabilistic answer sets to be created. For example, in majority voting, given a finite set of things to choose from, the highest-voted is the best answer. Repeated-labelling is a technique based on majority voting that takes uncertainty into account and is useful for estimating when an answer is good enough [17].

Weighted voting is similar to majority voting, but each vote is adjusted (or weighted) so that people who are most influential, most capable to answer or most popular (implemented differently in different systems depending on the output priorities) have more impact on the final decision [11]. The superuser reputation scoring model in the social gaming network *Foursquare*[6] hints at the commercial interest in weighting user contributions, and similar models are employed by other crowd-based datasets such as Stack Overflow [4].

The statistical probability of getting a correct solution shows that the worker rating (the assessed ability of the worker to provide the correct answer) will determine how many annotations you need per task. If we require a 99% probability of getting a correct solution from the workers and if each worker has a 90% chance of submitting a correct answer, only two annotations are needed. If the workers' ratings are less, say 70% chance, then four annotations are needed, and if less again at 50% then seven annotations are needed. A crowd with an average lower than 50% chance will take considerably more annotations.

This does not account for the variability in worker abilities, the order in which workers submit answers, the difficulty of the task, the possibility of having multiple correct answers or other confounding factors. However, it is important to estimate the number of annotations that are required; too few annotations and the correct solution for the task might not be discovered; too many annotations and the data collection will take longer than necessary, cost more and introduce more noise (incorrect solutions) that need to be filtered out.

Researchers investigating single-tier crowdsourcing systems, typified by microworking, make the assumption that if an answer is possible from the crowd then getting lots of annotations or labels, whilst applying filtering, will eventually lead to the best answer [19]. In some cases this may prove to be the case; however,

[6] http://engineering.foursquare.com/2014/01/03/the-mathematics-of-gamification

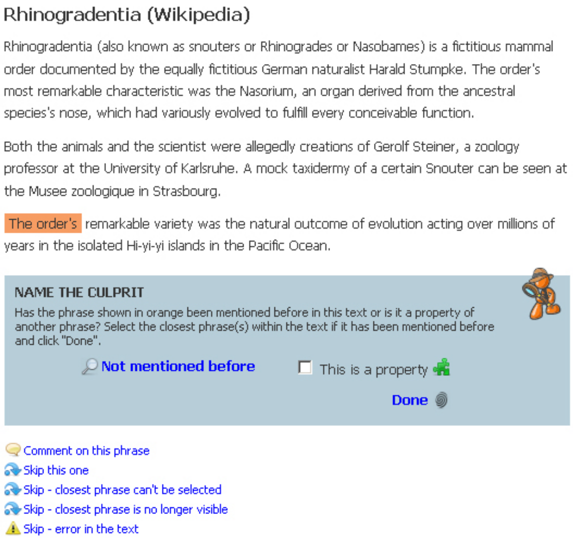
the caveat of getting more annotations is the chance of getting a more diverse range of answers or noise, from which the true answer cannot be extracted. There has been considerable research into optimising the data collection process by determining the most appropriate point to stop collecting data (based on the trade-off between cost, speed and quality), for example [9]. Additionally, attention has focused on the workflow of complex problems on microworking sites, for example [12, 14]. Such efforts have tested directive techniques on microworking sites but there has been little crossover of techniques into other crowdsourcing approaches nor simulations of data collection using large, undirected datasets.

2.1 Crowdsourcing with Validation

Validation of data usually occurs after data has been collected; the issue is whether those validations are part of the process that the workers are involved in, or whether it is a form of checking from the requester to ensure that a sample of the annotations is of a high enough quality. In systems such as Wikipedia, social networks and cQA, the verification of solutions is performed by the workers themselves. GWAP and microworking data are typically validated by the requester; however, some systems do use validation as an additional worker task to reduce the workload for the requester.

One example of a validation task, where the worker sees the solutions from the previous worker(s) and agrees with it or not, is seen in the Find-Fix-Verify approach, implemented in the crowd-based word processor called *Soylent* that enabled editing and summarising of text by the crowd [3]. The process breaks complex editing tasks into generative and review stages incorporating voting to produce a final result. In the *find* stage the users identify a section of text that needs work, in the *fix* stage users are asked to improve on the text and in the final *verify* stage the users vote on which improved text they prefer (or keep the original text).

The fundamental idea behind using validation as a supporting mechanism for crowdsourcing workflows is that it should be easier and faster for the worker to decide if a solution is correct rather than create a solution from scratch. An agreeing validation can be seen as another annotation in favour of the solution (if using majority voting to determine the best answer). A disagreeing validation provides less information, in that the worker is saying what the correct interpretation is not, rather than what it is. An agreeing validation says what the interpretation is and by inference what it is not (if we assume there is only one correct or best answer).



Rhinogradentia (Wikipedia)

Rhinogradentia (also known as snouters or Rhinogrades or Nasobames) is a fictitious mammal order documented by the equally fictitious German naturalist Harald Stumpke. The order's most remarkable characteristic was the Nasorium, an organ derived from the ancestral species's nose, which had variously evolved to fulfill every conceivable function.

Both the animals and the scientist were allegedly creations of Gerolf Steiner, a zoology professor at the University of Karlsruhe. A mock taxidermy of a certain Snouter can be seen at the Musee zoologique in Strasbourg.

The order's remarkable variety was the natural outcome of evolution acting over millions of years in the isolated Hi-yi-yi islands in the Pacific Ocean.

NAME THE CULPRIT

Has the phrase shown in orange been mentioned before in this text or is it a property of another phrase? Select the closest phrase(s) within the text if it has been mentioned before and click "Done".

Fig. 1: A task presented in Annotation Mode.

A validation step can be added in two ways: either synchronous, in which validation is completed after an initial data collection stage is complete, or asynchronous in which the task is annotated and validated together, such as a conversation thread on cQA or social networks.

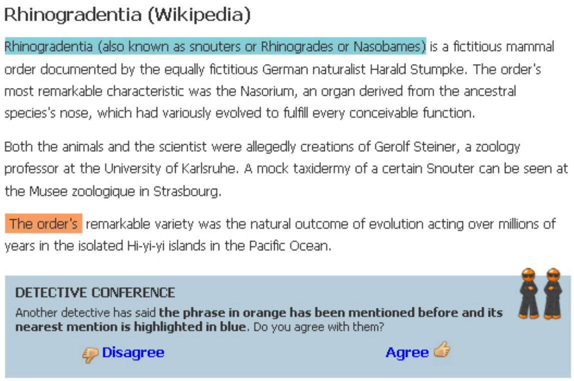
The question this research investigates is whether we can improve crowdsourcing efficiency (reduce the human effort required) and improve the final data quality by adding a crowdsourced validation stage.

3 Method

Phrase Detectives[7] is an online crowdsourcing game with a validation stage, primarily designed to collect data about English (and subsequently Italian) anaphoric co-reference [15].[8] The game uses two styles of text annotation for players to complete a linguistic task. Initially text is presented in **Annotation**

[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.'*



Rhinogradentia (Wikipedia)

Rhinogradentia (also known as snouters or Rhinogrades or Nasobames) is a fictitious mammal order documented by the equally fictitious German naturalist Harald Stumpke. The order's most remarkable characteristic was the Nasorium, an organ derived from the ancestral species's nose, which had variously evolved to fulfill every conceivable function.

Both the animals and the scientist were allegedly creations of Gerolf Steiner, a zoology professor at the University of Karlsruhe. A mock taxidermy of a certain Snouter can be seen at the Musee zoologique in Strasbourg.

The order's remarkable variety was the natural outcome of evolution acting over millions of years in the isolated Hi-yi-yi islands in the Pacific Ocean.

DETECTIVE CONFERENCE

Another detective has said **the phrase in orange has been mentioned before and its nearest mention is highlighted in blue**. Do you agree with them?

Fig. 2: A task presented in Validation Mode.

Mode (called Name the Culprit in the game, see Figure 1). This is a traditional annotation method in which the player makes an **interpretation** (annotation decision) about a highlighted **markable** (section of text). If different players enter different interpretations for a markable then each interpretation is presented to more players in **Validation Mode** (called Detectives Conference in the game, see Figure 2). The players in Validation Mode have to agree or disagree with the interpretation. Players may also make comments about the task and/or skip the task if they do not want to provide an interpretation.

Training texts show the players whether their decisions agree with the gold standard. Once the player has completed all of the training tasks they are given a user rating (the percentage of correct decisions out of the total number of training tasks). The user rating is recorded with every future annotation or validation decision. Players are given training texts until the rating is sufficiently high enough to be given real text from the corpus.[9]

Players could label markables as DN (discourse-new, where the markable refers to a newly introduced entity), DO (discourse-old, where the markable refers to an already mentioned entity in the text, NR (non-referring, where the markable does not refer to anything or PR (where the markable represents a property of a previously mentioned entity). Full details of the game's methodology, deployment and recruitment statistics are published elsewhere [15].

The dataset (`Phrase Detectives Corpus 1.0`) was used to determine what the collective quality of the players were, as well as the quality of individual

[9] A minimum rating threshold of 50% is set for the game.

decisions. Full details of the corpora, including processing pipeline, descriptive statistics and gold standard creation, are published elsewhere [7].

The quality of annotation and validation decisions are measured by agreement (the proportion of decisions that are correct compared to the gold standard) and each model's agreement score is statistically tested using a z-test. P values are reported unless they have an alpha level of $p<0.01$.

3.1 Agreement between expert annotators

Five documents from the Wikipedia corpus containing 154 active markables (W2) and one document from the Gutenberg corpus containing 57 active markables (G2) were selected. Each document was manually annotated by two experts operating independently.[10] The five documents from the GNOME (GN) corpus were annotated by e2 and compared to the consolidated annotations of the GNOME corpus (of which e18 was the main annotator). The GNOME annotations were recorded in *Phrase Detectives* under the expert ID e39181. In total there were 59 markables that e2 and GNOME produced an annotation for.

Overall, agreement between experts in the three corpora was very high although not complete: 93.2% (GN), 94.1% (W2) and 89.4% (G2), for a chance-adjusted κ value [1] of $\kappa = .93$, $\kappa = .88$ and $\kappa = .88$ respectively, which is considered extremely good. This value can be seen as an upper boundary on what we might expect from a crowdsourcing system on this type of data and task.

There was no significant difference between the inter-expert agreement of the three corpora (GN n(59) 93.2%; W2 n(154) 94.1%; G2 n(57) 89.4%; p=0.810, p=0.238, p=0.465, z-test) which shows that the expert annotations created by e2 are what could be considered a gold standard when compared to an existing gold standard and another linguistic expert. Expert annotator e2 also created the gold standard for two larger subcorpora from Wikipedia (W1) and Gutenberg (G1) data.

[10] The two experts were Jon Chamberlain (who developed the game and wrote the instructions) and Massimo Poesio (a linguistic expert in anaphoric coreference), called e2 and e18 respectively in the rest of this discussion.

3.2 Baseline measures of performance

Performance was measured by four variables: quality; cost; noise; and speed. These variables are of consideration when testing crowdsourcing models to assess quality as well as to reduce the cost, noise and speed of a crowd answer.

Quality is measured as the level of agreement between an expert and the highest-scoring system answer.

Noise is defined as the number of wrong interpretations per markable.

Cost is measured as the total number of decisions (annotations, validations or work) that are required to produce an answer set per markable.

Speed is defined as the time (in seconds) to create the game answer by summing all the response times of the annotations and validations per markable.

In the baseline validation model all annotations and validations for each interpretation of a markable were combined:

$$A + V_a - V_d$$

where A is the number of players initially choosing the interpretation in Annotation Mode, V_a is the number of players agreeing with that interpretation in Validation Mode, and V_d is the number of players disagreeing with it in Validation Mode. This formula is used to score each interpretation of a markable, with the highest scoring interpretation called the 'best' or game interpretation.

The baseline agreement in the three corpora in which two experts provided a gold standard show very high agreement, comparable to pairwise inter-expert agreement (see Table 1). Both W1 and G1 have lower agreement (quality) than W2 and G2, significantly so in the Gutenberg corpus (G1-G2, z-test, p=0.02; W1-W2, z-test, p=0.12). The baseline figures for the five gold standard corpora show high quality at near-expert annotator performance; however, the cost, noise and speed are high making this method too expensive via microworking, too noisy for extracting data in high-spam scenarios and too slow for short-term data collection projects.

4 Results

4.1 How many annotations are required?

With a majority voting annotation model there is an assumption that the larger the crowd, the more chance there is of getting the best answer to be in agreement (in this case) with an expert, which is the approach of microworking. It is also

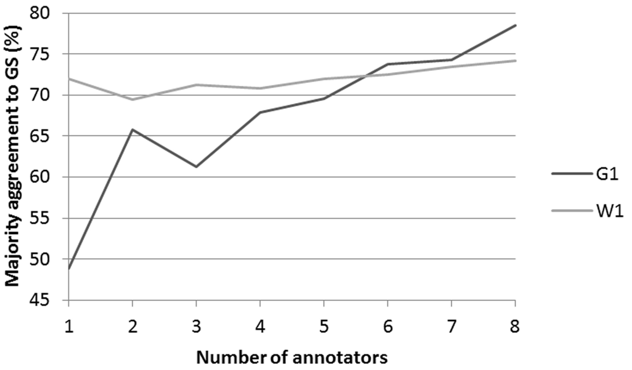


Fig. 3: Chart showing the majority voting agreement to the gold standard for different numbers of annotators for G1 and W1.

assumed that several annotators are superior to a single annotator, which is the approach of traditional, partly-validated expert annotation.

The expectation of diminishing returns from adding annotators past a certain point is simulated by comparing the agreement in the W1 and G1 corpora by using increasing numbers of annotators (in date order, oldest first). Due to the system's configuration, all tasks were shown to annotators at least eight times with no allowance for task difficulty. We observe only a small increase in agreement in the W1 corpus between one and eight annotators (A_1 to A_8), whereas the G1 corpus has a large, incremental increase of agreement (see Figure 3).

4.2 Improving quality with a crowdsourced validation stage

By adding the validation step to the eight annotations ($A_8 + V_a - V_d$), there is a significant increase in agreement in both corpora (G1 and W1, p<0.01, z-test, see Tables 2 and 3), whilst noise is not affected (as validation only votes up or down an interpretation). The validation stage will increase the cost and the time to complete the markable. The results show that the validation stage can increase the overall quality of an annotation-only crowd system without introducing more noise.

It is common on thread-based or cQA websites to feature validation functionality, but some may only have an upvote or 'like' button, the most notable example being Facebook. Here we test whether the same increase in agreement could be achieved by only using agreement validation (V_a) decisions.

Table 1: Baseline agreement $(A + V_a - V_d)$ between the two experts and the best answer from the game.

	GN		**W2**		**G2**		**W1**	**G1**
	e2	**e39181**	**e2**	**e18**	**e2**	**e18**	**e2**	**e2**
Markables	**264**	**61**	**176**	**160**	**63**	**58**	**3,729**	**1,844**
Agreement	**93.9%**	**85.2%**	**84.0%**	**81.8%**	**96.8%**	**93.1%**	**79.1%**	**86.6%**
Kappa κ	**0.86**	**0.85**	**0.63**	**0.59**	**0.96**	**0.92**	**0.52**	**0.85**
$Noise_{mean}$	**1.6**		**2.7**		**2.6**		**1.3**	**1.4**
	sd(2.0)		**sd(3.4)**		**sd(2.1)**		**sd(1.6)**	**sd(1.3)**
$Cost_{mean}$	**21.6**		**31.5**		**31.8**		**18.7**	**20.3**
	sd(15.0)		**sd(22.9)**		**sd(16.3)**		**sd(12.0)**	**sd(10.1)**
$Speed_{mean}$	**308.2**		**544.9**		**286.1**		**234.8**	**231.2**
$Speed_{median}$	**155**		**276**		**189**		**121**	**152**
	sd(471.1)		**sd(783.2)**		**sd(304.4)**		**sd(1,068.9)**	**sd(448.9)**

Table 2: Agreement between the expert e2 and the best answer (derived from different models) from the game in the G1 corpus.

G1 n(1,844)	$A + V_a - V_d$	A_8	$A_8 + V_a - V_d$	$A_8 + V_a$	$A_6 + V_a$	$A_6 + V_a$ **filtered**
Agreement	**86.6%**	**78.5%**	**86.0%**	**85.3%**	**84.1%**	**88.9%**
Kappa κ	**0.85**					
$Noise_{mean}$	**1.4**	**1.2**	**1.1**	**1.1**	**1.0**	**0.6**
	sd(1.3)	**sd(1.0)**	**sd(1.0)**	**sd(1.0)**	**sd(0.9)**	**sd(0.9)**
$Cost_{mean}$	**20.3**	**8**	**14.8**	**10.9**	**8.7**	**7.1**
	sd(10.1)	**sd(0)**	**sd(4.7)**	**sd(2.2)**	**sd(2.1)**	**sd(2.2)**
$Speed_{mean}$	**231.2**	**96.2**	**172.2**	**130.6**	**108.3**	**78.6**
$Speed_{median}$	**152**	**64**	**116**	**86**	**67**	**53**
	sd(448.9)	**sd(259.5)**	**sd(357.3)**	**sd(300.3)**	**sd(293.8)**	**sd(157.8)**

Table 3: Agreement between the expert e2 and the best answer (derived from different models) from the game in the W1 corpus.

W1 n(3,729)	$A + V_a - V_d$	A_8	$A_8 + V_a - V_d$	$A_8 + V_a$	$A_6 + V_a$	$A_6 + V_a$ **filtered**
Agreement	**79.1%**	**74.2%**	**79.2%**	**77.6%**	**76.9%**	**80.1%**
Kappa κ	**0.52**					
$Noise_{mean}$	**1.3**	**1.1**	**1.0**	**1.0**	**0.8**	**0.7**
	sd(1.6)	**sd(1.2)**	**sd(1.1)**	**sd(1.1)**	**sd(1.0)**	**sd(1.0)**
$Cost_{mean}$	**18.7**	**8**	**13.2**	**9.9**	**7.4**	**5.9**
	sd(12.0)	**sd(0)**	**sd(5.2)**	**sd(2.1)**	**sd(1.9)**	**sd(2.9)**
$Speed_{mean}$	**234.8**	**97.0**	**171.9**	**122.8**	**93.9**	**61.1**
$Speed_{median}$	**121**	**51**	**92**	**66**	**47**	**33**
	sd(1,068.9)	**sd(797.4)**	**sd(1,046.0)**	**sd(846.0)**	**sd(807.0)**	**sd(230.5)**

On both G1 and W1 corpora there is no significant difference in agreement between full validation ($A_8 + V_a - V_d$) and using agreement only ($A_8 + V_a$) validations (G1 n(1,844) p=0.542, z-test; W1 n(3,729) p=0.093, z-test), see Tables 2 and 3. This implies that a system that uses agreement validation or a like/upvote button such as Facebook can achieve the same level of quality for significantly less effort and time than using full validation.

4.3 Optimising and filtering the data

To discover a suitable stopping point for data collection, we determine how few annotations are required before most markables have been given the correct answer. Each markable in the corpus (when the correct interpretation was within the answer set) was measured to see how many annotations were required before the gold standard interpretation was introduced. This was averaged across all the markables in each corpus. According to these estimates, we require between 5.4 (G1) and 6.8 (W1) annotations before the gold standard interpretation is added to 97.5% of markables. Knowing most of the interpretations should be captured within approximately six annotations, and therefore further annotations were likely to introduce more noise, an optimised model ($A_6 + V_a$) was tested and showed agreement was not significantly reduced (G1 n(1,844) full 86.6%, optimised 84.1%, p=0.03, z-test; W1 n(3,729) full 79.1% optimised 76.9, p=0.02, z-test), but the noise and cost were.

Additionally, three types of anomalies were identified in the data that were considered worth filtering out because the data or data source were not what would be expected:

1. Recording a PR() interpretation should have been impossible to enter as an interpretation and is presumed to be caused by a technological issue;
2. A time of 0 (zero) seconds for an annotation or validation decision was presumed to be more likely a system error or spam response than a human response;
3. A method of profiling players was developed for the game to detect unusual or outlier behaviour. The profiling compared a player's annotations, validations, skips, comments and response times against the average for the entire game. Players with a proportion of DN responses greater than 90% or a proportion of DO responses below 10% were excluded with this filter.[11]

[11] Based on the profiles of confessed spammers *blbuc (946)* and *gully (1000)* unusual player behaviour was identified: selecting DN responses for almost 100% of markables

Table 4: Summary of agreement under different conditions, showing that the optimised and filtered validation model performs as well as the full baseline model.

	GN	G2	W2	G1	W1
Markables	**275**	**63**	**176**	**1,884**	**3,729**
Inter-expert	**93.2%**	**89.4%**	**94.1%**		
Baseline agreement ($A + V_a - V_d$)	**93.9%**	**96.8%**	**84.0%**	**86.6%**	**79.1%**
Baseline+filtered ($A + V_a - V_d$)	**93.9%**	**98.4%**	**85.2%**	**88.5%**	**79.4%**
Optimised+filtered ($A_6 + V_a$)	**93.5%**	**98.4%**	**84.6%**	**88.5%**	**80.1%**
Difference over baseline	**-0.4%**	**+1.6%**	**+0.6%**	**+2.3%**	**+1.0%**
p (z-test)	**0.849**	**0.555**	**0.881**	**0.077**	**0.285**

Filtering was applied to the baseline aggregation techniques and whilst it did increase the agreement in four of the five corpora (GN had no change) the change was not significant (G1 n(1,804) 86.6% pre-filtered, 88.9% post-filtered, p=0.03, z-test; W1 n(3,729) pre-filtered 79.1% post-filtered 80.1%, p=0.285, z-test). This is an indication that the aggregation methods used in the validation model were an effective, if not cost-efficient, way to remove spurious or malicious interpretations (see Table 4).

The optimised model was also filtered and, unlike the full validation model, was improved, with the agreement improved over the baseline, in addition to the reduced noise, cost and increased speed. With simple adjustments to the system (represented by the filtering), along with the optimised model, dramatic improvements to system performance can be achieved in all four key criteria.

5 Discussion

One of the simplest ways of reducing the costs of crowdsourcing is to increase the efficiency of the human computation. By optimising the data collection model, this research has shown that it is possible to maintain high-quality results whilst drastically reducing the amount of human effort required. By comparing the work of annotators against annotators supported by a validation stage, we showed that the latter can increase the overall quality of a crowd system without introducing more noise. The investigation showed that using agreement validation (instead

as this was the most efficient way to spam the game. Another unusual profile was few DO responses compared to DN, such as *Johnnickel (779)* or *askrukt (5970)*, which might indicate a technological issue.

of full validation) increases efficiency without reducing quality. Additionally, an optimised model reduces the number of annotations that are required, in a way so as not to affect quality significantly but also to reduce noise and cost. This reinforces the idea that understanding how many decisions need to be gathered is key to making a crowd-based system efficient. Finally, filtering the data in post-hoc analysis to remove spurious interpretations indicates that system testing and user training are essential for obtaining high quality results from a crowdsourcing system.

This experimental work, combined with the findings of directive methods in microworking, provide guidance for how to develop data collection methods using different types of crowdsourcing techniques, from games and gamification through to cQA and social networks.

A validation model is intuitive to users and features in some form on most social network platforms, allowing the community to show favour for particular content or solutions, and this method has been shown to be effective and efficient [6]. Other forms of voting exist, such as full validation (like and dislike) or graded voting (using a five star vote system) allowing for more fine-grained analysis of the community's preference; however, further research is needed to assess whether this is actually a waste of human effort and a simple like button proves to be the most effective.

In this research, users were rewarded for agreement and not punished for being disagreed with; however, other scoring models of this kind do exist [16]. The social network Facebook has resisted repeated calls from users to add a dislike button presumably because some of their content is linked to advertising. It may be that negative scoring would produce better results when using the model in post-processing or if the user did not know they were being punished. Social networks discourage the expression of negative views on other users' posts and it seems intuitive that positive behaviour be reinforced in crowdsourcing to encourage participation.

6 Conclusion

This paper investigated a validation model implemented in a game to test whether a validation step can provide higher quality results than just acquiring more annotations. The validation stage was shown to increase the overall quality without introducing more noise. The investigation showed that using agreement validation (instead of full validation) does not reduce quality but increases efficiency. Additionally, an optimised model reduced the number of annotations that were required, again not significantly affecting quality but reducing noise and cost. This reinforces the idea that understanding how many decisions need to be gathered is key to making a crowd-based system efficient.

Problem solving on social networks can be viewed in the same way as a crowdsourcing system with a validation stage and established techniques could be applied to make this an efficient, large-scale approach to human computation.

Acknowledgment: The authors would like to acknowledge the feedback from anonymous reviewers of this paper and to thank all the players who played the game. The creation of the original game was funded by EPSRC project AnaWiki, EP/F00575X/1. The analysis of the data was partially funded by an EPSRC Doctoral Training Allowance granted by the University of Essex, EPSRC grant ES/L011859/1 and European Research Council (ERC), Grant agreement ID: 695662.

References

[1] R. Artstein and M. Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008.

[2] Yochai Benkler and Helen Nissenbaum. Commons-based peer production and virtue. *Journal of Political Philosophy*, 14(4):394–419, 2006.

[3] Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. Soylent: A word processor with a crowd inside. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology (UIST'10)*, pages 313–322, 2010.

[4] Amiangshu Bosu, Christopher S. Corley, Dustin Heaton, Debarshi Chatterji, Jeffrey C. Carver, and Nicholas A. Kraft. Building reputation in StackOverflow: An empirical investigation. In *Proceedings of the 10th Working Conference on Mining Software Repositories (MSR'13)*, pages 89–92, 2013.

[5] Daren C. Brabham. *Crowdsourcing*. The MIT Press, 2013.

[6] J. Chamberlain. Groupsourcing: Distributed problem solving using social networks. In *Proceedings of 2nd AAAI Conference on Human Computation and Crowdsourcing (HCOMP'14)*, 2014.

[7] J. Chamberlain, M. Poesio, and U. Kruschwitz. Phrase detectives corpus 1.0 crowdsourced anaphoric coreference. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*, may 2016.

[8] Ido Guy, Inbal Ronen, Naama Zwerdling, Irena Zuyev-Grabovitch, and

Michal Jacovi. What is your organization ‘like’?: A study of liking activity in the enterprise. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI ’16, pages 3025–3037, New York, NY, USA, 2016. ACM.

[9] Matthias Hirth, Tobias Hoßfeld, and Phuoc Tran-Gia. Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms. *Mathematical and Computer Modelling*, 57(11):2918 – 2932, 2013.

[10] J. Howe. *Crowdsourcing: Why the power of the crowd is driving the future of business.* Crown Publishing Group, 2008.

[11] Faiza Khattak and Ansaf Salleb-Aouissi. Quality control of crowd labeling through expert evaluation. In *Proceedings of the 2nd Workshop on Computational Social Science and the Wisdom of Crowds (NIPS’11)*, 2011.

[12] Anand P. Kulkarni, Matthew Can, and Bjoern Hartmann. Turkomatic: Automatic recursive task and workflow design for mechanical turk. In *CHI ’11 Extended Abstracts on Human Factors in Computing Systems*, pages 2053–2058, New York, NY, USA, 2011. ACM.

[13] M. Lafourcade, A. Joubert, and N. Le Brun. *Games with a Purpose (GWAPS).* John Wiley & Sons, 2015.

[14] Greg Little, Lydia B. Chilton, Max Goldman, and Robert C. Miller. Turkit: Human computation algorithms on mechanical turk. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology*, UIST ’10, pages 57–66, New York, NY, USA, 2010. ACM.

[15] M. Poesio, J. Chamberlain, U. Kruschwitz, L. Robaldo, and L. Ducceschi. Phrase Detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Transactions on Interactive Intelligent Systems*, 3(1):1–44, April 2013.

[16] W. Rafelsberger and A. Scharl. Games with a purpose for social networking platforms. In *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia*, 2009.

[17] Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. Get another label? Improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD’08)*, pages 614–622, 2008.

[18] Brian Sidlauskas, Calvin Bernard, Devin Bloom, Whitcomb Bronaugh, Michael Clementson, and Richard P. Vari. Ichthyologists hooked on Facebook. *Science*, 332(6029):537, 2011.

[19] R. Snow, B. O’Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast - but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP’08)*, 2008.

7 Author biographies

Dr Jon Chamberlain is a web developer and lecturer in Human-Computer Interaction at the University of Essex with experience of industrial and academic computer applications (language processing, game design, social network analysis) in the domains of citizen science, marine ecology, and human rights observation. He was the lead developer of the Phrase Detectives project since its inception in 2007 and has continued investigating crowdsourcing using games and social networks for almost a decade.

Professor Udo Kruschwitz's research interests are in natural language processing (NLP), information retrieval (IR) and the implementation of such techniques in real applications. He is developing techniques that allow the extraction of conceptual information from document collections and access logs and the utilization of such knowledge in search and navigation contexts. Professor Kruschwitz was Co-PI in the original EPRSC project that developed Phrase Detectives.

Professor Massimo Poesio is a computational linguist. His work on anaphora is driven by the analysis of corpora and of disagreements in corpus annotation, most recently, using the Phrase Detectives game-with-a-purpose to collect such data. He is also a PI of the DALI project, an Advanced ERC grant; a supervisor in the IGGI Doctoral training centre in Intelligent Games and Game Intelligence; and a PI in the Centre for Human Rights and Information Technology in the Era of Big Data.

Fig. 4: Authors Jon Chamberlain, Udo Kruschwitz and Massimo Poesio.