Human Computation
AAAI Technical Report WS-12-08

Systematic Analysis of Output Agreement Games: Effects of Gaming Environment, Social Interaction, and Feedback

Shih-Wen Huang and **Wai-Tat Fu**
University of Illinois at Urbana Champaign
Department of Computer Science
201 N Goodwin Avenue
Urbana, IL 61801
{shuang51,wfu}@illinois.edu

Abstract

We report results from a human computation study that tests the extent to which output agreement games are better than traditional methods in terms of increasing quality of labels and motivation of voluntary workers on a task with a gold standard. We built an output agreement game that let workers recruited from Amazon's Mechanical Turks label the semantic textual similarity of 20 sentence pairs. To compare and test the effects of the major components of the game, we created interfaces that had different combinations of a gaming environment (G), social interaction (S), and feedback (F). Our results show that the main reason that an output agreement game can collect more high-quality labels is the gaming environment (scoring system, leaderboard, etc). On the other hand, a worker is much more motivated to voluntarily do the task if he or she can do it with another worker (i.e., with social interaction). Our analysis provides human computation researchers important insight on understanding how and why the method of Game with a Purpose (GWAP) can generate high-quality outcomes and motivate more voluntary workers.

Introduction

With the success of the ESP game(von Ahn and Dabbish 2004), there has been an increasing number of Game with a Purpose (GWAP) (von Ahn and Dabbish 2008) to harness the power of human computation. The main appeals of GWAP is that it provides a gaming environment to motivate people to engage in tasks that are otherwise difficult to perform by computers. One notable example of these tasks is labeling. Providing semantic labels to, for example, images or semantics of sentences, are known to be difficult for computers because it requires common sense knowledge. Yet, these labels are critical for improving performance of computations such as Web search of images, semantic analysis of documents, etc. With the enormous amount of time spent by people to play online games, it is possible for GWAPs to collect huge amount of human-generated labels with minimal cost. However, despite the apparent success of GWAP as a technique for harnesing human computation, it is still unclear which components in a GWAP is most useful for generating good labels (Robertson, Vojnovic, and Weber 2009). In order to better understand how GWAPs can outperform the traditional labeling process, a systematic analysis for GWAPs is necessary. The current paper focuses on two main reasons that make GWAP useful: (1) **how** and **why** GWAPs affect the quality of the collected labels, and (2) how can GWAP be designed to increase the motivation for players to label the data.

In this paper, we focus on output agreement games, one of the most popular and earliest form of GWAPs. A typical output agreement game has the following procedure: The game first randomly matches multiple players and provides them the same set of inputs, which can be images (von Ahn and Dabbish 2004), words (Seemakurty et al. 2010), or any data that the game designer wants to label. Then the players start to generate outputs that are related to the inputs. The players will be rewarded if the outputs generated by different players reach a certain level of agreement. When the game ends, some versions of the agreed outputs will be collected as the labels that describe the data.

Can output agreement games collect more high-quality labels?

In the original paper of the ESP game (von Ahn and Dabbish 2004), the authors provided three criteria to judge that the ESP game can collect labels with high quality. First, the labels collected by the game at least describe parts of the image. Second, at least 83% of the labels for each image generated by paid worker were covered by the labels collected from the game. Third, 85% of the labels collected from the game would be used to describe the image by other independent participants.

Recently, these criteria have been questioned. For example, in the paper of *Rethinking the ESP Game* (Robertson, Vojnovic, and Weber 2009), the authors built a robot that generated labels without the knowledge of the images at all. Instead, this robot only used the words that were already used to label the images, and used a language model to generate labels to play with other human players online. The result showed that the robot generated many labels that matched human players and thus earned high points. However, given that the robot apparently could not assign high-quality labels without knowledge of the images, high agree-

ment between players clearly did not imply high-quality labels. In fact, they used their results to argue that players in the ESP game usually produce obvious labels in order to match other player's labels, rather than high-quality labels that provide useful information about the images. Results from this study suggest that incentives that motive players to reach high agreement does not necessarily lead to high-quality labels.

Though different arguments have been proposed, without a carefully controlled experiment, it is still not clear whether output agreement games improve or harm the quality of collected labels. To the best of our knowledge, our research is the first one to compare the labels collected from output agreement games and the labels collected using traditional labeling methods.

In addition to quality, an important component of GWAP is that the gaming environment can easily motivate many people to participate, and thus a large amount of data can be collected in a short time. For example, (Seemakurty et al. 2010) showed that participants expressed the game is fun and would like to play the game multiple times. (von Ahn and Dabbish 2004) showed that more than 13,000 people played their game in a four-month period. To preview our results, we found that only 7% of the workers who work with the traditional labeling interface expressed they are willing to participate the task again for free while 37% of the participants play the output agreement game doing exactly the same task said they would love to do the task again even if there is no monetary reward. These results are in general consistent with previous results.

The current study

In the current study, we used an output agreement game to collect labels of semantic textual similarity [1], and compared the labels collected by the game with a traditional labeling interface. We are interested in the reasons that output agreement games can outperform traditional labeling interfaces. To answer this question, we decomposed an output agreement game into three components: Gaming environment (G), Social interaction (S), and Feedback (F). Based on these three components, we designed four different interfaces and implemented a baseline interface that imitate the traditional labeling interface. We conducted a systematic analysis that compared these five interfaces to find out the reasons for output agreement games to collect more high-quality labels and motivate voluntary workers.

Three Major Components of an Output Agreement Game

Different components of an output agreement game has been studied in previous studies. However, none of them has directly compared their effects on quality of labels and motivation to participate. The current study chose to compare three major components that are commonly used in GWAP: **Gaming Environment (G), Social Interaction (S), and Feedback (F)**. The use of these three components in previous studies are reviewed below:

[1] http://www.cs.york.ac.uk/semeval-2012/task6/

- **Gaming Environment (G)**: The gaming environment in an output agreement game is a scoring system that reward players who generate the same outputs. For example, in the ESP game (von Ahn and Dabbish 2004), players earn points if they enter the same word to describe the input image. *Jinx* (Seemakurty et al. 2010) has a more complicated bonus system that reward players who produce consecutive matched answers. To further increase players' incentives to earn high points, an output agreement game usually has a leaderboard that shows the scores of previous players. This design motivates players to beat the previous players by earning more points in the game. Some games (von Ahn and Dabbish 2008) have player skill level, which encourage players to play the game multiple times to reach higher level. Though there are many variation of the gaming environment, the main purposes of the gaming environment are:
  - Creating a fun environment for players to label the data while enjoying the game.
  - Encouraging players to generate agreed answers as many as they can.
  - Motivating players to engage in the task to generate high-quality outputs.
- **Social Interaction (S)**: Output agreement games randomly matched multiple online players to play the same game. Most output agreement games let players work together and reward them if their answers are agreed. For example, *Jinx* (Seemakurty et al. 2010) ask its two players in the same game to type synonym for the same word in a paragraph at each round. The players can provide answers multiple times until their answers are matched. This means the players' answers can be implicitly learned by other players, and their performance also depend on another online player's. It provides a social connection between players. This is much different from traditional labeling process, in which a single person labels the data without interaction with others.
- **Feedback (F)**: The other component that is relateively unexplored is feedback. Feedback is implied in the gaming environment and social interaction. In an output agreement game, players are informed if their answers are agreed or not at some point during the game. This helps players to self-evaluate their own answers and learn to provide better answers. For example, in the ESP game, if a player chooses to describe background objects of the image while the other player describes the foreground objects, the players might start to notice that their labeling strategies are not matching, and thus they may learn from the feedback to correct and change their labeling strategy to reach better agreement. Feedback therefore may provide crucial information for participants to improve their labels. Although intuitive, it is not clear to what extent the quality of feedback helps participants to learn, and what role does it play in generating high-quality labels in a GWAP.

The three components mentioned above make labeling using output agreement games different from that in traditional, single-person methods.

Experimental Design

In our experiment, we recruited 150 subjects and asked them to assign semantic similarity labels to 20 sentence pairs. The subjects were divided into five groups, with each group performed the task using five different interfaces. In other words, there were 30 subjects assigned to each interface.

The Semantic Textual Similarity Task

We used the semantic textual similarity task in the SemEval-2012 workshop[2] to evaluate the performance of the workers in our experiment. The semantic textual similarity task asks the participants to submit systems that examine the degree of semantic equivalence between two sentences. For each sentence pair, the system should find a label between 0 to 5 that best describe the degree of semantic equivalence between them. For example, given the following two sentences:

- John said he is considered a witness but not a suspect.
- "He is not a suspect anymore." John said.

Participants should choose a label using the criteria listed below:

- 5, if the two sentences are completely equivalent, as they mean the same thing.
- 4, if the two sentences are mostly equivalent, but some unimportant details differ.
- 3, if the two sentences are roughly equivalent, but some important information differs/missing.
- 2, if the two sentences are not equivalent, but share some details.
- 1, if the two sentences are not equivalent, but are on the same topic.
- 0, if the two sentences are on different topics.

In this example, the participant should choose label 3 because while the two sentences are similar, some important information is missing.

There is a publicly available training data that contains more than 1000 sentence-pairs with their gold standard labels. As described in the description of the task, the gold standard labels were generated using Amazon Mechanical Turk. These are the average of the labels given by 5 different workers for each sentence pair. To further verify the validity of the gold standard, we extracted the majority vote labels [3] from the labels assigned by the 150 subjects and compared them with the gold standard labels. We found that in 15 out of 20 sentence pairs, the majority vote labels of the 150 subjects are the labels that are closest to the gold standard labels.[4] In other 5 sentence pairs, the majority vote labels of the 150 subjects are also distant from the gold standard labels less than distance one. In our result analysis, we only keep the 15 sentence pairs that the majority labels generated by the 150 subjects are the closest labels to the gold standard labels.[5]

The reason we chose the semantic textual similarity task to evaluate the workers' performance is that it is a task with a validated gold standard. In addition, the gold standard is not obvious to the workers and thus the task requires the workers to put some effort into generating labels. It takes about 10 to 15 minutes for a worker to complete a task that consists of 20 sentence pairs. The workers have to carefully examine the sentences before they choose their answers. Therefore, it is an excellent test bed for us to examine the ability of different interfaces to motivate workers to generate high-quality labels.

Subject Recruitment on Amazon Mechanical Turk

In this experiment, we recruited 150 workers from Amazon Mechanical Turk. We published our HITs (Human Intelligent Task) on Amazon Mechanical Turk from 3/2/2012 to 3/16/2012. Each worker could earn $0.05 after he completed the HIT. In the HIT page we published, we provided a brief introduction to the task and a link to route the workers to our experiment website. Workers were shown an instruction page and started to perform the task on our website. In order to receive monetary reward, workers had to assign similarity labels to 20 sentence pairs. After a worker completed his job, the website would show him an unique completion code that he could enter in the original HIT page on Amazon Mechanical Turk. We then rewarded the worker based on the completion code he entered. The advantage of this approach is that we could build more flexible interfaces for our experiment.

Interface Design

Based on the three major components of output agreement games: **Gaming environment (G), Social interaction (S), and Feedback (F)**. We designed four different interfaces (F, F+S, F+G, F+S+G) and implemented a baseline interface (B) for worker to work with. This allowed us to test the effect of each component. The design of the five different interfaces are described below:

1. **Baseline (B)**: In the baseline interface, workers would be provided a page that showed a sentence pair at the upper half of the page and the choices of labels at the lower half of the page (Figure 1). This interface imitates the traditional labeling interface used to collect labels.
2. **Labeling with feedback (F)**: In this interface, we included the component of feedback of output agreement games. Compared to the baseline interface, workers who worked with this interface received a feedback from the system at each round of labeling that showed whether the label he generated at the last round was exactly the same as, similar to, or much different from[6] the label generated by a randomly selected previous worker.

[2] http://www.cs.york.ac.uk/semeval-2012

[3] Majority vote labels are the labels generated by most subjects for each sentence pair

[4] Because the gold standard is the average of 5 scores, some of the scores are decimal like 3.2. We regard the rounded gold standard score as the closest score to them.

[5] We also try to analyse the experimental result on 20 sentence pairs, the results are similar.

[6] The two answers were exactly the same if the labels were the same, were similar if they were in the range of +/-1, and were much different if they were not in the range of +/-1.

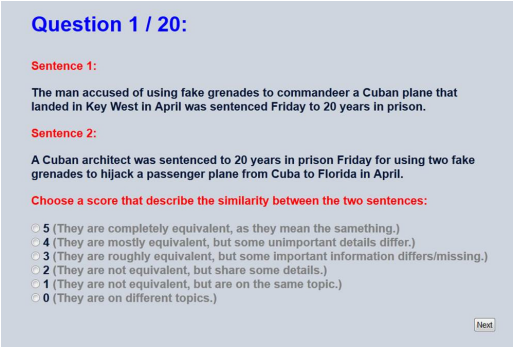

Figure 1: Baseline labeling interface.

3. **Labeling with feedback and social interaction (F+S)**: To test the effect of social interaction, we built a labeling interface with a matching page at the beginning of the task and a waiting page after each round of labeling. The matching page indicated that the system was trying to match the worker with another worker and the waiting page showed that the system was obstensibly waiting for the other worker to respond. These pages paused for a few seconds (the length depended on a random number generator) to make the worker feel that they were really working with someone else simultaneously. At each round, the system showed a message that told the worker whether the answer provided by the worker was exactly the same as, similar to, or much different from his teammate's[7].

   We implemented this interface by imitating the matching scenario so that participants were ostensibly connected to other participants during the task. We did this instead of actually matching two workers because it was not easy to recruit a pair of workers at the same time from Amazon Mechanical Turk. On the other hand, the way the interface was designed there is no way the participant could tell that they were not actually matched to a real person, as feedback was given to them based on actual players' answers. In other words, the interface allowed workers to work together asynchronously while maintaining the social interaction between them by creating the perception that they were connected to others during the task.

4. **Labeling with feedback in a gaming environment (F+G)**: This interface provided workers a gaming environment that awarded them if they generated answers that were the same as that from a randomly selected worker. The difference between this interface and a traditional output agreement game is that it told workers that it evaluated their answers by a previous worker instead of the answers provided by a concurrently working worker (teammate). Therefore, in this interface, we eliminated the effect of social interaction. However, we still included feedback such that the worker could tell whether the answers were the same.

   To further motivate workers, the game provided a leaderboard that showed the name of the top 5 workers and their scores. The instruction told participants that if they could earn more points than the top 5 workers, their names would appear on the leaderboard. The game also provided a graphical representation that showed the score of the worker compared to the score of the leaders on the leaderboard. (see Figure 2.)

   The rules of the game are summarized as follow:

   - A worker could earn 5 points for his first answer that matched exactly the answer from the randomly selected worker. For the next n consecutive matches, the worker could earn n*5 points. In other words, in the second consecutive match, the worker could earn 10 points; in the third consecutive match, the work could earn 15 points, and so on. Once there was a mismatch, n was reset to 0.
   - If the answer of a worker was similar to but not exactly the same as the answer of the previous worker, he could still earn 2 points. However, the consecutive match count n would be set to 0, such that the worker could only earn 5 points for his next matched answer.
   - The worker would lose 5 points if the answer provided by the worker was much different (i.e., larger than 1) from that of the selected worker.
   - The final 5 questions were bonus questions. All the points that were earned in these five questions would be doubled. (tripled in the last one). This mechanism was intended to increase workers' incentives to be engaged towards the end of the game.

[7]The teammate was a randomly selected worker from a previous trial.

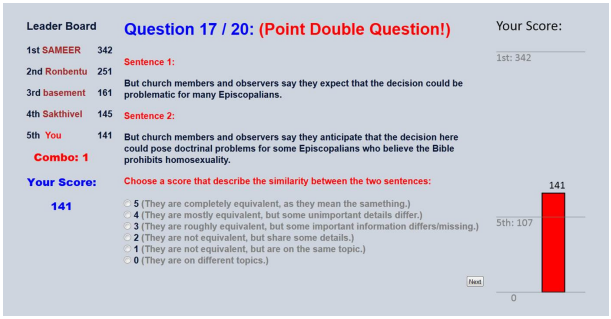

Figure 2: Labeling interface with gaming environment.

5. **Labeling with feedback and social interaction in a gaming environment (F+S+G)**: This interface included all the three components and was exactly the form of traditional output agreement games.

Results

We collected 3000 labels from 150 subjects. We discarded 5 of the 20 sentence pairs because the gold standard provided in the training data and the majority vote labels generated from our 150 subjects were different. In the remaining set of 15 pairs, we defined a high-quality label as one that had a difference of less than one from the gold standard. For example, if the label in the gold standard is 3.4, 3 and 4 will be regard as a high-quality label. This was done because the gold standard had fractions because of the averaging process, but the labels generated by participants were all integers. In our analysis, we used the number of the high-quality labels collected from different interfaces as a measure of performance.

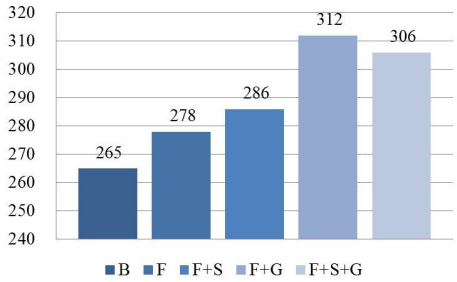

Figure 3: Amount of the high-quality labels collected by each interface

Can output agreement games collect more high-quality labels?

The amount of high-quality labels of the 15 sentence pairs collected from each interface are shown in Figure 3. We can see that the output agreement game (F+S+G) we developed collected 306 (68%) high-quality labels while the baseline interface (B) only collected 265 (59%) high-quality labels. A two-tails paired t-test on the labels collected from these two interfaces showed significant[8] differences (p=0.002). This shows that output agreement games are able to collect more high-quality labels compare to traditional labeling interface.

In addition, the amount of high-quality labels collected by the interface with gaming environment (F+G) also significantly higher than the number of baseline interface.(F+G : 312 (69%) vs B: 265 (59%), p=0.003) On the other hand, the interface with feedback (F) and the interface with social interaction (F+S) did not collect significantly more labels than baseline interface did. (F: 278 (62%) vs B: 265 (59%), p=0.27 and F+S : 286 (64%) vs B: 265 (59%), p=0.07)

This result indicates that when we included the gaming environment into the interface, the interface collected significantly more high-quality labels than baseline interface did. However, the effect of feedback and social interaction were not as significant as the effect of gaming environment on improving quality of collected labels.

Which component leads to more high-quality labels?

We further analyzed and compared the number of high-quality labels among the five different interfaces used in the experiment. The results are summarized below:

F vs F+S : Fixing the effect of feedback, we wanted to see if social interaction motivated workers to generate more high-quality labels. The result shows that the effect was not significant. (F+S : 286 (64%) vs F : 278 (62%), p=0.27)

F vs F+G *and* F+S vs F+S+G : In these two comparisons, we fixed other components and intended to see the effect of the gaming environment. The result shows there was a significant difference in the amount of high-quality labels collected between the interfaces with gaming environment and the interfaces without it. (F+G : 312 (69%) vs F : 278 (62%), p=0.01 and F+S+G: 306 (68%) vs F+S: 286 (64%), p=0.04) The result indicates that the gaming environment really made workers generate more high-quality labels no matter they were working with another worker or not.

F+G vs F+S+G : We also compared the performance of the workers who worked with **F+G** and **F+S+G**. The t-test statistic was not significant (p=0.58). This means social interaction did not significantly lead to more high-quality labels in a gaming environment.

To summarize the comparisons, we find that the reason our output agreement game could collect more high-quality labels is the gaming environment. On the other hand, feedback and social interaction did not significantly improve the ability to collect high-quality labels.

A game theoretic analysis of the results

Our experimental results have shown that the interfaces with gaming environment can collect more high-quality labels, here we present a simple game theoretic model to explain this phenomenon.

In this model, we consider each round of labeling as a game. First, we consider the case that there is no gaming environment (baseline). We assume the utility function of worker i labeling without gaming environment is:

$$U_i^B = I(x_i, e_i)$$

e_i indicates the effort worker i puts to solve the task:

$$e_i = \begin{cases} e_h, & \text{if worker } i \text{ chooses to put high effort} \\ e_l, & \text{if worker } i \text{ chooses to put low effort} \end{cases}$$

Because this is a task with a gold standard label, we assume that if the worker chooses to put high effort($e_i = e_h$), the worker can correctly select the gold standard label. On the other hand, if the worker chooses to put low effort($e_i = e_l$), which means the worker just randomly pick one label from the six options, the probability that the worker can choose the gold standard label is $1/6$.

x_i is a worker dependent positive utility for the worker puts high effort to solve the task (sense of accomplishment) and $I(x_i, e_i)$ is the intrinsic value for user i to perform the task:

$$I(x_i, e_i) = \begin{cases} x_i - e_h, & \text{if } e_i = e_h \\ 0, & \text{if } e_i = e_l \end{cases}$$

The assumption we make here is that if worker i chooses to put high effort to solve the task, the worker could get the positive utility x_i. Moreover, since the worker needs to put high effort to select the correct label, solving the task also incurs a negative utility e_h for him. On the other hand, if worker i chooses to randomly select one label from the options, he cannot get the sense of accomplishment of solving the task and there is also no cost for him ($e_l = 0$). Therefore, the worker would get zero utility if he chooses to put low effort.

Worker i would choose to put high effort if $U_i^B(e_h) = x_i - e_h > 0 = U_i^B(e_l)$. For simplicity, we assume that x_i is generated from the standard uniform distribution $U(0, 1)$ and $0 < e_h < 1$. This implies that the probability that a

[8] In this paper we regard $p < 0.05$ as significant

worker working with the baseline interface would choose to put high effort is:

$$P^B_{e_i=e_h} = P(x_i > e_h) = 1 - e_h$$

This shows that for an interface without gaming environment, the worker would choose to put high effort only if the positive utility he can get from accomplishing the task is higher than the cost for him to put high effort to work on it.

Then we consider the case that the gaming environment is included. The utility function of worker i that works with an interface with gaming environment becomes:

$$U^G_i = I(x_i, e_i) + P_m(e_i, e_j) \cdot G$$

$I(x_i, e_i)$ is the intrinsic value for the worker, which is the same as its counterpart in U^B_i. The second term in the formula is the expected utility that worker i could get from the gaming environment. We assume that if the labels generated from the two players (i and j) of the game are the same, player i can get a positive utility G (the enjoyment of earning the points of the game). The probability that the labels generated by the two players are matched is $P_m(e_i, e_j)$, which depends on the effort put by both of the players. We can summarize the outcomes of $P_m(e_i, e_j)$ below:

$$P_m(e_i, e_j) = \begin{cases} 1, & \text{if } e_i = e_j = e_h \\ 1/6, & \text{if } e_i = e_l \text{ or } e_j = e_l \end{cases}$$

If both of the players choose to put high effort, both of them would select the correct label. Therefore, the probability for them to generate matched labels is 1. In contrast, if one of the players chooses to put low effort (randomly choose one label from the six options), even if the other player puts high effort, the probability for their labels to be matched is $1/6$.

Therefore, the expected utility for worker i to choose to put high effort is:

$$E[U^G_i(e_h)] = x_i - e_h + P^G_{e_j=e_h} \cdot 1 \cdot G + (1 - P^G_{e_j=e_h}) \cdot 1/6 \cdot G$$

Where $P^G_{e_j=e_h}$ is the probability that the other player(j) chooses to put high effort to select the label. Moreover, if worker i chooses to put low effort to solve the task, no matter which effort level the other player chooses, $P_m(e_l, e_j) = 1/6$. As a result, the expected utility for worker i to choose low effort is:

$$E[U^G_i(e_l)] = 1 \cdot 1/6 \cdot G$$

Worker i would choose to put high effort if:

$$E[U^G_i(e_h)] > E[U^G_i(e_l)]$$

Which implies:

$$x_i - e_h + 5/6 \cdot P^G_{e_j=e_h} \cdot G > 0$$

We could rewrite the formula to:

$$x_i > e_h - 5/6 \cdot P^G_{e_j=e_h} \cdot G$$

Then we can compute the probability of worker i to put high effort in the gaming environment:

$$P^G_{e_i=e_h} = P(x_i > e_h - 5/6 \cdot P^G_{e_j=e_h} \cdot G)$$

At this point, we could already see that as long as G is positive (i.e. earning the points of the game is enjoyable for the workers) and the other player has a positive probability to choose to put high effort, then the probability of worker i to choose to put high effort in the gaming environment is higher: $P^G_{e_i=e_h} = P(x_i > e_h - 5/6 \cdot P^G_{e_j=e_h} \cdot G) > P(x_i > e_h) = P^B_{e_i=e_h}$. If we further assume that all the workers have the same utility functions(including the distribution of x_i) and this information is the common knowledge. We could compute $P^G_{e_i=e_h}$ using the following two equations:

$$P^G_{e_i=e_h} = P(x_i > e_h - 5/6 \cdot P^G_{e_j=e_h}) \cdot G \quad (1)$$

$$P^G_{e_j=e_h} = P(x_j > e_h - 5/6 \cdot P^G_{e_i=e_h}) \cdot G \quad (2)$$

Solve the equations above and we can get:

$$P^G_{e_i=e_h} = \frac{1 - e_h}{1 - 5/6 \cdot G}$$

This shows that when G (the enjoyment of earning the points of the game) increases, the gap between $P^G_{e_i=e_h}$ and $P^B_{e_i=e_h}$ also enlarges. This model shows that in order to earn higher points, the players would use the correct label as the protocol to generate matched labels in the gaming environment. Therefore, the incentives in the game not only encourage the workers to generate labels that are matched with the other worker, it also encourage the workers to generate the correct labels for the task. This explains why we can collect more high-quality labels from the interfaces with gaming environment.

Effect of feedback quality

One possible concern is that the reward system used in the gaming environment depends solely on consensus, and thus it sometimes penalizes workers even if they generate high-quality labels (closer to the gold standard). For example, it was possible that when the randomly selected worker had poor labels, feedback given based on answers from this worker would decrease the motivation for the workers to generate high-quality labels. To test this, for each worker, we extracted the labels given by the worker as well as the labels that were used to calculate feedback (i.e., the label created by the randomly selected worker). We then calculated a quality score for each label that was created by the worker and for the corresponding label used to calculate feedback based on how close it was to the gold standard. We then computed the correlation between the two sets of quality scores for each worker. If label quality was influenced by feedback quality, the correlation between the two sets of quality scores would be high. Interestingly, we found that this correlation was 0.04 in the (F+G) interface, and 0.07 in the (F+S+G) interface. The low correlations indicate that the quality of labels used to provide feedback did not affect the quality of labels provided by the workers. Rather, it seems that the mere fact that the workers knew that their labels would be evaluated was enough to encourage them to provide higher-quality labels.

This result is valuable because it means that we can evaluate a worker's output simply by another worker's output instead of the gold standard, which is often not directly available, or very costly to obtain, in many tasks. The results

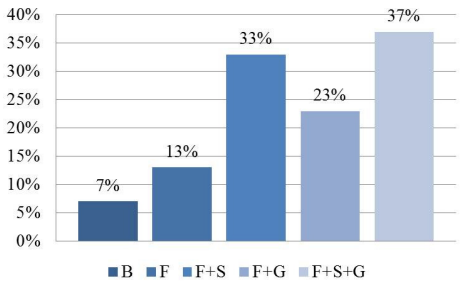

Figure 4: The ratio of workers report they will voluntarily do the task again in each group

therefore confirm that it is possible to design human computation systems that use the outputs from previous workers to evaluate the performance of the new worker to motivate workers to engage in the task to generate high-quality labels. Further research can be done to verify the generality of this finding in other tasks and other settings.

Which component motivates voluntary workers?

In our experiment, when workers completed labeling the 20 sentence pairs, we asked them how likely they would do the task again if they don't get paid next time. They could choose from 5 choices: I'd love to, Probably, Maybe, Unlikely, Never. We analyzed the number of workers who reported they would love to do the task again to see the ability to recruit voluntary workers of different interfaces. The number of the workers that reported that they would love to do the task again for each interface is shown in Figure 4. 11 out of 30 (37%) workers were willing to participate the output agreement game again even if there is no monetary reward. In contrast, only 2 out of 30 (7%) workers would do the labeling task again for free.

A more interesting fact we found in our result is that even if we took away the gaming environment from the output agreement game, simply making workers work with another online worker still motivated 10 out of 30 (33%) workers to do the labeling task again without monetary reward. This number is even higher than the number of voluntary workers (23%) motivated by the interface with gaming environment but without social interaction.

We can infer from this result that the key component for our output agreement game to attract voluntary workers was the social interaction between workers. If a worker was working with another worker as a team, he would have more incentives to do the labeling task.

Conclusion

In this paper, we conducted a systematic analysis of output agreement games, a popular form of GWAPs. The results show that the output agreement game we built not only motivated more voluntary workers, it also collected more high-quality labels compared to the traditional labeling interface. This promising result confirms that it is possible to use output agreement games to replace traditional labeling interfaces to collect high-quality labels.

We decomposed an output agreement game into three major components: Gaming environment (G), Social interaction (S), and Feedback (F). We found that the main reason our output agreement game could collect more high-quality labels was the gaming environment. There are two important implications from this result: First, the reward system used in output agreement games did motivate worker to engage in the game. Second, it is possible to utilize the labels generated by the previous workers to evaluate a new worker's performance to encourage them to generate high-quality labels. Even when the quality of collected labels are not perfect, feedback based on consensus does help to collect more high-quality labels.

The other interesting finding in our experiment is that when a worker thought he was working with another online worker as a team, he was highly motivated to voluntarily do the task. We found that even without the help of the gaming environment, simply matching two workers provided them much motivation to do the task voluntarily. This shows that it is possible to develop a system that matches workers (or even just make them think they are matched) to recruit more voluntary workers.

In future, we would like to conduct experiments to test the limit of the mechanism of output agreement evaluation. Although we believe that our results apply to data labeling tasks in natural language processing because they usually use interannotator agreement (ITA) to evaluate the collected labels.(Snow et al. 2008), it is still doubtful that whether this mechanism can apply to tasks whose gold standard labels are not that clear (e.g. image labeling) or even some more creative tasks. Knowing the limit of this mechanism could help people decide if they should include this mechanism when building human computation systems.

References

Robertson, S.; Vojnovic, M.; and Weber, I. 2009. Rethinking the esp game. In *Proceedings of the 27th international conference extended abstracts on Human factors in computing systems*, CHI EA '09, 3937–3942. New York, NY, USA: ACM.

Seemakurty, N.; Chu, J.; von Ahn, L.; and Tomasic, A. 2010. Word sense disambiguation via human computation. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, 60–63. New York, NY, USA: ACM.

Snow, R.; O'Connor, B.; Jurafsky, D.; and Ng, A. Y. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, 254–263. Stroudsburg, PA, USA: Association for Computational Linguistics.

von Ahn, L., and Dabbish, L. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '04, 319–326. New York, NY, USA: ACM.

von Ahn, L., and Dabbish, L. 2008. Designing games with a purpose. *Commun. ACM* 51:58–67.