Crowdsourcing Relevance Assessments through a Game-based Approach using Social Networking

Monisha Manoharan
Department of Computer Science
The University of Texas at Austin
monisha@utexas.edu

Madhura Parikh
Department of Computer Science
The University of Texas at Austin
mparikh@cs.utexas.edu

ABSTRACT

In this project, we use crowdsourced games to generate relevance assessments. Traditional methods for obtaining document relevance have typically used human experts. However, it quickly becomes infeasible for a limited group of judges to generate relevance judgments for large document collections. A more scalable alternative is to leverage crowdsourcing to generate these relevance judgments [1, 6]. However, as crowdsourced paradigms are inherently susceptible to cheating and poor quality of labor, we propose constructing a game-based approach for crowdsourced relevance assessments, that incentivizes good quality of labor via a reward and social recognition framework. We believe that such a game based task is more appealing to the worker, further improving the labor quality.

1. INTRODUCTION

The game based approach we propose may be considered to belong to a class of games called *Games With a purpose (GWAP)* [11]. People are always seeking ways of entertainment and one of the most popular means is via games. GWAPs aim to utilize this entertainment-seeking nature of humans constructively to solve large problems that are not easily solvable by machines.

One of the major drawbacks of crowdsourcing is that it is highly prone to cheating. Most workers are money driven often completing the work unsatisfactorily for the sake of making quick bucks. [3] studied some ways of making crowdsourcing more robust to such cheating. They reported that 'entertainment-driven' workers are less likely to cheat as compared to 'money-driven' workers. Moreover tasks that are more engaging are likely to result in better work quality.

There has already been some work that explores the use of games for improving the quality of crowdsourced relevance judgments [4]. They report that such a game based approach is indeed better at avoiding cheating in crowdsourced settings while also being much cheaper than typical models that are based on monetary payments. While the paper shows that this is a promising approach, they do mention some extensions, that can further strengthen it. For one, in their approach, the game is mostly a passive GWAP. Thus the worker blindly plays the game without much source of competition - while this could in fact be a significant motivating factor. They do have a leaderboard of high scores, but this mostly lists some other unknown workers, and so may not have a major impact.

1.1 Motivation

While we mentioned that there have already been approaches at gamifying relevance assessments, none of them have been much of a success. The goal of this project is to answer the following two research questions:

RQ1: Can crowdsourced games help in achieving higher quality relevance assessments –(while higher quality is a subjective term, the hope is that they are suitable to replace gold standard human judges) when compared to conventional crowdsourced techniques?

RQ2: Can crowdsourced games reduce the percentage of cheating and spamming as compared to conventional crowdsourced techniques?

The driving hypothesis of our project is that if a game can be designed to entertain, appeal or be useful to the crowdworker in learning or improving a new skill, then the answers to both **RQ1** and **RQ2** will most likely be a yes.

Before we discuss our design, we would first like to delve deeper into what are crucial aspects of designing a crowdsourced game.

1.2 Crowdsourced games : design principles

A good explanation of why crowdsourced games are likely to succeed over the conventional crowdsourcing techniques is provided by [8]. Whereas crowdsourcing platforms typically have a linear reward strategy (e.g pay-per-hit), the paper shows that other reward strategies might be more appealing to the workers and help get the maximum value for money. In particular, they propose competitive and randomized reward strategies. Another interesting point discussed in the paper is the concept of *information policies* - i.e providing workers with some information on how they stand with respect to other workers (in competitive strategies) or how much chance they have for winning (in a lottery based or randomized strategy). Another interesting insight they mention is that *workers typically work for fun and money.*

There are several important pros and cons of information and reward strategies which are useful guiding principles as we design our game. For instance a global leaderboard is likely to discourage lower ranked workers whereas a completely hidden policy might take away the competitive appeal of the game. They therefore suggest a medium approach where the workers may be shown their standing with respect to their k nearest neighbors, and show that it is much more effective.

Such information policies have also proven to be appealing on extremely popular platforms like Topcoder [7]. Topcoder offers a platform for various programmers to compete against each other in algorithmic matches. Here the players are randomly split into different groups - a.k.a *rooms* with a limit of 20 players per room. A per-room leaderboard is typically displayed throughout the match. While these Topcoder matches have nothing to do with crowdsourcing they do point towards effectively designing competitive games. Another very significant motivator in games is the social factor. For instance games on Facebook, such as QuizUp where people can compete with their friends are very popular. Similarly CodeJam, an annual coding competition held by Google, allows you to see how you fare with respect to your *friends*, in addition to the global leaderboard.

As a concrete example of these principles that work for a crowdsourced game, we would like to present a case study :

1.2.1 A case study : Duolingo

Duolingo [9] is a crowdsourced game that makes humans translate text while enabling them to learn a new language. Typically computers are much better at maintaining a dictionary, mapping words from one language to another, but in larger sentences, where context is important, they are no match for humans. So in Duolingo, the users start to learn the new language vocabulary and grammar and once they reach a particular skill level, they start applying these skills to help translate documents. Duolingo has recently become wildly popular having more than 12 million registered users [1] and voted as a best startup of the year.

We now describe some of the design principles in Duolingo.

As we mentioned earlier any crowdsourced game must have some payoff for it to be appealing. The payoff in Duolingo is that the user can learn a new language free of cost. Also Duolingo tries to make this process as much fun as possible - one user described it as *addicting*. To do this, it implements some of the design principles we mentioned earlier. For example to encourage quality submissions and avoid cheating, Duolingo introduces the idea of *streaks* where a player who maintains a continuous stretch of good and persistent work is awarded by *Lingots*. Lingots are a form of virtual currency in Duolingo and may be utilized for purchasing a language skill package that can enhance your learning. Similarly, social networking is a huge motivating factor in any game. Duolingo promotes this by awarding Lingots, every time a person invites her friend to join Duolingo. Further a person can follow her friends on Duolingo and track their progress. Duolingo awards badges/achievements that are displayed on the person's profile - this further adds to the excitement. A person can also challenge a friend to a *race* where they compete to prove their language skills. This leads to a rich, socially interactive and rewarding experience and also enhances learning. Duolingo also adopts a medium information policy wherein people can see their standing with respect to their friends but not globally.

So to summarize, here are a few of the key insights we have gleaned after surveying current work in the field:

I1 : The game can be made popular if it can provide entertainment to the user and/or helps her learn a useful new skill.

I2 : The quality of the submitted work can be improved if a continuous stretch of good work is rewarded, rather than each individual submission.

I3 : Games that have a competitive element are more likely to appeal to the user. Also global leaderboards may be demotivating so a local leaderboard design should be adopted.

I4 : Games that have a social networking element will be more appealing to a majority of the users. Encouraging people to invite their friends, will also improve the crowd participation.

I5 : People generally like to play for both fun and money. Thus virtual rewards like badges are motivating to the user, but rewards that can offer a new benefit to the user in the real world will likely have more appeal.

2. DESCRIPTION OF OUR GAME DESIGN

We believe that, success of a game mainly depends on its design [5] and based on the insights gathered above, we now describe the design of our game to generate crowdsourced relevance assessments. First, we have decided to make our game appealing by allowing the user to improve an important skill : English reading and comprehension. Every year nearly a million students appear for *Graduate Record Examination (GRE)*, out of which nearly $500,000$ are from the US itself. There are also other tests such as *Test of English as a Foreign Language (TOEFL)* that are very popular in non-US countries. These tests have a major section that tests the user's reading and comprehension skills by asking them to read paragraphs and draw various conclusions from them under time constraint. Clearly to succeed in these tests just being fluent in English is not enough. In these tests, deliberate practice can hugely boost a candidate's performance - thus a platform that supports improving reading speed, and the ability to comprehend text swiftly is highly desirable. There are many test-prep products that are available which offer these benefits to the buyers. However, these products are typically not cheap and more importantly they are not that much fun either. They will not help if for instance the candidate wants to work on her reading skills while on a long bus-commute or while waiting for an elevator. Learning

[1] `http://www.businessinsider.com/luis-von-ahn-creator-of-duolingo-recaptcha-2014-3`

can also be enhanced if the students can study in a group, challenge their friends in fun ways and keep themselves motivated via a healthy competition with their peers. This kind of social element is totally missing from most available test-prep suites.

Taking advantage of this gap in the area, we have designed GRE UP! (a play on 'gear up') - a game that will allow the players to improve their reading speed and comprehension capacities to perform better in an important test, while also enabling fun interactions with their friends and giving them the opportunity to win free swag. In what follows, we describe the finer points of our design.

2.1 The game - 'GRE UP!'

This will be a two-player game. In the game, both the players (say Alice and Bob) will be presented with 3 identical paragraphs from a document. After reading each paragraph, the player will need to enter a quick one line description of the paragraph they just read. Once all the three descriptions are available from Alice, their ordering will be randomized and these jumbled three lines will be presented to Bob, and similarly for Alice. The challenge now is that Bob will need to guess the correct sequence of the descriptions generated by Alice (i.e. match each description correctly to the paragraph for which it was generated). The descriptions will be mutually available only once both the players have completed their entries for all the three paragraphs. The rules are few and simple:

- The player who correctly guesses the ordering first will be awarded 25 points.
- A player who correctly guesses the ordering but doesn't come first will earn 0 points.
- If a player guesses the ordering incorrectly, her opponent will be awarded 25 points.
- If a player successfully guesses the ordering first and also if her opponent is unsuccessful, she will be awarded a bonus of 50 points.

Players will be encouraged to be creative in their descriptions. They could also obfuscate them as they see fit. The description doesn't need to be a one line summary of the paragraph - it should just be a clue about which paragraph it refers to and could be something that is relevant within a friend group. So for instance it could be something like `[lol, this reminds me of jimmy's dog]` or a twitter hashtag - `[#NobelPeacePrize]`. This could add a lot of fun factor to the game and make it more relatable. Also the goal of this entire step is to ensure that the players are engaged while they read the paragraphs and also for ensuring that they do not skip any paragraph. This will reduce any spamming or laziness.

An interesting question is on the choice of the three paragraphs to display from the document. The goal is to choose paragraphs that can be a good representation of the entire document and thus enable the players to make good relevance assessments. As a first cut heuristic, we will pick up the first and the last paragraph from each document and one more paragraph, uniformly at random, excluding these two. The first and last paragraphs have good markers indicating a document topic. The random paragraph will ensure that most of the document get covered in different instances of the game. We have decided to constrain the number of paragraphs to 3 per question, since a large number of paragraphs will likely reduce the interest and engagement of the players.

Once a player crosses a certain threshold of points, (we are maintaining this at 100 points), a new feature will be unlocked. Now, whenever the players complete a question, a quick dialog will be displayed asking the user *What was this about?* with 5 topics including a topic *Nonew*. The player should pick up one or more topics (or None) that she thinks the three paragraphs were relevant for. The correctness criteria will be, agreement with the majority of players, (based on inputs in other instances of the game). Each correct answer will be awarded 10 points and incorrect answers will lose 10 points. A player who maintains a streak of four consecutive correct answers will be entered to win an electronic gift voucher - this could be a gift card from Amazon, with a denomination as per the budget. Instead of the 'winner takes it all' strategy, it would be wiser to split the budget into small equal portions, each of which can be awarded to different players. To maintain active interest, the winners could be drawn from the ballot, at hourly intervals, during peak usage times. The winners will be announced in their friend group, thus also encouraging other people in their network. The winner will also have the option of posting this as a status update on Facebook. A player's name will be put in the ballot each time she creates a new streak of four correct answers, and the names will be drawn uniformly at random. Thus, a player who has a consistent good performance will theoretically be more likely to win a voucher.

In addition to these strategies, a leaderboard will be maintained for each friend group. The players will be displayed with the total points they have amassed so far and their mastery level. Currently we are maintaining three levels:

noob : The player has points $\leq$ 400 points.

hacker : The player has points $\geq$ 400 and $\leq$ 1000.

ninja : The player has points $\geq$ 1000.

As an incentive, hacker and ninja players will be able to save one miss in a streak of four by 'buying' it with their points. Thus, hackers can save one miss by 'paying' 200 points. Ninjas can save a miss by 'paying' 300 points. Thus the greater your mastery at the game, the more you would be penalized for being wrong. This kind of scoring should help prevent expert players from dominating over noobs.

Of course most of these parameters are experimental and there would need to be user testing and feedback before the exact values can be decided. There is plenty of work done in the research community on designing a reward system for games [12]. However, we plan to stick to this simple system for our project and believe that this game will motivate people to stay engaged in reading paragraphs of text. The game construction is such that the players will be examining the text several times, while trying to come up with creative descriptions as well as when trying to match a description with a paragraph. This can help in generating better quality of assessments. Further since players require a minimum of 100 points to become capable of making relevance judgments, it can reduce instances of spamming. Spamming is also reduced because a player needs to maintain a streak of good judgments, before she is awarded. The fact that players will lose points if they make a wrong assessment ensures that players do not become passive after accruing a certain amount of points. The social networking effect where people can relate their fun experiences to paragraph descriptions as well as the broadcast of winners in the network can help maintain popularity within the network.

In any game design, it is often difficult to balance the entertainment aspect with the fairness aspect. Ideally we would like the game to be such that it does not allow players to collude and cheat, since the original goal of the game was to reduce cheating. Here we look at some typical scenarios, where are game may be vulnerable to cheaters, unfair game-play and other weaknesses and discuss our defenses.

- **Scenario 1: Player 1 is very slow while player 2 is quite fast. So player 2 may get bored and quit the game** Clearly this is a possible situation, and also one we would like to avoid, since we do not want players to lose interest in our game. As a first line of defense, during the time when the player is waiting for responses from the other player, we display the interim screen (Figure 4). This is discussed in more detail later, but basically it will prompt the player to make the relevance assessments while waiting for the second player to complete. This should mask some of the latency. Moreover while we could not design a way to include this in our prototype, the real game will also have a timeout parameter. If any player fails to respond by then, she will just concede that round to her opponent, and the play will move on. If both the players are inactive, the game will not take any action. A player with more than two timeouts will lose 100 points. We prefer this alternative to some other choices - such as using a pre-recorded bot to play against a player. This is because we believe that people are much likely to enjoy playing with their friends. So a pre-recorded bot may not hold the same appeal and may also be easy to detect if it just gives canned replies and is not designed carefully.
- **Scenario 2: Player 1 writes irrelevant descriptions so that player 2 fails to match correctly and player 1 always wins** There is nothing implicit in our game that prevents this from happening. We mainly depend on the fact that our game is played within a friend group. Thus, people are generally less likely to resort to such techniques. A person may soon become unpopular in the network if her friends notice such behavior often. Thus in the future, other players from the network will be less willing to play with the person. Also if no one is ready to play with the person, the person has no chance of winning free swag, which is one of the main appeals of the game, as our evaluation shows later.
- **Scenario 3: Player 1 and player 2 collude on the answers via a second communication channel** There are games like ESP [10] which prevent this by randomly matching players against each other. Since the players don't really know who they are playing against, such an attack is avoided. Our game is slightly different. In ESP, if both the players match in their guesses, they both win points. Thus it is very important to prevent player collusion in this case. In our case, even if both the players cooperate, essentially it would benefit just one player - the player who gets the correct choices first. As per the rules we discussed earlier, if a player guesses correctly but does not come first (which would hold for a tie as well), both the players just get 0. Thus players would essentially need to agree to play 'winner' and 'loser' in the colluded game. Secondly, even if players collude in this way, they would just gain extra points. However, they are entered to win the lottery based on their agreement with the majority vote, on their relevance assessment submissions not the actual game score. Thus, this should not affect the quality of their relevance assessment submissions. Further players can save a miss on the relevance assessment by means of the points they have earned so far. However, the penalty is quite large, so players would need to play many rounds (at least 4 if they are a hacker and 6 if they are a ninja) just to save a single miss. Thus they would need to collude for a long stretch to see a substantial benefit.
- **Scenario 4: Players may get bored, waiting for the majority decision** The relevance assessments of the player are judged against the majority of the assessments to decide if they are correct. Since it may take a long time to get this majority dynamically, we will typically rely on the latest stored majority assessment for a particular document. This stored value will be updated periodically in the database, as more assessments become available.

  There may be several other subtle caveats in our game design. We believe that these can only be best realized when the game is put open to the public. However, we hope that this does address some of the more obvious attacks against the fairness of the game.

3. INTERACTIVE PROTOTYPE

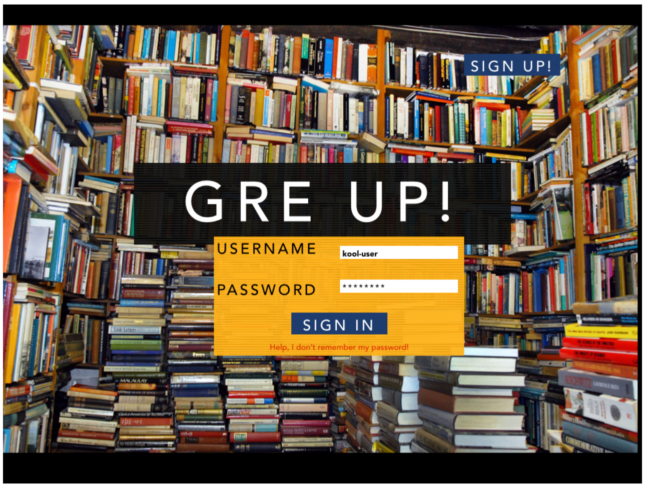
In this section, we describe the design of our prototype for GRE UP. For prototyping we used the *Keynote* app on Mac for designing the screens. For adding interactivity to our prototype, we used the Freemium version of Invision app [2]. Our game prototype is available online at `http://invis.io/281RS4H9W`. Currently the prototype supports the common happy-path scenario through the game. It does not encompass various invalid use cases. We also created a short video demonstrating our game, which is also available online [3]. In this section, we will focus on some of the important aspects of the prototype.

One of the major goals of the game is that it should be very easy and quick for people to sign up for the game. If they

[2] `http://www.invisionapp.com`

[3] `https://drive.google.com/file/d/0B7QjpnRS1mpmdlFuRHhDTnhpV00/view?usp=sharing`

have difficulty in signing up, they may likely not sign up at all. Our homepage is therefore very simple and clutter-free and allows the user to quickly locate the required option. A screenshot of our homepage is shown in Figure 1:

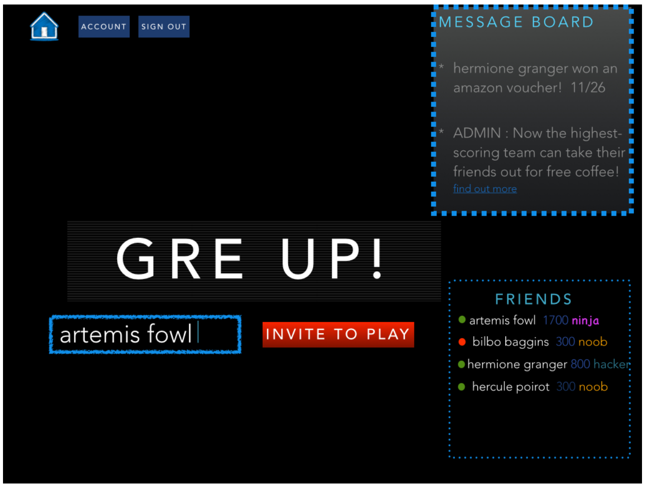


USERNAME kool-user

PASSWORD ********

Figure 1: GRE UP! Welcome screen

Once the user logs-in, she will be taken to her dashboard. The dashboard has several important links that can help her quickly navigate and perform a desired operation. For instance at the top-left of the screen, there are options to help the user change their profile settings, access their account, etc. The lotteries for which a user is currently entered as well as her past winnings, will all be available through her account. At the top-right of the screen, we have a message board.



* hermione granger won an amazon voucher! 11/26

* ADMIN : Now the highest-scoring team can take their friends out for free coffee! find out more

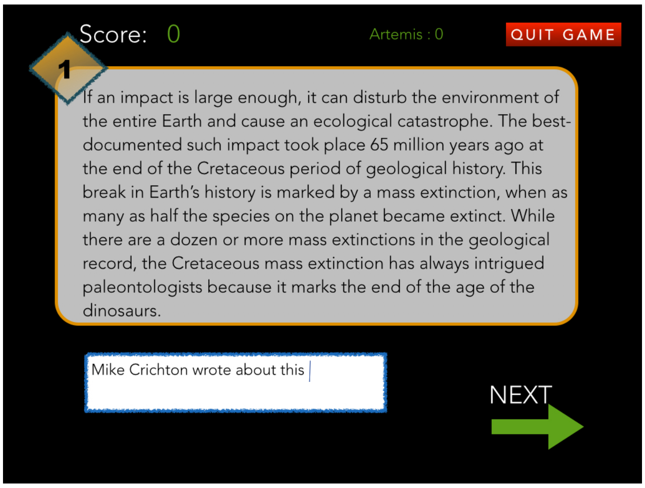
artemis fowl

- artemis fowl 1700 ninja
- bilbo baggins 300 noob
- hermione granger 800 hacker
- hercule poirot 300 noob

Figure 2: GRE UP! User dashboard

The message board is the means to communicate important announcements to the user from the game administrators (admin). However, it may also include messages from their friends such as - 'you played a fantastic game yesterday'. The messages from the admin are generally to inform the user if she or someone else in her network has just won a lottery. This will help in increasing the motivation of the users to play the game. Admins can also broadcast messages regarding policy changes, new upcoming prizes, etc. At the bottom-right of the dashboard, we display a list of friends, that are currently in the user's group. A status is displayed with the names, a 'red' indicating that the person is not available online, whereas a 'green' indicating that the person is online and available to play with. The list also displays the scores and the level for each friend. The user can quickly type in the name of the friend she wants to play with and click on the *Invite to play* button. Their friend will be notified, and once they accept, the user will be notified that the game is now ready to start. Again we have emphasized on keeping the design as simple as possible. To give a feel for the dashboard, we include the screenshot in Figure 2.

Once the game begins, both the players will be presented with a screen(Figure 3) that has a paragraph and a small text-box at the bottom, where they can type in their description of the paragraph. They can then click on *next*, to move to the next paragraph. Once they have finished viewing and entering descriptions for all the three paragraphs, they can send them to their opponent.

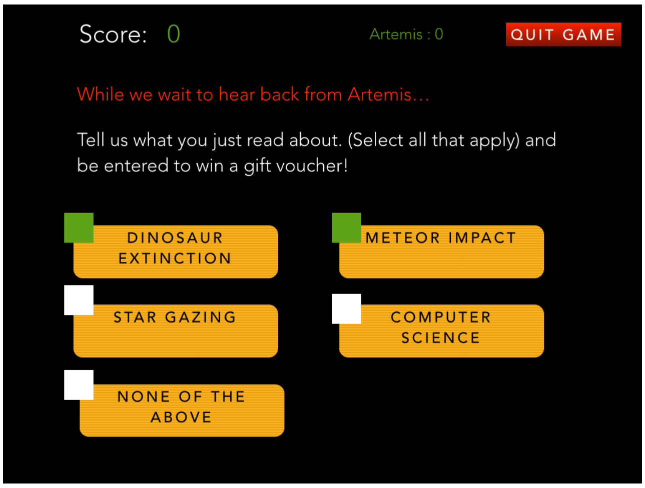


If an impact is large enough, it can disturb the environment of the entire Earth and cause an ecological catastrophe. The best-documented such impact took place 65 million years ago at the end of the Cretaceous period of geological history. This break in Earth's history is marked by a mass extinction, when as many as half the species on the planet became extinct. While there are a dozen or more mass extinctions in the geological record, the Cretaceous mass extinction has always intrigued paleontologists because it marks the end of the age of the dinosaurs.

Mike Crichton wrote about this

Figure 3: GRE UP : Entering a description for the given paragraph

Once the player has submitted her descriptions, and is waiting for the responses from the opponent, we display an interim screen (Figure 4), where she can select the topics that she felt were relevant to the three paragraphs she just read, The user may select more than one topic or the option *None of the above*, if she feels that none of the topics seem relevant to the read paragraphs. We could also provide the same screen to the user once she has finished matching all the descriptions from her opponent. The user could be asked if she wants to make any changes to her previous choices. This may help in boosting the overall accuracy of the task, since by this time the user has likely read the same paragraph a greater number of times. Currently our prototype does not reflect such a design, but we think it could be a worthwhile addition.



Score: 0

Artemis : 0

QUIT GAME

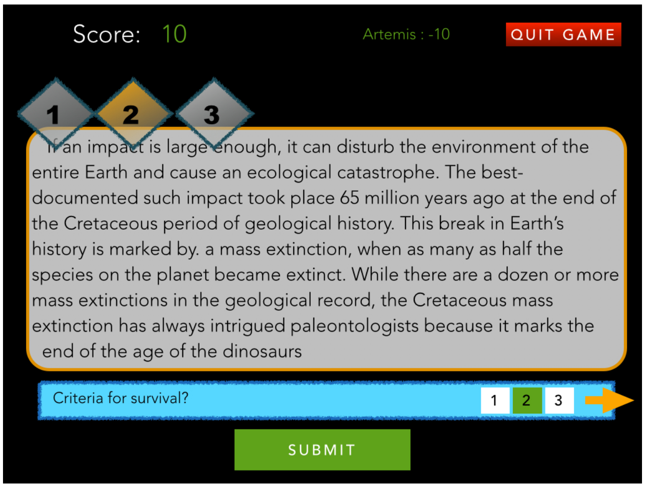
While we wait to hear back from Artemis…

Tell us what you just read about. (Select all that apply) and be entered to win a gift voucher!

- ☑ DINOSAUR EXTINCTION
- ☑ METEOR IMPACT
- □ STAR GAZING
- □ COMPUTER SCIENCE
- □ NONE OF THE ABOVE

Figure 4: GRE UP : Making relevance assessments for the given paragraphs.

Once the descriptions from the opponent are available, they will be made available to the user, though of course in jumbled order. The screen (Figure 5) allows the user to navigate quickly between the three paragraphs and the three descriptions, independently of each other. The user can select the sequence number of the paragraph, which she thinks matches a particular clue. For example in the Figure 5, the user has matched the currently displayed clue with the 2^{nd} paragraph, shown selected in green, at the bottom.



Score: 10

Artemis : -10

QUIT GAME

1 2 3

If an impact is large enough, it can disturb the environment of the entire Earth and cause an ecological catastrophe. The best-documented such impact took place 65 million years ago at the end of the Cretaceous period of geological history. This break in Earth's history is marked by. a mass extinction, when as many as half the species on the planet became extinct. While there are a dozen or more mass extinctions in the geological record, the Cretaceous mass extinction has always intrigued paleontologists because it marks the end of the age of the dinosaurs

Criteria for survival? 1 2 3

SUBMIT

Figure 5: GRE UP : Matching paragraphs to the correct descriptions

During all the stages, the user is constantly shown her score as well as her opponent's score. There is also an option to easily quit the game at any instant. After various stages, such as after the user has submitted her relevance assessments or her matched paragraph-description pairs, there will be a screen displaying the results. The complete walk through can be seen via the prototype available online.

4. EVALUATION

Ideally the following two metrics would be used for evaluating our approach with respect to both **RQ1** and **RQ2**. For **RQ1** compute the accuracy of the proposed method, given the gold standard judgments from TREC and compare it to those obtained by conventional crowdsourced techniques. For **RQ2** compare the percentage of cheat submissions in the proposed model vs the conventional model, using the definition in [4] where a player is regarded as cheating if 67% of his submissions disagree with the majority.

However, given the time constraint of a course project, we believe that a complete implementation of the described game is out of scope. As described earlier, we believe that a game that is accepted popularly by the users is likely to do well for both **RQ1** and **RQ2** [2]. Thus to evaluate our project we conducted a usability survey - by allowing our respondents to explore our interactive prototype. Our respondents were college students and served as a good representation of the actual population at which our game is targeted. Analyzing their feedback on the utility, appeal and other relevant aspects of our game serves as the current evaluation of our game.

4.1 Usability Survey

We conducted an online survey with 103 university students asking them about various aspects of the game. They were provided with a brief description of our game and also shown a video of the working prototype and then asked to answer a few questions. We present the questions and the results below.

4.2 How interested would you be in playing GRE UP! ?

Figure 6 is a pie-chart depicting the statistics of the answers obtained for the above question. When we asked students whether they would be interested in playing our game, we mostly got positive responses (65%). Only 12% of the students said they would not be interested. We believe this is mostly due to the lack of proper understanding of the game and/or limited exposure to the interface of the game. Some students however, specifically mentioned that they did not find any use for this type of game because they had already taken the GRE.

4.3 How often do you think you would find yourself playing this game?

We noticed that we had mostly mixed views on how often people would play this game. We attribute this to the fact that this game is not universally appealing, meaning one would not play such a game 24/7 because the purpose is mostly limited. Maybe, expanding the motivation of the game and not limiting it to just the GRE (and related exams) could help. We also feel that we may have obtained a different perspective, had this game be carried out in a population where a majority of the respondents were not fluent English speakers or struggled with English as a second language. A respondent for whom English is a second language told us that the game would be very popular amongst students from his country.

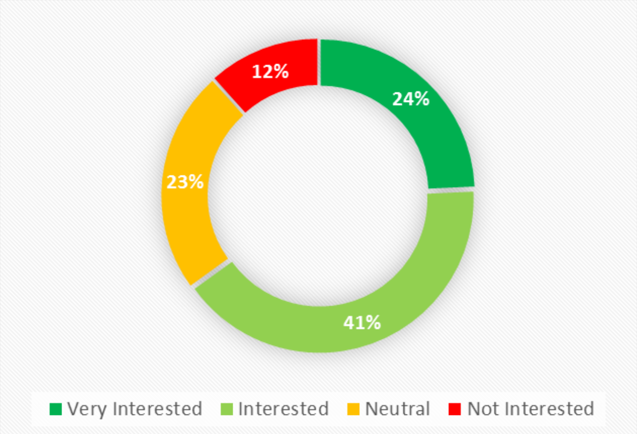


Figure 6: Interest in playing GRE UP!

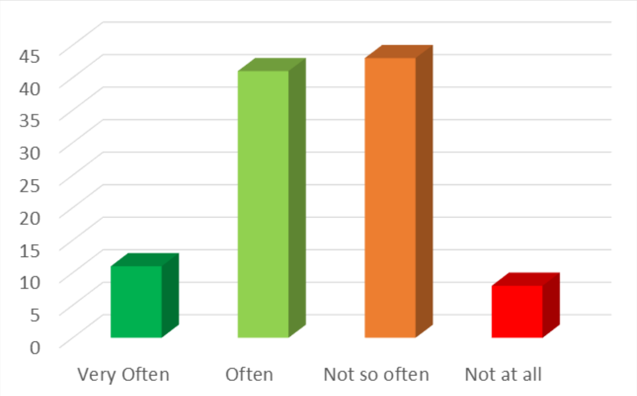


Figure 7: Frequency of playing GRE UP!

4.4 How do you like the design of the game?

80 out of 103 students like the design of our game. Only 4 of them thought it was a bad design. People particularly commented on the fact that they liked the colorful design, especially some of the memes, that would be displayed on the screen, whenever the player won the round.

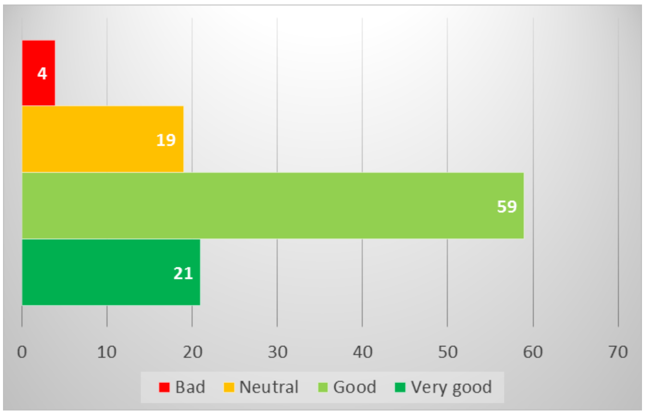


Figure 8: Game Design

4.5 What do you think about the idea of playing this game with your online friends?

65% of the students were for this idea. Of the 22% who were neutral seemed to be confused if they would want to play this game with known people or unknown users online. Amongst some of the comments, we found that people sometimes preferred to play with random people - in case if none of their friends were online - but they still wanted to play. However, people also noticed that opening the game could also increase the chances of unfairness, which we mentioned earlier (Scenario 2). For instance one of the comments goes as follows:

'I feel like there's a potential for a player to leave "troll" descriptions for paragraphs, similar to when people draw completely random objects in Draw Something. This could make the game unsuitable for play against random strangers, and might even make it frustrating within a friends circle'

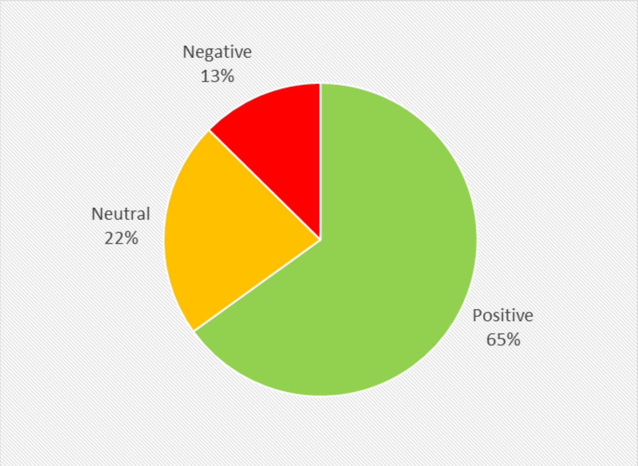


Figure 9: The idea of playing with online friends

4.6 How do you feel about the idea of winning free swag through the course of the game?

As expected, this feature really attracts many players for games. About 80% of the students were interested in winning free swag while playing.

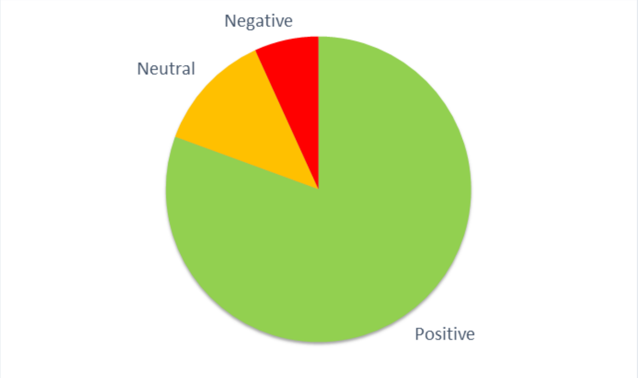


Figure 10: Views on winning free swag

4.7 How do you feel about the idea of preparing for GRE for free using such a game?

We received mixed opinions on this one. Some students were very enthusiastic about this aspect of the game and mentioned that they would play our game just for this. While, others were discouraged by this as they felt like they were being asked to study or do homework through a game. There were a few users who weren't sure how this game could really help in GRE preparation. For instance, one of the comments was:

'It sounds more like homework than a game. Sorry :(',

while another comment read:

'I think the main benefit would be in actually getting people to read (and hopefully think critically about?) several sections of text, which they otherwise wouldn't have done'.

Some other comments also said that they felt the game would be a great exercise in deductive reasoning as well!

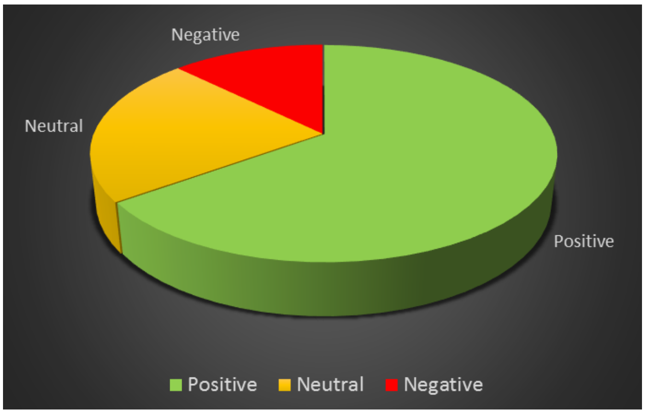


Figure 11: Views on the game helping GRE preparation

4.8 How do you find the difficulty level of the game?

It was very rewarding to learn that only 15% of our users found the game to be difficult. We believe this percentage could be further reduced by refining the description of our game and by letting users actually play the game, once fully developed. However, one of the students did complain that the rules were too complicated to understand and that this discouraged him from playing. We think that this issue may be easy to mitigate by designing a more understandable demo on the game website.

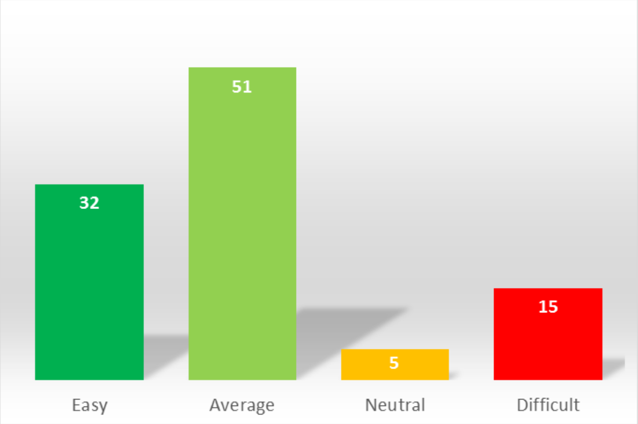


Figure 12: Diffculty level of the game

Overall, the survey results were very promising and in full support of going ahead with the plan for implementing GRE UP!. Many students were very interested in seeing more developments in the game in the future and expressed their willingness to play our game once it is out in the market.

5. FUTURE WORK

While the project helped us to work on several interesting game design issues, we believe that there are several issues that we can realize only when the game is publicly released. Here we describe our possible next steps:

- Implement the game from the prototype making sure that the implementation remains faithful to the design policies we discussed earlier
- We would like to test the implementation to see how the various incentives and scoring schemes in the game affect player behavior.
- Since we would now have a working implementation, we would like to compare the quality of our relevance assessments against the usual methods of crowdsourced relevance assessments, as well as against gold-standard human judges and see how our method fares against them.

6. CONCLUSION

In this work, we discussed our experience in designing a crowdsourced game that would help in making better quality relevance assessments. The major goal of our project was to design a game that would encourage players to make better quality relevance assessments, by making workers entertainment-driven rather than money driven. The goal was also to reduce the number of cheat submissions and make the work more appealing by setting it in a social network environment.

To this end, we described the design of GRE UP! - a crowdsourced game that will motivate users to perform relevance assessments by offering them the advantage of improving their language comprehension skills, while also socializing with friends and having the opportunity of winning free swag. We discussed the yin and yang of various policies in our game design.

Finally, to evaluate our game, we designed an interactive prototype and performed a usability survey amongst more than 100 college students. Our results are fairly encouraging and show that GRE UP! might have a good potential of improving the quality of crowdsourced relevance assessments

References

[1] Roi Blanco, Harry Halpin, Daniel M Herzig, Peter Mika, Jeffrey Pound, Henry S Thompson, and Thanh Tran Duc. Repeatable and reliable search system evaluation using crowdsourcing. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 923–932. ACM, 2011.

[2] Martin Caplan Desurvire, Heather and Jozsef A. Toth. Using heuristics to evaluate the playability of games. In *CHI'04 extended abstracts on Human factors in computing systems*, pages 1509–1512. ACM, 2004.

[3] Carsten Eickhoff and Arjen P de Vries. Increasing cheat robustness of crowdsourcing tasks. *Information retrieval*, 16(2):121–137, 2013.

[4] Carsten Eickhoff, Christopher G Harris, Arjen P de Vries, and Padmini Srinivasan. Quality through flow and immersion: gamifying crowdsourced relevance assessments. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 871–880. ACM, 2012.

[5] Chris Swain Fullerton, Tracy and Steven Hoffman. Game design workshop: Designing, prototyping, & playtesting games. 2004.

[6] Catherine Grady and Matthew Lease. Crowdsourcing document relevance assessment with mechanical turk. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's mechanical turk*, pages 172–179. Association for Computational Linguistics, 2010.

[7] David A. Garvin Lakhani, Karim and Eric Lonstein. Topcoder (a): Developing software through crowdsourcing. In *Harvard Business School General Management Unit Case 610-032*, 2010.

[8] Sergej Zerr Stefan Siersdorfer Markus Rokicki, Sergiu Chelaru. Competitive game designs for improving the cost effectiveness of crowdsourcing. *23rd ACM Conference on Information and Knowledge Management, Shanghai, China*, 2014.

[9] Neil Savage. Gaining wisdom from crowds. *Commun. ACM*, 55(3), March 2012.

[10] Luis Von Ahn and Laura Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326. ACM, 2004.

[11] Luis Von Ahn and Laura Dabbish. Designing games with a purpose. *Communications of the ACM*, 51(8):58–67, 2008.

[12] Henry Von Kohorn. System and method for playing games and rewarding successful players. pages 697–844.