SceneFinder: A collaborative game tha tacit knowledge from crowds

Neha Kalia Agathe Balayn Ujwal Gadiraju Jie Yang

Technische Universiteit Delft
June 26th 2021

The manual process of collecting and labelling data required for machine learning tasks is labour-intensive, expensive, and time consuming. In the past, efforts have been made to crowdsource this data by either offering people monetary incentives, or by using a gamified approach where users contribute to databases as a side-effect of playing an enjoyable game. However, most of these efforts focus on using a competitive setting to incentivize players. This sometimes results in users spamming the dataset for personal gains. Research is lacking in how a collaborative setup, where players work together to make decisions by consensus, can be used to source knowledge that is more accurate and reliable. This paper describes the design and evaluation of SceneFinder, a game that aims to crowdsource reliable and diverse textual data about scenes (such as rooms, parks, monuments, etc) and the tacit knowledge relevant to them, such as information about their contents, their purpose and their surroundings. SceneFinder makes use of a collaborative setup that elicits a relevance based ranking of facts about these scenes, that distinguishes it from existing games in the field.

1 Introduction

Machine learning focuses on the development of computer programs that can access data and use it to learn for themselves (Mitchell, 1997). Models trained by machine learning are increasingly gaining popularity in fields such as healthcare, agriculture, security etc. With this increasing popularity, it is necessary to build machine learning systems that are capable of executing tasks reliably, even in unseen situations. In order to accomplish this, the machine learning algorithms must be trained on large and varied datasets (Gong et al, 2019), which give them both tacit and explicit knowledge.

Explicit knowledge refers to information that can be found on the Web or in readily available databases while tacit knowledge is more intuitive and context-based. According to Rebber (1989, p1), tacit knowledge is "optimally acquired independently of conscious efforts to learn" and "it can be used implicitly to solve problems"and make accurate decisions about novel stimulus circumstances." Tacit knowledge can therefore be called common-sense knowledge, and consists of information that is learnt from context or experience. As a result, tacit knowledge can be challenging to acquire.

Crowdsourcing data in a gamified way can prove to be a cost-effective and time-efficient way to extract a large volume of information. This form of data elicitation revolves around designing a game that is both fun for the players and extracts useful knowledge for the developers. Game developers can use this concept of designing a 'game with a purpose' to "capture large sets of training data that express uniquely human perceptual capabilities" (von Ahn and Dabbish, 2008). The goal of this research is to find a way to crowdsource tacit knowledge in a gamified approach that engages participants - essentially designing a 'Game With a Purpose'. These games try to extract tacit knowledge from participants as a by-product of the participants playing a game. This knowledge can be later used to train machine learning algorithms.

An example of such a game is Verbosity (von Ahn et al, 2006) where relevant information is collected as a side effect of users playing a game they enjoy, or an app called Biotracker (Bowser et al, 2013) which aims to collect data from millennials to contribute to a plant phenology database. These games use a competitive setting to engage players to play more frequently, and contribute effectively to tacit knowledge databases. However, research is lacking in how collaborative games, where players work together and decisions are made by consensus, can be used in tacit knowledge elicitation. The main question that this research aims to answer is - "**How can we elicit tacit knowledge using a**

multiplayer, collaborative, text-based game?" For the purpose of this project, the aim will be to collect tacit knowledge for scene classification.

In this research, we build upon the basic structure of the Verbosity game, and attempt to increase engagement by adding collaborative elements that also contribute to increasing the reliability and diversity of data, and help to obtain a ranking in the relevance of contextual information.

The main question mentioned above will be answered by answering the following sub-questions -

a. How has knowledge been crowdsourced in the past (in a gamified as well as non-gamified way), and what are the benefits/ drawbacks of collecting tacit knowledge in these ways?

b. How can a multiplayer, collaborative, text-based game be designed to elicit information that is diverse as well as reliable?

c. How can the game be designed to be engaging enough to get a large volume of tacit information?

d. What benchmarks can be used to evaluate the quality of the collected data? And how well does the game perform its intended task?

This paper attempts to answer these questions and discusses the problem definition in detail in Section 2, followed by a discussion of related works in section 3. Section 4 elaborates on the design of the game, while Section 5 details the evaluation of SceneFinder and its results. Section 6 discusses the limitations of the game and suggests possible improvements, and Section 7 provides a review of the ethical concerns regarding this research.

2. Problem Definition

In order to gain an accurate and complete overview of scenes and the contextual relationships around data such as their contents and their purposes, SceneFinder must be designed keeping the following goals in mind -

a. Reliability - Getting reliable data from crowdsourcing tends to be a problem due to human error, as well as due to the presence of insincere/ malicious users who compromise the data quality due to either sloppiness or because they try to maximise their own profits by cheating (Eickhoff et al, 2012). SceneFinder must account for this and have design elements that ensure that the data is not polluted by these malicious users.

b. Diversity - The data collected needs to be diverse, and not limited to certain obvious and repetitive responses that a number of players might share. For example in the context of a 'kitchen', multiple sets of players are likely to say 'It is used for cooking'. This information is useful, however, we do not want our dataset to be limited to just this piece of concrete information, and would like to obtain a larger and more varied set of information.

c. Relevance based ranking - It should be easy to find out which information is more relevant to a given scene compared to others. For example, if our context is a kitchen, then 'food' should be ranked higher than 'table'.

d. Ease of integration into a database - Crowdsourced data often needs a significant amount of pre-processing in order to be usable in the context of machine learning. While data mining and natural language processing can be used to extract information from freely typed text, we intend that the information sourced from this game must be in a format that makes extracting relevant features as efficient as possible.

Keeping these considerations in mind, the detailed design of SceneFinder is explained in section 4.

3. Related Work

Numerous efforts have been made to crowdsource both tacit and explicit knowledge throughout the years. Games such as Verbosity, Peekaboom and ESP aim to obtain a large volume of knowledge by using a gamified version of crowdsourcing Verbosity (von Ahn et al, 2006) is a game specifically targeted to collect text-based tacit knowledge from users. It uses concepts from the traditional guessing game 'Taboo' to create a guessing game where users play in pairs to enable each

other to guess a phrase based on hints. These hints are proposed by a ‘narrator’ and used by a ‘guesser’.

Verbosity makes use of sentence templates in order to reduce data randomness and decrease post-processing required to use the collected knowledge. It focuses on collecting a large volume of ‘true’ facts. It tries to verify the usefulness and independence of the data collected by implementing a single player version of the game where a player must guess a phrase based on previously submitted hints. However, verbosity does not attempt to organize the collected facts by any metric, such as the relevance of collected phrases to each other. While it is capable of collecting a large volume of facts, these facts are not directed towards any specific purpose.

ESP (von Ahn and Dabbish, 2005) and Peekaboom (von Ahn et al, 2006) are games designed to label picture based data. The ESP game aims to pinpoint the objects contained in an image by using a game that displays random images from the web and outputs labels for the contents of these images. The Peekaboom game takes this a step further by determining exactly where in an image an object is located.

Peekaboom uses a collaborative setup where players play in groups of two and try to help each other guess an image by gradually revealing parts of the image. Though both these games rely on image based data rather than text based information that is the focus of our research, the inherent aspects of crowdsourcing remain the same, with the inherent purpose to collect large volumes of accurate data.

While all these games may be classified as ‘collaborative’, since players work with each other to reach an end goal (guessing an image or a word), they do not take into consideration the effect of collaboration with regards to teamwork. For example, what happens if instead of playing one-on-one, the users of these games play in groups and get a chance to interact with each other in some way that may prove beneficial to the information collected? Would this improve the quality of data?

Additionally, these approaches do not provide a way to compare the relevance of the collected data. If a large pool of information is collected via crowdsourcing, there is no concrete way to determine which points in this information pool are more relevant than others.

The research presented in this paper aims to tackle these two shortcomings of previously executed approaches in the context of obtaining a dataset to describe household and external scenes, by extending the template-based guessing-game concepts from the Verbosity game and extending them in a group-based collaborative setup.

4. Design of SceneFinder

SceneFinder is a text-based, multiplayer game, where players collaborate with each other to reach an outcome. This game was designed according to the specifications mentioned in section 2 to obtain a text-based tacit knowledge database for scene-based contexts, and has been structured as a guessing game, inspired from principles of games such as Taboo and Verbosity, as mentioned in section 3 of this paper.

4.1 Game Flow

SceneFinder is played in groups of four, where three players act as ‘narrators’ and one player acts as the ’guesser’. The narrators get access to a scene (for example - ‘garden’), and they must describe this scene to the ‘guesser’ by giving hints based on certain predefined sentence templates, as seen in figure 1.

When every ‘narrator’ has submitted a hint, all the hints are displayed to them, and they can vote for the hint they believe is the most relevant (however, a player cannot vote for a hint they themselves proposed). The hint with the most votes is sent to the ‘guesser’ who tries to guess the scene based on this hint. In case of a tie, an arbitrarily chosen hint is presented out of all submitted hints. This process (subsequently referred to as a ‘round’ of the game) is repeated until the ‘guesser’ correctly guesses the scene, or until 6 rounds of hints have taken place. In this way, the game follows the pattern laid out in the Verbosity game, where users aim to make each other guess objects based on hints derived from templates. However, SceneFinder differs from Verbosity in the sense that instead of being played in pairs with one guesser and one narrator, SceneFinder has three narrators that collectively contribute hints and determine the most relevant hint among themselves by using a vote-based consensus strategy. This enables us to determine a relevance-based ranking of how concepts relate to contexts.

4.2 Choice of sentence templates

Sentence templates are used for hints to ensure that the hints given can be easily used in machine learning tasks without the use of extensive post processing. Each narrator can submit a hint based on the following templates:

a. It contains …
b. It is used for …
c. It is surrounded by…
d. Is it indoor/ outdoor?

The first two of these templates are drawn from Verbosity. Out of the six sentence templates from Verbosity, we selected these two for the purpose of SceneFinder since they are especially indicative of tacit knowledge with regards to a specific scene. The 'it contains' template provides information about the contents of a scene. For example, a kitchen contains a stove. The 'it is used for' template contributes information about the purpose of a scene, such as the fact that a bridge is used to cross a river.

The 'is it indoor/ outdoor' and 'it is surrounded by' templates are also extensions of the verbosity template 'it is near/ in/ on'. Since this verbosity template was designed with random words in mind, instead of just scenes, a modification was necessary to ensure relevance of data.

The players can fill in these templates freely, upto a limit of 120 typed characters.

4.3 Scoring for the game

The 'narrators' get points based on the product of the number of votes they get and the weight for the current round of hints. The weights awarded for each round of hints decreases linearly with subsequent rounds. So hints in the first round get 10 points per vote, those in the next round get 9 points and so forth. The 'guesser' gets points when he correctly guesses the scene.

This scoring format ensures that the 'narrators' give the most relevant hints at first, since they get higher points if they get more votes in the initial rounds. It also incentivizes the guesser to make accurate guesses faster, since the longer s(he) waits, the more time the 'narrators' get to accumulate points.

This manner of a collaborative-competitive environment should return optimal results, since all parties involved are incentivized based on the speed and relevance of their contributions. The voting system also ensures that the players correct each other throughout the progress of the game. If a player enters an irrelevant or nonsensical hint, it is likely that this hint will not receive votes from the other players in the room, and will therefore be discarded and given low priority in the final ranking of information, as explained in section 4.5.

Figure 1: Template selection screen

4.4 Obtaining a ranking based on relevance

The hints submitted by the users of the game are ranked in order of relevance based on the points they obtain.

The simplest metric to calculate this ranking is to aggregate the data collected from different game sessions and sum up the votes collected for each unique hint. This would result in a ranking where highest ranked hints are the ones that got the highest approval from their peers and that were most intuitive to people (since hints collected in the first round of guessing are given higher weights than those collected in subsequent rounds).

4.5 Accounting for malicious inputs

Since the final ranking for the data is obtained by aggregating the hints from across different sessions of the game, it can be intuited that irrelevant data gets a lower rank, and we can define point thresholds below which hints can be discarded. For example, if for a scene 'kitchen', someone entered a hint 'It contains Donald Trump', there's a chance that 'narrators' in the same game room find this funny and therefore give it the highest number of votes, but when we aggregate the data from across rooms, the chances are quite low that someone else gave a hint with the same response, so this would not rank very high in the relevance based ranking. A relevant hint such as 'It contains a fridge', on the other hand, would be ranked higher since more people will likely enter this across game sessions.

The motives of the players must also be accounted for here. For example, if the 'narrators' get competitive with each other and notice that 'Narrator X' is getting more points than the others, they might choose not to vote for 'Narrator X's' hint in the next round. Keeping this in mind, we make sure that players only see the points they get, and do not get any prompts about which of the 'narrators' has a leading number of points. Additionally, the 'narrators' cannot see who the hints were proposed by, thus ensuring fairness of voting.

The players are randomly assigned to game rooms and cannot choose to play SceneFinder with their friends. While this reduces game engagement, it helps in avoiding inputs that might pollute our dataset. For example, if a group of friends plays this game, they may hint that 'This contains a dartboard' for a 'kitchen' scene, simply because one of their kitchens contains a dartboard. If such inputs occur in large volumes, this would greatly impact data quality. Therefore, for the purpose of this research, players are sequentially added to the next available room, and cannot choose to play SceneFinder with groups of people they are familiar with.

This sums up the design details of SceneFinder, along with explaining how it satisfies the requirements mentioned in section 2. The next section discusses the evaluation of the results of the game, and describes the post-processing of the data.

5. Evaluation of Results

The results of this research are evaluated in the context of engagement and data quality. Section 5.1 describes the setup under which the game was evaluated, followed by an explanation of the evaluation metrics in section 5.2, and a detailed discussion of the results in section 5.3.

5.1 Experimental Setup

For both these purposes, 3 sessions of the game were played by 12 users. These users were in the age group of 18-50, and were from different parts of the world with education levels varying from university students to medical practitioners.

Each game consists of a sequence of 12 scenes. In a real-world scenario, there would exist a much larger database of scenes, which are randomly assigned to players in the game. However, for the purpose of this evaluation, we stick to a sequence of 12 pre-selected scenes in order to analyze the data and compare player inputs across different games.

This setup is optimal for comparing data across games, since it helps us compare the hints given during different games for the same scene. We can therefore analyze the reliability and diversity of these inputs. If each game session got random scenes assigned to it, then these comparisons for the same scene would not be possible.

The scenes chosen for this evaluation can be categorized into the following classes:

a) Household (kitchen, living room, bathroom)
b) Recreational (club, restaurant, cafe, amusement park)
c) Educational (classroom, university, library)
d) Practical (bank, market)

These categories were chosen to evaluate subtle differences between concepts belonging to the same category (for example, a classroom and a university are both used for education, so what separates one from the other?), and to analyze the quality of the information obtained across categories.

5.2 Evaluation Metrics

Engagement was evaluated by asking the players to fill out a questionnaire at the end of the game, based on the Game Engagement Questionnaire (GEQ) developed by Brockmyer et al (2009). This scale aims to measure the flow, absorption and immersion of the participants while playing the game. The GEQ uses a 19-point questionnaire to measure engagement. However, since it was aimed to analyze the impact of violent video games, some of the questions used such as 'I feel scared' and 'the game feels real' have been omitted for the purpose of this research. The results from the questionnaire are summed up in Table 1.

The quality of the dataset obtained from these game sessions was evaluated based on the diversity and volume of data obtained per game. Some of the metrics used for this evaluation are:

1) Quantity of data obtained per game session (used to measure data volume)

2) Uniqueness of data obtained per game session (used to measure data diversity)

3) The accuracy of the submitted hints (used to measure if the data is actually reliable)

5.3 Evaluation Results

Out of the 12 players who filled out the game engagement questionnaire, 83.4% said they were likely to play the game again, while 58.4% said they were likely to recommend it to a friend.

The game ranked well in terms of immersion, flow, and presence as measured according to the responses on the 5-point GEQ scale. The absorption of the players into the game was comparatively lower. This could be a result of the fact that the game lacks a storyline and graphics that are included in a large variety of video games (for example mission driven first person shooter games). These factors are known to increase player absorption in video games. The responses of the players to the questionnaire have been summarized in table 1. The detailed list of questions that were included in this questionnaire are included in Appendix 1.

A total of 247 facts were collected across 3 game sessions (the time spent per game was approximately 18 mins). That brings the average number of facts collected per game to be 82.3

The accuracy of these facts was remarkably high. Out of 247 facts collected, we randomly sampled 50 facts. 98% of these randomly sampled facts were reported to be true when checked by us.

Metric	**Score (Normalized to 0-1)**
Flow	0.60
Immersion	0.85
Absorption	0.47
Presence	0.72

Table 1 - Results of Game Engagement Questionnaire

An interesting result obtained was the ranking of the hints in relation to the scene. Due to the scoring system, it was possible to obtain a ranking of relevant common sense facts about each scene. An example of the hints collected for the 'cafe' scene is recorded in table 2. Note that this table consists of a summarized set of hints that help to emphasize the point of a relevance based ranking (as discussed in the next paragraph), and the complete data can be seen in Appendix 2.

It can be seen in Table 2 that facts that are more relevant to cafes, such as the fact that they contain coffee and are used for coffee dates were suggested by multiple people, and therefore obtained a higher ranking. On the other hand, wrong facts such as 'it contains chapel' ranked comparatively lower. Messages that are incomplete such as 'it contains' (the last entry in the table) received no votes at all, and is therefore ranked last. This vote-based ranking can thus prove useful in determining which facts to take into account while utilizing the data collected from SceneFinder for machine learning tasks. Qualitative analysis can be used to set a benchmark for the number of votes below which facts will be discarded.

Facts collected for CAFE	Score
It contains coffee	45
It is used for coffee dates	38
It is surrounded by people	30
It is used for ordering coffee	30
It is used for socializing	20
It is used for socialising around a coffee	20
It contains idols	20
It contains coffee cups	20
It contains chapel	10
It is surrounded by trees	0
It contains	0

Table 2 - Summary of facts collected for Cafe

Overall, it can be concluded from these results that the game proves to be an engaging experience for the players. It is also an extremely reliable way of collecting tacit knowledge, with a high accuracy rate and a well defined ranking of the relationship between scenes and their related concepts. However, the diversity of the data collected over game sessions reduces as more and more sessions are played. We therefore end up with a small set of common sense facts that are accurate and reliable, but not extremely diverse.

6 Limitations and Future Improvements

The most important limitations of SceneFinder were in terms of the type and the diversity of knowledge collected. The game was designed to collect tacit knowledge over the course of it being played. We assumed that it would take players a number of rounds to guess each scene; a direct consequence of this would have been that the most obvious hints were taken out of the way in the initial rounds and succeeded by more subtle hints. However, most ‘guessers’ were able to guess the scene within the first 2 rounds, which meant that the results of the game sometimes failed to cross the barrier into more implicit knowledge. This results in a setup that is capable of collecting reliable data, however, the data is not diverse, and is not entirely composed of tacit knowledge. For example in table 2, it can be seen that the data collected includes tacit as well as explicit knowledge. Hints such as ‘it contains coffee’ and ‘it is used for ordering coffee’ refer to explicit knowledge, whereas information about how cafes are ‘used for coffee dates’ is an example of tacit knowledge.

A possible improvement to gear the game more towards collecting tacit knowledge could be introducing the concept of ‘taboo words’. This can be the list of the five highest ranked words in the database for a particular hint template. The ‘narrators’, while giving hints, will not be allowed to use these words. This could make the game more engaging for the players, since they need to be more creative with their answers. This can also improve the diversity of the data that we collect. Due to time constraints, this feature has not been included in the current version of SceneFinder.

To verify this assumption, we wanted to test the effect taboo words had on the quantity of tacit knowledge. To do this, we conducted an additional game session with 4 players and the ‘cafe’ scene, and instructed the ‘narrators’ to provide hints that did not include any words from the top 4 hints of table 2. The results of this ‘taboo-simulation’ game session are catalogued in Table 3.

Facts collected with taboo simulation for CAFE	Score
It contains non-alcoholic drinks	20
It is surrounded by confused tourists thinking they can buy weed here	19
It contains foamy drinks in cups	10
It is used for buying drinks to keep you awake	9
It contains pastries and drinks	0
It is used for meeting tinder matches in person	0

Table 3: Facts collected for cafe in taboo simulation

The results in table 3 indicate that using taboo words increases data diversity, as well as the quantity of tacit knowledge collected. The hints provided in his taboo simulation of SceneFinder contain more subtle,

context-based information than that obtained from the existing implementation of the game, suggesting that this could be a promising future improvement. However, more game sessions must be conducted to validate this concept further.

Another improvement could be in terms of game engagement for the 'guesser'. The 'guesser' only gets points when he guesses the scene correctly. To keep him/her motivated, it might be a good idea to implement SceneFinder in such a way that every game consists of multiple sessions, where each player gets to be the 'guesser' once. This evens out the playing field and ensures the players are motivated to answer correctly.

7 Data Integrity, Reproducibility and Ethics

To maintain the integrity of the data collected. all game sessions played so far have been logged and stored into a MongoDB database, meaning that it is possible to walk through every step taken by the players and derive insights from them if necessary. This leaves open the opportunity of using the information collected from this game for future analysis and interpretation.

The experimental setup described in section 5.1 is easily reproducible; the design of SceneFinder has been clearly and completely described in this paper, and can be implemented by future researchers. The source code for the game will be made publicly available on GitHub. While the experimental setup itself is simple to replicate, the exact data collected through the game will naturally vary depending on the players of the game, since they are the ones who contribute to the dataset.

Taking ethical constraints into account, we made sure that every player who participated in this game was aware of the purpose of the game, and of the fact that the data collected from them would be analyzed during this research. Additionally, the data was collected anonymously to avoid any privacy related concerns, and to remove any biases from the players regarding having their names associated with the facts contributed by them.

8 Conclusion

SceneFinder is a game designed to crowdsource reliable, diverse textual data in the domain of tacit knowledge regarding scenes. It builds up on the principles of Verbosity and incorporates elements of collaboration to improve data quality and increase its reliability.

SceneFinder proved to be an enjoyable game for the players, ranking high on engagement and immersion, and contributed a number of accurate facts over an extremely limited quantity of game sessions. While the quality of the obtained data somewhat lacked diversity, the improvements suggested in section 6 might prove to be a solution to this problem. SceneFinder can thus prove to be a useful tool to extract reliable data for a large variety of scenes.

9 References

Mitchell, Tom (1997). *Machine Learning*. New York: McGraw Hill. ISBN 0-07-042807-7. OCLC 36417892

Z. Gong, P. Zhong and W. Hu, "Diversity in Machine Learning," in *IEEE Access*, vol. 7, pp. 64323-64350, 2019, doi: 10.1109/ACCESS.2019.2917620.

Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General, 118*(3), 219–235. https://doi.org/10.1037/0096-3445.118.3.219

von Ahn, L., & Dabbish, L. (2008). Designing games with a purpose. *Communications of the ACM*, 51(8), 58–67. https://doi.org/10.1145/1378704.1378719

von Ahn, L., Kedia, M., & Blum, M. (2006). Verbosity. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '06. the SIGCHI conference*. https://doi.org/10.1145/1124772.1124784

Bowser, A., Hansen, D., He, Y., Boston, C., Reid, M., Gunnell, L., & Preece, J. (2013, October 2). Using gamification to inspire new citizen science volunteers. *Proceedings of the First International Conference on Gameful Design, Research, and Applications. Gamification '13: Gameful Design, Research, and Applications.* https://doi.org/10.1145/2583008.2583011

von Ahn, L & Dabbish, L. (2005). ESP: Labeling Images with a Computer Game.. 91-98.

von Ahn, L., Liu, R., & Blum, M. (2006). Peekaboom. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '06. the SIGCHI conference. https://doi.org/10.1145/1124772.1124782

Eickhoff, Carsten & Harris, Christopher & de Vries, Arjen & Srinivasan, Padmini. (2012). Quality through flow and immersion: Gamifying crowdsourced relevance assessments. *Signal Processing-image Communication - SIGNAL PROCESS-IMAGE COMMUN*. 871-880. 10.1145/2348283.2348400.

Brockmyer, J. H., Fox, C. M., Curtiss, K. A., McBroom, E., Burkhart, K. M., & Pidruzny, J. N. (2009). The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing. *Journal of Experimental Social Psychology,* 45(4), 624–634. https://doi.org/10.1016/j.jesp.2009.02.016

Appendix 1: Game Engagement Questionnaire

The following questionnaire was used to measure the game engagement of players with reference to SceneFinder. The players were given a set of questions and were ask to respond on a scale of 1 to 5 about how much they related to the sentence while playing the game, with one being ‘No’ and 5 being ‘Yes’.

1. I lose track of time
2. My thoughts go fast
3. Things seen to happen automatically
4. I play longer than I meant to
5. I feel spaced out
6. I lose track of where I am
7. Time seems to stand still or kind of stop
8. If someone talks to me during the game, I can’t hear them.
9. I get wound up
10. I don’t answer when someone talks to me
11. I can’t tell that I’m getting tired
12. Playing seems automatic
13. I play without thinking about how to play
14. I feel like I just can’t stop playing
15. I really get into the game
16. I am likely to play this game again
17. I would recommend this game to friends.

Appendix 2: Cafe Scene Complete Hint Table

Fact	Score
It contains coffee	45
It is used for coffee dates	38
It is surrounded by people	30
It is used for ordering coffee	30
It is used for hanging out with friends	30
It is used for socialising around a drink	20
It is used for socialising around a coffee	20
It is used for socializing	20
It is used for drinking coffee with friends	20
It contains idols	20
It contains coffee cups	20
It contains tables	18
It is used for coffee	10
Indoor/ outdoor indoor	10
It contains people that drinks hot stuff	10
It contains coffee machines	10
It contains chapel	10
It contains food	8
Indoor/ outdoor space to socialise and drink latte	0
It is surrounded by	0
It is surrounded by trees	0
It is used for dates	0
It is used for drinking coffee	0
It is used for ordering drinks	0
It contains cafeine	0
It contains	0
It is surrounded by hangouts	0

Appendix 3: Screenshots from the game

Figure 1: Narrator's screen to select hint template

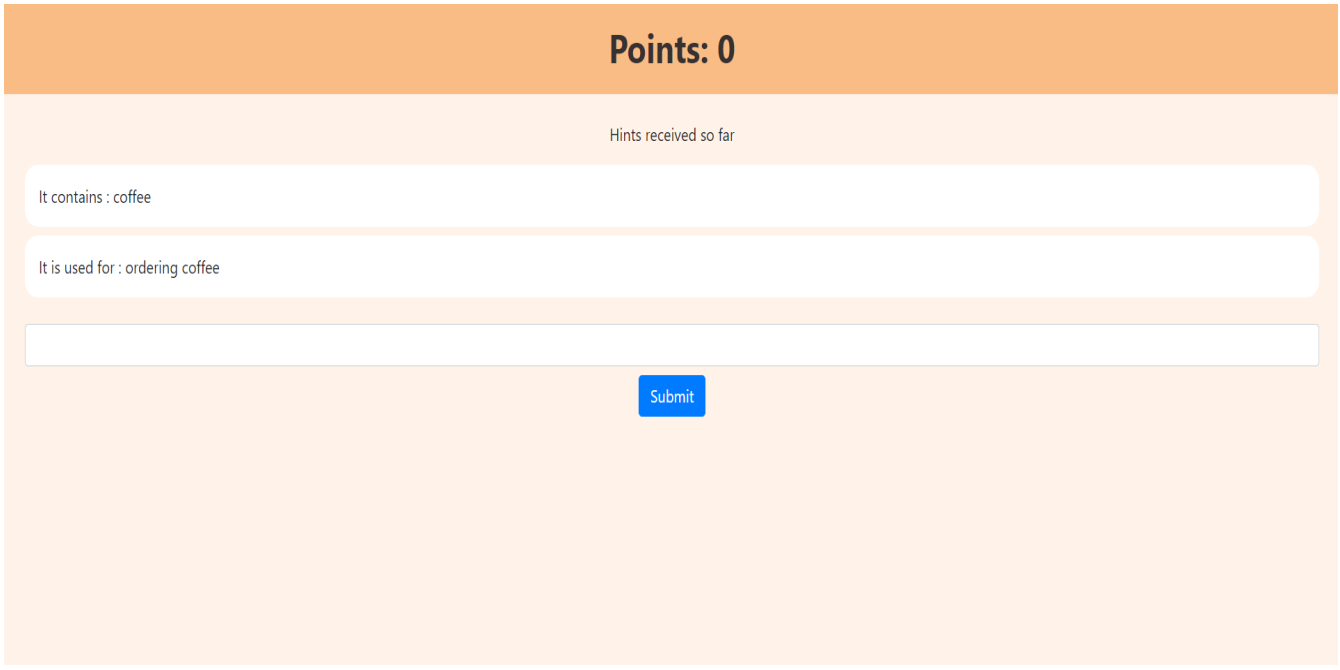

Figure 2: Guesser's screen to see hints and submit guesses

Figure 3: Narrator's screen to fill out a hint template

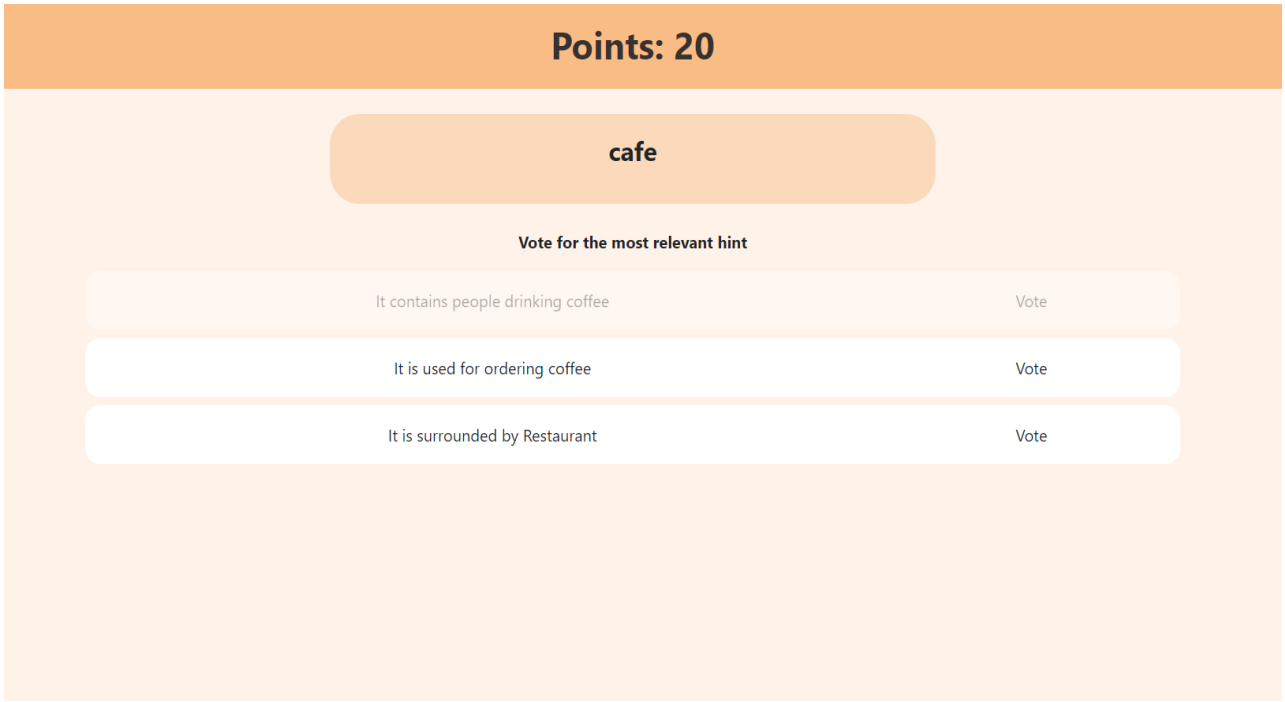

Figure 4: Narrator's screen to vote for most relevant hint