A Survey of Crowdsourcing Methods for Commonsense Knowledge Collection

Ilinca Elena Ioana Renţea
Supervisor(s): Gaole He, Ujwal Gadiraju, Jie Yang
EEMCS, Delft University of Technology, The Netherlands
22-6-2022

A Dissertation Submitted to EEMCS faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering

Abstract

Commonsense knowledge is information that all humans own and use to interpret common situations and react to them accordingly. This kind of information is necessary for the training of artificial intelligence models to reach a performance as close as possible to human performance. Researchers have developed methods that use crowdsourcing to collect this kind of knowledge from the general public. This research focuses on systematically surveying the existing literature about these methods. We created a taxonomy to describe and compare the existing work based on the following three measures that were the most common ones reported: efficiency, cost, and quality.

1 Introduction

Commonsense knowledge is a set of information typically possessed by all humans. This kind of knowledge helps people understand and be able to react to common situations. Examples of commonsense are "if you take the tray out of the heated oven with your bare hand you will get burned" or "if you stay outside while raining, you will get wet". The fact that this knowledge is considered widely known results in its exclusion from written or oral communication. [13] The main approach to collecting commonsense knowledge is through crowdsourcing, automatic or semi-automatic extraction. Since the knowledge is usually omitted, the first category is the preferred one because it relies directly on humans. As suggested in [8], commonsense knowledge is highly needed for machine learning models to be able to perform a task comparable to a human in domains such as "natural language processing, vision, and robotics" [8].

Although there has been advanced research in collecting commonsense knowledge using crowdsourcing methods, no existing work systematically surveys and compares the methods based on their efficiency, cost ,or quality of gathered knowledge. This gap in this research field motivates this research paper, as it contributes by systematically surveying the literature about the existing methods for commonsense knowledge collection with a focus on crowdsourcing methods.

Therefore, this research paper attempts to answer the following research questions:

- What do existing crowdsourcing methods do to collect commonsense knowledge?
- How efficient, costly, and accurate are existing crowdsourcing methods to collect commonsense knowledge?

In order to answer the mentioned research questions, the existing literature will be surveyed and the paper will create a taxonomy and describe the developed work in this field of collecting commonsense knowledge. Moreover, the methods will be compared using the following measures: efficiency, cost, and quality. The mentioned measures were selected for the analysis of the methods because they are the most common measures reported in the literature associated with the systems. Also, the efficiency of the methods influences the collection of commonsense knowledge and a more efficient system is preferred in order to collect as much data in as little amount of time as possible, while the quality should be as high as possible. More than that, the cost is considered since a lot of data has to be collected and a lower cost would be preferred.

The structure of this paper is the following: Section 2 describes the methodology used for the survey paper. Section 3 categorizes and gives details about the crowdsourcing methods that were found in the existing literature. Section 4 focuses on the analysis of the listed methods using the efficiency, cost and quality metrics. More than that, section 5 highlights the main findings, limitations and implications of the research. Section 6 identifies some future improvements that can be made to the paper, and draws conclusions. Lastly, section 7 details the ethical and responsible research implications of this research paper.

2 Methodology

This research proposes a systematic survey of previous research on crowdsourcing methods for commonsense knowledge collection literature that would be the base of the comparison between the methods. This section will detail the methodology of the systematic survey including search keywords identification and search strategy.

2.1 Search Keywords Identification

The first step in the systematic survey was to identify keywords to use in the search queries. The research is focused on commonsense knowledge collection through crowdsourcing methods and their comparison based on efficiency, cost, and quality. Therefore, the keywords used in the queries are variations of the terms: commonsense knowledge, crowdsourcing, efficiency, cost, performance, throughput, and quality. Using these terms, the following search query was constructed: *("commonsense* knowledge" OR "common sense* knowledge") AND ("crowdsourcing*" OR "crowd sourcing*") AND ("efficienc*" OR "cost*" OR "throughput*" OR "performance*" OR "qualit*")*. This query includes different spellings of the same words and takes into account different endings of possible words that can be relevant to the topic.

2.2 Search Strategy

After identifying the keywords and building the search query, these were used on Google Scholar, which returned 240 results for the mentioned query. The search results that the engine returned were manually checked to fit the research topic and were added to a list. Besides the search results, literature found in the list of references of selected papers was also considered and added to the list if they were relevant to the topic. This method of retrieving literature was especially practical in the case in which the initial paper was a literature survey as well.

After performing this search, the list of 53 references was added to the Mendeley reference manager. This system was chosen for managing the references because it generates the bibliography and it offers functionalities such as labelling the literature, reading the papers within the application, marking and making notes on different sections of the papers. This system was used to categorize the literature using some labels. Therefore, each paper was labeled using: *Background, GWAP, Knowledge Acquisition, Multiplayer, Single player, Knowledge manual extraction, Knowledge manual confirmation, Cost, Efficiency, Quality*. Some of these categories are sub-categories of others, such as Multiplayer with GWAP. This is a result of the categorization being done gradually, the highest levels being defined first and then divided into lower levels based on the identified crowdsourcing methods.

Labeling the identified literature was accompanied by excluding the papers that were added to the list, but were declared out of the scope of this research. Moreover, the labeling was followed by a detailed read of the papers, summarizing the findings and drawing observations based on the consulted literature.

3 Crowdsourcing Methods for Commonsense Knowledge Collection

Crowdsourcing methods are one of the possible approaches to collecting commonsense knowledge. In order to compare the crowdsourcing methods, these were split into multiple categories based on a closed sorting approach. The crowdsourcing methods can be split into two categories on the highest level of abstraction that were defined based on two aspects that differentiate the methods: the level of entertainment they provide to the users and if the purpose behind the system is presented to the user. Therefore, the systems that provide entertainment to the users and usually hide the purpose that represents the motivation of the system can be categorized as games with a purpose (GWAP). On the other hand, the methods that directly ask the users to provide knowledge tuples or confirm already collected tuples and have a clear purpose presented to the user from the start can be categorized as knowledge acquisition systems. These two categories were chosen as the highest level of abstraction based on the categorization of some literature [37], and because the level of entertainment and the purpose of the system could influence the analysis measure, especially the efficiency and cost of the systems. The level of entertainment directly influences the cost of the system since the games do not provide any financial compensation to the users. Also, if the purpose of the system is presented to the user, it might affect the efficiency and the accuracy of the systems. The efficiency could be influenced because some users might not be interested in helping collect commonsense knowledge and might prefer playing a game instead. Also, the quality of collected data could be influenced if the purpose of the system is mentioned because the users of a game could provide incorrect answers if they did not know that the data is collected. This categorization is reflected in figure 1, which also includes the lower levels of abstractions.

3.1 Games with a Purpose

Games with a purpose are computer games that try to entertain the players while using their knowledge to solve problems in an efficient way [37]. This method of collecting commonsense knowledge is based on people's desire to be entertained in order to attract users to contribute with knowledge. Therefore, even if this method usually does not give financial rewards to the users, it can be still considered attractive by those because of people's need for entertainment.

A further categorization of the collected games with a purpose can be done by splitting them into single-player or multiplayer games, as figure 1 suggests. This categorization is relevant for this research because the number of players can define some of the characteristics of the games, such as the type of knowledge collected or the purpose of the game. Therefore, this categorization does not affect the analysis between the methods, but it could affect the collected knowledge.

Single-player

Single-player games are similar to the previous category, the main distinction being that the games require only one player. However, this difference can play a role in the architecture of the game and the method in which it collects the knowledge. Therefore, the absence of a second player determines most of the games to be quiz-type games.

The identified examples of single-player games are:

- **Robot Trainer** [25] is a single-player game, where the user's goal is to teach a robot how to answer simple questions. The game includes three levels: *elementary* (creating CSK rules), *advanced* (choosing preferred rules for a given narrative) and *examination* (checking collected CSK and proposing changes which will be used in the first level).
- **Common Consensus** [19] is a single-player quiz game, where a player receives a question and has to give as many words associated with the question as possible.
- **Knowledge Coder** [26] is a single-player game where the users encode human knowledge in a structured way.
- **Concept Game** [12] is a single-player game where users have to verify already collected candidate assertions.
- **Virtual Pet Game** [15, 21] is a community-based game where users input commonsense knowledge using a predefined structure, ask questions ,or vote answers from other players.
- **Story Sense** [24] is an interactive learning environment, which gives story templates to children and collects commonsense knowledge from the received answers.

With this list of single-player games in mind, it is worth mentioning some characteristics that can be used to split the games into several categories.

Type of knowledge collected. Firstly, the type of knowledge that is collected is an important characteristic to analyze. All games, except for Story Sense, collect commonsense knowledge using a pre-defined template for the questions. For example, Common Consensus asks questions that are constructed from a term and a relation and expects the player to give the correct second term, while the Concept Game shows a knowledge tuple to the user and if the player considers it correct, it is collected. Although, Story Sense is using story templates to collect the knowledge from children.

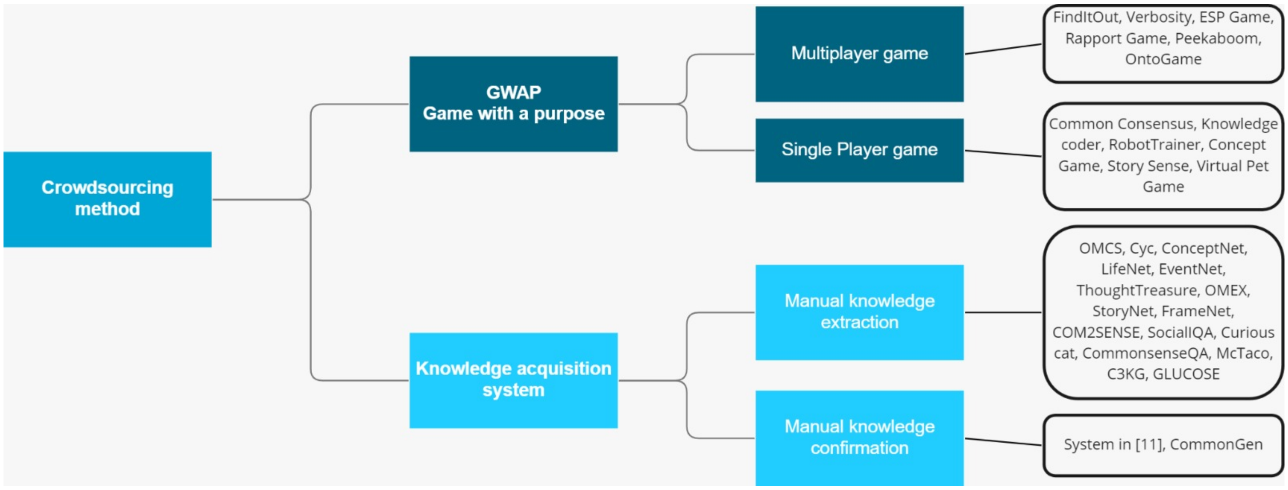
Crowdsourcing method
GWAP Game with a purpose
Knowledge acquisition system
Multiplayer game
Single Player game
Manual knowledge extraction
Manual knowledge confirmation
FindItOut, Verbosity, ESP Game, Rapport Game, Peekaboom, OntoGame
Common Consensus, Knowledge coder, RobotTrainer, Concept Game, Story Sense, Virtual Pet Game
OMCS, Cyc, ConceptNet, LifeNet, EventNet, ThoughtTreasure, OMEX, StoryNet, FrameNet, COM2SENSE, SocialIQA, Curious cat, CommonsenseQA, McTaco, C3KG, GLUCOSE
System in [11], CommonGen

Figure 1: Taxonomy of crowdsourcing methods for commonsense knowledge collection

Collection or confirmation of commonsense knowledge. More than that, the architecture of the game, the single-player mode, gives the option to both collect and verify collected knowledge. All games, except for Concept Game, include both the collection and confirmation of commonsense knowledge. For example, Robot Trainer offers this functionality by having three different levels, the first two taking care of the collection and the third one of confirmation where the user input is then used as a basis for the first level. However, the Concept Game generates random assertions and is only concerned with verifying if the tuples make sense or not.

Intended audience. Lastly, most of the games are intended to be used by adults or teenagers. In the mentioned list, an exception from this rule is the game Story Sense which intends to collect knowledge from children and have that knowledge be checked by adults who act as reviewers. Therefore, in this game, it is important to have adult reviewers since children might not own as much commonsense knowledge as desired, which should be visible in the quality of the collected tuples.

Multiplayer

Multiplayer games include two or more players that try to collaborate to finish a given task or to compete against each other to win the game.

The identified examples of multiplayer games are:

- **FindItOut** [5,6] is a game with two players that ask and answer questions to guess their opponent's card. The knowledge is collected based on the questions, answers and the removal of cards.
- **Verbosity** [2] is a game with two players, where one player gives descriptions about a word without including it and the other player tries to guess the word. The knowledge is generated based on the descriptions and guesses given by players.
- **The ESP Game** [1] is a game with two random players that try to type the same word for a given image, generating labels for images on the internet.
- **Rapport Game** [21] is a socialising Facebook game, where users can ask or answer questions, vote answers or follow other users.
- **Peekaboom** [3] is a game with two random players, where one player receives an image and reveals parts of it and the other tries to guess the word associated with the image.
- **OntoGame** [34] is a game attempting to develop and populate ontologies, using four different scenarios: turning Wikipedia into a domain ontology, annotating Youtube videos, mapping UNSPSC and eCl@ss and annotating eBay with eClassOWL.

Having mentioned this list of games, it is also worth detailing some similarities and differences between the games.

Type of knowledge collected. First of all, the type of knowledge that is collected differs by game. FindItOut and Verbosity collect commonsense knowledge tuples using a pre-defined format (FindItOut: IsA, HasA, AtLocation; Verbosity: “_is a kind of_”, “_is used for_”, “_is typically near/in/on_”). On the other hand, the ESP Game and Peekaboom are more concerned with visual commonsense knowledge. Both games try to collect labels for images, the latter being also able to collect labels for parts of images because of the game architecture. Although some of the collected knowledge might not fall under the category of commonsense, the game might still be considered a commonsense knowledge collection method. Commonsense knowledge in computer vision could be represented by humans reaching the conclusion that some objects exist in a context from looking at an image, even if the objects are not necessarily visible in the image [8]. For example, a person could look at an image with a table covered with a tablecloth and conclude that a table is there. Besides these two types of collected knowledge, Rapport Game is a game based on socializing and question asking and answering. Therefore, the Rapport Game collects text that might contain commonsense knowledge, but it is not limited to that. Lastly, OntoGame focuses more on developing and populating ontologies. The second one is more concerned with matching concepts from DBpedia ontology to ones from the ontology created using OntoPronto.

Game modes. Besides the type of knowledge, the games also differ based on the game modes they offer to the players. All listed games are intended to be played by two players, usually random ones. However, in [2] it is mentioned that Verbosity could also be offering a single-player mode by simulating the second player with already collected answers. The same principle applies to OntoGame, which requires a second player to reach an agreement. In both cases, if the game would be played as a single-player game, the purpose of the system would change from collecting commonsense knowledge to checking the quality of the collected knowledge.

Game location. Another aspect important to mention is the location of the games. All games are online web games, except the Rapport Game which is hosted on Facebook. This characteristic might be relevant during the analysis done in section 4 since the location of a game might influence its efficiency, cost ,or quality of the data.

3.2 Knowledge Acquisition Systems

Knowledge acquisition systems represent methods used for commonsense knowledge collection that are not attracting the contributors using a game interface. Therefore, because entertainment, the main aspect that attracts people to use the games with a purpose, is missing, the systems have to attract the contributors using another measure. Most of the time, this measure is money, which is efficient in motivating people to contribute to a project.

Based on the purpose, the knowledge acquisition systems can be split into manual knowledge extraction systems and manual knowledge confirmation systems, as figure 1 suggests. This categorization is similar to the single and multi-player games categorization because both determine the purpose of the system, to collect or verify knowledge.

Manual knowledge extraction systems

The manual knowledge extraction systems are built with the purpose of collecting commonsense knowledge using crowdsourcing. Sometimes, the quality of the collected knowledge can also be verified by somebody, but the system is still categorized as a manual knowledge extraction system. This happens because the system's main purpose is still collecting commonsense knowledge, not verifying it.

The methods identified with this purpose are:

- **Open Mind Common Sense(OMCS-1)** [10, 31] is a knowledge acquisition system designed to collect commonsense knowledge from humans over the web.
- **OMCS-2** [10, 31] is the second version of Open Mind Common Sense which collect commonsense knowledge using templates in English.
- **Cyc** [10, 16, 17] is the oldest project for commonsense knowledge acquisition from this list. It represents collected knowledge using formal logic.
- **ConceptNet** [10, 35] is a large-scale semantic network, gaining knowledge from OMCS corpus.
- **LifeNet** [10, 29, 32] is a commonsense knowledge base generated based on OMCS corpus and OMCSNet(an older version of ConceptNet). This method focuses on temporal reasoning, representing propositions as a probabilistic graphical model with temporal and atemporal relations.
- **EventNet** [9, 10] is a knowledge acquisition system focusing on commonsense temporal reasoning, which depends on LifeNet knowledge base.
- **ThoughtTreasure** [10, 23] was developed for reasoning purposes, natural language processing and computational linguistic tasks.
- **Open Mind Experiences(OMEX)** [10, 30] captures commonsense knowledge from humans through the internet. It collects story-like knowledge using pre-defined templates.
- **StoryNet** [10, 29] collects structured stories from humans to collect commonsense knowledge, using ConceptNet and LifeNet as knowledge resources.
- **FrameNet** [4, 10] focuses on human commonsense by building schematic conceptual scenarios in different domains. It extracts English words from electronic English corpora and extract knowledge using humans.
- **COM2SENSE** [33] collects complementary sentence pairs along the dimensions: knowledge domains and reasoning scenarios.
- **SocialIQA** [27] is a benchmark for commonsense reasoning focused on social situations. It collects commonsense questions and answers.
- **Curious cat** [7] is a knowledge acquisition system that collects the data from humans using a mobile application.
- **CommonsenseQA** [36] collects commonsense questions and answers based on a given concept extracted from ConceptNet.
- **McTaco** [38] focuses temporal reasoning by collecting commonsense questions and answers related to a given sentence.
- **C3KG** [18] is a Chinese commonsense conversation knowledge graph. It includes both dialog flow information and social commonsense knowledge.
- **GLUCOSE** [22] is a dataset of implicit commonsense causal knowledge. The knowledge is presented as causal mini-theories about the world using a narrative context.

Type of knowledge collected. Considering the mentioned list of methods, the type of knowledge that each of them collects can be described. There are methods that collect general commonsense knowledge, wanting to construct a knowledge base as complete as possible, such as OMCS-1, OMCS-2, Cyc, ConceptNet, ThoughtTreasure, OMEX, StoryNet, FrameNet, Curious Cat. Besides these, C3KG collects very similar knowledge, but it is focused on building a Chinese commonsense conversation knowledge graph. Besides this general knowledge, some methods are concerned with commonsense regarding temporal reasoning, such as LifeNet, EventNet. Also, COM2SENSE collects sentences that are complementary. SocialIQA collects commonsense knowledge regarding social situations. On the other hand, CommonsenseQA and McTaco focus on acquiring commonsense questions and answers, while the second one is only concerned with such data related to temporal reasoning. Lastly, GLUCOSE is a dataset of implicit commonsense causal knowledge, encoded as mini-theories about the world.

Source of commonsense knowledge. More than that, most of the mentioned methods collect the knowledge from crowdworkers themselves. However, other methods use knowledge bases created by existing systems to collect the needed knowledge. For example, ConceptNet depends on OMCS corpus, LifeNet on OMCS and OMCSNet, which is an older version of ConceptNet, EventNet on LifeNet, StoryNet on ConceptNet and LifeNet. There are also methods that extract knowledge as a base for crowdsourcing activity. CommonsenseQA uses concepts from ConceptNet as a base for the question and answers it generates, while McTaco takes sentences randomly selected from MultiRC [14]. Also, FrameNet extract information about English words from British National Corpus (BNC) and Concise Oxford Dictionary (COD), which serve as a base for the crowdworkers.

Expertise of crowdworkers. It is also worth noticing that most of the methods use non-experts as crowdworkers for collecting the knowledge. However, some of them perform a selection of the workers using some qualification tests. Such examples are COM2SENSE, McTaco, GLUCOSE. On the other hand, ThoughtTreasure and FrameNet require experts, knowledge engineers respectively lexicographers.

Confirmation of collected knowledge. Lastly, an important aspect of commonsense knowledge collection using crowdsourcing is the quality verification of the collected data. Regarding this, there are some methods that perform such verification using either other crowdworkers or experts. COM2SENSE, Curious Cat, McTaco, C3KG, and GLUCOSE check the collected knowledge during their process.

Manual knowledge confirmation systems

The manual knowledge confirmation systems are built with the purpose of verifying commonsense knowledge using crowdsourcing. These systems collect such knowledge using automatic, semi-automatic ,or other crowdsourcing methods and want to check if the collected knowledge can be considered correct.

The systems that fit the mentioned criteria are the following:

- System described in [11] aims to evaluate the output of KNEXT system [28] using Mechanical Turk to crowdsource the evaluation to non-experts with multiple rounds of tasks.
- **CommonGen** [20] collects references for the evaluation of the system from crowdworkers using Amazon Mechanical Turk to insure the best quality of the system.

Both methods generate commonsense knowledge automatically from text corpus and rely on humans for the verification of the knowledge. The first method uses crowdworkers to manually check the output of the system and the second one uses crowdworkers to generate the reference solutions for the evaluation of the system.

4 Analysis of Crowdsourcing Methods for Commonsense Knowledge Collection

After identifying a list of crowdsourcing methods to collect commonsense knowledge in section 3, a comparison between these systems is conducted. To conduct the comparison, a list of measures is required and the values for these are collected from the literature that introduces the methods.

4.1 Evaluation Protocol

Efficiency can be assessed using the throughput, number of knowledge tuples generated per unit time, of each method

Cost can be assessed by measuring the average cost for each knowledge tuple

Quality can be assessed using the average accuracy of each knowledge tuple.

Efficiency is considered because the methods should aim to collect as many knowledge tuples as they can in order to construct a database as extensive as possible. This measure might differ from method to method based on how attractive a system is and how interested and devoted are the crowdworkers.

Cost, in this context, represents the financial cost of collecting the knowledge. Although this aspect might not be as important for the games with a purpose, it might affect the collection of knowledge using a knowledge acquisition system.

The last item, quality, is especially important because crowdsourcing methods depend on human input. If you are assessing a game with a purpose, the knowledge will most likely come from a non-expert. Therefore, there can be errors in the collected tuples. To measure the quality of the collected data, a set of experts can check the data and confirm if it can be considered correct or not.

More than that, the main source of metrics used for comparing the found methods is the existing literature. Therefore, the chosen metrics reflect the most commonly used measures for the evaluation and analysis of crowdsourcing methods for commonsense knowledge collection.

4.2 Comparison of Crowdsourcing Methods for Commonsense Knowledge Collection

All results included in table 1 were extracted from the literature associated with each method. Therefore, the values fully depend on the experiments and results of the research papers that describe the methods.

Efficiency

The first aspect that has to be considered in the comparison of the mentioned methods is efficiency. The throughput included for all methods that mentioned this measure in the associated literature represents the total number of knowledge tuples generated per day. This method of displaying the

Method	Throughput	Cost	Accuracy
Robot Trainer	12	-	63.14%
Common Consensus	38880	-	-
Knowledge Coder	13	-	-
Concept Game	500	-	-
Virtual Pet Game	2796	-	92.07%
Story Sense	124	-	86.55%
FindItOut	20016	-	95.6%
Verbosity	1124	-	85%
The ESP Game	10337	-	85%
Rapport Game	76	-	-
Peekaboom	36225	-	100%
OntoGame	-	-	99%
OMCS-1	651	-	75%
OMCS-2	-	-	85%
Cyc	-	-	-
ConceptNet	-	-	68%
LifeNet	-	-	89%
EventNet	-	-	62%
ThoughtTreasure	28	-	-
OMEX	11520		62%
StoryNet	-	-	-
FrameNet	-	-	-
COM2SENSE	-	-	95%
SocialIQA	-	-	87%
Curious Cat	21	-	96%
CommonsenseQA	-	0.33	88.9%
McTaco	-	-	87.1%
C3KG	-	0.2	-
GLUCOSE	-	1.6	-
System from [11]	-	0.002	77%
CommonGen	-	-	-

Table 1: Table including values for comparison measures (throughput, cost, accuracy) for all methods discussed in section 3 collected from the referenced literature. The cells that are marked with a dash, "-", suggest that the literature did not include values for the selected measures. The throughput is measured in knowledge tuples generated per day and the cost in dollars per question.

throughput of the systems should be more relevant than displaying the throughput per user since it should also reflect the attractiveness of the system. However, some methods were not deployed to the general public and only show results that were obtained through an experiment in a closed, supervised environment. Therefore, some systems may have higher throughput than others without being more appealing to the general public because the users that participated in the experiment were instructed to review the system.

For the Virtual Pet Game, two papers reported the number of contributions per unit of time and the two results differ, so the bigger one was mentioned in table 1 since it reflected a longer period which should be closer to a real-world scenario.

Regarding the comparison of the collected values, the highest throughput values included in table 1 are associated with two games with a purpose: Common Consensus and Peekaboom. Therefore, collecting commonsense knowledge through games performs relatively well and such a system should be able to solve the problem of collecting such knowledge. However, most of the knowledge acquisition systems did not report their throughput and the comparison with the games would not be completely equitable. Most of the methods report the total number of knowledge tuples collected, but they do not mention the time frame over which they were collected. Therefore, the throughput could not be computed and included in the table.

Cost

The second measure that is considered during the comparison of the methods is the cost of collecting knowledge tuples. For the first category of crowdsourcing methods, the games with a purpose, the cost is not relevant and also not mentioned in the literature associated with the methods. The games are attracting the users to contribute with knowledge using the entertainment factor, so the users do not have to be compensated financially, since they might be open to contributing to their interest. However, the implementation of the method includes costs that were not reported in the papers, but since both categories of methods imply such costs, they can be ignored for this discussion.

As it can be seen in table 1, the cost was not reported in most of the literature. However, from the few values that were collected, the manual knowledge confirmation systems have lower costs than the manual knowledge extraction systems. This aspect should reflect the reality since crowdworkers that are asked to confirm some collected knowledge are probably paid less than workers that need to input knowledge.

Quality

The quality of the collected knowledge by each method represents the accuracy of the generated tuples. Since all methods rely on human input, it is expected for the accuracy to be relatively high because humans are the source of commonsense knowledge. The values showed in table 1 are between 63.14% and 100%.

The reported values could be considered to reflect the reality since the knowledge comes directly from the source and relatively high quality should be expected for the collected knowledge.

There are some methods for which the quality of the gathered knowledge was not reported in the literature. Systems such as C3KG, GLUCOSE selected the crowdworkers to ensure the best quality of knowledge possible. Also, other methods, such as ThoughTreasure and FrameNet, receive contributions from experts which should input high-quality knowledge since they are believed to have a higher level of overall knowledge in this domain.

5 Discussion

This section aims to highlight main findings, limitations and implications of the research conducted.

5.1 Main findings

The detailed findings of this research paper are presented in sections 3 and 4. The most important aspects are the created taxonomy together with the descriptions of each method and the analysis of these methods, which reflect the two research questions associated with this research.

The first contribution of this research is the taxonomy of the crowdsourcing methods for commonsense knowledge collection. The first level of abstraction is divided into games with a purpose and knowledge acquisition systems. Games with a purpose can be split into single-player and multiplayer games and the knowledge acquisition systems can be split into manual knowledge extraction and manual knowledge confirmation systems.

The second contribution is represented by the analysis in section 4. This section includes a table with values for the selected attributes: efficiency, cost, and quality. This table includes the values reported in the literature associated with each method. More than that, based on the values, some conclusions were drawn regarding the mentioned attributes. Therefore, based on the gathered data, some games with a purpose are more efficient than the knowledge acquisition systems that reported the throughput. Regarding the cost of the methods, games with a purpose have a lower cost associated with collecting the knowledge because they do not offer financial compensation to the users. Also, the quality of the gathered knowledge varies, but both categories of crowdsourcing methods have comparable results.

5.2 Limitations

The conducted research represents a systematic survey of literature about crowdsourcing methods for commonsense knowledge collection. Therefore, the findings of this research are limited by the collected literature on the mentioned topic. This aspect limits especially the analysis of the mentioned crowdsourcing methods because the measures that are used to compare the methods completely depend on the reported measures of the literature associated with each system. Some of the literature only reports some of the mentioned measures or none of them. Therefore, the analysis cannot be considered complete since it does not contain information for all of the methods.

More than that, another limitation of the comparison of the methods is the methodology behind the mentioned measures. Some of the reported values from the literature are obtained through a closed environment experiment, while others are obtained through a general public deployment. Also, especially for the accuracy of the collected knowledge, the values can represent slightly different measures. This aspect occurs because some papers take into consideration the correctness of the collected knowledge, while others measure the relevance of the data. Therefore, if a paper considers the relevance, it could lead to a lower accuracy value than a system that considers the correctness. This aspect limits the analysis section of this research because the measures are not consistent and concluding which methods are the best would not be possible.

5.3 Implications

The conducted analysis of the crowdsourcing methods for commonsense knowledge offers a general perspective on these systems. The insight given by comparing the methods based on the mentioned measures could contribute to research to develop a system that aims to collect commonsense knowledge. This system could be constructed similarly to the best methods mentioned so that it would be the most efficient, cost effective, and accurate system of commonsense knowledge collection. Therefore, the information collected from the surveyed literature contributes to a clearer understanding of crowdsourcing methods with the mentioned purpose.

6 Conclusion and future work

6.1 Conclusion

This research paper represents a systematic survey of literature on the topic of commonsense knowledge collection using crowdsourcing methods. The conducted survey gathered information about 31 systems, created a taxonomy and described each method mentioned. Also, a comparison based on common characteristics was done, while also a comparison using three common measures: efficiency, cost, and quality was detailed.

After investigating the mentioned literature, the games with a purpose showed promising results in efficiency, while also having lower costs since the users are not paid to contribute to the knowledge collection. Also, the quality of the gathered knowledge using games with a purpose is comparable to the one attained by using knowledge acquisition systems, even if the users of the first category are not experts and were not selected on any criteria in most cases. Therefore, we believe that games with a purpose are a good solution to the problem of collecting commonsense knowledge from the general public.

6.2 Future work

Regarding the future work that would improve this research paper, other methods for commonsense knowledge could be added in order for the survey to become more complete and to create a better comparison between the crowdsourcing methods that try to solve the commonsense knowledge collection problem.

Besides including more methods and trying to collect more literature related to the topic of this research, collecting more measures of comparison and values that are currently missing in the comparison would improve the overview. Currently, the comparison of the methods is limited by the information provided in the mentioned literature. However, some literature that was not collected and commented on might include values for the chosen measures that are omitted in this paper.

Another possible improvement that would be outside of the scope of this paper would be conducting experiments to collect missing results for the mentioned measures. This action would improve the completeness of the comparison but is outside of the scope of a systematic survey.

7 Ethical and Responsible Research

This section aims to highlight the possible ethical issues of the mentioned methods and to assess if the research done in the collected literature can be considered responsible research.

7.1 Ethical Research

The research that is described in the mentioned literature could present some ethical issues. These should be taken into consideration after systematically surveying the literature in order to make sure that the research that is used is taking measures against possible concerns.

Privacy. Since the methods that are described in the literature are crowdsourcing methods, they rely heavily on human input. Therefore, the privacy of the users should be considered and measures should be taken by the developers of the methods in order to protect their users. This aspect can be a problem, especially regarding the games with a purpose. As mentioned in section 3, the purpose of these methods of knowledge collection is usually hidden behind a game. Therefore, if the game does not explicitly mention the fact that it collects commonsense knowledge using the interaction with the game, then the users are not aware that some of their data could be collected.

Storage of data. Besides the privacy of the users, the storage of the collected data should also be considered. If the game collects personal information, this information should be stored in a secure way and only for the purpose of this research. The literature mentioned should take into account this ethical concern and describe how the research avoids a possible problem.

Vulnerability of users. For the previous concerns, an agreement to collect and store information from users is needed. This agreement can be easily set up if the users are adults. However, at least one of the methods presented is intended to be used by children. Story Sense is a game intended to be used for interactive education and the target audience is represented by children. Besides this example, some other games could also be used by children, but this aspect is not specified in the mentioned literature. However, in the case of Story Sense, the users cannot agree to have their information collected and stored since they are not legally allowed to take this decision. Therefore, the developers and researchers working on this method should take into consideration that the users are vulnerable and that their data should be handled differently than in a classic scenario.

Therefore, the privacy of the users, the storage of collected data and the vulnerability of the users should be taken into account while conducting the research. Some of the mentioned literature takes into account these possible ethical implications of the research and mentions the measures and methodology that were followed during the experiments.

7.2 Responsible Research

The main concern of this section is if the research conducted in the references papers is responsible. Regarding this aspect, the research should be reproducible by other researchers. In order for this characteristic to be met by the research, it should describe in a detailed manner the experiment that was conducted and the system that was implemented. Also, the sample size of the experiment should be high enough for the results to reflect the general truth and not to represent a small sample.

Reproducibility For the reproducibility of the research, the sample size is very important. There are methods that were mentioned in section 3 that have a small sample size, such as Common Consensus(with 11 participants). More than that, some of the mentioned systems are not in use anymore. Therefore, the research and experiments mentioned in the associated literature could not be reproduced as easily. Also, some systems were developed and analyzed more than 20 years ago. This aspect could influence the results that the systems would have now because the user behaviour might have changed in the meantime, especially if the systems were deployed to the general public.

References

[1] 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, 2004.

[2] Luis Von Ahn, Mihir Kedia, and Manuel Blum. Verbosity: A game for collecting common-sense facts. In *CHI 2006 Proceedings*, pages 75–78, 2006.

[3] Luis Von Ahn, Ruoran Liu, and Manuel Blum. Peekaboom: A game for locating objects in images. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 55–64, 2006.

[4] Collin F Baker, Charles J Fillmore, and John B Lowe. The berkeley framenet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, 1998.

[5] Agathe Balayn, Gaole He, Andrea Hu, Jie Yang, and Ujwal Gadiraju. Finditout: A multiplayer gwap for collecting plural knowledge. 2021.

[6] Agathe Balayn, Gaole He, Andrea Hu, Jie Yang, and Ujwal Gadiraju. Ready player one! eliciting diverse knowledge using a configurable game. pages 1709–1719. ACM, 4 2022.

[7] Luka Bradeško, Michael Witbrock, Janez Starc, Zala Herga, Marko Grobelnik, and Dunja Mladenić. Curious cat-mobile, context-aware conversational crowdsourcing knowledge acquisition. *ACM Transactions on Information Systems*, 35, 8 2017.

[8] Ernest Davis and Gary Marcus. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, 58:92–103, 9 2015.

[9] Jose Espinosa and Henry Lieberman. Eventnet: Inferring temporal relations between commonsense events. volume 3789, pages 61–69, 2005.

[10] Mohamed Gawish and Abdel-Badeeh Salem. A study on representation and reasoning techniques of commonsense episodic knowledge: Challenges and applications. *International Journal of Computers*, 3:145–158, 2018.

[11] Jonathan Gordon, Benjamin Van Durme, and Lenhart K Schubert. Evaluation of commonsense knowledge with mechanical turk.

[12] Amaç Herdagdelen and Marco Baroni. The concept game: Better commonsense knowledge extraction by combining text mining and a game with a purpose, 2010.

[13] Filip Ilievski, Alessandro Oltramari, Kaixin Ma, Bin Zhang, Deborah L. McGuinness, and Pedro Szekely. Dimensions of commonsense knowledge. *Knowledge-Based Systems*, 229:107347, 2021.

[14] Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, 2018.

[15] Yen-Ling Kuo and Jane Yung jen Hsu. Goal-oriented knowledge collection. In *AAAI Fall Symposium Series*, 2010. The Virtual Pet Game.

[16] Doug Lenat, Mayank Prakash, and Mary Shepherd. Cyc: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks. *AI Magazine*, 6:65–85, 1985.

[17] Douglas B. Lenat. Cyc: A large-scale investment in knowledge infrastructure. *Commun. ACM*, 38:33–38, 1995.

[18] Dawei Li, Yanran Li, Jiayi Zhang, Ke Li, Chen Wei, Jianwei Cui, and Bin Wang. C3kg: A chinese commonsense conversation knowledge graph. 4 2022.

[19] Henry Lieberman, Dustin A Smith, and Alea Teeters. Common consensus: a web-based game for collecting commonsense goals, 2007.

[20] Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. Commongen: A constrained text generation challenge for generative commonsense reasoning. 11 2019.

[21] Yen ling Kuo, Kai yang Chiang, Cheng wei Chan, Jong-Chuan Lee, Rex Wang, Edward Shen, and Jane Yung jen Hsu. Community-based game design: Experiments on social games for commonsense data collection. In *Proceedings of the ACM SIGKDD Workshop on Human Computation : 2009, Paris, France, June 28-28, 2009*, pages 15–22, 2009.

[22] Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. Glucose: Generalized and contextualized story explanations. 9 2020.

[23] Erik T. Mueller. A database and lexicon of scripts for thoughttreasure. *CoRR*, 2000.

[24] Ethel Ong, Kaizer Bienes, Nickleus Jimenez, Ephraim Miranda, and Gabriel Pascual. A system for collecting commonsense knowledge from children. In *DLSU Research Congress 2014*, 2014.

[25] Christos Rodosthenous and Loizos Michael. A hybrid approach to commonsense knowledge acquisition. volume 284, pages 111–122. IOS Press, 2016.

[26] Christos T. Rodosthenous and Loizos Michael. Gathering background knowledge for story understanding through crowdsourcing. In *OpenAccess Series in Informatics*, volume 41, pages 154–163. Schloss Dagstuhl-Leibniz-Zentrum fur Informatik GmbH, Dagstuhl Publishing, 2014.

[27] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions. 4 2019.

[28] Lenhart Schubert. Can we derive general world knowledge from texts? pages 94–97, 2002.

[29] P Singh, B Barry, and H Liu. Teaching machines about everyday life. *BT Technology Journal*, 22:227–240, 2004.

[30] Push Singh and Barbara Barry. Collecting commonsense experiences. In *Proceedings of the 2nd International Conference on Knowledge Capture*, pages 154–161, 2003.

[31] Push Singh, Thomas Lin, Erik T Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. Open mind common sense: Knowledge acquisition from the general public, 2002.

[32] Push Singh and William Williams. Lifenet: A propositional model of ordinary human activity, 2003.

[33] Shikhar Singh, Nuan Wen, Yu Hou, Pegah Alipoormolabashi, Te-Lin Wu, Xuezhe Ma, and Nanyun Peng. Com2sense: A commonsense reasoning benchmark with complementary sentences. 6 2021.

[34] Katharina Siorpaes and Martin Hepp. Ontogame: Weaving the semantic web by online games. In *The Semantic Web: Research and Applications*, volume 5021, pages 751–766, 2008.

[35] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. 12 2016.

[36] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. 11 2018.

[37] Man-Ching Yuen, Irwin King, and Kwong-Sak Leung. A survey of crowdsourcing systems. In *2011 IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing*, pages 766–773. Institute of Electrical and Electronics Engineers (IEEE), 1 2012.

[38] Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. ”going on a vacation” takes longer than ”going for a walk”: A study of temporal commonsense understanding. 9 2019.