A Crowdsourcing Approach to Finding Evidences

Panikka Mudiyansela Nadeesha Kumari Wijerathna

Graduate School of Library, Information and Media Studies

University of Tsukuba

September 2019

Contents

List of Figures

Chapter 1

Introduction

The Introduction chapter includes a background to the phenomenon of crowdsourcing including its history. Further, the problem discussion is featured in the chapter, along with the formulation of the purpose of the study and lastly, the chapter will be ended up an outline of this thesis.

1.1 Background Knowledge of Crowdsourcing

1.1.1 The Power of People

At the present time, without difficulty or effort, people can connect with others on the internet, and a set of new applications have emerged to share their knowledge, skills, experiences, and perception. Wikipedia [16] can be considered as a good example of this collaborative effort. Another example for this collaboration, well-known Open Street Map (OSM) [17] was started to make an online map of the world by volunteering that is free of use restrictions and open to all users. The software programs for question-and-answer (QA) forum such as Stack Overflow [18] and Quora [19] also the examples, where people give the programming questions to the public and support to the directed social graph to with the meaningful content respectively. This kind of applications demonstrates that people always willing and having the ability to collaborate in the online setting and this collaboration can also be identified as the acceleration of the latest breakthroughs in information technology.

Besides, humans often have the power to improve results considerably compared to computers. Although computer systems have made substantial progress in the past several decades, in many areas, computers have not been able to match the creativity, flexibility, and complexity of human intelligence [24]. Therefore, many attempts of involving humans in crowdsourcing research have been made to exploits human computational capabilities to search accurate information, to solve real-world problems and intelligent tasks. For example, the intelligent tasks on generating confer-

ence programs [1], image tagging for subjective topics [2,3], understanding topics in micro-blogs [4], finding approaches to collect data and understand human health [5,6], and finding evidences for proof in difficult problems [31] to names but a few. These factors show that 'power of people' and it is clear that crowdsourcing is a promising tool, involving multiple people to search for accurate information, give a solution for real-world problems and to do difficult tasks such as finding evidences.

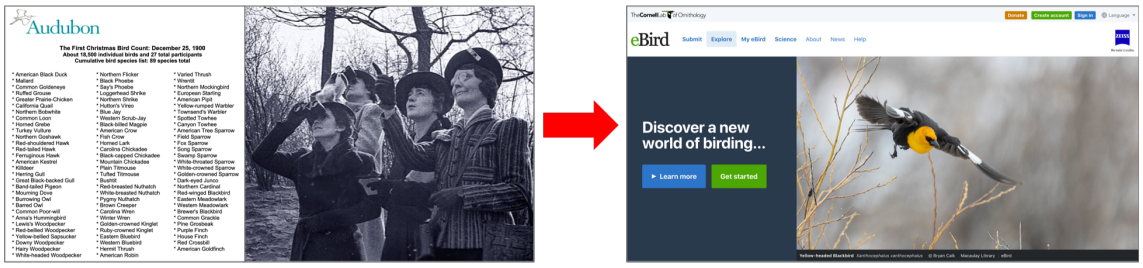
1.1.2 Crowdsourcing in History

The first form of Crowdsourcing was open in the competitions such as citizens arrange innovation competitions with the purpose of attracting the crowd to share their knowledge and skills for challenging problems. For example, the British government created a monetary award—known as the Longitude Reward—for the best proposal to measure a ship's longitudinal position in 1714 [20]. As same, the open competition stared for food preservation and the French army asked for solutions to food preservation to the purpose of supporting French soldiers In 1795. After a few years, the prize was won by a confectioner—Nicolas Appert [28,29]. In 1939, the automobile department of the Toyoda Automatic Loom Works, Ltd.'s, (today's Toyota Industries Corporation) logo was invented by means of a similar contest. A design for Toyoda was chosen among some 27,000 entries sent in from all over Japan [21]. Likewise, a competition to design a National Opera House for Sydney was also held to find the best architectural design in 1956 [22].

Another form of crowd-based events that started in the ealy centuries is engaging citizens to achieve a common goal. For example, the National Audubon Society (NAS), a non–profit environmental organization which was dedicated to conservation since 1900. It has conducted bird counts—known as Christmas Bird Count—in the Christmas season and counted the number of different types of birds by hobbyist bird watchers in home areas and send their performances to the NAS data center as a mail. The first Christmas Bird Count had just 27 volunteers. But it has grown to encompass more than 60,000 volunteers of all ages [23,35].

Despite the earliest crowdsourcing campaigns were successful, there were many limitations that they occurred. For example, organizers were not able to reach a large population, organizers could only access the local knowledge, and difficulties to participants to collaborate were the main limitations. Therefore, these campaigns unable to use the vast majority of distributed knowledge which is not yet exploited. However, due to technological development, new opportunities were accelerated and significantly reshaped in the previous attempts [25,26,27]. The example of these new directions shown in Figure 1.1 and Figure 1.2. Since 2002, as a new development, the eBird Web Portal [36] developed by the NAS and Cornell Lab of Ornithology to allow the crowd to

contribute their bird experiences to the public by using the server data-based electronically (see Figure 1.1). Another example of the new opportunities is after two centuries, the competition created by Netflix company to aggregate the new algorithms as an open competition worldwide during 2006 to 2009 [30,37] (see Figure 1.2).



Christmas Bird Count (since 1900) eBird Counting (since 2002)

Figure 1.1: Influence of technological development to achieve a common goal in Crowdsourcing; In 1900, Christmas Bird Count started as in the Christmas season and counted the number of different types of birds as a common purpose. The eBird Web Portal launched to communicate electronically with the server database to achieve the same objective Since 2002.

Open Competition for Food Preservation (Started in 1795 and Prize Won in 1810)

Netflix Prize Competition (Stared in 2006 and Prize won in 2009)

Figure 1.2: Influence of technological development in Crowdsourcing Competitions; In 1975, the open competition stared for food perservation. Two centuries later, in 2006, Netflix solicited new algorithms from the citizens to achieve the same kind of objective using Web technology.

Likewise, the earliest crowdsourcing related projects have engaged in different domains and a wide range of disciplines such as programming [30], ecological data [23,33,34], astronomical data [32], businesses [21] and so on. Lately, many of those projects have been developed with web technology. However, the term 'Crowdsourcing' is introduced in the 21st century. The phenomenon of modern Crowdsourcing and Human Computation will be briefly discussed in the next section.

1.1.3 Crowdsourcing in Modern History

The term Crowdsourcing is first coined by Jeff Howe in a Wired Magazine article in 2006 [38]. In later, the following definition was provided by Howe [39] on his website. According to his definition,

> *"Crowdsourcing represents the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call."*

In other words, the company or the organization invites large and diverse community—the crowd—to come with creative, flexible or complex ideas or solutions, the best ideas will be rewarded and produced. Howe was inspired and influenced by James Surowiecki and his book 'The Wisdom of Crowds' [43]. Therein Surowiecki describes the concept of collective intelligence in related to the concept of crowdsourcing in 2004.

In 2008, considering the specific applications of crowdsourcing, Daren C. Brabham [41]—the first person who published the scholar article using the word crowdsourcing—describes crowdsourcing as a process of online, distributed problem–solving and production model. According to Simula [40], crowdsourcing is a novel problem–solving approach that utilizes human intelligence to complete the complex tasks cost–effectively in the form of an open call. Likewise, after reviewing more that 40 definitions of crowdsourcing, Enrique and Fernando introduced a new integrating definition in 2012 [60] as cited in Wikipedia [16]:

> *"Crowdsourcing is a type of participative online activity in which an individual, an institution, a nonprofit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task; of variable complexity and modularity, and; in which the crowd should participate, bringing their work, money, knowledge and/or experience, always entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and use to their advantage that which the user has brought to the venture, whose form will depend on the type of activity undertaken."*

The concept of crowdsourcing is frequently used in relation to other collaborative innovation concepts like Human Computation, Social Computing and so on. Scholars such as Quinn and Bedersons [25] distinguish that the concept of Human Computation with versus related ideas like crowdsourcing and there is an overlap between Human Computation and Crowdsourcing. According to them, these two phenomena are interchangeable as well as interoperable.

In the last decade, crowdsourcing has become increasingly popular and the considerable success of crowdsourcing platforms. Such as Amazon Mechanical Turk (AMT), Free–Lancer, Yahoo

Crowdsourcing, CrowdFlower, and oDeck has been creating a new paradigm in the field of crowdsourcing. These online market places make the payment convenient for the workers who willing to solve a range of different tasks. In terms of research, the existence of these crowdsourcing markets make it possible for researchers to quickly recruit participants and conduct controlled experiments to investigate crowdsourcing itself and different subject areas via crowdsourcing.

On the other hand, online communities such as Wikipedia [16], Stack Overflow [18], and 99Designs [45], focus on building and maintaining a community of volunteers or on tempting experts with high rewards for complex tasks [44].

Further, the development of mobile techniques offered new possibilities for mobile crowdsourcing. The main advantages of the mobile applications are, it can gather real–time information and possibilities of implicit crowdsouring (described in the next section) by leveraging the fact that a significant amount of people always have their mobile devices with them. The example for such kind of project includes; Geo-social networking (Google Latitude and Loop generate social networks supported by geolocation features), Cooperative traffic such as Google Maps and Waze, product testing via mobile crowdsourcing such as uTest [46] and Mob4Hire [47] and so on.

However, in the world today, while online, mobile and many kinds of crowdsourcing have been extensively activated, in the same online setting billions of information is generating, sharing and using by the crowd. As results, misinformation, disinformation or ambiguate information have spread extensively on the web and social networks. The underlines story of this thesis will be the above challenge, and in this thesis, an empirical study on this issue has been addressed (see later sections).

1.1.4 Data Collection in Crowdsourcing

In crowdsorcing systems, data collection can occur using two types; implicit and explicit. Human can explicitly provide data in response to a given task or contribute data implicitly as a byproduct of their day–to–day work with using particular tools and applications. Quinn and Bederson [25] also have described above as types of crowdsourcing in their research on Human Computation: A Survey and Taxonomy of a Growing Field. According to them, implicit and explicit crowdsourcing can be recognized as unintentional and intentional crowdsourcing respectively.

With implicit crowdsourcing, users do not necessarily know that their actions are being crowdsourced. For instance, humans can implicitly contribute about their positions through various sensors such as the Global Positioning System (GPS) in the mobile phone. Traffic monitoring is possible using data from the GPS. Another example for the implicit crowdsourcing is reCaptcha—a CAPTCHA system designed to establish that a computer user is a person (normally in order to

protect websites from bots) and, at the same time assist in the digitization of books or improve machine learning [48].

In contrast, with explicit crowdsorucing, humans directly and consciously address the tasks. The size and complexity of the task can vary. For instance, asking people to provide keyword labels for image is a small task, while more involved task include designing logo [49]. Wikipedia [16] also being an example for explicit crowdsourcing. It includes collaborative composition and editing activities as well.

Data collection in crowdsourcing can be occur either types. In this thesis, explicit crowdsourcing will leverage; asking crowd to participate to finding evidences.

1.2 Research Problem

> A photograph was posted on Twitter by a marketer in Austin, Texas, the day after the presidential election in 2016 as shown in Figure 1.3. The tweet was a picture of buses with the caption *"Anti–Trump protesters in Austin today are not as organic as they seem...."* Although the person who posts the tweet had 40 followers, within half a day, it had nearly 17,000 retweets with more than 300 comments and became part of a national controversy. After 24 hours, president Trump also took a tweet to Twitter about professional protesters. Later, a Facebook post about the tweet also shared by more than 300,000 Facebook users. The tweet played a key role in making many fake news stories go viral by bolstering the emotional tenor of the lie. Finally, found that the tweet itself also a fake information and the marketer agreed that he posted this photograph without fact check and without any evidence [51].

The above example demonstrates that the extensive spread of misinformation (inaccurate information), disinformation (deceptive information)—sometimes known as fake information or news—and ambiguate information via the web and social networks have become a serious negative impact on society in the world today. While this fake information is not a new phenomenon [52], questions arise such as why is it appearing as a world topic and why is it enticing increasingly more public attention is particularly relevant at today. The leading cause is that fake information and ambiguate information can be created and published online faster and cheaper when compared to traditional media such as newspapers and television. Also, some people use fake information or ambiguate information to be famous or to their enjoyment and so on. In many situations, people who share misinformation also may not believe such information themselves [50]. In this context, the investigation process and searching accurate information are more challengeable.

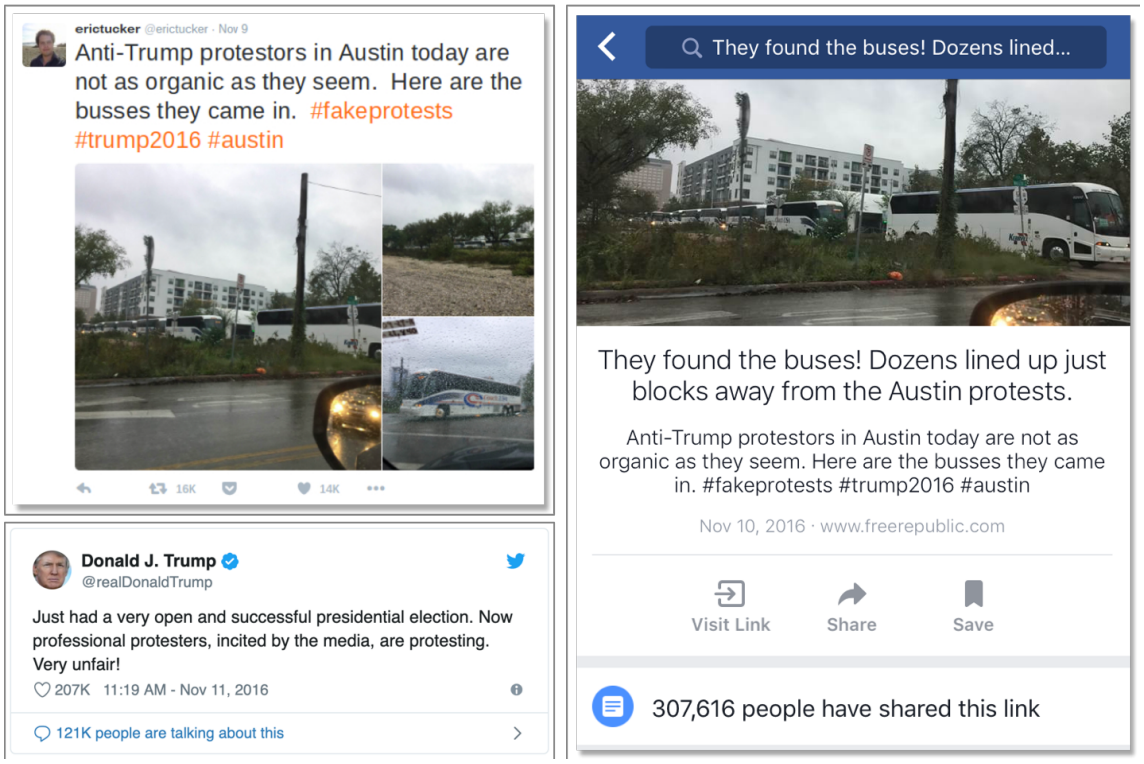


Figure 1.3: The famous example for fake information: A photograph was posted on Twitter by a marketer in Austin, Texas, the day after the presidential election in 2016 and it was a disinformation.

Over the last several years, there are growing efforts to develop the websites, the web services, and the systems to detect this misinformation, disinformation or ambigaute information. Figure 1.4 shows some examples of such kind of websites and services. Basically, many of them check political related facts. The approach of many of those websites and services can be summarized in generally two ways (for better understanding of these web services, Table 1.1 shows the list of websites and their fact check methods as a summary). First, they make a reference statement or the claim based on some incidents, then ask web users to measure degrees of true or false and publish the results. The second approach is, some services investigate the fact by their team collecting some evidences and publish generally accurate facts. Also few services follow both approaches together, in either way (see Figure 1.5).

However, the serious problems or the limitations of these web services are;

- Most of these websites do NOT evaluate the answers of people. In the first case, they just measure the degree of true or false and publish the results. In that case, they do not consider the evidences.

- Another limitation is, although some services try to find evidences by their team, they have

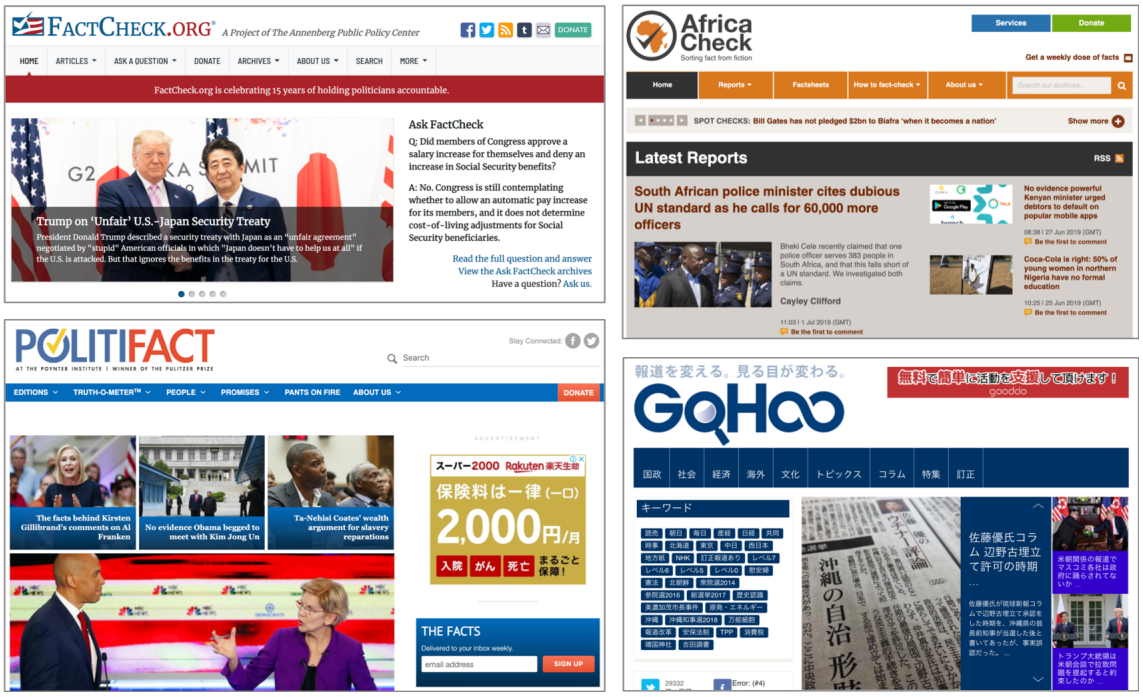


Figure 1.4: Famous fact check websites to detect this misinformation, disinformation or ambigaute information.

a serious problem of less manpower in the team and the limited budget. For example, according to the FactCheck.ORG [iv] service, they have only 4 team members to investigate all the claims.

- Also, there is no academic research has made using these web outputs.

To address the limitation of existing approaches of fake information, in this thesis, the researcher introduce 'Evidence-based' approach by crowdsourcing—ask crowd to search and provide possible evidences to prove the fact.

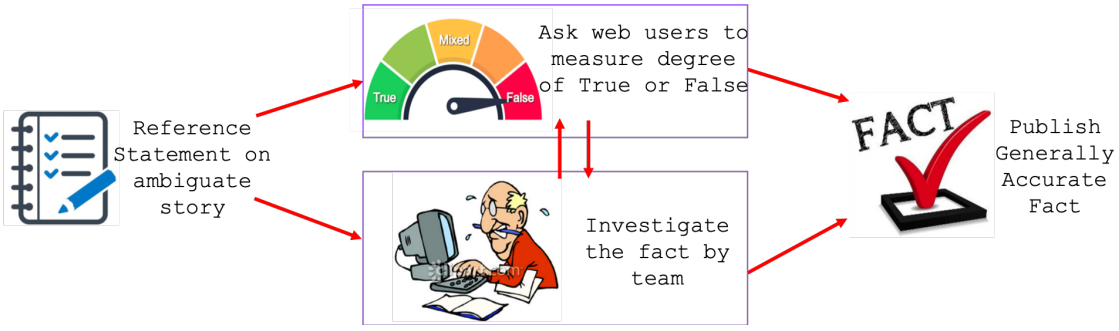


Figure 1.5: Approaches of the fact check by fact checking websites.

Table 1.1: The list of fact check websites and their fact check methods.

No.	Website	Fact check Method
1	Media Bias since 2015	Measure degrees of bias (especially political bias) by using ranking. Place a voting poll on each page for readers to vote on the bias of the source to examine if the public is agreeing or disagreeing with their ratings.[i]
2	Politifact since 2007	Select the most newsworthy and significant claim from TV and social media then made a statement to fact-check. When selecting the claim, they consider; A fact that is verifiable? Does it seem misleading or sound wrong? Is the statement significant? and so on. Measure the decreasing level of truthfulness by six ratings.[ii]
3	Snopes	Use a comprehensive list of ratings such as True, Mixture, False, Mostly True, Outdated, Misattributed, and so on.[iii]
4	FactCheck.ORG	Monitor the factual accuracy of what said major US political players in the form of TV, debates, speeches, and interviews. Basically, focus on claims that are false or misleading.Systematically go through transcripts and videos looking for statement based on facts. Once they find a statement, they will engage or attempt to engage with the person or organization that is being fact-checked. If the supporting material shows that the statement is accurate, they drop it and move to another. If the supporting material does not support the claim or if no evidence is provided, then they conduct research of their owns.[iv]
5	GoHoo (Japan)	Measure degrees by rating and false alarm level. False alarm means; Degree of whether the error is certain.[v]
6	Fact Checker.IN	As a fact-checking initiative, they pick claims/statements made by public figures or government reports and check for veracity and context. They refer to official data in the public domain to verify claims. Often, claims are found to be true, but misleading as the context and historical perspective is missing. In such cases, they provide context to the claim in question by their team.[vi]
7	Africa Check	Decide to rate to provide readers with the clearest understanding they can find of the accuracy of the statement- according to the best evidence publicly available at the time.[vii]
8	altnews (India)	Alt News continuously monitors social media and mainstream media for incorrect and/or dubious information.[viii]
9	Gomaneh (Iran)	Ask people to choose rumor or superstition by themselves and make a report with at least 600 words. it should be with sufficient references and scientific source, photos (if available). Then reports evaluate by team members.[ix]
10	Full Fact (UK)	When they see a misleading or unsubstantiated claim, they push for corrections and withdrawals. But they also want to prevent inaccurate claims from being made in the first place. Their long-term objective is to make sure there are systems in place so that only valid and validated claims are put into the public domain, and that the public is able to check for themselves.[x]

1.3 Research Goal and Contribution

In order to address the limitations of existing approaches, and the issues of fake and ambiguate information on the internet, this thesis explores a crowdsourcing approach—ask crowd to search and provide possible evidences to prove the given assumption which is adopting with a micro-task (a small unit of work that usually takes only minutes or less to complete) based general framework.

Crowdsourcing performs as a nascent tool to explores the power of humans, and unlike machines, since people have very different backgrounds, experiences, skills, and perceptions. This work clarifies that crowdsourcing workflows will also remain a good fit to investigate the power of the crowd.

The contributions of this thesis are as follows:

Crowdsource-based approach for evidence-based fact check. We explore a general crowdsourcing framework for collecting and selecting evidence-based proofs. We do not pose any particular form of evidences, and thus workers can input a variety of ways to prove. The proposed approach is useful especially in situations where many workers do not know answers or believe wrong answers.

Experiments with real-world crowd workers. This thesis presents the results of experiments with real-world crowd workers. First, we found that crowd workers can provide diverse methods for proofs, among which some were surprising as we did not factor those as proofs. This shows the power of crowdsourcing. It can provide different viewpoints. Second, we determined that the rating-based method worked well, whereas the voting-based method, which adopted a simple vote casting for a limited set of evidences shown in each micro–task. Thus allowing further improvement of the framework in several ways in the future studies.

Required a lower monetary cost and not require the expertise contribution: In order to find the most convincing evidence of given ambiguous entity, a framework that uses the multiple crowds with a lower monetary cost has been designed, and this approach does not require the expertise contribution.

1.4 Thesis Outline

The residue of this thesis is organized as follows.

Chapter 2 illustrate an overview of the related work. The chapter starts with a comprehensive literature review on academic researches of fact check and crowd-based studies on evidence verification.

Chapter 3 describes the crowdsource-based approach of this thesis. This chapter mainly focuses on the configuration of the experimental environment, along with a short description of the platform used for experiments, and lately, broadly describe the workflow design of the research.

Chapter 4 presents the experimental setting and the results of the entire work along with the tasks designs, a description of published images, requirements of workers, rewards and so on. Lately, the results with respected to the collecting and selecting phase.

Chapter 5 discuss the results and concludes the study. Finally, the chapter ends with presenting future directions.

Chapter 2

Related Work

The Related work chapter presents an overview of the relevant literature of crowdsourcing frameworks which are belonging of this thesis. The chapter consists of three sections; 1) crowdsourcing approaches for intelligent tasks, 2) approaches for a fact check and verification, and 3) quality control in crowdsourcing.

2.1 Crowdsourcing Approaches for Intelligent Tasks

Since the success of crowdsourcing techniques and sprawl of crowdsourcing during the past decade, many attempts of involving humans in crowdsourcing research have been made to exploits human computational capabilities to solve real-world problems and intelligent tasks. For examples, the intelligent tasks on generating conference programs [1], image tagging for subjective topics [2,3], understanding topics in micro-blogs [4], playing online crowdsourced games to help computers select discriminative features [8], editing a document with natural language understanding [9], finding approaches to collect data and understand human health [5,6], and finding evidences for proof in difficult problems [31], to names but a few. These factors show that 'power of people' and it is clear that crowdsourcing is a promising tool, involving multiple people to search for accurate information, gives a solution for real-world problems and to do difficult tasks such as finding evidences.

2.2 Approaches for Fact Check and Verification

Recently, efforts have been made to create web-based platforms and obtain data from the platform to verify the information. Popoola and the team [7] have developed a platform, calls 'Verily' to the purpose of produce a body to collect, share, evaluate, and prioritize of evidences. The focus of the platform is to verify information during natural disasters. According to them, this platform will handle as real-time crowdsourcing system for enabling humanitarian and governmental organizations. This study has been motivated by the Red Balloon Challenged with a mechanism of monitory reward. 'Verily' also provides incentives for referrals which is not monetary, and the crowd can get virtual points. In contrast to my approach, this platform only asks the crowd workers to submit a particular type of evidence related to the natural disaster and to verify the given evidence, the worker needs to give 'yes' or 'no' answer when departure from the platform. For example, when social media and mainstream media report the incident of the natural disasters and the information is not verified or high-resolution satellite images are not yet available, the government organization can be asked the crowd to verify the incident with answering the question and uploading the current situation.

Sethi [9] also proposed a prototype system that uses crowdsourced based argumentation structure to verify the proposed 'alternative facts' (the term used as the reference study) to help fake news detection. In this system, Sethi uses a graph-theoretical framework that incorporates the semantic web and linked open data. In particular, the system mediated by expert moderators. Therefore, in contrast to this thesis, the web system organizes the community (make groups) on the system. On the other hand, the system needs to have three kinds of members; users (requester), responders and moderators (expertise knowledge), who guide the question and answer flow.

Some other approaches have made to detect social media to identify, what fake news or the statement should be sent for the fact check by fact-checking systems. For example, Kaist et.al [10] proposed an approach to identify misinformation and detect what kind of fake news or reference statement should be sent for the fact check in social media[10].

There are some efforts made to integrate some AI techniques to identify the reference statement for the fact check, understand the content of the reference statement, and match statements against a library of claims already checked by one or more fact-checking organizations[58][59].

2.3 Quality Control in Crowdsourcing

Improve the quality of the results is the main focus area in crowdsourcing, and numerous studies have addressed this issue [12,11]. Some major approaches involve the assigning of the same task to multiple workers and then obtaining the results, to recruit the good workers by considering high approval rate or master's selection by the platform. Another approach involves the derivation of results from self-corrected individuals [13,14]. The approach of this thesis for quality assurance is to assign the same task to multiple workers and aggregate the possible evidences to support the given assumption.

Chapter 3

Approach of the Research

The Approach of the Research chapter focuses on the configuration of the experimental environment, along with a short description of the platform used for experiments described in this thesis, and lately, draw the workflow design which took place in order to complete the thesis.

3.1 Configuration of Experimental Environment

In this thesis, crowdsourcing is defined as an approach to solving difficult problems that are difficult to complete by an individual, a small team or a computer. In this crowdsourcing process, a requester publishes a task via a crowdsourcing platform—Amazon Mechanical Truk—to be available for all the undefined workers in the form of an open call. Afterward, any worker who is interested in a published task can submit an answer to the task. Finally, the requester approves the answers according to the approval mechanism and grants the rewards. Accordingly, this section discusses the crowdsourcing process of the thesis with relevant configurations.

3.1.1 Amazon Mechanical Turk (AMT)

The Amazon Mechanical Turk (AMT) is a popular labor marketplace for micro-tasks as well as for explicit. It connects the requesters with the workers who registered in the AMT. This platform used for the experiments reported in this thesis where the job postings, work pages, and compensation payments were made.

In AMT micro-tasks do not require particular subjective knowledge, expertise or training. However, user can be required high performing workers and specify the additional qualifications, workers must meet to work such as the location of the worker, born year, job title, number of HITs approved and so on.

Requester (service demanders); users who want to do assignments are called requester—who pays (with an AMT account and credit card) for the workers for satisfactory work. A requester can be a company, an organization or a person. Workers (service providers); the crowd consists of workers, also called 'turkers' performs the tasks specified (a number of workers and qualifications) by the requester can work on a task. According to AMT, a worker can work on each task only one time [53].

A human intelligence task, commonly called a HIT is a task that a requester submits to AMT workers to perform, and paid for. The lifetime of each HIT, specified by the requester which determines how long the HIT is available for workers. After completing a HIT, the requester of the task is able to review, approve and pay for (or reject and not pay).

3.1.2 Recruitment of Crowd Workers and Rewards

The workers were recruited as rewarding work on AMT. Since the AMT uses the English language, descriptions of the tasks were written in English. Workers who participated in the experiment till the end were paid the rewards, regardless of the quality of the answers.

3.2 Workflow Design

To the success of crowdsourcing research, workflows play a central role. Workflow is a kind of coordinational artifacts and it has been introduced over a century ago. Since then, workflows have been adopted as a dominant crowdsourcing infrastructure in the field of crowdsourcing today [54], which divide the research goal into small independent phases or tasks.

This section explains the workflow design of the research, carried out in two interdependent phases. In phase 1, calls 'collecting phase,' obtains possible evidences to support the given assumption. In phase 2, calls 'selecting phase,' selects the most convincing top-k (a set of K answer items with the highest ranking score according to the given scoring function or voting.) evidences from the collected evidences. Both phases were adopted by a microtask-based framework with the real crowdsourcing setting on the AMT.

3.2.1 Collecting Phase (Phase 1)

Figure 3.1 shows the crowdsource-based workflow, consisting of two phases. In the collecting phase, an image of uncertain fact (e.g., an image of scenery with an assumption of possible truth) were published, and crowd workers were asked to provide possible evidences to support the given assumption. In order to obtain the possible evidences, a scenery image was selected, which already

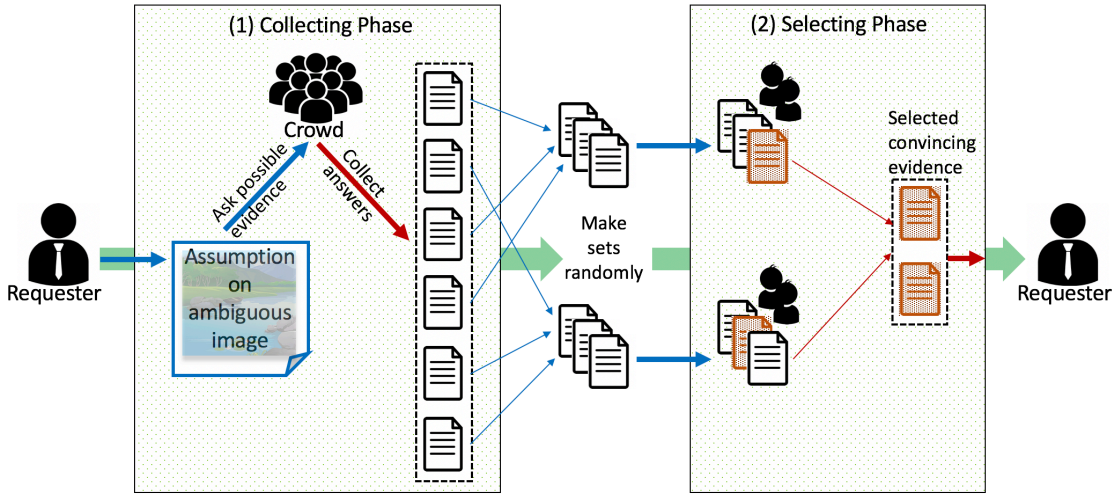


Figure 3.1: A crowdsource-based workflow consisting of two phases, from collection possible evidence (left) to selecting most convincing top-k evidence (right). Phase 1, crowd are asked to give possible evidences which support the given assumption of the image. Collected answers then used to make sets in randomly (middle). Phase 2, crowd are asked to select most convincing evidence from the sets which are having the possible evidence.

exists on the web, and then built an assumption based on the image. To understand the assumption clearly, the instructions were provided with an example of a task as well as an answer (see Figure 4.3). They have not been asked to follow any particular method to prove it, thus, allowing them freedom of providing any forms of evidences. For example, in order to prove that a photo is taken in the morning, one might say that the direction of the shadow is evidence with reference to the Google map, or another might say that she/he found the same picture in a blog article that states it was taken in the morning. However, the answers were required in the text and/or URL address. All the evidences obtained in the collecting phase were then used to select the most convincing top-k evidence by the crowd in selecting phase.

3.2.1.1 Characteristics of the Collected Answers

The possible evidences that could be able to support the assumption divided into two categories as 'strong possible evidence' and 'weak possible evidence' according to the reviewer's observation. The definition of the two categories are described below:

In the **Strong possible evidence - definition**

1. workers give an evidence/s to prove the fact with confirming the assumption is correct, and
2. input enough proofs to prove the given evidence and which is correct.

For example, some crowd workers found evidence on the web pages with finding the similar

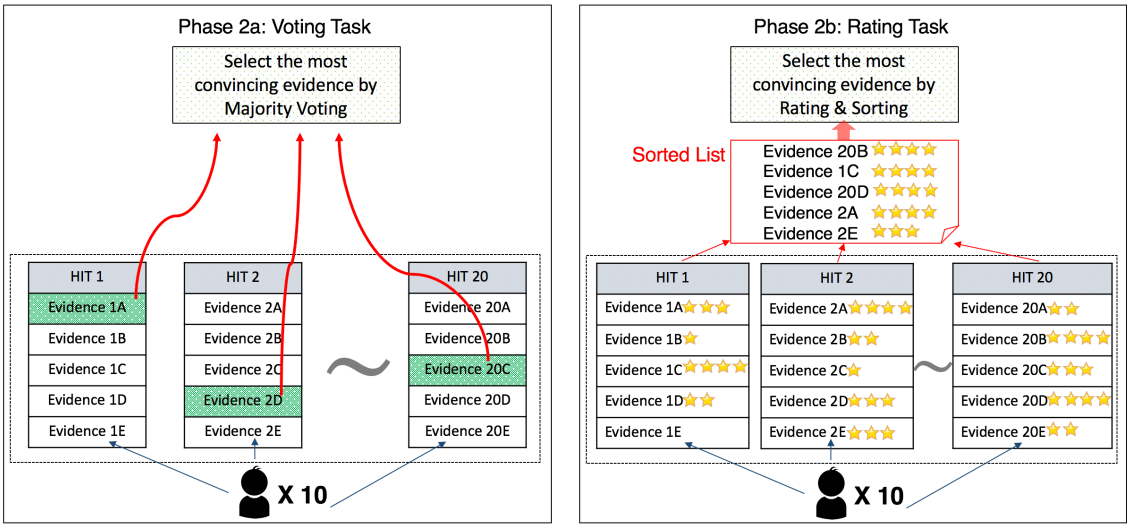


Figure 3.2: Task design of phase 2 to find the most convincing evidence by using two ways. The voting task (left) selects multiple evidence and asks 10 crowd workers to vote an evidences from each HIT. Rating task (right) also selects multiple evidences and again asks 10 crowd workers to rate for each evidences in a HIT. Rating scores then sort by mean and standard deviation to find the top-k evidences.

or the same image with the words of *'Morning'* or *'Dawn'*. Some crowd workers input the Google Map at Mt. Tsukuba with reference to a direction and the shape of the Mt. Tsukuba proving that the photo is taken at West side of the mountain.

In the **Weak possible evidence - definition**

1. workers give an evidence/s to prove fact with confirming the assumption is correct, but
2. DO NOT input an enough proofs to prove the given evidence and/or which is incorrect.

For example, some crowd workers describe the existing features on the photograph such as sky color, mist, and flowers as morning indicators without any proof to be a strengthened evidence.

3.2.2 Selecting Phase (Phase 2)

In the second phase, the collected evidence-based proofs were ranked according to the results of inputs from another set of crowd workers. To utilize the crowd workers to find top-k evidences, as shown in Figure 3.2, two widely-used methods were used. The first task is voting, calls phase 2a, which selects the most convincing evidences and asks the crowd to vote an evidence from each set of evidences. The second task is rating, calls phase 2b, which also selects the most convincing evidences and asks the crowd to assign a rate of individual evidences in the given set.

3.2.2.1 Voting tasks

First, the number of collected evidences were randomly divided into a set of HITs with each having 5 evidences. Note that, any of the collected answers have not been removed in this phase, and if there are occasional repetitions in the same HITs, crowd workers were asked to vote any of the convincing evidences from the repeated evidences. In order to select top-k evidences by majority voting, 10 crowd workers were assigned to complete each HIT.

3.2.2.2 Rating task

The same sets of HITs use for the rating task, and again, 10 crowd workers were asked to complete each HIT. The rating scale that award, fall from 0 to 4 as follows:

- Scale 4: Everyone agrees that the assumption is true
- Scale 3: Very likely that the assumption is true
- Scale 2: Some people may believe that the assumption is true
- Scale 1: This is related to the assumption but there is a large room for objection
- Scale 0: Nothing related to the assumption

To find the top-k evidence and data processing efficiency, all the answers were sorted by using the shell sorting method with respect to mean (μ) and standard deviation (σ).

Chapter 4

Experiment and Results

The Experiment and Results chapter presents the tasks designs, along with a description of published images, requirements of workers, rewards and so on. Each phases are ended by presenting the results with respected to the collecting and selecting phase.

4.1 Question to be Addressed

There are three questions that the researcher wanted to answer in the results as follows;

From the collecting phase,

- **Q1**: What are the distributions of collected answers among the three categories?
- **Q2**: How long it takes for them to reach the evidences?

From the selecting phase

- **Q3**: Which selecting method is most suitable for the selecting phase?

4.2 Collecting Phase

4.2.1 Overview of the Experiment of Collecting Phase

In the experiment of collecting phase, two cases were conducted using two different photographs via Amazon Mechanical Turk (AMT) as Case 1: Mt.Tsukuba and Case 2: San Diego. The differences between these two cases manly focus on the effort of complete the task and reach to the answer. Mt.Tsukuba photograph is online resource and worker do not need to follow long step to reach the evidences. In contrast, San Diego photograph is off-line resource (taken by researcher) and worker need to follow long step to find accurate answers (see Figure 4.1). The assumption for Mt.Tsukuba

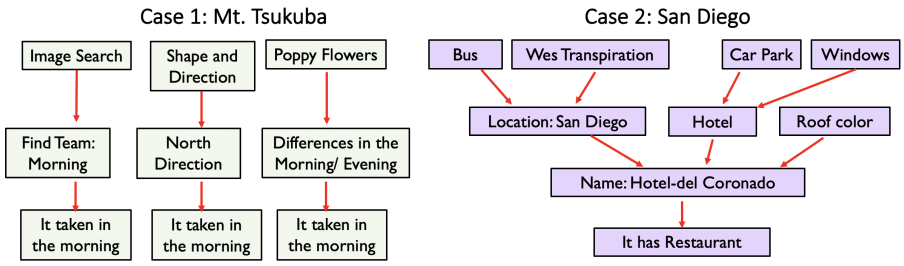


Figure 4.1: The differences between the two cases in collecting phase. Case 1 (left) is Mt. Tsukuba and Case 2 (right) is San Diego

case was "this photograph is taken in the morning" and for the San Diego it was "Inside the building in the shown picture has a restaurant.

The purpose of the experiment is to detect the potential of evidence-based proof by crowd-sourcing. Therefore, parameters were not necessarily determined in a systematic way. For each task, based on the objects and the scenery of the photograph, the assumption was built and asked the crowd to support the assumption by giving possible evidences. The following are some answers that the researcher expected from the crowd.

- Example 1: *"This YouTube video is presenting the sun rising at the Mt. Tsukuba in the same area. Therefore, I can say that the photograph is taken in the morning: (copy the URL)"*
- Example 2: *"Following URL to Google Map is showing the shape and the direction of Mt. Tsukuba. According to the shape of the mountain and the north point of the map, I can say that the mountain must be in East of the photograph. Therefore, it might taken in the morning: (copy the URL)"*

Finally, to understand the characteristics of the real answers from the crowd, the collected answers were carefully reviewed by three reviewers manually. The contents of each task will be described in detail in the next sections.

4.2.2 Procedure for Case 1: Mt.Tsukuba

For the case 1, a scenery photograph of Mt.Tsukuba along with poppy flower field at the Kokaigawa Fureai Park in Japan was selected [55], in which it is unclear *whether the photograph is taken in the morning or the evening* (see Figure 4.2). Based on this photograph, the assumption of *'This photograph is taken in the morning'* was built.

To obtain 100 possible evidences supporting the given assumption, 100 crowd workers were assigned to the task. They will be having a reward of about 0.30 USD when they complete the task. The task was available for the 10 days and crowd workers were asked to complete the task

Figure 4.2: The photograph of Mt. Tsukuba along with Kokaigawa Fureai Park, Japan (Case 1)

within two hours. Here, the researcher expected that the crowd worker should find some evidence within two hours of time.

Figure 4.3 shows the task design of Mt. Tsukuba. The task design consists of 4 parts; 1) Photograph, 2) Title of the instruction, 3) Instruction field and 4) Input field. In this task, a worker can simply perform a task with finding possible evidences and submit the answers into the input field as text and/or URL address. See the experiment result in the following section.

4.2.3 Results of Case 1: Mt. Tsukuba

One-hundred possible evidences were obtained in the first task, among which 41, 12 and 47 answers were obtained as a text description, only a URL link to the web page, and a text description with a URL link to the web page, respectively. To understand the characteristics of the real answers from the crowd, the collected answers were carefully reviewed by three reviewers manually and obtained results in majority voting among the three reviews.

Figure 4.4 displays a pie chart which shows the characteristics of the obtained possible evidences. Among 100 inputs, 43 were unable to support the given assumption—workers unable to input any possible evidence to prove facts and/or input wrong information. In contrast, 57 inputs were able to support the assumption in different levels.

The possible 57 evidences that could be able to support the assumption divided into two categories as ‘strong possible evidence’ and ‘weak possible evidence’ according to the reviewer’s

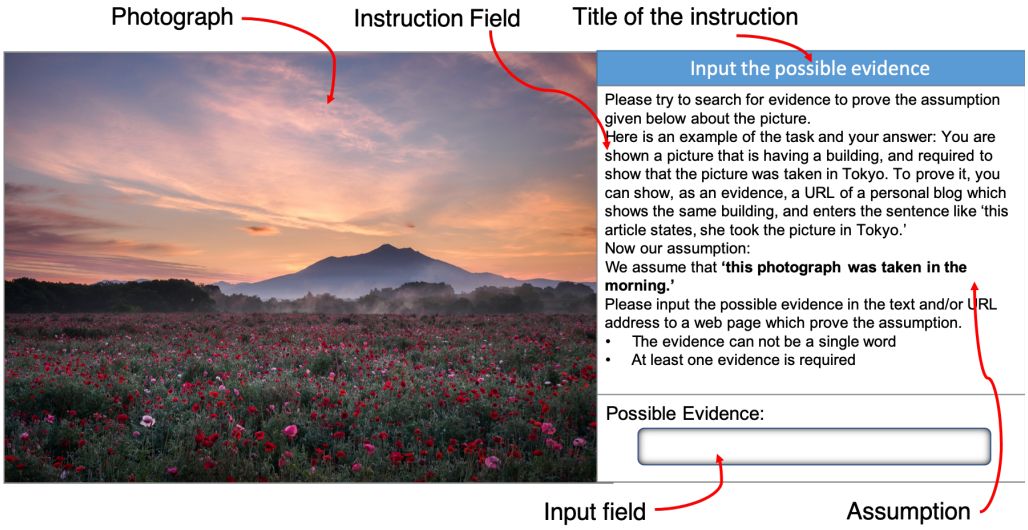


Figure 4.3: A task design for the collecting phase. The lower part of the instruction is an input field, crowd can utilize the space to submit the possible evidence.

observation. Accordingly, 44 and 13 out of 57 possible evidences were identified as a strong and weak evidences. In addition, apart from six evidences, all the strong possible evidence was straightforward, understandable and not contradictory assertion.

In order to understand the characteristics of the collected possible evidences, here, the collected answers in each category have been presented (see Figure 4.6).

Example 1: Some crowd workers found evidence on the web pages with finding the similar (see example 1.2) or the same image (see example 1.1) with the words of *'Morning'* or *'Dawn'* and input the URL address. Since the answer is matched with definition 1, this answer is recognized as strong possible evidence and assume that the function of 'reverse image searching' is used by workers. The workers might found the given photograph on the web because the Mt. Tsukuba image is already on the web.

Example 2: A few crowd workers input the Google Map at Mt. Tsukuba with reference to the direction and the shape of the Mt. Tsukuba proving that the photo is taken at West side of the mountain.

Example 3: This image shows an interesting example and it was strong possible evidence. In this case, the worker has considered the behavior of the poppy flowers in the morning and at night. According to the provided articles with evidences and some proofs, the worker said that the given photograph would be take in the morning.

When considering the diversity of the strong answers it is clear that workers have found diverse methods and key-works for proofs. For case 1, seven types of diverse answers found by works for

strong evidences.

Among the 44 strong possible evidences, 48% were found the very similar image from the internet with the title of 'Peak of Dawn'. Similarly, 25% and 9% answers were found a similar image from the internet with the title to 'At Dawn' and 'Early Morning' respectively. 7% of answers have claimed the answers with both Google Maps and term 'Morning' on the web page with the same image. Interestingly, 2% of the answers were focused on the objective of the given photograph and found the proud some surprising answers. Another 2% of the answers have given the video links which shows the sunrise of the Mt. Tsukuban (see Figure 4.6).

4.2.3.1 Reaction Time of Case 1

Figure 4.7, Box plot graph shows the elapsed time or the reaction time per HIT by a worker to complete one task. According to one-way ANNOVA test, no significant difference ($p<.05$) was identified between the three categories ($p= 0.249$), strong and weak categories ($p=0.136$), strong and no evidence categories ($p=0.445$) and weak and no evidence categories ($p=0.145$) for the time required per HIT. The comparison of the other statistics showed that 56%, 85% and 98% of the answers were given within 5 min, 15 min and 1 hour of the allocation time, respectively. Furthermore, for the many workers, the average elapsed time per task was 11 min 35 s. A maximum of 1 hour 57 min and a minimum 45 s were taken to complete the tasks. The average time allocation for the 'strong evidence' and the 'no evidence' were 10 min 12 s and 13 min 26 s, respectively. This implies that it is very difficult to analyse the relationship between finding evidences and reaction of workers. Although some answers did not provide any reasonable evidence, the workers have taken much time to complete the task. The reasons for these kinds of difference are going to discuss in the discussion section of the next chapter.

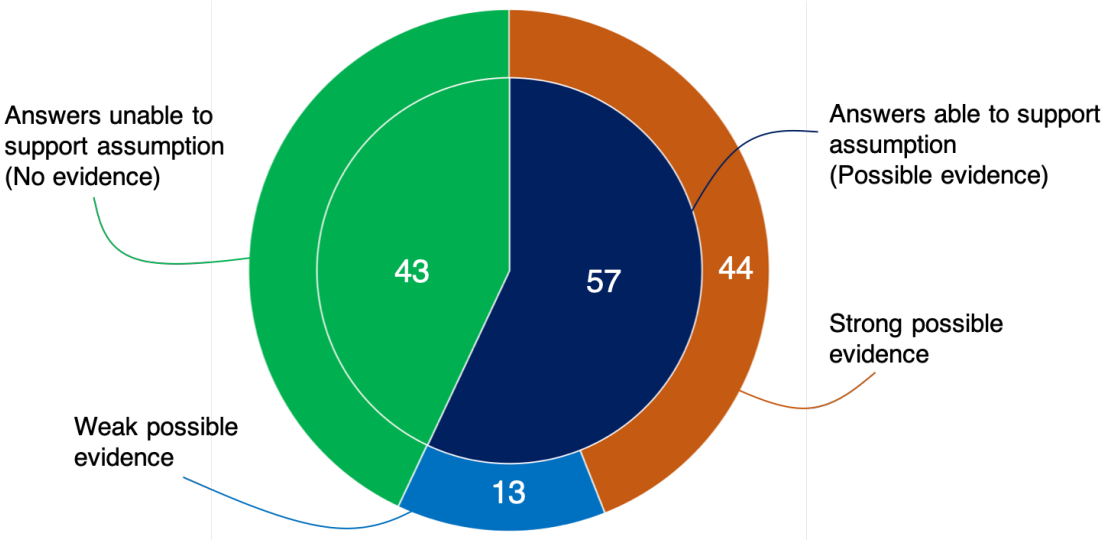


Figure 4.4: Results of Case 1—Mt. Tsukuba: Pie chart shows the characteristics of the obtained possible evidences. From 100 answers, 43 answers were unable and 57 possible evidences were able to support the given assumption. Possible evidences can be divided into two categories (1) strong possible evidence: 45; and (2) weak possible evidences; 13.

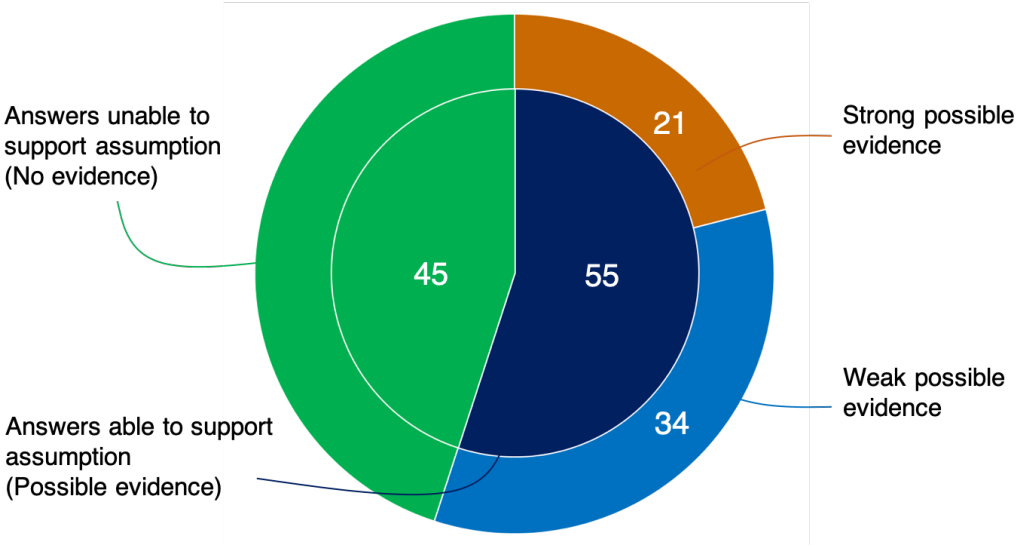
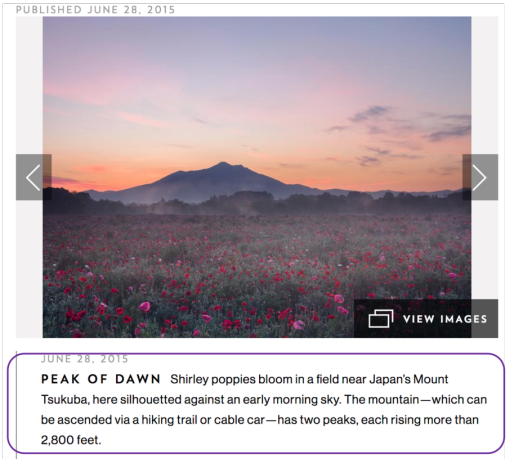


Figure 4.5: Results of Case 2—San Diego: This pie chart shows the characteristics of the obtained possible evidences in task 2. From 100 answers, 45 answers were unable and 55 possible evidences were able to support the given assumption. Possible evidences can be divided into two categories (1) strong possible evidence: 21; and (2) weak possible evidences; 34.

Example answer 1: ID 32
The photo description at
http://travel.nationalgeographic.com/photographer-of-the-year-2016/gallery/week-6-nature/19/
says the picture was taken at dawn.

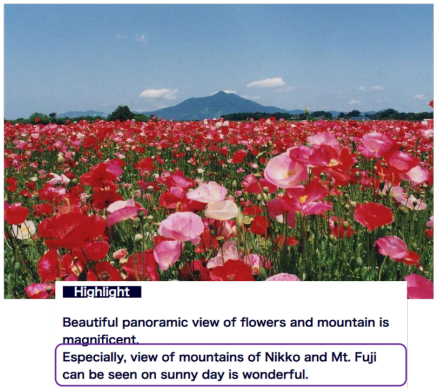
(a) Example 1.1

Example answer 1: ID 74
https://www.nationalgeographic.com/photography/photo-of-the-day/2015/6/mountain-view-poppies/
National geographic picture with explanation of the shot being early morning sky. The mountain is in the same silhouette



(b) Example 1.2

Example answer 2: ID 22
"The flowers are poppy flowers. Evidence: http://f-pedia.r-cms.biz/en/topics_detail8/id=738
Poppy Flowers bloom in the morning and close at night
Evidence: http://www.sciencefocus.com/qa/why-do-poppy-flowers-open-morning-and-close-night
Since all the poppy flowers are closed it can be assumed its morning since if it was evening at least some would be open.



(c) Example 2

Example answer 4: ID 53
"I CAN SEE THAT THE SUN IS RISING JUST BEHIND THE MOUNTAIN. THIS MEANS BEHIND THE MOUNTAIN IT SHOUL BE EAST.
I CHECKED ON GOOGLE STREET VIEW THE SHAPE OF THE MOUNTAIN FROM WEST AND THE SHAPE IS SIMILAR TO THE ONE IN THE PICTURE"
https://www.google.de/maps/@36.205271,140.0535805,3a,75y,90t/data=!3m6!1e1!3m4!1sdTF8ULCegS7Ggj9xdRmQlw!2e0!7i13312!8i6656!5m1!1e4

(d) Example 3

Figure 4.6: Inputs of the strong possible evidences by crowd workers.

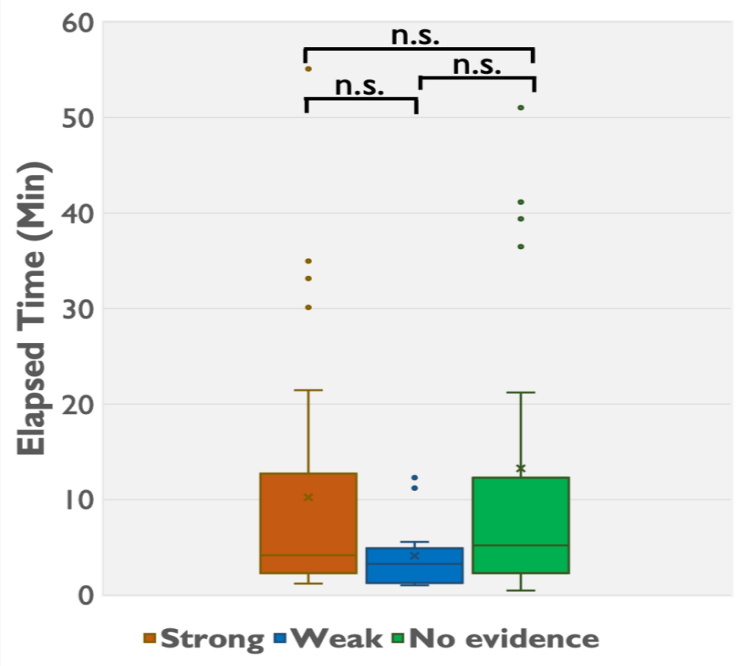


Figure 4.7: Box-Plot of Case 1 shows the required time of period for one worker to complete one task. The horizontal axis represents elapsed time in minutes up to one hour and the vertical axis represents the categories of collected answers. No significant difference ($p<.05$) was identified between three categories ($p= 0.249$)

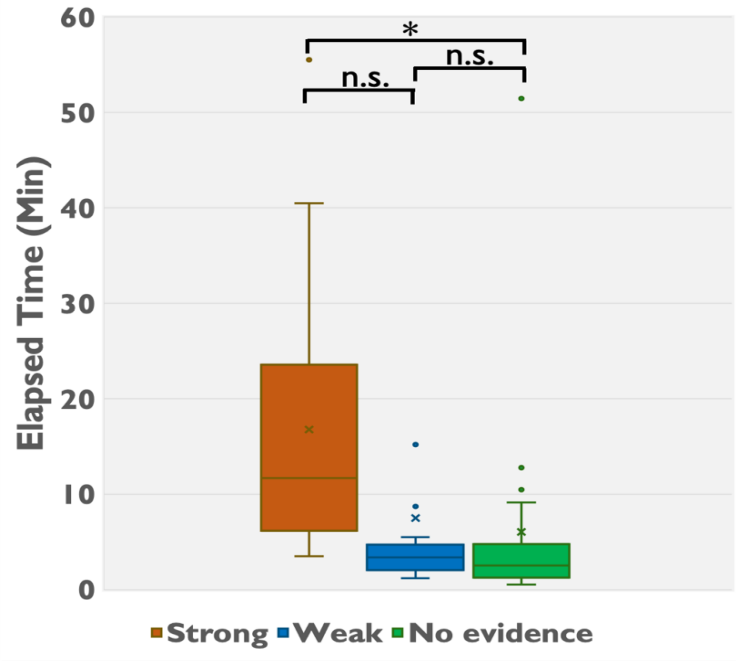


Figure 4.8: Box-Plot of Case 2 shows the required time of period for one worker to complete one task. The horizontal axis represents elapsed time in minutes up to one hour and the vertical axis represents the categories of collected answers. A significant difference ($p<.05$) was identified between three categories ($p= 0.045$)

4.2.4 Procedure for Case 2: San Diego

In the second case, a scenery photograph of Hotel Del-Coronado and the surrounding in San Diego, USA was selected. The difference between the two photographs is, although the first photograph exists on the web, the San Diego photograph has not been published since it was taken by the researcher and kept as off-line (see Figure 4.9). Based on this photograph, the assumption of *'Inside the building in the shown picture has a restaurant'* was built.

Figure 4.9: The photograph of Hotel Del-Coronado in San Diego, USA (Case 2)

As the purpose of the first case, to obtain 100 possible evidences supporting the given assumption, 100 crowd workers were assigned to the task and they will have to receive a reward of about 0.30 USD when they completed the task. The task was available for the 10 days and crowd workers were asked to complete the task within three hours. Here also, the researcher expected that the crowd worker should find some evidence within three hours of time. The task design of San Diego also, consists of 4 parts; 1) Photograph, 2) Title of the instruction, 3) Instruction field and 4) Input field. In this task, a worker can simply perform a task with finding possible evidences and submit the answers into the input field as text and/or URL address. The following section presents the results of task 2.

4.2.5 Results of Case 2: San Diego

Among the 100 possible evidences were obtained in the second task, 67, 3 and 30 answers were obtained as a text description, only a URL link to the web page, and a text description with a URL link to the web page, respectively. When compared to task 1, in this task, more than 50% of the answers were obtained as a text distribution. According to the observation of the answers, reviewers identified that in contrast to the Mt. Tsukuba photograph, workers were tried to provide the possible evidences with describing the object on the photograph. For example, 51 out of 67 text descriptions were considered the building size, architecture of the building, size of car parking and so on. Therefore, to understand the characteristics of the real answers from the crowd and for categorization, the collected answers were reviewed manually by three reviewers and obtained results in majority voting. The results of the reviews illustrated in Figure 4.5.

4.2.5.1 Characteristics of the Collected Answers

According to the categories, among 100 inputs, 45 were unable to support the given assumption—workers unable to input any possible evidence to prove the fact and/or input wrong information. In contrast, 55 inputs were able to support the assumption. Accordingly, 21 and 34 out of 55 possible evidences were identified as strong and weak evidences respectively (see Figure 4.5). The final input of the one-fifth answers was the official website of Hotel Del Coronado, saying that the hotel has many restaurants, therefore, the assumption should be correct. In particular, as expected, the name of the transportation system on the bus was noted by some workers. Afterward, they found the locational details and searched by different forms. Among them, two of the strong possible evidences and one weak possible evidence were chosen for detailed presentation in below.

Example 4: Most noticeably, 9 out of 21 answers were provided with the steps of the finding evidences by them, and it was interesting. The answers are also straightforward and understandable.The following two statements were reported by the workers in the worker's own words. Statement 1;

> *"First and foremost-using* ***what is written on the school bus*** *we can determine that this is taken somewhere in San Diego- http://www.certifiedsandiego.com/. Next, due to the layout of the* ***windows on the building, and all the cars in the parking lot****, it is safe to assume that this is a hotel of some sort. A quick google search of* ***'large white hotels in San Diego with red roofs'*** *reveals the Hotel-delCoronado, which looks remarkably similar to the building pictured on this Hit. See pictures here- https://www.google.com/:Once you look at the website for the Hotel del Coronado, we determine that* ***YES, it DOES have a restaurant inside****. That is how we find out that the building shown in the picture here does have a restaurant inside."*

Statement 2;

> *"I searched for he* ***name of the bus company****, it was from San Diego county, after that i searched for* ***'san diego county hotel red roof'****, I found a photo of a hotel with similar rooftops named* ***'Hotel del Coronado'****, made a quick search on Google Maps and found out the exact location the picture was taken from-* *https://goo.gl/maps/* *-* ***In the hotel webpage*** *-* *https://hoteldel.com/* *- you can find various restaurant options, but the hotel has different buildings, Inside the one from the photo there are at least 3 different restaurants listed on Maps: 'Crown Room Sunday Brunch', 'Sheerwater' and '1500 OCEAN"'*

Example 5: Three workers have had experiences in the area of the photograph, therefore those workers shared his/her experience as possible evidence. One noted below.

> *"Yes, there is a restaurant in this building. The building is the iconic Hotel Del Coronado.* ***While living in San Diego, we saw it multiple times and thought it was beautiful. The name on the bus is Wess Transportation, which services San Diego****. That fact confirmed which hotel this was for me and on the website for the Hotel del Coronado, there are multiple pictures to verify that it is the correct building. The website also shows that there are multiple restaurants and i will provide the link to that proof.* *https://hoteldel.com/restaurant-coronado/*

Example 6: In this example, worker has tried to indicate some possible evidences according to his/her observation and experience. The answer itself seem that it is not incorrect. The argument which is built by the worker can not be excluded. Therefore, although it does not have the proofs, these kind of answers were accepted as weak possible answers.

> ***"I would assume that this is a large hotel due to the fact it is well kept, has a car park, looks to have good transport links and the grounds look well tended****. This assumption leads me to support the assumption that this building must have a restaurant as its a sizable hotel so would need capabilities for catering for their guests."*

When considering the diversity of the strong answers of case 2 it is clear that workers have found diverse methods for proofs. Among the 21 strong possible evidences, 43% were following the steps from finding location name using 'Wess transportation' to reach the final answer. Similarly, another 43% of answers were found 'Hotel de Coronado' webpage by following some steps. But at this time, workers did not clearly describe the followed steps in their answers. Interestingly, 14% of the answers were based on their personal experiences in the location of the hotel.

4.2.5.2 Reaction Time of Case 2

Figure 4.8, Box plot graph illustrates the elapsed time or the reaction time per HIT by a worker to complete one task. In contrast to case 1, in this task, a significant difference ($p<.05$) was identified between the elapsed time of the three categories ($p= 0.045$). Although there is no significant difference was identified between the categories of strong and weak possible evidences ($p= 0.080$), there was a significant difference between the strong and no evidences categories ($p= 0.003$). Among the evidences of weak and no evidence categories ($p=0.713$) as shown in Figure 4.8, no significant difference was identified.

The comparison of the other statistics showed for each worker, the average elapsed time per task was 9 min 19 s, and a maximum of 2 hours 11 min and a minimum 55 s were taken to complete the tasks. Furthermore, The average time allocation for the 'strong evidence' and the 'no evidence' were 17 min 17 s and 6 min 06 s, respectively. In averagely, 59%, 82% and 98% of the answers were given within 5 min, 15 min and 1 hour of the allocation time period, respectively. This implies that the elapsed time differences between the categories of evidences cannot determine by the single experiment and the discussion of the two tasks will present the next chapter.

The observation of the two cases suggests that the further analysis is required to search for the most convincing evidences, because the results of collecting phase alone do not specify the most convincing evidences since it has a list of strong possible evidences. Therefore, the second experiment was conducted.

4.3 Selecting Phase

4.3.1 Overview of the Experiment of Selecting Phase

In the experiment of selecting phase, only the Mt. Tsukuba task was examined further. The task was conducted on AMT and the purpose of the experiment is to find the most convincing evidences from the collected answers. The task designs of the Selecting phase are show in the Figure 4.10 and in the Figure 4.11. To achieve the above aim, two tasks were designed by utilizing two widely-used methods (see Figure 3.2). The first task is voting, call phase 2a, which selects the most convincing evidences and asks the crowd to vote an evidence from each set of evidences. The second task is rating, call phase 2b, which also selects the most convincing evidences and asks the crowd to assign a rate of individual evidences in the given set. Afterward, both results were compared.

4.3.2 Procedure for Voting Task and Rating Task

In the voting task, the obtained 100 evidences were divided into 20 sets of HITs by each having 5 evidences. Among the 5 evidences, the crowd workers were asked to vote any of convincing evidences in the HITs. Ten workers were assigned to complete one HIT (see Figure 3.2). Accordingly, 200 votes (20 sets * 10 votes) were obtained from the voting task.

In the rating task, the same 20 sets of evidences were used as phase 2b, and 10 crowd workers were asked to rate all evidences in each HIT and 1000 rates were obtained by each evidence having 10 rates. The collected rating scores then sorted as described in chapter 3. In both tasks, each crowd worker receives a reward of about 0.10 USD. The tasks were made available for the 10 days and were to be complete within two hours.

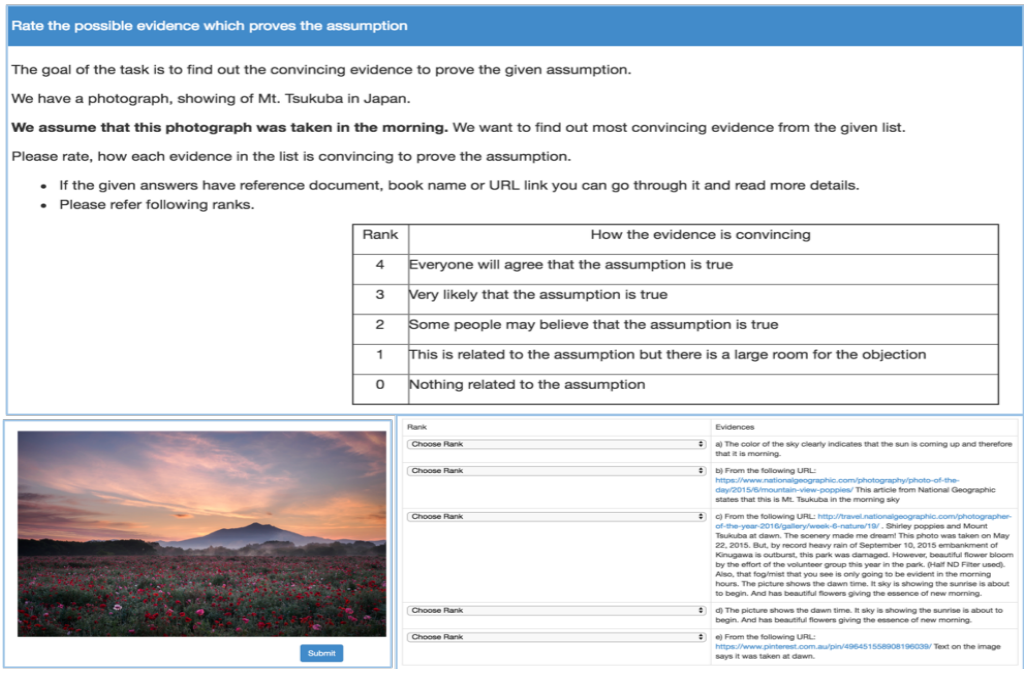


Figure 4.10: The Task Design of the Voting Task

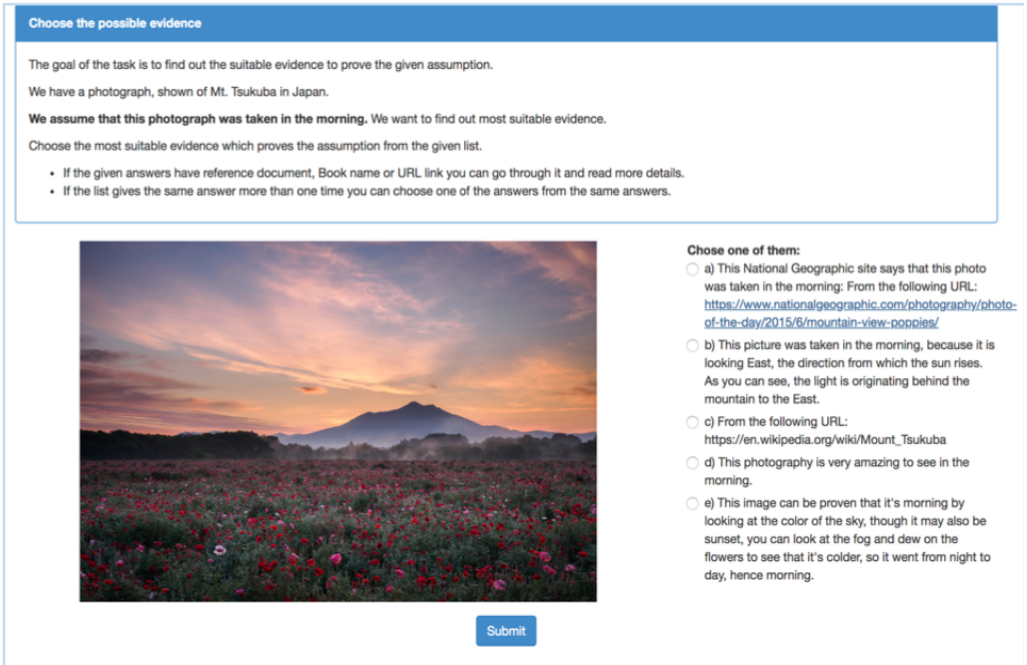


Figure 4.11: The Task Design of the Rating Task

4.3.3 Results of Selecting Phase

In the phase 2a, 200 votes were given by 200 crowd workers and the highest voting per HIT was counted. As a result of the highest similar number of votes in one HIT, the top-23 possible evidences were counted in the voting task. Then they have been compared with the highest rating and sorting top-23 possible evidences obtained from the rating task.

As detailed in the Table 4.1, the main observation of both voting and rating tasks in three categories can be presented as follows. Only 16 majority votes out of top-23 were selected from the strong evidences which are categorized by the reviewers. Among the remainder of the votes, 5 were identified as weak evidences and 2 answers were unable to support the assumption. Note that, the number of collected evidences were randomly divided into a set of HITs in this task and any of the collected evidences have not been removed in these tasks. If there are occasional repetitions in the same HITs, crowd workers were asked to vote any of the convincing evidences from the repeated evidences. The reason for having weak evidences and answers having no evidence as a top-23 is that in some HITs published with only weak evidences or incorrect answers in both task. Since the HITs were created randomly with having 5 answers, there was a chance to create a HIT, all having incorrect answers. Therefore, from the voting task, only 21 votes could be recognized as possible evidences to support the given assumption.

Table 4.1: Results of phase 2a and 2b

Task	**Answer category**		
	Strong evidence	***Weak evidence***	***No evidence***
Voting	16	5	2
Rating	22	1	0

In contrast, the top-23 evidences from the rating task be performed 22 strong evidences and only one weak evidence which was also found in the voting task. Therefore, it is interesting to examine the weak evidence itself, to understand why that particular evidence got highest voting and rating scores even though it is labeled as weak. The following statement was recorded in the workers' own words for the particular weak evidence.

> *"There appears to be fog at the base of the morning. Such fog typically appears in the dawn and early morning. It is highly unusual for there to be such fog later in the day, especially when there are high clouds and some blue sky above."*

Our argument is although the answer does not prove the difference on the fog between morning and evening through the Web link or any other, owing to the long explanation, the crowd workers assumed that fog and high clouds is a good reason for the image to be taken in the morning.

Chapter 5

Discussion and Conclusion

The final chapter, revisit the research goal put forth at the beginning of the thesis, discuss the results and limitations. Afterward, the thesis concludes with the brief introduction of the future directions.

5.1 Discussion

5.1.1 Revisiting the research Goal

The focus of this thesis was to introduce a crowdsource-based approach, in order to address the limitations and the gaps of the existing approaches and the issues of fake and ambiguate information on the internet. Since the almost every web-based approaches are adopting only the techniques of ranking to find the facts, and several academic research papers have done with the purpose of verifying fact and efforts made to understand the reference statement, and do not introduce a principle way to prove fact with finding evidences, the researcher motivated to find the straightforward approach with crowdsourcing. This study also provides insight into the crowdsourcing capabilities in difficult and complex tasks that are difficult to complete by an individual, a small team or a computer. Since crowdsouricing is a promising tool to explore the power of people, this work also clarified that the proposed crowdsourcing workflow remain a good fit to explore the power of people in this kind of problems.

It seems that people may also have different level of awareness and resourcefulness in finding evidences and presenting the results in the online market places. In addition to that, the crowd workers may have a different level of commitment and quality; some people might adopt at the task while others struggle with the finding evidences. Therefore, when developing an approach for

leveraging human computational capabilities and enhance the power of people, it needs to ensure the quality of performance of the workers by understanding the consequences of the spectrum of the crowd abilities.

To ensure the quality of the results, one of the major approaches is to aggregate answers from multiple people. By following this method, for the first part, the same task was assigned to multiple workers and obtain the possible evidence to support the given assumption which allows ensuring the quality of the results. Second, one of the standard quality control techniques of majority voting employed. In majority voting, generally, a requester asks same question from crowd workers and the answers that are given by a majority of the responses will be selected. On the other hand, when considering the quality matters, the issues can occur with the task itself such as question are difficult, subjective, ambiguous or open to various interpretations or quality control is more difficult to be ensured [56,57] in explicit crowdsourcing.

Another important aspect of the proposed crowdsource-based approach is, since the standard methods are not enough the subjective tasks or the open interpretation tasks are handled to address the real-world problem. For example, although there are many web-based fact check systems exist, they only take the majority voting or rating to assess the given statements and do not evaluate the subjective answers of web users or the discussion related to subjective answers will be happened in the small groups. Therefore, in this approach, tried to overcome these kinds of issues as well.

5.1.2 Discussion of the Results

When it comes to the discussion on the results of this study, first can consider the collecting phase. The results of the collecting phase demonstrated that this crowdsource-based approach is able to collect various possible evidences to support the given assumption and to prove the facts in the given situation to some extent. According to the results of both Mt.Tsukuba and San Diego cases, the researcher noticed that more than half of the collected evidences supported the assumption, and approximately one-third and one-fifth of the possible evidences were strong respectively.

Since the one photographs was founded from an online (Mt.Tsukuba photograph) and the other one from off-line (San Diego photograph) sources, the differences of the answering methods could be identified. As the steps of finding evidences have not been asked in Mt.Tsukuba case, can assume that many crowd workers tried to find the possible evidence by the image-searching option on the web. Among the 100 answers, 21 possible evidences were provided with the same URL link displaying a very similar image, stating that the photograph is showing the early morning. In contrast, in San Diego task, 18 out of 21 strong possible evidences were found by observing the photograph. For example, workers have searched the name of the bus company which shown in

the photograph and found the location. Afterward, according to the building structure and colors, the crowd found the evidences from the web. This indicates that, although the crowd workers were not the expertise of the related fields, they are powerful enough to search evidences with their own ways.

When comparing the two task, can see that, in both tasks, more than 50% of the answers were able to support the given assumption. But when considering the supported answers, it is clear that, if the given image is off-line resource, many workers tried to discuss the evidences as text description and did not try to search for proofs. This might happen due to many reasons. Especially those reasons unique to worker behaviors. The crowd may intentionally provide incorrect or poor answers since they work as spams or works only for the rewards. In that case, they may want to complete the task within a few minutes. Another thing is that the worker may perform unintentionally due to confusion or ignorance of the task. On the other hand, since the interpretation problems, task design problems or ambiguation problems can arise, it is very important that much consideration of these parts in my future research on this topic.

When examining the elapsed time for the finding evidences and completion of the task, can observe that some workers have spent plenty of time to reach the answers. For example, around 10 to 16 minutes were spent to search for strong possible evidences in Mt.Tsukuba and San Diego tasks respectively. This duration of the reaction can be considered as enough time to reach the quality and accurate answer in both cases. In contrast, when consider the duration of the reaction which were unable to support the assumption, in case 1, although the answers are incorrect, some workers have spent much time to reach the answers. There was no significant difference identified in case 1, while, there was a significant difference identified in case 2. This implies that it may be difficult to filter out strong/weak proofs based on the reaction time and difficult to come up with the concrete answer regarding the relationship between finding evidence and reaction time. Worker behaviors, online networks, location of the workers, spam workers, task design and many other reasons can be affected for the reaction time.

According to the consideration of the voting and rating tasks in the selecting phase, the rating method can be identified as a better method to find the most convincing evidence. The experimental results showed that with the use of the voting method, there is a possibility of obtaining weak evidence according to the number of sets and method of making sets. Further, the rating method shows very similar answers for the highest score.

These findings are also providing useful viewpoint, let requester understand how workers interpret the instruction, and where there is ambiguity on the task. Notably, this approach can be combined with any other approaches to improve the data quality of the work. Thus, the researcher believe that this design is an important step to develop a more powerful framework for proving the assumption of a given statement by crowdsourcing. The approach of utilizing the crowdsource-based framework has much room for improvement, and it is open for future studies.

5.2 Limitations

This study has several limitations that need to be further addressed in our future works. Due to the limited time, only the reference images were used for the experiments and investigate only the two cases for the collecting phase in this research. Additionally, the methodology for the collecting possible evidences were based on an assumption in this time and only the hundred workers were recruited for each case in the collecting phase.

Further research should attempt to address these limitations and to make the results more generalizable by increasing the case studies, the sample size and so on.

5.3 Conclusion

The explosive development of information techniques make humans and electronic equipments be tightly connected than ever before and humans can easily connect to one another on the internet. Through the several applications on the internet or the web, humans are willing and capable to share their knowledge, experience, skills, and perception. In the meantime, the concept and the techniques of the crowdsourcing have been developed and widely applied in various human intelligence tasks and solve real-world problems such as image tagging for subjective topics, data translations, generating programming codes, verify the natural disasters and so on. Therefore, crowdsourcing has become increasingly popular and considerable success in academic researches. This motivation has become an accelerator to introduce and investigate a crowdsource-based framework, in order to address the limitations of existing works of the spread of misinformation and disinformation—in the other word fake information on the online setting.

Over the last several years, although there are many web-based systems have been developed to detect misinformation or disinformation, some serious problems can be identified on those services. The approach of those systems is limited by getting only the rating on the statements from the users or investigate the fact by a few team members. Therefore, through this thesis, the researcher have tried to introduce a novel framework to detect accurate information by using the nascent tool of crowdsourcing by ensuring the quality of the results.

This thesis proposed a framework consisting of two phases and ask crowd workers to prove the fact by finding possible evidences. The framework uses multiple crowds with a lower monetary cost and does not require the contribution of expertise. The results of the thesis show that more than half of the answers from the crowd, were possible evidences, and approximately one-third of the possible evidences were strong. The study also demonstrated this crowdsource-based approach is able to collect various possible evidences on difficult problems and provide diverse methods for proofs, among which some were surprising as we did not factor those as proofs. In particular, the study implies that the crowdsourcing could help effectively to discover the possible evidences on inaccurate information or the given statements, that could be difficult for an individual, a small team or a computer. The framework also effective to choose relible proofs among the list of many possible evidences. This shows the power of crowdsourcing and the experiment shows that there is much room for further improvement of the framework in several ways. Thus, in the researcher's future work, it aims to be investigated effective and efficient workflows for the fact check underdeveloping a complete framework for crowdsource-based proofing system in crowdsourcing and address the limitation of the research.

Acknowledgement

I would like to express my sincere gratitude to the persons who supported me all the way through my master's career, and without them, this thesis work would not have been a success.

First of all, and most importantly I would like to express my sincere gratitude to my principle academic advisor, Professor Atsuyuki Morishima for providing me the opportunity to study this topic. His professional suggestions, guidance, support, feedback, and patient supervision kept me researching in the correct direction during my M.Sc. journey.

Thereby, I would also appreciate the Assistant Professor Masaki Matsubara for his guidance, support, and feedback in this thesis. My academic journey would not have been so rewarding without his kindness and wisdom. Furthermore, I would like to thank my second academic advisor, Associate Professor Hideo Joho for his feedback and advice during the complete research process.

In addition to my advisors, I am very happy to do worked with my Lab members (FusioncompLab) who provided feedback, conversations and help on this work. Fusioncomp Lab mates and fellow graduate students made my experience encouraging and exciting. I also thank the amazing administrative efforts of Fusioncomp Laboratory staff.

In particular, I would like to acknowledge to the Japan International Cooperation Agency (JICA) and the scholarship program of Innovative Asia for the grate support for my masters studies.

Most importantly, I would like to thanks my parents, Mr. Wijerathna and Mrs. Malani Amarasighe, my husband, Mr. Sirantha Athauda and my sister, Ms. Dhanushka Wijerathna. Their unconditional love, inspiration, support, and encouragement made me brave enough to accomplish this work and to pursue my aims.

Nadeesha Kumari Wijerathna

References

[1] N. Kobayashi, M. Matsubara, K. Tajima, and A. Morishima, "A crowd-in-the-loop approach for generating conference programs with microtasks," in: Proceedings of IEEE International Conference on Big Data 2017, pp. 4312-4314.

[2] T. Yan, V. Kumar, and D. Ganesan, "CrowdSearch: exploiting crowds for accurate real-time image search on mobile phones," in: Proceedings of MobiSys'10 2010, pp. 77-90.

[3] Z. Zhi, T. Kijewshi-Correa, D. Hachen, and G. Madey, "Haiti earthquake photo tagging: lessons on crowdsourcing in-depth image classification," in: Proceedings of Digital Information Management (ICDIM), 2012 Seventh International Conference on, pp. 357-364. IEEE, 2012.

[4] S. Ghosh, N. Sharma, F. Benevenuto, N. Ganguly, and K.P. Gummadi, "Cognos: crowdsourcing search for topic experts in microblogs," in: Proceedings of SIGIR'12, 2012

[5] K.E. Bevelander, K. Kaipainen, R. Swain, S. Dohel, J.C. Bongard, P.D.H. Hines, and B. Wansink, "Crowdsouring novel childhood predictors of adult obesity," in Joural of PLOS one 9(2), 2014, e87756.

[6] M. Swan, "Crowdsourced health research studies: an important emerging complement to clinical trials in the public health research ecosystem, in: Journal of Medical Internet Research 14 (2), 2012.

[7] A. Popoola, V. Naroditskiy, C. Castillo, and I. Rahwan, "Information verification during natural disasters," in: Proceedings of World Wide Web Conference, 2013.

[8] J. Deng, J. Krause, and L. Fei-Fei, "Fine-grained crowdsourcing for fine-grained recognition," in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2013.

[9] M.S. Bernstein, G. Little, R.C. Miller, B. Hartmann, M.S. Acherman, D.R. Karger, D. Crowell, and K. Panovich, "Soylent: a word processor with a crowd inside, in: Proceedings of ACM Symposium on User Interface Software and Technology, 2010, pp 313-322.

[10] J.K. Kaist, B. Tabibian, S.O. Kaist. B. Socholkopf, and M.Gomez-Rodriguez, "Leveraging the crowd to detect and reduce the spread of fake news and Misinformation," in: Proceedings of WSDW, 2018. pp. 324-332

[11] F. Daniel, P. Jucherbaev, B. Benatallah, and M. Allahbakhsh, "Quality control in crowsourcing: a survey of quality attributes, assessment techniques and assurance actions, in: ACM Computer Survey. 51(1), 2018, pp. 7:1-7:40.

[12] P. Venetis, and H. Garcia-Molina, “Quality control for comparison microtasks,” in: Proceedings of CrowdKDD, 2012.

[13] N. Shah, and D. Zhou, “No oops, you won‘t do it again: Mechanisms for self-correction in crowdsourcing.” in: Proceeding of the 33rd International Conference on Machine Learning, 2016.

[14] M. Kobayashi, H. Morita, M. Matsubara, N. Shimizu and A. Morishima, “An enpirical study on short-and long-term effects of self-correction in crowdsourced microtasks.” in: Proceedings of Conference on Human Computation and Crowdsourcing, HCOMP 2018.

[15] R.J. Sethi, “Crowdsourcing the verification of fake news and alternative facts.” in: Proceedings of HT‘17, 2017.

[16] Wikipedia Website. http://www.wikipedia.org, Retrieved June 4, 2019.

[17] Open Street Map Website. http://www.openstreetmap.org, Retrieved June 4, 2019.

[18] Stock Overflow Website. http://www.stackoverflow.com, Retrieved May 25, 2019.

[19] Quora Website. http://www.quora.com, Retrieved May 25, 2019.

[20] British Government Longitudinal Position Challenge. https://en.wikipedia.org/wiki/Longitude_$rewards$, Retrieved May 25, 2019.

[21] Toyoda Logo Contest. https://www.toyota-global.com, Retrieved May 27, 2019.

[22] National Opera House for Sydney Design Contest. https://www.sydneyoperahouse.com. Retrieved May 27, 2019.

[23] G. LeBaron, “audubon’s christmas bird count from 19th century conservation action to 21st century citizen science,” Citizen Science Toolkit Conference Report, USA, 2007.

[24] A. Kittur, J. V. Nickerson, M. Bernstein, E. Gerber, A. Shaw, J. Zimmerman, M. Lease, and J. Horton. “The future of crowd work.” in: Proceedings of the conference on Computer supported cooperative work, CSCW ’13, USA, 2013, pp. 1301-1318.

[25] A. Quinn and B. Bederson. “Human computation: a survey and taxonomy of a growing field.” in: Proceedings of the annual conference on Human factors in computing systems, CHI ’11, Canada, 2011, pp. 1403-1412.

[26] M. Raddick, G. Bracey, K. Carney, G. Gyuk, K. Borne, J. Wallin, and S. Jacoby. “Citizen science: Status and research directions for the coming decade.” ASTRO2010 Decadal Survey Position Paper, 2009.

[27] A. Quinn and B. Bederson. “A taxonomy of distributed human computation.” HumanComputer Interaction Lab Tech Report, 2009.

[28] French Food Preservation Contest. https://www.thevintagenews.com, Retrieved May 21, 2019.

[29] French Food Preservation Contest. https://www.history.com, Retrieved May 21, 2019.

[30] Netflix Programming Contest. https://en.wikipedia.org/wiki/Netflix$_{Prize}$, Retrieved May 25, 2019.

[31] N. Wijerathna, M. Matsubara, A. Morishima, "Finding evidences by crowdsourcing," in: Proceedings of IEEE International Conference on Big Data 2018, pp. 3559-3562.

[32] M. Raddick, G. Bracey, P. Gay, C. Lintott, P. Murray, K. Schawinski, A. Sza- lay, and J. Vandenberg. "Galaxy Zoo: Exploring the motivations of citizen science volunteers." Astronomy Education Review, 9(1), 2010.

[33] D. Estrin. "Participatory Sensing: Citizen Science, Scientific Citizens, Computational Thinking." in NSF Broader Impacts for Research and Discovery Summit, USA, 2010.

[34] C. B. Cooper, J. Dickinson, T. Phillips, and R. Bonney. "Citizen science as a conservation in residential ecosystems." Ecology and Society, 12(2), 2007.

[35] The First Christmas Bird Count: December 25, 1900 About 18,500 individual birds and 27 total participants Cumulative bird species list: 89 species total, National Audubon Society, 2010. PDF: https://www.audubon.org. Retrieved May 25, 2019.

[36] e Bird Web Portal. https://ebird.org/home. Retrieved June 10, 2019.

[37] Netflix Competition. https://www.ejumpcut.org. Retrieved May 25, 2019.

[38] J. Howe, "The Rise of Crowdsourcing." Wired Magazine, 2006, 14(6): pp1-5, https://www.wired.com. Retrieved May 27, 2019

[39] J. Howe, "Crowdsourcing: A Definition," in Howe's website. https://crowdsourcing.typepad.com. Retrieved May 27, 2019.

[40] H. Simula. "The rise and fall of crowdsourcing?" in: 46th Hawaii International Conference on System Sciences, USA, 2013, pp. 2783-2791.

[41] D. C. Brabham. "Crowdsourcing as a model for problem solving: An introduction and cases." Convergence, 14(1), 2008. Pp. 75-90.

[42] E. Enrique, G. Fernando, "Towards an Integrated Crowdsourcing Definition," Journal of Information Science, 38(2), 2012, pp. 189-200.

[43] J. Surowiecki. "Wisdom of crowds," Doubleday; Anchor, USA, 2004, ISBN:0385721706.

[44] K. Heimerl, B. Gawalt, K. Chen, T. Parikh, and B. Hartmann, "CommunitySourcing: Engaging local crowds to perform expert work via physical kiosks," in: Proceedings of the IGCHI Conference on Human Factors in Computing Systems (CHI '12), USA, 2012, pp.1539-1548.

[45] 99Designs website. https://en.99designs.jp. Retrieved June 14, 2019.

[46] Utest website. https://www.utest.com. Retrieved June 14, 2019.

[47] Mob4Hire on Wikipedia. https://en.wikipedia.org/wiki/Mob4Hire. Retrieved June 14, 2019.

[48] reCAPTCHA website. https://www.google.com/recaptcha. Retrieved June 14, 2019.

[49] Logo Design Crowdsourcing website. https://www.designcrowd.jp. Retrieved June 14, 2019.

[50] K. Natascha, and F. Karen, “Plz RT: A Social Diffusion Model of Misinformation and Disinformation for Understanding Human Information Behaviour,” in: Proceedings of the ISIC, Japan, 2012, pp.1-17.

[51] S. Maheshwari, “How fake news goes viral: case study,” The New York Times, 2016. https://www.nytimes.com. Retrieved May 29, 2019.

[52] H. Allcott and M. Gentzkow, “Social media and fake news in the 2016 election,” Journal of Economic Perspectives 31(2), 2017, pp.211-36.

[53] Amazon Mechanical Turk Guide Book. 2017, PDF: https://docs.aws.amazon.com. Retrieved May 11, 2019.

[54] D. Retelny, M. Bernstein, and M. Valentine, “No Workflow Can Ever Be Enough: How Crowdsourcing Workflows Constrain Complex Work,” Proc. ACM Humman computation. Interact. 1, 2, Article 89, 2017, p.23.

[55] Mt. Tsukuba Photography, Katsuyoshi Nakahara Photography, https://www.katsuyoshinakahara.com//spring/ Retrieved April 22, 2018.

[56] M. Allahbakhsh, B. Benatallah, A. Ignjatovic, H. Nezhad, E. Bertino and S. Dustdar, “Quality control in crowdsourcing systems: issues and directions,”in: IEEE Internet Computing, 2013, pp. 76-81.

[57] W. Lee, C. Chang, P. Yang, C. Huang, M. Wu, C. Hsieh, and K. Chuang, “Effective quality assurance for data lables through crowdsourcing and domain expert collaboration,” in: Open Proceedings, 2018, pp. 646-649.

[58] L. Graves, “Understanding the promise and limits of automated fact-checking,” Factsheet 2018, https://reutersinstitute.politics.ox.ac.uk.

[59] A.T. Nguyen, A. Kharosekar, S. Krishnan, Siddhesh K. E.Tate, B.C. Wallace, and M. Lease, “Believe it or not: designing a human-AI partnership for mixed-initiative fact-checking,” in: Proceeding of UIST, 2018.

[60] E. Enrique, and G. Fernando, “Towords an integrated crowdsouring definitio,” Journal of Information Science, 2012, pp 1-14.

[i] https://mediabiasfactcheck.com
[ii] https://www.politifact.com/
[iii] https://www.snopes.com
[iv] https://www.factcheck.org
[v] http://gohoo.org
[vi] https://factchecker.in
[vii] https://africacheck.org
[viii] https://www.altnews.in
[ix] https://gomaneh.net
[x] https://fullfact.org